|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 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.  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. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.  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 | | |  |
|  | |  |