Artificial Intelligence for Biotechnology

Summer 2023

Prof. Dr. Dominik Grimm

Exam/Final Project

26. June 2023 - 18. July 2023

Important Dates

- Final project submission: **18th of July at 11:59 p.m.** (CEST - Central European Summer Time).
- Final presentations and oral exams: **20th and 21th of July** (schedule for the groups will be released approximately two weeks before the presentations)

Formal Instructions

- You have to submit a single executable and runable Jupyter Notebook that shows the code for your experiments. All documents have to be submitted via the Moodle submission form until the **18th of July at 11:59 p.m.**! Please upload a single ZIP file, including your Jupyter Notebook and other relevant material. **Late submissions will not be considered and will lead to the failure of the exam!**
- You must provide a signed statement (use the provided Word document in Moodle), in which your group confirms that all the work was done by the group, that you did not copy code from the others or the internet and that all used resources are properly cited and mentioned in your Jupyter Notebook. **All forms of plagiarism will lead to the failure of the exam**.
  The statement must also contain a section in which you describe the author contributions of each group member. Distribute the work equally between all team members. The group as a whole is responsible for the project. Author contributions could be described as in the following example: *"DG preprocessed the data; DG implemented the logistic regression and evaluated its performance; CT implemented the ....; HG analysed the results; All authors helped to compile the Jupyter demo; All authors approved the final demo"*
- **Presentations and Oral Exam**
  - All members of the group must be present for the code presentation. The Jupyter Notebook code presentation should take exactly 15 minutes.
  - Each member of the group must present, so distribute your presentation evenly among the group.
  - Presentations are followed by 15-30 minutes of questions on the content of your code and general understanding of the lecture.
  - Finally, each member of the group must complete a 10-minute multiple-choice test on the content of the course.

- **We would like to draw your attention to the fact that the texts and data of the examinations are protected by copyright as written works, § 2 Para. 1 No. 1, Para. 2 UrhG.** The author (creator of the task) is also entitled to the sole right of reproduction and distribution of his work, § 15 No. 1 and No. 2 in conjunction with. §§ 16, 17 para. 1 UrhG. The copy handed over to you is therefore exclusively for your personal use. Any transfer to third parties is not permitted. Any infringement of copyright shall give rise to claims for injunctive relief and damages pursuant to Section 97 (1), (2) UrhG and may also give rise to criminal prosecution, Section 106 (1) UrhG.

Grading Schema

- **50% Code and Questions**: You must submit a single functional Jupyter Notebook showing how you trained your (final) model(s), including data pre-processing, training and model evaluation. The notebook should be well structured and commented. Make sure that the steps you have taken and the reasons for each step are clear, e.g. by using Markdowns for documentation purposes (see for example the `Example_notebook.ipynb` on Moodle). The notebook should not just be an unstructured compilation of every single line of code your team has produced for this project. We will be asking each team member questions about the code and the general understanding.

- **50% Multiple Choice Test**: The code presentation will be followed by a 10 min multiple choice test on the course content.

The Exam can only be passed if both parts are passed (code & multiple choice test). If you can not answer a single question or if you have not contributed to the overall progress of the project, you will fail the exam!

Project Description

Protein Thermophilicity Prediction

Thermostable proteins are important tools in many biotechnological domains, such as enzyme engineering. For instance, enzymes working at higher temperatures can potentially accelerate chemical reactions. Thus, it is essential to properly identify naturally occurring or artificially generated thermostable proteins. However, testing the thermostability of proteins is cost- and time-intensive. To accelerate the identification process, machine learning techniques can be used for predicting thermophilic proteins (Zhang and Fang, 2006; Lin and Chen, 2011; Ahmed et al., 2022).
The main objective of this exam/project is to build predictive models to discriminate between thermophilic and non-thermophilic proteins. For this purpose, you will use an unpublished dataset consisting of more than 4500 proteins. We provide nearly 1000 physicochemical and structural measurements calculated from the protein sequence. A summary of all pre-computed descriptors can be found in Table 1.

Table 1: **Physicochemical and structural descriptors:** The descriptors can be divided into four general groups shown in the first column. For each descriptor the total number of features and the prefix used in the data file is shown.

Descriptor group	Descriptor	Prefix	Features
Residue composition	Amino acid composition	AA_	20
	Dipeptide composition	DP_	400
Sequence order effects	Pseudo amino acid composition	pAA_	23
	k-spaced pseudo amino acid property	k_pAA_	400
Physicochemical properties	Composition	CTDc_	21
	Transition	CTDt_	21
	Distribution	CTDd_	105
Basic descriptors	Weight	molW	1
	Charge	Ch_	3
	Polarity	pol, nonp	2
	Aromacity	arom	1
	Hydrophobicity	hydro	1
	Van der Waals Volume	vdWV	1

Available data: We provide a single ZIP file `data.zip` which contains all necessary data for this project. All data files are comma separated. The following files are included in the ZIP archive:

`protein_data.csv` This file consists of data for more than 4500 proteins. The first column contains the dataset specific unique protein identifier. The remaining columns contain almost 1000 pre-computed features which measure physicochemical and structural properties of the protein sequences. The first row is a header containing the names of all descriptors, which consist of the prefix given in table 1 and some amino acid or group identifiers if applicable (e.g., `AA_C`, `CTDd_solv_g3r100`, `k_pAA_CA`).

`thermophilicity_labels.csv` This file contains the labels for all available proteins. The first column of the data is the identifier of the protein and the second column contains the label (1 for thermophilic, 0 for non-thermophilic).

`unknown_protein_data.csv` Additional test dataset of proteins for which the true labels are hidden. The file has the same data format as the original `protein_data.csv` file. This file can be used to evaluate the predictions of your model by submitting the predictions to the AI leaderboard (see below).

`random_predictions.csv` Demo file to illustrate the data format for the leaderboard submission. The file has no header. The first column is the protein id and the second column your prediction. Columns are comma seperated.

AI leaderboard: We have created a webserver (http://10.152.16.10/ai/) that can be used to evaluate the predictions of your model based on the additional test dataset (the webserver is only available via VPN). Each group will be given a unique verification token during the course, which can be used to upload the predictions to the webserver. We then calculate various metrics, including accuracy, precision, recall and MCC. Teams are then ranked based on their MCC. This allows you to compare your team's performance with other groups, as well as how well your model generalizes to unseen data. **However, each team is limited to 10 submissions**. This is to avoid overfitting and over-engineering your model to the unknown data. Thus, the goal is to first train, evaluate and compare different models with the labeled data you have available.

Tasks & Objectives

The following bullet points should serve as a proxy of what to do. This is not an exclusive list nor a comprehensive list of tasks. They should help you to get started with the project:

1. Download the training data from Moodle and familiarize yourself with the dataset, its features and additional meta-information. Create some basic summary statistics and plots to get a better understanding of the data. Not all features might be useful or needed to fit the model.

2. Think about necessary data pre-processing steps.

3. Train and compare different models, tune your hyperparameters and evaluate how well your models generalise. How good is your model in terms of predictive power?

4. Evaluate and compare your final models and predictions on the unknown dataset using the webserver and leaderboard: `http://10.152.16.10/ai/`. Each team will be assigned a unique identification token for submitting predictions to the webserver. Please note that the leaderboard only accepts a total of 10 submissions per team. This should prevent that models are over-engineered towards the unknown dataset. We therefore recommend that you do most of your analysis based on the data you have. Once you think you have a good model you can test your model using the unknown dataset and submit your predictions to the leaderboard. The leaderboard ranks submissions based on MCC. The data format for submissions is a plain CSV file, without a header and should contain the following information: The first column is the unique identifier of the protein, the second column is your prediction (either 1 or 0) for thermophilicity. Columns are separated with a comma, as illustrated in the following example:

```
p_SD43,0
p_B6A9,1
p_6O4x,0
...
```

5. Write a final (single) Jupyter Notebook to demonstrate what you did. This notebook should not contain all the code you have produced during the project. It should serve as a demo, demonstrating what you did and why you did certain steps in a documented manner (e.g. showing the data pre-processing, training and evaluation of your comparison of models). You can use the `Example_notebook.ipynb` as an inspiration on how to structure your notebook.

6. Illustrate your results appropriately (e.g. in a plot).

7. Interpret/discuss your results.

8. Prepare your team to present the code. (Keep in mind that your presentation should take exactly 15 minutes!)

9. Prepare for the multiple choice test by studying the content of the lecture.

References

Ahmed, Z., Zulfiqar, H., Khan, A. A., Gul, I., Dao, F.-Y., Zhang, Z.-Y., Yu, X.-L., and Tang, L. (2022). ithermo: A sequence-based model for identifying thermophilic proteins using a multi-feature fusion strategy. *Frontiers in microbiology*, 13:790063.

Lin, H. and Chen, W. (2011). Prediction of thermophilic proteins using feature selection technique. *Journal of microbiological methods*, 84(1):67–70.

Zhang, G. and Fang, B. (2006). Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. *Process Biochemistry*, 41(3):552–556.