ML Engineer: Coding challenge

Introduction

The ML Engineer coding challenge involves investigating and modelling a dataset of sequence activity data, which is typical of the datasets that we generate during our enzyme engineering process.

We have developed methods to introduce amino acid mutations to our protein sequences to produce protein variants. We can develop libraries of these protein variants and screen them for an activity of interest.

This ultimately produces a dataset of sequence-activity relationships, and the dataset enclosed is a mapping of (amino acid / protein) sequence to thermostability; thermostability being a measure of the temperature at which a protein unfolds and ceases to be active. One of our engineering goals is to maximise thermostability, allowing our protein(s) to be used in industrial processes, where the operating temperature is higher than that at which a protein operates in the wild.

Goals

The goals for this task are as follows:

- Pre-process the data. What steps would need to be taken to generate an ML model to predict thermostability from the sequences? Are the sequences the same length? Do they need aligning or clustering, how would one encode these? Does the thermostability data need normalisation? Is there sufficient data to model? What clarifications would you need from the team that generated the data? These are just suggestions: the important thing to convey is how you would deal with such a dataset if you saw it for the first time?
- If possible, try to model the sequence-activity data. What methods, frameworks, tools would you use? Would you need additional data, and if so, how much and in what form? Again, the important thing to explain (with coded justifications) is the approach that you use / would use. It is not expected that you will generate a fully predictive model in the time available.

Deliverable

The deliverable for this task is a short report and examples of Python code that you have used in this investigation. This can take the form of either a short document and associated code, or a Jupyter Notebook explaining the steps that you take. An important consideration is the packaging of your work: if a readme is required to provide instructions on how to install dependencies and run the code, then please add this, along with any requirements.txt or environment definitions if using Conda, for example. The final deliverable can be provided as a zip file or a link to a publicly available Github repository.

General

We are very appreciative on the time that you have spent on your application, and do not wish to overburden you. Please try to limit yourself to around 4 hours of work. Again, just to emphasise, providing some well packaged code, and some means of understanding the approach that you have followed / would follow if more time was available is the most important take-home for us.

If you have any questions throughout this, then please feel free to contact us any time from Monday – Friday and we will get back to you as soon as possible. Mail Neil at neil@epochbiodesign.com if any questions arise.

The deadline for this is the end of Wednesday 24 April. Please submit your work to the above e-mail address by this time. Good luck!