Analyze the genome of a virus to identify mutations and predict their potential impact

As part of this project, you will -

1. Find the genes contained in a viral genome

2. Try to find the function of those genes

3. Identify the closest relatives of the virus

4. Identify if the viral genome contains any mutations, and

5. Try to predict if those mutations are likely to affect the function of the virus

Duration : **10 Days**

In the first task, you must identify the conceptual knowledge, methods and tools, and resources required for completing the research project. Carefully read the project statement and description and try to answer the following questions:

What are the milestones and deliverables of the project?

What are the concepts that you need to know for different steps of the project?

What are the resources you need, in terms of expert help, books, online resources etc?

What are the tools and methods you may have to use?

# Timeline

| Days | Task Description |
| --- | --- |
| 1 | * Decide the Virus and find the sequence   + Website/Online Tools   + Background Research |
| 2-3 | * Identify the genes from the genome   + Online Resources/Softwares * Find out organization method for gene information |
|  | * Gene to Protein and Output   + Find software to predict protein from gene   + Optionally:     - Manually write a sequence     - Write a program to get an AA sequence |
| 4-6 | * Go from Protein to function   + Google search:     - AA sequence to function       * Finding similar AA sequence from other organisms     - Find websites which have databases of protein and their function.   + Find the protein from AA sequence? |
| 7 | * Find the Closest Relatives of the Virus   + Find similar/ relatives of the virus     - Google search     - Use similar AA seq. To get to similar virus |
| 8-9 | * Identify Mutations   + Study same species in large number   + Identify genes and compare them to find changes |
| 8-9 | * Effect of mutations and prediction of function   + Use the AA seq. Changes for protein function prediction. |

# Milestones and Deliverables:

| Milestone | Deliverable | Resources We Need |
| --- | --- | --- |
| A List of Genes(Genome Sequence) | RNA or DNA seq., Gene Sequence, Protein Sequence, Total Number of Genes, Repetitive genes, location of genes, length of genes and proteins | NCBI, pubmed,pubchem, Interpro |
| Functions of Genes | Function of Protein, About the protein: Structure, Function, Interaction Network, Physical Structure Diagram | Codon chart, protein analyzing softwares. |
| Closest Relatives of the Virus | Blast result of the protein sequence and multiple alignment file. |  |
| Mutation and effect on Function | Name of the relative virus, Compare mutated gene, Gene Seq., Introns and Exons, AA Seq., Protein, Function of Protein, Difference/Mutation of Gene, Predicted function |  |

# Online Tools:

1. Pymol
2. PDB
3. Pubchem/Pubmed
4. ORF finder
5. BLAST

# Executive Summary

The report studies the strain of marburg virus named Marburg Marburgvirus isolate PREDICT\_SLAB3960 Kakbat\_SL\_2017. The report comprises genomic analysis of the virus, identification of mutations and their effect on the function of proteins. The genome sequence for the virus is fetched from the national library of medicines in fasta format. To locate the genes a tool called an ORF finder was used. The ORF finder gave many genes as per its written program. To cross-check all the genes given by ORF finder are actual genes or actual protein coding genes we used BLAST. BLAST is a tool which gives us the closely matched aligned sequences as a result from which we can say that whatever sequence we have provided may exist. The total genes after cross-examining were seven. Those protein coding genes are RNA directed RNA polymerase L, nucleoprotein, glycoprotein, polymerase cofactor VP35, matrix protein VP40, Transcriptional activator VP30 and Membrane-associated protein VP24. From the results of the BLAST search we tried predicting the function of the protein using the Uniprot database. For identifying the mutations the closest relatives of the virus were chosen. Those were chosen on the basis of Expect value, Query coverage and percentage identity in the BLAST results. The E value was equal to 0, query coverage was 100% and the percentage identity varies between 100 - 97%. Those sequences then aligned using a multiple alignment tool called Muscle whose results were viewed using a software called Jalview. Jalview helped in finding the mutations at specific locations and changes in nucleotides or amino acids. Those mutations were then studied to give out the prediction of the effect of mutation on the function of protein. The mutations of RNA polymerase L and polymerase cofactor VP35. There were 203 nucleotide variations in pol L and 39 amino acids mutations. The structure of RNA pol L was unavailable thus we studied another protein VP35. There were 23 nucleotide variations with 4 amino acids change. All those four changes were studied and found out that those changes in amino acid occurred due to the change in first and second nucleotides of the codons.

# Introduction

The goal for this task was to identify the genes from the genome and then be able to get the amino acid chain from the gene to be able to identify the protein. Our overall approach for this task involves the following steps: obtain genome data, use ORFfinder to find genes and amino acid sequences, analyze the amino acid sequences to predict the encoded proteins using Interpro. The databases we used include the National Library of Medicine, and tools such as ORFfinder and Interpro.

## 

## **Virus chosen : Marburg marburgvirus (MARV)**

# Marburg Virus and it’s Genome:

## Marburg virus(MARV)

Marburg is a highly pathogenic virus which induces a disease called marburg virus disease(MVD). Marburg is the member of the Filoviridae family.

Marburg virus is known as one of the most dangerous viruses with a fatality rate of 88%. Its natural host is the Rousettus aegyptiacus fruit bat, after which it then transfers to humans. The incubation period lasts somewhere between 2 to 21 days. The Marburg virus causes sudden onset illness characterized by a high fever, severe headaches, and malaise. On the third day, severe watery diarrhea, cramps and pain in the abdomen, nausea, and vomiting are all possible.

Between five and seven days after the onset of symptoms, severe hemorrhagic signs occur, and fatal cases typically involve some degree of bleeding, frequently from numerous sites. Death in fatal instances typically occurs eight to nine days after the onset of symptoms, and is typically preceded by significant blood loss and shock.

Marburg spreads from person to person by direct human-to-human contact (via broken skin or mucous membranes) with the blood, secretions, organs, or other body fluids of infected persons and even those items that came into contact with bodily fluids, surfaces and items such as bedding or clothing.

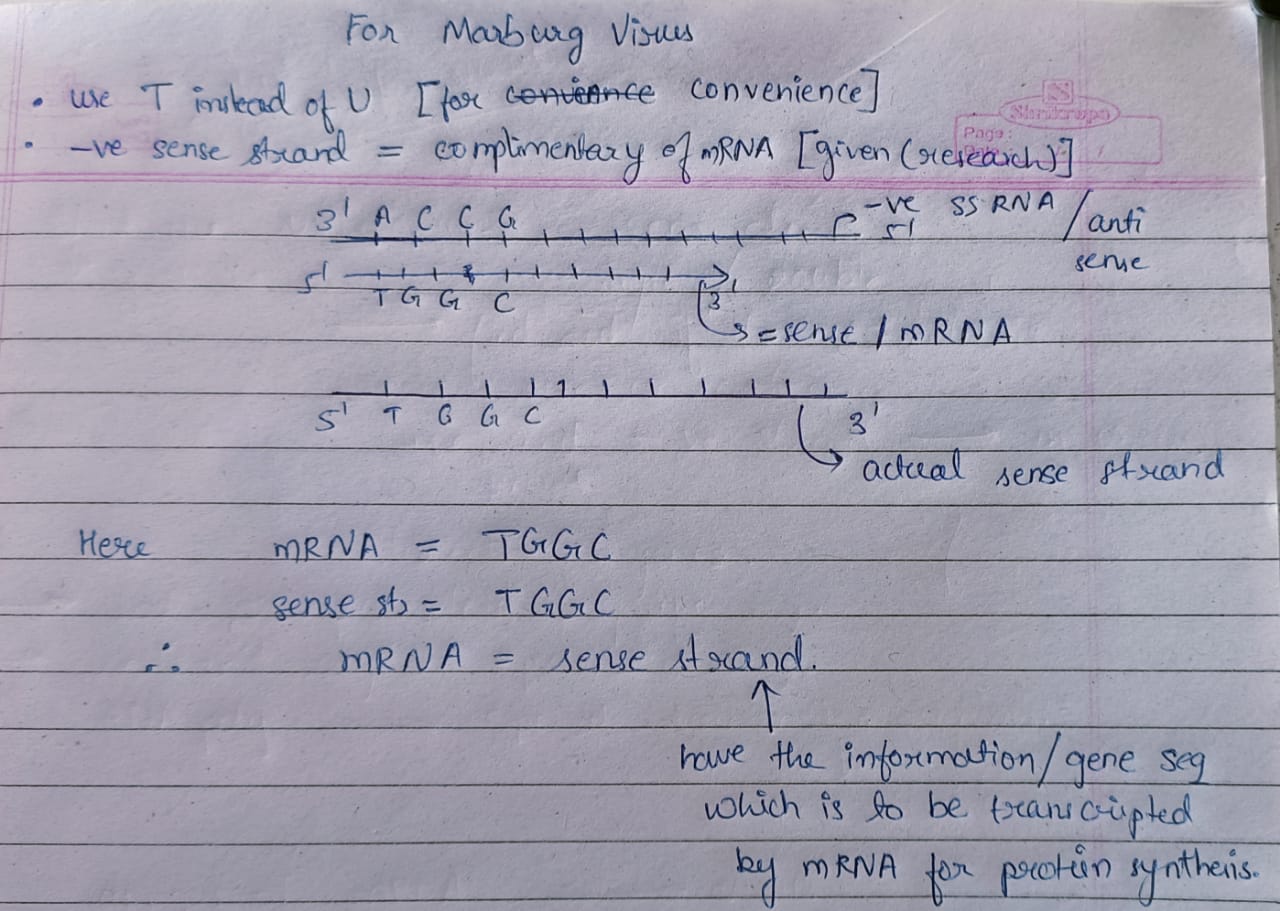
The Marburg Virus was first identified in 1967 after there were simultaneous outbreaks in Marburg and Frankfurt in Germany, as well as Belgrade, Siberia. The Marburg virus itself was first isolated in 1967 by the scientists of the University of Marburg, Germany in cooperation with specialists for viral electron microscopy at the Bernhard Nocht Institute in Hamburg, Germany.

## Marburg genome

Marburg Strain : [Marburg marburgvirus isolate PREDICT\_SLAB3960 Kakbat\_SL\_2017](https://www.ncbi.nlm.nih.gov/nuccore/1811087178)

Marburg is a single stranded RNA (ssRNA) with a negative sense strand.

**Negative sense RNA strand :** The negative sense strand is basically the antisense strand. This negative sense makes up the genome sequence of marburg virus. The marburg virus genes are complementary to mRNA sequence and exact same to the sense strand with the difference of Uracil -U instead of Thymine-T.



Since we have a negative or complementary strand, when the mRNA goes through the process of protein synthesis, it uses the complementary (or Antisense) strand to get the sequence of the sense strand and go through the process of protein synthesis. Thus we don’t need to change the DNA sequence we have to its complementary.

### Marburg genome description

A scientific article published on July 31st, 2019 by the One Health Institute at the University of California, Davis provides information on the Marburg marburgvirus. This virus has a genome length of 19,100 nucleotides and belongs to the Filoviridae family in the order Mononegavirales. Marburg marburgvirus is classified under the Monjiviricetes class and Negarnaviricota subphylum in the Riboviria kingdom. The genome source of the virus is Marburg marburgvirus. The research article provides valuable information on the genome structure and classification of the Marburg marburgvirus.

# 

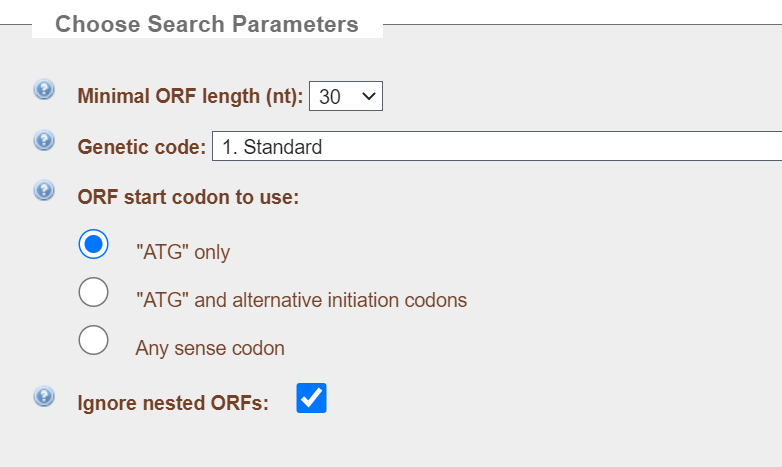
# ORF Finder

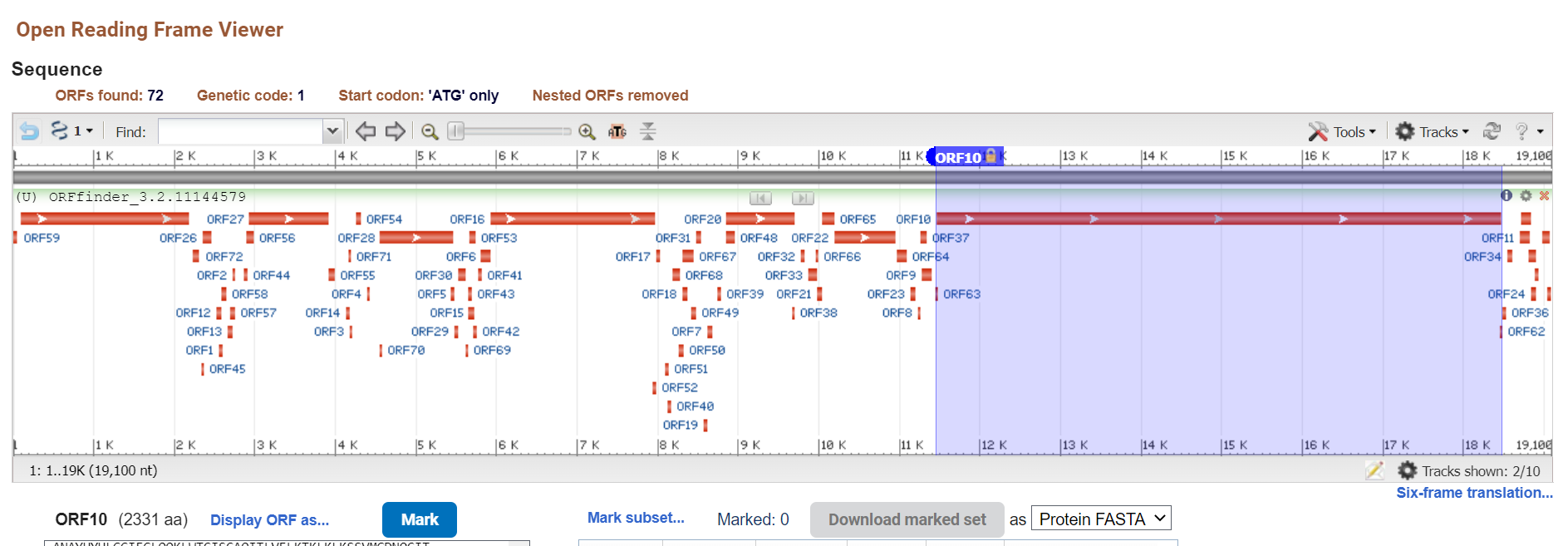
ORF finder is an online bioinformatics tool used to identify open reading frames (ORFs) in nucleotide sequences. An ORF is a region in a genome that has the potential to encode a protein. ORFfinder helps to identify potential genes and corresponding amino acid sequences in a genome. The tool provides 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 ORF finder uses the start and stop codon sequence and it lays out all the possible groups of start and stop codons. The start codons which it detects are ATG, GTG,CTG and stop codons are TAA, TAG, TGA [[1]](http://ubwp.buffalo.edu/wnygirp/wp-content/uploads/sites/5/2016/07/Module_5.pdf).

The input for ORF is the marburg genome sequence. We are using Standard for the genetic code. In ORF finder the only mention of a virus for the Genetic Code is for prokaryotic viruses and the Marburg virus is a Eukaryotic Virus. The Starting codon for marburg virus is AUG/ATG [[2]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668471/). The minimal ORF length we choose was 30 nucleotides, originally we decided 30 nt to see how many genes are there in the genome and 30 was the smallest threshold for the ORF finder to work. Later on in the results we find out 51 is the cut out and there was only 30 and 75 ORF length so we let it to be 30 nt and covered it while doing the protein prediction.

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





| Label | Strand | Frame | Start | Stop | Length(nt|aa) |
| --- | --- | --- | --- | --- | --- |
| ORF10 | + | 1 | 11467 | 18462 | 6996 | 2331 |
| ORF25 | + | 3 | 90 | 2177 | 2088 | 695 |
| ORF16 | + | 2 | 5927 | 7972 | 2046 | 681 |
| ORF27 | + | 3 | 2931 | 3920 | 990 | 329 |
| ORF28 | + | 3 | 4554 | 5465 | 912 | 303 |
| ORF20 | + | 2 | 8855 | 9700 | 846 | 281 |
| ORF22 | + | 2 | 10193 | 10954 | 762 | 253 |

## 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.

### Transcription start site and Termination sequence for marburg virus

The transcriptional start codon is (3'-NNCUNCNUNUAAUU-5') and the termination codon is (3'-UAAUUCUUUUU-5'). [[21]](https://pubmed.ncbi.nlm.nih.gov/1626422/)

## How will you know the gene is making the predicted protein?

To verify the predicted protein is making the actual protein, we can use tools where we can compare the pre-existing protein sequences in the protein databases. If we find a match with an acceptable threshold of proteins which is similar to the predicted protein then we can say that the protein has the high possibility to exist.

# 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.

## How does blast work?

The inserted sequence is compared with the pre-existing sequences and matched to give out the results with matching percentages.

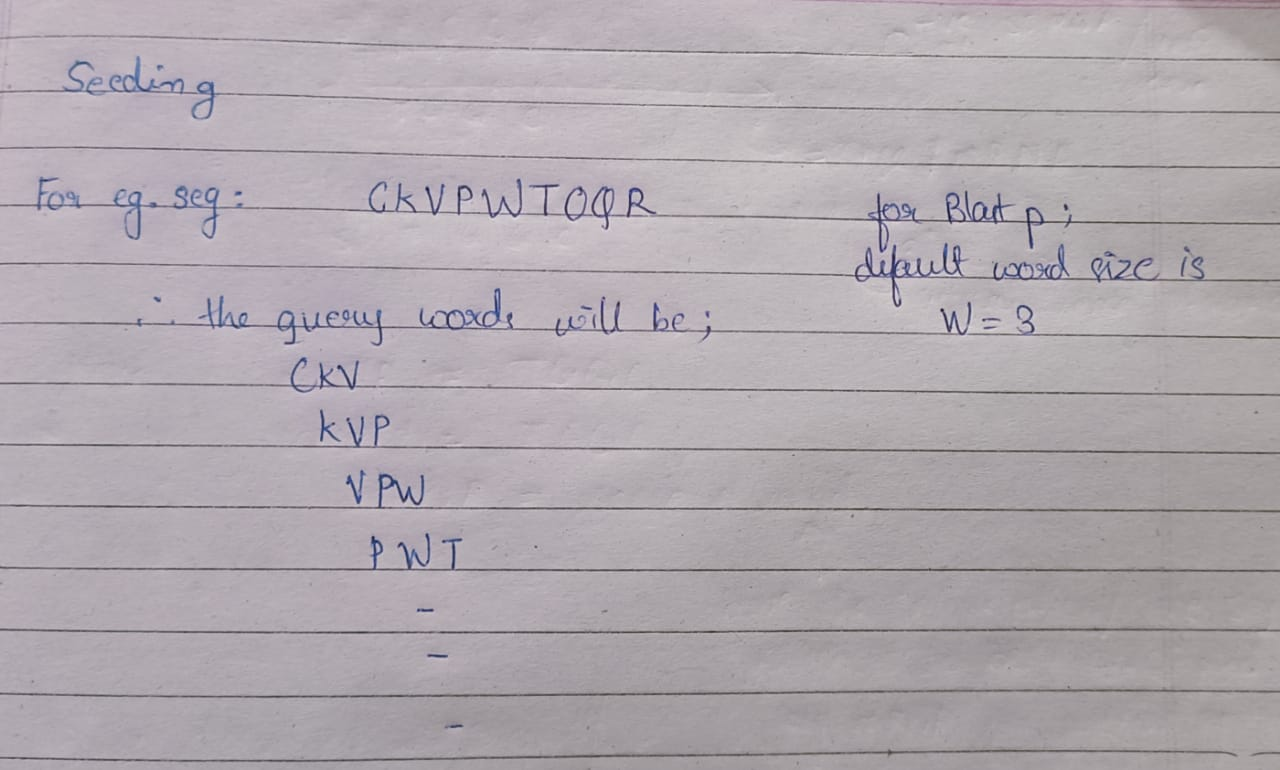
Blast uses homologous sequences using the heuristic method. In Bioinformatics, the heuristic method refers to the approximation of a particular algorithm. In Blast this is the approximation of the Smith–Waterman algorithm that this software uses.

The blast does not take the entire sequence into consideration, instead it breaks down the sequence into tiny fragments and then tries to match it with the database sequence. [[3]](https://www.youtube.com/watch?v=NwiETShpIDo)

When BLAST compares two sequences, it first looks for similar regions and starts matching them. However, it may not always find the best possible match, which could cause some similar sequences to be missed.

BLAST divides the sequence into tiny fragments which are called “**seed**.”

**Seeding** is the process of finding matches between the inserted/given sequence and the database sequence. The word or seed is matched by looking for sequences or letters which are common in both the sequences. The default word /seed size for BLASTp is 3 letters.



The query words which are described in the picture above are then matched . The matched words are remembered and highlighted so that the program does not have to go through the entire sequence again when it tries to match things. Only the highlighted portions will be used to build an alignment. This helps in elimination of an extra step of finding things in a massive sequence.

BLAST being a computer program it will give all the similar matching sequences which it finds. To decrease the error a threshold T is set up to cut out the more dissimilar options. BLASTp uses a scoring matrix called “**BLOSUM62**” where the threshold value is T=13. The matrix score assigned to the similar words below 13 is eliminated and above are allowed to proceed.

Then the matched words are then extended from both directions by the size of 1 word that is by three letters. Each time this happens, the matrix score increases or decreases. This again passes through the threshold and unacceptable words are eliminated. The score or alignment above the threshold is included in the final blast results. [[4]](https://www.ccg.unam.mx/~vinuesa/tlem/pdfs/Bioinformatics_explained_BLAST.pdf)

## Limitations of BLAST:

1. It uses a heuristic method which does not provide 100% match, there is a possibility that during alignment it may miss some sequences.
2. All the words which pass the scoring matrix are included in the final blast so there is a possibility that it will give error/wrong results.

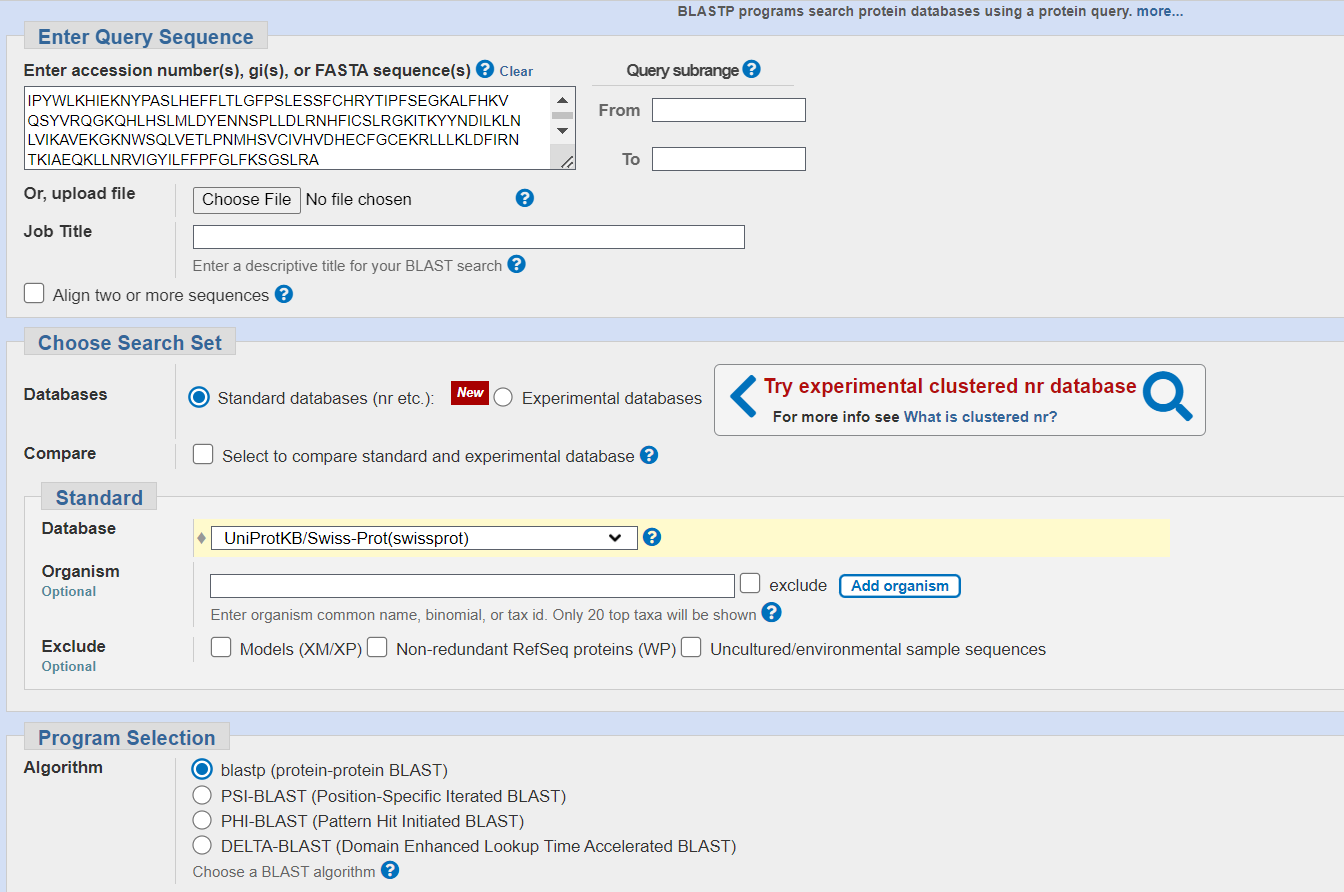
## Inputs given to Blastp:

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 10**. As we need the most aligned sequences to get the highly matched sequences.

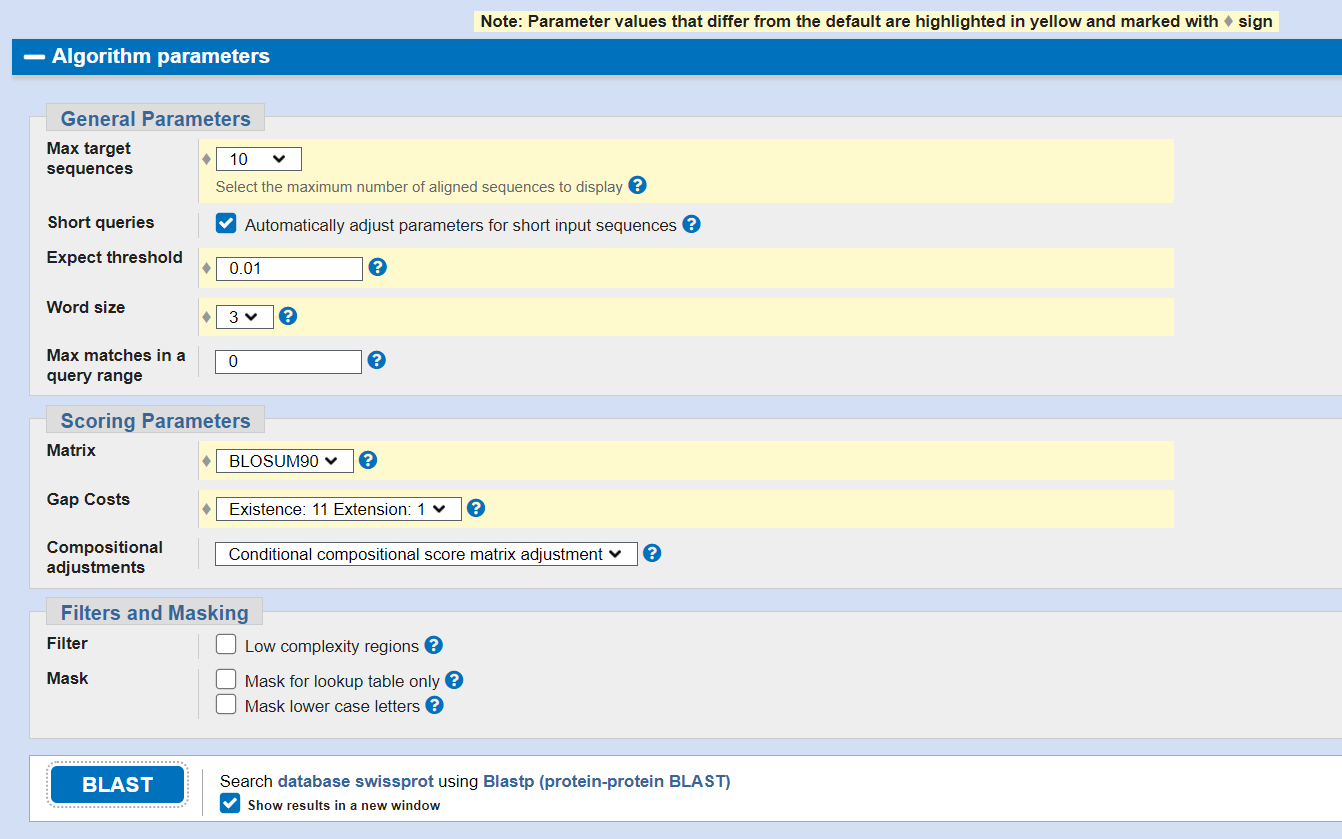
The **expected threshold or E value is set low to 0.01** to avoid the number of matches found by accident. The lower the E-value, or the closer it is to zero, the more "significant" the match is. [[5]](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ)

The **word size is set to 3.** Word size is the size of seed to initiate the alignment.The word size 6 is not chosen as it can decrease the speed of the process but it will be more specific. Whereas Seed size of 3 provides the balance between sensitivity and speed of the blastp.

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.**



## Outputs received from Blastp:

### Output results page:





### 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.[[6]](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.)

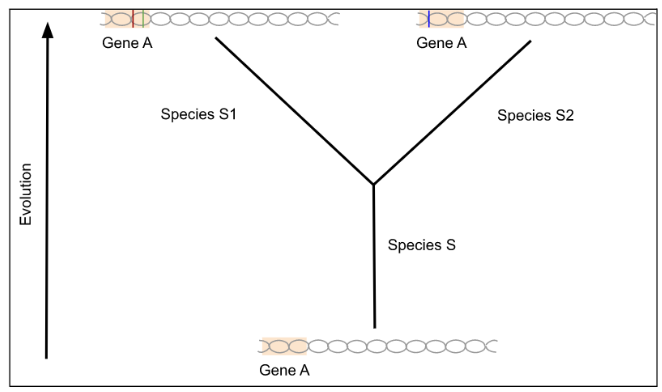
**Query cover** is the percentage of the query aligned with the database sequence. High query cover is the indication that the result sequence has highly aligned with the query sequence. **E value** is the value or score of how much the aligned sequence is by accident. Lesser the E value the better the 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). [[7]](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.

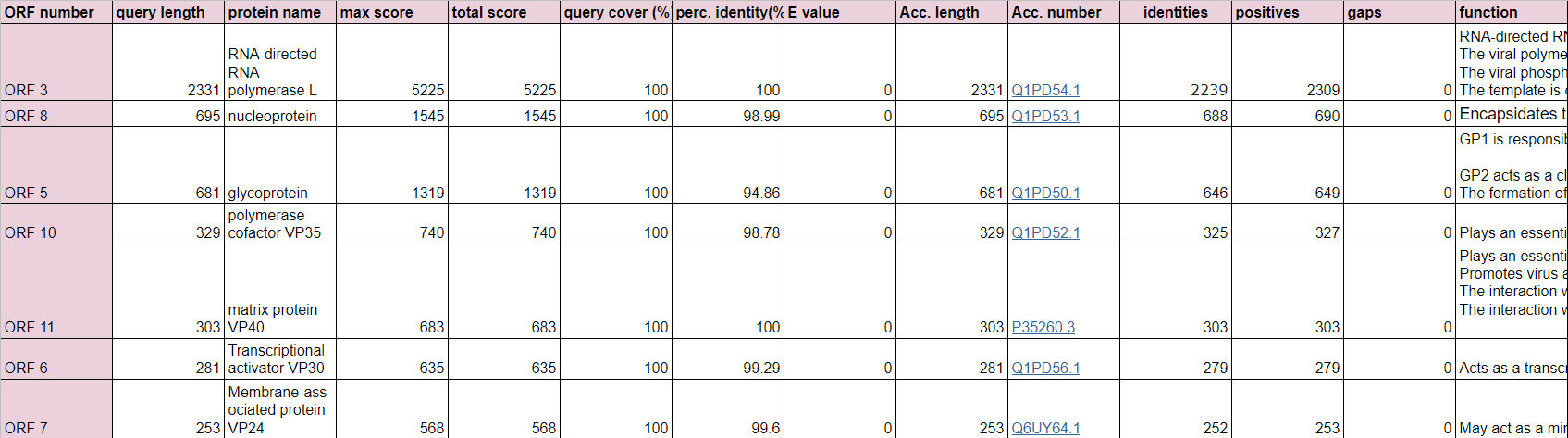
#### Why are we using the matched result from blast to describe the function of our protein?

**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

[**Note** : The amino acid sequences below 51 are not taken into consideration for protein prediction. The interproscan is unable to identify the small sequences. ]



[Excel link](https://docs.google.com/spreadsheets/d/1GZTAFv_0cc1dlm9hhTjShajdjrVu4NmPVrW2bZXB2qM/edit?usp=sharing)

# Proteins and their functions

## RNA-directed RNA polymerase L

**Interpretation** :

The polymerase L protein is correctly identified by the blast as the length of the query exactly matches with the accession number with 100% coverage and 100% percent identity. But the total identities are 2239 not 2331 then how come percent identity be 100%? The error value is zero thus are no chance or very very less chance that this alignment happens by chance.

**Function** :

RNA-directed RNA polymerase that catalyzes the transcription of viral mRNAs, their capping and polyadenylation. The template is composed of the viral RNA tightly encapsidated by the nucleoprotein (N).

The viral polymerase binds to the genomic RNA at the 3' leader promoter, and transcribes subsequently all viral mRNAs with a decreasing efficiency. The first gene is the most transcribed, and the last the least transcribed.

The viral phosphoprotein acts as a processivity factor. Capping is concomitant with initiation of mRNA transcription. Indeed, a GDP polyribonucleotidyl transferase (PRNTase) adds the cap structure when the nascent RNA chain length has reached a few nucleotides.

The template is composed of the viral RNA tightly encapsidated by the nucleoprotein (N). The replicase mode is dependent on intracellular N protein concentration. In this mode, the polymerase replicates the whole viral genome without recognizing transcriptional signals, and the replicated genome is not capped or polyadenylated. [[9]](https://www.uniprot.org/uniprotkb/Q1PD54/entry#structure)

## Nucleoprotein

**Interpretation**:

The nucleoprotein with query length of 695 and total and maximum score of 1545 with 100% query cover. Nucleoprotein is 98.99% identical to the resultant sequence. The total identical bases are 688 and positives that are with similar properties are 2 more than the identities which are 690 with no gaps in the sequence. The E value is zero. Hence we can confidently say that this is a real sequence is not by- by chance.

**Function**:

Encapsidates the genome, protecting it from nucleases. The encapsidated genomic RNA is termed the nucleocapsid and serves as template for transcription and replication. During replication, encapsidation by NP is coupled to RNA synthesis and all replicative products are resistant to nucleases. [[10]](https://www.uniprot.org/uniprotkb/Q1PD53/entry)

## Glycoprotein

**Interpretation** :

The glycoprotein with query length of 681and total and maximum score of 1319 with 100% query cover.

It has a 94.86 percent identity where there is no error value or gaps. The total number of identities are 646 and positives are 649.

**Function**:

GP1 is responsible for binding to the receptor(s) on target cells. Interacts with CD209/DC-SIGN and CLEC4M/DC-SIGNR which act as cofactors for virus entry into the host cell. Binding to CD209 and CLEC4M, which are respectively found on dendritic cells (DCs), and on endothelial cells of liver sinusoids and lymph node sinuses, facilitate infection of macrophages and endothelial cells. These interactions not only facilitate virus cell entry, but also allow capture of viral particles by DCs and subsequent transmission to susceptible cells without DCs infection.

GP2 acts as a class I viral fusion protein. Under the current model, the protein has at least 3 conformational states: pre-fusion native state, pre-hairpin intermediate state, and post-fusion hairpin state. During viral and target cell membrane fusion, the coiled coil regions (heptad repeats) assume a trimer-of-hairpins structure, positioning the fusion peptide in close proximity to the C-terminal region of the ectodomain.

The formation of this structure appears to drive apposition and subsequent fusion of viral and target cell membranes. Responsible for penetration of the virus into the cell cytoplasm by mediating the fusion of the membrane of the endocytosed virus particle with the endosomal membrane. Low pH in endosomes induces an irreversible conformational change in GP2, releasing the fusion hydrophobic peptide. [[11]](https://www.uniprot.org/uniprotkb/Q1PD50/entry)

## Polymerase cofactor VP35

**Interpretation** :

VP35 has the query length of 329 and the matched protein sequence has exactly the same length which increases the possibility of this protein to be a real protein. It is 98.78% identical and the expected value is zero. The identical bases are 325 and similar bases are 327.

**Function**:

Plays an essential role in viral RNA synthesis and also a role in suppressing innate immune signaling.[[12]](https://www.uniprot.org/uniprotkb/Q1PD52/entry)

## Matrix protein VP40

**Interpretation** :

VP40 with a query length of 303 and found results of length 303 bases. This protein is exactly matched to the protein existing in the database showing that it is a real protein. It has 100% identity and coverage with 0 error value. The total number of positives and identities are equal to the query length with no gaps.

**Function** :

Plays an essential role in virus particle assembly and budding.

Promotes virus assembly and budding by interacting with host proteins of the multivesicular body pathway.

The interaction with host E3 ubiquitin ligase SMURF2 facilitates virus budding.

The interaction with the nucleocapsid and the plasma membrane may also facilitate virus budding. Specific interactions with membrane-associated GP and VP24 during the budding process may also occur. May play a role in genome replication. [[13]](https://www.uniprot.org/uniprotkb/P35260/entry)

## Transcriptional activator VP30

**Interpretation** :

VP30 is 99.29% identical to the resultant sequence. Error value for VP30 is 0 with 0 gaps. The 279 bases are identical from 281 with the same number of positives.

**Function**:

Acts as a transcription anti-termination factor immediately after transcription initiation, but does not affect transcription elongation. This function has been found to be dependent on the formation of an RNA secondary structure at the transcription start site of the first gene.[[14]](https://www.uniprot.org/uniprotkb/Q1PD56/entry)

## Membrane-associated protein VP24

**Interpretation** :

The query length of VP24 is 253 which is matched to a 253 long sequence. It has a percentage identity of 99.6%. Total number of identities and positives are 252 and 253 respectively.

**Function** :

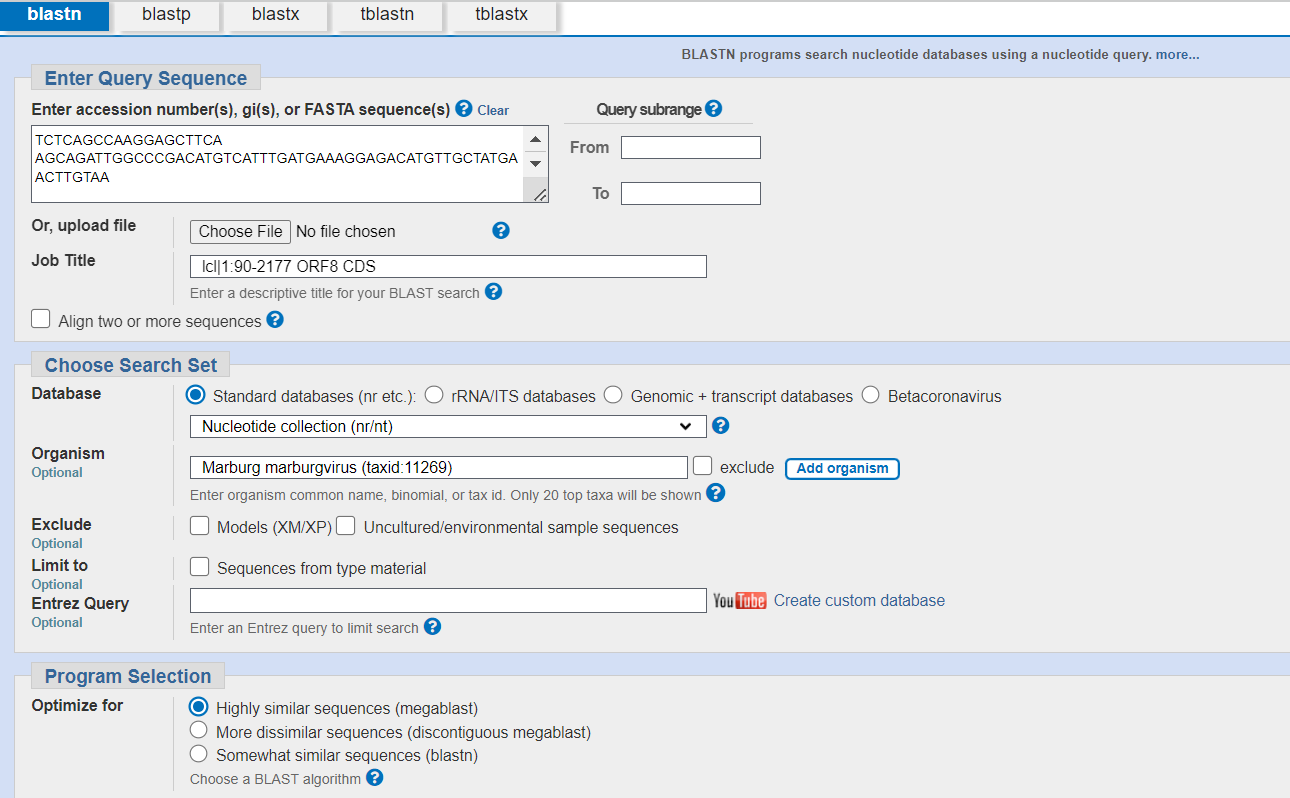
May act as a minor matrix protein that plays a role in assembly of viral nucleocapsid and virion budding. Unlike Ebola VP24, mVP24 has no measurable impact on host dendritic cell function. [[15]](https://www.uniprot.org/uniprotkb/Q6UY64/entry)

# Identifying Mutations

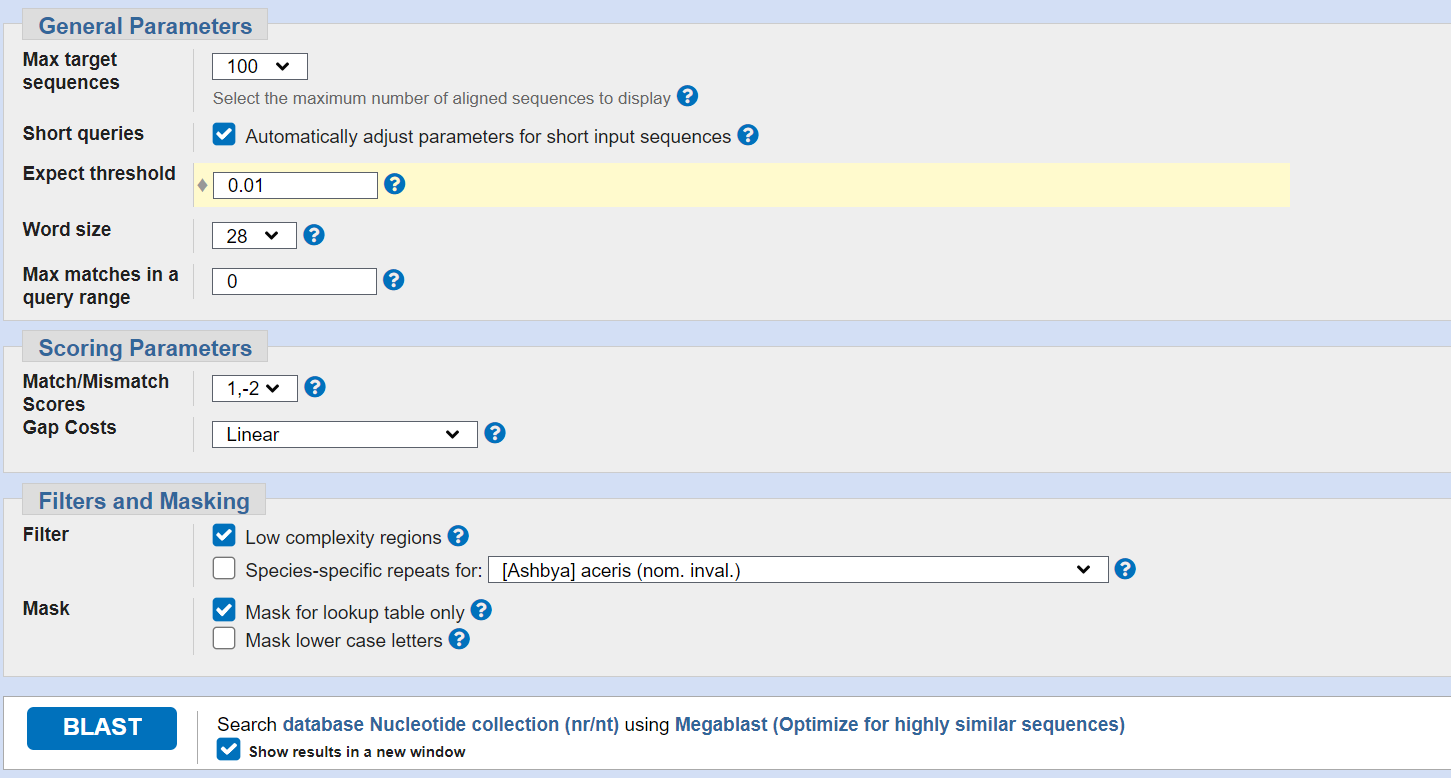
## Input

The blast used is “blastn”- nucleotide to nucleotide matching. The entered query is the gene nucleotide sequence of the orf found by the orf finder. The job title is the title of the orf sequence only. The database chosen is nucleotide collection(nr/nt) because it has a nucleotide collection consisting of GenBank+EMBL+DDBJ+PDB+RefSeq sequences.

As we are looking for the relative and very closest relatives of the virus we specifically look for the different species or strains of the same virus. Thus the organism chosen is “marburg marburgvirus.”

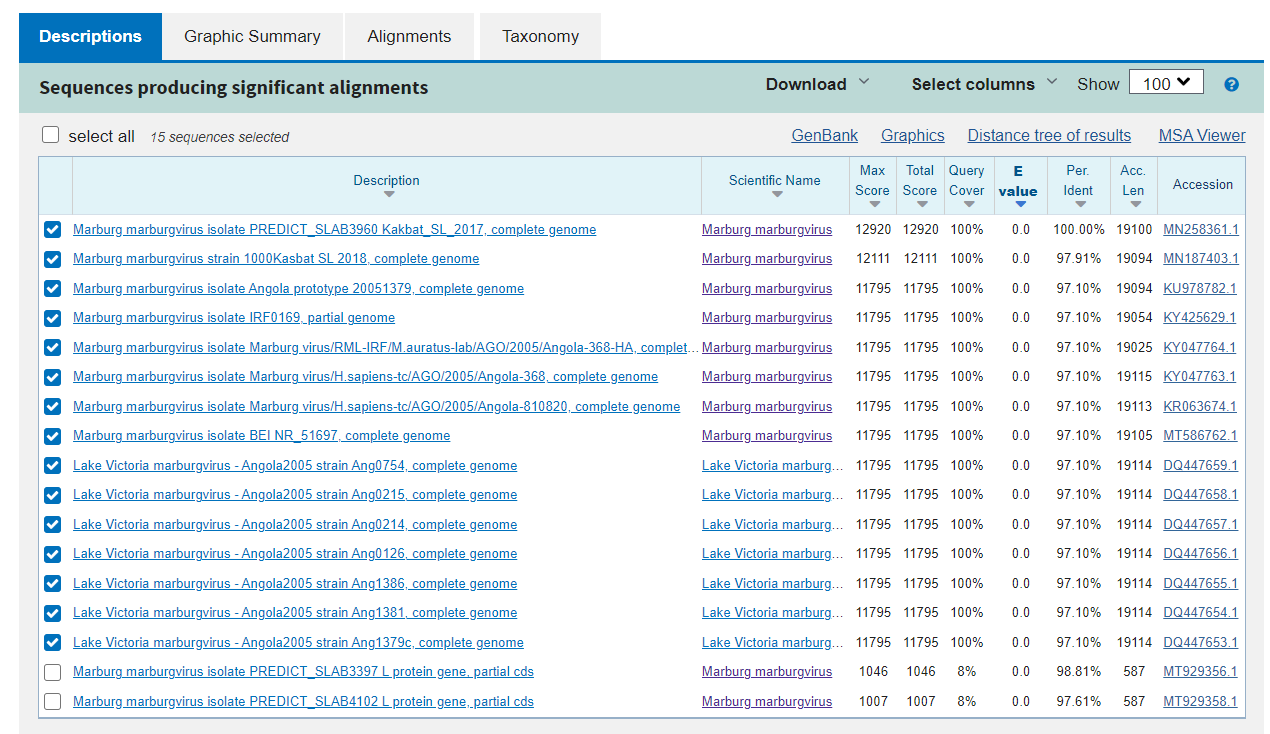


In general parameters, the maximum number of target sequences were set up to 100 to get as many as possible relatives we can find. The word is automatically set up to 28 to have highly specific results. The match score is 1 and the mismatched score is -2.



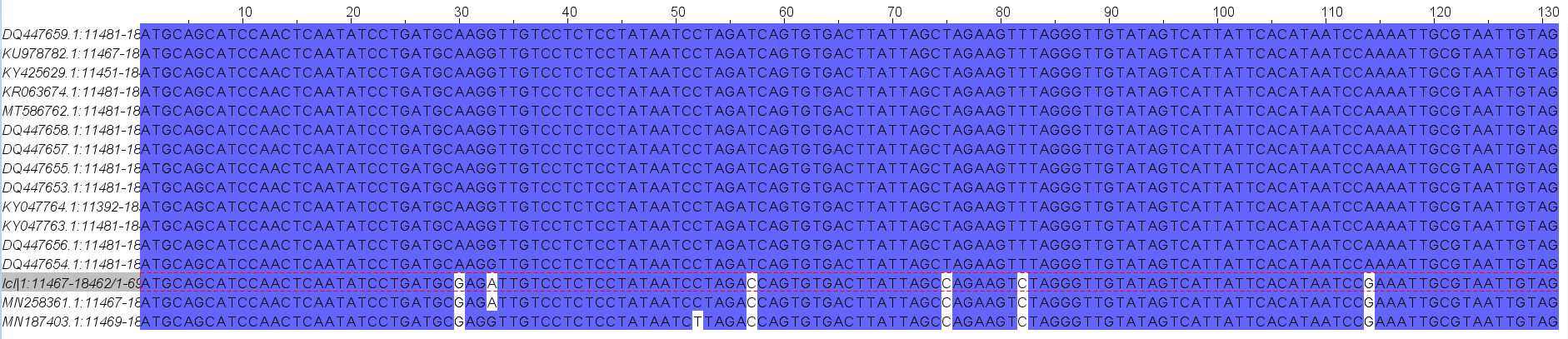
## Output

For Orf 3 ( polymerase L protein) all the sequences with 100% coverage and 100 - 97.10% identity are selected for multiple alignment.



Orf 3 :

Jalview result page :

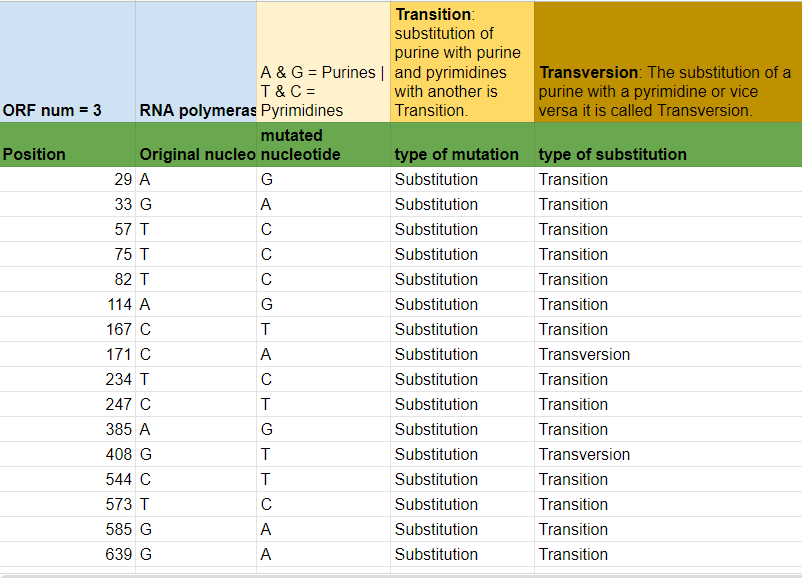


The gray sequence id is the inserted query sequence of our marburg virus.

## ORF 3- RNA-directed RNA polymerase L

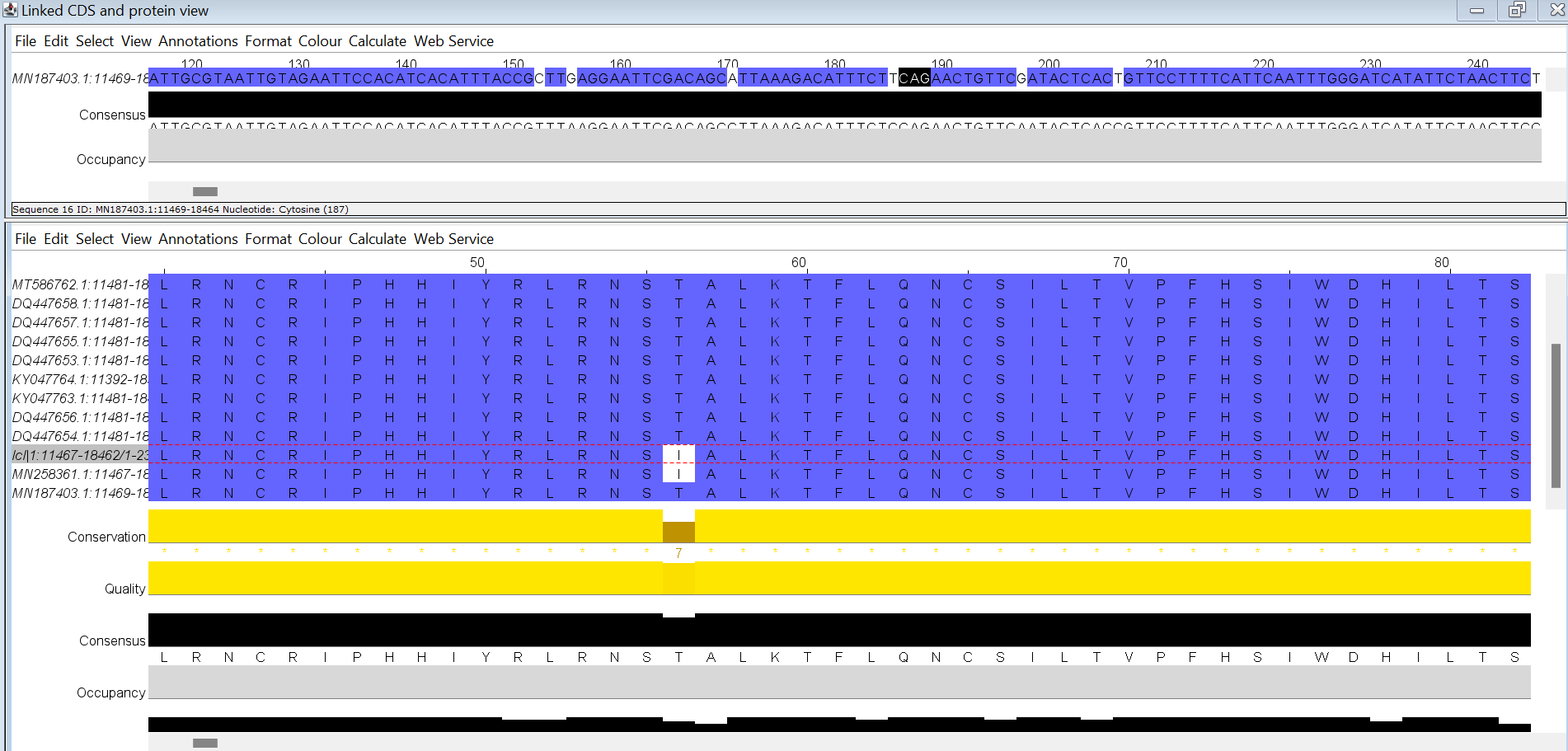
**Variations observed:**

There are a total number of 42 mutations, all of them are substitutions and of type transition. The mutations are only seen in bottom three sequences out of which the last third is the input query sequence. By parsimony’s principle we can say that all the above sequences are existing before those three or those bottom three are evolved from them.



[Excel link](https://docs.google.com/spreadsheets/d/1NTWB7NOieER3krnj7AmJL1A6E3WrgUTtoVN1i9P1gXM/edit?usp=sharing)

Identifying mutations in protein.



[Excel link](https://docs.google.com/spreadsheets/d/1GZTAFv_0cc1dlm9hhTjShajdjrVu4NmPVrW2bZXB2qM/edit?usp=sharing)

### 

# Predict effect of mutations on protein function

Some of the genes found in our virus structure for those proteins were not available in the database of uniprot. We have mapped the structure from blast result to the uniprot through accession number but there was no structure available for the specific strain of the marburg virus. To cross verify, we have directly inputted the name of the protein in uniprot as well as in PDB and tried searching for it but it was not there.

Then we looked for all the seven proteins that we found in our virus to see which all protein structures are available.

Below is the list of proteins whose structure we found in the database and which we do not:

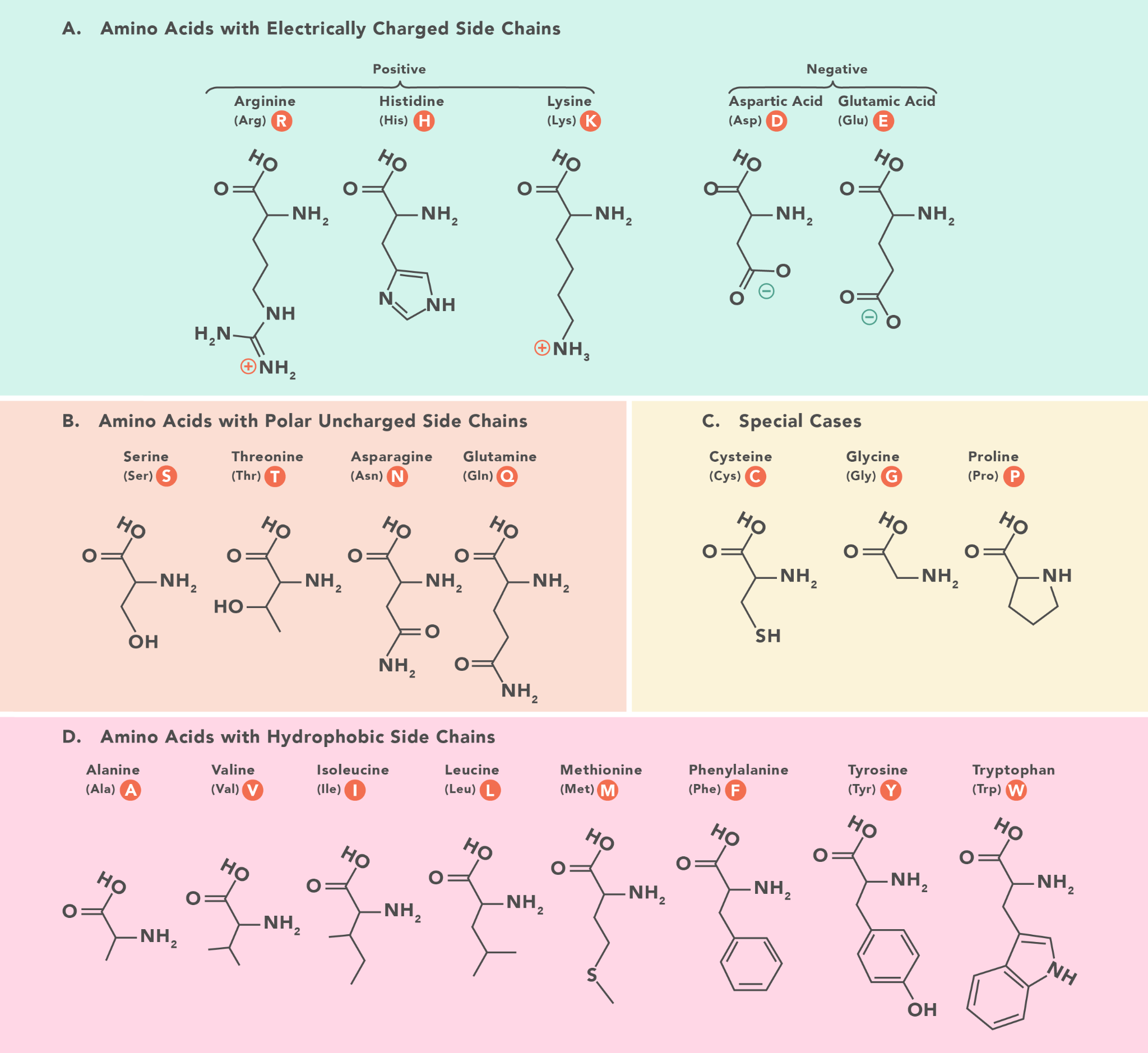
### Missing protein structures :

* RNA polymerase.
* Nucleoprotein
* Glycoprotein
* Transcriptional activator VP30

### Structure found :

* Polymerase cofactor VP35
* Matrix protein VP40
* Membrane-associated protein VP24

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



Above is the codon table to study the properties of different amino acids and their structures.

## Polymerase cofactor VP35

| **ORF num = 10** | **polymerase cofactor VP35** |  |  |  |
| --- | --- | --- | --- | --- |
| **position** | **original amino acid** | **changed amino acid** | **original codon** | **changed codon** |
| 67 | V | I | GTC | ATC |
| 69 | R | K | AGG | AAG |
| 80 | N | S | AAT | AGT |
| 88 | S | G | AGT | GGT |

### Function of Polymerase cofactor VP35

VP35 is a multifunctional protein which acts as a polymerase cofactor, a viral protein chaperone, and an antagonist of the innate immune response.

VP35 contains a central oligomerization domain with a predicted coiled-coil motif. This domain has been shown to be essential for RNA polymerase function. [[16]](https://pubmed.ncbi.nlm.nih.gov/27847355/)

Plays an essential role in viral RNA synthesis and also a role in suppressing innate immune signaling.[[17]](https://www.uniprot.org/uniprotkb/Q1PD52/entry)

### Changes in structure and function of a protein due to change in amino acid

#### **At position 67**

The amino acid V is not exactly at the 67th position; instead it is shifted by 5 amino acids.

The amino acid **Valine-V is mutated to Isoleucine-I.** The original codon was GTC whose first base got mutated to A and became ATC.

Valine is a neutral amino acid with a hydrophobic side chain similar to isoleucine. Valine is shorter in length compared to isoleucine as isoleucine has one extra carbon atom.

The oxygen atom of Valine is interacting with another oxygen atom and other amino acids in the beta sheet. If we mutate V to I, the size of the interaction length will increase as I is slightly longer. The structural angles may change to fit in the amino acid.

As both of them share similar properties hence there will be no major change in the functioning of the protein but they may not behave exactly identical.

All the amino acids with hydrophobicity tend to be found in the core of the protein as the protein molecules are suspended in aqueous environment in the body hence to hold and interact with the H2O molecules, the protein structures have hydrophilic amino acids on the outer surfaces.



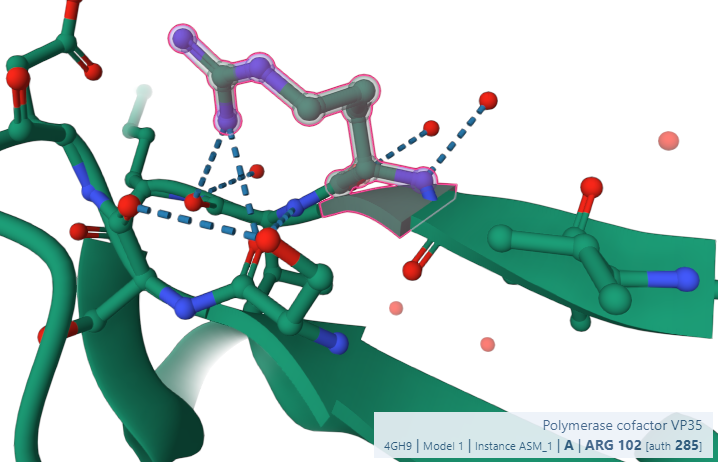
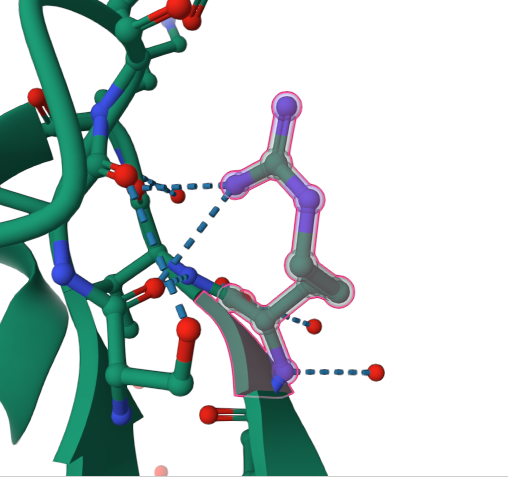
Structure of Valine

#### **At position 69**

The amino acid Arginine-R is replaced with Lysine-K. Both of them are electrically positive amino acids, the net charge on the amino acids is positive. Arginine has extra NH and double bond NH2. Thus arginine is longer than lysine. If R is substituted with K then the resultant structure may have more space or shrink in size and shape. In this the middle nucleotide changes from G to A changing the resultant amino acid. The N of NH2 of R is interacting with two other oxygen atoms.

In K, there is no NH2 hence there is a possibility that it may not interact with the surrounding oxygen molecule.

The arginine is coming out of the beta sheet in an upward direction and interacting with the O2 molecules which are in turn interacting with amino acids present on other beta sheets.



Structure of Arginine

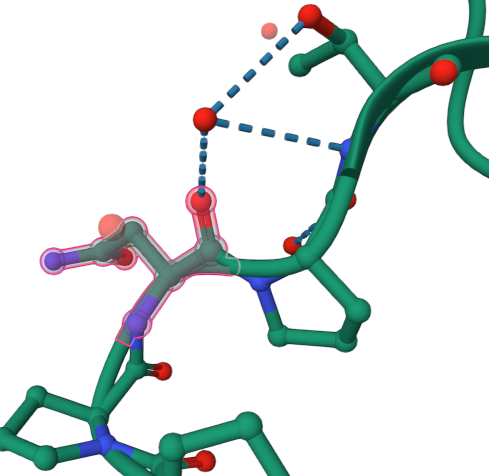
#### **At position 80**

Asparagine-N is changed to Serine-S. Here also the second nucleotide changed to lead to a mutation. A is changed to G both are purines hence transition has occurred there.

Serine is a smaller molecule whereas asparagine is longer in size. Serine has a C-OH bond which is replaced by O=C-NH2 .

Both of them belong to the class of polar uncharged molecules or chains which means the net charge on the molecule is zero or neutral but the distribution of positive and negative charge among the length creates slightly positive and negative charge.

If we replace N with S we may be able to see the changes in size and shape of the molecule. There are chances that the molecule shrinks as serine is smaller in size. The interactions may remain somewhat as earlier. But if the shape remains the same then the distance between the neighboring interacting atoms the distance increases thus decreasing in bond strength.



Structure of Asparagine

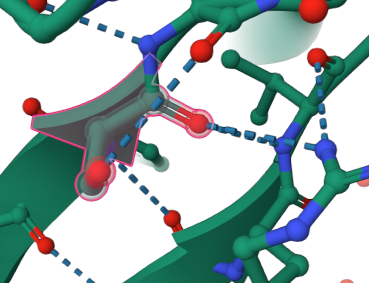
#### **At position 88**

The amino acid S-Serine changed to G-Glycine. The first nucleotide of the codon- A changed to G.

Serine is a neutral polar molecule whose charge is unequally distributed whereas in the case of glycine it is a neural non-polar molecule.

Glycine is a smaller molecule than serine. It does not have OH-C-CH3 bond.

The molecule's alterations in size and structure may be visible if S is substituted with G. As glycine becomes smaller, there is a possibility that the molecule will contract. The interactions might resemble earlier ones in certain ways. Yet, if the shape does not change, the distance between the interacting adjacent atoms grows, resulting in a reduction in bond strength.

****

Structure of Serine

# Conclusion

In conclusion, the report provides a detailed analysis of the Marburgvirus isolate PREDICT\_SLAB3960 Kakbat\_SL\_2017, including the identification of its genomic sequence, protein-coding genes, and mutations. The report utilized various bioinformatics tools such as ORF finder, BLAST, Muscle, and Jalview to analyze and predict the function of the proteins and their mutations. The study identified seven protein-coding genes, including RNA directed RNA polymerase L, nucleoprotein, glycoprotein, polymerase cofactor VP35, matrix protein VP40, Transcriptional activator VP30, and Membrane-associated protein VP24. Additionally, the report identified mutations in two proteins, RNA polymerase L and polymerase cofactor VP35, and predicted their effects on the function of these proteins. Overall, this study provides important insights into the genomics of Marburgvirus and can aid in the development of effective treatments and preventive measures against this deadly virus.

# References :

1. <https://www.ncbi.nlm.nih.gov/nuccore/1811087178>
2. <https://www.ncbi.nlm.nih.gov/nuccore/1811087178>
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668471/>
4. <http://ubwp.buffalo.edu/wnygirp/wp-content/uploads/sites/5/2016/07/Module_5.pdf>
5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668471/>
6. <https://www.ccg.unam.mx/~vinuesa/tlem/pdfs/Bioinformatics_explained_BLAST.pdf>
7. <https://www.youtube.com/watch?v=NwiETShpIDo>
8. <https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ>
9. <https://guides.lib.berkeley.edu/ncbi/blast#:~:text=Acc%5Bession%5D%20Len%5Bgth,records%20in%20the%20NCBI%20databases>
10. https://xcelerator.ninja/myQuests/186985/myLearning/477738/resources/37378
11. <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>.
12. <https://www.uniprot.org/uniprotkb/Q1PD54/entry#structure>
13. <https://www.uniprot.org/uniprotkb/Q1PD50/entry>
14. <https://www.uniprot.org/uniprotkb/Q1PD53/entry>
15. <https://www.uniprot.org/uniprotkb/P35260/entry>
16. <https://www.uniprot.org/uniprotkb/Q1PD52/entry>
17. <https://www.uniprot.org/uniprotkb/Q6UY64/entry>
18. <https://www.uniprot.org/uniprotkb/Q1PD56/entry>
19. <https://www.technologynetworks.com/applied-sciences/articles/essential-amino-acids-chart-abbreviations-and-structure-324357>
20. <https://www.uniprot.org/uniprotkb/Q1PD52/entry>
21. <https://pubmed.ncbi.nlm.nih.gov/27847355/>

# Appendix

1. Finding protein function from amino acid sequence:   
   **Sequence Similarity Can Provide Clues About Protein Function**Because amino acid sequence determines protein structure and structure dictates biochemical function, proteins that share a similar amino acid sequence usually perform similar biochemical functions, even when they are found in distantly related organisms.  
     
   Source:  
   [Analyzing Protein Structure and Function - Molecular Biology of the Cell - NCBI Bookshelf](https://www.ncbi.nlm.nih.gov/books/NBK26820/#:~:text=Sequence%20Similarity%20Can%20Provide%20Clues%20About%20Protein%20Function&text=Because%20amino%20acid%20sequence%20determines,found%20in%20distantly%20related%20organisms).
2. **Protein finder website :** <https://www.ebi.ac.uk/interpro/result/InterProScan/#table>
3. ~~Online tools for protein analysis :~~[~~https://guides.lib.byu.edu/c.php?g=216337&p=1428367~~](https://guides.lib.byu.edu/c.php?g=216337&p=1428367)[~~Panzer 2~~](http://ekhidna2.biocenter.helsinki.fi/sanspanz/)[~~PredictProtein~~](https://predictprotein.org/)
4. Marburg marburgvirus:  
   <https://www.ncbi.nlm.nih.gov/nuccore/1811087178>
5. Genome sequence :

AAGAGATCTTGTTTTTGTGTATCATATAAATAAAGAAGAATATTAACATTGACATTGAGACTTGTCAGTC

TGTTAATATTCTTGAAGAGATGGATTTACATAGTTTGTTAGAGTTGGGTACAAAACCTACTGCTCCTCAT

GTTCGTAATAAGAAGGTGATATTATTCGACACAAATCATCAGGTTAGTATCTGTAATCAGATAATAGATG

CAATAAACTCAGGGATTGATCTTGGAGATCTTCTAGAAGGGGGTTTGCTGACGTTGTGTGTTGAACACTA

TTATAATTCCGACAAAGATAAATTCAACACAAGTCCCATCGCAAAATACTTGCGTGATGCGGGTTATGAA

TTTGATGTCATCAAGAATGCAGATGCAACCCGCTTTCTGGATGTGATTCCCAATGAACCTCATTACAGCC

CTTTAATTCTGGCCCTTAAGACATTAGAAAGTACTGAATCTCAGAGGGGGAGGATTGGGCTCTTTTTGTC

ATTCTGCAGTCTTTTTCTCCCAAAACTTGTCGTCGGAGACCGGGCCAGTATCGAAAAGGCCTTAAGACAA

GTGACAGTACATCAAGAACAGGGAATTGTCACATACCCTAATCATTGGCTTACTACAGGTCACATGAAAG

TAATTTTTGGGATTTTAAGGTCTAGCTTTATTTTAAAGTTTGTATTGATTCATCAAGGAGTAAATTTGGT

GACAGGCCATGATGCATATGACAGTATCATTAGCAATTCAGTAGGTCAAACTAGATTCTCAGGACTTCTT

ATTGTGAAAACAGTTCTCGAGTTCATCTTGCAAAAAACTGATTCAGGGGTGACACTACATCCCTTGGTGC

GGACCTCCAAAGTGAAAAACGAAGTTGCTAGTTTCAAGCAGGCGTTGAGCAACTTAGCCCGACACGGAGA

GTACGCACCATTTGCACGGGTTTTGAATTTATCAGGGATTAACAATCTTGAACATGGACTCTATCCTCAG

CTCTCAGCAATTGCGCTGGGTGTAGCAACAGCACATGGCAGTACATTGGCTGGTGTCAATGTCGGCGAAC

AATATCAACAGCTACGAGAGGCAGCACATGATGCAGAAGTAAAACTACAAAGACGACATGAACATCAGGA

AATTCAAGCTATTGCCGAGGACGACGAGGAAAGGAAGATATTGGAACAATTCCACCTTCAGAAAACCGAA

ATCACACACAGTCAGACACTTGCCGTCCTCAGCCAGAAACGAGAAAAATTAGCCCGTCTTGCTGCAGAAA

TTGAAAACAATATTGTGGAAGATCAGGGATTTAAACAATCACAGAATCGGGTATCACAGTCGTTTTTGAA

TGACCCCACGCCTGTGGAAGTAACGGTTCAAGCCAGGCCCATAAATCGACCAACTGCTCTGCCTCCTCCA

GTCGACGACAAAATTGAGCGTGAATCTACAGAAGATAGCTCTTCTTCAAGTAGCTTTGTTGACTTGAATG

ATCCATTTGCACTGCTGAACGAGGACGAAGACACTCTTGACGACAGTGTCATGATCCCGAGCACAACATC

GAGAGAATTTCAAGGGATTCCTGAACCGCCAAGGCAATCTCAGGACCTCGATAACAGCCAAGAAAAGCAG

GAAGATGAATCAACAAACCTGATTAAGAAACCGTTTTTGCGATATCAAGAATTGCCTCCTGTCCAAGAGG

ATGATGAATCGGAATACACAACTGACTCTCAGGAAAGTATTGACCAGCCAGGATCCGACAATGAACAAGG

GGTTGATCTTCCACCTCCTCCGTTGTACGCTCAGGAAAAGAGACAGGATCCAACACAGCACCCAGCAGTA

AGCTCTCAGGATCCCTTTGGCAGTATTGGTGATGTAAATGGTGATATCTTAGAACCCATAAGGTCACCCT

CTTCACCGTCTGCTCCTCAGGGAGACACAAGGGCAAGAGAGGCCTATGAATTGTCGCCTGACTTCACAAA

TTATGAGGATAATCAGCAGAATTGGCCACAAAGAGTGGTGACAAAGAAGGGTAGGACTTTCCTTTATCCT

AATGATCTTCTGCAGACAAATCCTCCAGAATCACTTATAACAGCCCTCATTGAGGAATACCAAAATCCTG

TCTCAGCCAAGGAGCTTCAAGCAGATTGGCCCGACATGTCATTTGATGAAAGGAGACATGTTGCTATGAA

CTTGTAATCCAGATAACACAGCACGTTTACCCACTCATCCACTTCAATACGATTCACCCTCAGATCACAG

CAATCAACCTTATTTGAATATCCAAACCACTATTTAGTATCCTATCACTTGTTACTATTGTGTGAGACAA

CATAAGCTATCAAATAATAATCACGGGCAAGAATTGGGCATACGATGGTGATTTGCGAGCATTGTTCAGT

GCTACAAATTCTTTTTTCAATTGCTATAATGATACAACCACAAACCTCCGTACACTTGCCGCAATACTGT

AATTAACACTGCTGTATCTCTTCTTCAAGCCATCTGATTTAACTCAATAAATATGACTTGATTCAGAGAA

TATGCTGAAAATGTTATTGATTGAACTTCTCAAATGGTGCACTATCCTACTGTTTTGCTCAGCTTAGTAT

ATTGTAACATATAAGTGGACTCTCCACTTCTCTTCTTGTGTATTCTCTACAAGTCATTTACTTGATAGAA

TGTCAAGTCTACTGGTTCAGAGTCTCCTTACTCTAATGAATGTAATAATTAACTGTTAGCTTAGATGATA

ACAGATATGAGGTTATATAATTACTCATAGTGTAAAGTGTAATTCTTACCTCTGTTTCTTCTGTTTTCCC

TCTCCTTTATAATATGCCAATTAAGAAAAACTAAAAATCGAAGAATATTAAAGGTTTTCTTTAATATTCA

GAAAAGGTTTTTTATTCTCTTCTTTCCTTTTACAAACATATTGAAATAATAATTCTCACAATGTGGGACT

CATCATATATGCAACAAGTCAGCGAAGGGTTGATGACTGGAAAAGTACCCATAGATCAAGTGTTTGGTAC

CAATCCCTTAGAGAAGCTATACAAGAGAAGAAAACCAAAAGGAACAGTTGGGCTACAATGTAGCCCTTGT

CTAATGTCAAAGGCAACAAGTACTGATGATATTATTTGGGACCAACTGATCGTGAAGAAAACACTAGCTG

ATCTACTTATACCGATAAGTAGGCAGATATCGGATATTCAGGGTACTCTAAGTGAAGTAACAACAAGAGT

CCATGAAATTGAGCGGCAATTACATGAGATCACCCCAGTTTTGAAAATGGGAAGGACACTGGAAGCAATC

TCCAAGGGGATGTCAGAAATGTTAGCTAAATATGACCACCTTGTAATCTCAACTGGAAGAACCACTGCAC

CAGCTGCTGCCTTTGATGCTTACTTAAATGAGCATGGTGTCCCTCCTCCTCAACCCGCGATCTTCAAGGA

TCTCGGGGTTGCTCAACAAGCTTGTAGTAAGGGGACTACGGTCAAAAATGCAACAACAGATGCAGCCGAC

AAGATGTCGAAGGTCCTTGAACTCAGTGAGGAAACGTTCTCTAAGCCAAACCTCTCAGCCAAGGATTTAG

CCCTTTTATTGTTTACCCATCTACCCGGCAACAACACTCCATTCCACATCCTGGCTCAGGTCCTTTCAAA

AATTGCTTACAAGTCAGGAAAATCCGGGGCATTTTTGGATGCATTTCACCAGATTCTGAGTGAAGGAGAG

AATGCTCAGGCGGCATTAACTCGACTAAGCAGAACATTTGATGCTTTCCTCGGAGTGGTTCCTCCAGTGA

TAAGAGTCAAAAACTTCCAAACAGTCCCTCGCCCATGTCAAAAAAGTCTTCGGGCTGTCCCTCCAAATCC

AACAATTGACAAAGGATGGGTCTGTGTTTATTCATCTGAGCAAGGTGAGACACGGGCTCTTAAAATCTAA

TTCTCATTGTTCAAAGTTGCAAGGGAAGTTATATTTCCGAGTTGATACAAAGACACTAAACATTTCAGAA

GCATGTATGTGGGCAAAACATAACTAGACCATCTTAATCGAAGTAGTAATTTATTTCTGTCTTAAATGTG

ATTTTTACTTCAAAAGAGTTAAATGGTGATAGATTAATCTTTGAAATTACCCATTTTTACATTATAGAGA

AACTAATGTTACTAACAAAAGGGGTCTACCTAACAGGTATGATTGAATGATCAGTATATCTCATAGACCA

AACAATTGACTTCTTACTTTTTAAGAATCAACTAACAGCATAGGAAACACATTTATCTTTACGTAATCCT

TGGCTTAGTTGGAATTAACTTTTGTTGTAATTCAATACGCTTAGTCATAATATATTGTATGATTTTTTAT

AAGTTTAGAATATCTTAAATTATACCCACAAAAAATACTGTTTTAATTAAGAAAAACTATGAAGAACATT

AAGAAGATCTTTCTCTCATAGTGTTCTTTTACCAGAAGAAGTATTCTAATTTTAGCTTGTTGGATCAGTT

GTCACTTAAGTTGTTCTTTTTGAAATTAATTCACACAAAGTAGTTTAAATTTGTATCTAGAACAAATTTT

AATATGGCCAGTTCCAGCAATTACAATACATACATGCAATACCTGAACCCCCCTCCTTATGCTGACCACG

GTGCAAACCAGTTAATCCCGGCGGATCAGCTATCAAATCAGCAGGGCATAACTCCAAATTATGTGGGTGA

TTTAAACCTAGATGATCAGTTCAAAGGGAATGTCTGCCATGCTTTCACTTTAGAAGCAATAATTGACATA

TCTGCGTATAACGAGCGAACAGTCAAAGGCGTCCCGGCATGGCTGCCTCTTGGGATTATGAGCAATTTCG

AATATCCTTTAGCCCATACAGTGGCTGCATTGCTCACAGGCAGCTATACAATCACCCAATTTACTCATAA

TGGGCAAAAATTCGTCCGTGTCAATCGACTCGGTACAGGAATCCCGGCACACCCACTCAGGATGTTGCGT

GAAGGAAATCAAGCTTTCATTCAGAATATGGTGATCCCCAGGAATTTTTCCACTAATCAATTCACATACA

ATCTCACTAACTTAGTATTGAGTGTGCAAAAACTTCCTGATGATGCCTGGCGTCCATCCAAGGACAAATT

AATTGGAAACACCATGCATCCCGCAGTCTCCATTCACCCGAATTTGCCACCTATTGTTCTACCAACAGTC

AAGAAGCAGGCTTATCGTCAGCACAAAAATCCCAACAATGGTCCACTGCTGGCCATATCTGGCATCCTTC

ATCAACTGAGGGTCGAAAAAGTCCCAGAAAAAACAAGCCTGTTTAGGATTTCGCTTCCTGCTGACATGTT

CTCAGTAAAAGAGGGTATGATGAAGAAAAGAGGAGAAAATTCCCCGGTGGTTTATTTTCAAGCACCTGAG

AACTTCCCTTTGAATGGCTTCAATAATAGACAGGTTGTACTAGCGTATGCGAATCCAACACTCAGCGCCG

TTTAAAATGATGCTCAAATGAAACAGAAGTCCATCAATATAAGGGGTATAGCTTAAATGGATATTTGTCA

GATTCTTAAGAGATTAGTTTTTATTGATTTCAATAATGCTTTAACCTTACATTGCTGCTTTAAATGGTTG

ATTAAGCTGATTAGCTTGTAATATGTAATCTCTTTTGGGCCATCAGATCTGTAATGGGTTTACTAAATTA

TATACAAGAAACAGTAACATTTTATAAACAATTCTTGCTCAGTTTTACTTTGATTTAATAACATACATCA

TTGTGCCCTTCATTGCTAGGTCAACTCAACTGATAATGATATTCCTTCTGAAATAGTAAGAAAAACTAAT

GAAGAACATTAATTGCTGGGTAAAAGTGATTAAGTTCTTTAAATTTGATCAAAGTAATATTTTGTCAGTG

AATACATTCCCATATCACTTGATTAAAAACAGGAAATTATCCTGACATGAAGACCATATGGCTCCTTATC

AGTCTTATCTTAATCCAAGGGGTAAAAACTCTCCCTATTTTAGAGATAGCCAGTAACAATCAACCCCAAA

ATGTGGATTCAGTATGCTCCGGGACCCTCCAGAAGACAGAAGACGTTCATCTGATGGGATTCACACTGAG

CGGGCAAAAGGTTGCCGATTCCCCTTTAGAGGCATCCAAACGATGGGCCTTCAGGACAGGTGTACCTCCT

AAGAATGTTGAGTATACAGAAGGGGAAGAAGCTAAAACATGTTACAATATAAGTGTAACGGATCCCTCTG

GAAAATCCTTGCTGTTAGATCCTCCCACCAACATTCGTGACTACCCTAAATGCAAAACTATCCATCATAT

TCAAGGTCAAAACCCTCATGCACAGGGAATCGCTCTCCATTTGTGGGGAGCATTTTTCTTGTATGATCGC

ATCGCCTCCACAACAATGTACCGAGGCAAAGTCTTCACTGAAGGGAACATAGCAGCTATGATTGTCAATA

AGACAGTGCACAAAATGATTTTCTCGAGGCAAGGACAAGGATACCGTCACATGAATCTAACTTCTACTAA

CAAATATTGGACAAGTAGCAACGGAACGCAAACGAATGACACTGGATGCTTCGGCACTCTTCAAGAATAC

AATTCTACAAAGAACCAAACATGTGCTCCGTCCAAAAAACCTCCACCGCTGTCCACAGCCCACCCAGAGGTCAAGCCCACCAGCACCTCAACTGACGCCACCAAACTCAATACCACAGACCCAAACAGTGATGACGAAGACCTCACAACATCCGGCTCAGGGTCTGGAGAGCAAGAGCCCCACACAACTTCTGACGCGACCACCAAGCAAGGGCCCCCATCAACAATGCCGCCCACCCCCTCACCACAACCAAGCACGCCACAGCAAGGAGGAAACAACACAAGCCATCACCAAGGTGCTGCGACCGAACCCGACAAAACCAACATAACTGCGCAACCGTCCATGCCCCCCCACAACACCACCACAATCCCTCCTAACAACACCTCCAAGCACAACCCCAGCACTCCCCCTGCACCAACACAAAACGCCACCAATTACAACACACAGAGCACGGCCACCGAAAACGAGCAAACCAGTGCCCTCTCGAAAA

CAACCTTGCCTCCAACAGAAAATCCCACAACGGCAAAGAGCACCAACAGCACAAAAAGCCCCACCACAAC

AGCACCAAACACGACAAATGAGTATTCCACCAGTCCCTCCCCCGCCCCCAACCTGACTGCACCACACCTT

GTATATTTCAGAAGGAAACGAAGTATCCTCTGGAGGGAAGGCGACATGTTCCCTTTTCTGGATGGGTTAA

TAAATGCCCCGATTGACTTTGATCCGGTTCCAGATACAAAGACAATCTTTGATGAATCCTCTAGTTCTGG

TGCTTCAGCTGAGGAAGATCAACATGCCTCCCCTAATATCAGTTTAACTTTATCTTACTTCCCTAAGGTA

AATGAAAACACTGCCTACTCTGGAGAAAATGAAAATGATTGTGATGCAGAGTTAAGAATTTGGAGTGTTC

AGGAGGATGACCTGGCAGCAGGGCTCAGTTGGATACCGTTTTTTGGCCCTGGGATCGAAGGACTTTATAC

TGCTGGCTTAATTAAAAATCAAAATAATTTGGTTTGCAGGTTGAGGCGTCTAGCCAATCAAACTGCCAAA

TCCTTGGAACTCCTATTGAGAGTCACAACCGAGGAAAGAACATTTTCCTTAATCAATAGACATGCCATTG

ATTTTTTACTCACAAGGTGGGGAGGAACATGCAAAGTGCTTGGACCTGATTGTTGCATTGGAATAGAAGA

CTTGTCCAGAAATATTTCAGAACAAATTGATCAAATCAAAAAGGACGAACAAAAAGAGGGGACTGGTTGG

GGTCTGGGTGGTAAATGGTGGACATCAGACTGGGGTGTTCTTACTAACTTGGGCATCTTGCTACTACTGT

CCATAGCTGTCTTAATTGCTCTATCCTGTATCTGTCGTATTTTTACTAAATATATTGGATAACATTAAAT

GTGCAATGATGAGAACTTTAAGATAATTGCTACTGAGCCCTTCTTCTGATCTACTGAAATCAATTTGGGA

GATTTTTGAGAAGCTGATAACTTAATGTGAATCAACAATTTACGTATTATCGATTATTATGGTTTGAGAT

TCAACTGTTATCATCATCAAGAGTGACCTGTCCTATTTGACGCATTAATGTTTTAAACTACCTCTTAAGC

CTTTATGGATGGTCTCAATATGCGCGTAGGGGTTAATTTAAAGAGATTTCTTGTTGCATAGTTTTTTGTA

TTACTCATTTGGGCTTGAAGACATAGTGAAGACTGGCCGAAAATGCTCTCCAGTCAACTCCATTCCCTCT

CAGAGGAGACGTGCTGTCCAAAGAGTCTTAATTTATAACCAACCATTGCAAGAATTAATTTACTTTTTCC

GTTATACTTATCTACATTAATTCTTTGAATATCCAGCATCATTAACGACTTGCCTTATTTCAATCTTTTG

GATGCAAATCATAAGGAAAGATAAGCCACTTTCCCTTTACTCTGAGCTAAGGAAATTTCTCTTATCAGCT

TAAAATATGATTCATCAGGTCACGGGCCCTTCATGATCTATTTGAGCGTGAATATTGATCAAAAGACCAA

ATGATGTTGCATTCGTATAGACTCAATTATCCTTTATTAAGAAAAAGACAAATAGAACACAAATAGTTGA

CAAAACTTTACTTTGATAAATTTCGCGAGGAATTATAAAAATCTTGAAGGATAAATTATTGTAGAGTAGA

GTCGAAGAACATTAAGTGTTCTTTGTTAGAATTATTCATCCTAGTTATTTTTGAGTATATTCGCTTCAAT

ACAACTTCTCTTTATATTTGATTTAAGATTTAAAATGCAACAACCTCGGGGAAGAAGTCGAACCCGTAAC

CACCAAGCCACACCGTCTATATATCATGAAACTCAATTGCCCTCTAAACCTCATTACGCAAGTCATCATC

CACGTGCAAGATCTATGAGCTCAACCCGTAGTAGTGCAGAAAGCAGTCCCACCAATCACATTCCTCGTGC

TCGACCACCCTCAACATTCAACTTATCGAAACCCCCTCCTCCCCCAAAGGACATGTGTAGGAACATGAAA

ATTGGACTGCCGTGCACTGATCCTACTTGTAATAGGGATCATGACCTTGATAACCTAACAAATCGTGAAC

TTTTGCTATTGATGGCCCGAAAAATGCTCCCCAATACAGATAAGACATTTAGAAGTCCGCAGGACTGTGG

GTCACCGTCTCTTTCTAAAGGGCTCTCAAAAGATAAACAGGAGCAAACGAAGGATGTGTTGACCTTGGAA

AATCTAGGGCACATCCTGAGCTACCTCCACAGATCAGAAATTGGGAAATTAGATGAGACATCACTTCGTG

CAGCATTAAGTTTGACGTGTGCTGGAATTCGAAAGACGAATAGATCTTTGATCAACACCATGACAGAATT

ACACATTAACCACGAAAATCTCCCTCAAGACCAAAATGGTGTTATCAAGCAGACATACACAGGTATTCAC

CTTGATAAAGGAGGTCAATTCGAAGCCGCCTTATGGCAAGGTTGGGATAAGAGATCGATATCTCTATTTG

TACAAGCAGCTTTATATGTGATGAACAATATCCCCTGTGAATCATCGATCAGTGTGCAAGCCTCATACGA

TCATTTTATTCTTCCTCAAAGTCAAGGCAAAGGACAATGATCATTGTTCAAAGATTGACAACCTGATCAC

TTTCAGTTTTCAATTTCAACTCTTATCGCGAGATTTGAACATGACCCTACTAACTTCAATAATTGACCCC

AAATTCAAGTTCACTGAAGGCTAAGATGACAATAATCACTAGTTTATTGTAAATTACTCTATCAAAATGT

TCTTAAGCTATCTTAAACTTAATGATGCAACTTTATTTCACCTTTCTTTTAATGTCAATGTCACAGCTGT

ATCCAAGTGTCTAATTAACAACTTGTACCTCTAAGGAAAACCATGAAGAACATTAAGAAAAAGGATGTTC

TGATTCTTCAATTGAACTTGTATATCTCTTACTGATACCTTTGAGAGACAACTATTGACGCTAGATCACG

GATAAAGCATATTTCAGTCAAACATCCTAAATTTTCAATCATACACATAATAACCATTTTGATAGCATTA

CTGTTCAGTATAGGATAGGTGATCGTGGAAAGACTTCTAAGCATGGCAGAATTATCAACGCGTTACAACT

TGCCTGCAAATGTTACAGAAAAGAGCATAAATCTTGACCTCAATTCCACAGCACGATGGATTAAAGAACC

CAGTGTTGGGGGCTGGACAGTGAAGTGGGGAAACTTTGTTTTCCACATACCAAATACTGGAATGACATTG

TTGCATCATTTAAAGTCTAATTTTGTTGTTCCAGAGTGGCAACAAACAAGGAATCTATTCTCCCATCTCT

TTAAAAATCCGAAATCAACAATTATAGAACCGTTCTTGGCTCTGAGGATCTTGCTTGGAGTTGCTTTGAA

GGACCAAGAATTACAGCAATCATTAATTCCTGGATTTAGATCTATTGTTCATATGCTTTCAGAATGGTTG

CTCCTGGAGGTCACGTCGGCAATCCACATTAGTCCCAATCTGTTAGGGATCTATTTGACTTCAGATATGT

TTAAGATTCTGATGGCAGGTGTGAAAAATTTCTTTAATAAGATGTTCACTCTTCATGTTGTAAATGACCA

CGGGAAACCCAGCAGTATTGAAATAAAGTTAACTGGGCAACAGATCATTATCACTCGTGTTAATATGGGG

TTTCTAGTGGAAGTCAGGAGGATTGACATAGAACCTTGTTGTGGTGAGACAGTCCTTTCAGAATCAGTTG

TTTTTGGGCTAGTGGCTGAGGCAGTTCTAAGAGAACATAGTCAAATGGAGAAGGGCCAACCTCTCAATCT

GACACAATACATGAACAGCAAAATTGCTATATAATTGGCTTAAACTAGCATGAGTATTTCCAGTTCAACT

ACATAATAATGCTGGAGACACAGTACATTATAGTTGATTATCCTGTATAATAAAGGATATACCTCCCCTG

ATCTATATTTACTGGTATAAAATAGCAGCATCATCTTATCAAATAGTTGTCATACAATAGGCTGTTCCTA

TAATCTGATGGTGAAAGTATATGCTTGTAGAATTACTGTGAGTCACAACTGTTGTATATTGTCCTAAATA

TATCTTGTGCAAGTGATGTGTGCTTAAATATGCCGATGCAATACGCACTAATAACGATTGATTAAGAAAA

ATCAATGATGGATATTAAATGTCCATCAAGCAGGTATTGTAGAATATTAGGGGTTTCAAATGCTGCCATA

CTTACTAAACTTTACATAGGATTATATAATTCTCTTCGATACACGTTATATCTTTAGCAAAGTAATGAAA

ACAGCTTTATCATGTTAGACGTCAATTATTCATCTTAAGTGAATCCTTTCTTCAATATGCAGCATCCAAC

TCAATATCCTGATGCGAGATTGTCCTCTCCTATAATCCTAGACCAGTGTGACTTATTAGCCAGAAGTCTA

GGGTTGTATAGTCATTATTCACATAATCCGAAATTGCGTAATTGTAGAATTCCACATCACATTTACCGTT

TAAGGAATTCGATAGCATTAAAGACATTTCTCCAGAACTGTTCAATACTCACCGTTCCTTTTCATTCAAT

TTGGGATCACATTCTAACTTCTATTCAATATGATGCAATTAATCATGTTGATGATTTTAAATACCTATTG

CCATCTGAGCTAGTCAAGTATGCAAATTGGGACAACGAGTTCTTAAAGGCATATCTTAATAAGATCTTAG

GACTTAACCATGTTTTTCCAGCTTCTGCAAGGTCACAATGTGATGATTTTTCTCCTAAGGAAAATCCTTA

TTATTGGGGGATGTTGTTACTCGTACATTTGTCCCAACTTGCCAGGAGGATAAAAGGACAAAGAGGGTCA

TTAAGAAGTAACTGGAAGTTTATAGGAACAGATTTGGAGTTGTTTGGAATAGCAGATTTTGTTATTTTCA

AGGTTCCAGTAAAAACAATAATCCGAAATGCTGTAAGCTTACAAGCTTCAAAACCAGGGTTAAGAGTATG

GTACCGTGACCAAAACTTGACCCCTTATCTATGCGACGATGAGTTTATTGTAAGCGTCGCTAGTTATGAA

TGTTTTATTATGATTAAAGATGTCTTCATTGAGAGGTACAACACATGGGAGATATGTGCCCGCGCTTGGC

TCGAAGACAGTGATGGAGCTGATTACCCTCCTCTTGATGTGTTAGGTGAGTTATACAACCAAGGAGACCA

AATTATTGCTATGTACCTGGAGGACGGTTTCAAATTGATTAAACACTTAGAACCTTTATGTGTTAGCTGT

ATACAAACACATGGTATTTTTACACCAAGAAAATACTGGTTCCAATCACAGATGATTAAGTCATATTATG

ATGAACTCTGTTGTCTCAACTTGAAGCTTCAAATTTCAGACAATAAATCCGAGTGTGCCCAAAACTTCAT

TAAAACTATAATTCAGGCGAAATTGACCCCTCAACAATACTGTGAATTATTCTCTCTACAAAAGCATTGG

GGTCATCCCGTCCTATACAATGATGTTGCACTGGATAAGGTCAAAAAACATGCACAATCTACAAAAATCT

TAAAACCTAAAGTCATGTTTGAAACTTTTTGTGTTTTCAAGTTCATAGTAGCAAAGAATCATTATCATTC

TCAAGGATCATGGTATAAGACCACACATGATTTGCATTTAACACCATATCTTAGACAACATATTATGTCA

AATTCATTTCCATCACAAGCTGAAATTTATCAGCATCTTTGGGAGTGGTATTTCGTGGAGCATGAACCTC

TTTTTTCAACTAAAATAATAAGTGATTTAAGTATTTTCATAAAAGACAGGGCCACTGCTGTGAACCGGGA

GTGTTGGGACAGTGTCTTCGATAGAAGTGTGTTAGGATATAACCCTCCTGTTAGATTTCAGTCAAAGAGA

GTGCCGGAGCAATTTTTGGGTCAAGCAGACTTTTCCTTGAATCAAATATTGGATTTTGCTGAAAAATTAG

AGTATTTGGCTCCTTCTTATAGGAATTTTTCTTTCTCATTAAAAGAAAAAGAGTTGAATATAGGAAGGAC

TTTTGGGAAATTACCGTATCGTGTCAGAAATGTCCAAACACTCGCAGAAGCCTTGTTAGCAGATGGACTA

GCAAAAGCATTCCCTAGTAATATGATGGTTGTCACTGAGAGGGAACAGAAAGAAGCATTGTTGCATCAGG

CCTCTTGGCACCATAATTCAGCAAGCATAGGGGAAAACGCTATAGTGAGAGGTGCGAGTTTTGTTACTGA

TCTTGAGAAATACAATCTTGCCTTCCGATATGAATTTACACGGCATTTTATAGACTACTGTAATCGATGT

TATGGTGTGAAGAATTTATTCGATTGGATGCACTTTTTAATACCACTATGTTACATGCATGTCAGTGATT

TTTATAGCCCACCACATTGTGTGACAGAAGATAACCGAAATAACCCACCTGATTGTGCCAATGCTTATCA

TTATCACTTAGGAGGTATAGAGGGACTTCAACAGAAATTGTGGACATGTATATCATGTGCCCAAATCACC

CTTGTGGAGTTAAAGACTAAATTAAAATTAAAGTCTAGTGTCATGGGTGATAATCAATGTATAACAACTC

TAAGTCTGTTTCCAGTTGATGCTCCCAATGATTATCAAGAGAACGAAGCTGAATTGAATGCAGCGCGAGT

TGCTGTTGAATTAGCTATTACTACGGGTTATAGTGGTATATTTTTGAAGCCTGAAGAAACATTTGTCCAT

TCAGGGTTCATTTATTTTGGTAAAAAGCAATATCTCAACGGTGTTCAACTGCCACAATCATTGAAAACAA

TGGCAAGATGTGGACCCTTGTCTGACTCTATTTTTGATGATCTTCAAGGCTCTCTAGCCAGTATTGGTAC

ATCATTTGAGAGAGGAACAAGTGAGACACGGCACATTTTTCCAAGTCGTTGGATAGCTTCATTTCATTCA

ATGTTAGCAATAAATTTATTAAATCAGAATCACCTTGGATTTCCCCTAGGGTTCAATATTGATATTTCTT

GTTTCAAAAAGCCTCTTACCTTCTCGGAAAAATTAATTGCTCTCATAACACCCCAAGTTTTAGGAGGGTT

ATCATTTTTGAATCCAGAGAAATTGTTCTACCGGAACATAAGTGATCCTCTCACTTCAGGTCTATTTCAA

CTCAAGAATGCATTGGAATTTCTTGAAAAGGAAGAATTATTCTATATCTTAATTGCTAAAAAACCTGGTT

TAGCAGATGCCTCAGATTTCGTCATGAATCCTTTAGGCTTAAATGTACCAGGATCAAGGGAAATAATAAC

GTTCCTTAGACAGACAGTTCGTGAAAATATCACGATCACGTCGCAAAATAGAATAATAAATTCTCTTTTT

CACATAGGCTCTGATTTAGAGGACCAAAGGGTGTGTGAATGGCTTTTATCATCAAACCCCGTAATGAGCC

GATTTGCGGCTGACATCTTTTCAAGAACACCCAGTGGAAAACGGCTTCAAGTCTTAGGCTATCTGGAGGG

AACAAGAACATTGCTAGCCTCTCGGACAATCAGTTTAACTACAGAAGGGACAATGTTGATGAAATTAAGG

GAATTAACAAAAAACCGATGGAAAAGCTGGTTTTCTTATATTGATGCACTGGACGATGATTTATCTGAGT

CTTTAGAAAAATTCACATGCACAGTTGATATAGCTAATTTCTTGAGGGCATATTCATGGTCTGACGTCTT

AAAAGGGAAAAGGCTAATTGGTGCCACATTACCGTGTTTACTAGAGCAATTTAAGGTAAAGTGGATTAAT

TTATCTGAGGATTTAAGGGAACAATTTAATCTGTCTTCAAACTCAGAGTCAACTATAAATTTGTTGCCAT

ATGACTGCAAGGAACTACGACTTGGAGGAAGCAATGACACAGAGTTGAATTATGTCAGTTGTGCTCTCGA

CCGGAAAGTTGTCCAGAAACATCCCTCTGTTAATCGTCTTGCTTGGACGATAGGAAATCGAGCACCGTAT

ATAGGCTCACGGACAGAAGATAAGATCGGTTATCCTCCCTTGAGAGTAAACTGCCCATCAGCGGCACTTA

AAGAAGCTATTGAGATGGTTTCTAGATTATTATGGGTGACTCAAGGCACTGCAGACCGAGAGAAATTGCT

TATTCCTCTTCTCAATTCGAGAGTAAATCTGGACTATCAGACAGTACTTAACTTTTTACCTACACACTAC

TCAGGCAATATAGTTCATAGATATAATGATCAATATGGACAACATTCCTTTATGGCAAATAGGATGAGTA

ATACTTCTACACGTGCAATTATATCAACTAACACACTGGGGAAATACGCTGGGGGAGGTCAAGCTGCTAT

TGATAGTAATATAATCTTTCAAAATACTATTAATTTAGGGGTTGCAGTTTTAGATATTGCATTATCTCTT

GCTAAATTGTCGTCATCATCAAATGTCACTTTCCGTTTAATGTTAAATAAATGCTGCACACGGCATGTAC

CGTCCGAATACCTATTTTTTGATAAACCTTCAGATGTGGATCTGAACAAGTATATGGACAATGAACTGGT

TTATGACAATGACCCCCTTTGCAGTGGGATTAAAGGGAGATTAGGCAGAGTATCTCGATCAACACTATCA

TTGAGTTTGAATGTCAGTGACATTGGTTCTTATGACTTTCCAACTATTGCTGCATGGACACTAGGAGAAA

CTATAGTCGGAAGCATTTTTTCTGATGAGTCTTCTCAAAGCACAGATCCAATAAGTTCAGGTTGCACAAA

AACTTTCGTCACACATTTCCTTGTGTATCCAGTTGAGAGTATTTTTTATGCATTTGGAGCTAACTTAATA

GTGGAAAGTTTAAGTCCAAATAGAATCAAATCAATCAAGAACCTCTCAGATTTAACATTCCTTATATCAT

CCACAATCAGGAATTTATCACATAGATCACTTCGGATTCTTCAATCTACTTTCCGACATGAATTGGTGCT

TACCCGACTAGCCCATCATATACCGTTAATTTCTTTGATGCTAGGGGGCTCTGCAGGAGAAAAAAGTTCA

TCAGATGCCGTCCGGCTATTCCTTACAGCAAGTTATCAGAATTTCATCAACAATTTCAGTTGCCTGATAA

AAAAGGGTCAATCATCATTACCAGTTTGGCTTTACTTTCCTAGCGAAGGGCAACAGTTGAAACCTATATT

GAAAATCTTACAGAAGTTATCAGACTTGTTTTCACCTGACAAAGTTCAAAAGCGTAAAATTTTAGCTGAC

ACCTGTTATCCAATTGATTGCTTTTGGGTTTATCCAAGCAAATCCACAAGGACTAATCATTATTATGCAA

GCCTTAATTATTGGAGGGATAAAGCTAATAAGGTCAAGAATACTTCTTTTTCGCATTTGATAAATTGCTC

ATTTCCCGAATTTTCTTCACACACCAGTTCAGTCCCTTCTAATCAACAAATGACCAATTCAAAATATATT

GTTCATCCAGAAAATATCCCTGAAATGAATGCAAGGACCGAATTAATAAATTATGGAACAACAACTTTAC

AGGGGATGGACATCAAGATACCACTCTCGGAGCAAAATCTGGTCGAAAATTGTCGACCATCAAAGGGCAC

CAGATGCAAGGACAATCAAAAGATAACAAAACATGACCAGAGATATGAGAAGAAGGAGTCCTCATCGCGA

CAAATGCCCCCCGAAGACAACATGAAAACTCCTGCGTACATACATGGCTCCTCTCCGTCTCAAACCCTTA

TAAAATCACTAGATGTACATGAGGACTTTGATGCCCCAAAGGTAATCTTAAATTCTGAAATAAATAATCT

TAACCTCACAGATTGTACCCTTAATACAAAGTTATTGACAACTCTTACCGGGACAGAAATTTTAGGTACA

AGTCCATTACAACCCTCTAGATATTCCTCAACTTCCAAGGAACGGTCTCTACTATCTAGAGAACAAGCTT

CATATTTATATGTTGATTGCAGCAATATTCCTTCTATCTCTCTAGACCCGGGTTTTCGGAATATGTCTGA

TCAGAATCAAGTTCAAATGTTAATCAATGCCTACAAACGTGATTTACATGCTTGTTTCGACAGCAATCAA

TTCTGTCGGTTTACGGGGGTAGTCTCATCAATGCATTACAAGCTTTATGATCTCTTGCCTCCAGGTGAAT

TGAGAAAGGCAATTTGCTTGGCTGAAGGGGAAGGGAGTGGCGCTCGGTTACTTTTGAAGTGGAAGGAAAC

TGATTATTTATTCTTCAACACTCTAGCTACAGATTCACAGCAAGAGGCAGAGATCTTGAGTGGCCGGGTA

ATACCGAGAATGTTGTATAACATAGACAAATTAAGTGTTTTGCTTGAATCAAGAAGACTAATATTGAACA

ATCTAACTATCCAAATTACAGATATTACAAATCCATTATGGCTAGATTCTGTAATACAATATTTACCTGA

AGATAGTGACATTCTTACAATGGATGCAGAGACCACTAAGGATGAAACAAGGGAACAGCTTTATAAGACC

ATTGTGAATATTTGGACACGTACTTCTCCTAATATCCCAAAAATTAGTATCATCAAGGTGTTCTTATTAG

ACTATGATGGGACTTTATTCTTAATGAGGAATGCCATTCAGTATTATGGGCAAGTTCAACTCAAGAAACC

ATATAGCTCAAATGCGAAAAACTCAGAATGGTACTTGTGTTGCGGTAAACGAAGGATTCAACGACTCCAA

ATTGATTTCTCAGACCAGGTGGGAATTTTTTTGATTTGTAAAGCAATGTCGCGCCAGAGACAAGCAATTC

CTTACTGGTTAAAGCATATAGAAAAGAATTATCCTGCTTCATTACATGAGTTTTTCCTAACTTTAGGTTT

TCCTTCTCTAGAGTCATCTTTCTGCCATCGTTACACTATTCCATTCAGTGAAGGAAAGGCTCTTTTTCAC

AAGGTCCAGTCTTATGTTCGTCAAGGCAAACAACATTTACATTCTCTTATGTTGGACTACGAAAACAATT

CCCCTTTACTAGACTTGAGAAACCACTTTATTTGCTCATTAAGAGGAAAGATAACTAAATACTACAATGA

TATCTTAAAGTTAAATCTAGTCATCAAGGCCGTAGAAAAAGGTAAAAATTGGTCACAACTTGTTGAGACC

CTTCCTAATATGCATTCAGTATGCATAGTACACGTAGATCATGAGTGTTTTGGGTGCGAAAAACGGTTAT

TACTTAAATTGGATTTTATTAGAAATACAAAGATCGCAGAACAAAAATTGCTTAATAGGGTAATCGGGTA

TATCCTATTCTTTCCGTTCGGTCTGTTTAAGTCTGGATCATTAAGGGCATAACTTTAACAGAGAGAACTT

CATTTAATTTACAAAAACAATCTATTTAAAAGTGGGGGTTGCATTTTCTAGGGTATTGTATGAGAAATAA

TAAAATAAACAAGAAAAAGAAAAAGCTATTAGACAGCTTACTTTACAGAAGATAATCTTATAACGCCTTA

GGCCTTACACAGGTTAGAAAATTGCGAGCACAGATTAACTTGTGATTGAACGTTCAGTCACAACAATGGT

GACTTTTTAATGTTGGTTACTCAAACATTATTGATCATAATTAGTATTAACATTAGCACATTGGGTGAAT

CCTTGAGCTTTATCCCTAATATAAGGCGGGAGATTAAGAAAATACTGAGATATACTAGTTGAATTGAGTT

ATGACATGCCATATATCATAAATATAAAAGAACTGTCTGCTGTAATCTATAAGCATCTCTTTTATATACA

TTGAAGAAAGAACGAGATTATCATTGAGATCAAAAAAACTACGTTACATTTTCTCTGATGACAAGTGATA

AAATTTAGTAATTAAATTTCTAGAATGTCAATGTGAATGTAAATTAAGAAAAACCAATATATAAAATTAA

AAAAGTAAAAAGCTTTGATACAAGTAACACAAAACATTCTTCATCTTTTTAGTGTGTCCA