# Maximum Mismatch between Six-Residue Sequences of SARS-CoV-2 Spike Proteins and Human Proteome

# Maximum Mismatch between Six-Residue Sequences of SARS-CoV-2 Spike Proteins and Human Proteome

Mohammadali Towhidlou

A Thesis in the Field of Bioengineering and Nanotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2024

# Abstract

A first step toward finding effective antivirals for Covid-19 may be the identification of immunogenic, conserved, short and accessible epitopes on the surface molecules of human coronavirus. An important requirement, however, would be to be sure these epitopes do not align well with human proteins, thereby reducing the possibility of off-target interactions by the antiviral. So far, the focus of the medical and scientific communities has been on finding vaccines that can combat the disease preemptively, but this approach might fall short upon the next mutation of the virus. Would there be a safe and sustainable treatment strategy for a viral infection such as Covid-19 when new variants or strains of the virus are evolved? In this study, it was hypothesized that an immunogenic, conserved, short and accessible epitope of the virus with maximum mismatches when aligned with human protein sequences might be a good target for sustainable therapeutics such as antivirals. This hypothesis was based on the premise that antivirals are fabricated to bind strongly and selectively to a target molecule in the form of paratope-epitope complex while skipping all other undesirable bindings. This exquisite affinity and specificity ultimately result in maximizing the antiviral's precision and minimizing adverse immunological reactions. To achieve this goal, bioinformatic approaches such as BLAST querying and utilizing different NCBI databases were used. In this study, the results were narrowed down to one six-residue sequence, IKWPWY, in the spike protein of SARS-CoV-2, with two mismatches against

human protein sequences. Finally, IKWPWY, being a conserved sequence among all strains of human coronavirus, was demonstrated to be likely a reliable and steady target for therapeutics. The conclusion from these findings was that by employing exhaustive bioinformatic algorithms, we might not be too far away from discovering an antiviral that can be used against all strains and variants of human coronavirus provided that the hypothesis of ‘Maximum Mismatch’ is true.

## Table of Contents

## List of Figures

# Chapter I

## Introduction

Covid-19 has revealed our vulnerabilities, such as high morbidity and mortality, and the demand for more assuring treatment strategies for future pandemics. Although there is no FDA-approved antiviral that can claim clearing the Covid-19 infection, the current preventative care, the mRNA vaccine, has proven to be somewhat effective as a prophylactic measure in fighting and preventing the spread of Covid-19. The major disadvantages of vaccines are three-fold: 1) There is no guarantee that a vaccine can work against a new variant of a virus; 2) Booster shots are hypothesized by many scientists to be needed periodically for many kinds of vaccines; and 3) The Covid outbreak unveiled people's reluctance to mandatory inoculation perhaps due to the fear of receiving the alleged side effects i.e. Serious Adverse Reactions. This pushback can potentially put our society at risk. Nonetheless, it is a fact that unlike targeted antivirals that can be used once and on an emergency basis, hence lowering the chances of receiving the possible Serious Adverse Reactions, healthy people are urged to be immunized to halt a potential epidemic.

Undoubtedly, to circumvent the mentioned complications with vaccines in general, and the Covid-19 vaccine in particular, an alternative approach such as developing efficient antivirals needs to be considered. The goal is to find stable, accessible epitopes on the virus that are conserved among variants, and to evoke strong immune response with antiviral ligands that ignore human proteins selectively. Klase (as cited in Avril, 2020) in an article entitled “Why the coronavirus and most other viruses have no cure” stated that the reason viruses are so hard to treat is their wide variety. Viruses tend to mutate much more rapidly than bacteria, and that is why a drug to suppress the viral infection is likely to be highly effective only against the variant for which the drug was made.

To show the depth of the variation among viruses, coronaviruses can be investigated. Coronaviruses, divided into four genera, infect many species (hosts) such as mammals and birds. According to Centers for Disease Control and Prevention (2020), there are seven strains of human coronavirus: Four HCoV strains (causing common cold), SARS-CoV (causing SARS), MERS-CoV (causing MERS), and SARS-CoV-2 (causing Covid-19). A variant is the result of a single or multiple point change(s) in the nucleotide sequence of the original strain. A lineage is a collection of sub-variants that define a specific line of the original variant. A lineage may further divide into sub-lineages comprised of similarly identified sub-variants. For example, SARS-CoV-2 as a strain, stemmed from the *Betacoronavirus* genus, has Omicron as a Variant of Concern that itself has many lineages and sub-lineages such as A and A.1 respectively (Figure 1).

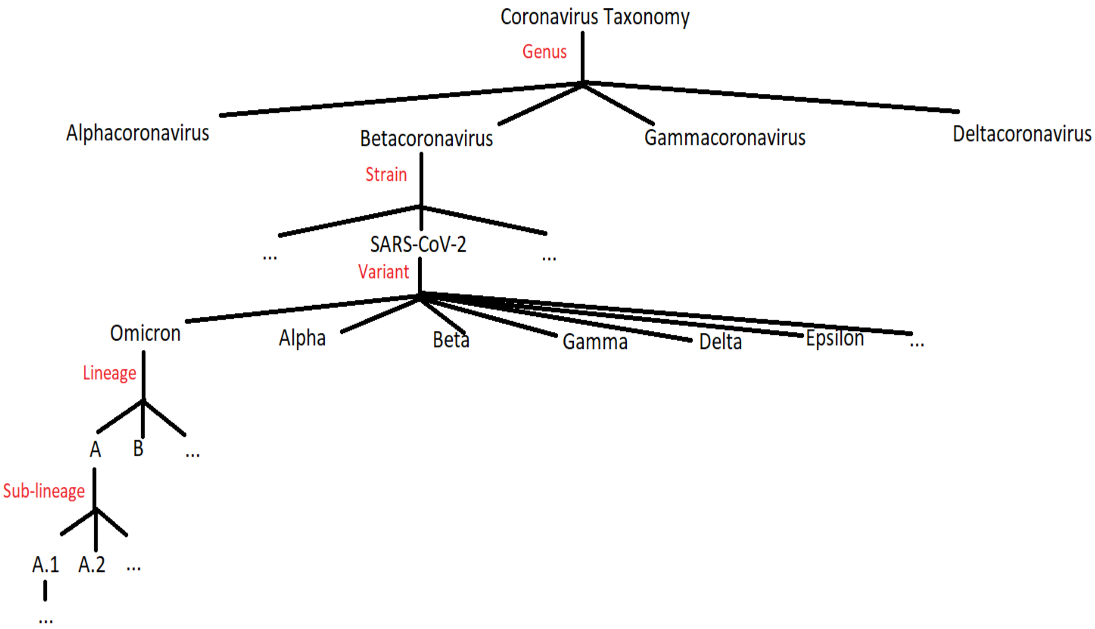


Figure 1. *Coronavirus Taxonomy*

To elucidate the complexity of dealing with a virus variant further, the structure of SARS-CoV-2 can be investigated. The SARS-CoV-2 genome has two main Open Reading Frames (ORFs), ORF1a and ORF1b, which contain two-thirds of the genome and ultimately encode 16 non-structural proteins. The remaining one-third, ORF2-ORF10, encodes four structural proteins including spike (S), nucleocapsid (N), membrane (M), and envelope (E) proteins, in addition to nine accessory proteins (Figure 2). The non-structural proteins have crucial roles such as viral replication and methylation. The S protein is responsible for viral attachment and entry into host cells. All S, N, M, and E proteins play a major role in pathogenesis and viral assembly, and the three surface proteins S, M, and E may be useful as vaccine or antiviral targets. The accessory proteins play a very crucial role in viral replication ("Tools for COVID-19 Research," 2022).

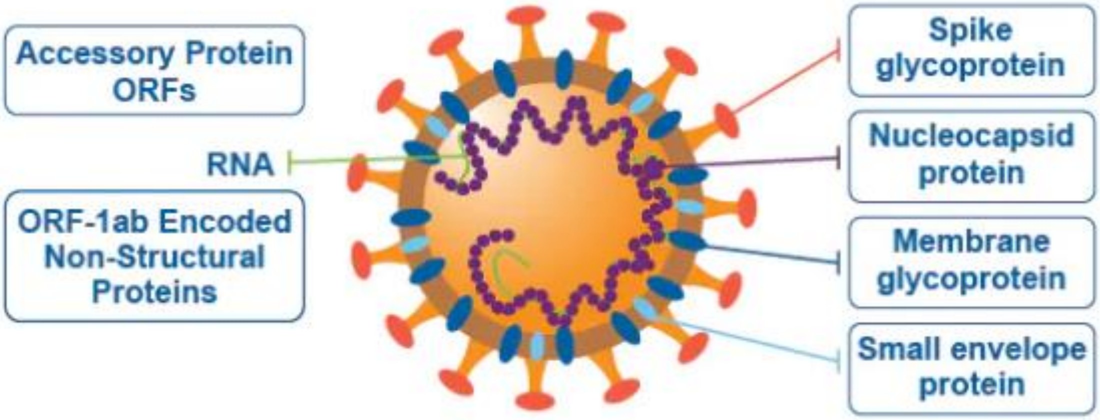


Figure 2. *Genome and Proteome Organization of SARS-CoV-2*

*Note.* From *Tools for COVID-19 Research,* (https://www.novusbio.com/support/sars-cov-research-resources)

Given that the commonly used therapeutics (vaccines and antibody cocktails) can target the S protein, a single mutation leading to a single amino acid change in the S protein creates a new variant that can potentially escape from vaccine or antibody-mediated immunity. Current vaccinology and pharmacology approaches do not consider a one-size-fits-all treatment to circumvent the mentioned complications with current therapeutics. Toward this goal, this study investigates whether there is a sequence that is particular to SARS-CoV-2 surface proteins, conserved among all variants, and as a bonus, conserved among all seven human coronaviruses, but with little or no similarity to *Homo sapiens* proteome. For this sequence to be immunogenic and responsive to drugs as an epitope, it must have certain properties such as being short and accessible. Per Atassi et al.'s discovery (1975), the antigenic reactive regions, i.e. epitopes of an antigen, are surprisingly small and about 6-7 residues. Sigma-Aldrich (2023) specifies the size of antigenic determinants, aka epitopes, as five to eight amino acid residues on the surface of the antigen. Epitopes and paratopes are commonly described as the unique amino acids, known as binding sites, of an antigen and antibody respectively. A paratope binds to an epitope and neutralizes the antigen that usually exists on the surface of a pathogenic entity such as coronavirus. According to Akbar et al. (2021), the most reliable way to identify a paratope-epitope pair is by solving its 3D structure and determining which amino acids contact each other. This paratope-epitope affinity relies on the quality of the match, like a lock and a key, and therefore they both should have approximately the same number of residues. To show how effective short polypeptide sequences are, Qi et al.'s research (2017) demonstrates that the hexapeptide PGPIPN can function as a paratope and bind to epitopes in alcohol-induced human liver cell lines and liver tissues of model mice. This therapeutic short string of amino acids can ultimately prevent and cure alcoholic fatty liver disease by affecting the expressions of genes and oxidative stress.

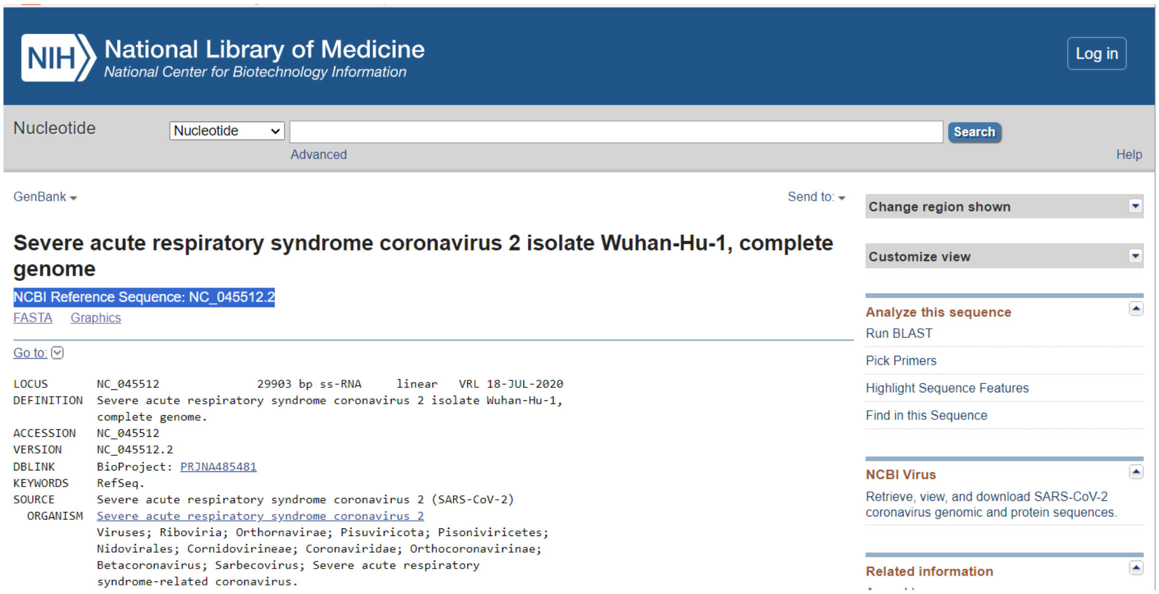
Conceptually, using this short conserved (resistant to mutation) SARS-CoV-2 sequence as an antigenic target for therapeutics may enable us to address all three vaccine complications with eradicating Covid-19. By having a single antiviral that is administered once and only when infected, a drug that can effectively bind only to the virus, it can be posited that upon a new surge of a virus variant, we can expect to readily have a safe effective therapy at our disposal.

# Chapter II

# Materials and Methods

## Step 1

The RefSeq entry of the complete genome of Severe Acute Respiratory Syndrome 2 was searched in the NCBI Entrez (Figure 3).



**Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome**

NCBI Reference Sequence: NC_045512.2

FASTA Graphics

Go to:

```
LOCUS       NC_045512               29903 bp ss-RNA     linear   VRL 18-JUL-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus; Severe acute respiratory
            syndrome-related coronavirus.
```

Figure 3. *SARS-CoV-2 NCBI RefSeq: NC_045512.2*

Further, all other six human coronavirus accession numbers were examined and a list of what seemed to be the most curated entries was made.

Step 2

NCBI BLASTP queries on all three SARS-CoV-2 surface molecules were run against the SARS-CoV-2 Non-redundant Protein Sequences (nr) database to see if the surface molecules were highly conserved across all variants of SARS-CoV-2 or not. The choice of the surface molecules and nr database was backed by the fact that epitopes exist on the surface of antigens and that all known variants/subvariants of the molecules should likely exist in the nr database. The goal in this step was to ensure about conservation across variants of SARS-CoV-2 which is a must-have feature and not across strains of human coronavirus which is a nice-to-have feature.

Step 3

All three S, E, and M surface molecules were segmented into six-residue sequences with five residue overlaps and the repeated sequences were excluded (Appendix 1). The rationale was that short surface molecule epitopes are easily accessible to antivirals and can be immunogenic i.e. successfully evoke an immune response.

Step 4

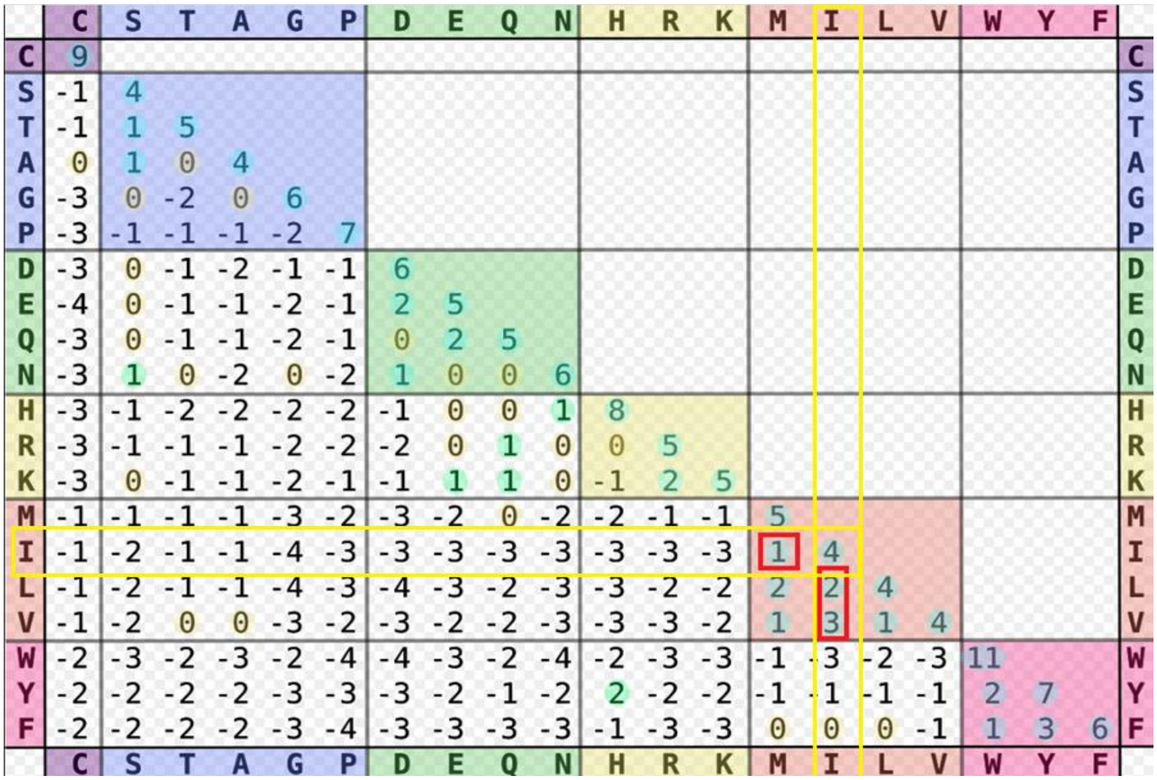
NCBI BLASTP queries on each six-residue sequence were run against the *Homo sapiens* nr exhaustively to identify which sequences would have the maximum number of mismatches when aligned with human protein sequences. The rationale for this step was as follows: The more an epitope mismatches with human proteins, the lower the chances of adverse immunological reactions to an antiviral administered to bind to that epitope selectively to neutralize the virus. Moreover, the choice of the nr database was justified by the fact that it is the broadest database, and it would be very unlikely for a discovered human sequence (such as immunoglobins which are very variable) to get neglected in the search.

In brief, random six-residue sequences were BLASTP'ed and it was noticed there were several of them with zero and one mismatch. This helped formulate less complex automation later. Then the aggregated six-residue sequences of each surface molecule of Step 3 in FASTA formats were submitted to the BLASTP form of NCBI webpage separately. For example, for the S protein, all 1268 FASTA'ed sequences were copied and pasted directly into the webpage form and the BLASTP was run in one submission. For these BLASTPs, the Max Target Sequences was set to 100 to ensure adequate hit coverage is sought and 20000 was entered for the E value to ensure the threshold is less stringent, leading to more chance hits being returned (Appendix 2-A). Then the results of BLASTPs for three surface molecules were downloaded in three separate text files using the 'Download All' dropdown menu (Appendix 2-B).

Furthermore, MATLAB was used to clean the noise from the text files, leaving them with batches of 100 hits for each sequence, with each hit including a Query line and a subsequent line i.e. the result of the alignment (Appendix 2-C). Considering the earlier findings about hits with zero and one mismatch, the MATLAB code was defined in such a way to skip the batches that had at least one subsequent line with > 4 matches. Also ignoring hits that had a gap(s) was defined in the code. If a batch was not skipped, its sequence was printed in the output file to later manually scan the sequence alignment of the entire batch for any gaps. This manual, visual, and automatic method of carefully examining the gaps in the text file also helped avoid executing complex logic to cover different scenarios of a gap in the code (Appendix 3).

### Step 5

Further it was examined if altering amino acids of IKWPWY by substituting them would optimize the mismatch composition or not. To achieve this goal, the matrix that was used for BLASTP queries, BLOSUM62 (Figure 4), was applied as follows:



| | C | S | T | A | G | P | D | E | Q | N | H | R | K | M | I | L | V | W | Y | F | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 9 | | | | | | | | | | | | | | | | | | | | C |
| S | -1 | 4 | | | | | | | | | | | | | | | | | | | S |
| T | -1 | 1 | 5 | | | | | | | | | | | | | | | | | | T |
| A | 0 | 1 | 0 | 4 | | | | | | | | | | | | | | | | | A |
| G | -3 | 0 | -2 | 0 | 6 | | | | | | | | | | | | | | | | G |
| P | -3 | -1 | -1 | -1 | -2 | 7 | | | | | | | | | | | | | | | P |
| D | -3 | 0 | -1 | -2 | -1 | -1 | 6 | | | | | | | | | | | | | | D |
| E | -4 | 0 | -1 | -1 | -2 | -1 | 2 | 5 | | | | | | | | | | | | | E |
| Q | -3 | 0 | -1 | -1 | -2 | -1 | 0 | 2 | 5 | | | | | | | | | | | | Q |
| N | -3 | 1 | 0 | -2 | 0 | -2 | 1 | 0 | 0 | 6 | | | | | | | | | | | N |
| H | -3 | -1 | -2 | -2 | -2 | -2 | -1 | 0 | 0 | 1 | 8 | | | | | | | | | | H |
| R | -3 | -1 | -1 | -1 | -2 | -2 | -2 | 0 | 1 | 0 | 0 | 5 | | | | | | | | | R |
| K | -3 | 0 | -1 | -1 | -2 | -1 | -1 | 1 | 1 | 0 | -1 | 2 | 5 | | | | | | | | K |
| M | -1 | -1 | -1 | -1 | -3 | -2 | -3 | -2 | 0 | -2 | -2 | -1 | -1 | 5 | | | | | | | M |
| I | -1 | -2 | -1 | -1 | -4 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | 1 | 4 | | | | | | I |
| L | -1 | -2 | -1 | -1 | -4 | -3 | -4 | -3 | -2 | -3 | -3 | -2 | -2 | 2 | 2 | 4 | | | | | L |
| V | -1 | -2 | 0 | 0 | -3 | -2 | -3 | -2 | -2 | -3 | -3 | -3 | -2 | 1 | 3 | 1 | 4 | | | | V |
| W | -2 | -3 | -2 | -3 | -2 | -4 | -4 | -3 | -2 | -4 | -2 | -3 | -3 | -1 | -3 | -2 | -3 | 11 | | | W |
| Y | -2 | -2 | -2 | -2 | -3 | -3 | -3 | -2 | -1 | -2 | 2 | -2 | -2 | -1 | -1 | -1 | -1 | 2 | 7 | | Y |
| F | -2 | -2 | -2 | -2 | -3 | -4 | -3 | -3 | -3 | -3 | -1 | -3 | -3 | 0 | 0 | 0 | -1 | 1 | 3 | 6 | F |
| | C | S | T | A | G | P | D | E | Q | N | H | R | K | M | I | L | V | W | Y | F | |

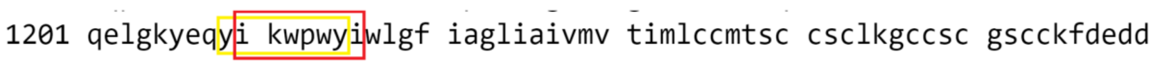
Figure 4. *BLOSUM62*

*Note.* Blosum62 scoring matrix is a quantitative approach that is used as a default by NCBI BLASTP to show the substitution score of an (i)th amino acid by a (j)th. Amino acids are categorized into groups with similar physicochemical properties while a positive score is assigned to more likely substitutions and a negative score to less likely substitutions. For instance, the cross section of I row and I column outputs I (+4), V (+3), L (+2), and M (+1) as valid positive substitutions for I. Conversely, replacing I with K for example has a BLOSUM62 score of -3.

**Firstly**, each amino acid of IKWPWY was substituted, one residue at a time, with an amino acid with a positive score. For example, 'I' was substituted with M, L and V, BLASTPs with the same parameters in Step 3 were run, and it was observed if the mismatches increased or at least a match lost its place to a partial match or not. The justification for this substitution was that an amino acid can be conveniently substituted with another amino acid with a positive score without sacrificing the properties of the polypeptide sequence as a whole. When an antiviral ligand is engineered for an epitope, the paratope-epitope binding affinity might not be lost if one of the residues of the ligand paratope, which is supposed to target a residue in the epitope, is substituted with another physicochemically close residue. Herein, the abstract substitution in the epitope IKWPWY, attempting to increase the mismatches against human proteins, demands a real substitution for the mirroring amino acid of the paratope. In other words, an actual substitution in a paratope residue targets an epitope residue that could have the potential of substitution retrospectively. Ultimately, this change of amino acid in the paratope can result in decreasing of undesirable 'antiviral-human proteins' complex formations.

To perform this test properly, the substitution was limited to only one residue for each BLASTP query and without replacement; if one residue was substituted, it had to be placed back before the next residue was substituted. This was because simultaneous substitutions, no matter how high the score of the substituted residues are on the matrix, could compromise the overall binding affinity of the paratope-epitope interface.

**Secondly**, using the adjacent residues, the size of IKWPWY was increased to two seven-residue sequences, YIKWPWY and IKWPWYI, to examine whether the three mismatches goal could be achieved or not (Figure 5).



```
1201 qelgkyeqyi kwpwyiwlgf iagliaivmv timlccmtsc csclkgccsc gscckfdedd
```

Figure 5. *Residue-added IKWPWY*

Specifically, this test was performed because three mismatches out of seven residues would result in 43% mismatch threshold which is superior to what was already discovered for two mismatches out of six residues i.e. 33%. Moreover, two mismatches and a partial match across all hits would also be superior to two mismatches only. The attempt was to optimize the mismatch composition and therefore increase instability of the undesirable ‘paratope-human proteins’ complexes.

## Step 6

A BLASTP was run on IKWPWY against SARS-CoV-2 nr to confirm the sequence conservation across variants of the virus like what was done in Step 2 for the surface molecules. Moreover, a multiple pairwise BLASTP was run on IKWPWY of SARS-CoV-2 as the query against the S protein of the other six human coronaviruses as the subjects to see if the query is also conserved among all strains of the virus.

## Step 7

The S protein of SARS-CoV-2 was viewed in JSmol to locate the position of the IKWPWY sequence on the protein. The sequence which includes coordinates 1210-1216 of the protein did not appear in the 3D structure.

# Chapter III

## Results

Step 1 resulted in the following list of accession numbers:

NC_045512.2: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NC_004718.3: SARS coronavirus Tor2, complete genome

NC_019843.3: Middle East respiratory syndrome-related coronavirus isolate HCoV-EMC/2012, complete genome

NC_002645.1: Human coronavirus 229E, complete genome

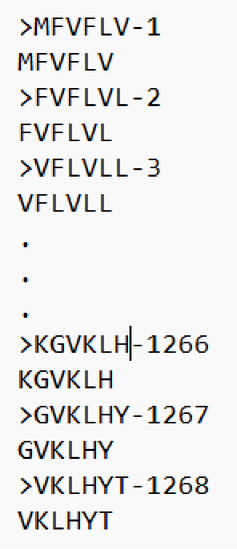
NC_006577.2: Human coronavirus HKU1, complete genome

NC_006213.1: Human coronavirus OC43 strain ATCC VR-759, complete genome

NC_005831.2: Human Coronavirus NL63, complete genome

In Step 2, as can be expected for variants of a virus, the first 5000 hits of the BLASTP showed that the 'query coverage' was maintained in the 99th percentile, confirming that all three S, E and M were highly conserved among variants of SARS-CoV-2. This liberated the sequence discovery from restricting it to any specific segment of the molecules until further verification (in Step 5).

The quest for small sequences in Step 3 resulted in 1268 six-residue sequences for the S protein (Figure 6). The other two surface molecules E and M (Appendix 4) were also segmented.



```
>MFVFLV-1
MFVFLV
>FVFLVL-2
FVFLVL
>VFLVLL-3
VFLVLL
.
.
.
>KGVKLH-1266
KGVKLH
>GVKLHY-1267
GVKLHY
>VKLHYT-1268
VKLHYT
```

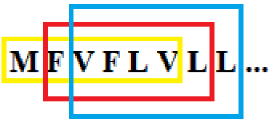
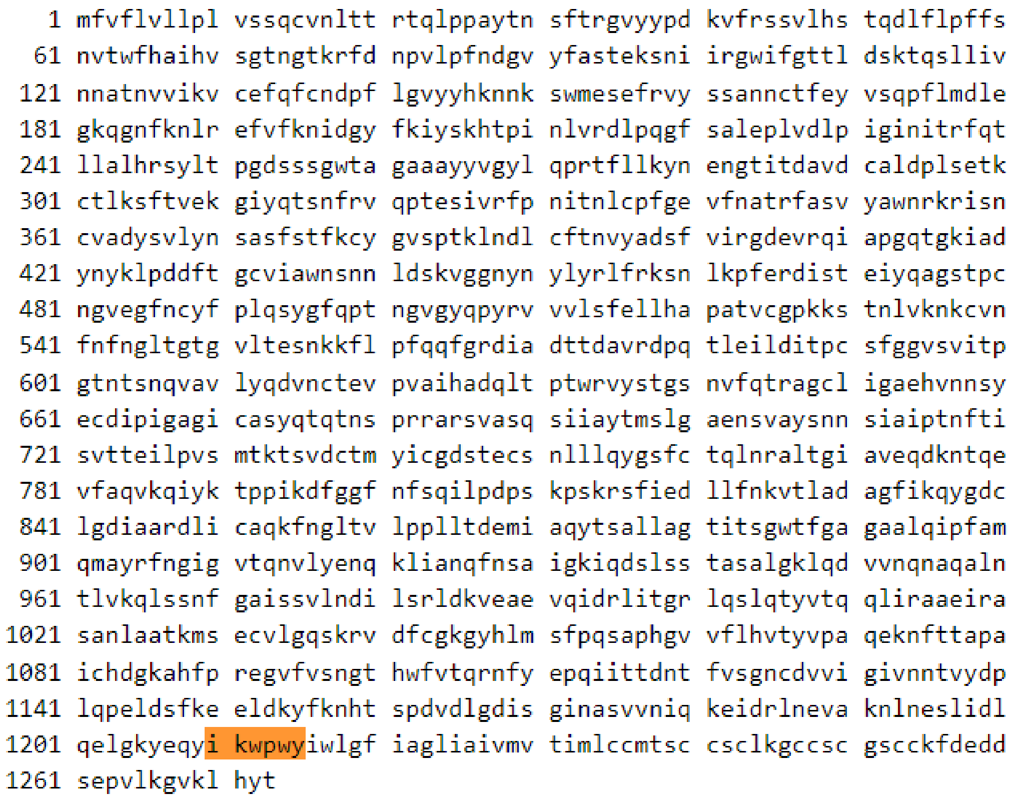


Figure 6. *Snapshot of FASTA Sequences for S Protein*

Step 4 resulted in many sequences with one mismatch but only one sequence, IKWPWY, with two mismatches across all hits, and in the S protein (Figure 7).



```
   1 mfvflvllpl vssqcvnltt rtqlppaytn sftrgvyypd kvfrssvlhs tqdlflpffs
  61 nvtwfhaihv sgtngtkrfd npvlpfndgv yfasteksni irgwifgttl dsktqslliv
 121 nnatnvvikv cefqfcndpf lgvyyhknnk swmesefrvy ssannctfey vsqpflmdle
 181 gkqgnfknlr efvfknidgy fkiyskhtpi nlvrdlpqgf saleplvdlp iginitrfqt
 241 llalhrsylt pgdsssgwta gaaayyvgyl qprtfllkyn engtitdavd caldplsetk
 301 ctlksftvek giyqtsnfrv qptesivrfp nitnlcpfge vfnatrfasv yawnrkrisn
 361 cvadysvlyn sasfstfkcy gvsptklndl cftnvyadsf virgdevrqi apgqtgkiad
 421 ynyklpddft gcviawnsnn ldskvggnyn ylyrlfrksn lkpferdist eiyqagstpc
 481 ngvegfncyf plqsygfqpt ngvgyqpyrv vvlsfellha patvcgpkks tnlvknkcvn
 541 fnfngltgtg vltesnkkfl pfqqfgrdia dttdavrdpq tleilditpc sfggvsvitp
 601 gtntsnqvav lyqdvnctev pvaihadqlt ptwrvystgs nvfqtragcl igaehvnnsy
 661 ecdipigagi casyqtqtns prrarsvasq siiaytmslg aensvaysnn siaiptnfti
 721 svtteilpvs mtktsvdctm yicgdstecs nlllqygsfc tqlnraltgi aveqdkntqe
 781 vfaqvkqiyk tppikdfggf nfsqilpdps kpskrsfied llfnkvtlad agfikqygdc
 841 lgdiaardli caqkfngltv lppltdemi aqytsallag titsgwtfga gaalqipfam
 901 qmayrfngig vtqnvlyenq klianqfnsa igkiqdslss tasalgklqd vvnqnaqaln
 961 tlvkqlssnf gaissvlndi lsrldkveae vqidrlitgr lqslqtyvtq qliraaeira
1021 sanlaatkms ecvlgqskrv dfcgkgyhlm sfpqsaphgv vflhvtyvpa qeknfttapa
1081 ichdgkahfp regvfvsngt hwfvtqrnfy epqiittdnt fvsgncdvvi givnntvydp
1141 lqpeldsfke eldkyfknht spdvdlgdis ginasvvniq keidrlneva knlneslidl
1201 qelgkyeqyi kwpwyiwlgf iagliaivmv timlccmtsc csclkgccsc gscckfdedd
1261 sepvlkgvkl hyt
```

Figure 7. *Position of IKWPWY in S Protein*

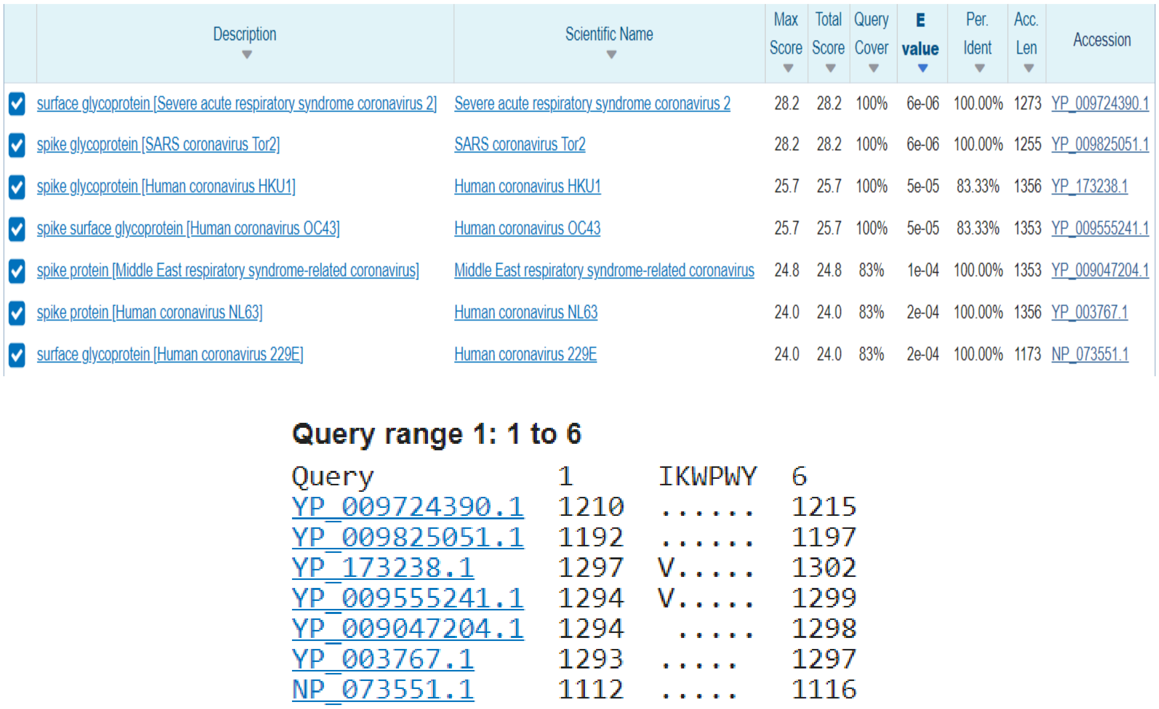
Nonetheless, no sequence with greater than two mismatches was found, confirming that an optimal sequence with six mismatches, i.e. zero similarity to human proteins sequences, would be impossible to find. A sample of zero and one mismatched sequences among all three surface molecules is depicted in Appendix 5.

As an illustration, a BLASTP was run on IKWPWY as the query and human proteins as the subjects for the first 100 hits (Appendix 6). Even one hit with fewer than two mismatches could have disqualified IKWPWY as the candidate sequence to conduct the remainder of the study on. Also, gaps in the alignment could qualify or disqualify a candidate depending on where the gap is. For example, a gap between the third and fourth residues on a hit for IKWPWY would mark it as a hit with greater than two mismatches and qualify it immediately but not necessarily a gap between the fourth and fifth residues. That was why a careful visual examination of the winning candidates was foreseen in the automation.

Looking at the description table of the BLASTP (Appendix 7), it was confirmed that no fewer than two mismatches were possible across all hits, which warranted 66% query coverage and 33% mismatch threshold. There were a few hits however with 83% query coverage due to the partial matches as the E value increased.

In Step 5, running all different possibilities that the BLOSUM62 matrix permitted for all six residues, no sequence with three mismatches or at least two mismatches and a partial match across all hits was found (Appendix 8). Moreover, as can be seen in Appendix 9, the extra tests that included adjacent amino acids were not successful either and the mismatch composition was not optimized.

In Step 6, the first 5000 hits showed 100% query coverage for all hits, indicating IKWPWY can be used as a stable target. Moreover, as demonstrated in Figure 8, at least five out of six residues of IKWPWY were aligned with all strains of human coronavirus, which made us confident that upon an emergence of a new variant or strain, IKWPWY would probably remain conserved and be used as a steady target for the same therapeutics.



| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | surface glycoprotein [Severe acute respiratory syndrome coronavirus 2] | Severe acute respiratory syndrome coronavirus 2 | 28.2 | 28.2 | 100% | 6e-06 | 100.00% | 1273 | YP_009724390.1 |
| ☑ | spike glycoprotein [SARS coronavirus Tor2] | SARS coronavirus Tor2 | 28.2 | 28.2 | 100% | 6e-06 | 100.00% | 1255 | YP_009825051.1 |
| ☑ | spike glycoprotein [Human coronavirus HKU1] | Human coronavirus HKU1 | 25.7 | 25.7 | 100% | 5e-05 | 83.33% | 1356 | YP_173238.1 |
| ☑ | spike surface glycoprotein [Human coronavirus OC43] | Human coronavirus OC43 | 25.7 | 25.7 | 100% | 5e-05 | 83.33% | 1353 | YP_009555241.1 |
| ☑ | spike protein [Middle East respiratory syndrome-related coronavirus] | Middle East respiratory syndrome-related coronavirus | 24.8 | 24.8 | 83% | 1e-04 | 100.00% | 1353 | YP_009047204.1 |
| ☑ | spike protein [Human coronavirus NL63] | Human coronavirus NL63 | 24.0 | 24.0 | 83% | 2e-04 | 100.00% | 1356 | YP_003767.1 |
| ☑ | surface glycoprotein [Human coronavirus 229E] | Human coronavirus 229E | 24.0 | 24.0 | 83% | 2e-04 | 100.00% | 1173 | NP_073551.1 |

**Query range 1: 1 to 6**

```
Query           1     IKWPWY  6
YP_009724390.1  1210  ......  1215
YP_009825051.1  1192  ......  1197
YP_173238.1     1297  V.....  1302
YP_009555241.1  1294  V.....  1299
YP_009047204.1  1294   .....  1298
YP_003767.1     1293  .....   1297
NP_073551.1     1112  .....   1116
```

Figure 8. *Multiple Pairwise BLASTP of IKWPWY and S Protein of Human Coronaviruses*

In Step 7, the lack of a 3D structure for IKWPWY (tagged as an unmodeled region in JSmol) could be interpreted as excessive conformational disorder in the IKWPWY region. This yielded some uncertainty as to whether the residues of IKWPWY were structured linearly, at least had a 2D distribution, or they were distributed in a 3D format.

# Chapter IV

# Discussion

The attempt in this study was to support the proposed hypotheses of 'Maximum Mismatch', with bioinformatic data, as a method of discovering effective antivirals for viral infections such as Covid-19. This hypothesis was based on the premise that if a short immunogenic sequence as an epitope, a sequence unique enough to not align well with human proteins, can be found on the surface of all variants of an antigenic invader, it would then be possible to discover a common targeted drug to bind selectively to that sequence. Of course, the ideal sequence would be one whose BLASTP against human proteins would result in hits with the highest possible mismatch threshold across all hits.

Toward this goal, SARS-CoV-2 surface molecules were segmented to six-residue sequences and BLASTP hits were narrowed down to a unique 100% conserved sequence across all SARS-CoV-2 variants: IKWPWY with at least two mismatches when aligned with human protein sequences. It was further tried to optimize the mismatch composition to no avail. It would be reasonable to predicate that a paratope designed for a six-residue epitope, an epitope that aligns with the maximum of four out of six residues of all human protein sequences, should have a lower chance of forming undesirable 'paratope-human proteins' complexes. Ultimately, this should result in the ability for the antiviral to avoid nonspecific interactions with human proteins and instead selectively bind to the target viral antigen. Finally, it was demonstrated that IKWPWY was highly conserved among all strains of human coronavirus and thus upon the next emergence of a new strain, we might have the same target and the same antiviral readily available for treatment.

Nevertheless, the scope of this study did not cover the other contributing factors that qualifies a sequence as an immunogenic epitope. These factors may range anywhere from the spatial structure of the sequence all the way to the properties of the amino acids of paratope-epitope side chains. To this note, the 3D structure of IKWPWY was not viewable in JSmol. One of the assumptions for finding mismatches, considering the short size of the sequences, was that we would be probably dealing with both linear viral sequences and linear aligning human protein sequences. The other assumption was that selecting the right amino acids for paratope should provide high enough specificity to target only a certain epitope in the vast biological milieu of molecules. Although there is complete confidence that, with the parameters used in this study, the highest mismatched surface sequence for SARS-CoV-2 was discovered, it cannot be verified that the best performing target sequence *in vivo* was determined. Further studies, using a scripting programming language, can be conducted exhaustively to see if there exists a larger than six-residue epitope with a greater mismatch threshold than what was discovered in this study; an epitope whose 2D sequential or 3D conformational structure can be examined.

# Appendix 1

## Pseudocode for SARS-CoV-2 Surface Molecules Segmentation

Assume the name of the text file containing the strings of amino acids (each surface protein) is data.txt:

```
input = fopen(data.txt);   % opening data.txt in MATLAB workspace
A = fscanf(input);    % reading all characters of data.txt into A
fclose(input);                                 % closing data.txt
len = strlength(A);                         % finding length of A
size = len/6;                  % calculating size of output array
B(1) = A(1:6);        % B is output array and B(1) is first entry
for i = 1 to size
       B(i+1) = A(6*i: 6*i+5):
end
output = fopen(out.txt);            % creating out.txt for output
fprintf(output,B);                       % writing B into out.txt
fclose(output);                                 % closing out.txt
```

The above code generates the first series of 6-tuple amino acids. This is repeated five times: For the second series, the first character (amino acid) read from data.txt is omitted and the code is run on the file again. For the third series, the first two characters are omitted and so on.

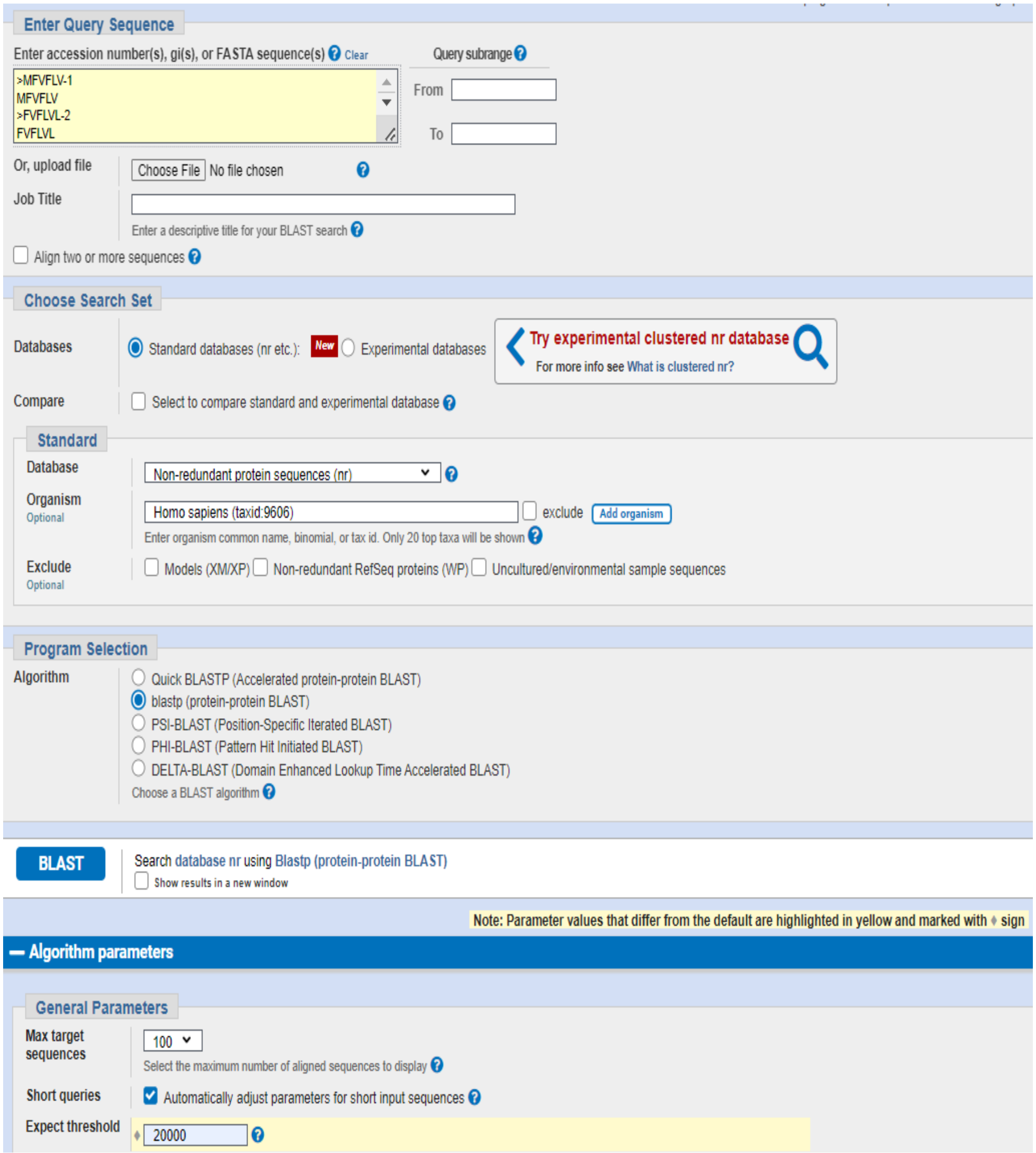
```
input = fopen(data.txt);   % opening data.txt in MATLAB workspace
AA = fscanf(Input);  % reading all characters of data.txt into AA
A = AA(j: ); % j=1 generates first series, j=2 second series, etc
```

```
fclose (input);                                % closing data.txt
len = strlength(A);                         % finding length of A
size = len/6;                  % calculating size of output array
B(1) = A(1:6);        % B is output array and B(1) is first entry
for i = 1 to size
       B(i+1) = A(6*i: 6*i+5):
end
output = fopen(out.txt);            % creating out.txt for output
fprintf(output,B);                       % writing B into out.txt
fclose(output);                                 % closing out.txt
```

# Appendix 2

## Discovering Maximum Mismatch

**A.**



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>MFVFLV-1
MFVFLV
>FVFLVL-2
FVFLVL
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Homo sapiens (taxid:9606) □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

— Algorithm parameters

General Parameters

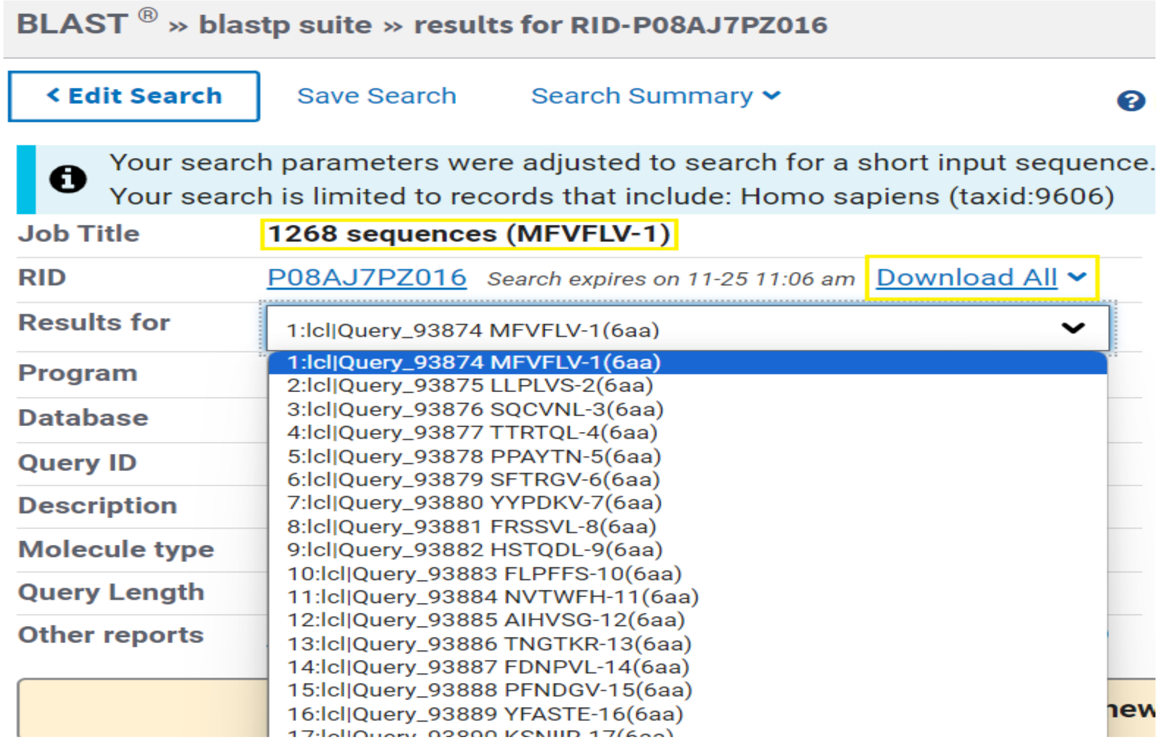
Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: ♦ 20000

**B.**



BLAST ® » blastp suite » results for RID-P08AJ7PZ016

Your search parameters were adjusted to search for a short input sequence.
Your search is limited to records that include: Homo sapiens (taxid:9606)

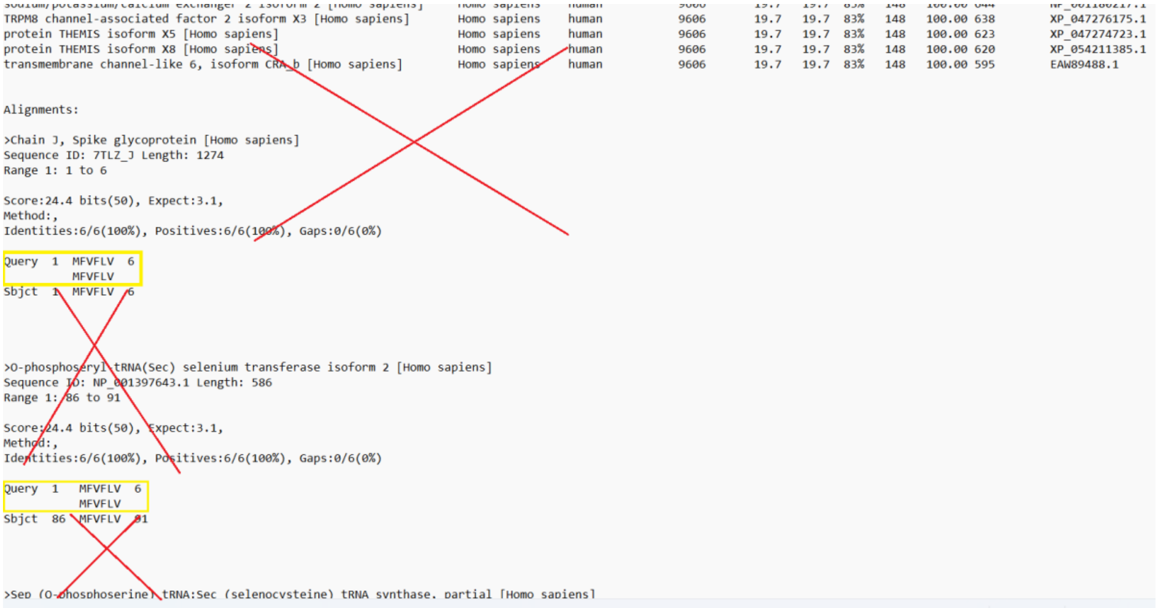
Job Title: 1268 sequences (MFVFLV-1)

RID: P08AJ7PZ016 *Search expires on 11-25 11:06 am* Download All

Results for: 1:lcl|Query_93874 MFVFLV-1(6aa)

- 1:lcl|Query_93874 MFVFLV-1(6aa)
- 2:lcl|Query_93875 LLPLVS-2(6aa)
- 3:lcl|Query_93876 SQCVNL-3(6aa)
- 4:lcl|Query_93877 TTRTQL-4(6aa)
- 5:lcl|Query_93878 PPAYTN-5(6aa)
- 6:lcl|Query_93879 SFTRGV-6(6aa)
- 7:lcl|Query_93880 YYPDKV-7(6aa)
- 8:lcl|Query_93881 FRSSVL-8(6aa)
- 9:lcl|Query_93882 HSTQDL-9(6aa)
- 10:lcl|Query_93883 FLPFFS-10(6aa)
- 11:lcl|Query_93884 NVTWFH-11(6aa)
- 12:lcl|Query_93885 AIHVSG-12(6aa)
- 13:lcl|Query_93886 TNGTKR-13(6aa)
- 14:lcl|Query_93887 FDNPVL-14(6aa)
- 15:lcl|Query_93888 PFNDGV-15(6aa)
- 16:lcl|Query_93889 YFASTE-16(6aa)

**C.**



```
sodium/potassium/calcium exchanger 2 isoform 2 [Homo sapiens]      Homo sapiens   human      9606     19.7   19.7  83%   148   100.00 644      NP_001186217.1
TRPM8 channel-associated factor 2 isoform X3 [Homo sapiens]        Homo sapiens   human      9606     19.7   19.7  83%   148   100.00 638      XP_047276175.1
protein THEMIS isoform X5 [Homo sapiens]                           Homo sapiens   human      9606     19.7   19.7  83%   148   100.00 623      XP_047274723.1
protein THEMIS isoform X8 [Homo sapiens]                           Homo sapiens   human      9606     19.7   19.7  83%   148   100.00 620      XP_054211385.1
transmembrane channel-like 6, isoform CRA_b [Homo sapiens]         Homo sapiens   human      9606     19.7   19.7  83%   148   100.00 595      EAW89488.1


Alignments:

>Chain J, Spike glycoprotein [Homo sapiens]
Sequence ID: 7TLZ_J Length: 1274
Range 1: 1 to 6

Score:24.4 bits(50), Expect:3.1,
Method:,
Identities:6/6(100%), Positives:6/6(100%), Gaps:0/6(0%)

Query  1  MFVFLV  6
          MFVFLV
Sbjct  1  MFVFLV  6



>O-phosphoseryl-tRNA(Sec) selenium transferase isoform 2 [Homo sapiens]
Sequence ID: NP_001397643.1 Length: 586
Range 1: 86 to 91

Score:24.4 bits(50), Expect:3.1,
Method:,
Identities:6/6(100%), Positives:6/6(100%), Gaps:0/6(0%)

Query  1   MFVFLV  6
           MFVFLV
Sbjct  86  MFVFLV  91



>Sep (O-phosphoserine) tRNA:Sec (selenocysteine) tRNA synthase, partial [Homo sapiens]
```

## Appendix 3

## Pseudocode for Finding Sequences with Maximum Mismatches

Assume the name of the original file is “file.txt”. Now refine file.txt and extract those lines starting with the word “Query” and the subsequent line into a file named “refined.txt”:

```
input=fopen(file.txt);     % opening file.txt in MATLAB workspace
refined=fopen(refined.txt);        % opening refined.txt to store
                   refined lines
while ~feof(input)         % checking file.txt lines till the end
     tline=fgetl(input);             % reading a line of file.txt
     if (tline starts with word “Query”)
         temp=fgetl(input);            % reading subsequent line
         fprintf(refined, tline /n);      % writing line starting
                    with “Query” into refined.txt
        fprintf(refined,temp /n);  % writing subsequent line into
                      refined.txt
     end if
end while
fclose(input);
fclose(result);
```

Now we have refined.txt where its lines look like the following (no more noise):

```
Query  1  YTWEW  5
          YTWEW
```

Query 1 YTWEW 5
        YTE EC
Query 1 …..

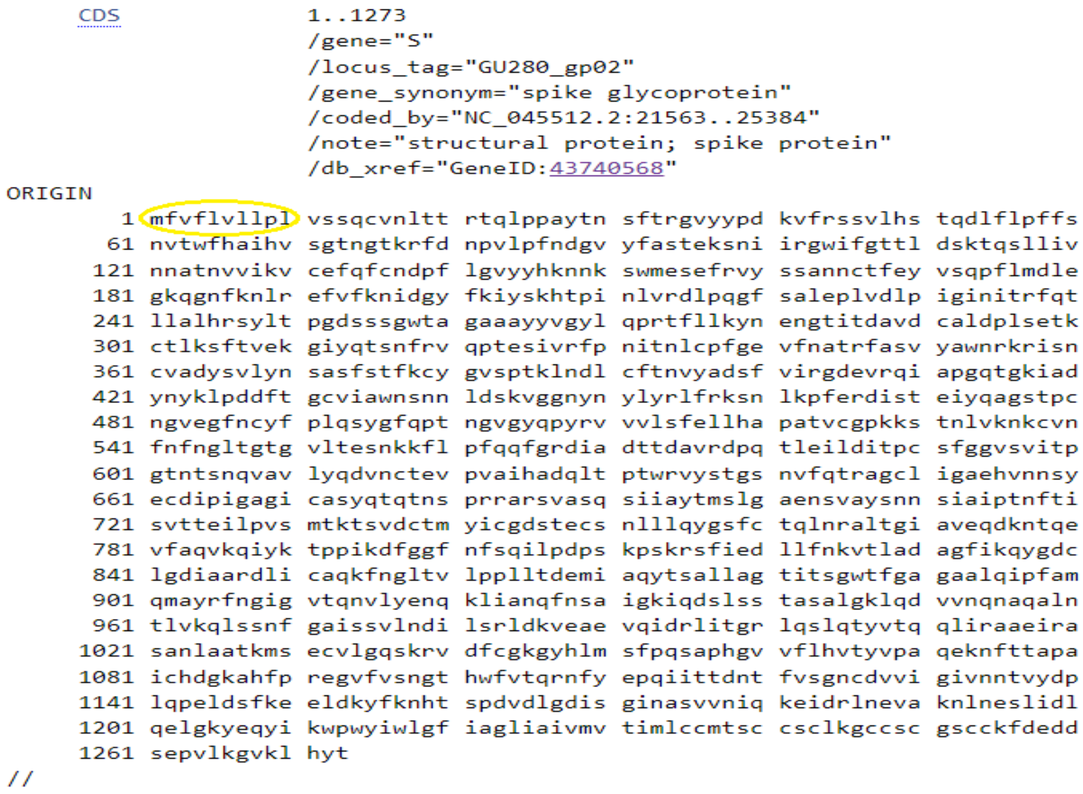
Now run the second code on this refined file which only contains lines starting with the word Query and the subsequent line:

```
input= fopen(refined.txt);
output=fopen(final.txt);
while ~feof(input)      % checking refined.txt lines till the end
     flag=0;           % resetting flag at beginning of each batch
     for i=1 to 100          % checking 100 entries of each batch
          Sequence=fgetl(input);          % reading line containing
                          word “Query” into var Sequence
          Temp=fgetl(input); % reading subsequent line into var Temp
          if (number of alphabetical characters in Temp > 4 && no
                          ‘-‘ in Sequence)
              increment flag        % flag content will change if
                     condition is met
          end if
     end for
     if flag = 0
          fprintf(Sequence, output);    % if condition was met in
                above batch, line containing word Query will be
                written into final.txt
     end if
end while
fclose(input);
fclose(output);
```

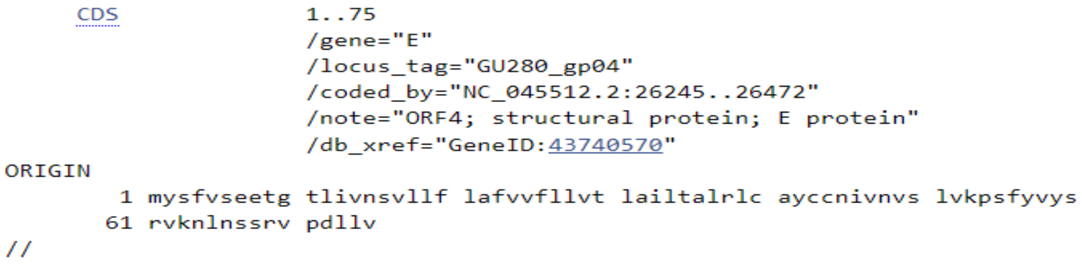
Now final.txt contains the desired results, if any.

# Appendix 4

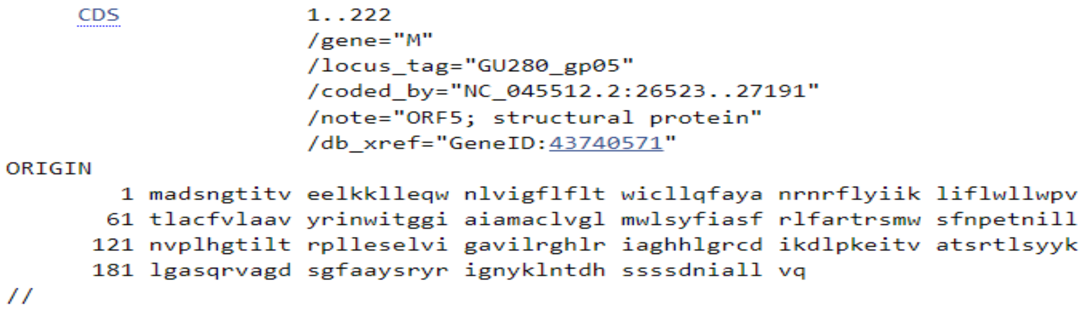
## S, E and M Surface Molecules



```
     CDS             1..1273
                     /gene="S"
                     /locus_tag="GU280_gp02"
                     /gene_synonym="spike glycoprotein"
                     /coded_by="NC_045512.2:21563..25384"
                     /note="structural protein; spike protein"
                     /db_xref="GeneID:43740568"
ORIGIN
        1 mfvflvllpl vssqcvnltt rtqlppaytn sftrgvyypd kvfrssvlhs tqdlflpffs
       61 nvtwfhaihv sgtngtkrfd npvlpfndgv yfasteksni irgwifgttl dsktqslliv
      121 nnatnvvikv cefqfcndpf lgvyyhknnk swmesefrvy ssannctfey vsqpflmdle
      181 gkqgnfknlr efvfknidgy fkiyskhtpi nlvrdlpqgf saleplvdlp iginitrfqt
      241 llalhrsylt pgdsssgwta gaaayyvgyl qprtfllkyn engtitdavd caldplsetk
      301 ctlksftvek giyqtsnfrv qptesivrfp nitnlcpfge vfnatrfasv yawnrkrisn
      361 cvadysvlyn sasfstfkcy gvsptklndl cftnvyadsf virgdevrqi apgqtgkiad
      421 ynyklpddft gcviawnsnn ldskvggnyn ylyrlfrksn lkpferdist eiyqagstpc
      481 ngvegfncyf plqsygfqpt ngvgyqpyrv vvlsfellha patvcgpkks tnlvknkcvn
      541 fnfngltgtg vltesnkkfl pfqqfgrdia dttdavrdpq tleilditpc sfggvsvitp
      601 gtntsnqvav lyqdvnctev pvaihadqlt ptwrvystgs nvfqtragcl igaehvnnsy
      661 ecdipigagi casyqtqtns prrarsvasq siiaytmslg aensvaysnn siaiptnfti
      721 svtteilpvs mtktsvdctm yicgdstecs nlllqygsfc tqlnraltgi aveqdkntqe
      781 vfaqvkqiyk tppikdfggf nfsqilpdps kpskrsfied llfnkvtlad agfikqygdc
      841 lgdiaardli caqkfngltv lppllddemi aqytsallag titsgwtfga gaalqipfam
      901 qmayrfngig vtqnvlyenq klianqfnsa igkiqdslss tasalgklqd vvnqnaqaln
      961 tlvkqlssnf gaissvlndi lsrldkveae vqidrlitgr lqslqtyvtq qliraaeira
     1021 sanlaatkms ecvlgqskrv dfcgkgyhlm sfpqsaphgv vflhvtyvpa qeknfttapa
     1081 ichdgkahfp regvfvsngt hwfvtqrnfy epqiittdnt fvsgncdvvi givnntvydp
     1141 lqpeldsfke eldkyfknht spdvdlgdis ginasvvniq keidrlneva knlneslidl
     1201 qelgkyeqyi kwpwyiwlgf iagliaivmv timlccmtsc csclkgccsc gscckfdedd
     1261 sepvlkgvkl hyt
//
```



```
     CDS             1..75
                     /gene="E"
                     /locus_tag="GU280_gp04"
                     /coded_by="NC_045512.2:26245..26472"
                     /note="ORF4; structural protein; E protein"
                     /db_xref="GeneID:43740570"
ORIGIN
        1 mysfvseetg tlivnsvllf lafvvfllvt lailtalrlc ayccnivnvs lvkpsfyvys
       61 rvknlnssrv pdllv
//
```



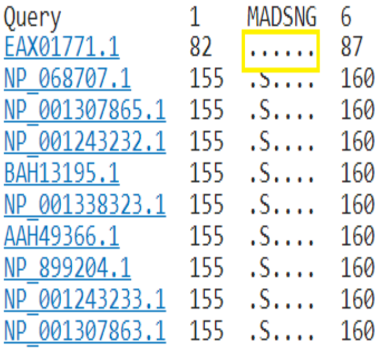
```
     CDS             1..222
                     /gene="M"
                     /locus_tag="GU280_gp05"
                     /coded_by="NC_045512.2:26523..27191"
                     /note="ORF5; structural protein"
                     /db_xref="GeneID:43740571"
ORIGIN
        1 madsngtitv eelkklleqw nlvigflflt wicllqfaya nrnrflyiik liflwllwpv
       61 tlacfvlaav yrinwitggi aiamaclvgl mwlsyfiasf rlfartrsmw sfnpetnill
      121 nvplhgtilt rplleselvi gavilrghlr iaghhlgrcd ikdlpkeitv atsrtlsyyk
      181 lgasqrvagd sgfaaysryr ignyklntdh ssssdniall vq
//
```

## Appendix 5

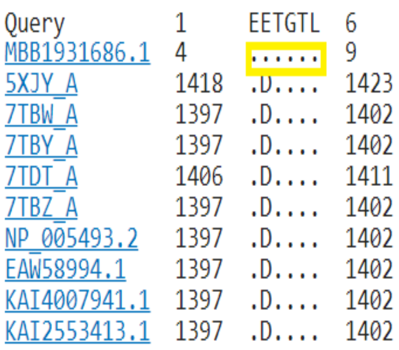
### Samples for Mismatched Sequences

### Zero mismatch

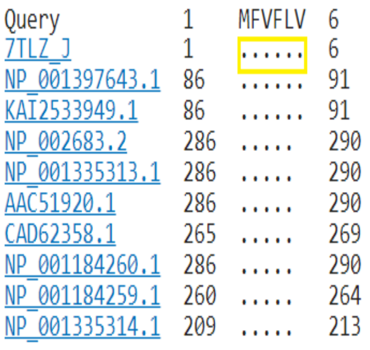
| MADSNG: M Protein | EETGTL: E Protein | MFVFLV: S Protein |
|---|---|---|



```
Query            1     MADSNG  6
EAX01771.1       82    ......  87
NP_068707.1      155   .S....  160
NP_001307865.1   155   .S....  160
NP_001243232.1   155   .S....  160
BAH13195.1       155   .S....  160
NP_001338323.1   155   .S....  160
AAH49366.1       155   .S....  160
NP_899204.1      155   .S....  160
NP_001243233.1   155   .S....  160
NP_001307863.1   155   .S....  160
```



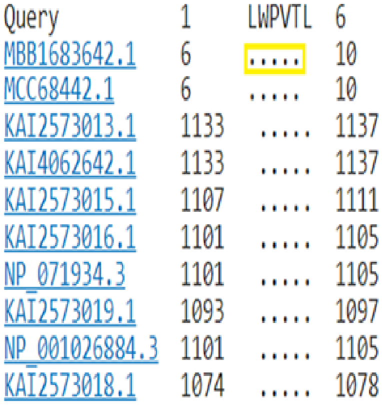
```
Query            1      EETGTL  6
MBB1931686.1     4      ......  9
5XJY_A           1418   .D....  1423
7TBW_A           1397   .D....  1402
7TBY_A           1397   .D....  1402
7TDT_A           1406   .D....  1411
7TBZ_A           1397   .D....  1402
NP_005493.2      1397   .D....  1402
EAW58994.1       1397   .D....  1402
KAI4007941.1     1397   .D....  1402
KAI2553413.1     1397   .D....  1402
```



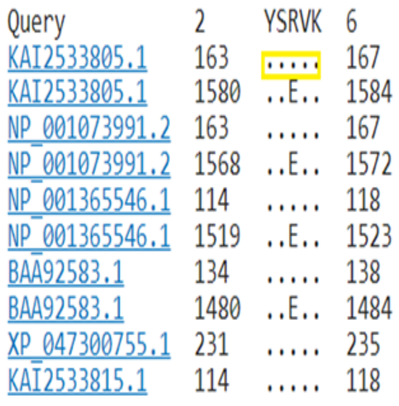
```
Query            1     MFVFLV  6
7TLZ_J           1     ......  6
NP_001397643.1   86    ......  91
KAI2533949.1     86    ......  91
NP_002683.2      286   .....   290
NP_001335313.1   286   .....   290
AAC51920.1       286   .....   290
CAD62358.1       265   .....   269
NP_001184260.1   286   .....   290
NP_001184259.1   260   .....   264
NP_001335314.1   209   .....   213
```

### One mismatch

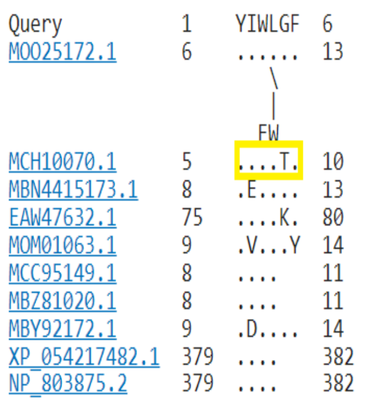
| LWPVTL: M protein | VYSRVK: E Protein | YIWLGF: S protein |
|---|---|---|



```
Query            1      LWPVTL  6
MBB1683642.1     6      .....   10
MCC68442.1       6      .....   10
KAI2573013.1     1133   .....   1137
KAI4062642.1     1133   .....   1137
KAI2573015.1     1107   .....   1111
KAI2573016.1     1101   .....   1105
NP_071934.3      1101   .....   1105
KAI2573019.1     1093   .....   1097
NP_001026884.3   1101   .....   1105
KAI2573018.1     1074   .....   1078
```



```
Query            2      YSRVK  6
KAI2533805.1     163    .....  167
KAI2533805.1     1580   ..E..  1584
NP_001073991.2   163    .....  167
NP_001073991.2   1568   ..E..  1572
NP_001365546.1   114    .....  118
NP_001365546.1   1519   ..E..  1523
BAA92583.1       134    .....  138
BAA92583.1       1480   ..E..  1484
XP_047300755.1   231    .....  235
KAI2533815.1     114    .....  118
```

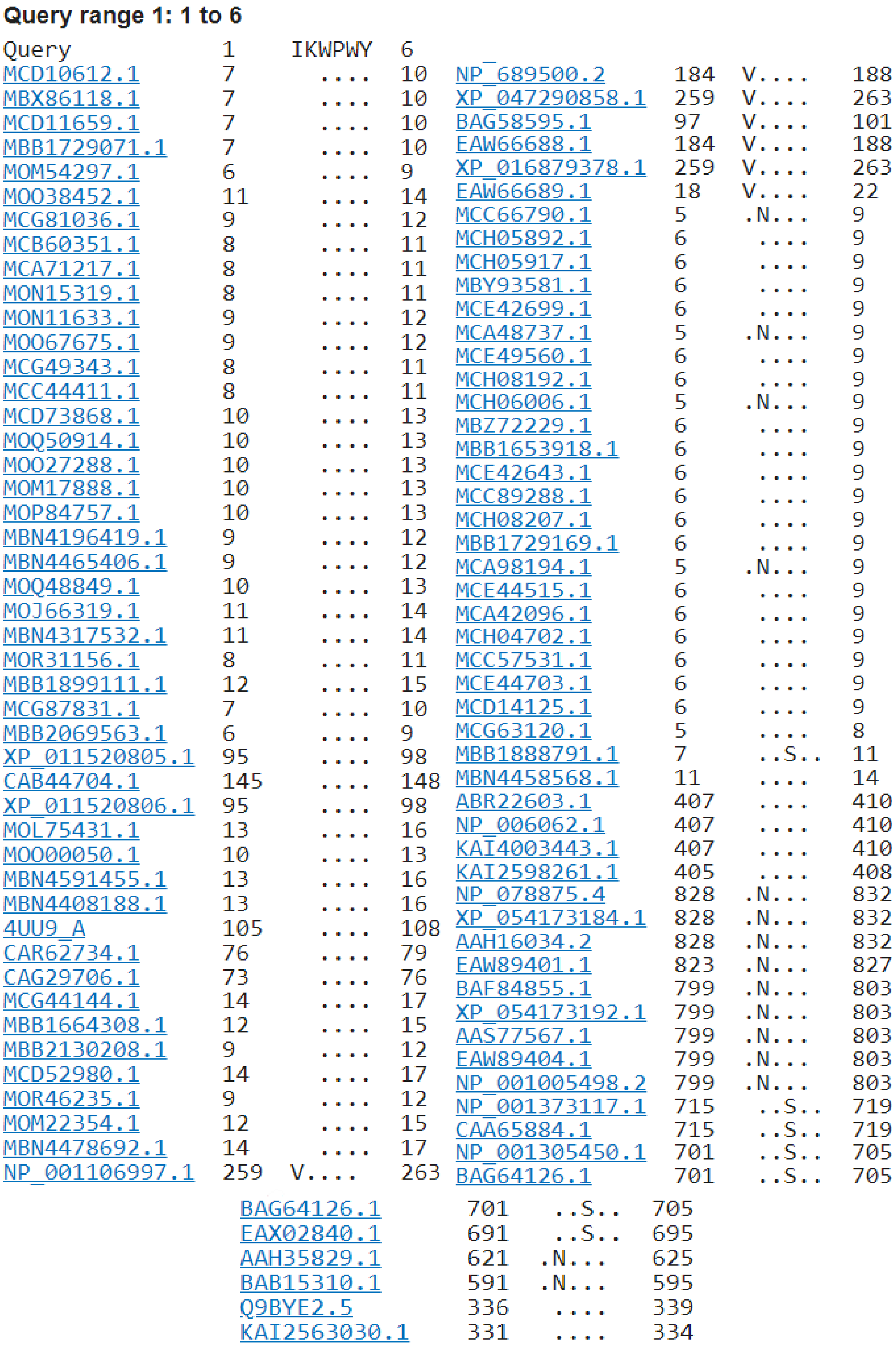


```
Query            1     YIWLGF  6
MO025172.1       6     ......  13
                          \
                          |
                          FW
MCH10070.1       5     ....T.  10
MBN4415173.1     8     .E....  13
EAW47632.1       75    ....K.  80
MOM01063.1       9     .V...Y  14
MCC95149.1       8     ....    11
MBZ81020.1       8     ....    11
MBY92172.1       9     .D....  14
XP_054217482.1   379   ....    382
NP_803875.2      379   ....    382
```

## Appendix 6

## IKWPWY Alignment against Homo sapiens nr

In a ‘query-anchored with dots for identities’ view, identities appear as dots (.), mismatches as blank, and partial matches as single letter abbreviations. Slashes (/) indicate gaps in the alignment. Gaps represent parts where Query or Subject have no counterpart. IKWPWP has maintained the minimum of two blank spots or at least one blank spot and one letter throughout. This is the only six-residue sequence of the SARS-CoV-2 surface molecules that has no more than four dots when aligned with human proteome.



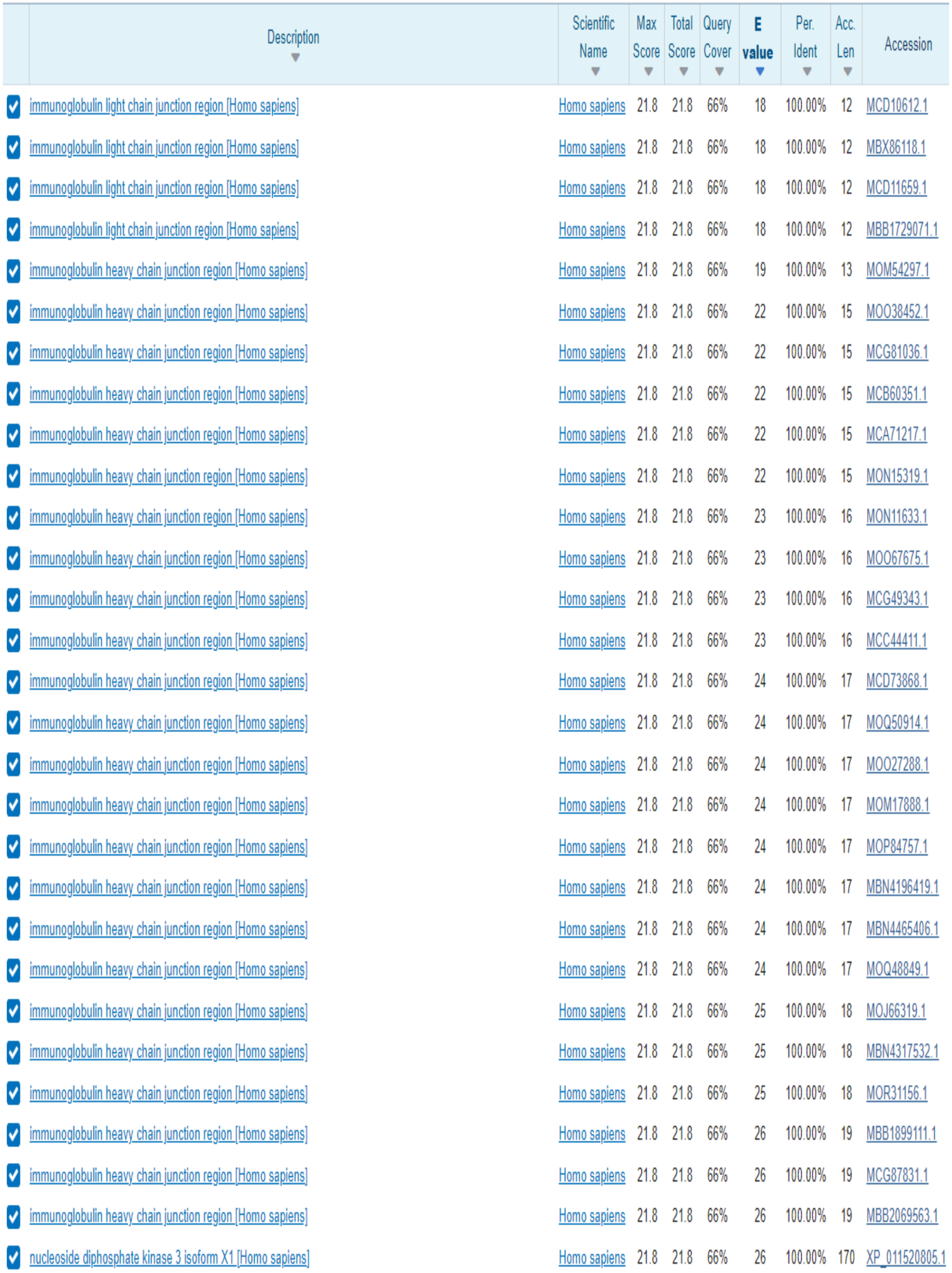
**Query range 1: 1 to 6**

```
Query            1     IKWPWY  6
MCD10612.1       7      ....   10
MBX86118.1       7      ....   10
MCD11659.1       7      ....   10
MBB1729071.1     7      ....   10
MOM54297.1       6      ....   9
MOO38452.1       11     ....   14
MCG81036.1       9      ....   12
MCB60351.1       8      ....   11
MCA71217.1       8      ....   11
MON15319.1       8      ....   11
MON11633.1       9      ....   12
MOO67675.1       9      ....   12
MCG49343.1       8      ....   11
MCC44411.1       8      ....   11
MCD73868.1       10     ....   13
MOQ50914.1       10     ....   13
MOO27288.1       10     ....   13
MOM17888.1       10     ....   13
MOP84757.1       10     ....   13
MBN4196419.1     9      ....   12
MBN4465406.1     9      ....   12
MOQ48849.1       10     ....   13
MOJ66319.1       11     ....   14
MBN4317532.1     11     ....   14
MOR31156.1       8      ....   11
MBB1899111.1     12     ....   15
MCG87831.1       7      ....   10
MBB2069563.1     6      ....   9
XP_011520805.1   95     ....   98
CAB44704.1       145    ....   148
XP_011520806.1   95     ....   98
MOL75431.1       13     ....   16
MOO00050.1       10     ....   13
MBN4591455.1     13     ....   16
MBN4408188.1     13     ....   16
4UU9_A           105    ....   108
CAR62734.1       76     ....   79
CAG29706.1       73     ....   76
MCG44144.1       14     ....   17
MBB1664308.1     12     ....   15
MBB2130208.1     9      ....   12
MCD52980.1       14     ....   17
MOR46235.1       9      ....   12
MOM22354.1       12     ....   15
MBN4478692.1     14     ....   17
NP_001106997.1   259   V....   263

NP_689500.2      184    V....   188
XP_047290858.1   259    V....   263
BAG58595.1       97     V....   101
EAW66688.1       184    V....   188
XP_016879378.1   259    V....   263
EAW66689.1       18     V....   22
MCC66790.1       5      .N...   9
MCH05892.1       6       ....   9
MCH05917.1       6       ....   9
MBY93581.1       6       ....   9
MCE42699.1       6       ....   9
MCA48737.1       5      .N...   9
MCE49560.1       6       ....   9
MCH08192.1       6       ....   9
MCH06006.1       5      .N...   9
MBZ72229.1       6       ....   9
MBB1653918.1     6       ....   9
MCE42643.1       6       ....   9
MCC89288.1       6       ....   9
MCH08207.1       6       ....   9
MBB1729169.1     6       ....   9
MCA98194.1       5      .N...   9
MCE44515.1       6       ....   9
MCA42096.1       6       ....   9
MCH04702.1       6       ....   9
MCC57531.1       6       ....   9
MCE44703.1       6       ....   9
MCD14125.1       6       ....   9
MCG63120.1       5       ....   8
MBB1888791.1     7       ..S..  11
MBN4458568.1     11      ....   14
ABR22603.1       407     ....   410
NP_006062.1      407     ....   410
KAI4003443.1     407     ....   410
KAI2598261.1     405     ....   408
NP_078875.4      828    .N...   832
XP_054173184.1   828    .N...   832
AAH16034.2       828    .N...   832
EAW89401.1       823    .N...   827
BAF84855.1       799    .N...   803
XP_054173192.1   799    .N...   803
AAS77567.1       799    .N...   803
EAW89404.1       799    .N...   803
NP_001005498.2   799    .N...   803
NP_001373117.1   715     ..S..  719
CAA65884.1       715     ..S..  719
NP_001305450.1   701     ..S..  705
BAG64126.1       701     ..S..  705

BAG64126.1      701     ..S..  705
EAX02840.1      691     ..S..  695
AAH35829.1      621    .N...   625
BAB15310.1      591    .N...   595
Q9BYE2.5        336     ....   339
KAI2563030.1    331     ....   334
```

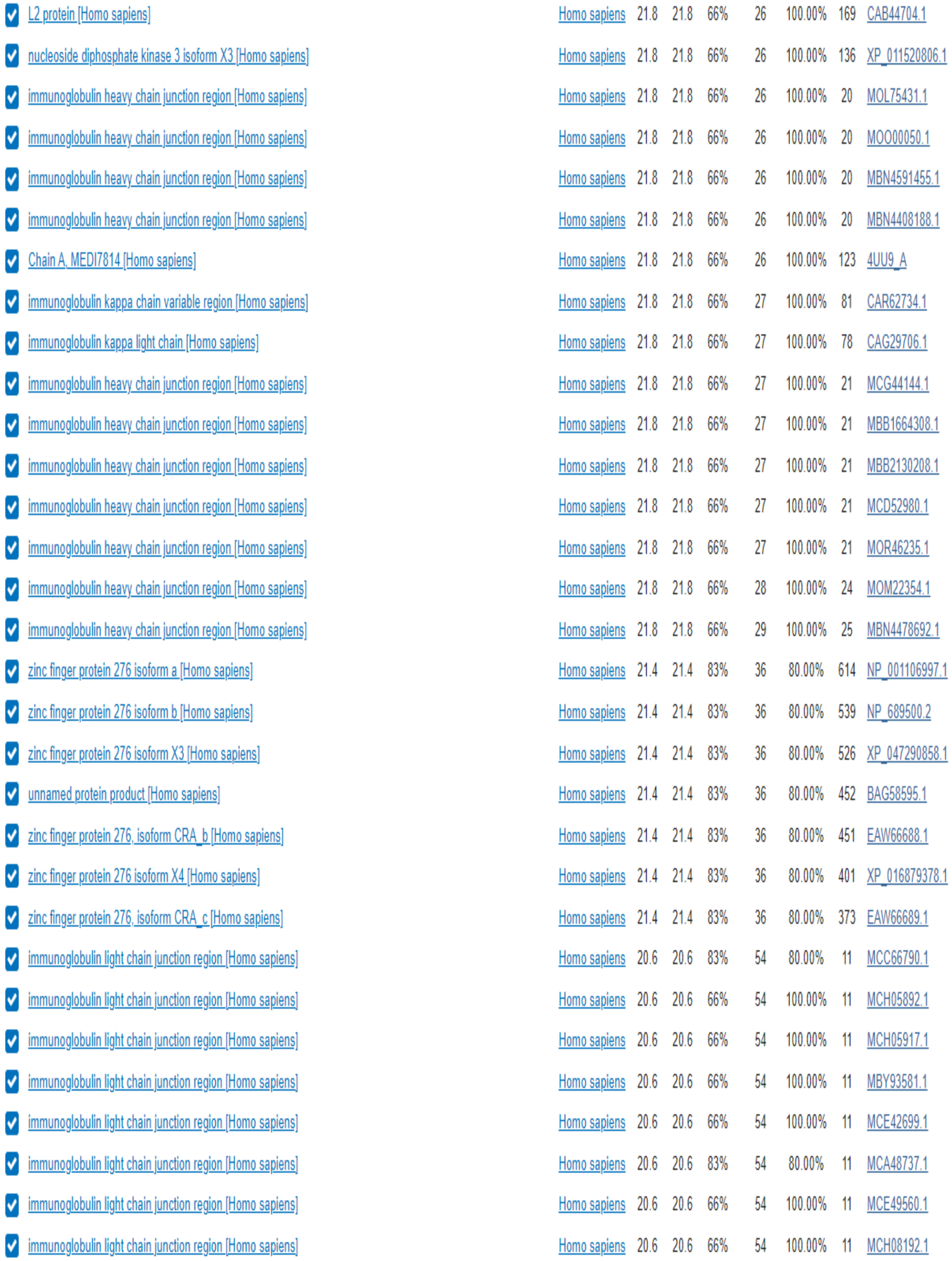
## Appendix 7

## Tabular Description Display of BLASTP of IKWPWY against Homo sapiens nr

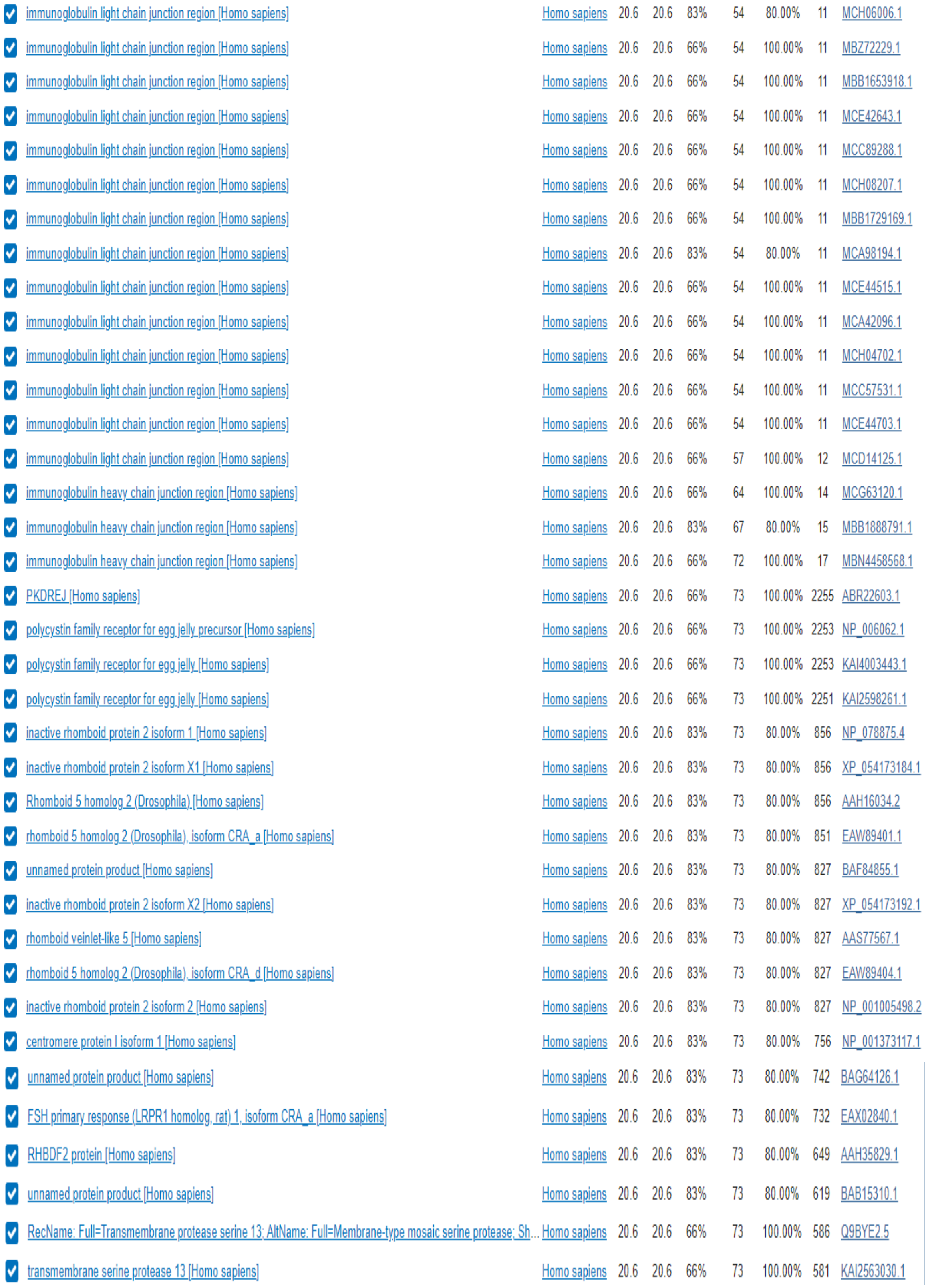
BLASTP description table lists the significant hits along with accession numbers and statistical measures of significance. If BLASTP can align all six amino acids of IKWPWY against a hit, that would be 100% coverage. IKWPWY has maintained the maximum of 83% query coverage when all six residues are identified by BLASTP. This is the only six-residue sequence of the SARS-CoV-2 surface molecules that warrants two mismatches or at least one mismatch and one partial match when aligned with human proteome.



| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 18 | 100.00% | 12 | MCD10612.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 18 | 100.00% | 12 | MBX86118.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 18 | 100.00% | 12 | MCD11659.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 18 | 100.00% | 12 | MBB1729071.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 19 | 100.00% | 13 | MOM54297.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 22 | 100.00% | 15 | MOO38452.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 22 | 100.00% | 15 | MCG81036.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 22 | 100.00% | 15 | MCB60351.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 22 | 100.00% | 15 | MCA71217.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 22 | 100.00% | 15 | MON15319.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 23 | 100.00% | 16 | MON11633.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 23 | 100.00% | 16 | MOO67675.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 23 | 100.00% | 16 | MCG49343.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 23 | 100.00% | 16 | MCC44411.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MCD73868.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MOQ50914.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MOO27288.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MOM17888.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MOP84757.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MBN4196419.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MBN4465406.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 24 | 100.00% | 17 | MOQ48849.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 25 | 100.00% | 18 | MOJ66319.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 25 | 100.00% | 18 | MBN4317532.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 25 | 100.00% | 18 | MOR31156.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 19 | MBB1899111.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 19 | MCG87831.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 19 | MBB2069563.1 |
| ☑ | nucleoside diphosphate kinase 3 isoform X1 [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 170 | XP_011520805.1 |



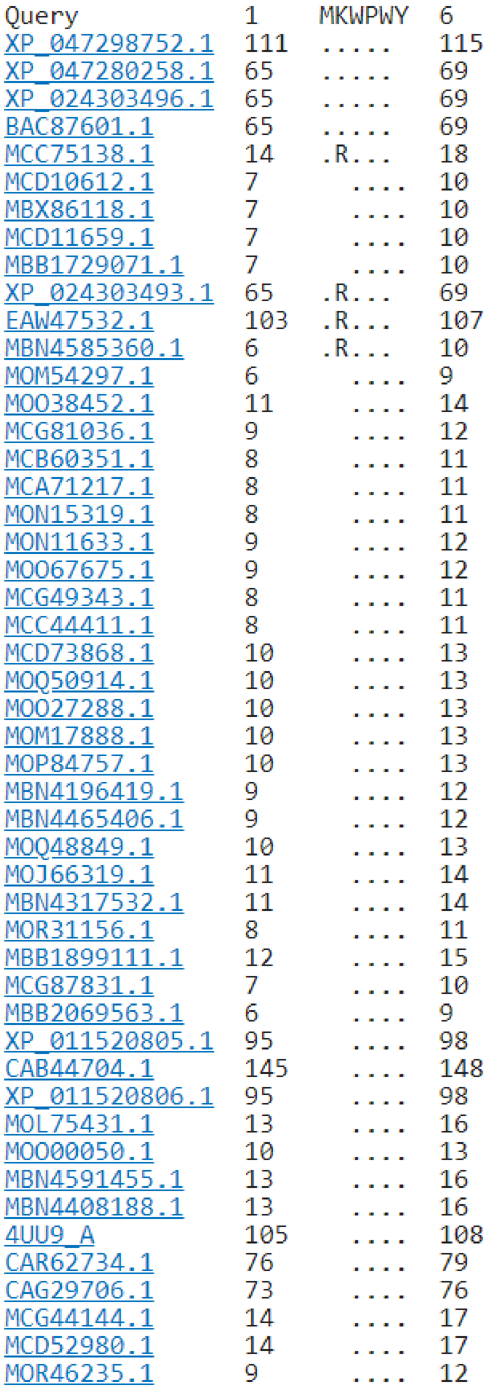
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | L2 protein [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 169 | CAB44704.1 |
| ☑ | nucleoside diphosphate kinase 3 isoform X3 [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 136 | XP_011520806.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 20 | MOL75431.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 20 | MOO00050.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 20 | MBN4591455.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 20 | MBN4408188.1 |
| ☑ | Chain A, MEDI7814 [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 26 | 100.00% | 123 | 4UU9_A |
| ☑ | immunoglobulin kappa chain variable region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 81 | CAR62734.1 |
| ☑ | immunoglobulin kappa light chain [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 78 | CAG29706.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 21 | MCG44144.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 21 | MBB1664308.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 21 | MBB2130208.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 21 | MCD52980.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 27 | 100.00% | 21 | MOR46235.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 28 | 100.00% | 24 | MOM22354.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 21.8 | 21.8 | 66% | 29 | 100.00% | 25 | MBN4478692.1 |
| ☑ | zinc finger protein 276 isoform a [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 614 | NP_001106997.1 |
| ☑ | zinc finger protein 276 isoform b [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 539 | NP_689500.2 |
| ☑ | zinc finger protein 276 isoform X3 [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 526 | XP_047290858.1 |
| ☑ | unnamed protein product [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 452 | BAG58595.1 |
| ☑ | zinc finger protein 276, isoform CRA_b [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 451 | EAW66688.1 |
| ☑ | zinc finger protein 276 isoform X4 [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 401 | XP_016879378.1 |
| ☑ | zinc finger protein 276, isoform CRA_c [Homo sapiens] | Homo sapiens | 21.4 | 21.4 | 83% | 36 | 80.00% | 373 | EAW66689.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 54 | 80.00% | 11 | MCC66790.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCH05892.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCH05917.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MBY93581.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCE42699.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 54 | 80.00% | 11 | MCA48737.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCE49560.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCH08192.1 |



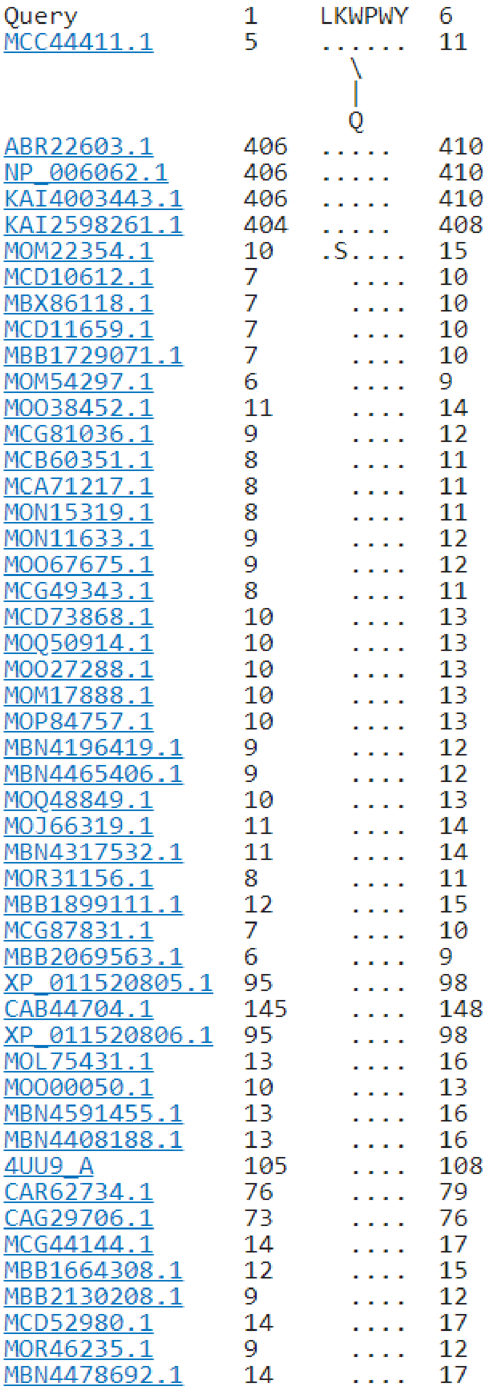
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 54 | 80.00% | 11 | MCH06006.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MBZ72229.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MBB1653918.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCE42643.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCC89288.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCH08207.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MBB1729169.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 54 | 80.00% | 11 | MCA98194.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCE44515.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCA42096.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCH04702.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCC57531.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 54 | 100.00% | 11 | MCE44703.1 |
| ☑ | immunoglobulin light chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 57 | 100.00% | 12 | MCD14125.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 64 | 100.00% | 14 | MCG63120.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 67 | 80.00% | 15 | MBB1888791.1 |
| ☑ | immunoglobulin heavy chain junction region [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 72 | 100.00% | 17 | MBN4458568.1 |
| ☑ | PKDREJ [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 2255 | ABR22603.1 |
| ☑ | polycystin family receptor for egg jelly precursor [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 2253 | NP_006062.1 |
| ☑ | polycystin family receptor for egg jelly [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 2253 | KAI4003443.1 |
| ☑ | polycystin family receptor for egg jelly [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 2251 | KAI2598261.1 |
| ☑ | inactive rhomboid protein 2 isoform 1 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 856 | NP_078875.4 |
| ☑ | inactive rhomboid protein 2 isoform X1 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 856 | XP_054173184.1 |
| ☑ | Rhomboid 5 homolog 2 (Drosophila) [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 856 | AAH16034.2 |
| ☑ | rhomboid 5 homolog 2 (Drosophila), isoform CRA_a [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 851 | EAW89401.1 |
| ☑ | unnamed protein product [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 827 | BAF84855.1 |
| ☑ | inactive rhomboid protein 2 isoform X2 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 827 | XP_054173192.1 |
| ☑ | rhomboid veinlet-like 5 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 827 | AAS77567.1 |
| ☑ | rhomboid 5 homolog 2 (Drosophila), isoform CRA_d [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 827 | EAW89404.1 |
| ☑ | inactive rhomboid protein 2 isoform 2 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 827 | NP_001005498.2 |
| ☑ | centromere protein I isoform 1 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 756 | NP_001373117.1 |
| ☑ | unnamed protein product [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 742 | BAG64126.1 |
| ☑ | FSH primary response (LRPR1 homolog, rat) 1, isoform CRA_a [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 732 | EAX02840.1 |
| ☑ | RHBDF2 protein [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 649 | AAH35829.1 |
| ☑ | unnamed protein product [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 83% | 73 | 80.00% | 619 | BAB15310.1 |
| ☑ | RecName: Full=Transmembrane protease serine 13; AltName: Full=Membrane-type mosaic serine protease; Sh... | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 586 | Q9BYE2.5 |
| ☑ | transmembrane serine protease 13 [Homo sapiens] | Homo sapiens | 20.6 | 20.6 | 66% | 73 | 100.00% | 581 | KAI2563030.1 |

Appendix 8

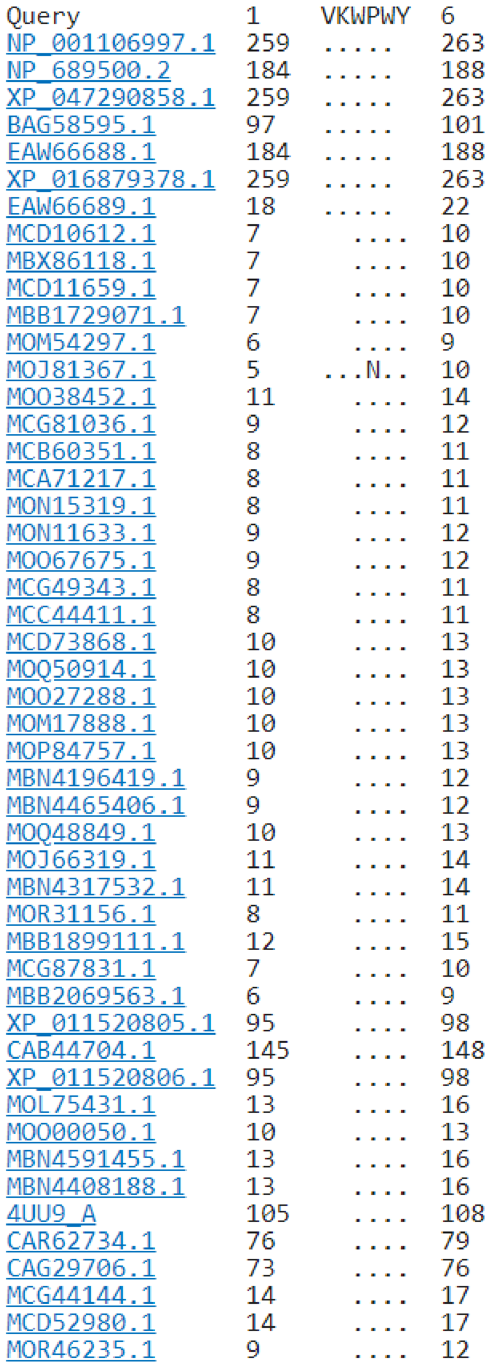
BLASTPs of Residue-substituted IKWPWY



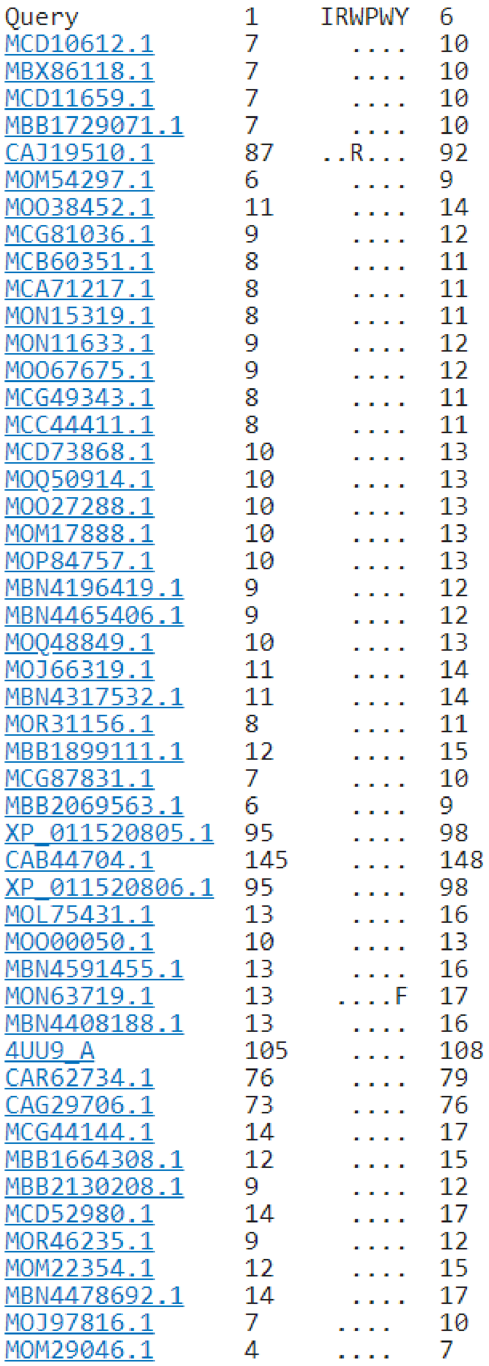
```
Query            1     MKWPWY  6
XP_047298752.1   111   .....   115
XP_047280258.1   65    .....   69
XP_024303496.1   65    .....   69
BAC87601.1       65    .....   69
MCC75138.1       14    .R...   18
MCD10612.1       7       ....  10
MBX86118.1       7       ....  10
MCD11659.1       7       ....  10
MBB1729071.1     7       ....  10
XP_024303493.1   65    .R...   69
EAW47532.1       103   .R...   107
MBN4585360.1     6     .R...   10
MOM54297.1       6       ....  9
MOO38452.1       11      ....  14
MCG81036.1       9       ....  12
MCB60351.1       8       ....  11
MCA71217.1       8       ....  11
MON15319.1       8       ....  11
MON11633.1       9       ....  12
MOO67675.1       9       ....  12
MCG49343.1       8       ....  11
MCC44411.1       8       ....  11
MCD73868.1       10      ....  13
MOQ50914.1       10      ....  13
MOO27288.1       10      ....  13
MOM17888.1       10      ....  13
MOP84757.1       10      ....  13
MBN4196419.1     9       ....  12
MBN4465406.1     9       ....  12
MOQ48849.1       10      ....  13
MOJ66319.1       11      ....  14
MBN4317532.1     11      ....  14
MOR31156.1       8       ....  11
MBB1899111.1     12      ....  15
MCG87831.1       7       ....  10
MBB2069563.1     6       ....  9
XP_011520805.1   95      ....  98
CAB44704.1       145     ....  148
XP_011520806.1   95      ....  98
MOL75431.1       13      ....  16
MOO00050.1       10      ....  13
MBN4591455.1     13      ....  16
MBN4408188.1     13      ....  16
4UU9_A           105     ....  108
CAR62734.1       76      ....  79
CAG29706.1       73      ....  76
MCG44144.1       14      ....  17
MCD52980.1       14      ....  17
MOR46235.1       9       ....  12
```



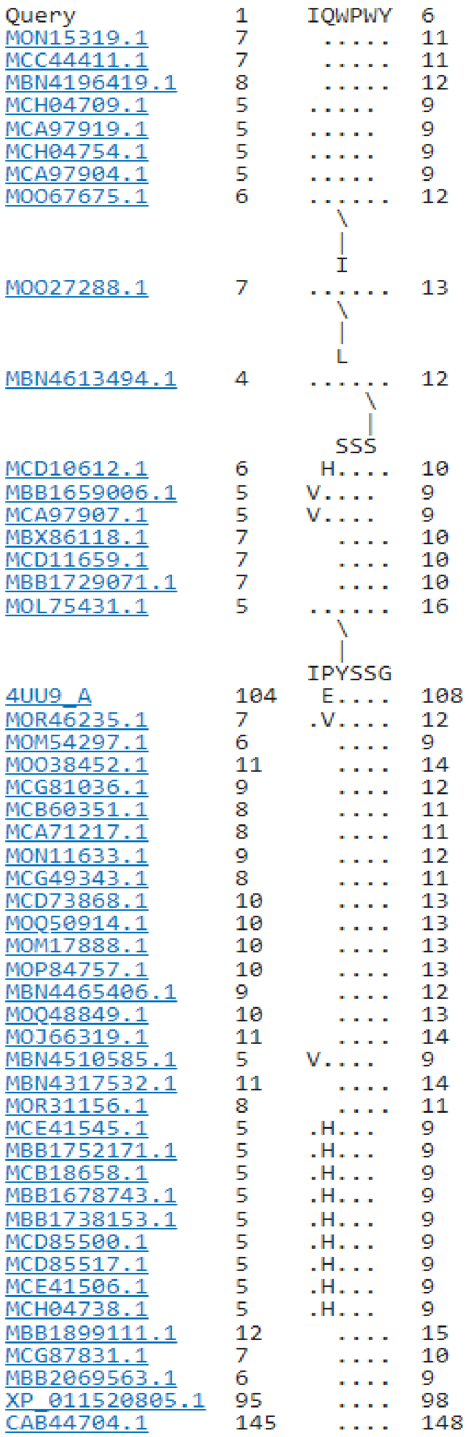
```
Query            1     LKWPWY  6
MCC44411.1       5     ......  11
                         \
                         |
                         Q
ABR22603.1       406   .....   410
NP_006062.1      406   .....   410
KAI4003443.1     406   .....   410
KAI2598261.1     404   .....   408
MOM22354.1       10    .S....  15
MCD10612.1       7       ....  10
MBX86118.1       7       ....  10
MCD11659.1       7       ....  10
MBB1729071.1     7       ....  10
MOM54297.1       6       ....  9
MOO38452.1       11      ....  14
MCG81036.1       9       ....  12
MCB60351.1       8       ....  11
MCA71217.1       8       ....  11
MON15319.1       8       ....  11
MON11633.1       9       ....  12
MOO67675.1       9       ....  12
MCG49343.1       8       ....  11
MCD73868.1       10      ....  13
MOQ50914.1       10      ....  13
MOO27288.1       10      ....  13
MOM17888.1       10      ....  13
MOP84757.1       10      ....  13
MBN4196419.1     9       ....  12
MBN4465406.1     9       ....  12
MOQ48849.1       10      ....  13
MOJ66319.1       11      ....  14
MBN4317532.1     11      ....  14
MOR31156.1       8       ....  11
MBB1899111.1     12      ....  15
MCG87831.1       7       ....  10
MBB2069563.1     6       ....  9
XP_011520805.1   95      ....  98
CAB44704.1       145     ....  148
XP_011520806.1   95      ....  98
MOL75431.1       13      ....  16
MOO00050.1       10      ....  13
MBN4591455.1     13      ....  16
MBN4408188.1     13      ....  16
4UU9_A           105     ....  108
CAR62734.1       76      ....  79
CAG29706.1       73      ....  76
MCG44144.1       14      ....  17
MBB1664308.1     12      ....  15
MBB2130208.1     9       ....  12
MCD52980.1       14      ....  17
MOR46235.1       9       ....  12
MBN4478692.1     14      ....  17
```



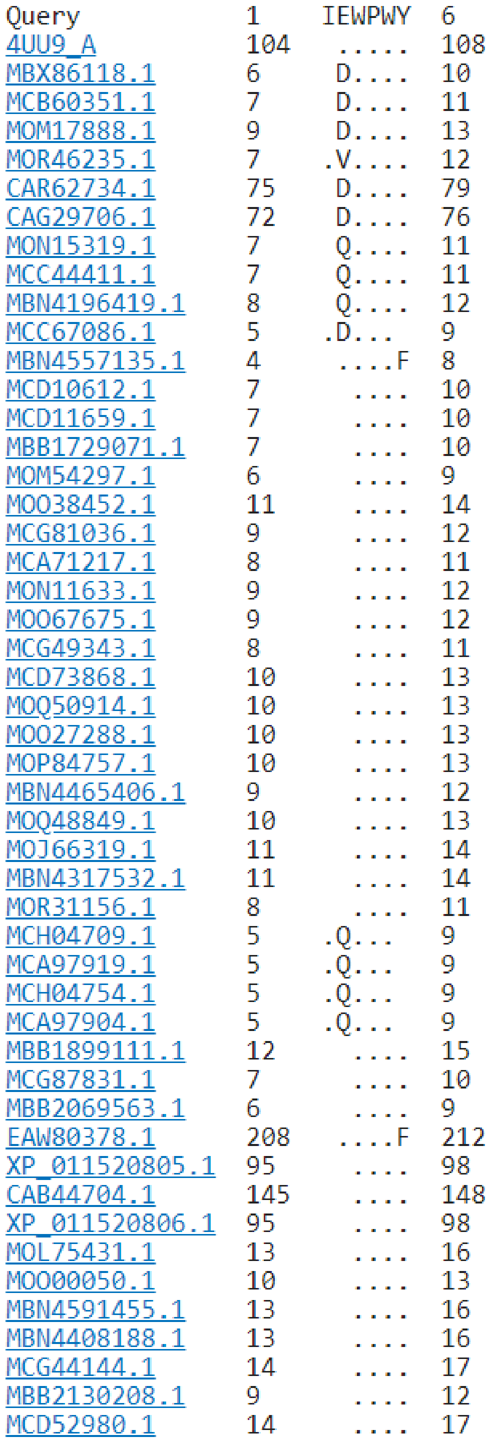
```
Query            1    VKWPWY  6
NP_001106997.1   259  .....   263
NP_689500.2      184  .....   188
XP_047290858.1   259  .....   263
BAG58595.1       97   .....   101
EAW66688.1       184  .....   188
XP_016879378.1   259  .....   263
EAW66689.1       18   .....   22
MCD10612.1       7     ....   10
MBX86118.1       7     ....   10
MCD11659.1       7     ....   10
MBB1729071.1     7     ....   10
MOM54297.1       6     ....   9
MOJ81367.1       5    ...N..  10
MOO38452.1       11    ....   14
MCG81036.1       9     ....   12
MCB60351.1       8     ....   11
MCA71217.1       8     ....   11
MON15319.1       8     ....   11
MON11633.1       9     ....   12
MOO67675.1       9     ....   12
MCG49343.1       8     ....   11
MCC44411.1       8     ....   11
MCD73868.1       10    ....   13
MOQ50914.1       10    ....   13
MOO27288.1       10    ....   13
MOM17888.1       10    ....   13
MOP84757.1       10    ....   13
MBN4196419.1     9     ....   12
MBN4465406.1     9     ....   12
MOQ48849.1       10    ....   13
MOJ66319.1       11    ....   14
MBN4317532.1     11    ....   14
MOR31156.1       8     ....   11
MBB1899111.1     12    ....   15
MCG87831.1       7     ....   10
MBB2069563.1     6     ....   9
XP_011520805.1   95    ....   98
CAB44704.1       145   ....   148
XP_011520806.1   95    ....   98
MOL75431.1       13    ....   16
MOO00050.1       10    ....   13
MBN4591455.1     13    ....   16
MBN4408188.1     13    ....   16
4UU9_A           105   ....   108
CAR62734.1       76    ....   79
CAG29706.1       73    ....   76
MCG44144.1       14    ....   17
MCD52980.1       14    ....   17
MOR46235.1       9     ....   12
```



```
Query            1    IRWPWY  6
MCD10612.1       7      ....  10
MBX86118.1       7      ....  10
MCD11659.1       7      ....  10
MBB1729071.1     7      ....  10
CAJ19510.1       87   ..R...  92
MOM54297.1       6      ....  9
MOO38452.1       11     ....  14
MCG81036.1       9      ....  12
MCB60351.1       8      ....  11
MCA71217.1       8      ....  11
MON15319.1       8      ....  11
MON11633.1       9      ....  12
MOO67675.1       9      ....  12
MCG49343.1       8      ....  11
MCC44411.1       8      ....  11
MCD73868.1       10     ....  13
MOQ50914.1       10     ....  13
MOO27288.1       10     ....  13
MOM17888.1       10     ....  13
MOP84757.1       10     ....  13
MBN4196419.1     9      ....  12
MBN4465406.1     9      ....  12
MOQ48849.1       10     ....  13
MOJ66319.1       11     ....  14
MBN4317532.1     11     ....  14
MOR31156.1       8      ....  11
MBB1899111.1     12     ....  15
MCG87831.1       7      ....  10
MBB2069563.1     6      ....  9
XP_011520805.1   95     ....  98
CAB44704.1       145    ....  148
XP_011520806.1   95     ....  98
MOL75431.1       13     ....  16
MOO00050.1       10     ....  13
MBN4591455.1     13     ....  16
MON63719.1       13    ....F  17
MBN4408188.1     13     ....  16
4UU9_A           105    ....  108
CAR62734.1       76     ....  79
CAG29706.1       73     ....  76
MCG44144.1       14     ....  17
MBB1664308.1     12     ....  15
MBB2130208.1     9      ....  12
MCD52980.1       14     ....  17
MOR46235.1       9      ....  12
MOM22354.1       12     ....  15
MBN4478692.1     14     ....  17
MOJ97816.1       7     ....   10
MOM29046.1       4     ....   7
```



```
Query           1    IQWPWY  6
MON15319.1      7     .....  11
MCC44411.1      7     .....  11
MBN4196419.1    8     .....  12
MCH04709.1      5    .....   9
MCA97919.1      5    .....   9
MCH04754.1      5    .....   9
MCA97904.1      5    .....   9
MOO67675.1      6    ......  12
                       \
                       |
                       I
MOO27288.1      7    ......  13
                       \
                       |
                       L
MBN4613494.1    4    ......  12
                         \
                         |
                       SSS
MCD10612.1      6     H....  10
MBB1659006.1    5    V....   9
MCA97907.1      5    V....   9
MBX86118.1      7      ....  10
MCD11659.1      7      ....  10
MBB1729071.1    7      ....  10
MOL75431.1      5    ......  16
                       \
                       |
                     IPYSSG
4UU9_A          104   E....  108
MOR46235.1      7    .V....  12
MOM54297.1      6      ....  9
MOO38452.1      11     ....  14
MCG81036.1      9      ....  12
MCB60351.1      8      ....  11
MCA71217.1      8      ....  11
MON11633.1      9      ....  12
MCG49343.1      8      ....  11
MCD73868.1      10     ....  13
MOQ50914.1      10     ....  13
MOM17888.1      10     ....  13
MOP84757.1      10     ....  13
MBN4465406.1    9      ....  12
MOQ48849.1      10     ....  13
MOJ66319.1      11     ....  14
MBN4510585.1    5    V....   9
MBN4317532.1    11     ....  14
MOR31156.1      8      ....  11
MCE41545.1      5    .H...   9
MBB1752171.1    5    .H...   9
MCB18658.1      5    .H...   9
MBB1678743.1    5    .H...   9
MBB1738153.1    5    .H...   9
MCD85500.1      5    .H...   9
MCD85517.1      5    .H...   9
MCE41506.1      5    .H...   9
MCH04738.1      5    .H...   9
MBB1899111.1    12     ....  15
MCG87831.1      7      ....  10
MBB2069563.1    6      ....  9
XP_011520805.1  95     ....  98
CAB44704.1      145    ....  148
```



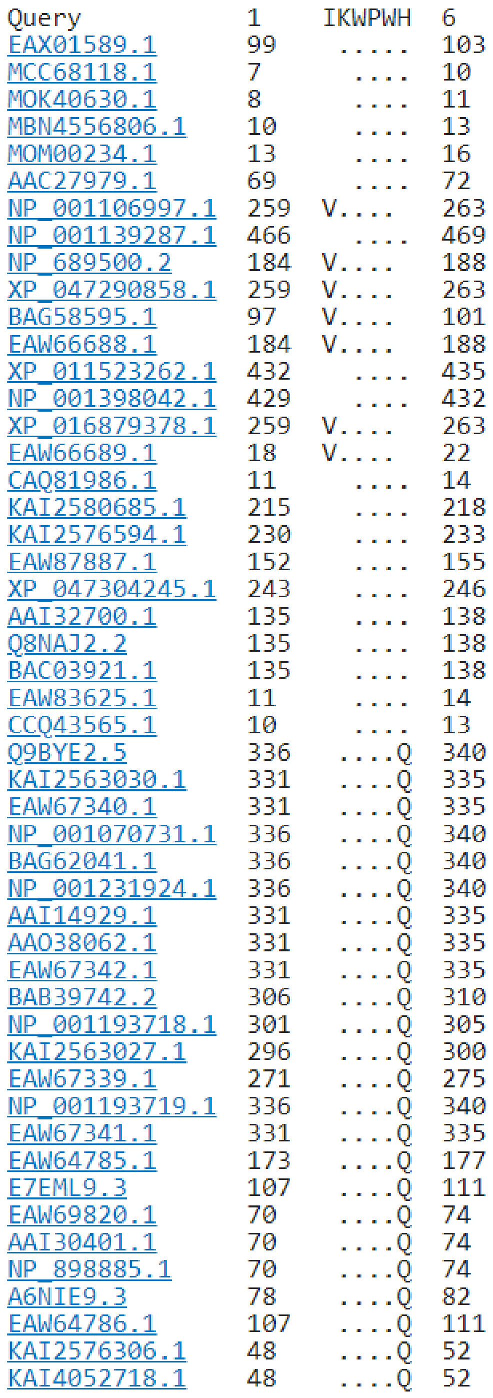
```
Query           1    IEWPWY  6
4UU9_A          104   .....  108
MBX86118.1      6     D....  10
MCB60351.1      7     D....  11
MOM17888.1      9     D....  13
MOR46235.1      7    .V....  12
CAR62734.1      75    D....  79
CAG29706.1      72    D....  76
MON15319.1      7     Q....  11
MCC44411.1      7     Q....  11
MBN4196419.1    8     Q....  12
MCC67086.1      5    .D...   9
MBN4557135.1    4     ....F  8
MCD10612.1      7      ....  10
MCD11659.1      7      ....  10
MBB1729071.1    7      ....  10
MOM54297.1      6      ....  9
MOO38452.1      11     ....  14
MCG81036.1      9      ....  12
MCA71217.1      8      ....  11
MON11633.1      9      ....  12
MOO67675.1      9      ....  12
MCG49343.1      8      ....  11
MCD73868.1      10     ....  13
MOQ50914.1      10     ....  13
MOO27288.1      10     ....  13
MOP84757.1      10     ....  13
MBN4465406.1    9      ....  12
MOQ48849.1      10     ....  13
MOJ66319.1      11     ....  14
MBN4317532.1    11     ....  14
MOR31156.1      8      ....  11
MCH04709.1      5    .Q...   9
MCA97919.1      5    .Q...   9
MCH04754.1      5    .Q...   9
MCA97904.1      5    .Q...   9
MBB1899111.1    12     ....  15
MCG87831.1      7      ....  10
MBB2069563.1    6      ....  9
EAW80378.1      208   ....F  212
XP_011520805.1  95     ....  98
CAB44704.1      145    ....  148
XP_011520806.1  95     ....  98
MOL75431.1      13     ....  16
MOO00050.1      10     ....  13
MBN4591455.1    13     ....  16
MBN4408188.1    13     ....  16
MCG44144.1      14     ....  17
MBB2130208.1    9      ....  12
MCD52980.1      14     ....  17
```

```
Query           1    IKYPWY  6    Query           1    IKFPWY  6
NP_001386388.1  130  VN....  135  NP_057086.2     310  .....F  315
NP_001387512.1  127  VN....  132  AAH26185.1      310  .....F  315
AAH49388.1      110  VN....  115  AAD34044.1      309  .....F  314
NP_001387513.1  107  VN....  112  NP_079413.3     655  .....   659
KAI2554336.1    99   VN....  104  AAI53880.1      655  .....   659
NP_001387514.1  82   VN....  87   BAH58760.1      655  .....   659
NP_001387511.1  81   VN....  86   NP_001398061.1  655  .....   659
AAH29924.2      48   VN....  53   KAI2573960.1    655  .....   659
EAW52858.1      39   ..C...  44   NP_001153699.1  655  .....   659
BAB15724.1      42   VN....  47   AAI50641.1      655  .....   659
MCB74633.1      7      ....  10   KAI4057566.1    655  .....   659
MCB74390.1      7      ....  10   KAI2573962.1    655  .....   659
MOJ89056.1      9      ....  12   XP_047289100.1  454  .....   458
MOR15618.1      6      ....  9    KAI2573964.1    655  .....   659
MBX76455.1      6      ....  9    KAI4057568.1    655  .....   659
MBN4395206.1    6      ....  9    EAW77272.1      655  .....   659
MCB59478.1      7      ....  10   XP_016878124.1  655  .....   659
MBN4603152.1    11     ....  14   XP_006720764.1  655  .....   659
MBN4603712.1    9      ....  12   XP_047289101.1  655  .....   659
MOK12147.1      9      ....  12   CAH10686.1      572  .....   576
MCC48956.1      9      ....  12   KAI4057572.1    655  .....   659
MBB1967028.1    9      ....  12   KAI2573968.1    655  .....   659
MBN4463881.1    9      ....  12   EAX03332.1      112  .....   116
MOQ53273.1      9      ....  12   BAC05215.1      112  .....   116
MCG31942.1      9      ....  12   MBN4386014.1    13   .T....  18
MBB2014497.1    9      ....  12   EAW68810.1      3    ...E..  8
MOM96109.1      9      ....  12   MOL91922.1      7    .G....  12
MOK42276.1      9      ....  12   MBN4367036.1    7      ....  10
MBB1944464.1    9      ....  12   MON85597.1      7      ....  10
MBN4260109.1    9      ....  12   MOL70367.1      7      ....  10
MOQ39959.1      9      ....  12   MCB92776.1      7      ....  10
MBB1695933.1    10     ....  13   MBN4387292.1    8      ....  11
MOP14825.1      9      ....  12   MBB1965638.1    9      ....  12
MOQ42815.1      10     ....  13   MBN4487645.1    9      ....  12
MCA63677.1      10     ....  13   MBN4206519.1    9      ....  12
MBN4346513.1    8      ....  11   MOM03516.1      9      ....  12
MBB2075739.1    10     ....  13   MBB1705963.1    6      ....  9
MBN4386820.1    10     ....  13   MBN4487644.1    9      ....  12
MCC40383.1      10     ....  13   KAI2578896.1    1483   ....  1486
MBN4250717.1    10     ....  13   AAH40523.1      1483   ....  1486
MBN4334666.1    9      ....  12   NP_996882.1     1483   ....  1486
MOL17051.1      7      ....  10   XP_054231829.1  717    ....  720
MCB67554.1      10     ....  13   EAW81212.1      717    ....  720
MBB1657165.1    13     ....  16   NP_001035197.1  717    ....  720
XP_047275450.1  32     ....  35   EAW81211.1      717    ....  720
NP_001337461.1  171    ....  174  NP_055196.2     717    ....  720
AAH68589.1      150    ....  153  XP_054231830.1  717    ....  720
AAH95419.1      150    ....  153  AAF23904.1      717    ....  720
NP_722523.1     150    ....  153  AAF23905.1      717    ....  720
NP_001337462.1  149    ....  152  XP_054231831.1  717    ....  720
```

```
Query           1    IKWPYY  6    Query           1    IKWPFY  6
MOK27403.1      9    .R....  14   MBB1669096.1    6     .....  10
MBN4470190.1    5     .....  9    MOQ31127.1      10    .....  14
MBN4470187.1    5     .....  9    MBB2077643.1    11   ...T..  16
MOQ53641.1      3     .....  7    MOK27403.1      9    .R..Y.  14
MBN4632641.1    8     .....  12   MBB1700426.1    7      ....  10
MCG46075.1      7    .N....  12   MCA38891.1      7      ....  10
MOP17821.1      8     .....  12   MBB1720344.1    7      ....  10
ANT83437.1      98    .....  102  MCE41972.1      7      ....  10
MOO35630.1      5     .....  9    MOP26219.1      8      ....  11
MBN4606059.1    22    .....  26   MBB1683199.1    7      ....  10
MCC68048.1      5    .....   9    MCA98894.1      7      ....  10
MBB1664993.1    9    .T....  14   MBN4470190.1    5     ...Y.  9
MBZ57577.1      7    .S....  12   MCD10694.1      7      ....  10
MBN4625487.1    9    .S....  14   MOQ83110.1      6      ....  9
MBN4190070.1    4    ...E..  9    MBN4470187.1    5     ...Y.  9
MOO09741.1      4    VR....  9    MOK54079.1      10     ....  13
MBN4463232.1    6    .P....  11   MCG59213.1      10     ....  13
QEP20047.1      101  VR....  106  MOM63246.1      7      ....  10
MBB1942701.1    11   .I....  16   MCG18092.1      8      ....  11
MOQ61661.1      4    .P....  9    AAO22237.1      374    ....  377
MBB2048377.1    14   ...E..  19   NP_001333799.1  169    ....  172
MOK42213.1      8    .G....  13   KAI2537826.1    169    ....  172
MOL69998.1      7    ...G..  12   XP_047299234.1  169    ....  172
MBB1737947.1    7      ....  10   XP_005267462.1  377    ....  380
MOO31001.1      8      ....  11   XP_054231544.1  377    ....  380
MOM79833.1      5      ....  8    NP_001366220.1  374    ....  377
MOO10905.1      5      ....  8    XP_005267463.1  377    ....  380
MBB1704280.1    6      ....  9    XP_054231545.1  377    ....  380
MBB1743434.1    6      ....  9    XP_047299235.1  169    ....  172
MOM80055.1      9      ....  12   KAI2537825.1    169    ....  172
MOM97106.1      6      ....  9    XP_047299236.1  169    ....  172
MBB1707334.1    6      ....  9    XP_047294497.1  278    ....  281
MOK24293.1      6      ....  9    AAQ16198.1      306    ....  309
MBN4400616.1    6      ....  9    NP_001229735.2  297    ....  300
MBN4359714.1    6      ....  9    KAI4081437.1    297    ....  300
MBN4410860.1    6      ....  9    KAI2517882.1    297    ....  300
MBB1661912.1    6      ....  9    BAH13876.1      297    ....  300
MOO08866.1      6      ....  9    KAI2517883.1    293    ....  296
MBX78272.1      6      ....  9    AAB87862.1      293    ....  296
MCG22469.1      7      ....  10   NP_001229734.2  293    ....  296
MBN4374350.1    7      ....  10   XP_054194452.1  293    ....  296
MBN4383619.1    7      ....  10   KAI2517885.1    247    ....  250
MOR37977.1      7      ....  10   NP_001229736.2  247    ....  250
MOK30773.1      7      ....  10   KAI4081436.1    247    ....  250
MOK51571.1      7      ....  10   BAH13797.1      247    ....  250
MOP37685.1      10     ....  13   KAI2517894.1    220    ....  223
MOO61658.1      7      ....  10   NP_001229739.2  220    ....  223
MOP50547.1      6      ....  9    EAW73109.1      220    ....  223
MOQ22829.1      7      ....  10   KAI4017837.1    370    ....  373
MOM22489.1      7      ....  10
```

```
Query           1    IKWPWF  6
MBB1696440.1    4    ......  11
                         \
                         |
                        GD
MCD13305.1      7      ....  10
MBN4557135.1    5      ....  8
MCD12083.1      7      ....  10
MOR88251.1      8      ....  11
MBB2045168.1    8      ....  11
MBN4467872.1    11     ....  14
MOQ44538.1      11     ....  14
NP_001106997.1  259  V....   263
NP_689500.2     184  V....   188
XP_047290858.1  259  V....   263
BAG58595.1      97   V....   101
EAW66688.1      184  V....   188
XP_016879378.1  259  V....   263
EAW76639.1      371    ....  374
EAW66689.1      18   V....   22
EAW80378.1      209    ....  212
CAD97673.1      227    ....  230
4YQM_A          181    ....  184
NP_899062.1     179    ....  182
AAP47743.1      179    ....  182
EAW49599.1      179    ....  182
5UEH_A          179    ....  182
3VLN_A          178    ....  181
BAG36430.1      178    ....  181
NP_004823.1     178    ....  181
AAO23573.1      177    ....  180
4IS0_A          178    ....  181
6PNM_A          177    ....  180
3LFL_A          177    ....  180
3Q18_A          179    ....  182
EAX05304.1      217    ....  220
NP_001177943.1  151    ....  154
KAI4077377.1    151    ....  154
NP_001177932.1  150    ....  153
NP_001177942.1  145    ....  148
NP_001177931.1  145    ....  148
KAI2557190.1    150    ....  153
NP_001177944.1  117    ....  120
KAI2557189.1    117    ....  120
ABV49422.1      115    ....  118
ABS19011.1      115    ....  118
KAI2576387.1    82     ....  85
MON63719.1      14     ....  17
MCG57608.1      14     ....  17
MCG83042.1      14     ....  17
MBN4520411.1    19     ....  22
AAD14258.1      3      ....  6
MCH08192.1      6     ....   9
MCH06006.1      5    .N...   9
```

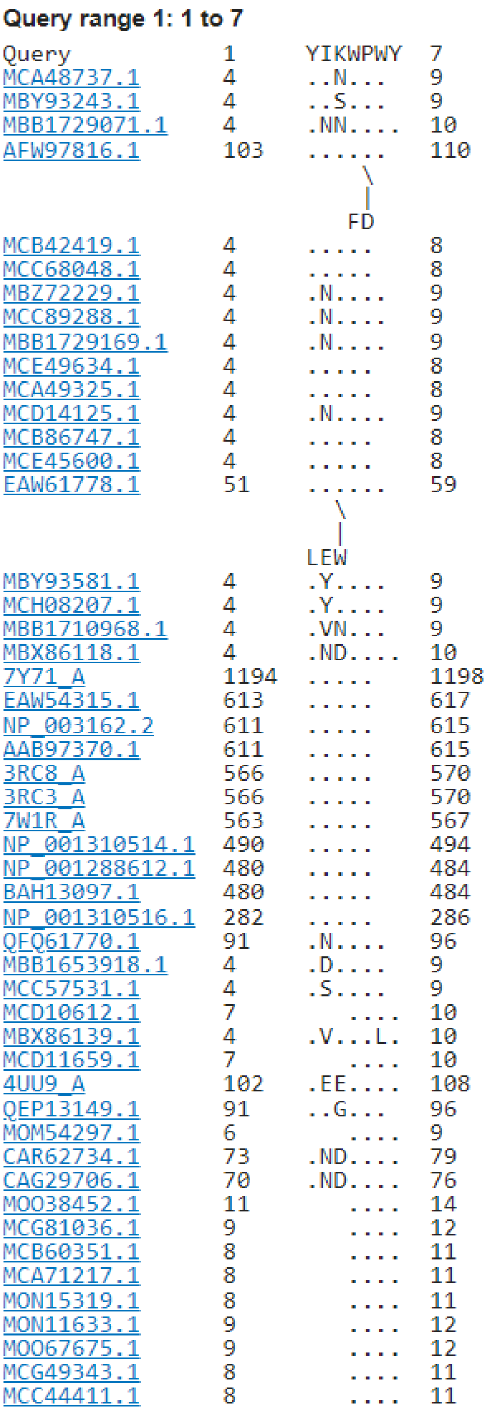
```
Query           2     KWPWW   6
EAW61778.1      56    .....   60
MCB39795.1      7      ....   10
MOQ77216.1      9      ....   12
MCB75966.1      7      ....   10
MBX85160.1      7      ....   10
MCE41663.1      7      ....   10
MBZ69186.1      7      ....   10
MBB1654651.1    7      ....   10
MCC48117.1      13     ....   16
MBY89686.1      10     ....   13
XP_054213301.1  4820   ....   4823
XP_047275875.1  4820   ....   4823
XP_054182016.1  1425   ....   1428
XP_011528762.1  1425   ....   1428
XP_011528760.1  1425   ....   1428
XP_054182012.1  1425   ....   1428
XP_054182018.1  1425   ....   1428
XP_016884503.1  1425   ....   1428
XP_054182017.1  1424   ....   1427
XP_011528766.1  1424   ....   1427
XP_054182019.1  1424   ....   1427
XP_047297507.1  1424   ....   1427
KAI2597058.1    1383   ....   1386
KAI4002244.1    1383   ....   1386
NP_001305174.1  1383   ....   1386
AAL75811.1      1383   ....   1386
KAI2597059.1    1382   ....   1385
BAC16363.1      1382   ....   1385
CAC70712.2      1382   ....   1385
NP_115997.5     1382   ....   1385
CAC70714.3      1382   ....   1385
KAI4002242.1    1382   ....   1385
XP_016884504.1  1382   ....   1385
AAI44598.1      1382   ....   1385
XP_054182020.1  1382   ....   1385
EAW59707.1      1263   ....   1266
XP_054182021.1  1264   ....   1267
EAW59703.1      1264   ....   1267
XP_011528767.1  1264   ....   1267
EAW59706.1      1263   ....   1266
EAW59702.1      1263   ....   1266
XP_011528768.1  1425   ....   1428
XP_054182022.1  1425   ....   1428
EAW59704.1      895    ....   898
BAB55550.2      895    ....   898
NP_001333694.1  1253   ....   1256
EAW51184.1      1237   ....   1240
NP_510880.2     1234   ....   1237
KAI2582130.1    1234   ....   1237
NP_001333695.1  1246   ....   1249
```



```
Query           1    IKWPWH  6
EAX01589.1      99    .....  103
MCC68118.1      7      ....  10
MOK40630.1      8      ....  11
MBN4556806.1    10     ....  13
MOM00234.1      13     ....  16
AAC27979.1      69     ....  72
NP_001106997.1  259  V....   263
NP_001139287.1  466    ....  469
NP_689500.2     184  V....   188
XP_047290858.1  259  V....   263
BAG58595.1      97   V....   101
EAW66688.1      184  V....   188
XP_011523262.1  432    ....  435
NP_001398042.1  429    ....  432
XP_016879378.1  259  V....   263
EAW66689.1      18   V....   22
CAQ81986.1      11     ....  14
KAI2580685.1    215    ....  218
KAI2576594.1    230    ....  233
EAW87887.1      152    ....  155
XP_047304245.1  243    ....  246
AAI32700.1      135    ....  138
Q8NAJ2.2        135    ....  138
BAC03921.1      135    ....  138
EAW83625.1      11     ....  14
CCQ43565.1      10     ....  13
Q9BYE2.5        336   ....Q  340
KAI2563030.1    331   ....Q  335
EAW67340.1      331   ....Q  335
NP_001070731.1  336   ....Q  340
BAG62041.1      336   ....Q  340
NP_001231924.1  336   ....Q  340
AAI14929.1      331   ....Q  335
AAO38062.1      331   ....Q  335
EAW67342.1      331   ....Q  335
BAB39742.2      306   ....Q  310
NP_001193718.1  301   ....Q  305
KAI2563027.1    296   ....Q  300
EAW67339.1      271   ....Q  275
NP_001193719.1  336   ....Q  340
EAW67341.1      331   ....Q  335
EAW64785.1      173   ....Q  177
E7EML9.3        107   ....Q  111
EAW69820.1      70    ....Q  74
AAI30401.1      70    ....Q  74
NP_898885.1     70    ....Q  74
A6NIE9.3        78    ....Q  82
EAW64786.1      107   ....Q  111
KAI2576306.1    48    ....Q  52
KAI4052718.1    48    ....Q  52
```

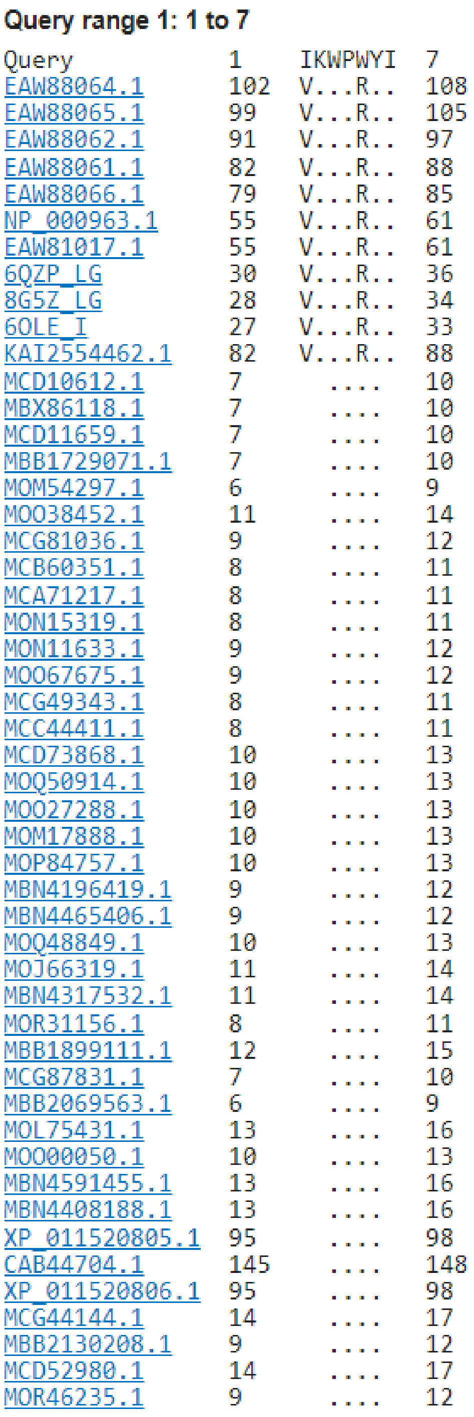
## Appendix 9

### BLASTPs of Residue-added IKWPWY



**Query range 1: 1 to 7**

```
Query             1      YIKWPWY  7
MCA48737.1        4      ..N...   9
MBY93243.1        4      ..S...   9
MBB1729071.1      4      .NN....  10
AFW97816.1        103    ......   110
                             \
                             |
                            FD
MCB42419.1        4      .....    8
MCC68048.1        4      .....    8
MBZ72229.1        4      .N....   9
MCC89288.1        4      .N....   9
MBB1729169.1      4      .N....   9
MCE49634.1        4      .....    8
MCA49325.1        4      .....    8
MCD14125.1        4      .N....   9
MCB86747.1        4      .....    8
MCE45600.1        4      .....    8
EAW61778.1        51     ......   59
                           \
                           |
                         LEW
MBY93581.1        4      .Y....   9
MCH08207.1        4      .Y....   9
MBB1710968.1      4      .VN...   9
MBX86118.1        4      .ND....  10
7Y71_A            1194   .....    1198
EAW54315.1        613    .....    617
NP_003162.2       611    .....    615
AAB97370.1        611    .....    615
3RC8_A            566    .....    570
3RC3_A            566    .....    570
7W1R_A            563    .....    567
NP_001310514.1    490    .....    494
NP_001288612.1    480    .....    484
BAH13097.1        480    .....    484
NP_001310516.1    282    .....    286
QFQ61770.1        91     .N....   96
MBB1653918.1      4      .D....   9
MCC57531.1        4      .S....   9
MCD10612.1        7         ....  10
MBX86139.1        4      .V...L.  10
MCD11659.1        7         ....  10
4UU9_A            102    .EE....  108
QEP13149.1        91     ..G...   96
MOM54297.1        6         ....  9
CAR62734.1        73     .ND....  79
CAG29706.1        70     .ND....  76
MOO38452.1        11        ....  14
MCG81036.1        9         ....  12
MCB60351.1        8         ....  11
MCA71217.1        8         ....  11
MON15319.1        8         ....  11
MON11633.1        9         ....  12
MOO67675.1        9         ....  12
MCG49343.1        8         ....  11
MCC44411.1        8         ....  11
```



**Query range 1: 1 to 7**

```
Query             1      IKWPWYI  7
EAW88064.1        102    V...R..  108
EAW88065.1        99     V...R..  105
EAW88062.1        91     V...R..  97
EAW88061.1        82     V...R..  88
EAW88066.1        79     V...R..  85
NP_000963.1       55     V...R..  61
EAW81017.1        55     V...R..  61
6QZP_LG           30     V...R..  36
8G5Z_LG           28     V...R..  34
6OLE_I            27     V...R..  33
KAI2554462.1      82     V...R..  88
MCD10612.1        7        ....   10
MBX86118.1        7        ....   10
MCD11659.1        7        ....   10
MBB1729071.1      7        ....   10
MOM54297.1        6        ....   9
MOO38452.1        11       ....   14
MCG81036.1        9        ....   12
MCB60351.1        8        ....   11
MCA71217.1        8        ....   11
MON15319.1        8        ....   11
MON11633.1        9        ....   12
MOO67675.1        9        ....   12
MCG49343.1        8        ....   11
MCC44411.1        8        ....   11
MCD73868.1        10       ....   13
MOQ50914.1        10       ....   13
MOO27288.1        10       ....   13
MOM17888.1        10       ....   13
MOP84757.1        10       ....   13
MBN4196419.1      9        ....   12
MBN4465406.1      9        ....   12
MOQ48849.1        10       ....   13
MOJ66319.1        11       ....   14
MBN4317532.1      11       ....   14
MOR31156.1        8        ....   11
MBB1899111.1      12       ....   15
MCG87831.1        7        ....   10
MBB2069563.1      6        ....   9
MOL75431.1        13       ....   16
MOO00050.1        10       ....   13
MBN4591455.1      13       ....   16
MBN4408188.1      13       ....   16
XP_011520805.1    95       ....   98
CAB44704.1        145      ....   148
XP_011520806.1    95       ....   98
MCG44144.1        14       ....   17
MBB2130208.1      9        ....   12
MCD52980.1        14       ....   17
MOR46235.1        9        ....   12
```

# References

Akbar, R., Robert, P. A., Pavlović, M., Jeliazkov, J. R., Snapkov, I., Slabodkin, A., Weber, C. R., Scheffer, L., Miho, E., Haff, I. H., Haug, D. T., Lund-Johansen, F., Safonova, Y., Sandve, G. K., & Greiff, V. (2021). A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, *34*(11), 108856. https://doi.org/10.1016/j.celrep.2021.108856

An introduction to antibodies: Antigens, epitopes and antibodies. (n.d.). https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/elisa/antigens-epitopes-antibodies

Atassi, M. Z. (1975). Antigenic structure of myoglobin: The complete immunochemical anatomy of a protein and conclusions relating to antigenic structures of proteins. *Immunochemistry*, *12*(5), 423–438. https://doi.org/10.1016/0019-2791(75)90010-5

Avril, T. (2020, March 19). *Why the coronavirus and most other viruses have no cure*. Medical Xpress - medical research advances and health news. https://medicalxpress.com/news/2020-03-coronavirus-viruses.html

Centers for Disease Control and Prevention. (n.d.). *Human coronavirus types*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/types.html

Qi, N., Liu, C., Yang, H., Shi, W., Wang, S., Zhou, Y., Wei, C., Gu, F., & Qin, Y. (2017). Therapeutic hexapeptide (PGPIPN) prevents and cures alcoholic fatty liver disease by affecting the expressions of genes related with lipid metabolism and oxidative stress. *Oncotarget*, *8*(50), 88079–88093. https://doi.org/10.18632/oncotarget.21404

*Tools for covid-19 research*. Novus Biologicals. (n.d.). https://www.novusbio.com/support/sars-cov-research-resources