**LABORATORY OF BIOINFORMATICS 2018-12-21**

*A sequence is a sequence.* --ok, I guess?

*We live in a society.*

The problem of describing a protein: the protein in known in terms of molecular orbitals of the main chain (*perhaps*?), at least we have some knowledge. It depends on what we want to do: what we do here is an approximation that is typical of bioinformatics, but THE SERIOUS THINGS lie in biophysics. In which reference frame am I? In structural biochemistry (and bioinfo) we are going on with reduced representations, starting from the pdb file. Pdb file: a representation of atoms as points in a cartesian 3D-”plain”.

Since 5 years, the PDB requires also the electron density (the original data). From the so called diffraction pattern of the molecule, I have to understand how to reconstruct the protein structure (very complicated mathematical procedures: through the Fourier’s synthesis). Then the atoms and bonds are placed depending on the goodness of your data (=> resolution).

Again: going on this (referring to image on protein representation slide) is a representation of the main chain of the protein (its backbone). This is totally invented (there is no folding), because it represents the chain in the unfolded representation). A good way to describe a disordered protein is not with disorder, but with “less folded” proteins that can have different conformation in solution. Dihedral angles -> angles between planes in space (not represented within a single plane, as we are used with angles)

Psi == at the level of C-alpha (the only flexible carbon -- the others are not because of the specific electron density of the peptide bond).

Phi == at the level of N in the peptide bond.

Reminder that the peptide bond lies within a single plane

Physical nature of the hydrogen bond -> fluctuation of a hydrogen between two electronegative centres, which **must** be coplanar! 

People through the years took advantage of this knowledge to describe the alpha helices considering the phi & psi angles.



The alpha-helix is a geometrical entity in which all the residues have similar values of the two torsion angles phi and psi. So, I may individuate a region in the Ramachandran plot and state that all residues that fall in that region belong to alpha-helices.

So in short, certain regions of the Ramachandran plot can identify motifs of local order in proteins.

Chi angles -> dihedral angles between the various atoms in the lateral chains.

Omega angle -> practically constant because it’s referred to the rotation of the N-C bond, which has partial characteristics of a double bond and hence does not rotate.

Representing atoms as points (such as it is done in a .pdb file) is a very strong reduction, because we lose the richness of information specific of that atom

Beta factors, temperature factors: one refers to the occupancy of the electron density, the other is more related to the flexibility of the atom in that position in the crystal (how the coordinates of that atom vary in the copies of molecules in the crystal).

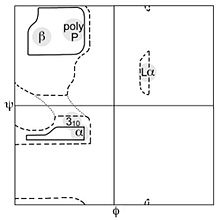
There is a strict distribution of Calpha-Calpha values (bimodal).

→ [SCOP](http://scop.mrc-lmb.cam.ac.uk/scop/) : PDB-derived reorganization of the structure in terms of the secondary structure content of the proteins (all alpha, all beta, alpha+beta, alpha/beta… then for multidomain proteins there is a further classification which we are not mentioning)

A protein is a protein when there is a certain proportion of hydrophobic-hydrophilic residues with properties that allow the protein to cope with electromagnetic field distortions.

Hydrogen bonds stabilize the protein and local motifs are present. Thus, an all-alpha may have up to 5% of beta, all-beta can have a small content of alpha.

All of the above pertain to the description of the **protein topology**.



The Ramachandran plot has precise regions determining the secondary structure of the residues that fall within them. We have a region for right-hand alpha helices + 3-10 helices, a region for left-hand alpha helices, a region for beta sheets/strands and poly-proline.

Comparison of a computed model to these allowed/forbidden regions in the Ramachandran lets us infer considerations about the quality of our model.

In NMR experiments, the data type is not electron densities. They are retrieved from the distance maps. No resolution data! So for now, when we do modelling by homology, let’s stick to X-ray resolved structures. Regions far from the solvent -> x-ray and nmr give identical results. Regions exposed for the solvent -> NMR gives you more than 1 configuration, because it captures the flexibility at the level of the coils in touch with the protein surface. From the various configuration the “consensus structure” is computed (so that our pdb files derived from NMR won’t be an absolute mess).

Temperature factor -> is that position of the protein flexible? And if yes, how much?

in structural biochemistry we are going on with a reduced representation starting from a pdb file, in which we represent atoms as points in a cartesian space with a set of coordinates; we have other 2 columns to give an approximate representation of the electron density => B-factor, temperature (see above)

a disordered protein is described with the less folded proteins => unstable in a given configuration

electric permanend dipole

O,C,N lie in a plane => backbone

2 rigid planes => the flexible point of contact/linking atom that is the Calfa, that carries along all the C-beta(residue)chains

omega => another torsion angle not considered much => the peptide bond is very rigid

alpha helixes and beta sheets => how the backbone is organized within protein space

strict distribution value (bimodal) of distances among all proteins that are generally considered templates on the pdb

coil => non ordered portion of the backbone 

all this stuff can be deleted if already written

50% of the residues are in ‘coil’, 25% are in alpha, and 25% are in beta. That means that after all no matter what there is also a certain proportion of, first of all, H-bonds that stabilize the protein, and secondly motifs.

The poor Ramachandran was able to invent this representation. *Ramachandran being unamused stares at you very sternly →*

a way of checking the protein stability is the possibility of reconciling the model you compute and the structure you generate in a crystallography experiment with whatever is in the pdb

stick to templates derived from x-ray crystallography

What is the difference b/w a structure resolved by x ray crystallography or by NMR? *Hopefully no difference.*

When we are thinking in terms of regulating the functioning of a protein, we are in fact thinking of all the TF that go back to the nucleus and “clump” and “cluster” around DNA, modulating transcription and ultimately protein synthesis.

How many types of secondary structures? At least 8, according to the [DSSP](https://www.wikiwand.com/en/DSSP_(hydrogen_bond_estimation_algorithm)) program.

*it is in the presentation*

*what is a stabilizing interaction? (typical test question)*

*The major source of stability in proteins, but also in DNA, RNA and their interactions/complexes/organization, are H Y D R O G E N B O N D S*

Biological molecules are always either hydrophobic or hydrophilic.

Whatever can be stabilized by the solvent is capable of hydrogen bonds or dipole-dipole interactions with the solvent. Vice versa, hydrophobic molecules do not make interactions with water.

The generation of compartments is paramount to the existence of life and according to many theories it has been also the requisite for the origin of life.

… But this is just a theory. *A GAME THEORY.*

In spite of calling them “hydrogen bonds” they are not really bonds (as far as we are concerned everything that is not ionic or covalent is not a bond, although chemists may disagree)

AGAIN THE EQUATIONS. Molecular dynamics approaches for description of the interactions depicted. These models are simplified to the only question easy to answer for us: which is the distance between the two atoms?

I may claim that the distance between two charges in a protein is whatever. The ion-ion distance is consistent with th protein dimension. These types of interactions are suitable for long distances (as long as 5 Angstrom in the context of the protein)

Stabilization of cysteines with disulfide bridges (= weak covalent bonds) occurs at most 2 A bond distance, even better if the distance is slightly less (1.8, 1.7 Angstrom).

The lower the energy of a bond, the higher the distance!

distance between two atoms can be easily computed from pdb file

cysteins => at most 2 A bond distance => disulphide bridge is a weak covalent bond; weak is related to energy: the lower the energy the higher the distance

functional annotation wins over seq id => consider this while selecting the template

function is known before you start modelling => for our projects its in the title

moonlight proteins can change conformation

Moonlighting proteins -> they change their conformation based on their subcellular location or on other conditions (and in each conformation they perform a different function, this is why they are called moonlighting, because they don’t have an active and inactive form but in a way both forms are active). (this is just a trivia fact)

poor enzyme

list of things that are poor today

* hydrogen
* Ramachandran
* laccase enzyme

How far can we go computationally with the study of the surface of a protein? By studying the surface accurately, one may even understand things about the reaction process, how and where in the protein the binding of the substrate occurs… but to do these kinds of studies it takes years and they pertain more to quantum chemistry than to bioinformatics.

What is an alignment? First of all, it’s a method.

Why alignment and not just comparison? Because while performing the alignment we are preserving the topology of the two entities under consideration.

You have to define a metric

Any alignment method around can be clustered into: sequence vs structure; global vs local.

Then, for practical reasons, you need alignment methods suited to retrieve data from a database, which means they need to be faster and less complex (such as the BLAST algorithm).

Dynamic programming is involved in the alignment procedures. Then, we have to decide how an algorithm is developed: definition of a metric (that may or may not include indels)

Alignment methods are completely and always heuristic.

how is an alignment algorithm developed? defining a metric that may or may not include **INDELS** (( ←we sure about this? yes, we asked.))

*((( in this context, I think, insertions can be represented*

*as gaps in the template, whereas deletions*

*produce gaps in the target sequence )))*

alignment is an operation based on heuristic approaches => optimized only when you have templates that already give you the right answer => no real theory behind them => for this reason whatever method you use you retrieve hte optimal design but behind we have many different results

in proteins when we do struc alignment when we sueprimpose the main chain of 2 proteins and the RMSD is within 1 A or at the most 2 A (with 3 its pretty bad)

i can have the correct sequence alignment as derived from the structural alignment: no alternatives => im superimposing the backbone, and so each calpha will give me the superimposition between two residues that may or may not be conserved

we have the problem of reliable seq alignment methods => compare original seq align with the one derived from struc alignment => if seq id is high not much difference, with low seq id its another story

When we do structural alignment in proteins (although ti may be affected by several problems), when we superimpose the main chain of two proteins and the RMSD of the main chain is within 1-2 Angstrom, I can have the correct sequence alignment as derived from the structural alignment. Each c-alpha will give me a correspondence between two residues (same position in the two proteins) that may or may not be conserved.

Reliable sequence alignment methods are necessary because we don’t have structural data for most of the sequences. When the sequence identity is very high there is no big difference between sequence-sequence alignment and sequence-from-structure alignment, but if identity gets lower this is absolutely not granted.

For proteins, to define a metric requires to go back to biology and relate again to the physical, chemical and evolutionary properties of the amino acids. This leads to the definition of scoring matrices.

How many matrices do we know? PAM, BLOSUM, McLachlan...

Scoring matrices are there to *score*  the alignment (in any scoring matrix there are weights allowing to build up a total sum of the scores point by point). Each substitution is weighted, as much as the gap/indels.

The length of the alignment is completely different from the length of the sequence. The length of the alignment is (((visible confusion))) (((visible colpo della strega))) (((drumroll because nobody ever told us what the length of the alignment is)))

The sequence alignment length is the overall length of the aligned regions, considering gaps.

Gap modelling (reminder from Martelly): linear model vs affine model.

Sometimes the global alignment is necessary to understand the coverage of a sequence versus the template (as in our case). But sometimes both types of a lignment can give us information, so it’s not that one precludes the other. If two sequences have more or less the same length, and they have a high sequence identity, the same properties will emerge from both global and local alignments. Sequences of different length/different families… lead to more problems and global vs local makes a huge difference. Statistical scores most used around:

* a

What is a score? It’s a number used to assess the biological relevance and reliability of a result. Score = sum of all the single substitutions (given by the scoring matrix) minus gap penalties.

Why do we introduce gap penalties? Because sometimes, the score given by a gap opening is greater and more convenient than a substitution of a certain kind.

***\*Suppose that you are the algorithm\****

*dO NOT FEAR THE ALGORITHM*

*bE THE ALGORITHM*

t h e f a s t e r t h e b e t t e r

working harder make it better do it faster makes us stronger hour after hour working our work is never over

BLAST is conceived to optimize the search against a database, with a fast execution time and a medium reliability. The tHrick is to cut sequences into words and look for occurrences of those words in the database. You can select the length of the BLAST word to adjust precision/speed. If BLAST finds a hit, it tries to extend the alignment.

BLAST has I don’t know how many papers (Author of BLAST: Althshul)

Every alignment method should have a means of statistical validation (cause otherwise you wouldn’t go along with it huhhh)

Bit-score: rescaled version of the score depending on two parameters, lambda and k, depending themselves on the dimension of the problem. Lambda depends on the dimension of the query sequence. K depends on the substitution matrix and gap penalties.

Knowledge of p-value and E-value is *required*.

Statistics of local alignments -> you stop the operation of searching for the optimal alignment when these segments of local alignment can’t be extended without incrementing or lowering the score (it doesn’t lower because at the extremities if you try to open a gap you get a reset of the score)

Maximal segment pairs = locally aligned segments of query vs entry. When we are presented with BLAST result, we are starting with the highest homologies but if we go down the list of results we see that only fragments are aligned.

Main difference between global and local alignment -> in global you try to span the entire coverage of the two sequences and so you accept easily gap opening. Local alignment is a superimposition of fragments. Altshul’s paper is very difficult to read but it’s in the dropbox.

Statistics -> originally designed for alignments without gaps, but it was established that it’s valid for gapped alignments as well.

the usual question is “what is the E-value?, from where does it come from?”

“Hello hello hello? Good for you”

“There is a demonstration, on the Altshul paper, that is quite… prrrt - quite SOMETHING”

MsM you really are an excellent curator

I like kekcellent

The maximum of a large number of independent identically distributed random variables tends to an extreme value distribution.

The E-value is some measure of reliability of the result related to the dimension of the search space and to the scoring matrix used (parameters k and lambda)

Lambda -> number of expected sequences that can be randomly generated given the frequencies of amino acids

**E is the expected number of sequences generated randomly that score equal or higher than a threshold S.**

**the expected number of HSPs with score at least S is given by the E-value for the score S:**

**E=Kmn exp(-(lambda)S)**

P-value associated with the score S

Lalign was invented by William Pearsons. It’s an implementation for both global and local alignments in a more accurate way than with BLAST. BLAST is just for database search.

The algorithm implemented gives a better alignment. Different matrices can be employed. Everything is easy when the two sequences are very much identical.

Lalign also allows DNA and RNA alignments.

Reminder that A:T has two Hbonds and C:G has three.

Lalign outputs MANY local subalignments. The optimal alignment is the one that goes along with the parameter that you are selecting. E-value threshold in LALIGN: changing the number of sequences that you are considering to retrieve from the operation (???). If you think that you are comparing things that are very far apart, better to keep this value high.

In general in LALIGN default option it’s gap opening that is heavily punished.

Projects should be delivered as three separate files for the three proteins, each containing the whole workflow.

Alignments in the project can be decorated/color coded to show conservation of residues and other features that you may want to highlight.

----- Second part of the lecture

[Swiss-model](https://swissmodel.expasy.org) -> “*SWISS-MODEL is a fully automated protein structure homology-modelling server*”. It’s a graphical web interface implementation of Modeller. One of the most user-friendly ones, implemented at the Swiss Bioinformatic Institute, University of Basel (good center for structural bioinformatics). {quoting from the site: SWISS-MODEL is a fully automated protein structure homology-modelling server}

This implementation may be different from the Salilab version of Modeller. How can we consider this? Quick structure prediction of a protein -> SwissModel is the place to go. They both find a template and compute the target structure.

You still have to perform structural superimposition, functional annotation is your responsibility or you could compare this structure with the one you computed. You can download the Swiss-model structure.

(“You might want to go to Switzerland for a couple of reasons…”)

SwissModel, if you don’t specify it, automatically searches the template for you. Be aware that the program only evaluate the template on the basis of sequence identity and template resolution → so it doesn’t necessarily pick the best one for functional annotation. The whole point of swiss model is just to produce a model. Swiss model could even use multiple templates to produce a model.

Solvation (solvatazione) -> interaction/exposure to the solvent of the protein.

In RasMol you can display the water-exposed surface (computed as → stereochemical hindrance of the model of the protein and colour it according to electronegativity. Blue is positive charge while red represents negative charge.

On SwissModel you can also upload your own model and evaluate it.

From the Swiss-model webpage

*“GMQE (Global Model Quality Estimation) is a quality estimation which combines properties from the target–template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, reflecting the expected accuracy of a model built with that alignment and template and the coverage of the target. Higher numbers indicate higher reliability. Once a model is built, the GMQE ((1) in the figure above) gets updated for this specific case by also taking into account the QMEAN score of the obtained model in order to increase reliability of the quality estimation.”*



why is she showing us this things? So slowly? Is she hoping to ship us (-> I just thought of aristogatti when they get delivered to timbuctu, yes it sound like she would be happy to do that) off to switzerland next year?

(edgar is happy because he is going to inherit all the go terms)

*And then you can annotate the moleculeeeeeeeeeee*

it’s like to enter in a convent

you are firmly believing that for a year and a half you don’t have anything else to do

well I guess np I never had a life anyway



## Gene ontology reviewing

What is GO?

Most friendly concept for computer scientists, in the sense that it’s a computer-readable translation of functions

It’s a consortium that originates the GO on its own. Ontology: ? Gene Ontology: ? Organization of the GO: ?

To make a text **computable → a very current issue with biological data in DBs**

[Ontology](https://en.wikipedia.org/wiki/Ontology_(information_science)) -> it’s basically a vocabulary.

The more we get to know about the items around us, the more we need to express words and enrich the ontology we are working on.

Each word is represented by a number → a mathematical descriptor which make it possible to compute

You also need a key

directed acyclic graph = HIERARCHICAL STRUCTURE of the terms! From more general to more specific. Molecular function on its own is the most general GO term from which many branches and lower hierarchical levels depart.

THEY will have an APERITIF?

I ADVISE AGAINST.

Do you think that she improves he she’s drunk? or does she get worse?

can’t really say but I would be curious, yeah me too, but like it should be in a safari fashion

it’s nice to see lions but not too up and close

←improves dramatically IMHO, but only if she’s not speaking to you

*good for them*

*g o o d f o r u*

What is gene ontology? GO is a consortium that originates the gene ontology on its own

Also a db

Also a vocabulary that allows to make computable knowledge associated to biological molecules

Its a dynamical procedure

most comprehensive resource for gene functions and etc

aim: how to describe the function of a molecule (in medical areas gene=protein)

we want to make this knowledge computable => we may relate each term to a random number (biunivocal correspondence) => with a mathematical descriptor we may start computing

in any file i have lines with a certain header => in order to be able to address specifically a term i should be able to have a key that connects me with the mathematical descriptor

The GO knowledgebase is composed of two primary components:

* the [**Gene Ontology (GO)**](http://www.geneontology.org/page/ontology-documentation), which provides the logical structure of the biological functions (‘terms’) and their relationships to one another, manifested as a directed acyclic graph
* the corpus of [**GO annotations**](http://www.geneontology.org/page/go-annotations), evidence-based statements relating a specific gene product (a protein, non-coding RNA, or macromolecular complex, which we often refer to as ‘genes’ for simplicity) to a specific ontology term

hierarchical structure: goes from more general to more precise terms

when i say molecular function i also say that molecular function is the most general go term addressing the problem of molecular function

laccase => is a molecular function that goes in at the edge (“leaf” from the roots of molecular function)

reactions are described by the EC numbers (also hierarchical numbers)

as of yet we don’t have all the enzymes to cover all the EC numbers

EC2PDB

when we say molecular function we are not addressing the problem of a specific molecular activity/function

graph resembling an upside down tree => from roots to leaves without any connection between leaves (DIRECTED ACYCLIC GRAPH)

it was found more convenient to distinguish the GO into 3 major clusters : a protein is a heteropolymer capable of auto-organization in the space of a solvent => one cluster of GO considers **molecular function,** another considers the fact that any protein is involved in a biological process (**biological processes**) and **cell component**( where the protein in the cell is located as a functional entity)

600 000 experimentally supported GO terms

=> **the core of the GO terms are derived from experiments**

inferences computationally obtained

Each term of the controlled vocabulary is biunivocally associated to a GO number.

GO annotations -> evidence-based statements

Enrichment analysis:

Considering a set of genes, each with its own go terms associated, is there something in common among all the genes in the set? This is done specifically to understand whether there is something significant to note in a large number of items.

*And we consider, ELÒELÒ, molecular function*

ECO:NNNNN = EVIDENCE CODE! (inferred, experimental…)

Majority of the UniProt entries are not even real proteins (do you even protein VRO)

The later you take the LB1 oral session the worse it’s going to be

no pressure

Bioinformatish Greatest Hits (2018), by artist Rita Casadio

Track listing:

* papapapaam
* I HAVE DATA
* ta-da-da-daaan *[papapapaam alt. take]*
* this is fantastish
* good for you
* *dududùn (x3)*
* so far so good
* (this is also) g o o d
* *ELÒELÒ*
* *[pernacchia]*

B-sides:

* this is populism
* (do not be) an end user
* what you want to do in your life
* (((generic screaming)))
* then you know everything
* very good indeed (rare bonus track)

Let’s ask her if we can join them tonight STRONG AGREE I AM IN

biological processes and metabolic pathways: the latter are a subset of the former. Biological processes can involve the cross-talk of many pathways. (Ensemble of all the reactions, including various metabolic pathways, to which a certain protein can participate).

Sometimes, if the model is not conceived well, heteroatoms/cofactors are not retained, so you cannot identify the structure of the active site.

***pfam is her favourite, don’t forget***

wait what

what did she say about maladies and mutations? she was joking or i will kill her

KEGG lists all the possible pathways where a protein can be involved. (Kyoto Encyclopedia of Genes and Genomes)