First of all, one a n n o u n c e m e n t, and the announcement is as follows - two dots:

timetable -> in february Winter School will be shifted to week 18-22 feb. Oral section of Casadio exam is shifted from 18th to 7th February. Last section on 28th Feb.

We still have the week with Allegra Via (<3) in week 11-15 Feb

Sequences for the project are just 3 and not anymore 4.

The deadline to turn in the projects is the 31st of January, if you can turn them in sooner do so, in this way you can have access to the first oral session

what is modeller?

It is a program for structural modeling by satisfaction of spatial restraint

before the exam it would be a good idea to read around the various website sections for modeller

**Modeller => comparative protein structure modeling by satisfaction of spatial restraints**

Modeller was conceived in 1993 and it had 10-15 different releases, refining it through the years, it is still one of the best programs around provided that you are able to handle the spatial restraint. Not much used for membrane proteins though.

in principle even knowing only secondary structure prediction, we can simply align the two sequences on the basis of the secondary structure.

**Spatial restraints:**

proteins in the space of the solvent are entities capable of autorganization.

a protein in the space of the solvent has a fold (in order to support the active site or the protein function), that it’s stable in its solvent. It is stable enough to support catalytic activity thanks to its structure. Something supports this structure: they are pairwise bonds that are established between the atoms of the protein itself. The folding is driven by the interactions between the residues and by the frustration of the system (?).

the protein is a flexible system capable of dynamics:

Molecular reasons for stability but at the same time flexibility of the protein?

The protein is capable of dynamics, simulable by programs. Even modeller implements a few steps with molecular dynamics.

other approaches for membrane proteins

protein is a flexible system => capable of dynamics, whose simulation is studied by certain programs

A protein is a stable entity : why?

when a protein is a protein? when the heteropolymer has enough heterogeneity .

The heteropolymer should have enough heterogeneity in terms of lscs to support folding, interactions with the solvent and to be a dynamic entity in the space of the solvent

A polymer of Ala molecules => hard to define a protein, more like a “structure”

The folding process(effect) has a major thermodynamic driving force(cause) => the polar environment, and the many opposite tendencies at the level of the lscs (frustrated)

no matter the protein topology (how it’s organized in the space of the solvent and how the diff sec structures are related) => the deltaG related to the protein folding ranges from up to 0 to -50kcal/mol

an overall free energy variation of 50kcal/mol should tell you that it’s a flexible/dynamic molecule

by changing the temperature or bumping into other molecules this can reorganize in order to interact, or undergo conformational changes that are rather frequent wheny tou think of a protein that goes from a ….^??

I have a template => it’s a stable molecule(i have the crystallized structure)

at the end of the day i have a target that finally has an associated folding

template => we have to understand which are the molecular reasons for its stability

at the backbone level we have all the possible covalent bonds

there is a plethora of known covalent bonding interactions that make the protein a stable entity=> H bonding, ion-ion interactions, dipole-dipole interactions, induced dipole-dipole, induced dipole-induced dipole.

we are using spatial restraints to make it simple => we are extracting from the protein geometrical restraints/constraints/distances/angles that characterize all those interactions that are causing the protein stability

we are not using pairwise interacting equations for each interacting atom in the protein=> we are describing the interactions with partial geometrical constraints

folding process has as a major thermodynamic driving force: the interaction with the solvent , the fact that the environment is polar…

no matter which type of secondary structure has the protein and its topology (how it is organized in the solvent) ...the delta g related to the protein folding ranges from about zero to -50 kcal/mol

the protein is flexible since we know The free energy change associated to a single C-C covalent bond is -100 kcal/mol!

This molecule can not only reorganize the surface for the sake of interaction, but also undergo conformational changes that are rather frequent in the case of e.g. proteins going from inactive to active form.

in this general concept: i have the template ( it is a stable molecule), which are the molecular reason for its stability?

at the backbone there are the covalent bond, but about the non covalent bonding interaction they also make the protein stable.

we can use which equation?pairwise equation between two centres

we can use also spatial constraints (characterize all the interactions that cause the protein stability, along with the distances and angles/orientations that make these interactions possible. In the protein that has to be modelled, in order for those interactions to happen, there have to be the same chemical groups in the same spot and with the same orientation -> hence the spatial constraints)

(see the file in Dropbox for all the list of non bonding interactions)

1. Ion-ion interactions
2. Dipole-dipole
3. Ion-Dipole
4. Hydrogen bond)

The key point for modeling is understanding what makes the template protein stable: which kind of interactions are in place, where do they act, which of them are key to the function of the template? Starting from this knowledge we derive rules in terms of bond distance and angles, in order to apply the to the template protein => We call this rules  *s p a t i a l r e s t r a i n t s*. These restraints are defined first of all in terms of dihedral angles, and later on in terms of distances between side chain atoms.

Modeler simplifies everything in terms of geometric properties, it doesn’t deal with pairwise interactions, tht is the subject of molecular dynamics.

when we’re thinking geometrically on the protein we’re schematizing the protein in terms of dihedral angles, torsion angle, distances between atoms..lateral chain..

we’re not using pairwise interactions equations but we are using geometrical properties in order to describe this interaction.

torsion angles: angle among place in the protein.

Dihedral angles or torsion angles

es dihedral angles: are (phi, psi and ki)

**phi and psi:** torsion angle between the plane of the peptide bond, that is described by the coplanarity of all the 3 atoms of the peptide bond. This bond is a point of rigidity for the statistical mechanics: it is a constraint in the heteropolymer because it is planar, due to the dual nature of the bond which is both single and double and therefore doesn’t allow rotation of its atoms.

constraints: divided at the level of the alpha carbons that is the carbon in which the radicals are bond.

*psi* angle: is between the c carbon and the c alfa in the peptide bond.

*phi*: between the c alpha and the nitrogen of the previous peptide bond.

*Why are torsion angles important ?*

Torsion angles are important because they allow to define local arrangement of the peptide (the secondary structure like beta-sheets, turns) on the basis of specific set of values of phi and psi.

8 different types of local motifs:ALFA HELIX BETA STRAND AND COIL, sometimes also turns

Basically for simplifications we only refer to three (alpha helix, beta strand, coil) (some people also name turns, because they are points of stabilization via h-bonds inside the loops of the proteins)

we have dipole moment at the level of the peptide bond, .. we have the bond length more or less fixed and then 1,33 amstrong between the carbon and the nitrogen, and then the distance between the carbon and the carbon alpha..

these distances are spatial restraints: the geometrical characteristics among the atoms that participate to the formation of the peptide bond and the backbone.

instead of referring to the plane of the peptide bones we can also think to represent the protein considering the distance between carbon alpha and carbon alpha (“diagonal” of the plane where the peptide bond lies).

why? Since the peptide bond plane is rigid we don’t have to describe the position of all the three atoms we just describe the position of the C-alpha at the tips of the plane because they are the only atoms that can move. The position of the peptide bonds is represented through the distance of the two C alpha. Modeller bases its analysis on this Calpha-Calpha distance.

in other words: i have at first to extracts from the template the spatial restraints, how? defining the one that can be derived from the template ( es in the original template are 20).

(non bonding interaction that allows us to describe them in term of geometrically characteristics like distance..).

all this geometrical restraints have to be satisfied, the program(modeller) has to satisfy them.

i have : modeller

1. extracts from the template the spatial restraints that make the template stable=> how? it defines a priori the restraints one should derive from the template (20 in the original paper)
2. try to optimize this restraints considering the different sequence. on the basis of the alignment it tries to optimize the restraints considering a different sequence. spatial restraints are fixing the skeleton of the protein but now the input is an alignment. The sequence differs between the two proteins so modeller takes into account how different residues respond to the spatial restraints???????
3. optimization procedure: if there is the change of a residue in a certain position, modeller has to ???

flexible molecules shift between multiple different stable configurations, they are not unorganised. In modeling you consider just one stable conformation, we don’t have to consider the interaction with the solvent, so we can define the scaffold in a geometrical manner. Is the scaffold still stable if the residues change? it requires an optimisation procedure, called conjugate gradient (wiki reference: <https://en.wikipedia.org/wiki/Conjugate_gradient_method> ).

modeller always work always returns an output but the quality of the biological information depends on the quality of the initial alignment. So we are always responsible for the final quality of the data we produce. Modeller is built to model proteins coming from the same family

I am assuming that the template is just one, thereby not considering the interaction with the solvent. I use geometry to model the scaffold

protein is always folded => less or more flexible, but still folded

what are the reasons for the stability of the protein?

the protein is a closed compartment but it can interact positively with water, proteins are less or more flexible but the reason of the stability of the protein are so many…

i’m able to describe all these interaction . but here i’m not using these equation, here the important thing is adopt a geometrical view of the protein.

the interaction with the solvent is done, i’m assuming that the template is the one that is in that specific conformation because of the solvent.

i’m considering the protein .

i’m producing the geometrical picture of the stability of the protein.

then i’m introducing an alignment.

if i change one residue is the scaffold still stable?

as soon as modeller has an alignment it works, it has a predictive result.

in case the two sequences have a very low sequence identity, then you may align on the basis on secondary structure prediction..

we need to use modeller in the best optimal way to do later on functional annotation => modeller was conceived to build models of proteins within the same family

modeller version => has to be in the methods , with the website i downloaded it from

Andrej Sali → was a PhD student when he first published Modeller in 1993.

Starting from the atms coordinates in the pdb file Modeller computes the spatial restraints(ie the Calpha-Calpha distances) → the whole process is rather fast.

⇒ since the atom coordinates of the pdb are already a reduced representation of the electron density data, there is already a simplification

**when mentioning modeller you should quote the original paper from 1993 (in the dropbox)**

Our references are the first few pages of the manual

From the section 1.1 of the Modeller manual

“ Modeller is a computer program that models three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints.

Modeller is most frequently used for homology or comparative protein structure modeling: The user provides an alignment of a sequence to be modeled with known related structures and Modeller will automatically calculate a model with all non-hydrogen atoms (these structures are often homologs, but certainly don’t have to be, hence the term “comparative” modeling).

More generally, the input to the program are restraints on the spatial structure of the amino acid sequence(s) and ligands to be modeled. The output is a 3D structure that satisfies these restraints as well as possible. Restraints can in principle be derived from a number of different sources. These include related protein structures (comparative modeling), NMR experiments (NMR refinement), rules of secondary structure packing (combinatorial modeling), cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, intuition, residue–residue and atom–atom potentials of mean force, etc. The restraints can operate on distances, angles, dihedral angles, pairs of dihedral angles and some other spatial features defined by atoms or pseudo atoms.

there are geometrical restrains. Presently, Modeller automatically derives the restraints only from the known related structures *(the pdb file that you adopted)*and their alignment with the target sequence. *nb: pseudoatoms: can be present in a certain point of a protein structure. They are placed there in order to have a point of reference for the analysis. Once the spatial restraints have been derived and optimized, the 3D model is obtained by optimization of a molecular probability density function (pdf).*

The molecular pdf for comparative modeling is optimized with the variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing. *this is the computational part of model*

Modeller can also perform multiple comparison of protein sequences and/or structures, clustering of proteins, and searching of sequence databases. The program is used with a scripting language and does not include any graphics. *For this reason either rasmol, chimera, gmol are used to visualize the output of Modeller. Both the input and the output of modeller are a pdb file..* It is written in standard Fortran 90 and will run on Unix, Windows, or Mac computers. ”

From section 1.5 of the Modeller manual

“Modeller implements an automated approach to comparative protein structure modeling by satisfaction of spatial restraints (FiModeller implements an automated approach to comparative protein structure modeling by satisfaction of spatial restraints (Figure 1.1) [Sali & Blundell, 1993]. The method and its applications to biological problems are described in detail in references listed in Section 1.2.

Briefly, the core modeling procedure begins with an alignment of the sequence to be modeled (target) with related known 3D structures (templates). This alignment is usually the input to the program. The output is a 3D model for the target sequence containing all mainchain and sidechain nonhydrogen atoms. *\*all this main chain we derive from the template*

Given an alignment, the model is obtained without any user intervention. First, many distance and dihedral angle restraints on the target sequence are calculated from its alignment with template 3D structures (Figure 1.2).

*(i have the alignment and i can say that i should have a certain interaction between two residue, i have the same interaction in the template and also in the main chain ).*

*the value of these restrains: what can i do to obtain the restrain: i have the protein family, the template for the protein family, and the sequence with a certain variation . in any protein family provided that the homology is >50% the structure has to be conserved.*

*i may derive the distribution of value considering the parameters of a template that are known for a given protein function because of the structure conservation.)*

the form: means the distribution of this restreins.

The form of these restraints was obtained from a statistical analysis of the relationships between many pairs of homologous structures.

*here we have what we can extract from the experiment.*

for ex: This analysis relied on a database of 105 family alignments that included 416 proteins with known 3D structure [Sali & Overington, 1994]. By scanning the database, tables quantifying various correlations were obtained, such as the correlations between two equivalent Cα – Cα distances, or between equivalent mainchain dihedral angles from two related proteins. *since we are reducing all the different fold to a given family..The data has been derived empirically*

These relationships were expressed as conditional probability density functions (pdf’s) and can be used directly as spatial restraints. For example, probabilities for different values of the mainchain dihedral angles are calculated from the type of a residue considered, from mainchain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the pdf for a certain Cα–Cα distance given equivalent distances in two related protein structures (Figure 1.2).

An important feature of the method is that the spatial restraints are obtained empirically, from a database of protein structure alignments. *i’m having my model computed* *.*

*now what can i do to adjust the system?*

*suppose that you have the protein with all the non bonding interaction in place, so that the protein is stable. Now suppose to shake the protein and try to understand how much the protein is resistant to this shake.*

*if the protein tolerate shakes then the protein is supposed to be a stable entity.*

*Inside modeller there is CHARMM, to implement molecular dynamics, because in order to shake the protein you need the force speed: number of molecular interaction used to describe the stability of the protein .*

*the idea is: shake in a force field the protein structure, when you relax the perturbation you want to see where the molecular position of the protein are restored. You retain the few models that stand shaking the best, because this are the models that are more stable*

*since you have very different models you want to verify which one is the best in terms of molecular dynamics.*

*optimization: perturbing the system and then relaxing the perturbation.*

*relative position of lateral side chain respect to the backbone: they are identical to the native whe the residue is conserved.*

*Conserved residues-> identical (or very similar) to the native when the residues are conserved.*

*but when there are changes of residue -> Non conserved residues -> after finding optimal satisfaction of spatial restraints, nobody guarantees that you can have a probability value associated to the lateral side chain orientations with respect to the backbone. Modeller allows you to find a dihedral angle for these lateral side chains (chi angle). There is no means to reach a conclusion for the “trustworthiness” of this computation, because the value of this dihedral angle is optimal according to the procedure. We have no clue at all what the actual phi angle is going to be in a crystallized protein.*

Next, the spatial restraints and Charmm energy terms enforcing proper stereochemistry [MacKerell et al., 1998] are combined into an objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method [Braun & G˜o, 1985] employing methods of conjugate gradients and molecular dynamics with simulated annealing (Figure 1.3). Several slightly different models can be calculated by varying the initial structure. The variability among these models can be used to estimate the errors in the corresponding regions of the fold.”

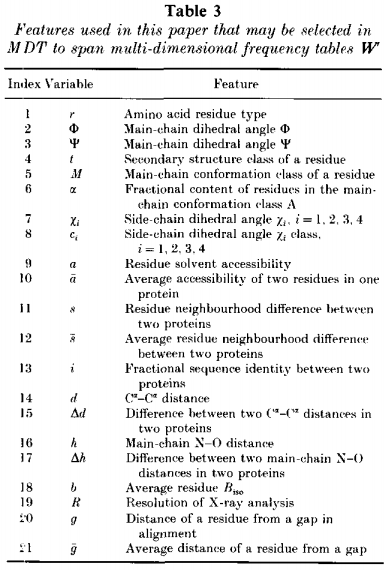
i have the protein family => templates

i have sequences that allow a certain variation per position   
structure is conserved in protein families

100 -200 different forms

Relative position of lateral side chains w/resp to the backbone

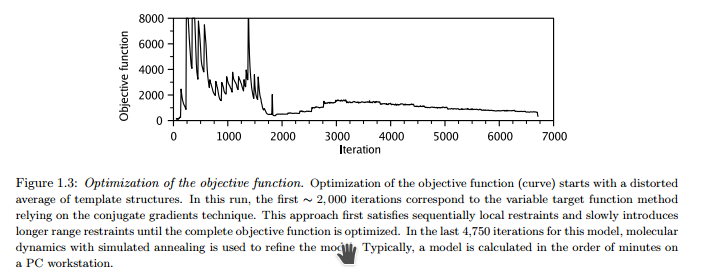
Table 3 -> table of spatial restraints, there are 21, derived from the original database. 21 in the sense of 21 distinct variables for each residue. Taken from the original Modeller article 1993.



mainly juristic parametre: ou need the statistics in order to realize the distribution, and then apply the algorithm

objective fuction: smooth with a few steps the molecular dynamics

So the aim is to apply these molecular dynamics and then refine the angles and then apply molecular dynamics over again, until the parameters stay the same, right?



Taken from the Modeller manual

Running Modeller

Modeller issues

On the coppers → deleted all heteroatoms aside from Cu, cannot visualise Cu in Rasmol

You have to adapt the length of the alignment to include the extra Cu

Also inside Rasmol check if you’re visualising the heteroatoms

USE THE SEQUENCE DERIVED FROM PDB FOR THE TEMPLATE NOT FROM UNIPROT

Delete the signal peptide

To align → run lalign >local if you are only modeling one region /domain

>global if you are trying to model the whole target on one Template

Lalign\_output is used to create a Pir file → alignment format that contains relevant informations for Modeller

>P1;name\_of\_target

sequence::::::::: #there are 9 separators → 10 fields

>P1;name\_of\_template

structureX:name\_of\_pdbfile.pdb:first\_aa\_present\_in\_alignment:chain:last\_aa\_present\_in\_alignment:chain::

#name of pdb file also works without the extension but you have to be in the same directory as the script, X stands for Xray crystallography, first\_aa\_present\_in\_alignment != alignment length

We can delete a gap at the end of the protein, provided that I trimm in an equivalent manner both target, template and its pdb file. *This is manual curation of pir file and alignment*

To add the Cu atoms

We add a number of dots == number of atoms we are adding

Good practise → change the number of the Cu atoms in the pdb file, as if they were residues of the protein. This is t make sure that Modeller will model the heteroathoms

anlfile=’alignment.pir’(NAME OF THE PIR FILE)

knowns=’TEMPLATE’-> same name of first line after your first column

sequence=’TARGET’ ->

EXAMPLE -> given a .pir file that looks like this (filename = pir\_alignment.pir)

on my Python script for modeller, I will set

alnfile = “pir\_alignment.pir”

knowns = “T\_vers” (---> same name as header of my template sequence in the pir file, in other words the identifier that I put after “>P1;...” in the sequence that corresponds to my template, the one with all the fields filled in)

sequence = “P\_eryn” (---> same name as header of my *target* sequence in the pir file)



Discussing the Modeller output

To compare different models of the same protein derived from the same alignment look at the energy function value. It is associated to the stability of the structure.

molpdf is useful to compare models of the same run of modeler (coming from the same input files). It’s a score without a real measure, so it doesn’t make sense to compare models from different runs. It’s not an absolute parameter, it’s relative.

compare models of a single run.

which model would you select,

You can use jCE and compare each model to the original template.

The final parameter that will be indicative of which model is best is the RMSD.

(based in the molpdf?) The lowest value, because it grows with every violation of a constraint

Lowest molpdf value -> best model.

What is the best criteria? There is no absolute answer, *but RMSD is suggested by Casadio*

Selection of the result is up to the end user (T R I G G E R W A R N I N G)