Haute École en Hainaut
Av. Victor Maistriau 8 – 7000 Mons

MA2 Ingénieur industriel – Life Data Technologies

Computational biology project

Practicals description

Vincent Branders

Academic year 2023-2024

Version of 21 septembre 2023

Table des matières

Project description

Introduction

Welcome to the 'Bioinformatics Project' course. This project is the ideal opportunity to develop crucial skills while exploring fascinating aspects of structural bioinformatics

Project objective

This project is divided into four chapters, each offering a unique and stimulating experience in the field of structural bioinformatics. You will have the opportunity to deepen your understanding of topics such as the classification and prediction of biological macromolecule structures, as well as the study of enzyme-ligand complexes.

Key skills

Use of LaTeX

Beyond technical skills, you will discover the importance of LaTeX for scientific document writing. As engineering students, the ability to create technical reports in LaTeX is a valuable skill that will serve you throughout your career.

Critical thinking

This project emphasizes critical thinking. You will need to make significant decisions throughout the process and explain your choices transparently. This reflects the engineering approach you aim to develop.

Analyzing structural bioinformatics results

You will learn to analyze and interpret results in structural bioinformatics, a fundamental skill for any future bioinformatician.

Assessment

At the end of the year, you will be evaluated based on several essential criteria :
LaTeX Usage : Your ability to write a quality report in LaTeX will be assessed. This skill will enable you to effectively communicate your scientific findings.
Readability : The organization and clarity of your report will be considered because scientific communication is just as important as the research itself.
Clarity of Responses : The quality of your answers to project questions will be evaluated, showcasing your understanding of the concepts covered.
Reflection and Analysis : The analysis of the choices you have made and the quality of your explanations will be key elements of the evaluation, reflecting your development as engineers and bioinformaticians.

Practical 1

Visualization and analysis of biomolecular structure

1.1 Quality of experimental structures

Exercice 1.1 Find on the Protein Databank website (`http://www.rcsb.org`) the proteins with the PDB code *1N0S* and *1T0V*. Which experimental method has been used to obtain them ? What are the values of the resolution and of the R factor (R-values) of the X-ray structure ? What is the difference between R-value observed and the R-value free ? How much conformers are there in the PDB file of the NMR structure ? What are conformers ?

Exercice 1.2 *Compare the quality of both structures.* Use for that purpose the full report of the "Structure validation" section ; focus on sections "Overall quality at a glance", "Residue property plot" and "Model quality / Torsion angles / Protein backbone" of this report. In the "Residue property plot" section of the report for *1N0S*, what does a red dot on the figure mean ? Which residues are marked with a red dot ?

Analyze also the Procheck report of each structure. This Procheck report is available on PDBSum (`http://www.ebi.ac.uk/pdbsum`).

1.2 Visualization of structures

The Pymol software will be used to visualize the structures. The Pymol software and a user manual are provided on `https://pymol.org/2/` and `http://pymol.sourceforge.net/newman/userman.pdf`.

Exercice 1.3 Download the PDB files of *1N0S* and *1T0V* (menu on the right, "Download files PDB text"). Open *1N0S* in Pymol. Test the menus on the right (A-S-H-L-C). Show the protein with the "Cartoon" representation (menu "S"). Select the chain A of the protein ("select xxx, chain A", where xxx is the name of the menu that will be created for the selected items), and represent this chain in a different color (menu "C"). Hide the "Cartoon" representation (menu "H") and show the sticks representation.

Exercice 1.4 Download the electronic density file of *1N0S* from the "Structure Summary" menu, then "Electron Density". This file has been obtained from the "Electron Density Server". Open this file in Pymol. In the "A" menu of *1N0S*M1, choose "Mesh", with the level 1.0. Then, modify the level to 1.5 ("A" menu of the "_mesh" menu that has been created). Select one of the amino acids that is marked by a red dot in the "Residue property plot" section of the validation report (see section 1B) and zoom on this residue (Menu "A", "zoom"). Is there a good conformity between the electron density of this residue and the position of its atoms ?

Exercice 1.5 Erase the electronic density and open the *1T0V* file in Pymol. Align *1T0V* and *1N0S* (menu "A" of *1T0V* align to molecule *1N0S*). Zoom on the superimposed structures (menu "A" zoom). Erase *1N0S*. Show the "Cartoon" representation of *1T0V*. Analyze the structure of the 20 conformers of this NMR structure (menu at the bottom right of the Pymol window, use the arrows to visualize the different conformers).

1.3 Enzyme-ligand complex

Erase the previous work in Pymol : "File reinitialize".

Exercice 1.6 Download from the PDB the structure file of dihydrofolate reductase in complex with trimethoprim (PDB code *1DG5*). Submit this PDB code to the PDBSum website (http ://www.ebi.ac.uk/pdbsum/). Which amino acids are in interaction with trimethoprim (the code "TOP" corresponds to this ligand ; go in the left menu, section "Ligand").

Exercice 1.7 Open the PDB *1DG5* in Pymol. Show the surface of the protein. Select the amino acids that are involved in the interaction with the ligand (see section 3A) : "select inter,resi x+y+z", where x, y, z, ... are the number of the identified amino acids and "inter" is the name of the menu that will correspond to the selection (you can choose this name). Color these amino acids. Select the trimethoprim ("select trim, resn TOP") and show this selection by using the "Sticks" representation and color this selection. Analyze the surface of the protein and the position of the ligand.

Exercice 1.8 Show the protein in "Cartoon" and the amino acids that interact with the ligand in "Sticks". Analyze the distance between them (take the atoms that interact) : use the menu "Wizard" (at the top of the Pymol window) and then "Measurement".

1.4 Compactness of the protein core

Exercice 1.9 Select the amino acid number 100 of *1DG5* and use a "Sticks" representation (give a name to this selection). Select then the surrounding of this residue : "menu A of the selection modify around residues within 5 Å". Show these amino acids in "Sticks". We will mutate this core amino acid to illustrate the protein compactness. Use the menu "Wizard Mutagenesis". In the new menu on the right of the Pymol window, choose a mutation into phenylalanine (Phe), then click on the amino acid number 100. You can test the different possible conformers of the side chain with the arrows on the bottom right of the Pymol window. The red disks show the steric clashes.

Practical 2

Classification and protein structure alignment

2.1 Search for protein domains and their structure

Exercice 2.1 Search in the Uniprot database (`www.uniprot.org`) the protein with the code Q12923.

- What is the sequence length of this protein ?
- Cite some domains found in this protein ?
- To which domain does the PDB code structure 3PDZ correspond ?
- Are there disordered regions in this protein ? If so, were the annotations made automatically or by experimental measurement ?
- Is there an experimental structure for the whole protein ?

2.2 Classification database

Exercice 2.2 Find on the PDB website (`http://www.rcsb.org`) the proteins with the following PDB code : 3PDZ, 1Z86 and 1RGW. Download the sequences in FASTA format (menu "Download Files" on the top right of the web page).

Exercice 2.3 Click on the "Annotations" tab at the top of the PDB web pages

- What is the CATH classification of these domains ?
- What is the Protein Family Annotation of these domains ?
- What do CATH classification and Protein Family Annotation refers to ?

Exercice 2.4 Go to the InterPro database (`https://www.ebi.ac.uk/interpro/`). Perform a text search for this domain in the InterPro database. What is the biological function of the domain ?

2.3 Structure superimposition

Pymol will be used to visualize the superimposed structures. Pymol will not be used to superimpose the structures.

Exercice 2.5 Use ClustalOmega (`http://www.ebi.ac.uk/Tools/msa/clustalo`) to perform a multiple sequence alignment of the sequences downloaded in first exercice of previous section (paste all the sequences in FASTA format). Analyse the sequence alignment, and focus on the conserved amino acids at the different sequence positions. Are these sequences highly conserved ?

Exercice 2.6 Needle (`https://www.ebi.ac.uk/Tools/psa/emboss_needle`) is a tool for performing global sequence alignment, and Matcher (`https://www.ebi.ac.uk/Tools/psa/emboss_matcher/`) for performing local sequence alignment.

Perform a global sequence alignment between 3PDZ and 1Z86 with Needle ; the sequences must be provided in FASTA format (see first exercice of previous section).

Exercice 2.7 Use PDBeFold (`http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html`) to align the structures 3PDZ and 1Z86. Choose in the "Query" and "Target" fields the option "PDB entry" as a source and give the PDB codes. What are the values of the z-score and of the rmsd, and what do they mean ?
Click on the "1" in the "##" column and then download the superimposed structures (click on the 2 "download" tabs). Use Pymol to visualize these superimposed structures. The aligned sequences according to the structure superimposition are available lower on the web page.
Compare the aligned sequences obtained by sequence alignment (Needle) and by structure superimposition.

Exercice 2.8 Download the sequence of 1FCF from the PDB website.

- Identify from the PDB website the Uniprot code of 1FCF. Use Uniprot to identify the limits of the PDZ domain of this protein.
- Use Matcher and Needle to perform a local and a global sequence alignment between 1FCF and 3PDZ. Comment the results : are the PDZ domains aligned properly ?
- Use PDBeFold to superimpose the structures of 1FCF and 3PDZ. What are the alignment scores ? Visualize the superimposed structures in Pymol. Analyse the sequence alignment obtained from the structure superimposition, from the PDBeFold website. What are your conclusions ? Compare these results with those obtained with the sequence alignment.

Practical 3

Secondary structure prediction – conformational diseases

3.1 Analysis of secondary structure prediction programs

Exercice 3.1 The programs that will be used to predict the secondary structure are GOR IV, HNN (see `http://npsa-pbil.ibcp.fr/`, "secondary structure prediction" section) and Sympred (`http://www.ibi.vu.nl/programs/sympredwww/`). Describe briefly the approach used by each program.

3.2 Comparison of the performances of secondary structure prediction programs

Exercice 3.2 Search for the human thymidylate kinase (PDB code 1E2F) in the protein databank (`www.rcsb.org`). Search for the secondary structure limits of the protein in PDBSum (`http://www.ebi.ac.uk/pdbsum`; "Protein" tab, "7 strands" and "11 helices" in the menu on the left). This secondary structure assignation is provided in the file 1e2f.xls (on Moodle). A secondary structure assignation is also provided in the PDB file, sections "HELIX" and "SHEET" (from the PDB website, menu "Display file", "PDB file", find the "HELIX" and "SHEET" sections). Compare both secondary structure assignations (from PDBSum and from the PDB file). Why are they slightly different ?

Exercice 3.3 Perform a secondary structure prediction of 1E2F by using the three programs GOR IV, HNN and Sympred. Show in the .xls file the results. Analyse and comment these results.

3.3 Analysis of two sequences with particular properties

Pymol will be used to visualize the superimposed structures. Pymol will not be used to superimpose the structures.

Exercice 3.4 Predict the secondary structure of the sequences unknown1.fasta and unknown2.fasta (available on Moodle), by using the 3 programs used in the previous sections. Record the predictions in the files unknown1.xls et unknown2.xls that already contain the secondary structure of the experimental structure of these proteins. Use BLAST (`http://www.ncbi.nlm.nih.gov/blast/`; "Protein Blast") to identify the proteins corresponding to these sequences. Analyse the results, make assumptions and interpret the results.

Practical 4

Three-dimensional structure prediction

This practical focuses on the modelling of protein structures by comparative modelling and by fold recognition. You will create different models for the acyl carrier protein of *Rhodospirillum centenum*. The quality of your models will be evaluated with the global Qmean score and Procheck. These tools are available here :

- Qmean : `https://swissmodel.expasy.org/qmean/`
- Procheck : *via* "PDBSum Generate" : `https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html`

What is the Qmean score based on and how to interpret its value ?

4.1 Comparative modelling : manual approach

The uniprot code of the acyl carrier protein of *Rhodospirillum centenum* (ACP) is **B6IN76** ; its sequence is available in FASTA format on Uniprot (`http://www.uniprot.org`). In this first section, you will do manually each step of a comparative modelling :

- search of possible templates,
- selection of a template,
- sequence alignment between the target and the template,
- modelling and evaluation of the quality of the model.

Exercice 4.1 Perform a Blast (`http://www.ncbi.nlm.nih.gov/blast/` ; use "Protein blast") on the ACP sequence to identify a template to model the structure of this protein (**choose the appropriate database to scan with BLAST** ; see alignment parameters).

Exercice 4.2 Select a template among the hits that have been identified by Blast. For that purpose, take into account the percentage of sequence identity, the percentage of query cover and the quality of the structure of the template (see the tools used in Practical 1). Download the sequence of the template from the PDB website.

Exercice 4.3 Perform a sequence alignment between the template and ACP. For that purpose, use the Needle global alignment program (`https://www.ebi.ac.uk/Tools/psa/emboss_needle/`, choose the "Markx3" output format in the "More options" menu).

Exercice 4.4 Submit the sequence alignment obtained in the section 1.C. to the Modeller server (`https://toolkit.tuebingen.mpg.de/#/tools/modeller`). The license key to use Modeller is "MODELIRANJE". The sequence alignment must be provided in PIR format. On Moodle there is a document that explains how to convert the Markx3 format obtained from the sequence alignment into a PIR format that must be submitted to Modeller.
Save the PDB file of the model.

Exercice 4.5 Analyze the quality of this model.

4.2 Comparative modelling : semi-automatic approach

The HHPRed server combined to Modeller (`https://toolkit.tuebingen.mpg.de/#/tools/hhpred`) will be used.

Exercice 4.6 Submit the ACP sequence to the HHPred server. Describe the first step performed by HHPred. Several templates are proposed. Compare these templates to those identified with Blast in section 1.B. (sequence identity, quality of the template structure, ...).

Exercice 4.7 Select the first template and click on "Model using selection". HHPred will align the 2 sequences. Then click on "Forward to Modeller", use the Modeller-key "MODELIRANJE" and click on "Submit job".
Save the PDB file of the model (you will find a "Download PDB file" tab).

Exercice 4.8 Analyze the quality of this model.

4.3 Comparative modelling : automatic approach

Exercice 4.9

- Use the SwissModel server (`https://swissmodel.expasy.org`) to build a model of ACP.
- Select a different template than that used in sections 1 and 2.
- Identify the template that has been used. Save the PDB file of the model.

Exercice 4.10 Analyze the quality of this model.

4.4 Fold recognition

Exercice 4.11 Submit the sequence of ACP to genTHREADER (`http://bioinf.cs.ucl.ac.uk/psipred/`, choose only GenTHREADER-Rapid Fold Recognition). Select the first template proposed (GenThreader Table) and build a model. For that purpose, provide the Modeller license key (MODELIRANJE), scroll to the right of the result page and click on "Model".
Download the PDB file of the model.

Exercice 4.12 Analyze the quality of this model.

4.5 AlphaFold model

An AlphaFold model is available on the Uniprot page of the protein ("Structure" section).

Exercice 4.13 Download the PDB file of the AlphaFold model from Uniprot.

Exercice 4.14 Evaluate the quality of this model.

4.6 Comparison of the models

You have now 5 models for ACP : 3 models obtained by comparative modelling, 1 model obtained by fold recognition and 1 model from AlphaFold.

Exercice 4.15 Compare the quality of these models.

Exercice 4.16 Open the 5 PDB files of the models in a same window in Pymol.
Compute the rsmd between all pairs of structure. For that purpose, you can use the "super" command in pymol. To superimpose the main chain atoms of struc1 to those of struc2 and to compute the rmsd, for instance, use the command "super struc2 and name ca+c+o+n,struct1 and name ca+c+o+n,cycles=0". Read the documentation about the "super" command.
Write a script to perform this repetitive task. Pymol can execute scripts written in Python. To execute a given script, "scr.py" for instance : write "run scr.py" in the command line of Pymol. The commands that could be useful in the script are : `cmd.get_names()`, `cmd.super()`, `len()`. You can also use the Pymolwiki to get more information.
Group the models according to their structural similarity (evaluated by the rmsd value).

Exercice 4.17 Use the "super" command, to superimpose all the models on one the 5 models and to identify the regions where the structural differences are the largest.