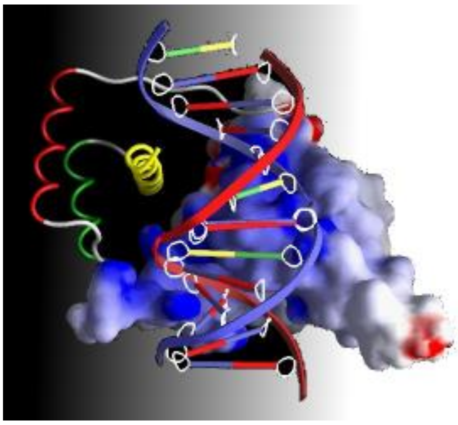


TP 6 : Synthesis of the practical courses

Charlotte Nachtegael

MA1 – Bioinformatics and Modelling

Biophysics and structural Bioinformatics I

2015-2016

1. L-lactate dehydrogenase A-like 6A (*Homo sapiens*)

1.1. Prediction of secondary structure

```
ccchhhhhhhhhhhhhhh?ccceeee?cccc?hhhhhhhhhhcccche
eeeeecccccccce?eecccccceccccc?e?ccceeeecccceeeeee
ccccccccc?hhhhh?chh??eeecccc?cccccceeeeeccc?????
?hhhhcccccceeeecccccchhhhh???hc?eee?ccccc?eeeecc
cccc?eeee?c?eccccccccccccccccccchhhhhhhhhh??cceee
eeeecccccceeeeehhhhhhhhhhhhhhcccccc?ccc?cccccceee
e?cceeccccc?hhhhhhc?hhhhhhhhhhhhhhhhhhhhhcc
```

Different methods to predict secondary structure of protein exist. In this case, we used a consensus of different approaches. Each method gave his results which are compared and finally give a final result. If the results of the methods were equally divided the final result is a '?'. We preferred to use consensus method, so we combine different approaches. The final results obtained can be trusted more implicitly as it is a result confirmed by different methods.

The algorithms used for the consensus were GOR IV, HNN, SOPMA and PHD. GOR IV is a statistical method, based on the already known secondary structures of proteins. It predicts the secondary structure of an amino acid based on all pair frequencies in a 17 amino acids window. HNN (Hierarchical Neural Network) is a learning method. It is made of two networks, the first one is a sequence to structure and the second one a structure to structure. This type of algorithm is first train with a dataset of protein with known protein structures, until their parameters give the best output compared to the real result. SOPMA (Self-Optimized Prediction Method with Alignment) is based on sequence alignment and then predict the secondary structure by homolog method. The PHD method combines the neural network method and the alignment method.

The limits for the different structures were compiled in the aim to compare them with the limits obtained with the assignation of secondary structure from the predicted 3D structure of the same protein (see point 1.3).

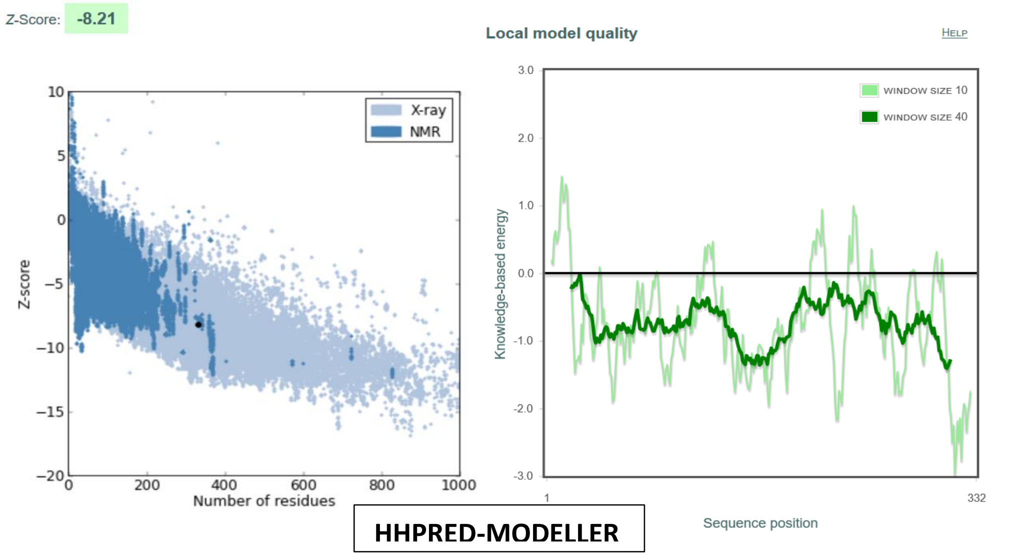
1.2. Prediction of tertiary structure

For the prediction of the 3D structure, two approach can be used. The first one is by comparative modelling method, based on sequence alignment, and the second one is by fold recognition method, based on the fit of the sequence of the target protein with structural templates. This second method is generally used when no homologous protein can be found for the target protein by sequence alignment, thusly resulting in a quite poor quality for the 3D predicted structure found with the first method.

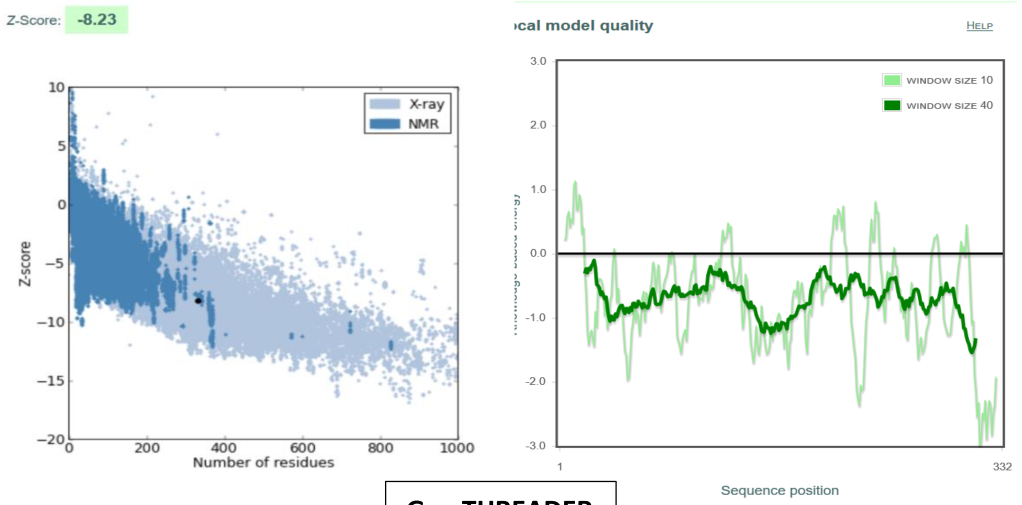
The tool used for the comparative modelling method, was HHPRED-modeller. HHPRED allows to search a wide variety of database for sequence alignment. We can then select or let the program automatically select, the best homologous protein found by sequence alignment which will then be used by modeller to build a 3D structure for our target protein. The template chosen here was a rat LDHA complex (4AJ2).

The tool for the second method was Gen THREADER. It is based on profile-profile alignment of predicted secondary structures, which select then 3D templates. The algorithm then fit the sequence of our target protein to the template and calculate the compatibility of the sequence with the structure. The template found with this method was the chain H of the complex of the human heart lactate dehydrogenase (1I0Z).

To evaluate the models, we use the website ProSa-web, which calculates the quality of the models with an energy function, based on the distance between residues. The overall quality is determined by the z-score, which can be used to check whether the *z*-score of the input structure is within the range of scores typically found for native proteins of similar size. The local model quality is built by plotting energies for each amino acid, based on the average energy over each 40-residue window which is then assigned to the 'central' residue of the window. Another plot with a 10-residue window is also shown. In general, positive results for the local model quality are translated into problematic or poor parts of the model.



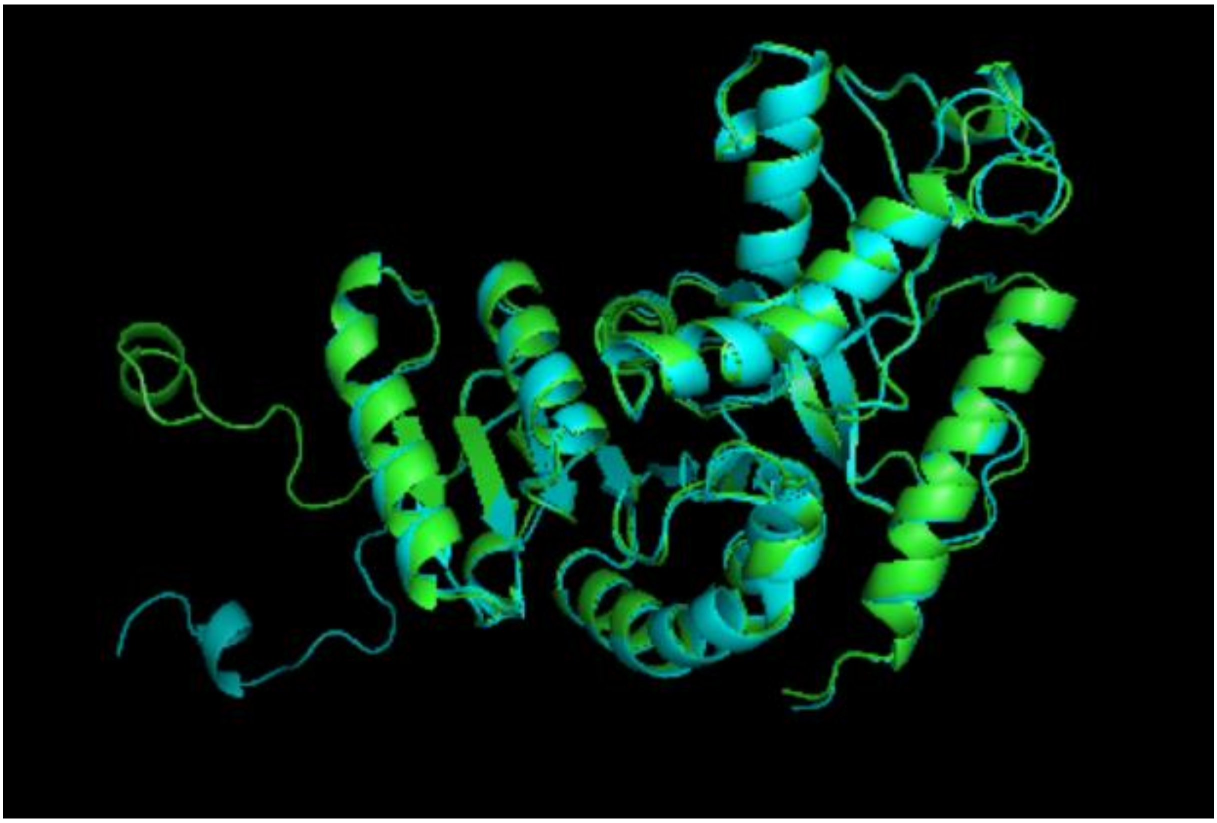
HHPRED-MODELLER



Gen THREADER

The two models were found to have a z-score found within the range of scores typically found for native protein structures. Both the local model quality presented a similar profile, with all the plot for the 40-residue window beneath 0, although the 10-residue window plot had some parts above 0. The similarity of both plots calculated from the models obtained by different 3D structure prediction methods and their z-score within range of scores for native proteins allowed us to determine the overall good quality of the models.

Finally, we compared the models obtained themselves by aligning them with PyMol. We could already see with the plots of ProSa-web that the structures were close, which is confirmed by PyMol. We can just say with this kind of results that the models are close structurally, but this does not say anything about the trust we can have in these models or their quality.



1.3. Assignation of secondary structures

With STRIDE server, we are able, from a 3D model of coordinates, to assign secondary structures to the sequence of the protein. The model used with STRIDE was the one obtained with HHPRED-Modeller. We compared the limits of the secondary structures obtained with this method to the ones obtained with the algorithm of secondary structures prediction.

We could find for most of the alpha helices predicted a correspondence within the limits found with STRIDE. The same could not be said for the beta strands, where a greater number of them could not be found with STRIDE. Altogether, it shows a large number of consensus, which could give a greater weight to the 3D model.

Alpha Helix		Beta strand	
STRIDE	CONSENSUS	STRIDE	CONSENSUS
4-8	4-18	22-26	24-26
30-42	34-42	47-51	48-53
	47		62-65
55-68			72
110-127	107-111	76-80	78
	114-116		83-86
140-151	146-150	91-94	91-96
163-178	167-171		118-120
	175	132-135	132-136
211-214		158-160	156-159
227-245	225-234		177-179
250-265	256-269	185-190	187-190
	302-307	198-206	198-201
310-329	310-330		205
		209-210	
			238-244
			250-254
		269-276	
		288-296	286-289
			293-294

2. Lysozyme structure

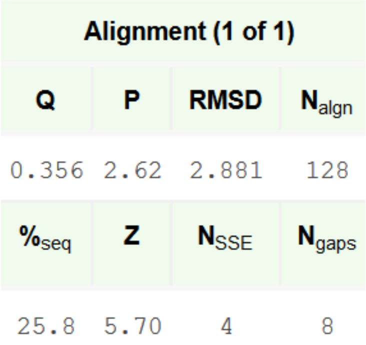
2.1. Quality of experimental structure

On the profile of 1LYD on PDB, we can find several informations about the quality of the structure obtained. This structure was obtained by X-Ray diffraction, with a resolution of 2 angström and a R-value of 0.191. R-value is the measure of the quality of the atomic model obtained from the crystallographic data. When solving the structure of a protein, the researcher first builds an atomic model and then calculates a simulated diffraction pattern based on that model. The R-value measures how well the simulated diffraction pattern matches the experimentally-observed diffraction pattern. A totally random set of atoms will give an R-value of about 0.63, whereas a perfect fit would have a value of 0. Typical values are about 0.20. There is one potential problem with using R-values to assess the quality of a structure. The refinement process is often used to improve the atomic model of a given structure to make it fit better to the experimental data and improve the R-value. Unfortunately, this introduces bias into the process, since the atomic model is used along with the diffraction pattern to calculate the electron density. The use of the R-free value is a less biased way to look at this. Unfortunately in this case, we did not have this R-free value.

A complete report of the quality of the structure can also be found on PDB. We could not only find in this report the presence of only one outlier in Ramachandran plot, but also that no expected bond length was above the limit. Overall, the structure seems to be of a good quality.

2.2. Superimposition of chain A 1LYD and of 2ANV

To superimpose two structures, we used PDBefold. The statistical results found for this action are as below :



Alignment (1 of 1)

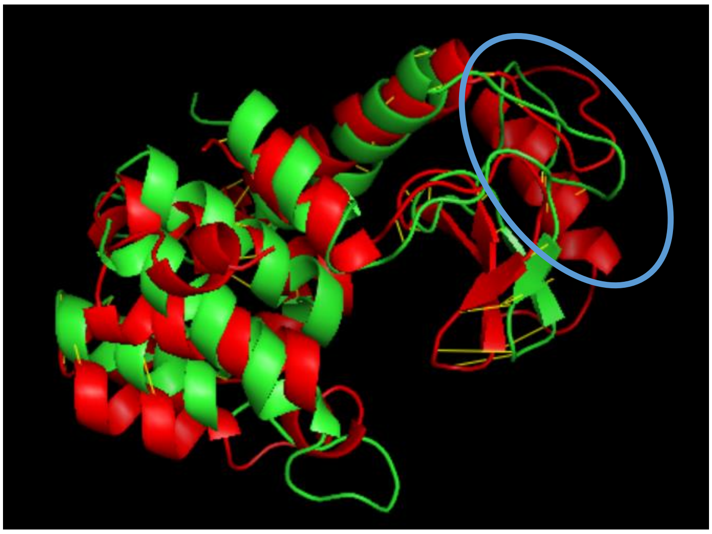
Q	P	RMSD	N_{algn}
0.356	2.62	2.881	128
$\%_{seq}$	Z	N_{SSE}	N_{gaps}
25.8	5.70	4	8

- Q-score : gives the altogether quality of the alignment, based on the rmsd, the number of matched residues (Nalign), and is equal to 1 if the structures are identical.
- P-score : probability to achieve the same quality of match, when it is lower than 3, it indicates a statistically insignificant match.
- RMSD : Minimal mean distance between residues, equal to 0 if the structures are identical.
- Z-score : statistical significance of match with the help of a Gaussian distribution. The higher the score, the more significant eh result.
- NSSE : number of secondary structures matched
- Ngaps : indication of the quality of the alignment.

It was quite visible that the significance of the match was not high, as indicated by the P, Q and Z-scores. However, the sequences seemed to be quite well aligned (Nalign high and low Ngaps), although the percentage of sequence identity was not high. Despite these results, we had 62% of the 1LYD secondary structures which could be matched to 89% of the secondary structures of 2ANV.

2.3. Secondary structure not matched

We could find in the match of the two structures, more than 10 residues in 1LYD not matched in 2ANV. These amino acids can be found from 34 to 47 and correspond to a coil and the majority of an alpha helix (in red inside the blue circle on the figure below). No specific function could be attributed to this region.



References

Tools

- PDBeFold : http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html
- Network Protein Sequence analysis : https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html
- PDB : http://www.rcsb.org/pdb/home/home.do
- HHpred : http://toolkit.tuebingen.mpg.de/hhpred
- Gen THREADER : http://bioinf.cs.ucl.ac.uk/psipred/
- ProSa-web : https://www.came.sbg.ac.at/prosa.php
- PyMol : http://pymol.org/ep/
- Stride : http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py

Documentation

- Help on scores of PDBeFold : http://www.ebi.ac.uk/msd-srv/ssm/ssmresults.html
- Secondary structures Prediction Algorithms : http://genamics.com/expression/strucpred.htm
- Validation report of 1LYD : http://ftp.wwpdb.org/pub/pdb/validation_reports/ly/1lyd/1lyd_full_validation.pdf
- Jones, D. T. 1999. « GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences ». *Journal of Molecular Biology* 287 (4): 797-815. doi:10.1006/jmbi.1999.2583.
- Söding, Johannes, Andreas Biegert, et Andrei N. Lupas. 2005. « The HHpred interactive server for protein homology detection and structure prediction ». *Nucleic Acids Research* 33 (Web Server issue): W244-48. doi:10.1093/nar/gki408.