Final project

# Problem statement

Antimicrobial or antibiotic resistance (AMR) is a serious world-wide problem. Microbes are becoming resistant to the antimicrobials and antibiotics that we use against them. Typically a microbe becomes resistant to antimicrobials and antibiotics by acquiring new mutations or new genes.

The following story is based on real and imagined events. Some scientists in an animal farm in country X were monitoring antibiotic resistance in bacteria isolated from animals. They found one bacterial isolate which was resistant to Colistin - an antibiotic used as a last resort after bacteria have become resistant to all other antibiotics! This is a serious problem…

The scientists wanted to study which genes/proteins were responsible for the antibiotic resistance. So they isolated the DNA from the bacteria, and found a type of DNA known as plasmid. They got it sequenced in a lab, and that sequence is given to you. You have to find if there is a gene in the plasmid that may be responsible for the resistance to Colistin.

Report all your analysis with justification of the methods used, and appropriate technical parameters and interpretation.

# Executive summary

The report investigates naturally occurring colistin-resistant bacteria. The plasmid sequencing of the given bacteria is examined in the report to see if the bacteria have a gene or protein that makes them resistant to colistin. In order to do so, analyze the full genome sequence using the ORF finder to identify the genes that are present. According to the parameters set, a total of 30 genes were found by the ORF finder.

To confirm the legitimacy of the genes discovered through orf sequencing, we utilized BLAST to determine whether they actually coded for proteins. Our findings indicated that not all genes were responsible for protein production, as evidenced by our alignment with the protein database. Using the BLAST results, we linked the proteins to their functions in the uniprot database. Our aim was to identify the gene responsible for colistin resistance, so we researched colistin and its resistance. Our investigation revealed that colistin is a type of polymyxin antibiotic, and we used this information to pinpoint the gene responsible for resistance, which we named "Phosphoethanolamine transferase EptA or Polymyxin resistance protein PmrC." The name itself implies that this gene plays a role in polymyxin resistance. Colistin functions by interacting with the membrane and entering bactericidal cells, so Polymyxin resistance protein PmrC modifies the bacterial outer membrane by altering the LPS and reducing its negative charge, which is essentially involved in the membrane's biogenesis.

# Overall Approach

The objective is to determine the function of a protein based on its plasmid DNA sequence. To accomplish this, the given plasmid DNA sequence was analyzed using the ORF finder to extract the genes. The ORF finder is designed to identify the start and stop codons of genes, allowing for the identification of potential genes in the sequence.

To verify that the predicted genes would actually produce functional proteins, we utilized BLAST to compare the gene sequence against a database of known protein sequences. It is possible for the query sequence and the aligned sequence to share common ancestry, resulting in shared sequences.

Once verified, we predicted the function of the protein by linking the accession number to the uniprot database.

# Introduction

The rise of antibiotic-resistant bacteria is a growing concern in public health. In particular, colistin-resistant bacteria pose a significant threat as it is considered a last-resort antibiotic for multidrug-resistant bacterial infections. This report investigates the naturally occurring colistin-resistant bacteria and examines the plasmid sequencing of the given bacteria to identify the gene or protein that makes them resistant to colistin. The analysis was carried out by analyzing the full genome sequence using the ORF finder to identify the genes present. The report also explores the confirmation process of the legitimacy of these genes through BLAST to determine their protein-coding ability. The ultimate goal is to identify the gene responsible for colistin resistance and understand how it operates. The findings reveal that the named gene "Phosphoethanolamine transferase EptA or Polymyxin resistance protein PmrC" plays a crucial role in polymyxin resistance by modifying the bacterial outer membrane through alteration of the LPS and reduction in its negative charge, which is involved in the membrane's biogenesis. The report provides insight into the mechanisms of antibiotic resistance and highlights the importance of continued research in this area.

# Colistin

Colistin is a polymyxin antibiotic used to treat bacterial infections caused by susceptible Gram negative bacteria.

**Colistin is a polymyxin antibiotic agent.** Colistin is a Cyclic polypeptide antibiotic from Bacillus colistinus. It is composed of **Polymyxins E1 and E2** (or Colistins A, B, and C) which act as detergents on cell membranes.

Colistin is polycationic and has both hydrophobic and lipophilic moieties. It interacts with the bacterial cytoplasmic membrane, changing its permeability. This effect is bactericidal. There is also evidence that polymyxins enter the cell and precipitate cytoplasmic components, primarily ribosomes.

Except for the treatment of multidrug-resistant pulmonary infections in individuals with cystic fibrosis, parenteral polymyxin use was mostly abandoned with the development of less toxic medications, such as extended-spectrum penicillins and cephalosporins.

Yet more recently, the rise of multidrug-resistant gram-negative bacteria like Pseudomonas aeruginosa and Acinetobacter baumannii, as well as the scarcity of new antimicrobial drugs, have brought the usage of polymyxins back into fashion. [[1]](https://go.drugbank.com/drugs/DB00803)

# Colistin resistance

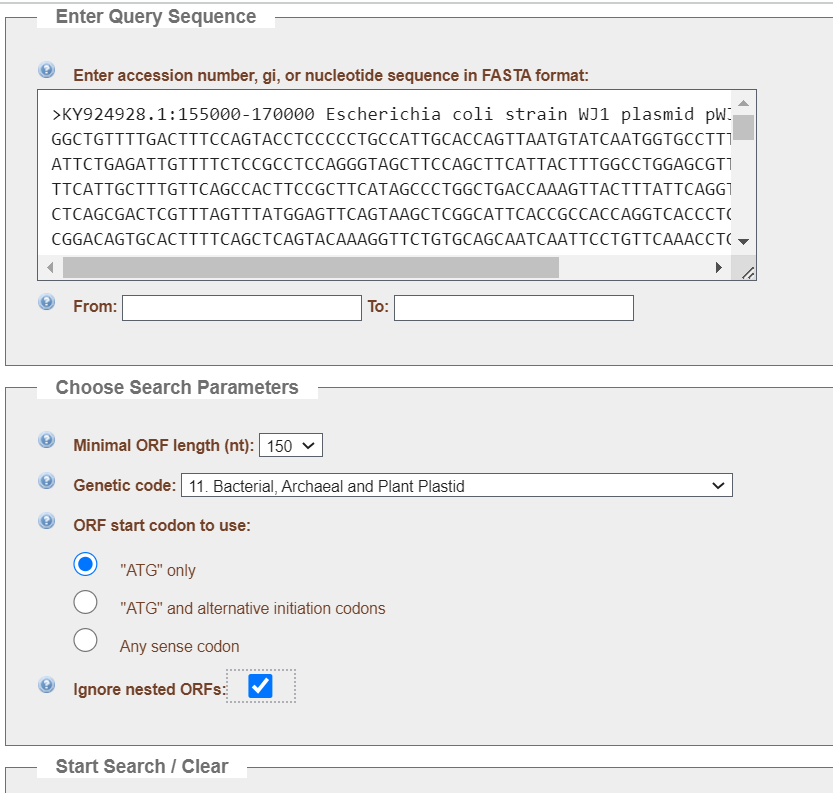
Colistin resistance mechanism occurs by chromosomal modulations similar to bacteria that are naturally resistant to colistin. The various molecular mechanisms have been determined, and the most common modifications occur via cationic groups (l-Ara4N and pEtN) to the lipid membrane of bacterial strains. [[2]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8406936/#:~:text=Colistin%20resistance%20mechanism%20occurs%20by,strains%20(14%2C%2035).)

# Orf finder

ORF finder is an online bioinformatics tool used to identify open reading frames (ORFs) in nucleotide sequences. An ORF is a region in a genome that has the potential to encode a protein. ORFfinder helps to identify potential genes and corresponding amino acid sequences in a genome. The tool provides information on the location and orientation of the ORFs, which can be used to extract the nucleotide and amino acid sequences of the gene.

The input for ORF is the bacterial plasmid genome sequence. I am using genetic code 11- Bacterial, Archaeal and Plant plasmid. The given plasmid sequence is of a bacterial hence using the genetic code 11. The minimal ORF length chosen was 150 nucleotides just to minimize the ORF results.

We ignored the nested ORFs because it was counting the small ORFs which were detected in the large genes.



Total ORFs found are 30 where after the amino acid length of 94 there are no blast results hence excluding all the sequences which lie below the threshold of 95 amino acids.

## Why are all ORF sequences not genes?

All the orf finder predicted genes are not considered as genes because the orf finder uses the start and stop codon sequences as parameters to predict the genes. We can not be hundred percent sure that all the orf sequences are genes as the gene comprises many other parts such as promoters, enhancers, transcription start site (TSS)and termination signals or sites. There is a possibility that even if the ORF predicted gene has start and stop codon they may not have promoter or TSS or termination site which will not lead to a gene sequence.

If we ensure the predicted gene has a promoter region or termination site then we can say with more certainty that the predicted gene is an actual gene.

## How do you ensure the predicted ORF gene will provide the real protein?

We are not sure that the predicted gene will make the real protein in the cell. For protein synthesis one of the main mechanisms is ribosome machinery but for the machinery to work it has to come and attach to the strand. The ribosome determines the signal which is present on the strand called the “kozak sequence”. If we detect the kozak sequence in the beginning of the sequence then we can be more confident that the ribosome will attract and if the ribosome attaches then it will continue the protein synthesis.

This prediction we can make on the basis of ORF sequences.

# BLAST

BLAST stands for Basic Local Alignment Search Tool which uses local alignments or similarity between the query and database sequences to provide the closely related sequence.

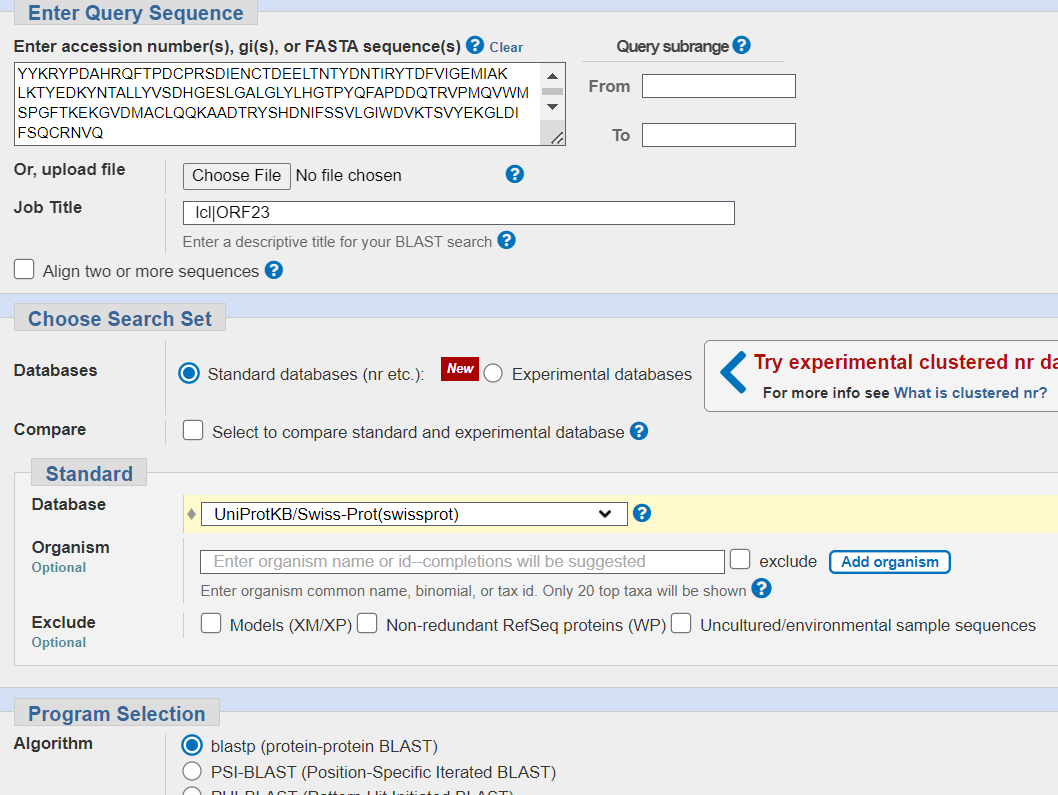
## Inputs given to Blastp:

The main data input was the protein sequence obtained from ORF predicted gene sequences.

The database we chose was “**UniProtKB/Swiss-Prot**”. The data in UniProtKB/Swiss-Prot is the protein sequence from Uniprot which is from translations of CDS submitted to the EMBL-Bank/GenBank/DDBJ nucleotide sequence resources (International Nucleotide Sequence Database Collaboration (INSDC)). Those sequences are generated using gene prediction programs or are experimentally proven.

Swiss-Prot (created in 1986) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.

Thus UniProtKB/Swiss-Prot is a correct database for proteins which are found from experimental results.



Other algorithm parameters are specified such as **max target sequences are 100**. As we need all the aligned sequences and there is no mid way option between 10-100. Thus choosing the max target to be 100 which will be helpful later during the identification of mutations.

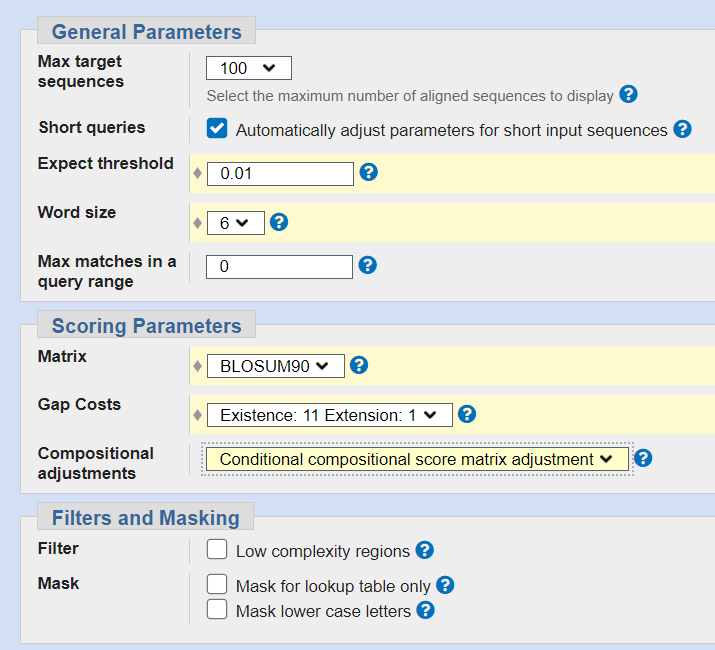
The **expected threshold or E value is set low to 0.01** to avoid the number of matches found by accident. The lower the E-value, or the closer it is to zero, the more "significant" the match is. [[3]](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ)

The **word size is set to 6.** Word size is the size of seed to initiate the alignment.The word size 6 will provide more accurate results.

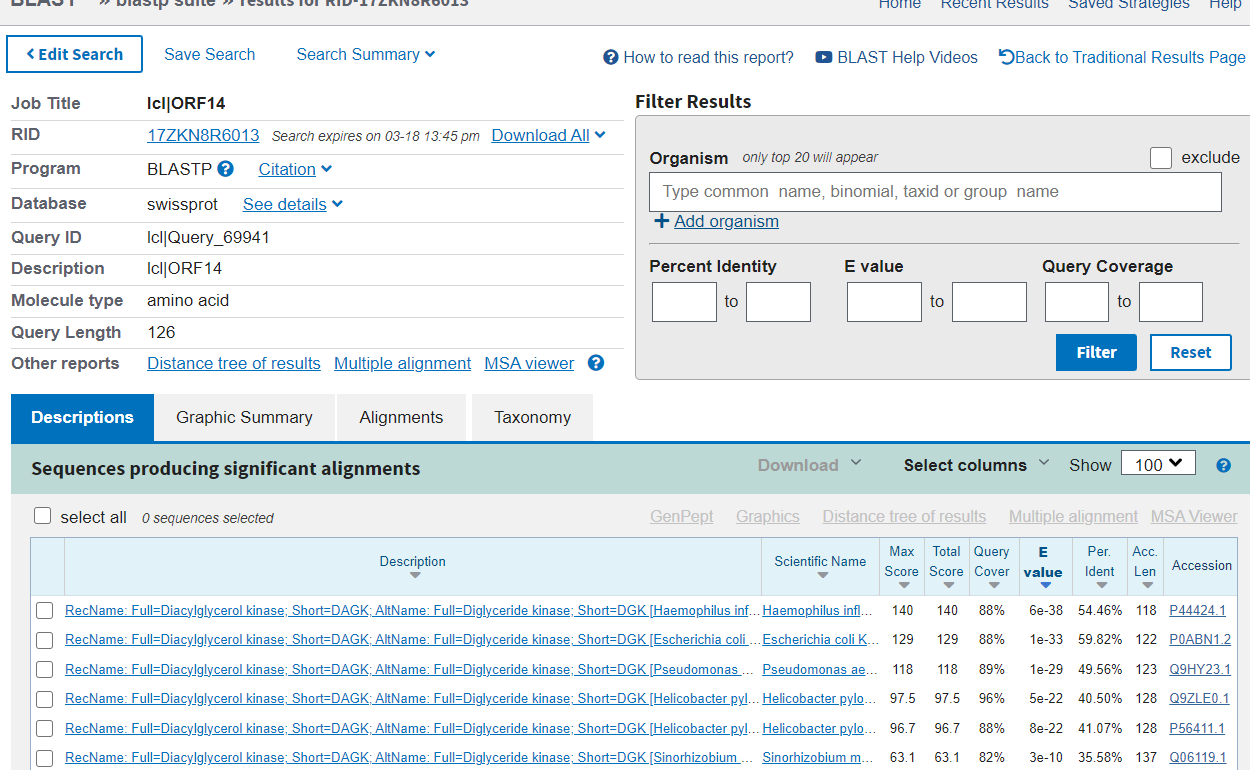
The **scoring matrix chosen for the search is BLOSUM90.** This matrix is chosen to get more specific results as close and accurate as possible sequences will be shown.

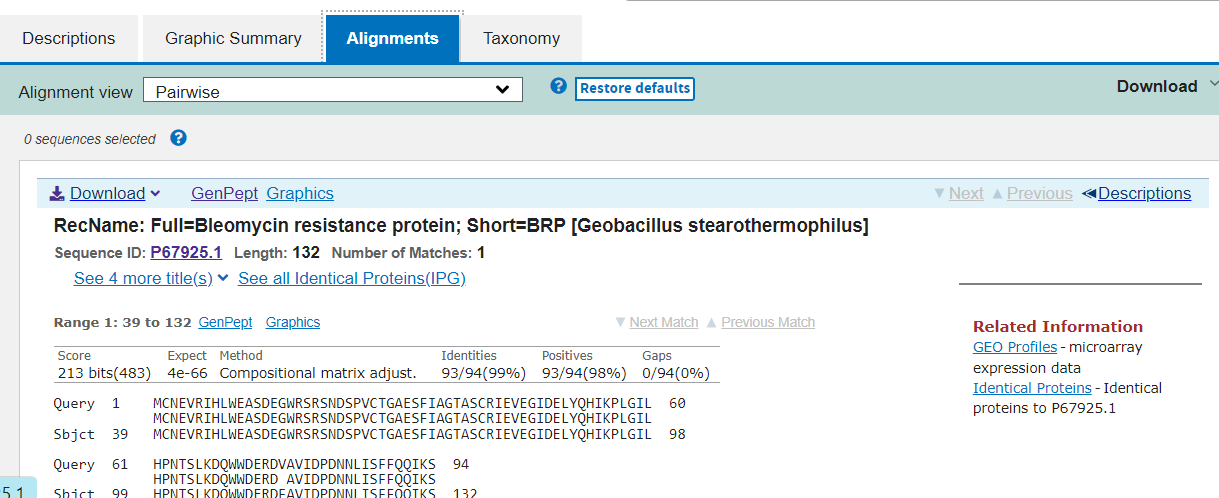
Gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. The gap cost helps to keep in check the alignment score. Higher the gap cost lower the alignment score. The **inserted gap score is the maximum which is Existence:11 Extension:1.**

**The compositional adjustments are conditional compositional score matrix adjustments.**



**Output**





### Output result page description:

The output results is a well designed report of different constituents of the blast protein program. The result page has a **description column** which is the description of the organism matched from the database followed by the **scientific name** of the matched organism.

There is a column of max score and total score where **max score** is the highest bit score that is calculated from alignment matches and mismatches.The bit score represents the size of a sequence database needed for the current match to be expected by chance. It is based on the initial alignment score and reflects the quality of the alignment, with higher scores indicating better alignment.The **total score** refers to the sum of the alignment scores of all of the segments from the sequence.[[4]](https://chanzuckerberg.zendesk.com/hc/en-us/articles/360050963352-A-guide-to-BLAST#:~:text=Max%20score%3A%20the%20highest%20bit,score%2C%20the%20better%20the%20alignment.)

**Query cover** is the percentage of the query aligned with the database sequence. High query cover is the indication that the result sequence has highly aligned with the query sequence. **E value** is the value or score of how much the aligned sequence is by accident. Lesser the E value the better the results, higher the E score it tells that the aligned sequence has more chances that they are accidentally aligned which may lead to conclude that the sequence may not belong to the same taxon.

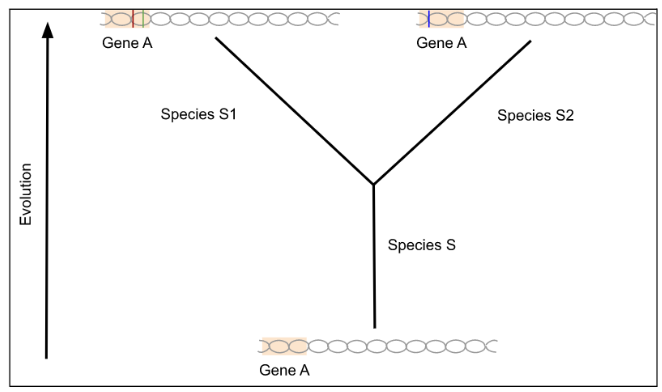
**Percentage identity** is the measure of how much percent of the query sequence is identical to the sequence in the database. **Accession length** is the number of nucleotides or amino acids in the result sequence identified by the **accession number**(a unique identifier assigned to records in the NCBI databases). [[5]](https://guides.lib.berkeley.edu/ncbi/blast#:~:text=Acc%5Bession%5D%20Len%5Bgth,records%20in%20the%20NCBI%20databases)

**Identities** give the total number of letters which are exactly matched to the sequence. **Positives** is the measure of the total number of residues which are exactly identical or have similar properties. **Gaps** is the total number of gaps in the amino acid sequence.

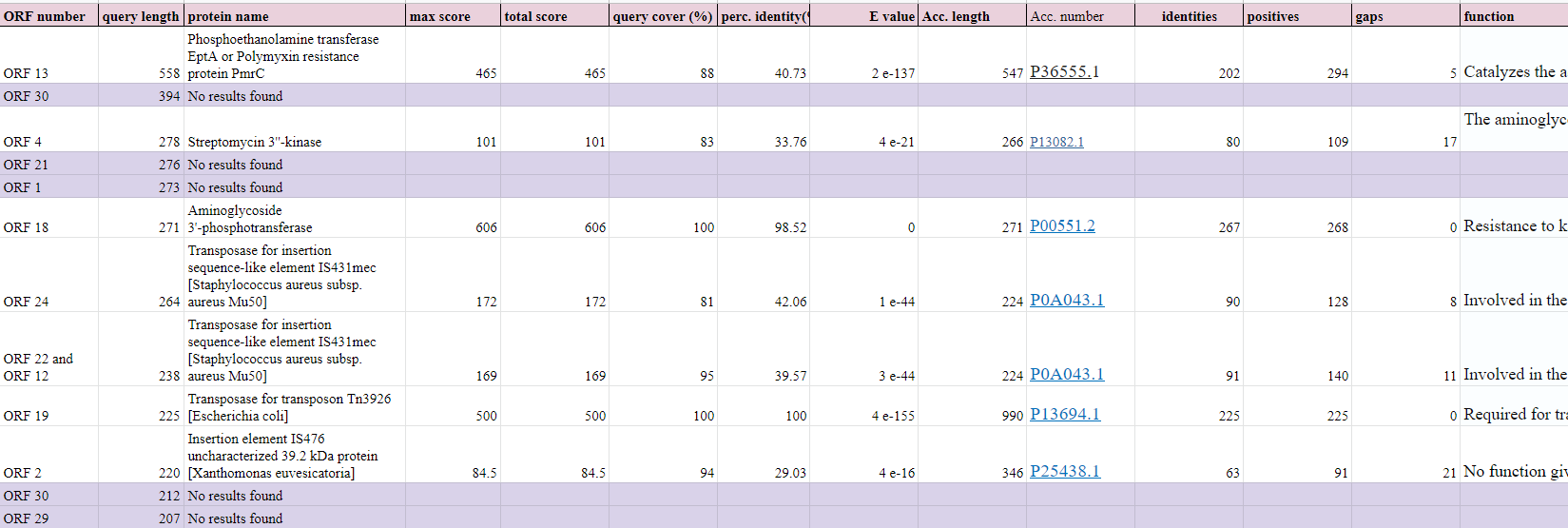
# Why are we using the matched result from blast to describe the function of our protein?

**Orthologs** are genes that developed through speciation from a shared ancestral gene. This means that when a certain species evolved throughout the years thereby having its genes vary throughout their progeny there may be variations of the same genes as the ancestor. For example, as given below, Species S(the ancestral species) has a gene by the name of Gene A. Over millions of years as it evolves, even if Species S1 and Species S2 are physiologically very different, we can tell that they are related to Species S and each other by the presence of Gene A(Gene A may have it’s variations as well, but overall should be the same).

By applying this concept, we are using the results from blast because we are trying to find related genes and subsequently proteins from our origin since the virus and its relatives may share a common ancestry and thus variations of the same genes and proteins. Allowing us to predict these proteins’ function and structure.



## Protein prediction Results -Blastp



[Excel link](https://docs.google.com/spreadsheets/d/1SdEraRpHOGelFLm2TxgXxhzneAFKres8elnVWcAbKZ0/edit?usp=sharing)

# Protein Responsible for Colistin Resistance

From all the proteins, the protein which can be responsible for colistin resistance is the polymyxin resistance protein. All the other proteins found their functions were different and do not showcase any visible or direct relationship in acquiring resistance.

The protein responsible for colistin resistance in bacteria is **Phosphoethanolamine transferase EptA or Polymyxin resistance protein PmrC .**

The function of the protein is to catalyze the addition of a phosphoethanolamine moiety to the lipid A. The phosphoethanolamine modification is required for resistance to polymyxin. It biogenesis the Bacterial outer membrane; LPS lipid A biosynthesis. [[6]](https://www.uniprot.org/uniprotkb/P36555/entry#function)

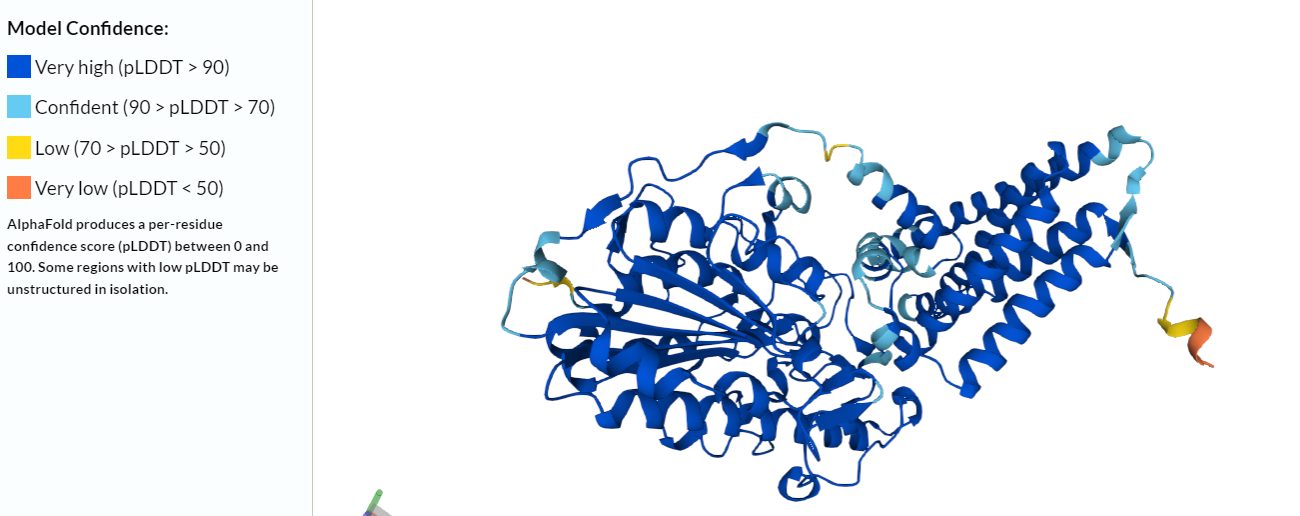
Resistance to colistin occurs with LPS modification via different routes. This protein is responsible for the modifications of the bacterial outer membrane through alteration of the LPS and reduction in its negative charge. [[7]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519339/#:~:text=Resistance%20to%20colistin%20occur%20with,reduction%20in%20its%20negative%20charge.&text=The%20other%20strategy%20is%20the%20overexpression%20of%20efflux%2Dpump%20systems.)

## Plasmid -mediated colistin resistance

Plasmid-mediated colistin resistance is the resistance acquired by the transfer of plasmid from one bacteria to another.

The mcr genes are responsible for horizontal transfer of colistin resistance. These plasmid-mediated genes were first reported in E. coli. Later on three mobile colistin-resistance genes (mcr6, mcr7, and mcr8) were discovered in 2018. Out of which mcr8 was similar to mcr1, as PETN leads to the addition of **phosphoethanolamine** to lipid A, followed by colistin resistance. [[8]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519339/)

**Alpha fold predicted structure of the protein**



# Conclusion

In conclusion, this report provides a detailed analysis of naturally occurring colistin-resistant bacteria and the mechanisms behind their resistance. Through the use of plasmid sequencing, ORF finder, and BLAST analysis, we were able to identify the specific gene responsible for colistin resistance, as well as the other genes and proteins present in the bacteria. The findings of this study highlight the importance of understanding antibiotic resistance mechanisms and the need for continued research in this area to develop effective treatments. Ultimately, this research may contribute to the development of new antibiotics or alternative treatments to combat antibiotic-resistant bacteria and improve public health.

# References

1. <https://go.drugbank.com/drugs/DB00803>
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8406936/#:~:text=Colistin%20resistance%20mechanism%20occurs%20by,strains%20(14%2C%2035)>.
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519339/>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519339/#:~:text=Resistance%20to%20colistin%20occur%20with,reduction%20in%20its%20negative%20charge.&text=The%20other%20strategy%20is%20the%20overexpression%20of%20efflux%2Dpump%20systems>.
5. <https://www.uniprot.org/uniprotkb/P36555/entry#function>
6. <https://guides.lib.berkeley.edu/ncbi/blast#:~:text=Acc%5Bession%5D%20Len%5Bgth,records%20in%20the%20NCBI%20databases>
7. <https://chanzuckerberg.zendesk.com/hc/en-us/articles/360050963352-A-guide-to-BLAST#:~:text=Max%20score%3A%20the%20highest%20bit,score%2C%20the%20better%20the%20alignment>.
8. <https://chanzuckerberg.zendesk.com/hc/en-us/articles/360050963352-A-guide-to-BLAST#:~:text=Max%20score%3A%20the%20highest%20bit,score%2C%20the%20better%20the%20alignment>.

Part II

# Problem statement

1. Now that you have identified the gene(s) which may be responsible for the resistance to Colistin, can you find which bacteria have a similar gene?
2. The scientists in the previous study sequenced some more DNA samples, and found that they also contained one of the genes related to Colistin resistance. They aligned the gene sequences using MSA programs. The alignment is given to you. Can you find if the gene sequences have any variation, and which of the variations should be prioritized for further study at structural level?

For both - justify your approach, tools and databases used, and interpretation of results.

# Overall approach

Using BLASTn, all potential aligned sequences with related genes were obtained in order to determine whether bacteria have similar colistin resistance genes. This is done under the assumption that the evolved species may have shared a common ancestor with similar genes during speciation. By examining certain indicators like the E value, percentage identity, and query coverage, we can infer from the BLAST findings that these bacteria may or may not share comparable genes. Lowest E value and highest % identity and query coverage are used to group similar bacteria.

Then, we utilized the Jalview tool for MSA visualization to investigate the variation. Observed variations were noted and then ranked based on the characteristics of amino acids and change in type of substitution such as transition and transversion.

# BLASTn

The "blastn" program searches for nucleotides using BLAST[Basic local Alignment Sequence Tool] . The sensitive general-purpose "blastn" tool can be used to align tRNA or rRNA sequences, as well as mRNA or genomic DNA sequences comprising a mixture of coding and noncoding regions.

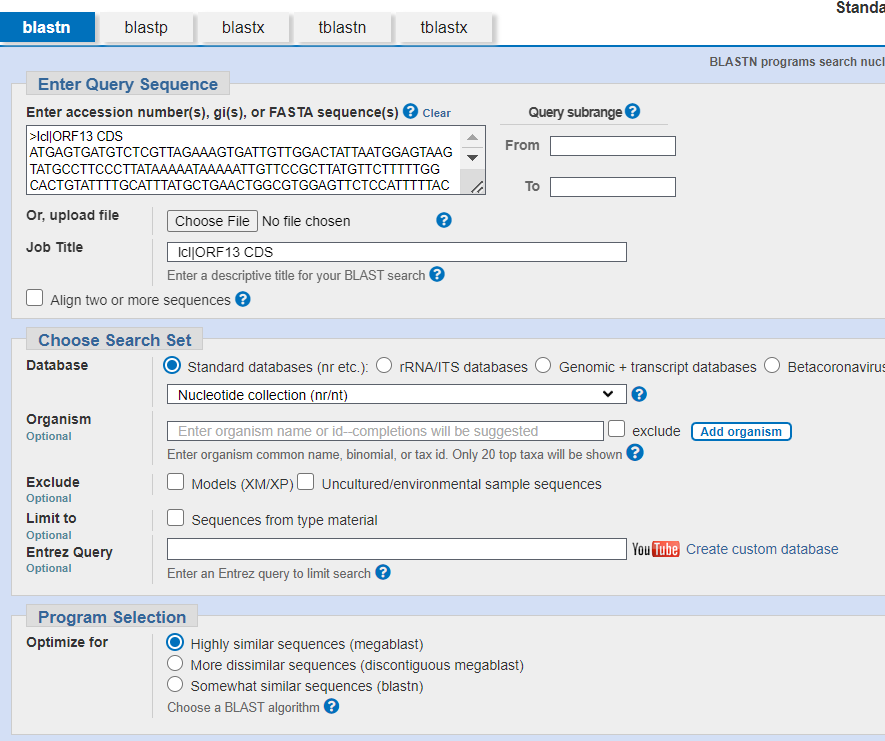
### Inputs :

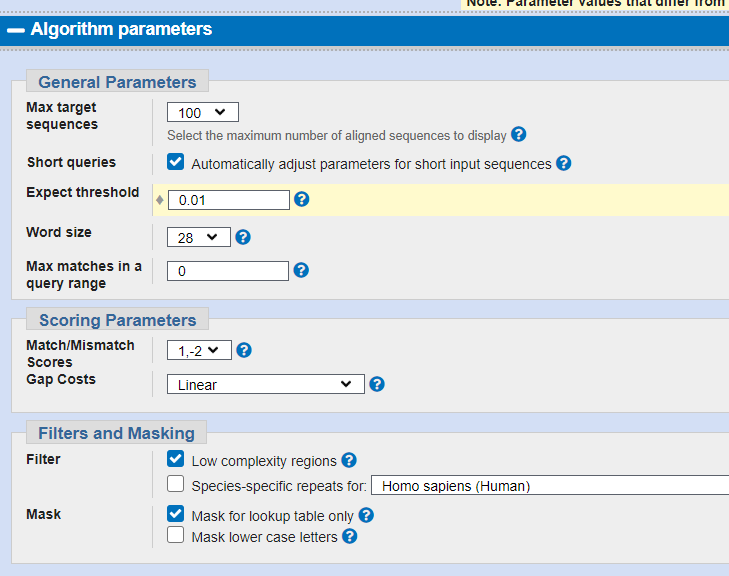
The input sequence given was the nucleotide sequence of the polymyxin resistance protein PmrC(protein responsible for colistin resistance). The **database used was Nucleotide collection (nr/nt)** with **max target of 100** sequence As we need all the aligned sequences and there is no mid way option between 10-100. Thus choosing the max target to be 100 which will be helpful later during the identification of mutations.

The **expected threshold or E value is set low to 0.01** to avoid the number of matches found by accident. The lower the E-value, or the closer it is to zero, the more "significant" the match is. [[3]](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ)

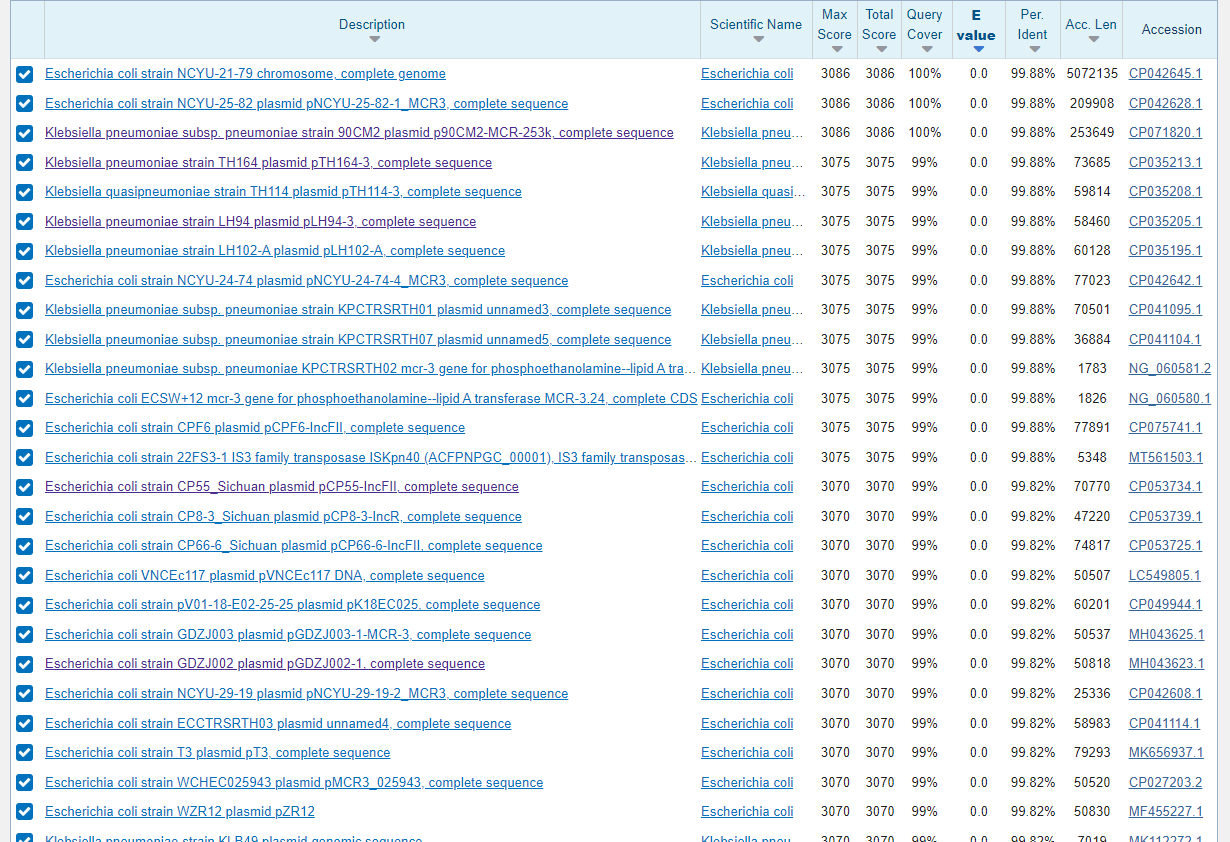
**Word size was chosen as default- 28** as for nucleotide sequence accuracy the number should be higher as there are only four letters possible - ATGC which is very likely to find easy combinations. 28 is a best size to create a balance between accuracy and speed of the BLAST.

The match score was 1 and mismatch score was -2 with gap cost as linear.





Output



## Colistin -resistance similar gene possessing bacteria

There are many bacterial strains from Escherichia coli and Klebsiella pneumoniae that have similar colistin resistance genes.

List of all the bacteria which have similar colistin-resistance gene:

* Escherichia coli strain NCYU-21-79 chromosome
* Escherichia coli strain NCYU-25-82 plasmid pNCYU-25-82-1\_MCR3
* Escherichia coli strain CP55\_Sichuan plasmid pCP55-IncFII
* Escherichia coli strain GDZJ002 plasmid pGDZJ002-1
* Klebsiella pneumoniae subsp. pneumoniae strain 90CM2 plasmid p90CM2-MCR-253k
* Klebsiella pneumoniae strain TH164 plasmid pTH164-3
* Klebsiella pneumoniae strain LH94 plasmid pLH94-3
* Klebsiella pneumoniae strain KLB38 plasmid genomic sequence

[**Note**: all the bacteria having similar gene sequence their file has been attached with the deliverable.]

## How do we decide that all the above mentioned bacteria share similar colistin resistance genes?

### Why do we use BLAST?

The BLAST tool allows us to match our query sequence against a database of pre-existing sequences. This comparison enables us to identify genes and their sequences that closely match our query, giving us information about our closest genetic matches.

Orthologous genes are those that share a common ancestor through the process of speciation, and often have similar functions in different organisms. BLAST can detect homologous sequences, such as orthologs, by assessing their similarity to a given query sequence.

### What are the criteria to state the similarity of genes on the basis of BLAST’s result?

There are three criterias on the basis of which we can confidently state the similarities:

1. **E value or Expect value**

E value is the value or score of how much the aligned sequence is by accident. Lesser the E value the better the results, higher the E score it tells that the aligned sequence has more chances that they are accidentally aligned which may lead to conclude that the sequence may not belong to the same taxon.

1. **Percentage Identity**

Percentage identity is the measure of how much percent of the query sequence is identical to the sequence in the database.

1. **Query coverage**

Query cover is the percentage of the query aligned with the database sequence. High query cover is the indication that the result sequence has highly aligned with the query sequence.

For “Escherichia coli strain NCYU-21-79 chromosome” and “Escherichia coli strain NCYU-25-82 plasmid pNCYU-25-82-1\_MCR3, ” the query cover is 100% , percentage identity is 99.88% and E value is 0 which states that the gene sequence of those two matches to the query sequence with 100% accuracy stating that hit by chance is zero and it is 99.88% identical to the pre-existing sequence and all the nucleotide of the query sequence is covered in the resultant sequence.

Whereas all the remaining sequences have 99% query coverage with zero E value and percentage identity in the range of 99.88 to 99.70 %. Those values are reliable and can be considered as similar genes possessing bacteria.

# Identifying Variations

# Jalview

The aligned mutation sequence was analyzed using jalview. Jalview is a platform for MSA editing, visualization and analysis. [[1]](https://www.jalview.org/)

**Allotted search Position: 900 - 1000**

There were no nucleotide variations observed in samples 1 to 6. Only samples 7 and 8 have variation in nucleotides. Those variations are mentioned below.

| **Sample 7 and Sample 8** | | | | |
| --- | --- | --- | --- | --- |
| **Position** | **Original nucleotide** | **mutated nucleotide** | **original amino acid** | **mutated amino acid** |
| 909 | C | T | G | G |
| 910 | C | T | L | L |
| 913 | C | T | L | L |
| 927 | A | G | Q | Q |
| 937 | A | G | I | V |
| 993 | A | T | V | V |

# Variation Prioritization

## Nucleotide variation prioritization

The nucleotides are prioritized on the basis of purines and pyrimidines and transition and transversion.

**Purines** : They are compounds which consist of two rings. Adenine and Guanine are purines.

**Pyrimidines** : They are compounds which consist of one ring. Cytosine and Thymine are pyrimidines.

**Transition**: substitution of purine with purine and pyrimidines with another is Transition.

**Transversion**: The substitution of a purine with a pyrimidine or vice versa is called Transversion.

From the above variation table :

The variation at position 993; A —-> T or Adenine to Thymine is Transversion. As purine is getting converted into pyrimidine.

Here the size of the molecule is changing. From two rings it changed to a single ring molecule.

In all the other variations observed they are transitions where purines are changing to another purine and pyrimidine to another pyrimidine.

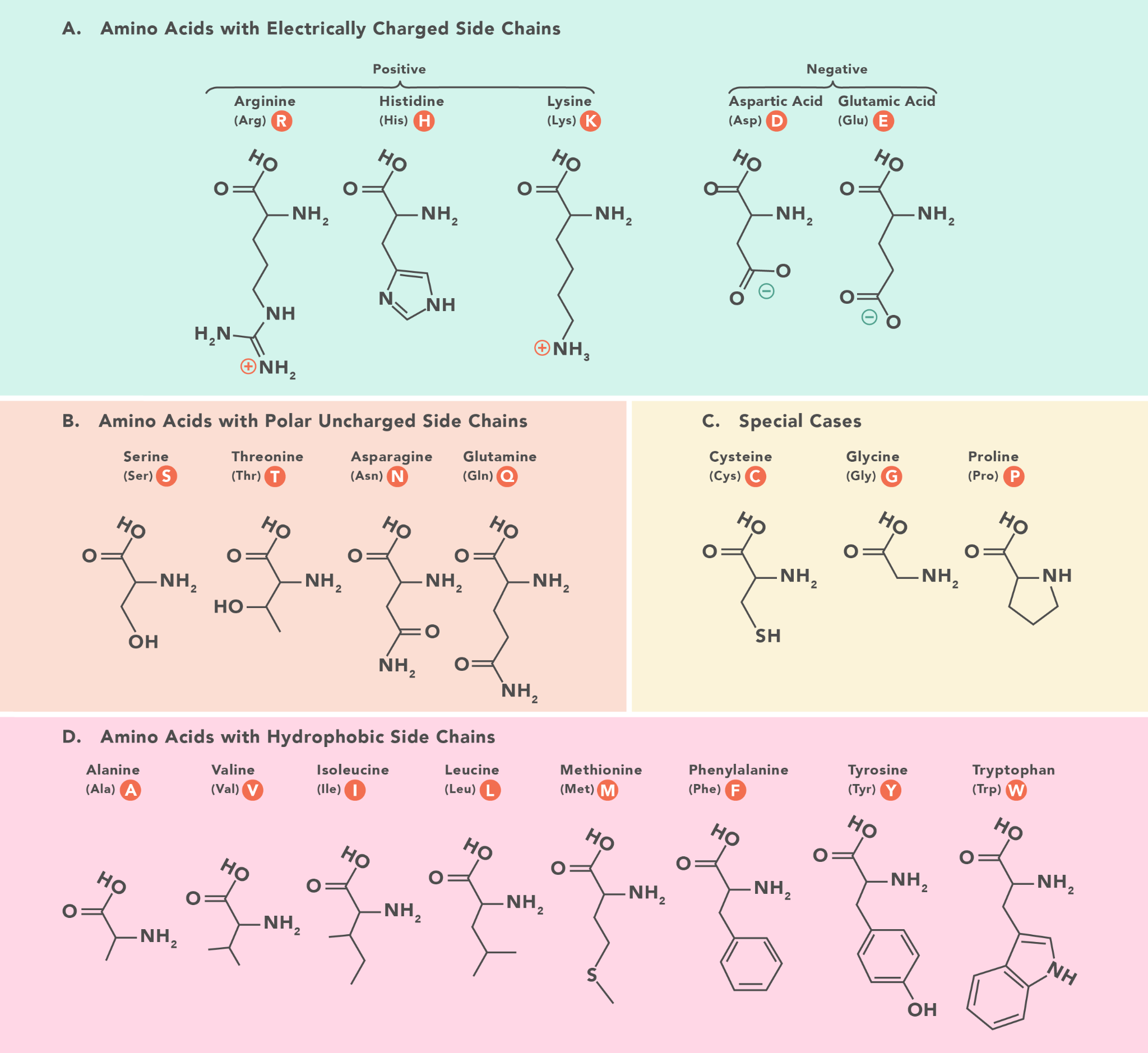
So on priority basis we can study :

**A to T - Transversion variation and other transition variation later.**

## Amino acid variation prioritization

The amino acid variations can be prioritized on the basis of their charge and properties, shape and size.

Using the below mentioned chart for studying the properties of amino acids.



[Source](https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357)

In case of amino acids, the variation is only observed at one position that is at 937 where amino acid Isoleucine is changing to Valine.

Whereas in others even though there is change in nucleotide the resultant amino acid is the same.

So the first position which should be studied must be 937 on the basis of nucleotide and amino acid change. Here purine is changing to another purine but still leading to change in amino acid.

Valine and Isoleucine share similar properties as they both have hydrophobic side chains. Valine and isoleucine both are neutral amino acids. [[2]](https://teaching.ncl.ac.uk/bms/wiki/index.php/Amino_acids)

Second we should study the variation at position 993 where A is changed by T. Here purine is changing to pyrimidine but the amino acid is valine.

All the remaining four positions can be studied in any order.