**Data Science and Machine Learning Assessment Report** [**– GitHub Link**](https://github.com/Sachacharlton/Data_Science_Assessment)

**Introduction**

Mupirocin is a *trans-*AT polyketide produced by soil-dwelling bacterium *Pseudomonas fluorescens* NCIB 10586*.* It is used as a topical antibacterial agent and is active against Gram-positive bacterial infections including methicillin-resistant *Staphylococcus aureus* (MRSA).1

Mupirocin consists of a mixture of pseudomonic acids (A-D), in varying degrees. Pseudomonic acid A, the major component consisting 90% of mupirocin, contains a monic acid moiety esterified by 9-hydroxynonanoic acid (9-HN), a saturated fatty acid (Figure 1).1 While the biosynthesis of mupirocin has been studied extensively over the years,2,3 there is still wide conjecture on the mechanism of 9-HN esterification. On all accounts, it is still not fully understood whether the 9-HN side chain is fully assembled pre-attachment or if a precursor 3-hydroxypropionate (3-HP) could be esterified to the monic acid and sequentially elongated. Divulgence of this esterification step is the focus of my research.



***Figure 1.*** *Structures of pseudomonic acids (A-D) that comprise mupirocin antibiotic*

A candidate enzyme that could be explored for esterification is MupB, encoded on the mupirocin biosynthetic gene cluster. In this assessment, working code has been generated to explore the *mupB* DNA sequence, as well as design primers for the expression of MupB in *Escherichia coli*, followed by using the code to calculate the GC content of *MupB* and design primers for PCR amplification.

**Methods and Results**

Using Biopython,4 the MupB FASTA file2 (accession no. AAM12910.1) was read in by employing the SeqIO.read function. An alternative FASTA file could be used in this code if the user desired a different protein.

Expression of proteins is obfuscated by the redundancy, also known as degeneracy, of codons; whereby the same amino acid residue can be encoded by synonymous codons.5 Therefore, after defining the MupB protein sequence (mupB\_seq), a dictionary was set up of *E. coli* optimised codons, codons that can be translated faster by the ribosome than others,5 for MupB expression in *E. coli.* An empty list of mupB\_codons was filled by a loop function running through the different amino acids in the mupB\_seq and then assigning them an optimised codon, with a new string (codon\_output) defined by joining all mupB\_codons generated.

A picture containing text

Description automatically generatedTo check that these optimised codons returned by the code would be correctly transcribed and translated to the amino acids present in the desired MupB protein sequence, an ‘if’ statement was generated, using the mupB\_protein defined by transcribing the codon\_output to RNA, and then translating this RNA to protein (Figure 2).

Graphical user interface, text

Description automatically generated with medium confidenceBy defining a function to determine GC content, the MupB GC content was found to be 60% (Figure 3).

***Figure 2.*** *Output of the ‘if’ statement showing that the optimised codon dictionary and loop have correctly assigned codons for mupB expression*

***Figure 3.*** *Output of GC content function defined for MupB, found to be 60%*

Code was then used to compliment the generated optimised codon\_output, to find the reverse strand for MupB expression (compliment\_mupB). The forward and reverse MupB DNA strands were then used to find suitable primer sequences, by selecting the first and last 6 codons on the forward and reverse strand, respectively. These primer sequences can be edited/appended in later lines of code to include enzyme restriction sites. By using the same GC content function described above, the GC content of the forward and reverse primer strands were determined as 56% and 67%, respectively.

Finally, by setting up a dictionary of the different amino acid molecular weights, and setting up a loop filling the empty list mupB\_mw, the total molecular weight of MupB was determined. Initially, the code generated an error as the list of molecular weights in mupB\_mw were not defined as integers but after transforming this list to integers, the total sum of mupB\_mw was found to be 38354 Da.

**Conclusions and Outlooks**

Working code was successfully generated to explore the protein MupB and design suitable DNA sequence and primer strands for MupB expression in *E. coli.*

Using a defined GC function, the GC content of MupB and the pair of primers were determined. Finally, by using a dictionary of amino acid molecular weights and setting up a loop function, the total molecular weight of MupB was determined.

Future work could focus on editing the primer sequences in the code to find strands with more complimentary GC contents, as this can affect the melting temperature (Tm) of the primers. Additionally, code that employed the MupB protein sequence in a list, and could edit the individual amino acids (by position i.e. [0]) could be used to generate libraries of MupB mutants for expression, which could then also employ the above code for PCR amplification.

**References**

1 C. M. Thomas, J. Hothersall, C. L. Willis and T. J. Simpson, *Nat. Rev. Microbiol.*, 2010, **8**, 281–289.

2 A. K. El-sayed, J. Hothersall, S. M. Cooper, E. Stephens, T. J. Simpson and C. M. Thomas, *Chem. Biol.*, 2003, **10**, 419–430.

3 S. Gao, J. Hothersall, J. Wu, A. C. Murphy, Z. Song, E. R. Stephens, C. M. Thomas, M. P. Crump, R. J. Cox, T. J. Simpson and C. L. Willis, *J. Am. Chem. Soc.*, 2014, **136**, 5501–5507.

4 P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck and F. Kauff, *Bioinformatics*, 2009, **25**, 1422–1423.

5 P. S. Spencer and J. M. Barral, *Comput. Struct. Biotechnol. J.*, 2012, **1**, e201204006.