Content of the lecture → Protein family/Report writing/Running Modeller/Discussion of the Laccase paper

Why perfect the art of modelling procedures? -> functional annotation.

Have clear what a protein family is in bioinformatic. (how are gene families and protein families related?)

Proteins and genes are mixed up concepts in medical doctors’ field. Correct? If we say that 1 gene = 1 protein, it might be, but in reality the concept of “gene” is much more blurred than that! So we can’t mix up gene and protein concepts.

Problem => functional annotation => biggest problem of a database => each file has a series of associated features that we call annotations => they can come from experimental/computational approaches or inferred information from similar proteins

Protein/gene differences => protein family not a gene family

=> relation between gene and gene family?

=> proteins are often named genes in medicine => gene concept is blurred

=> there might be some overlap between gene and protein

=> but there are isoforms

=> we know that depending on the tissue we have different levels of protein expression

How are all files in the biological databases annotated? How was the annotation done?

Often the annotation procedures that appended information to the different dbs are automated. It is important then to determine whether this information are reliable or not, and in order to do this we have to understand how the process works.

Much better to understand the process from scratch and then can evaluate the annotations present in a file

Proteins investigated in terms of their behaviour in solution, biomedical/biotech characterization (experimentally characterized) -> proteins that *exist*

VS

Proteins that are predicted, inferred by homology or by seq similarity, or totally unknown -> proteins that have not been “touched” experimentally (have not entered in contact with any kind of experimental procedure) -> it’s very questionable whether the protein is a real entity or not.

Interpretation of genomic data → sometimes we can be led astray if we can’t place a precise distinction between what is accurate and what is not (between, for example, exons and introns -- a process that is automatic, so if exons and introns are predicted wrongly, does the sequence of the translated protein make sense?)

If the protein existence is predicted on the basis of seq similarity, inferred by homology or in any case has no experimental to prove its existence => questionable if its a real entity

When we interpret the genomic data we may be far away from reality => its a problem

If we are wrong in positioning the distinction between exons and introns as has been done in an automatic way then the sequence of the protein derived might be wrong

Behind functional annotation there is the big and unique concept of protein family

Family => cluster of proteins that share an omology higher than 30% for more than 250 aligned residues

Series of objects called proteins =>function rather independent of the organisms => it was noticed that proteins have the same function in organisms far away related

It has been demonstrated through experimental data that proteins from distantly related organisms were performing the same molecular function (which is strictly related to the protein structure) => FIRST i cluster according to function (main feature of protein family)

EC clustering => old

Today we define a protein family according to the concept of molecular function.

As long as the superimposition of the backbone is within 2 amstrong of RMSD in principle to similar structures correspond to the same function.

As long as the sequence alignment length after structure alignment is 200 residues with a seq id of 30% we are describing the whole cluster

We also have to consider the distantly related homologues => proteins in the yellow region of the famous graph that have the same function

**Protein family => set of proteins that share the same function**

We can understand the role of structure => with seq alignment we can say that whatever statistics you are using for aligning you cannot apply it when the seq id is below 30%

Everything depends on how you start the procedure of alignment

To retrieve homologues you cannot use only alignment, but also need stuff like HMM

We are restraining ourselves to the region where sequence similarity implies also structural similarity and the same function => not the whole situation

Annotation => predict the function from the sequence

=> most straightforward one is to revert the experiment related to the pairwise structural alignment on the pdb

We have data in the PDB

Sequence alignment after structure alignment

=>very general clue to go backwards

Provided that the target sequence finds a template in the family with 200res/30%id

=> i may inherit the function (after putting it in the family)?

FUNCTION <--> PROTEIN FAMILY

same (molecular) function <--> same protein family

NOTE THAT a protein family includes also distant homologues (less than 30% identity) that share the same function!

The field of proteins is quite older than the field of genes and messengers.

Independently of organisms people were able to isolate proteins with same function, even if the organisms were very distantly related. It was also demonstrated. First proteins were clustered according to function, and they were identified through EC number. then after data accumulation, structures were compared: it was observed that similar structures have the same function. generally speaking they had 30% homology.

Cluster of protein that share the same function → Protein family

if sequence homology between 2 structures is lower than 30% cannot infer structure (and therefore cannot infer function)

From similarity >30% I can infer structure, but from similar structure I am not necessarily able to infer similar function. So the pairwise comparison PDB experiment gives me a general rule that is necessary, but not sufficient, to do a traceback operation in order to figure out function from sequence

Molecular function is a broader concept than the EC number (it includes it but it’s not limited to it)

So

Protein family, functional alignment, BUT WAIT A MOMENT

In order to assign function we first want to know the structure

In life there are always complications => same function can be performed by different structures => protein superfamily

there are families whose function is proven to be supported by different structures

Biology cannot be reduced to a scheme => the more you go inside the more data you find, the more you have to adjust your point of view to accommodate the increasing complexity resulting from the data

Protein SUPERFAMILY -> same functions achieved by different protein families

## WRITING THE REPORT

(something similar to an article)

Why do I care about the protein that I am taking into consideration?

What is the most correct procedure available to perform functional annotation?

Why should I ? I want to understand how a db works, but first i want to understand the most correct procedure to functionally annotate as a bioinformatician

We have to follow a [frame](https://www.dropbox.com/s/z0p14wyx0wvk89t/How%20to%20write%20a%20report_extended.docx?dl=0) => a typical frame in which we may start editing our own report

Report -> one or two columns, no matter.

Title -> it is indicative of what we want to do. The title should answer to the question “I did the following”

Author and affiliation -> *each* person is responsible for whatever he or she will present, so in the projects there should be just our names, no matter if we work in small groups or what.

Introduction -> biological interest, state of the art.

* What are the biological features of the template (rather than the target sequence) we are basing our model on?
* What makes the protein family in our work relevant?
* ACTIVE SITE <---> FUNCTION | Depending on the protein I may have to describe an active site, but I will need to zoom in the *template*, not the *target*).

After all the computation is done, we need to be able to understand if the situation is exactly the same when it comes to protein functioning (so the active site of the target must have a very similar structure to that of the template)

Methods-> the resources like databases (UniProt, PDB) ! how do you mention them? You have to put the [www.etceteraetcetera.AHHH](http://www.etceteraetcetera.ahhh) in the report → either list of reference or quote the site in square brackets

In general how to do quotations/citations -> either you put it into round brackets in the text, or you put numbers in square brackets and then at the end of the report you put a list of all the corresponding papers/resources/sites/references

Computational methods → blast jce modeller, visualization software (rastop, pymol, rasmol, chimera) Also have to quote the *original article* related to the program (usually each program has an article related to it, which is used to “present” the software to the scientific community), again with numbers or with the first authors according to the standards of a bioinformatics journal.

Paragraph -- template selection:

**MIND THE SIGNAL PEPTIDE**! pdb doesn’t always contain signal peptides(I think it depends on what sort of signal peptide it is, not all of them are cleaved) so aligning along with the rest of the sequence it will lower the score of a good template.

how did you choose them and why did you select them. Attach a figure related to the alignment Target-Template aligned with laligne

What motivation in order to select one template other than some other template?

Figure -> target/template alignment, which is the INPUT TO MODELLER! So it’s like a sort of pipeline

**Lalign** -> algorithm of alignment that unlike BLAST does not return *only* the *best* scoring alignment for a given pair of proteins, but also return suboptimal-scoring alignments! So even if the best scoring alignments lead us to scrambled/horrible modeller results, we have some backup plans (or at least I think, but I don’t dare asking because *of course* we didn’t explain it yet so we are not allowed to *know*)

Coverage = % of residues that are “covered” by the template considering the seq length of the target.

Paragraph -- Modeller at work: modeller outputs 5-10 models, you have to choose just a few ones to carry on along the pipeline. You have many different possibilities depending on on the different parameters we use to run the simulation. This parameters are user-set and so depend on our knowledge!

Different parameters -> different outcomes! So instead of adjusting parameters 852 times, we just allow for sub-optimal solutions.

Each solution that we carry along the pipeline must be validated statistically, since there is no experimental proof it is very important.

Target annotation -> we have to check if our model/structure really compares to the original structure! At least we have to superimpose target structure computed with modeller with the original template structure, and we have to zoom in the active sites/binding sites and realize whether it is possible that the active site is really concerned at the level of the 3D structure.

*composition and architecture* of the active site

***Iff***I establish that there is sufficient similarity and “comparability” at the level of the active site, I can infer same function and therefore *transfer* my information from template to target.

provide **RMSD** along with some comment.

realign the sequence according to structural superimposition → compare the sequence alignment performed at the start (modeller input) and the one derived from the structure superposition (final output)

**comparison with swiss-modeller**

In principle, one can go to swiss-modeller (database of …?)

SW-MOD -> place to go if you don’t have enough time to do your process at home

Possible to obtain a draft for the structure of a protein but it takes a while

Figure 4 -optional

zoom on the structural superimposition, particularly on the active site → graphical representation

13th December extra hour of class to study visualizers

Blas blas bla

LB1-A/tutorial/Modeller/Homework → has fast format sequences of proteins we are supposed to model

1. title => all the general keywords referring to what you are doing
2. author and affiliation => each person is responsible for whatever he/she will present
3. introduction => since we are thinking in terms of templates for a certain protein family we have to emphasize the important features of the sequence we want to model and the corresponding protein family

characterize and describe carefully whatever makes the template functioning => WAIT A MOMENT => depending on the protein i may have to describe an active site => read the article to get the needed information (not cut&paste) => characterization of the active sites in relation to the protein function => WHY? cause later on we need to understand if we are exactly in the same situation (otherwise the protein isn’t functioning)

1. methods => 1st of all databases => mention uniprot, pdb for sure => you have to mention the release and a www reference in brackets or to use numbers for the final references in which you also quote the site so that everybody can retrieve the info
2. computational methods => BLAST, modeller, procheck, JCE, a visor(rasmol, rastop, chimera, pymol) GOOD => friends dear, you also have to quote the original article => either with numbers for the references list or quote 1st author etc according to the standard rules for reference papers in bioinformatics => we need a template => important to put references in the right way=> in the list of references we should also put the title of the quoted paper
3. she suggests that the schemes go on with a paragraph and title in which you explain why you ended up with the template/s => figure in which you show the sequence alignment of the target to the template => best way is to once you selected the template with blast you download the sequence of the template and align pairwise with the target with lalign => sehr gut (but mention what you do) you are generating the input that later on you give to modeler => if you have a sequence with a signal peptide either you use it on the template too or you take it out (highly recommended, maybe also mention that you do it) => ON PDB THE SEQUENCE IS WITHOUT THE SIGNAL PEPTIDE (lost during life/experiments) no matter what i have to use the sequence of the original template => if by chance there is only a long fragment (coverage length => choose high coverage => ratio of residues covered by the template to the sequence length) unfortunately uniprot doesn’t compute coverage
4. section devoted to modeller => modeller releases a table that gives the best model => why different models? in life you never get a single result => THE PARAMETERS STRONGLY DEPEND ON OUR KNOWLEDGE => problem of standard deviation => associate error to measure => you should decide how many different solutions to keep, and statistically evaluate them
5. Last important story of target annotation => check if our model really compares to the original structure of the template => superimpose the modeller computed structure with the original one and zoom on the active sites and realize whether it is possible that the active site is really conserved => no guarantee that we end up with the right torsion angle of the lateral side chains UUUNLEEEES we have a very high similarity => each computer structure is or not a stable structure (measure whether the stability is pertinent to the model we did)
6. functional annotation => i can derive all the GO terms of the molecular function and transfer them from the template to the target => also a few comments on the problem of biological processes and cell compartments , also check if the protein is in a specific db on biological processes
7. rmsd with some comments, are the alignments compatible? comparison between figure 3 and 1????
8. COMPARISON WITH SWISS MODELER => EMPHASIZED => SHES NOT HIDING HERSELF => one can go to swissmodeler in switzerland => database of structures precomputed and stuff => same workflow is computed => if you dont have time to do it at home => AUTOMATIC PROCEDURE => depends on what you want to do with the model => implements modeller, not the original version => they keep it secret => possible website to go to if we need a draft of our protein => takes some time (remote)
9. fiberfour???? => you either zoom on the superimposition and show to her that the active sites are exactly the same

Fasta format of a specific sequence that is laccase 4 from pleurotus

means that you have a line that starts with > and then whatever you want , > is important

FASTA parsing -> FASTA files always start with the “>” symbol.

program of translation applied to the gene => sequence is deposited on the db

pipeline that identifies via blast => optimized for db search => it works based on comparison of words => its fast but basic => statistics attached to blast to define the E-value,score etc => t

basic because the k-uples used for search startup are a quite strict condition, we need to find an exact occurrence of a k-uple in a sequence of the database in order to extend that alignment and potentially include the result in the output. A bit harsh so we have the risk of leaving behind some potential distant similarities, but at least the algorithm is faster.

Almost aaany bioinformatics algorithm is heuristic (= *not exhaustive*)

the statistic of blast is based on extreme value statistics

density function (??????????

we have the infrastructure

i have sequence => i t h r o w my sequence against a database for example the BLAST interface from UniProt

guys should we ask her about the fact that BLAST from UniProt gives some result and BLAST from NCBI gives different ones? (oops sorry ida) [i disrupted your braketz it’s fine i am too slow to write what she blabbers anyway]

she loves uniprot and pdb

mlmlml

guys remember the question is ***how to write the introduction -> read the right article to justify why you want to model that specific sequence, and what is the protein family about***

threshold status:very stringent

Retrieve information for the introduction: go to PDB, because here you will find the paper regarding that structure. In the end the answer however she might dislike it is always l i t e r a t u r e → PUBMED

from pdb get paper

from paper get structure

how get paper? not money, use scihub => pirate articles | also the UNibo proxy service can work for some articles | pirate is cooler | nobody wants to go to jail

1.9 A resolution => good resolution?

## Discussion of the Laccase paper & additional pointers on the report

Quote articles

The lines I write (no more than 10-15 lines) have to be short but *very* informative!

Relevance of our protein => read abstract => derive information from it, summarise it, do not copy and paste.

Laccases are oxidases and there are two types: high or low redox potential.They belong to the family of blue multi-copper oxidases. Laccases are involved in degradation of lignin. It is interesting from the point of view of paper industry(pulp delignification) and bioremediation of water and soil, due to the wide range of substrates. <- !

Lignin degradation => important as an industrial process => paper industry etc

blue multi-copper oxidases -> the proteins are just scaffold for the metal ions that are the proper actors of the redox process.

the coordinating amino acids and their exact positioning/orientation are key to the properties of the enzyme. Even a small difference can alter the electron density of Cu.

The active site can allow the formation of coordination bonds of metal ions: the side chains are mainly His. If they are missing the copper site will be altered and it is impossible to tunnel the electrons in order to degrade lignin

Correct orientation of the residues that coordinate the copper --->>> correct tunneling of the electrons in order to achieve the redox.

If I don't have histidines i will not be able to keep in place the coppers

Coppers: 3 (trinuclear cluster) + 1

It is up to us to emphasize what we think is really important

From paper derive information on the active site structure

laccases cannot synthesize lignin can it? it is involved only in the remodeling of lignin in wound responses, but that doesn’t mean it *synthetizes* lignin pls answer me

little review on laccases => try to derive some info

What leads to high or low potential? To find out compare structures of laccases from Cc(Coprinus cinereus) and Tv (Trametes versicolor), and also with more distant homologous like ceruloplasmin and AO.

Attention to which organisms’ homologues the protein is being compared with. Here we are comparing fungi, mammals (ceruloplasmin)... there is an evolutionary relation to it

Figure one→ shows how the electron density allowed to identify every amino acid with a good level of confidence, even if they weren’t in accordance with the primary sequence.

Figure 2 → ribbon representation of the protein structure, domains are color coded (portion of protein that can fold in a definite structure, has its own independence)

glycosylation sites -> specific residues can be glycosylated.

disulphide bonds -> redox environment: there is a strong influence on protein stability. Stable in oxidizing environment. (We can seen the disulphide bonds in the protein because is in its oxidised form)

DOMAIN => portion of the protein that can fold in an “autonomous” way with respect to the rest of the protein => theoretical definition

disulphide bonds => strongly dependent on the solvent redox potential(environment)

=> stable in oxidizing environment

Our template

**https://www.rcsb.org/structure/1gyc**

propanol to mimic the reduced center

glucosamine mimics(it doesn’t mimic in this case because they are the native glucose groups, they weren’t eliminated during the purification process) the glycosylation

Try writing introduction for this case study

things to include:

* what database, what is the database dimension, version of the db
* parameters
* substitution matrix used
* everything necessary to make the experiment replicable, as it is always good practice in articles

METHODs =>also point out the way of running blast => highlight the db towards which it runs (UNIPROTKB\_…..Generated…), E threshold value,

Sequence alignment of query-TvL: extended informations on gaps and matches. Is the sequence of the TvL correct or does it contain mistakes? What effect does it have on the alignment.

Careful with Modeller input → if some parts of the sequence have no structure Modeller has no reference for that part and it will give no structure (a squiggle cit.) to the target protein.

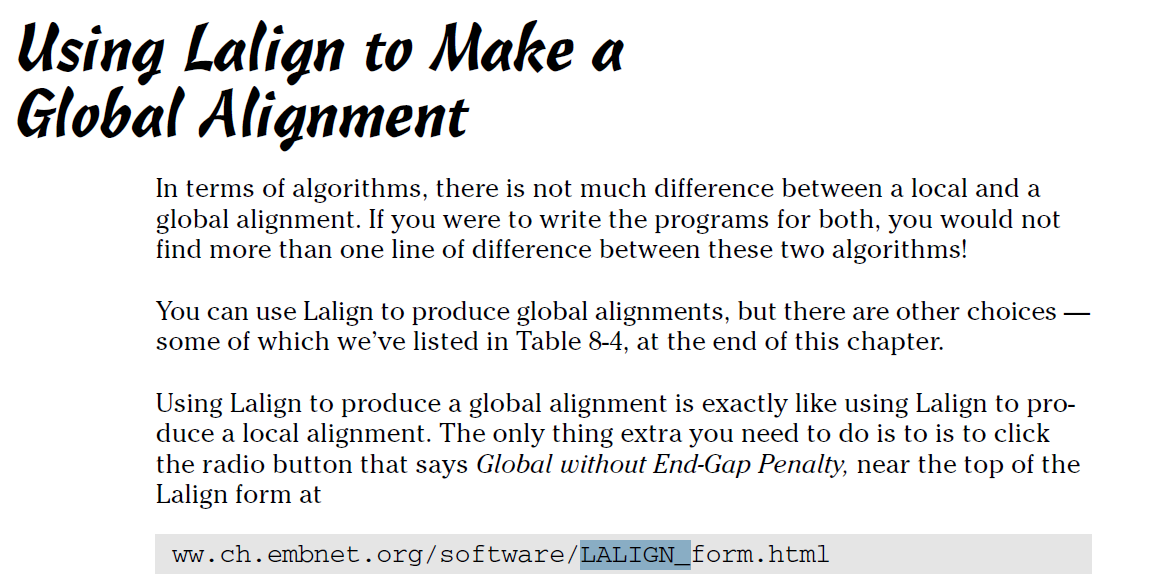
We already have the sequence alignment with our query => all ids are highlighted; gaps; is this the same sequence we have in the PDB?

LALIGN → another alignment tool for global or local alignment from Expasy Y should we use it??\*\*\* refer to below It has its own paper

modeller is exactly computing what you give it => prepare the input; to be fast we have to prepare it in the best possible way(exclude what can be excluded)

Solution -> download fasta and run lalign does laligne exclude the bits that have to be excluded for us??

\*\*\*

<--- some Lalign trivia if you are curious, kindly brought by ferré-senpai <3 *secret word of the trentino massoneria*

*-- About default parameters --*

They are good in a very high percentage of the cases, so this is why you are proposed those. If you are aware of what you are doing, you are encouraged to adjust them to your needs

heuristic rules to select the templates: what to prefer? coverage or identity? (**identity**)

building by homology, when I want to compare two proteins and want to understand (or think I have understood) -> whatever is identical has good chance to have same orientation with resp to the backbone. We need to be led by seq identity. In this way we have highest possibility of ending up without misleading stuff that leads us to wrong conclusions/bad modelling. This errors could eventually affect the active site and completely disrupt our model, so we must minimise them.

One rule of thumb is to go for highest identity (ok, as stated before). Few percentage points of difference in coverage don’t really bother us. In case we don’t care too much we may enter into modeler with a MSA!!! (MORE THAN A TEMPLATE c: ) As long as the templates are not generating noise!!!

you may even enter into modeller with a msa as long as it’s not generating noise => multiple templates (inputs)

We can take 4 model and make a pairwise alignment (no, it’s multiple sequence *alignments* in the sense that i do a *pairwise* alignment with query vs one of my templates and I align pairwise manyfold).

**Task for tonight** -> read the article about this other laccase. Can we use it as template?

“Kinetic analysis and structural studies of a high-efficiency laccase from Cerrena sp. …”

PDB code -> 5Z1X

Ex to do tomorrow morning -> compare with rasmol or sth the two protein structures that we found what did she just say about tomorrow morning??? do we have to be here at 10? (ok good yes she said quarter past ten)

Is the structure of this new template superimposable?

“So we have to read another article”

I want to learn from her so bad how to speak *italics* like she does

I mean she can *speak i t a l i c s*

you can tell the words she is saying, if they were written, they would be *italics* with no doubt

Original template of the family -> TvL (the laccase from Trametes versicolor)

Tomorrow we decide how to compare this structure. Two templates to have a model.

We have for sure the original template for the family. The higher sequence identity is the main value, then the coverage, and then the others things.

Look at all the possible ligands.

## How to run Modeller

Software (better to say *program*) that will allow us to built a structure given the alignment between the template and the target sequence

The templates are in the dropbox LB1-A/tutorials/Modeller

To use modeller need 3 files:

1. pairwise alignment in a specific format (pir, the PIR format is a format for alignments, more info in the manual found at https://salilab.org/modelle/9v6/manual/
   * PIR has a line >P1; protein\_name
   * structureX:pdb\_file\_name(has to be the same as the pdb code otherwise error arise):n° of first residue of alignment: n° of last aligned residue: other\_fields
   * actual\_sequence ends with a \*
   * then information for second protein
   * >P2;protein\_name
   * EXAMPLE:  
     >P1;protein1  
     structureX:5fd1.pdb:1:A:106:A::(other fields delimited by colons, no matter if empty)::  
     seqseqseqseq\* (\* = stop codon)  
       
     >P2;protein2  
     sequence::::::: (all empty fields, we should leave empty)  
     seqseqseqseq\*  
       
     see manual for other specifications and questions about the PIR format  
     <http://salilab.org/modeller/9v6/manual/node441.html>
2. pdb template
3. modeller script, written in Python → allows to set parameters and to run modeller
   * premade form, compile with names of file (pir and pdb) as in your computer and names given to proteins in PIR file, max n° of models (start low! no more than 5 for now)
   * with the help of the manual can make more complex script

**All 3 files should be in one folder**, then run Modeller by typing in the terminal

$ python model.py

Output → there will be many outputs files, we’re interested in the computed 3D structure (PDB format!!! also we may be interested in the log file, we will understand it better when we use modeller in more detail)

How to change our alignment in a PIR format?

BLAST search of target to find the right template -> find template -> align with LALIGN -> output of lalign is NOT IN PIR FORMAT! We should edit the format starting from the LALIGN output which is going to be a text file anyway.

Script -> lalign2pir.sh (premade by the tootorz) we can write our own python or bash script to

change the format. If trouble just use the premade skrypt

So she wants us to turn the report in within the next few weeks?? This is the level of depth with which they cover this topic?? no the report we should write now now is something meant for the next few *hours* and it’s about laccase for everyone, our report for the project needs to be about one unique protein that we annotate