laboratory of bioinformatics(3/12):

we compare structure within the same protein family.

when the function is conserved ,the structure is also conserved.

when sequences are at least similar: when they have at least 30% of sequence identity.

foundation?after having identified a cluster where the function is conserved, then we can say that sequence diverge but the sequence identity is always 30%.

Structure comparison within the same protein family -> the structures are conserved

There is a “grey zone” where we can’t infer structure similarity from sequence similarity => we can’t infer the protein family of belonging.

Protein families => superimposition of backbones with RMSD < 3Å, same function.

We may also identify similar structures with sequence identity <30%! Distantly related homologs belong to the same protein family, because their function is conserved as well as their structure, but their sequence is very different.

So starting a systematic pairwise comparison at the level of the structure among all the PDB entries, we get to this rule (evolution-did-it graph). These experiments of structural comparison were carried out at the beginning of PDB and are repeated periodically with all the new entries.

from the pdb by doing pairwise comparison i will obtain the evolution-did-it graph

in reality it is a very simplified simplification of the experiment.

in one axis we have the length of the alignment residues that are similar after the structural comparison.(the *only* correct sequence alignment is the one that is obtained after a structural alignment/superimposition).

template (prototype for the protein family)and target(the one that has to be compared with the prototype).

we start by the superimposition of the structure.

the folding of the protein is the same.

structures: output of the experiments (especially wet lab ones that lead to crystallization of a protein)

and i have to compare them

then i have to do the correct alignment of the sequence .

when i say sequence i’m losing the information about the dihedral angle of the backbone of the protein.

length of alignment: measured in terms of the number of aligned residues.

this graph is bidimensional:

Graph: percentage seq identity (y axis) / n° of residues aligned (x axis).where in blu there are protein that are similar

In the yellow region of the graph, the sequences may or may not be similar in structure. We can’t infer it.

provided that two sequences are at least aligned with a number of 250 residues and the percentage of identity is >=30% we can say that sequence identity implies structural similarity.

es: number of residues is 50, we have to find a high percentage of identity

The evolution-did-it graph gives us some rules and constraints that allow us to identify the famous proteins family.

i have the sequence: how do i know the function? => how can I annotate a sequence, in terms of its function? If I have a sequence and I am aligning it against a database of structures; if I find a template that my target sequence is at least 30% similar with an alignment length of 250 residues, I may *transfer* the information related to the template to the target I am trying to annotate. => I can assign a **function** to my target protein!

to annotate: to endow a biological entity with structural and functional features (protein sequence in this case)

**The protein-folding problem**

may i find a method to assign protein function also to sequences of proteins that I can’t crystallize? Can I take advantage of this clustering?

may i take advantages of this clustering of pdb?

i can derive set of rules and then use them for solving the problem of functional annotation (going through the structural, the assignment of a given function to a certain sequence and then take advantage of it)

functional annotation of a sequence through the function:

when i identify a protein family it has as a general feature: the function

so if i’m able to assign a sequence to a family/cluster, that sequence will inherit the function although we don’t have any structural information about it.

if the function is the general feature, someone has to define it and prove it in an experimental way.

people use to isolate the protein and characterize it in terms of function

Nowadays, when we say “function” we don’t just refer to enzymes, but in general also to biological processes in which proteins are involved, and, if available, the compartments in which the proteins are carrying out their roles.

Recalling the definition of functions in bioinformatics (....)

There is a sort of belief that in principle proteins start in bacteria and eventually, to justify the complexity of superior organisms, a hypothesis of gene fusion,

the organization of protein became more complex, more complex structure generation (multi-domain proteins) is inferred. This is what evolution did. So basically according to this hypothesis, our proteins are evolved, more complex forms of bacterial proteins.

in the pdb we have many fragments (especially with human proteins).

Our goal is given a sequence to be able to retrieve a suitable template, so that we can infer function.

**Not all GO terms can be transferred!**

template allows you to model the target and assign the function.

GO terms: if you have a protein from the chloroplast and you have the go term for a mammal, you can’t transfer that term about the taxonomy because of course it would not make sense.

Whatever is automatically done, may be not exactly accurate and may need curation (manual).

may be not exactly perfect

nb:

the sequence length of alignment is at least 250 residues and the percentage of identity is 30%.

distant homologues: for now, we don’t take them into consideration

i have a sequence that non have the function and the structure (i have the organism).

instead of using structure alignment now i have to trust structural alignment.

if i find a sequence that has a structure following the feature i’m asking for, I will have a template protein to align my target to.

provided that i’m able that the binding site is conserved I may also transfer the GO terms of the template to the target, and the problem of functional annotation is solved because I can say that I have identified the family to which my target belongs (ok we have stated this 2000 times now, what else?)

I’m not allowed to take a seq with 20% similarity with a certain template and infer structural similarity/same family belonging. I might be right, but I have no guarantee. So for these cases other procedures will need to be designed.

any sequence alignment methods fails when the sequence identity is lower than 30%

the more than the structure diverge the less the sequence is conserved.

heuristic solutions to the folding problem -> take into account some assumptions given by common good sense (protein modeling by comparison, building by homology)

think about list of problem that can be solved with the statistics of pdb:

Which problems can be solved using as the only source of information the PDB?

* protein folding
* functional annotation
* phylogenetics
* active site characterization/extent of conservation
* architecture of the active site (in terms of best *orientation* of side chains) → problem! when can I say that the orientation is conserved? The conformation of the active site needs to be **stable**.

how they are strictly oriented around themselves.

a protein has to be flexible but stable(in order to maintain the function )

if the protein is unstable, in vivo, it will start assuming disordered conformations, and so it will be destroyed by proteolysis/proteases/proteasome.

{unfolded and hydrophobic protein segments stick together and accumulate}

* principles of protein stability -> free Gibbs energy difference (from zero to about 50kcal/mol). Protein folding is spontaneous in polar solvents.
* drug design in silico , i can generate drugs molecules that have high affinity to the protein
* protein-protein interaction, protein-molecules interactions. The more we have crystallized complexes in PDB, the better we are able to derive general rules that govern how proteins interact with other entities.

i may derive at the level of pairwise interaction, all the interaction that makes a proteins stable(charge -charge interaction, charge non polar interactions, dipole non dipole interaction, hydrogen bonds, van der waals interaction)... for each type of protein folded i may understand which are the main interaction that make the protein stable,which are the interaction that are responsable of that specific protein folding.

What are the interactions that make the protein a stable entity? -> charge-charge, charge-dipole, dipole-dipole, charge-nonpolar, dipole-non-dipolar, hydrogen bonds, two non-polar molecule interactions (Van der Waals).

For each type of protein folding, I may understand which are the basic interactions that make the protein stable. Thermodinamically speaking, the biggest responsible is the hydrophobic effect, but what is responsible for that precise folding?

Protein prediction *ab initio* (given the sequence, predicting the 3D structure) is the biggest challenge.

Drug design -> generation of molecules with a very low binding energy to the active site. Foundation of molecular dynamics and mechanics, foundation of drug design… this knowledge derives from our production of protein structure information at the electronic level

protein protein interaction: complexity:possibility of the protein to interact to themselves or with the rna or dna…

Starting from the data in the PDB, we can derive general rules, from which we can implement an algorithm for predictions (?).

protein-protein interaction:two protein surface that interact.

g o o d

Van der Waals interactions are always about induced dipoles that contribute to protein stability.

Another very important problem to solve is: how are the proteins/mRNA expressions regulated?( we have to move on how the the expression of the mRna is regulated: cell transcription factories and interaction with transcription factor and how they recognize the binding site..)

a protein is stable in a polar solvent because of electrostatic interaction, including Van der Waals forces. Understanding electron densities and how they are in place gives me the knowledge about what are the forces that keep the protein together.

(this type of knowledge is possible to be derived by the electron density)

protein comparison:what can we learn from 3d superimposition.

our goal is protein functional annotation of protein sequences via structural alignment.

to be able to do the project we need to know:

* browse the pdb
* browse uniprot
* what is information
* GO database
* EC numbers vs GO terms (in other words, how to label function)
* how to do sequence alignment

protein comparison:

functional annotation in silico by homology search

we have the sequence and then we compare it with other sequences with the algorithm BLAST

First thing to do with an unknown sequence -> go and align it against a database to see if that sequence is already present in some database if it’s relatable to some other known protein.

Multiple sequence alignment =/= pairwise sequence alignment

sequence alignment pairwise is different from multiple sequence alignment!! this is performed with different algorithms:

which alignment methods do you know:

* blast, alignment method to do search all over the database
* lalign
* clustal W (omega) ( for multiple sequence alignment)

if the protein has experimental go terms, it is sufficient to look at the UniProt file (remember that we can understand whether the annotations are assigned experimentally or manually)

our problem is functionally annotating the protein: we are not so sure that even a similar 3d structure can produce similar function.

Template perfectly characterized -> variant with 99% seq identity -> may I infer that the proteins have similar function? No! The only 1% variation could fall in the catalytic site, disrupting the function. Before transferring the GO terms, I must check where the variations are located in the sequence of the protein. Sequence alignment has many answers, of which we try to find the optimal one. So maybe the alignment we have from the sequences alone is not sufficient to estimate functional annotations.

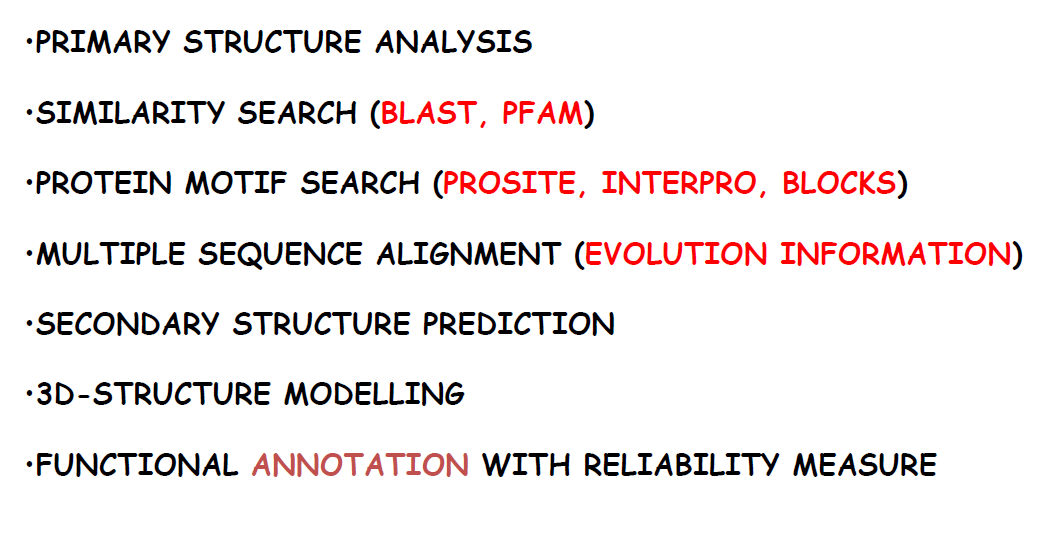
i have a protein template perfectly known from the structure and function point of view. now i have a target protein that has the 99 % of identity with the template.

do they have the same function? i can’t know the answer if i don’t know if the only different residue is not in the catalytic site of the protein.

Pfam is useful to assign distantly related homologues, but we’re not dealing with those now (we are in the yellow region of the graph).

pfam: database of models, so you go there and paste your sequence. you may assign distant homologues (with lower of 30% of identity percentage).

framework for functional annotation through the web:



proteins are heteropolymers, complex systems capable of autoorganization in the polar solvent and are social entity.

When it comes to the problem of folding, we can prove theoretically that proteins are **frustrated** (actually it’s a real physics term) systems, because there are too many different tendencies constraining the backbone in solution, the too many tendencies are due to the different physico-chemical properties of the residues.

Small hydrophobic regions flanked by charged regions -> charges and solvent will have a local effect in letting the region organize itself

the interaction with the polar solvent will have its local effect in letting this region organize itself : the protein is forced to have a specific configuration in solution , it is frustrated.

Local interactions promote the hydrophobic effect.

protein folding problem: if i have a sequence is there a method that allows me to know its tridimensional folding? due to the generation of alpha elixes .

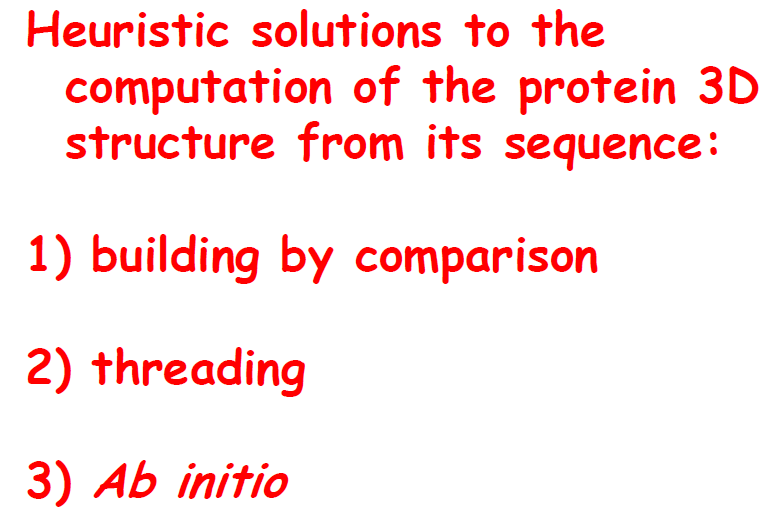
secondary structure: conformation at the level of the backbone.

It doesn’t make sense to talk about secondary structure of the residues.

Lateral side chains may or may not contribute to overall protein stability (e.g. loops can be substituted, enlarge or shrink and they don’t affect protein stability that much).

We may also understand which are the residues that are mostly responsible for protein stability.

the protein folding is an enourmus and relevant process for the cell cycle, since cells rely in the protein function. A protein is functional when it is stable in the context of a cell.



how can i compute the protein tridimensional structure starting from the sequence?

* building by comparison -> quite successful compared to the other two methods. => sequence comparison. If similarity is 30% or above, I can apply it.  
    
  {{{There is a sort of international experiment called Critical Assessment of protein Structure Prediction methods (CASP) that is carried out every two years}}}

we still have the problem that so many proteins don’t have the template to be compared with. so we can also use other methods.

* threading methods => if I have a similarity of less than 30%
* ab initio => I have no proteins to compare my target protein, I try to do a simulation in silico of what the forces among the residues will be. starting from all the residue sequence, and from what i know about stabilizing effect , and applying the pairwise interaction we end up with the protein structure.

we align our target to the pdb database:

sequence comparison : we have to pay attention to the different residue and where they are localized.

our project is :

try to understand how to apply comparative modelling? How to annotate functionally something that is out of our comparative modelling procedure?

(we will work with protein that have an high percentage to sequence identity in order to predict the function.)

comparative model:

start with a target sequence, try to recognize why the sequence is worth to be modelled (why is the problem biologically interesting for us)?

and then we should have an algorithm that works from a sequence alignment and then on the basis on the folding of the template i will assign a three-dimensional structure..

Modeller: program, input -> seq alignment, output -> structure of target on the basis of the structure of the template.

what modeller does? modeller is used to check a transplantation of the coordinates of the template to the target, allows you to map the electron density of the template to the target.

model takes as input the sequence alignment of the target.

We also have to evaluate the model that we obtained! is this a protein?Does it have the characteristics of a protein? => Ramachandran plot of the model, distribution of the dihedral angles with respect to the pdb database distribution (so we can see if it’s something plausible). (--> PROCHECK)

each blocks is a series of operation :

* selection of the template
* alignment of the target sequence with template
* modelling of the target on the template
* evolution of the model

which is the limiting step ? the input, the alignment of the sequence. have i used the right alignment as input of the algorithm? the alignment is the end user responsibility.

alignment of the target sequence with template

If the model is not good, we have to question the quality of our alignment.

threading: procedure that relies on different models and tries to understand if a sequence fits a model (which is the model in which the sequence fits best?). Sequence-model fittings are associated to a certain *score*, the higher the score the fitter the model.

I can be sure about the backbone, but I don’t know if the dihedral angles have been computed correctly for the side chains (in other words, if my protein is actually a protein).

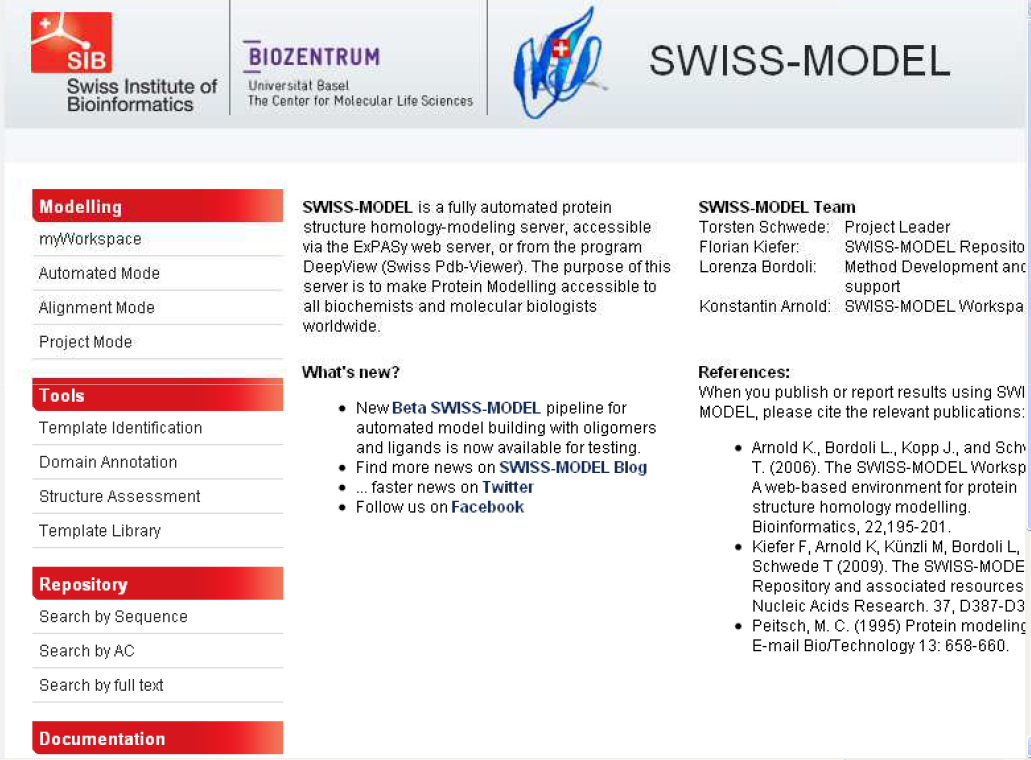
what i will do with the model?

with building by homology if my template is at least 60% identical to the target i’m using the structure that i derived for docking and..???is it equivalent to low resolution NMR? yes so it says on the slide

Below 60% -> used for supporting site-directed mutagenesis

Below 30% -> threading region, we may use the model to refine NMR structures, finding binding/active sites, annotating function by fold assignment (HMM)

SwissModel ->???



ModBase -> models generated automatically everytime a sequence enters UniProt on the basis of structure in the PDB. Many times, this automatic procedure compute their own structure (precomputed structures) since the sequence identity is lower than 30%.

-------------------

Tutorial -> protein modelling

1) 3D modelling of the Laccase 4 from Pleurotus eryngii

file uniprot:

http://www.uniprot.org/uniprot/B0JDP9

>tr|B0JDP9|B0JDP9\_PLEER Laccase OS=Pleurotus eryngii GN=ery4 PE=2 SV=1

MAVAFIALVSLTLALVRVEASIGPRGTLNIANEVIKPDGFSRSAVLAGGSYPGPLIKGET

GDRFQINVVNKLADTSMPVDTSIHWHGIFVRGHNWADGPAMVTQCPIVPGHSFLYDFEIP

DQAGTFWYHSHLGTQYCDGLRGPFVVYSKNDPHKRLYDVDDESTVLTVGDWYHAPSLSLS

GVPHPDSTLFNGLGRSLNGPASPLYVMNVVKGKRYRIRLINTSCDSNYQFSIDGHAFTVI

EADGENTQPLQVDQVQIFAGQRYSLVLNANQAVGNYWIRANPNSGDPGFANQMNSAILRY

KGARNVDPTTPERNATNPLREYNLRPLIKEPAPGKPFPGGADHNINLNFAFDPATVLFTA

NNYTFVPPTVPVLLQILSGTRDAHDLAPAGSIYDIKLGDVVEVTMPALVFAGPHPMHLHG

HSFAVVRSAGSSTYNYENPVRRDVVSIGDDPTDNVTIRFVADNAGPWFLHCHIDWHLDLG

FAVVFAEGVNQTAVANPVPEAWNDLCPIYNSSNPSKLLMGTNAIGRLHAPLKA

Why do I care about the laccase of a fungus?

1)try to understand what we have in uniprot for this protein

Submitted name: **Laccase**

Experimental evidence at transcript level:Protein existence

This indicates the type of evidence that supports the existence of the protein. Note that the ‘protein existence’ evidence does not give information on the accuracy or correctness of the sequence(s) displayed.

type of evidence of the existence of a protein : the sequence can’t be taken as granted

given the gene annotation the first problem is to locate..

de novo assembly: (assembling a genome from scratch, without a reference)

the location of the loci (region that are transcribed)

when i’m translating a mRna there can be mistakes also in the translation of mrna.

While it gives information on the existence of a protein, it may happen that the sequence slightly differs from genomic sequences, especially for sequences derived from gene model predictions.

how many types of evidence do we have in uniprot?

In UniProtKB there are 5 types of evidence for the existence of a protein:

1. Experimental evidence at protein level (when a protein has been purified and it has been analyze with x ray diffraction)

2. Experimental evidence at transcript level (you can sequence the transcript)

3. Protein inferred from homology

4. Protein predicted (you may have automatic evidence about the protein)

5. Protein uncertain

3,4 and 5 are not experimentally

The value **'Experimental evidence at protein level'** indicates that there is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies.

The value **'Experimental evidence at transcript level'** indicates that the existence of a protein has not been strictly proven but that expression data (such as existence of cDNA(s), RT-PCR or Northern blots) indicate the existence of a transcript.

The value **'Protein inferred by homology'** indicates that the existence of a protein is probable because clear orthologs exist in closely related species.

The value **'Protein predicted'** is used for entries without evidence at protein, transcript, or homology levels.

The value **'Protein uncertain'** indicates that the existence of the protein is unsure.

**GO - Molecular function**

* [copper ion binding](https://www.ebi.ac.uk/QuickGO/term/GO:0005507) Source: InterPro
* [oxidoreductase activity](https://www.ebi.ac.uk/QuickGO/term/GO:0016491) Source: UniProtKB-KW

Unreviewed proteins that have evidence at transcript level follow an automatic pipeline for automatic annotation.

E.g. the oxidoreductase activity is inferred from .

there’s nothing about the localization of the proteins.

Gene Ontology (GO)

The [Gene Ontology (GO)](http://www.geneontology.org/) project provides a set of hierarchical controlled vocabulary split into 3 categories:

* Biological process
* Molecular function
* Cellular component

UniProtKB lists selected terms derived from the GO project. The GO terms derived from the ‘Biological process’ and Molecular function' categories are listed in the ‘Function’ section; the GO terms derived from the ‘Cellular component’ category are listed in the ‘Subcellular location’ section.

**electronic annotation**

**BRENDA ->** database that is cross-referenced with uniprot, contains information about the enzymes and their functions

|  |
| --- |
| [1.10.3.2](http://www.brenda-enzymes.org/enzyme.php?ecno=1.10.3.2&UniProtAcc=B0JDP9&OrganismID=4910) |

^ example of an EC (enzyme commission) number.

Hierarchical level of organizations in clusters of enzymes -> each number in the EC number gives me information, respectively, about:

* major cluster
* major activity (chemical categorization)
* type of substrate
* type of product (specificity of the reaction)

Question ----> how is the EC number related to the GO terms for a given enzyme?

these numbers are known for each enzyme.

GO terms -> molecular function, biological processes, cellular components... EC number provides a very narrow information about a protein and its interactions.

four digit: gives you the specificity of the reaction .

How can an uncharacterized protein have an enzyme numbers?

we have a proteins with a level of evidence at the transcriptomic level, so how can it be possible that we have detailed information about the type of reaction that they perform?

ec number: categorizes univocally the function of a protein ( an enzyme)

The Enzyme Commission number (EC number) is a [numerical classification](https://en.wikipedia.org/wiki/Numbering_scheme) scheme for [enzymes](https://en.wikipedia.org/wiki/Enzyme), based on the [chemical reactions](https://en.wikipedia.org/wiki/Chemical_reaction) they [catalyze](https://en.wikipedia.org/wiki/Catalysis).] As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme.

the second digit tells you about the function of the enzyme

Example -> if first number is 1 (oxidoreductases), second digit is the type of donor, third digit is the type of acceptor (of electrons in the redox reaction), fourth digit points to the specific enzyme.

What are laccases? -> copper-containing oxidase enzymes, found in many plants, fungi and microorganisms. They act on phenols, cross-linking by means of single-electron transfers. => formation of lignin.

so this protein is an [oxidoreductases](https://en.wikipedia.org/wiki/Oxidoreductases), [Acting on diphenols and related substances as donors](https://en.wikipedia.org/wiki/List_of_EC_numbers_(EC_1)#EC_1.10_Acting_on_diphenols_and_related_substances_as_donors), [With oxygen as acceptor](https://en.wikipedia.org/wiki/List_of_EC_numbers_(EC_1)#EC_1.10.3_With_oxygen_as_acceptor).

this database performs an annotation mainly based on detection of sequence similarity.

so there is a template for this sequence that allows you to predict the function of the protein

if i know the ec number i know something about the chemical behavior of the protein.

enzymes with the same ec number should have the same function (=> same GO terms for molecular function).

ec number is a little bit more general then the go term.

There is a database called EC to PDB, nowadays for the same EC number there might be different structure and functions.

ex: try to evaluate via comparative modelling if it is correct or not.(since it is a partial annotated protein).

Is it legitimate for the target sequence to inherit all these GO terms? Can we find a template to evaluate this?

we know that the Length of the protein is533

can we prove from here that this protein has a template??

PTM/molecular processing ->

Signal peptide is a specific sequence at the n terminus, and may help to leading the protein in a localization inside the cell, but after the transfer the signal peptide is cleaved!

In the case of our protein, the signal peptide has only been predicted according to rules.

this signal peptide has been just predicted.

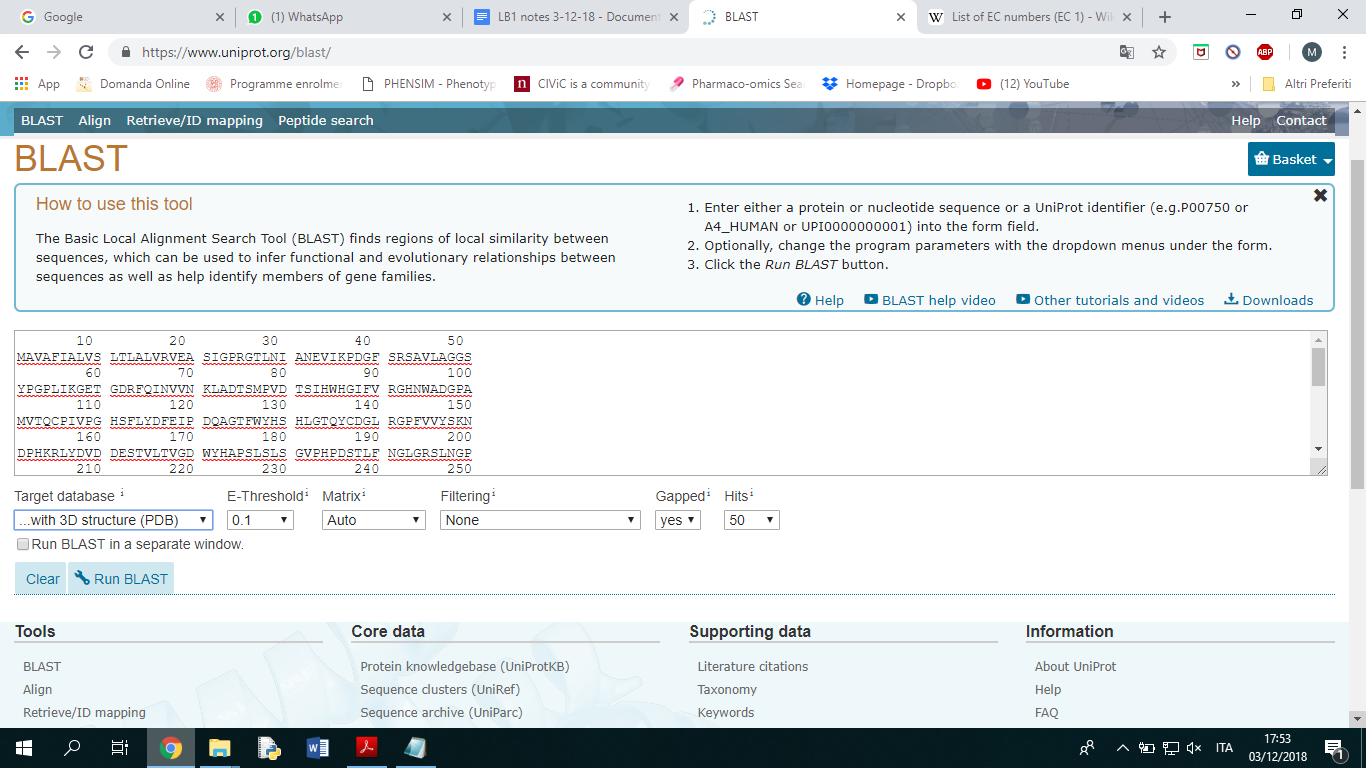
so the n terminus will not be aligned to a PDB structure ---> since our goal is to align our sequence to some template, we should not consider the amino acids belonging to the signal peptide when performing the alignment. These extra 20 aa can alter the alignment.

so cleave the signal peptide before perform the alignment because the presence of this residue can influence the alignment.

Next step -> BLAST against the PDB database (from the UniProt interface)

looking for the template on the blast

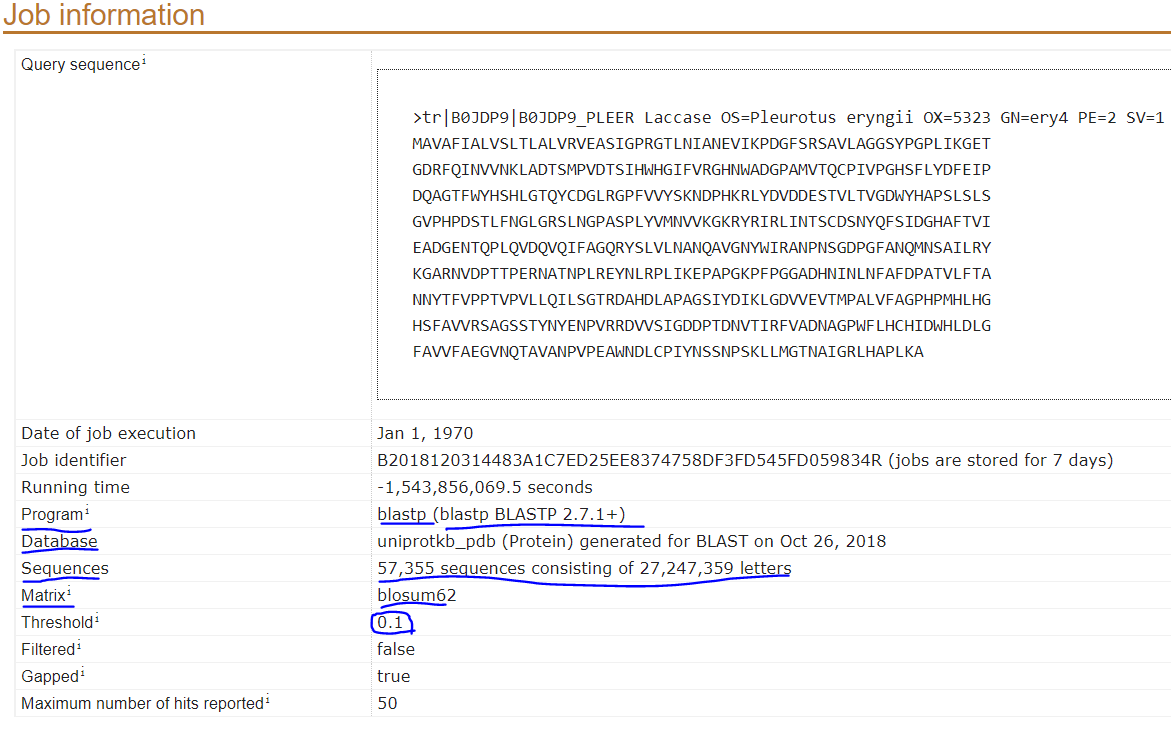
minimum e -value to the alignment to be accepted: in this case we choose 0.1 as E-threshold



For the first sequence alignment (aligning with BLAST our target sequence against all the UniProt entries with a PDB structure associated) we needn’t remove the first 20 amino acids. It will matter much more when we will input the alignment for Modeller.

Belonging to materials and methods in the report: it is really important to mention the relevant details of the BLAST job (matrix used, version of BLASTp, number of sequences!!!).

methods:



In front of different possible template sequences, always pick the most reviewed/annotated one. The more the sequence identity %, the better.

* the most review files
* then i will look at the sequence identity (62 and 58 are in the same range)

laccase 2: alignment has to be reviewed, E-value:0

Experimental evidence at protein level with 1GYC structure

We find a protein that is reviewed, LCC2 from Trametes versicolor. (UniProt entry Q12718)

*task:*

*read the article : biological motivation*

*LB1A\3D-modelling\Biological motivations\J. Biol. Chem.-2002...*

If the protein LCC2 needs four coppers to function, we need to understand where they are located, where the catalytic sites are.