|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Assignment  Cedric Hermans / MLOps@Home / 06-02-2022 |  |
|  |  |  |  |
|  | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  |  | TABLE OF CONTENTS |  |
|  |  | Introduction ..………....….….…………….….…….3  Method ..………..…..…..….….…….............…….4  Discussion …………..……..….….……..…….…….5  Results ..………..…....……..….….…..……….…….6 |  |



|  |  |  |
| --- | --- | --- |
| INTRODUCTION |  | |
| This project focuses on an AI model that can differentiate two different protein families based on the occurrence of amino acids in the protein sequences. The two proteins that will be studied in this assignment are of the non-ribosomal peptide synthase (NRPS) family and of the polyketide synthase (PKS) family.  Proteins are some of the basic building blocks of living organisms. They are in most cases biologically encoded by the DNA. The DNA first gets transcribed into RNA and finally is translated into proteins. Proteins consist out of 20 different amino acids each with specific properties. In bioinformatics, we write down the amino acids in a one-letter code. An example would be the amino acid valine that has the one-letter code V. An overview of all basic amino acids with their letter is given in Table 1.  Table 1: an overview of the 20 basic amino acids with their one-letter abbreviation   |  |  | | --- | --- | | **Amino acid** | **Single letter abbreviation** | | **Alanine** | **A** | | **Arginine** | **R** | | **Asparagine** | **N** | | **Aspartic acid** | **D** | | **Cysteine** | **C** | | **Glutamine** | **Q** | | **Glutamic acid** | **E** | | **Glycine** | **G** | | **Histidine** | **H** | | **Isoleucine** | **I** | | **Leucine** | **L** | | **Lysine** | **K** | | **Methionine** | **M** | | **Phenylalanine** | **F** | | **Proline** | **P** | | **Serine** | **S** | | **Threonine** | **T** | | **Tryptophan** | **W** | | **Tyrosine** | **Y** | | **Valine** | **V** |   Table 1: an overview of the 20 basic amino acids with their one-letter abbreviation   |  |  | | --- | --- | | **Amino acid** | **Single letter abbreviation** | | **Alanine** | **A** | | **Arginine** | **R** | | **Asparagine** | **N** | | **Aspartic acid** | **D** | | **Cysteine** | **C** | | **Glutamine** | **Q** | | **Glutamic acid** | **E** | | **Glycine** | **G** | | **Histidine** | **H** | | **Isoleucine** | **I** | | **Leucine** | **L** | | **Lysine** | **K** | | **Methionine** | **M** | | **Phenylalanine** | **F** | | **Proline** | **P** | | **Serine** | **S** | | **Threonine** | **T** | | **Tryptophan** | **W** | | **Tyrosine** | **Y** | | **Valine** | **V** |   The two protein families that will be used here are synthases. Meaning that they synthetize a specific molecule. PKS enzymes will produce polyketides. These are a large class of secondary metabolites that are mainly produced in bacteria, fungi and plants but have also been observed in a few animal lineages. Secondary metabolites are molecules that do not have an immediate role in the normal growth of the organism. They often play an important role in the defense against other organisms. A common example of secondary metabolites are toxins.  The NRPS family are a class of peptide secondary metabolites that are usually produced by microorganisms like bacteria or fungi. Normally, to produce a peptide, an mRNA molecule is needed that will be processed by a ribosome. As the name of the NRPS implies, there is no ribosome (or mRNA) required for the NRPS to produce a peptide. The non-ribosomal peptides that get formed by NRPS enzymes often have different properties than the peptides produced by a ribosome. For example, cyclization or specific modifications can occur what is normally not the case for normal peptides.  Both families often consist out of multiple domains. These are regions of the protein that are self-stabilizing and fold independently from the rest of the protein. This is often the case for parts of the protein that have a specific function in the workings of the protein. A 3D structure for one of the proteins of both families is given in Figure 1.  Due to an ongoing research project in the Bioinformatics Knowledge Center (BiKC). We are interested in the different domains that these PKS and NRPS families have and the substrates and products that each of these enzymes have. However, due to limited time this project will first set a baseline for a potential future application that can predict this based on the peptide sequence. Here the goal will be to be able to upload a protein sequence to predict if the sequence belongs to the PKS or NRPS family.  The model will be a basic dense Neural Network that was first trained on a personal laptop and afterwards transferred to Azure Machine Learning Studio to test out multiple hyperparameters so the best model can be picked. The best model will be picked and exported to the personal laptop where it will be integrated with an API (implemented with the FastAPI framework). Finally, it will be converted to a docker and stored on GitHub where it can be used to be deployed on a Kubernetes cluster. | |  |
| Figure : 3D structure of one protein from both protein families of interest. Left: a PKS from Streptomyces albus (PDB: 4OPF). Right: a NRPS from an unnames Streptomyces specie (PDB: 6LTA) | |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | |  | |
| Data retrieval  The data was downloaded from NCBI using a custom script called download\_sequences.py. All relevant scripts/documents can be found on GitHub through the following URL: <https://github.com/CedricHermansBIT/assignment_MLOps-Home>. NCBI is one of the largest and most important databases for biological data. It provides an API which we can access using URLs. There are some modules available that should provide us with easy access, however they do not provide all the options available to us if we would use the URLs directly. For that purpose, a custom script was made which allows us to download protein sequences regarding a specific keyword. Furthermore, the sequences were filtered so only Streptomyces proteins would be returned that belong to the RefSeq section of NCBI. This means that the sequences have gone through some form of quality control. We also do not allow partial sequences. The API requires us to first search for the IDs that are linked to our search term whereafter we can use these IDs to fetch the actual sequence.  The sequences are returned in FASTA format, the most common sequence format in bioinformatics. This means we always have a header line starting with the “>” symbol that contains information about the sequence (like organism or the name of the protein). The following lines contain the actual sequence data. An example of a part of such FASTA file is shown below.  >WP\_003946496.1 SDR family NAD(P)-dependent oxidoreductase, partial [Streptomyces albidoflavus]  EGADPAPCRGRTAPSPATADTTARAGALTRAVLEVPGVGDAVVLPGPDGEPATVYVVPNRAGAADRTEQV  VSSLAPGTRVVAISGLPRTAEGGLDEGALKDLPVIDQVAAGAWRERLARLPGVREAEVVLEEVPEELERR  HVGRPRAAGGAAEPDAPSVERPASVPALSEGPALPEPSVSGWAEALLRAAGRPDGEVVHVRADGSETRRS  YASLVPEASRVLAGLRRRGLRPGDRVILQCDDTEDFVATLWGCVLGGFVAVPLTVPVSYATTSAAVSKLE  Due to the sheer number of files that were used, these are not included in the repository. There are some sample files provided to test the API later on.  Each of the sequences was processed by first removing the header lines of the FASTA and then counting the occurrence of each amino acid in the sequence. This results in a single list of 20 values that can be used as input values for the model. An example of the results of the partial FASTA from above is given below. Note that the amino acid list is for your information, and normally not really show/saved.  ['A','R','N','D','C','E','Q','G','H','I','L','K','M','F','P','S','T','W','Y','V']  [ 44, 28, 1, 15, 3, 24, 3, 29, 2, 3, 25, 2, 0, 2, 27, 16, 15, 3, 3, 35]  Model structure and training (using Azure Machine Learning Studio)  Before starting to use the Azure Machine Learning Studio, the model was first constructed and tested on a local laptop (since I actually really dislike the AMLS because it’s so slow. I wanted to punch my monitor a few times, but I managed to not actually do it). The model that was used is a simple Dense Neural Network with three hidden layers each containing twenty neurons with relu activation functions. The input layer consists out of twenty values as well, each representing the count of one of the twenty amino acids in the peptide sequence. The output consists out of two output neurons each representing one of the two protein families. The activation function used here was sigmoid. The model was evaluated using the categorical crossentropy loss function. The structure of the model is shown in Figure 2.  Using the local machine, the adam optimizer was used, but this was switched out for an SGD optimizer on the AMLS. An additional change of the model on the AMLS was the addition of different callbacks. First off, a callback for saving the best model was added. Next, early stopping was implemented and finally reducing the learning rate on a plateau was added. The files used to use the AMLS were adapted from the files that were used for one of the assignments.    Figure : representation of the structure of the neural network used in this assignment. The opacity of the edges represents a fictional value for the weights and is not representative of the actual used model.  As the local machine was mainly used as a proof of concept, this won’t be discussed anymore in the further parts.  Data upload  The data was uploaded the AMLS using the interface. A blob storage was used for the FASTA files. For the NRPS family, 3224 sequences were used. These were uploaded individually. For the PKS family, 16413 sequences were used. However, due to crashing and slowing down of the web browser (and frustration because of this) it was decided that for PKS, the sequences would be “cat” to a single file that was uploaded to the dataset. This meant that the data preprocessing will have to be adapted for both families. The resulting datasets are shown in Figure 3.    Figure : The datasets that were obtained after uploading the data.  Data processing  For the data processing on the AMLS, the notebook 01\_DataPreparing.ipynb  Subheading | | |  |
| * Lorem ipsum dolor sit amet, consectetur adipiscing elit * Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. | * Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. * Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur | |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Discussion | | |  | |
| Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi | | | |  |
| Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Purus in massa tempor nec feugiat nisl pretium. Aliquam vestibulum morbi blandit cursus risus at ultrices mi. Ut morbi tincidunt augue interdum velit euismod in. Tortor posuere ac ut consequat semper viverra nam libero. Egestas sed sed risus pretium quam vulputate. Nunc mattis enim ut tellus elementum sagittis vitae et. Tristique senectus et netus et. Pulvinar neque laoreet suspendisse interdum consectetur libero id faucibus. Lacus laoreet non curabitur gravida arcu ac. Odio facilisis mauris sit amet massa. |  | Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. | |  |
| Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Purus in massa tempor nec feugiat nisl pretium. Egestas sed sed risus pretium quam vulputate. Nunc mattis enim ut tellus elementum sagittis vitae et. Tristique senectus et netus et. Pulvinar neque laoreet suspendisse interdum consectetur libero id faucibus. | | | |  |

|  |  |  |
| --- | --- | --- |
| Results |  | |
| Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.  Lorem ipsum dolor sit amet, cons ectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.  Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Purus in massa tempor nec feugiat nisl pretium. Aliquam vestibulum morbi blandit cursus risus at ultrices mi. Ut morbi tincidunt augue interdum velit euismod in. | |  |
| |  |  |  |  | | --- | --- | --- | --- | |  | Class 1 | Class 2 | Class 3 | | Experiment 1 | 90 | 70 | 85 | | Experiment 2 | 70 | 65 | 85 | | Experiment 3 | 85 | 80 | 60 | | |  |
|  | |  |