Laboratory of Bioinformatics 1

Saul Pierotti

December 29, 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 09/01/20 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

Things to have in mind

- H bonds are 3 Å long, they require planarity and their E = -4 kcal/mol
- Salt bridge -5 kcal/mol
- Lennard-Jones interactions have E = -1 kcal/mol
- SS bonds
- Covalent bonds 2-3 Å, -100 kcal/mol

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
- We will focus on proteins because we have many structures available, and we can establish a clear relationship among protein structure and function
- 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
- Functional annotation is not prediction: we use data to infer properties
- Functional genomics is the process of how the genome gives rise to the proteome and metabolome, which in turn influence gene expression
- 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

- A relational database is organized in tables in relation with each other
- A file is a container of data
- A file format is the specific organization of a collection of data
- In Hamburg there is an EMBL faicility that uses X-rays in a flow citometer
- A protein is stable because of H-bonding in the backbone

X-ray cristallography

- A molecule is carachterized by its electron density
- Affinity among molecules comes from the electron densities
- A PDB file is not the protein structure, but a way to represent it
- A protein is
  - A polymer formed of amminoacil residues joined together by peptidic bonds
  - A frustrated system capable of auto-organization in the solvent
  - A social entity that interacts with other molecules
- 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 refer to the reciprocal space of the Fourier transform
  - 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
- Errors in the electron density mainly derive from errors in the phases
- The resolution of a diffraction map depends on how ordered the crystal is
  - It is typically 0.5-5 Å
- 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
  - This is the R-value (!)
- In order to facilitate crystallization, protein can be stabilized by crosslinking
  - A crosslinker is a molecule that forms chemical bonds with the protein and stabilizes it

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 file

- 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
- An heteroatom is any atom in a PDB file that does not belong to the primary sequenc eof the protein
- 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 (B size)
- Occupancy is how well the atom fits the electron density
  - It is usually 1
  - If it is less than 1, there are more records for the same atom with occupancies that add up to 1
  - The same atom represented more than once has the residue name changed as ARES, BRES, CRES, ...
- The temperature factor refers to the mobility of the position
  - Tends to be higher at the surface and lower in the core
- 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 Ramachandran plot of a structure can be generated with Procheck (EMBL)
  - It is much more informative than the 3d view generated with Rasmol
- 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
- In protein structure we do not have signal peptides because they are cleaved post-translationally
- For every PDB entry, there is a validation report with all the statistics
  - This is summarized by percentile ranks on the PDB webpage
- Atom coordinates have 5 digits, xx.xxx
  - The significance of the digits depends on the resolution of the structure
- mmCIF is a different file format that is used for representing protein structures

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 Ramachandran 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

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
    - It should be below 3 Å
- The raw score (a dimensionless metric) 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 structure, ...)
  - 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 superimpoisition
- 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
- When alligning structures, it is better to use structures taken with the same method

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 classification is generally conservative: when in doubt we make a new division
- Protein families were discovered by M. Dayhoff
- Multiple structural allignment allows to define protein families
- We know around 14000 protein families
- Multiple allignments are the result of repeated pairwise allignments
- We are doing multiple alignments in order 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 consensus sequence
- A protein family is a set of proteins charachterised by structural superimposition
- Protein families are important because they allow us to cluster PDB data
  - Proteins in the same family typically have >30% identity, but there are exceptions (globins)
  - They are constructed starting by comparing proteins with the same function
  - It can be then computationally described with structural allignment
  - A protein family is described with an HMM (hidden markow model)
  - Hidden Markow Models are also called Pfam domains
- 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 possibly different foldings that can perform the same function
  - The functional similarity suggest common origin, but this is not certain
- 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
- A protein domain is a protion of the sequence that harbours a function
  - It is also defined as indipendently folding, but are we sure they did experiments about it?
- 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
- Margaret Dayhoff studied cytochrome c and determined that it is similar in many organisms
  - She was the first to relate structure and function
- Protein in the same family with less than 30% homology are called distantly related homologs

Sequence allignment

- It is our only way to compare proteins for which I do not have structures
- I have around 180,000,000 proteins in UniProtKB
  - 99% are only predicted

  - 561,568 in Swiss-Prot
  - 179,250,561 in TrEMBL
- There are at least 3 orders of magnitude among the number of strucutres and sequences in databases
  - Actually more because PDB is redundant and UniProt no
- A gene is a transcribed locus
- A sequence allignment is a continuous stretch of residues of any lenght
- Sequence comparison can be pairwise or database search
- Database search is an extension of pairwise sequence alignment
- Sequence allignment can be local or global
- Multiple alignments are based on many pairwise alignments
- 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
- 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
  - They are based on ungapped multiple alignments in short regions of related sequences
  - The different matrices were built using allignments with different thresholds of sequence identity
  - Lower matrices are more permissive since are built with sequences that have less than a treshold of sequence identity
- A dotplot is a plot that gives an overview of the similarity between 2 sequences
  - It is also based on scoring schemes
  - Dotlet is a Java tool for dotplot analisys

  - It is useful for finding repeated portions and for finding intron-exon boundaries
- Dynamic programming optimizes the solution of subproblems in order to find a global solution
  - It gives the correct solution provided that all the subproblems are independent, but it is computationally expensive
- Dynamic programming approaches are Needelman-Wunsch (global) and Smith-Waterman (local)
- 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
- For database searches, we use methods based on words (K-tuples), also called heuristic
  - Heuristic means approximate, it does not give an optimal solution
  - It also means empirical, not based on theory
- BLAST is an heuristic local allignment method used for database search
  - Originally described by Altschull in 1990 in J. Mol. Bio.
  - It is 1 order of magnitude faster than other heuristic methods
- 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
- An expected result is the result of experimental approaches, which is well accepted in the scientific community
  - It is high quality data
- 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
- In BLAST I chop the query in K-tuples and make a list of words, that I use to scan the DB
  - Word and k-tuples are the same thing
  - A k-tuple is an ordered set of k values
  - At this phase BLAST searches for exact matches of words in the list with DB entries
  - Any of these local alignments can form a maximal segment pair
  - 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
  - In biology we care for all conserved regions, not only the best scoring one
    - To take care of this, a segment pair is defined as locally MSP if its score cannot be improved by extending or shortening both segments
  - BLAST filters for all local MSP that score above a cutoff
  - I want to retrive from a database all the sequences with MSP score above a cutoff T
  - The greatest advantage of MSP is that we have the matematical tools to determine its statistical significance
- In BLAST, sequences that score far above the cutoff are almost definitely biologically relevant, while borderline matches can be evalued considering the biological context
- The behaviour of BLAST can be tweaked with some parameters
  - I can search for exact matches or allow for gaps
  - I can choose scoring matrice and gap penalty
- 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
- BLAST then tries to extend the MSPs in both directions
- During the extension phase, if the score falls below a certain treshold below the score of 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
- A sequence allignment method uses its algorithm and substitution matrices to give a result that maximizes the score of the allignment
- Sequence allignment methods are less stable than structural ones, more sensitive to lenght of the sequences and other variables
- The raw score of a sequence 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
  - It can be demonstrated that raw score follows an extreme value distribution
- The bit-score is the Log scaled version of the raw score
  - It is measured in bit, and it is a metric for the search space
  - Each unitary increse in bit score doubles the search space
  - A bit score of 30 means that we expect that score to be observed once in 2^30 comparisons
  - 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, and on the size of query and database
  - It is independent on the size of the search space (dimension of the database), beacuse it corrects for it
- 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
- MegaBLAST is an implementation of BLAST optimized for very long and very similar sequences, such as those differing only for sequencing errors
  - It uses a greedy algorithm
- PsiBLAST (position specific iterated)
  - It takes a single protein sequence as input, and compares it to a protein DB with a normal BLAST search
  - Given a threshold, it builts a multiple allignment with all the local allignments above the threshold
  - From the multiple allignment, a profile is built for any local alignment using the query as a base
  - The profile has the same lenght as the query
  - The profile is used again for DB searches using a slight modification of the BLAST algorithm
  - The statistical theory developped for BLAST is also valid in profile searches
  - The algorithm then iterates the process by building another profile from the allignment of the new hits
  - The process is repeated a fixed number of times, or until convergence

Distantly related homologs (30% sequence identity)

- Burkhard Rost is a professor in Munich who first published a graph showing the confidence in sequence allignment as sequence identity against number of residues aligned

  - In this graph, it marks the region where we are confident to have evolutionary relationships
  - The line is the best fit deriving from the data points of structural allignments of all the PDB structures
  - The region below the best fit is where I cannot be confident that the allignment reflects a structural relationship
  - The horizontal asymptote is around 30% sequence identity +- something, so irrespective of sequence lenght, under 30% identity we cannot imply strucutural similarity
  - read paper!
- Also on a statistical standpoint, the allignment is not significant under 30% indentity
- Sequence allignment methods are reliable and will give a similar result to structural allignment only when sequence identity is above 30%
- Distantly related homologs are proteins that have the same folding and perform the same function, but have a really different sequence
  - We cannot do sequence allignment with them (!)
  - We can use the multiple sequence allignment (derived from structure) and the HMM of a protein family to model them (!)
- We can have proteins that have the same domains but shuffled in a different order
  - In this case structural allignment is problematic

Ramachandran 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 Ramachandran 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 Ramachandran 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 Ramachandran plots from PDB files
- A Ramachandran plot is a bidimensional map of a protein structure where the torsion angles of the backbone are reported
  - Don't say residues, they are in the backbone
- The expected values are determined by measuring torsion angles from a set of well charachterized proteins
- The main regions are alpha (A), beta (B), 3-10 helices and left-handed helices (l), and proline region
  - In the top left quadrant (negative ϕ, positive ψ) we have beta-strands
    - Note that beta-strands can be also partially allowed in the extreme bottom left because the angles are circular (180 = -180 !)
    - Same thing in the extrem top right and extreme bottom right
  - Alpha-helices are on the left, vertically centered but more towards the bottom quadrant (negative ϕ, negative to slightly positive ψ)
  - In the top right quadrant we have left-handed 3-10 helices (slightly positive ϕ and ψ)
  - Proline is special because of its cis peptide bond and has a specific area on the rigth, at the extreme bottom (very negative ψ, positive ϕ)
  - Glycine has really low steric hindrance and can be practically everywhere in the plot
- A good model has at lest 90% of the residues in the most allowed regions
  - This is based on the analysis of 118 structures of at least 2 Å resolution and R-factor less than 20%
- The G factor, for the different angles, measures how unusual a structure is
  - It is a log-odds based on the observed distribution of stereochemical properties

Protein structural prediction

- The goodness of a protein strucuture can be determined by comparison with a set of optimal conditions, determined by analysis of the PDB database

- The strongest interaction in proteins are H bonds
  - The electronegative atoms that partecipate in H bonds are O and N in proteins
  - The main source of H bonding interactions is the backbone
  - The length of an H bond is around 2Å between the 2 electronegative atoms
  - The bond can happen only if the atoms are in plane
  - The strenght of an H bond is around 10 Kcal/mol
- Charge-charge particles are really dependent on the environment
- Lennard Jones interactions have an energy around 1 Kcal/mol
  - They are described by the 6-12 potential, since they have a repulsive term (R^12) and an attractive term (R^6)
- SS bonds have energy of 30-40 Kcal/mol and a lenght of 2Å
  - Their presence depends on the redox ambient potential
- A protein family is charachterized by different sets of GEO terms
- Given an unknown sequence, I can allign it to different protein families and eventually translate the GEO terms of the family to it
  - I cannot go below a treshold of 30% of sequence identity with the family template
- It is one of the main shortcut used for predicting the folding of a protein sequence
- If in my model I have even a small difference in the active site from that of the family, the GEO terms cannot be traslated
- If we don't have the structure of the protein and we cannot assign it to a protein family, we don't know anythong about it
- If we do have the structure, we can try to compute the function by programs of theoretical chemistry
- Pfam is a DB of protein families that should be used for those sequences that don't find a template in the PDB with sequence allignment
- Protein folding usually starts with seeds of alpha-helices that then elongate and promote the final folding
- We have 3 methods for computing protein structure
  - Building by homology
  - Threading
  - Ab initio
- Which method is better to use depends on the availability of strucutures that have a sequence identity over a threshold with my sequence
  - Above 30% I can use the concept of protein family (building by homology)
  - Below 30% I need to use threading, fold recognition, machine learning
  - If I have a new folding, I need ab initio or machine learning
- If I get a model, I can understand its validity by checking if it would be stable
- Comparative modelling is a procedure that, starting from a sequence
  - Selects a template by alligning (BLAST) against all the PDB structures
  - Once found a good template, I can use NW (global allignment) to improve my allignment with the target
  - I then use modeller for modelling my sequence on the template
  - The output of modeller is a PDB file containing the model
  - I check with Procheck the plausibility of my model
  - The goodness of my model depends on the goodness of the initial allignment
  - If I am not satisfied I allign again and model again, until I am satisfied
- Threading procedures are based on the modelling of my sequence on different folds
  - In the PDB we have folds, that are self-stable and can be astracted from the protein
  - The main difference among different folds is the secondary structure, hence the topology of my model
    - In essence, the difference is in the pattern of H bonds
  - I have a scoring function that considers the stability of my model (H bonds, other interactions)
    - I get a likelyhood value for every model
  - I select the best model among the computed ones
  - Different models can better cover different portions of my sequence

- Ab initio minimizes the energy of the system by computing all the pairwise interactions
  - It is feasible only on small sequences
- There is a paper written by the author of the Rosetta method (Sanchez, 2000) that analised the validity of the different methods
  - If my starting sequence identity is above 60% my model is good for docking and it is comparable to a low quality experimental structure
  - In the 60-30% range we have a rough idea of the organisation of the backbone
  - Under 30% we can be lucky, or completely wrong
- The coverage of a local allignment is its lenght compared to the sequences
- How to select the best template
  - Highest sequence identity
  - Highest coverage, at least 70%
  - Highest template resolution
- Modeller was written by Sali, a PhD student, who became really famous
  - It is written in Fortran
  - It transplants the coordinates of the template to the target and it seeks for protein stability
  - It checks if the pairwise interactions are conserved (spatial restraints)
- Once we get the model, if it does not make sense we can try to tweak the initial sequence allignment
- If the model respects the Ramachandran plot, we can allign it to the template structure
- When trasferring functional annotation, we have to be careful of the meaning of what we are doing
  - The GO Cellular component can be different even if the structure is conserved
- Swiss-model runs modeller in remote, and it has many pre-computed models
- We need to be carefull with the pre-calculated models because they do not have any quality check (!)
- Modeller is a software that models the 3D structure of proteins by satisfaction of spatial restraints
- The input for modeller are the set of spatial restraints on the structure of the protein and of the ligand to be modelled
- The output is a structure for the target that satisfies the restraints as well as possible
  - Restraints are distances, angles, pairs of angles, ecc.
  - The restraints are automatically derived from the allignment with the template
  - A restraint is defined in term of a probability density function
- In modeller I can ask as many models as I want, and they are scored from best to worse

Modeller (from manual and other source)

- Modeller is a computer program that models three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints
- The input to the program are the restraints on the spatial structure of the aminoacid sequence to be modeled
- The output is a 3d structure that satisfies the restraint as well as possible
- Restraints can be related structures, NMR experimental data, rules of secondary structure packing, other experiments
- The model is computed by optimization of a the modeller objective function called molpdf
- The sequence alignment of target and template must be given as input
- Given an alignment, the restraints on distances, angles and other features are automatically derived from the staistical analisys of the relationships between many pairs of homologous structures
  - This was based on 105 families that included 416 proteins of known structure
- From the statistical distribution of a restraint (e.g. distance of related C-C bonds) I can derive the pdf for that specifc property
- Spatial restraints and CHARM energy terms are combined into an objective function
- The objective function is minimized in a Cartesian space via conjugate gradients, molecular dynamics and simulated annealing
- Different models can be obtained by varying the initial structure
  - This is done in a randomized way

- Variability among models can be interpreted as error in a specific region of the fold
- Some regions use a specialized modeling protocol (e.g. loops)
- The optimization is iterated, first satisfying short-range restraint and then long-range
- Every model produced is carachterized by molpdf, DOPE score and GA341 score
- The DOPE (discrete optimized protein energy) is a statistical potential described by Sali in a paper (doi: 10.1110/ps.062416606)
  - It approximates the free energy of a protein, since the native conformation was demonstrated to have the lowest free energy (Anfinsen, 1972)
  - It is too costly to compute the free energy
  - DOPE is constructed from a set of crystallographic structures, so it is knowledge-based
- GA341 (doi: 10.1110/ps.062095806) is a function of the statistical potentials, z-score, compactness and sequence identity with the template
  - It ranges from 0 to 1, where 1 is a good model
  - It is based on machine learning