Computational modeling of macromolecules: Developing a model for the structure of an enzyme

We know the nucleotide sequence of the gene and want to know the structure of the enzyme complex.

1- *in-silico* translation of the nucleotide sequence of the genes to the amino acid sequence of the protein. (Primary structure)
2- Search for other proteins with high homology and known structure.
3- Model the amino acid sequence of the query protein onto the fold of the homologous template protein, use simulated molecular dynamics to allow the new amino acid side chains to adjust to their folded environment and to allow loops and secondary structures to adjust to their new lengths. (Secondary and tertiary structures)
4- Test to see if function is likely to be supported by the model structure, on the basis of fit with cofactors. (Tertiary structure and function)
5- Look for signatures of protein-protein interactions and use these to guide docking together of the different subunits or domains of the complex (Quaternary structure).

Example: bifurcating hydrogenase, in which moderately-reducing electrons are drawn from NADH. Of each electron pair one electron undergoes a very favourable transfer that ‘pays for’ thermodynamically unfavourable transfer of the other electron, with the result the 2 H⁺ are reduced to H₂.

\[
\text{NADH} + 2\text{Fd}_{\text{red}} + 4\text{H}^+ \Leftrightarrow \text{NAD} + 2\text{Fd}_{\text{ox}} + 2\text{H}_2
\]

For the organism *Thermotoga maritima* this activity is encoded by a set of three genes that occur together in the chromosome: Tm 1424, Tm 1425 and Tm 1426.

Deletion of any portion of these genes abrogates activity. Transfer of this cluster of genes to another organism confers this activity. This activity would be valuable in fuel production. How is this activity produced by these proteins ???

Adams, King, Maness et al.
Computational modeling of macromolecules:

We know the nucleotide sequence of the gene and want to know the structure of the enzyme complex.

Goals

1- in-silico translation of the nucleotide sequence of the genes to the amino acid sequence of the protein. (Primary structure)
2- Search for other proteins with high homology and known structure.
3- Model the amino acid sequence of the query protein onto the fold of the homologous template protein, (simulated molecular dynamics to adjust sequence of one protein to structure of the other (Secondary and tertiary structures).
4- Test to see if function is likely to be supported by the model structure, on the basis of fit with cofactors. (Tertiary structure and function)
5- Look for signatures of protein-protein interactions and use these to guide docking together of the different subunits or domains of the complex (Quaternary structure).

Tools (just a small sampling of the many)

   (the parent site provides access to many more tools: http://www.expasy.org/)
2- BLAST of various sorts:
   (parent site provides access to many more tools: http://www.ncbi.nlm.nih.gov/ )
3- SwissModel
   http://swissmodel.expasy.org/?pid=smd03
   (again, this is just one of a whole family of tools.)
4- Download the template protein from the PDB and transfer the cofactor coordinates into your model.
   http://www.rcsb.org/pdb/home/home.do
5a- Homology must be obtained within a curated set of proteins that retain the same activity as the model, in order to identify features of interest and interaction patterns, since modules are reused in many different contexts. I have done this for you.
2- BLAST of various sorts: http://blast.ncbi.nlm.nih.gov/Blast.cgi
3- SwissModel  http://swissmodel.expasy.org/?pid=smd03

4- Download the template protein from the PDB and transfer the cofactor coordinates into your model.  http://www.rcsb.org/pdb/home/home.do
5a- Homology must be obtained within a curated set of proteins that retain the **same activity** as the model, in order to identify features of interest and interaction patterns, since **modules are reused in many different contexts**.

Beware: gene sequences are commonly annotated on the basis of sequence homology, creating a dangerous cyclic logic. You cannot use the stated names to know more than the vaguest of ‘type of function’. Good ol’ fashioned (skilled effort in a laboratory) is required to know what an enzyme or complex actually accomplishes.

We will stand on the shoulders of giants: Adams, Harwood, JanSchut, King and Maness who have studied H$_2$ producing organisms and the pathways by which they do so.

Thermotoga maritima Bif-H$_2$ase, HydABC
TM1426-1424 ([FeFe], 6 x [4Fe-4S]), 4 x [2Fe-2S]), FMN)

And away we go:
Begin with step zero: getting the genes in the first place.


Accession name of protein
NP_229226.1
Use this to search for the protein in NCBI’s ‘Protein’ database (see below).
from N-terminus:
TRX_Fd: binding site for [2Fe-2S] cluster, peptide fold corresponds to thioredoxin-type ferredoxin. Structural model is 1M2A.pdb.
NuoF is Nqo1, the FMN- and N3 [4Fe-4S] binding domain of complex I (in its entirety). Structural model is chain 1 or chain B of 3IAM.pdb.
(SLBB domain soluble beta-grasp domain binds diverse soluble ligands.)
Fer4_10 includes two [4Fe-4S] binding domains one after the other. 1Y4Z.pdb
from N-terminus:
Fer2 binds an [2Fe-2S] cluster. Structural model is 1A70.pdb.
NADH-G-4S_3 residues 104-144 of NuoG chain 3 in 3IAM.pdb.
HCP is hybrid cluster protein, including a [4Fe-4S] cluster as in the active sites of CODH and AcetylCoA synthase. This is not coming up for me. FeFeH2ase large subunit, Cterm. 1FEH.pdb
FeFeH2ase sm subunit

0- Get the genes
Hover over bars and boxes for other fun information

Protein accession #

Diaphorase component
FeFeH2ase component A

1- Get the amino acid sequence (No need for Expasy translate.)
Use the accession name of the protein in the /protein/ data base of NCBI to get the amino acid sequence directly. The amino acid sequence is way down at the bottom of the page.
NP_229226.1
Begin a txt file to house all the aa sequences.
Then use BLAST to find homologues. Do this directly from the Entrez/protein page.
2. The NCBI BLAST (Basic Local Alignment Search Tool) window opens. Because our objective is to identify proteins that can serve as templates for our model structure, search only among proteins with known structures (the pdb).

This is one of a family of structures determined on the same protein but with different bound ligands. They are all of the same protein and therefore are equally plausible templates on the basis of the amino acid sequence.
3- SwissModel http://swissmodel.expasy.org/?pid=smd03

You can either provide your choice of template structure to be used or you can let SwissModel choose. SwissModel will not know if strong homology in one region of the protein is more important than overall homology, so if you have specific questions or regions of interest, do your homework, learn what potential templates exist and determine if some will be better than others. You can ALWAYS generate models with each of several templates and determine later if one proves to be more consistent with other knowledge than the others. Eg., one may be able to explain the severity of certain mutations whereas other models may place the mutated residues in places that ought not to matter.

In class and for this exercise you can run the fully automated version if you prefer. We are in the situation where all but one of the possible models are pretty closely related and the two I tried gave indistinguishable results. (You can also look in the ‘repository’ for models already generated in case someone else has posed your problem.) In fully automatic mode, copy the FASTA format amino acid sequence of your protein and paste into the box. The FASTA formatted amino acid sequence is obtained from the NCBI Protein page by clicking the ‘FAST’ button (green box, above).

3- SwissModel results

Template used

N → C colouring

VERY VALUABLE figures of merit and help buttons to provide explanations of each measure of model quality.

Download your model structure!
**3- The model structure**, coloured in accordance with the quality of the model for each amino acid’s conformation and environment.

Anolea

The atomic empirical mean force potential ANOLEA (Melo et al.) is used to assess packing quality of the models. The program performs energy calculations on a protein chain, evaluating the "Non-Local Environment" (NLE) of each heavy atom in the molecule.

The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

QMEAN is a composite scoring function for both the estimation of the global quality of the entire model as well as for the local per-residue analysis of different regions within a model.

Gromos

GROMOS (van Gunsteren et al.) is a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems and can be applied to the analysis of conformations obtained by experiment or by computer simulation.

Save your model using an informative name.

Blue: more reliable, red= less reliable

modelhydrogenase_QMEANlocal_in_Bfactor_on3c8y.pdb

**3- the quality of the model**

In the SwissModel Workspace the QMEAN4 score is used to evaluate the generated models. The global QMEAN4 scoring function (Benkert et al. 2009) is a linear combination of four structural descriptors using statistical potentials: The local geometry is analysed by a torsion angle potential over three consecutive amino acids. Two distance-dependent interaction potentials are used to assess long-range interactions: the first is a residue-level implementation based on C-beta atoms only and the second an all-atom potential which is able to capture more details of the model. A solvation potential investigates the burial status of the residues. The global QMEAN6 score uses two additional terms describing the agreement of the predicted (from sequence) and the calculated secondary structure and solvent accessibility of the model.

QMEAN4 is a reliability score for the whole model which can be used in order to compare and rank alternative models of the same target. The quality estimate ranges between 0 and 1 with higher values for better models. Additionally, the pseudo energies of the four contributing statistical potential terms are provided. The comparison of the differences of the terms among the models may help understanding the reason for the differences in the estimated model quality.

For the quality estimation of multiple models of the same protein, please visit the QMEAN server (Benkert et al. 2009, NAR Web Server Issue) which allows to process sets of models (submitted as compressed archives) and pools the results: http://swissmodel.expasy.org/qmean

In addition to the raw scores, Z-scores of the QMEAN composite score as well as all terms are provided relating the quality estimates to scores obtained for high-resolution reference structures solved experimentally by X-ray crystallography (Benkert et al. 2017). The QMEAN Z-score represents an measure of the absolute quality of a model by providing an estimate of the ‘degree of nativeness’ of the structural features observed in a model and by describing the likelihood that a given model is of comparable quality to experimental structures. Models of low quality are expected to have strongly negative QMEAN Z-scores (i.e. the models’ QMEAN score is several standard deviations lower than expected for experimental structures of similar size). The analysis of the Z-scores of individual terms may help identifying the geometrical features responsible for an observed negative QMEAN Z-score. A more detailed explanation on how the Z-score are calculated can be found in the help of the QMEAN server.
4. **Test the model** by determining whether it can accommodate the cofactors we know it must have in order to support its known function.

(Many different tests are possible at this stage, the cofactor test is just an example that is appropriate for our objective but you could (and should) think of lots of useful tests.)

Download the template structure since it will contain coordinates for the cofactors, whereas the SwissModel structure will be of amino acids only.

Save with an informative name such as 
H2aseTemplate_3C8Y.pdb

---

4. **Test the model:** Test integrity of the active site by determining whether it can accommodate the cofactors we know it must have in order to support its known function.

Use a text editor such as textedit or microsoft word, open the pdb file of our model and the pdb file of your template (modelhydrogenase_QMEANlocal_in_Bfactor_on3c8y.pdb and H2aseTemplate_3C8Y.pdb respectively).

Scroll to the end of the lines describing amino acids in the model file (the end of the file) and to the end of the lines describing amino acids but before the lines describing water molecules in the template file.

Locate the lines describing the cofactor atoms’ positions and confirm that the atoms’ numbers are not used by any other atoms in the model file. Copy these lines into the model file and save with a new name (Do not corrupt the file produced by SwissModel. That file is your result, analogous to data, don’t modify it.)

Cofactor residues are FES, HCN and FS4 in this case.

Be sure that you save your modified file with a new name and as plain text as .txt or .pdb.

Copy these to the model
4. Test the model. Test integrity of the active site by determining whether it can accommodate the cofactors we know it must have in order to support its known function. Open your model that has been augmented with cofactors, using Chimera. Chimera is very ‘clever’ and should display as atoms any residues that are not one of the familiar 20 amino acids. (In this case it displays the residues FES, HCN and FS4.)

We know (for example by reading the paper describing the 3C8Y structure) that all the FeS clusters are ligated by S atoms from Cys or N atoms from His side chains. We will test whether the model also retains these correct chemical connection. We will accept the model if we see S or N atoms from Cys or N at 2.5 Å or less from each Fe atom.

Select > residue > FES to select one cluster select one Fe atom and ^ to get the whole cluster. Then select > zone and choose < 2.5 Å (note that the example chooses < 2.4 Å). Look up the actual Fe-S bond lengths, I am giving us a bit of a break, what should those distances really be? Once you have selected all atoms within 2.5 Å display them. Make a table of all four Cys-S-Fe distances (or His-N-Fe), for each cluster. These are experimental results from the computational modeling exercise performed by SwissModel.

The H-cluster does not have any bonds to protein residues, but check and comment, are there any VdW conflicts?
5- Look for signatures of protein-protein interactions and use these to guide docking together of the different subunits or domains of the complex (Quaternary structure).

Coming Soon !!!
The next lecture will review step 4 and cover this as well as the actual docking (step 5).