Laboratory of Bioinformatics 1

Saul Pierotti

November 30, 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 is the core of this course
- 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
- In Hamburg there is an EMBL faicility that uses X-rays in a flow citometer

X-ray 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-electron 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 and other DBs and tools

- 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
- A database can be defined by its statistics

- Data in a DB can be distributed in categories that are relevant for the interpretation of data
- The space group refers to the simmetries of the unit cell
- The Ramachadran plot of a structure can be generated with Procheck (EMBL)
  - It is much more informative than the 3d view generated with Rasmol
- The PDBsum is a pictorial database that provides an at-a-glance overview of the contents of each 3D structure deposited in the Protein Data Bank
  - It is offered by EMBL
  - It shows also the Procheck output for the structure
  - It shows the biological unit instead of the unit cell

PDB101

- A crystal is composed of the unit cell, that is translationally repeated in the crystal
- The unit cell is composed of asymmetric units, that rotated and translated form the unit cell
- The asymmetric unit is the unique part of the crystal structure
- The biological assembly is the biologically relevant form
- Occupancy of an atom is the fraction of times that atom is in the specified position in the crystal
  - The occupancies for an atom always sum to 1, giving the possible alternate conformations
- The R-value is the fit between the theoretical diffraction pattern of the model and the experimental one
  - 0 is a perfect fit, 0.63 is a the fit of a random diffraction pattern
- R-free is another statistic that avoids the bias introduced in the refinement step

Nuclear magnetic resonance

- It does not require the crystal, it is performed in an homogeneous really concentrated protein solution
- The frequencies used are on the radiowave lenght, so really low energies
- NMR measures contacts among atoms in solution, it measure frequence changes when nuclei come close to each other
- The data produced is a contact map
- The only nuclei considered are mainly O, H, C
- From NMR I get many conformation for each molecule, and if I visualize it in a molecular visualization software I see many superimposed structures
  - The core is usually stable and in agreement with X-ray data, while regions with high temperature factor have many configurations
  - From the many structures I can recover a consensus structure

Time resolved X-ray crystallography

- I use pulsed X-ray instead of a continuous beam
- I can observe the conformational changes of the protein between different pulses

Protein structure allignment

- We wan to compare the structure of 2 proteins
- Hydrogens bond are mainly located at the level of the backbone
- I keep 1 protein fix as a template, and I try to superimpose the other backbone
  - There are many different algorithms to do this
- The first reasonable step would be to compare the position of the backbone of the 2 proteins
- Sequence similarity is meaningful only with an underlying structural similarity
- The reduce representation of a protein contains only the $C - C\alpha - N$ elements of the backbone
- The simplest measure of structural similarity is the root mean squared deviation (RMSD) among atomic coordinates

- One of the algorthm for strucutural alligment used in the PDB is called java combinatorial extension (JCE)
- JCE was written by Philippe Bourne, the director of the PDB
  - It is one of the best-performing algorithms
- The lenght of an allignment is the lenght of the protein sequence, plus the gap introduced
- Sequence identity is a score between 0 and 1 that gives the number of corresponding residues after the allignment
- Sequence similarity is a score between 0 and 1 that gives the number of similar residues after the allignment
- Structure is conserved more than sequence (!)

Protein families

- Multiple structural allignment allow to define protein families
  - We know around 14000 protein families
- They were discovered by M. Dayhoff
- A protein family is the set of proteins that perform the same function in different organism
- Proteins in the same family can also be really different in sequence, but their structure is really similar
- We cannot detect sequence similarity when under 30%, but we can detect structure similarity in those cases
  - Under 30% the result is not statistically significant
- A superfamily is a set of protein families with different foldings that can perform the same function

Domains

- A protein domain coincides with the folded protein for small globular proteins (150 aa)
- When the PDB grew, we realised that multi-domain proteins share domains with small globular proteins

Sequence allignment

- It is our only way to compare proteins for which I do not have structures
- Sequence comparison can be pairwise or database search
- Database search is an extension of pairwise sequence alignment
- Sequence allignment can be local or global
- A global allignment optimizes pairing over the whole sequences by introducing gaps
  - A global allignment has a lenght that is at least as long as the longest sequence
- A local alignment stops the allignment if continuing it makes it worse
- A metric is a set of rules that allow us to define the distance between strings
- The Hamming distance is, for a pair of sequences equal in lenght, the number of mismatching positions
  - It is used for ungapped allignments
- The Levensthein or edit distance of 2 strings is the minimal number of edits necessary to change 1 string into the other
  - It is suitable for gapped allignments
- A scoring scheme is a measure of sequence similarity
  - It is a substitution matrix
- Sequence alligment algorthms seek to maximize a scoring function or minimize a dissimilarity measure
- For aminoacids, we have the PAM and BLOSUM matrices, and matrices derived from structure allignment
- 1PAM (Dayhoff) is 1% of accepted mutation
- PAM corrects for the relative abundance of aminoacids
- PAM250 is constructed with sequences of 20% of sequence identity
- Odds are ratios of probabilities

- The log-odds of a mutation can be used as a score
  - It is the log of the probability of the substitution normalized for the frequencies of the residues
  - It is often multiplied by 10 because it is a really small value
- The PAM is a symmetric matrix, therefore sometimes only half is shown
- Conservation is always positively score, but with different scores depending on aa abundance
- The BLOSUM are a family of matrices that also use the log-odds for the substitutions
  - They were produced in the 1990, where there where many more sequences available
  - Lower matrices (es. BLOSUM40) are more permissive
- Dynamic programming optimizes the solution of subproblems in order to find a global solution
  - It gives the correct solution, but it is computationally expensive
- For database searches, we use methods based on words (K-tuples), also called heuristic
  - Heuristic means approximate, it does not give an optimal solution
- Gaps are introduced in order to optimize the number of identities
- Dynamic programming approaches are Needelman-Wunsch (global) and Smith-Waterman (local)
- BLAST is an heuristic local allignment method used for database search
- FASTA is another heuristic algorthm but is no longer used
- An heuristic method is optimized for the expected result, therefore it does not have any intrinsic validity
- In BLAST I chop the query in K-tuples and search for matching K-tuples in the database
  - The K stands for the number of fragments in which I chopped my query
  - I can search for exact matches or allow for gaps
  - Then it tries to extend the match until its score is lower than a threshold
  - Parameters that can be chosed are scoring matrices and gap penalty
- We are doing multiple alignments to build a sequence profile
- A sequence profile is a matrix with residues in the y axis and the position in the alignment in the x axis
  - It is a compressed way to describe a protein family
- A protein family is a set of proteins charachterised by structural superimposition
- Affinity comes from electron densities
- Protein families are important because they allow us to cluster PDB data
  - They are constructed starting by comparing proteins with the same function
  - It can be then computationally describe with structural allignment
  - A protein family is described with an HMM (hidden markow model)
- A heuristic method is not based on theory, while QED is firmly based on theoretical ground
- Whatever is heuristic is at the core data-driven
- A sequence alligment mixes algorthm and substitution matrices to give a result
- Sequence allignment methods are less stable than structural ones, more sensitive to lenght of the sequences and other variables
- Hidden Markow Models are also called Pfam domains
- When alligning structures, it is better to use structures taken with the same method
- The score of an alignment is the sum over its lenght of the score for each match
  - It uses a score substitution matrix to determine the score of each match
- The bit-score is the Log scaled version of the score
  - It is used by BLAST and it uses a formula that is a bit complex
    - $S' = \frac{\lambda S - \ln(K)}{\ln(2)}$
  - The bit-score S' depends on the parameters λ and K
  - The 2 parameters depend on the substitution matrix and on the gap penalty
  - It is independent on the size of the search space (dimension of the database)
- The E-value is a correction of the p-value for multiple testing
  - It is the expected number of matches of that score that I expect in a random database
  - It depends on K, λ and the size of the database

BLAST (original paper)

- Originally described by Altschull in 1990 in J. Mol. Bio.
- It is 1 order of magnitude faster than other heuristic methods

- Global alligment methods (NW) optimize the allignment over the whole sequence, and can include low-similarity regions
- Local allignment methods (SW) can yeld multiple allignments from a single comparison
  - Low similarity regions do not affect the allignment score
  - Local allignment is preferred in DB searches
- Many similarities methods start with a substitution matrix
  - It used PAM120 for proteins and a -5 identity -4 mismatch matrix for DNA
- A sequence allignment is a continuous stretch of residues of any lenght
- The similarity score of an allignment is the sum of similarity values for each pair of alligned residues
- A maximal segment pair (MSP) is defined as the highest scoring pair of identical segments chosen from 2 sequences
  - It can be of any lenght, so to maximize the score
  - It provides a measure of local similarity
- Since in biology we care for all conserved regions, not only the best scoring one
  - Therefore a segment pair id defined as locally MSP if its score cannot be improved by extending or shortening both segments
- BLAST seeks all local MSP that score above a cutoff
- The greatest advantage of MSP is that we have the matematical tools to determine its statistical significance
- I want to retrive from a database all the sequences with MSP score above a cutoff S
  - Sequences that score far above the cutoff are almost definitely biologically relevant, while borderline matches can be evalued considering the biological context
- BLAST speeds up DB search by avoiding to spend time in sequences that are unlikely to give high MSP scores
  - Given a fixed word lenght w, BLAST seeks only segment pairs with a word of score at least T
  - When a matche is found, BLAST tries to extend the segment to see if it reaches the desired final cutoff score S
  - The lower the value of T, the more probable that a segment of score >S will contain a word with score >T
  - However, the lower the value of T the higher the number of hits, and therefore the execution time
  - Random simulations allowed to determine an optimal T value for various conditions
- The algorithm first makes a list of words that score >T when compared with some word of the query
  - The time of list generation is linearly proportional to the lenght of the query
- During the extension phase, if the score falls below a certain treshold lower than the original MSP, it is discarded
  - It loses in accuracy, but in a negligible manner
- Theoretical results on the distribution of MSP scores of random sequences allow the following determiantion
  - Given a set of probabilities for the occurrence of each residue and a scoring matrix
  - The theory gives the parameters λ and K for evaluating the statistical significance of MSP scores
  - With 2 random sequences of lenght m and n, the probability of finding an MSP with score equal or better than S is $1 - e^{-y}$, with $y = K * m * n * e^{-\lambda S}$