Laboratory of Bioinformatics 1

Saul Pierotti

December 1, 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

- Hydrogens bond are mainly located at the level of the backbone
- 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-rays are used because they have a wavelenght comparable to an atomic bond, and therefore they can resolve with that level of detail
- X-ray sources used for studying protein structure are rotating anode tubes or synchrotrons
- In the rotating anode tube electrons are accelerated towards the anode in a vacuum

  - The anode rotates, so to expose a different portion to the electron beam at each moment and avoid damage
  - X-rays are releades when the electrons collide with the anode, as a way to get rid of their kinetic energy
  - Only 1% of the kinetic energy is converted to photons, the rest heats up the anode
- Sychrothron source give an anisotrope beam, which gives semicircles in the diffraction map instead of focused spot
  - They are located in the US, Europe or Japan
  - X-rays are produced by charged particles moving in a magnetic field
- 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
- It allowed to obtain the structure of entire ribosomes
- It cannot be used with small proteins that cannot whitstand the electron bombardment
- It is possible to capture the protein in different conformations, and so understand how it works
  - It was used by Rubinstein to obtain a movie of V-ATPase (!)
- It is really costly, around 6000€/day

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

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 17Rasmol
- Some PDB statistics
  - 158180 macromolecular structures
  - 76380 enzymes
  - 48974 distinc protein sequences
  - Resolution range from < 1Å to > 4.6Å, with a peak around 2Å
    - Distribution not normal with a long right tail
  - 1702 source organinsms

PDBsum

- It 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 hosted by EMBL-EBI
- It shows also the Procheck/Procheck NMR Ramachadran plot 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

Ramachadran plot

- α carbons in proteins are 3.8Å apart
- The ϕ angle is the dihedral angle between N-Cα, ψ is between Cα-COOH
- The Ramachadran plot graphs the ϕ angle of a residue against its ψ angle
- Some regions of the plot are really common and allowed, some are not because of steric hindrance
- The Ramachadran plot of a protein is a scatterplot of its dihedral angles superimposed on a color code for the allowed conformational spaces
- Procheck and Procheck NMR calculate Ramachadran plots from PDB files

Protein structural allignment

- Rigid superimposition requires the knowledge of at least 3 non-allineated equivalent residues, while structural allignment requires no previous knowledge of equivalent positions
- The output of a structural allignment is a set of superimposed 3D coordinates, one for each input structure
- A structural allignment implies a corresponding sequence allignment, from which we can calculate sequence identity and similarity
    - Sequence similarity is meaningful only with an underlying structural similarity
    - 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
    - Residues are considered similar if they belong to the same chemical class (polar, non polar, cationic, anionic,. . .)
    - Structure is conserved more than sequence (!)
- Generally, I keep 1 protein fix as a template, and I try to superimpose the other backbone onto it, allowing the introduction of gaps
- There are many different algorithms to do structural allignment
- The reduce representation of a protein contains only the $C - C\alpha - N$ elements of the backbone
    - Structural allignemnt usually only considers the position of the backbone, so it works on the reduced representation of the protein
- After the allignemnt, it is possible to derive various measures of strucutral similarity
    - The simplest metric is the root mean squared deviation (RMSD) among atomic coordinates
    - The raw score can be normalized by subtracting the mean and dividing by the standard deviation, so to get the z-score
- One of the most famous structural allignment alogorithms is jCE (Java Combinatorial Extension), written by Philippe Bourne, the director of the PDB

  - It is one of the best-performing algorithms
  - It breaks down the proteins in fragments, and it tries to allign the structure of these by several methods (RMSD, secondary strucutre, . . .)
  - It forms a series of alligned fragment pairs (AFPs) and filters them, retaining only those that respect a given measure of local similarity
  - It generates an optimal path among AFPs, that yelds the final allignment
  - The first AFP that nucleates the allignment can occur at any position
  - The size of AFP and the maxium allowed gap are parameters, usually set to heuristic optimal values
  - An important drawback is that It does not deal well with flexible regions that can have different conformations, since it is based on rigid superimpoisitio
- FATCAT is another algorithm that deals better with flexible regions, but can also give spuorios allignments among unrelated regions
- Many algorthms cannot recognize structural similarities that are not sequence order dependent
- Triangle Match deals with sequence order independent relationships
- The lenght of an allignment is the lenght of the protein sequence, plus the gaps introduced
- In the PDB, all possible pairwise structural allignments are pre-calculated and stored in xml files
  - The database is updated weekly

Protein structural classification

- A protein family is the set of proteins that perform the same function in different organism, and therefore share a similar structure
- Protein families were discovered by M. Dayhoff
- Multiple structural allignment allows to define protein families
- We know around 14000 protein families
- 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 of a sequence allignment is not statistically significant
- Pfam categorizes all the entries in the PDB in protein families, clustering for strucutral similarity
  - The Pfam database was built by performing pairwise comparison of all the PDB entries
  - A protein family is described by a Hidden Markow Model (HMM)
- A superfamily is a set of protein families with different foldings that can perform the same function
- 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
- SCOP categorizes proteins in superfamilies, Pfam families and fold
- The SCOP fold can be all alpha, all beta, alpha+beta, alpha/beta, small proteins
  - Alpha+beta has distinct alpha and beta regions
  - Alpha/beta has mixed alpha-beta structures
- Proteins in the same family have clear common evolutionary origin, and usually have >30% sequence identity
- Proteins in the same superfamily have low sequence identity, but common structural and functional features suggest evolutionary relationships
- Proteins are said to have the same fold if they have the same secodary structres in the same arrangement and with the same topology
  - If 2 proteins have the same fold they do not need to be evolutionary related: it can be a case of converging evolution

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 its score lower
  - From a pairwise comparison, I can get many local allignments
- 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
  - An edit operation is defined as delition, insertion or alteration of a single charachter
- A scoring scheme is a measure of sequence similarity
  - It is a substitution matrix where each possible substitution has a score
  - The matrix is symmetric, so it is often reported only half of it
- Sequence alligment algorthms seek to maximize a scoring function or minimize a dissimilarity measure
- For nucleic acids, there are substitution matrices that only consider match vs mismatch, and matrices that give different scores to transitions and transversions
- For aminoacids, we have the PAM and BLOSUM matrices, and matrices derived from structure allignment
- The PAM matrices were developped by M. Dayhoff and are based on the observed frequencies of mutation of 1 aa into another in alligned proteins of the same family
  - 1 PAM is 1% accept mutation, so 2 sequences 1 PAM apart have 99% sequence identity
  - The matrices were built using closely related sequences 1 PAM apart, so that multiple substitutions were unlikely
- The PAM1 matrix was built by collecting statistics on substitution frequencies in pairwise comparison of sequences 1 PAM apart and correcting for relative aminoacid abundance
  - The score of the mutation i->j is the log-odd of the mutation
  - $S = \log \frac{p(i,j)}{p(i)*p(j)}$
  - p(i,j) is the observed i->j mutation rate while p(i) and p(j) are the relative aminoacid abundances
  - Note that p(i)*p(j) is the expected mutation rate if all mutations are equally likely, it is a correction factor for aminoacid frequencies
  - Since the score is a really small number, it is usually multiplied by 10
- Other PAM matrices are built as powers of PAM1
- PAM250 is used for comparing sequences with 20% 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 $y = K * m * n * e^{-\lambda S}$
  - In a similar way, we can calculating the probability of having c MSPs with score greater than S
  - This result is the p-value of the MSP score