laboratory of bioinformatics 13/12

From here to Christmas -> only one more lesson next Thursday (also friday we have all day)

20th December: we will start discussing the project. Have something written by then not something but more like the whole project n°1).

The other two projects will be carried out and written by ourselves.

reconsider what we learned together in order to do the projects (of course). For project issues we can contact Casadio. Since in January there are no lessons we will need to make an appointment. Deadline: end of January. Mark of Allegra Via week will be consider for the final mark in June.

To attend the first oral session send in projects well ahead of the deadline so to have time to oversee the corrections.

* write project n1
* read the article <https://salilab.org/pdf/Sali_JMolBiol_1993.pdf>

modeller compute the distance between Calpha-Calpha to do a model of the backbone of protein. Force field (or energy field) by changing relative atomic coordinates is used to represent a protein instead of using geometrical restrain. You have to adopt all the equations for representing pairwise interactions and atomic interactions. When there is identity with template you have precise position and angle because you transfer the information from template itself. If you change something you can only have a statistical precision.

Given our task (functional annotation [experimental function]) there are different methods to choose a template. First step is to consider the protein family of the sequence I want to analyze. I want a coverage of at least 70%. As soon as the atomic resolution is not so bad (2 A). Experimental GO terms is the way to choose the best model.

1)Coverage  
2)select templates that have experimental determination of GO terms

When you have a fragment may or may not represent a protein. Be aware that use a fragment of 30% coverage it is a possible representation of a protein.

3.8 angstrom is the distance, in general, between two Calpha. Is also a method to compute the probability, is the average value of this distance.

There is an older database that helps modeller in terms of normal value.

With different distribution we can compare feature of our sequence. There are distributions for many features. This features describes protein stability, not structures.

In swiss modeller you can upload the sequence and it predict its structure. But we don’t want to predict structures, we want to understand when specific functions could be associated to a sequence. This is the f u n c t i o n a l a n n o t a t i o n problem. When the sequence identity is low we have a lot of problems.

We find the restrains as probability density function (PDF). It is optimized with the variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing.

Problem of statistical evaluation of model:

We have a model with his score. We are computing an objective function (His parameter is the important one). The best solution is not always the best model for modeller. To be sure that the active site is annotate we have to see the literature. I am able to predict the active site of a protein? (Is the question that all the Bioinformaticians do to themselves every night).

Print screen the table in the paper and comment this.

Phi and Psi angle: THE PROBLEM.

features are derived from a protein structure representation in a geometrical way and they are somehow describing all the possible sources of protein stability= > as soon as modeller starts changing gears and applying some molecular dynamics steps, all these details are abandoned in favor of an imaginary force field in which the protein moves and it may change the conformation => we are using an energy field (Charmm) in which by changing the energy you may even change a little bit the protein structure

adopting all the equations in terms of electrostatic pairwise interactions in which you have numerator and denominator depending on the structure => all those equations (conjugate gradient) are necessary to describe all the atomic interactions in the protein => perturb and then relax the protein to find the exact position of each atom

not even molecular dynamics can fix for you the optimal torsion angle of the lscs => whatever angle you get may be good(identity) or, if you change the lsc, you will have no clue on the correctness of the position of the atoms in that part of the protein => so i have to statistically compare to obtain a score of how good is my model compared to structures in the pdb

i have a template for the protein family => how do i select it? there are several procedures to adopt = > given our task(functional annotation) we don’t want to transfer guessed functions, but experimental functions that have been associated to the protein structure in the pdb

if i predict, i might be wrong (cit. radiohead) (best cit)

i should go for at least a coverage of 70%

atomic resolution => as long as it’s not very bad ( in the pdb you may find resolution of on average 2 Å)

don’t be confused my friends => fancy techniques

don’t go just for the resolution, go for the fact that whatever in the uniprot is associated to the template as experimental determination/ GO term => i will export it later => associate functional annotation to ??

**basic rules for selecting template:**

1. **coverage => atleast 70%**
2. **select templates that have experimentally determined GO terms => check the labels (we will do it together 2day)**
3. **use good common sense if you have it => check yourself what you do, you have to be able to criticize**

Template selection =>

i prefer to have one step of transferring and

in our case low/middle is in any case around 2A res => any res above it has lscs poorly organized => we dont want it

how do you select the template for your building by homology in order to do functional annotation?

a fragment may or may not represent the overall protein folding

a fragment with 30% coverage is a poor representation

in general alfa carbons of residues in proteins are about 3.8 A apart

having defined the spatial restraints through pure statistics from databases, they have to extract the same features from the template and compare them with the db

density probability function => will have an interval

Modeller uses the pairwise forces we have seen many times (expressed as equations) to describe all the atomic interactions

All the equations we formerly saw regarding electrostatic pairwise interactions are implemented in a conjugate gradient algorithm. The field perturbs the system and hence the protein structure. When the perturbation is relaxed, the protein may assume a slightly different conformation, that is optimized

The optimal torsion angles cannot be defined by molecular dynamics (or by any other algorithm really). The only way to have very high probability of considering optimal the torsion angles of the template is only to have identity of sequence between template and target (you can transfer the structure perfectly to the target)

then you start comparing your structure to what is in the pdb.

We have to select a template according to our task: functional annotation, so we should use protein that are well characterised, whose function have been experimentally evaluated and confirmed. We try to minimize the steps of prediction (we aren’t predicting function, we want to transfer function experimentally collected ).

**Criteria for template-picking**:

0. To model sidechains the resolution has to be under 2A

1. The coverage should be at least 70% to obtain a good model
2. Experimentally determined GO terms → check them yourself

[beware that GO terms are not a golden standard]

If the coverage of 2 templates is similar you should prefer the one that is better annotated (criteria 2), not the one with better resolution (unless the worst one is higher that 2A).

functional annotation of a given sequence: with the concept of the protein family.

a fragment may or may not represent the protein folding of the whole native protein, to use a fragment that possibly has a coverage of 30% of your protein is not so good, so you have to use others approaches if you want to retrieve the protein structure.

having defined the spatial restraints, the extract the spatial restraints of a sort of database to which you can refere.

you can extract the same feature for the template .

then you can compute the density probability distribution for each feature.

Calpha-Calpha distance for adjacent residues in a protein is about 3.8 Angstroms:

This mean distance comes from a golden standard of selected proteins



when you have computed all these restraints , this equation is telling you that if you want to derive spatial restrains you have to focus on a certain average measure of how pairwise interactions work (e.g. the hydrogen bond is a bond between two electronegative atoms that “share” a proton and on *average*, the bonding distance is of 2.2-2.5 Angstrom

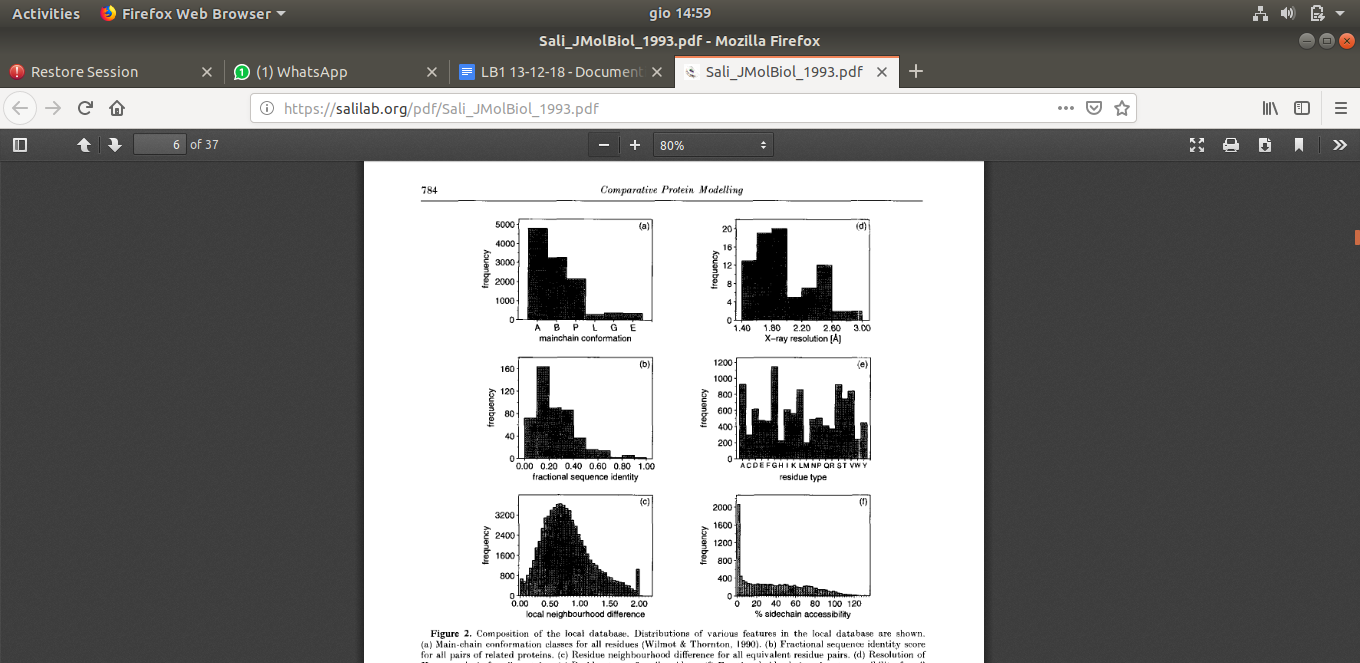
you compute a pdf for each feature

derive spatial restraints: measuring the density probability distribution of a given property.

local database behind model and helps modeller fixing all the parametres.

where did the 2 clusters come from???

The structural comparison of two members of a same superfamily doesn’t make sense, because when we compare two members of a superfamily (i.e. performing the same function, but with a different architecture of the active site/different mechanism/different overall structure) the two proteins might be quite distant and the results that we get will be misleading.



The *features* Modeller implements are exclusively related to a geometrical description of the protein **stability** (not even just protein structure)

All these protein are stable by definition, they are derived by xray diffraction

we want to understand when we can associate the function of a given family to a sequence.

if we don’t understand the problem of functional annotation in silico we will never have an answer to this big problem => cafa? big experiment, with a bunch of proteins whose structures get predicted and compared

under 30% seq id we can’t use automatic annotation => we have to be sure that the active site is there => warm suggestion of preliminary seq alignment to understand if the critically important residues are conserved

annotation on databases relies on manual procedure.

when the sequence identity is getting lower the problem of functional annotation becomes critical and can’t be left to automatic processes

problem starts to be effective since you have to be sure that you active site is there!

since critical residues at the active site have to be conserved

Demonstration of the functional annotation validity ---->>> presence of the active site in the modelled structure. Verify carefully validate the initial alignment that we “feed” modeller with. When we are aligning one residue of the target with one residue of the template, we are basically fixing the spatial restraints for that position.

**The restraints are fixed as probability density functions.** I don’t want a certain value of a distance or a geometrical parameter to exceed the extreme values of that distance as normally occurring in real proteins. (back to the example of the hydrogen bond).

The molecular pdf is optimized through conjugate gradients (in some way when we adjust the parameters we try to go along the gradient, not against gradient, so towards a more probable configuration) and molecular dynamics with simulated annealing.

After this step there is a smoothing process with CHARMM, which is the most suited for protein dynamics(???)

how do they define the restraints? as the probability density function => i don’t want that specific value to exceed the extremes of the distribution,should be in that range of value.

you start with each value and you have a frequency distribution and then you have to evaluate this probability density function that allow the variable to shift between max and min

on top of these functions modeller also employs some molecular dynamics => force field (charmm)

we have models => each model gets a score which is computed against a db => in any case i have to understand on my own if the protein is stable or not =>

we have to optimize the value of the models, if we are searching for the optimal value we have to go along the conjugate gradient.

**Statistical validation of the model:**

At the end of the procedure Modeller returns a table of three models, each one with its score, computed against a database; beware: this table is in the final output in the terminal but it is not saved automatically so take care to save it. At this point we have to understand whether a protein is stable or not and there are different ways to do this.

assess\_methods = (assess.DOPE, assess.GA341)

-> line to insert in the python script for Modeller, in order to retrieve some more statistical scores

scoring parametres: one is based on objecting function

The default parameter is the molpdf (that parameter that we mentioned last time saying that it rose when we broke constraints, so the lower the better). It is written inside each pdb file in a REMARK line

as long as the sequence identity is high => all the models are more or less the same => around 30/40% seq id => the numbers differ instead of being very close =>

When Modeller spits out the final table with the results filename and the associated molpdf parameters, save that final table in order to put it in the project later.

DOPE can compare proteins derived from different modeling procedures(??), it’s based on the energetical optimisation.

The best solution of Modeller may not be the best solution, we are looking for a model with a very low RMSD when compared to the template AND we want it to *overlap in the active site/critical regions and amino acids*.

In order to find it you have to do structural superimposition, then zoom in on the active site to ensure that the model matches well the template. As we saw in the paper related to the template, by changing even just 0.1 Angstrom those distances, the activity of the enzyme is going to be impaired. So our model needs to satisfy those sort of restraints, *in order for us to be able to transfer function*.

Attach a picture in the report with distances measured.

how to be sure that thàe active site is conserved here?

there are the actual geometric distances between coppers. Those distances should match!

best model might not coincide with what you are looking for for functional annotation => do structural superimposition => zoom on the active site => give the file an answer with a transfer of annotation if possible => how to be sure the active site is conserved? there are the actual geometric distances between the coppers =>

3 main issues regarding structural bioinformatics:

1. am i able to predict the structure of protein?
2. am i able to predict the structure of an ACTIVE protein?
3. at the end of the structure prediction, is the protein that i computed a stable protein (and hence stable enough to perform a function)?

print figure 2 => zoom on active site and highlight the fact that the active site geometry in terms of the pyramids of the copper atoms is consistent with that of the template

since the active site geometry in terms of piramid generated with the coppers atoms have to be consistent with the one of the template.

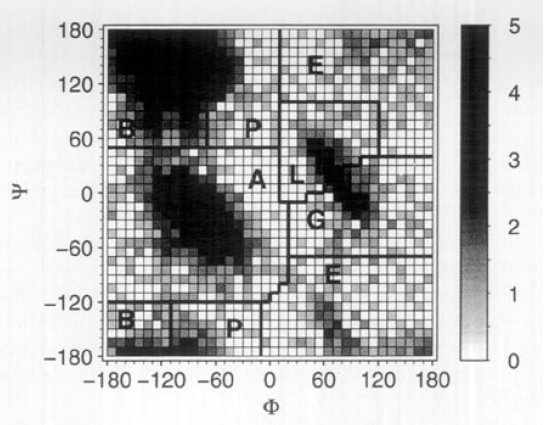
**Results**:

1. modeller output => print screen table in the end and select a model among the 3
2. comparison of template outcome with target seq => superimposition of structure and show relative conservation of topology
3. sequence alignment after 2) to be compared with sequence alignment derived from structure superimposition (best alignment is sequence alignment form structure alignment) show alignment of target to template sequence(input to modeller) it may happen again if the two sequences have poor seq id the alignment is different from the one we start with => adopt the ?st seq alignment
4. zoom in on the active site and comment on its conformation (adding copper and ligands if you want); show the superimposition of the histidines with perfect conservation by “hand” (on RasMol/RasTop), measure the relative distance in the template and in the target model => know in advance that ***the difference shouldnt exceed 0.1 A***
5. transfer of GO terms => **functional annotation**
6. optional(for the best mark or she might ask this question during the oral exam): discuss stability of the target model (after 1st step) (using ramachandran plot and procheck)

when the protein is a stable entity? we are implementing an objective function that contains all the possible non bonding interactions at the level of the different atoms

=>

We said that the best sequence alignment is the one we derive from structural alignment. It may happen that at the end of our computation, we can derive a sequence alignment from the superposition of our modelled target over our template. This sequence alignment may be different from the one we gave to modeller at the beginning! (This concept has already been stated so it should be somewhere in the previous google docs)



mnRAMACHANDRAN PLOT

It’s a plane in which each residue is represented as a point, with coordinates equal to their phi and psi angles.

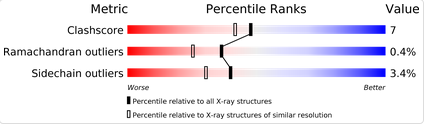
(<- Sali et al., JMolBiol, 1993)

Each region in the plot is associated to a structure, but not all regions are equally likely.

(((for a good explanation refer to this [link](https://proteinstructures.com/Structure/Structure/Ramachandran-plot.html), already linkedin previous doc)))

All new structures in the PDB can be analyzed in terms of how their dihedral angles match the “allowed” regions of the R. plot. It can be visualised in PDBsum entry.

as soon as you have a new structure in the pdb you can compute several scores : metric scores ecc...



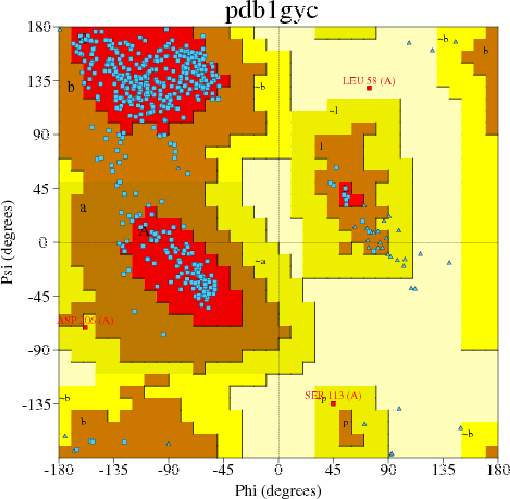
for each file deposited people write algorithms to understand whether the new structure is compliant or not with the majority of the values

i have values , percentile ranks (red => disaster, blue => optimal)

i have allowed regions for phi and psi values in pair that define alpha helix or beta strands Red indicates low-energy regions, brown allowed regions, yellow the so-called generously-allowed regions and pale-yellow marks disallowed regions.

if you want to see the RAMACHANDRAN PLOT you can go in uniprot and the in PDB sum.

RAMACHANDRAN PLOT



<https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=1GYC>

In this figure, blue dots are the amino acids represented as phi-psi angles. The red, brown, yellow, light yellow regions are respectively the strictly allowed, allowed, generously allowed, forbidden regions of the plot. The Leu 58 falls in the light yellow region, that’s why we can’t comment it.

all the blue dots representing the residues

leucin 58 out of order => can’t comment on it in the project, it would be based on wrong assumptions

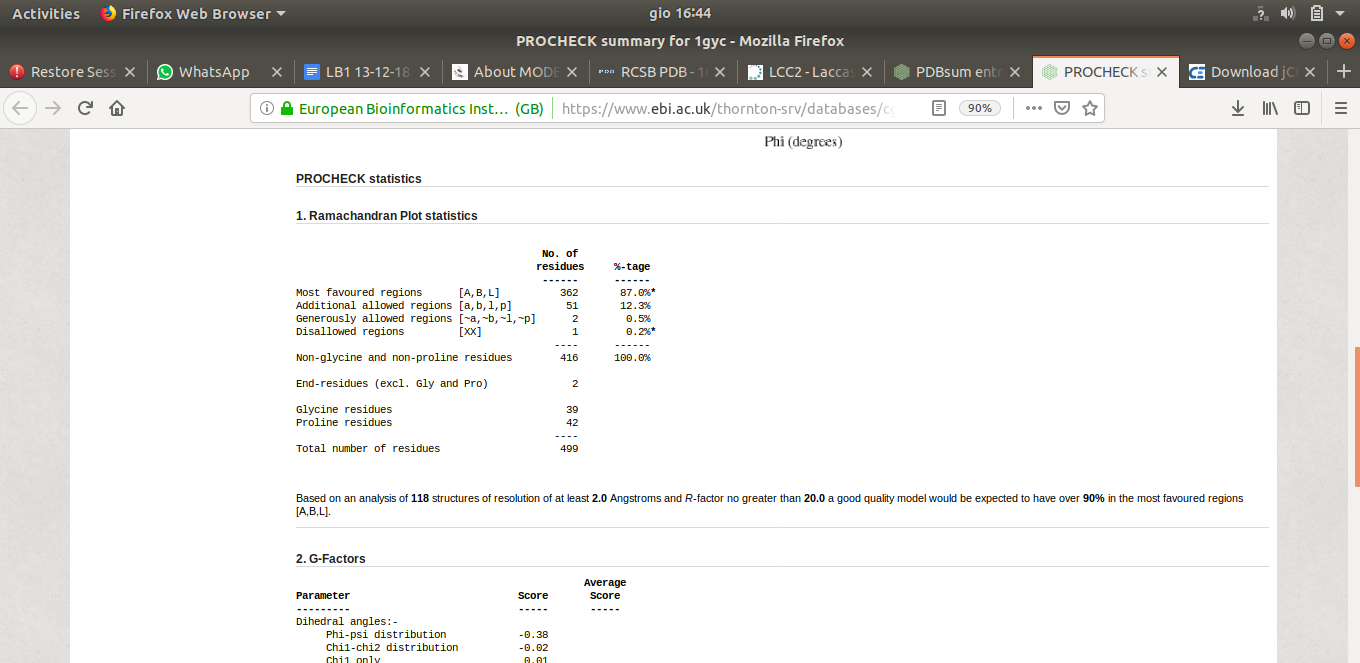
Ideally, I would like my model to present a Ramachandran distribution that is similar to the template. If I have a large fraction of residues in the allowed region, I can safely say that my modelled protein is plausible and might be a stable entity.

These regions were derived from the natural variability of a set of fairly-well solved proteins. The forbidden regions correspond to values of phi and psi that are never taken by an amino acid.

Another strategy could be to compare the Ramachandran plot of 1GYC to the PROCHECK of our derived model.

PROCHECK is an algorithm that allows to do comparisons in terms of pairwise values of dihedral angles, computed starting from the PDB file of our model.

Also PROCHECK needs to be referenced in the project report (in terms of original article), (along with its release?)

Check the R

plot for predicted structure → how many residues are in the generously allowed region? there shouldn’t be too many

compare with the template R plot → the predicted one will be worse

This comparison is done by ProCheck MUST QUOTE IT

between -45 psi and -50 phi are the best coordinates for alpha helixes

less populated regions are not typical organizations of the backbone

the stability of the target model: my objecting function established if the protein is stable

another possibility to answer is to check to compare the ramachandran plot of the template to the procheck of you own project.

**PROCHECK** checks the stereochemical quality of a protein structure, producing a number of PostScript plots analysing its overall and residue-by-residue geometry. It includes **PROCHECK-NMR** for checking the quality of structures solved by NMR.

procheck installed

quote the use of procheck ,either remote or locally installed

Procheck can be found online at <http://servicesn.mbi.ucla.edu/PROCHECK/>

=> perform PROCHECK analysis on all the three

ramachandran plot of the model >= higher number of residues in best position => molecular dynamics helped => model is stable

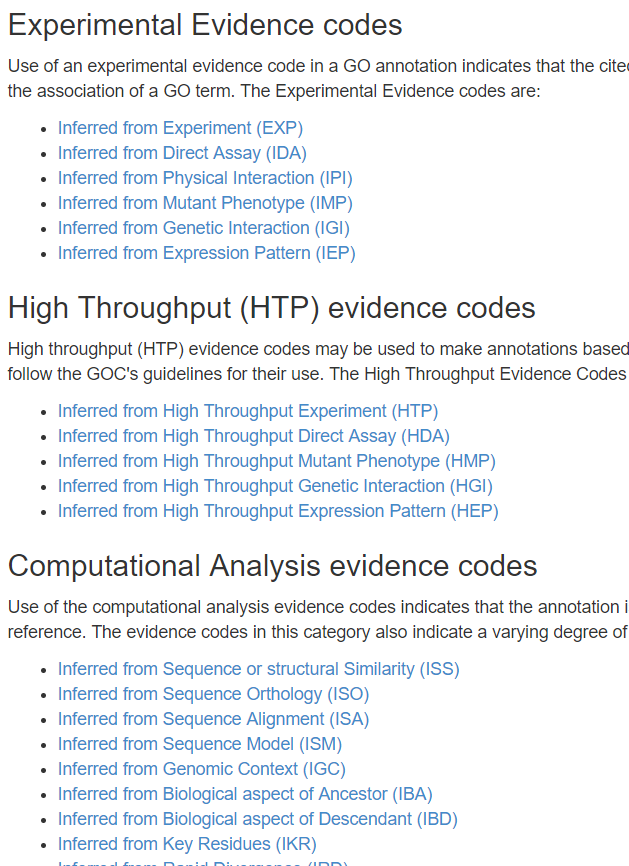
To zoom in on the superimposed active site: use jCE

actually it is better to superimpose in jCE, the save this as a separate pdb file and then open in on RasMol which is better for visualisation. Only then can you zoom in on the active site.

Finally we check for disulfide bridges: are they well superimposed?

GO evidence codes

labels that are independent on what kind of function I am taking into consideration. They are the same for all GO terms. They are written in a specific page of the Gene Ontology organization website (-> <http://www.geneontology.org/page/guide-go-evidence-codes>)



check the Go terms of the entry → View the complete GO annotation on QuickGO

The relevant terms are the F ones

ECO:(numbers numbers) -> this is the tag we are interested in, because it encodes the kind of evidence, or in other words the logical operation that has been used to associate a certain GO level to a certain protein sequence

Evaluate the ancestor chart

<https://www.ebi.ac.uk/QuickGO/annotations?geneProductId=Q12718>

may i transfer these go terms to my model?

are these derived from originally experiment?

what IEA stands for? electronic annotation → does not mean that no experiment was performed

It could derive from the fact that the sequence had been annotated before the structure was obtained.

you can copy and paste these go terms to your protein considering that the copper ion binding is in a hierarchical relationship with other GO terms (such as oxidoreductase activity, which is a more broad and general, so if I transfer the term copper-ion binding, I will automatically transfer the term oxidoreductase activity)

we have also another crystal:

<https://www.rcsb.org/structure/5z1x>

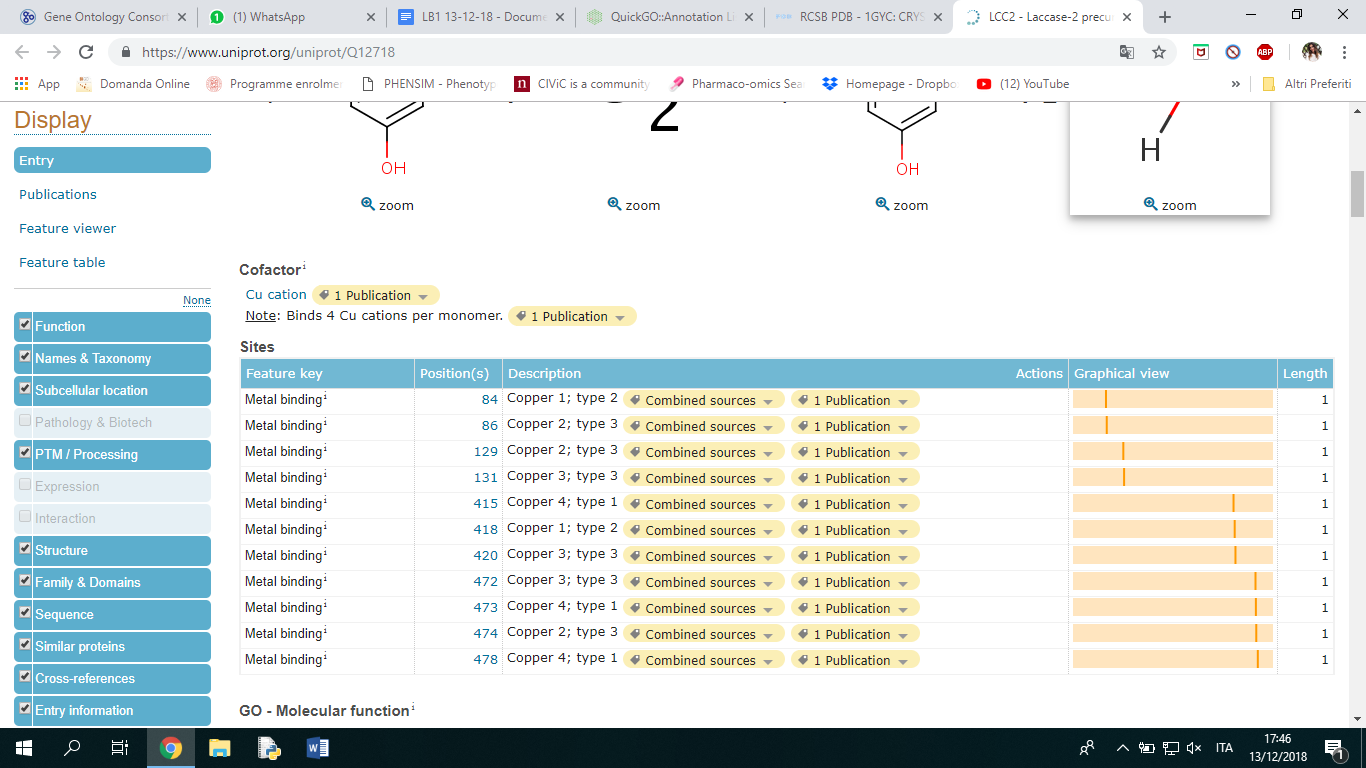
we can not use a protein without a link to uniprot for functional annotated protein

WE ARE DOING FUNCTIONAL ANNOTATION NOT STRUCTURAL PREDICTION!!! <- this remark is less trivial than you would think

please explain the implications

when you just predict a structure you can’t annotate shit

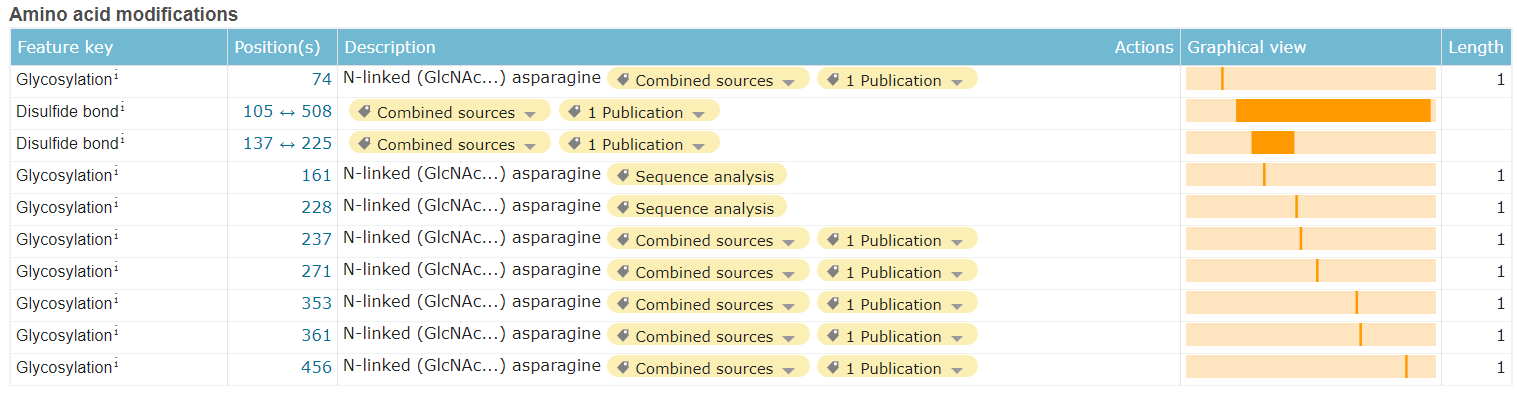
“cofactor” section of the UniProt entry: in the table are described the lateral side chains interacting with cofactors, which are their positions, and which are the cofactors they are interacting with. In this case we can see almost all of these atoms are histidines.



In the PTM section you find glycosylation sites and disulfide bonds

Glycosylation occurs only on Asparagines so you have to evaluate how many aspargines are in your model

i have 4 position for 2 different disulphide bonds, where cysteine have a distance of 1,8 amstron



Should we be asked how many laccases are on the PDB, what would be the answer?

EC NUMBERS in EC2PDB

<https://www.ebi.ac.uk/thornton-srv/databases/enzymes/>

functional annotation via building by homology

Pay attention when transfering GO terms, if you go from eukaryote to prokaryote

there are terms that are specific for one group