

Project Number: MQP-DK2-AAQM

MOLECULAR EVOLUTION OF EBOLA: A BIOINFORMATICS APPROACH

A Major Qualifying Project Report  
Submitted to the Faculty  
Of  
Worcester Polytechnic Institute  
in partial fulfillment of the requirements for the Bachelor of Science degree  
by

---

Yifan Zhao

September 11, 2015

---

Professor Dmitry Korkin, Project Advisor

## Abstract

An Ebola outbreak in spring and summer 2014 drew a lot of public attention because of the staggering amount of deaths it caused. In this project, the proteomes of Ebola and its deadly relatives from the *Filoviridae* (Filovirus) family are studied to better understand the evolution of these viruses that may help in developing better antiviral drugs and vaccines. Specifically, we construct three-dimensional models for the seven proteins in the proteome using the homology modeling approach. Then, we find the conserved and diverse regions in each protein across species and map the regions back to the models we have constructed using trident score. Finally, we find the functional areas in each homology model, specifically those involved in virus-virus interactions and in human-virus interactions, and compare these areas with the conserved and diverse regions. Since our approach relies highly on the existing related structures, some models are not complete. Based on the results, we draw the conclusion that highly conserved regions are linked to the areas involved in virus-virus interactions, while the highly diverse regions are linked to those involved in human-virus interactions.

## Table of Contents

<b>Abstract</b> .....	<b>2</b>
<b>Chapter 1: Introduction</b> .....	<b>5</b>
1.1 Motivation .....	5
1.2 Previous Studies .....	7
1.2.1 Sequence Information .....	7
1.2.2 3D Structures of the Proteome .....	7
1.2.3 Previous Study on Sequence Evolution .....	7
1.2.4 Previous Study on Structural Bioinformatics.....	7
1.3 Our Approach.....	7
<b>Chapter 2: Methods</b> .....	<b>10</b>
2.1 Data Collection.....	10
2.2 Comparative Modeling.....	10
2.2.1 Pre-Processing Steps.....	10
2.2.2 Basic Modeling Method.....	12
2.2.3 Example of Basic Modeling Method .....	13
2.2.5 Example of Advanced Modeling Method .....	16
2.3 Family-Wise Sequence Alignment .....	17
2.3.1 Sequence Selection.....	18
2.3.2 Trident Score .....	19
2.3.3 Mapping Conserved and Diverse Regions.....	20
2.4 Clustering and Function Mapping.....	20
<b>Chapter 3: Results</b> .....	<b>22</b>
3.1 Structural modeling of viral proteins .....	22
3.1.1 NP .....	22
3.1.2 VP35 .....	23
3.1.3 VP40 .....	24
3.1.4 GP.....	24
3.1.5 VP30 .....	25
3.1.6 VP24 .....	25
3.1.7 L.....	26
3.1.8 Summary of Models .....	26
3.2 Family-Wise Sequence Alignment of <i>Filoviridae</i> Proteomes.....	27
3.2.1 Data Selection .....	27
3.2.2 Trident Score Distribution.....	27
3.2.3 Mapping Conserved and Diverse Regions.....	29
3.3 Result for Clustering and Function Mapping.....	33
3.3.1 NP .....	34
3.3.2 VP35 .....	34
3.3.3 VP40 .....	35
3.3.4 GP.....	36
3.3.5 VP30 .....	38
3.3.6 VP24 .....	39
3.3.7 L.....	40
<b>Chapter 4: Conclusions</b> .....	<b>41</b>
<b>Chapter 5: Potential Problems and Future Studies</b> .....	<b>42</b>
<b>Reference</b> .....	<b>43</b>

<b>Appendices</b> .....	<b>45</b>
Appendix A. Protein Sequences Found in UniProt .....	45
Appendix A1. Protein Sequences Found in UniProt (NP) .....	45
Appendix A2. Protein Sequences Found in UniProt (VP35) .....	48
Appendix A3. Protein Sequences Found in UniProt (VP40) .....	50
Appendix A4. Protein Sequences Found in UniProt (GP).....	52
Appendix A5. Protein Sequences Found in UniProt (VP30) .....	57
Appendix A6. Protein Sequences Found in UniProt (VP24) .....	59
Appendix A7. Protein Sequences Found in UniProt (L).....	61
Appendix B. Modeller Scripts .....	64
1. <i>align2d.py</i> .....	64
2. <i>model-single.py</i> .....	64
3. <i>evaluate_model.py</i> .....	65
4. <i>plot_profiles.py</i> .....	65
5. <i>salign.py</i> .....	66
6. <i>align2d_mult.py</i> .....	67
7. <i>model_mult.py</i> .....	68
Appendix C. Genome Selection for Family-Wise Sequence Alignment.....	69
1. <i>NP</i> .....	69
2. <i>VP35</i> .....	69
3. <i>VP40</i> .....	69
4. <i>GP</i> .....	69
5. <i>VP30</i> .....	70
6. <i>VP24</i> .....	70
7. <i>L</i> .....	70
Appendix D. Conserved and Diverse Residues Mapping .....	71
Color Code:.....	71
1. <i>GP</i> .....	71
2. <i>L</i> .....	71
3. <i>NP</i> .....	72
4. <i>VP24</i> .....	73
5. <i>VP30</i> .....	73
6. <i>VP35</i> .....	73
7. <i>VP40</i> .....	74
Appendix E. Model Check and Functional Annotation .....	75
1. <i>VP30</i> .....	75
2. <i>VP40</i> .....	76
3. <i>VP35</i> .....	79
4. <i>VP24</i> .....	84
5. <i>GP</i> .....	85
6. <i>NP</i> .....	91
7. <i>L</i> .....	91
8. <i>RNA-Protein Interactions</i> .....	91
Appendix F. Functional Residues Mapping.....	92
<i>VP35</i> .....	92
<i>VP40</i> .....	92
<i>GP</i> .....	92
<i>VP30</i> .....	92
<i>VP24</i> .....	92

# Chapter 1: Introduction

## 1.1 Motivation

In the year of 2014, Ebola has drawn people's attention because of the large number of death it has caused. As of May 6, 2015, there are 11007 deaths in the countries with widespread transmission, specifically Guinea, Liberia, and Sierra Leone (CDC, 2015). However, this is not the first time that Ebola has caused problems in the world. In 1976, Ebolavirus was first recognized for its outbreak in Yambuku and surrounding area. It was reported that 318 people were infected, and among them, 280 people died (CDC, 2015).

Similarly, other members belonging to the same family, *Filoviridae*, can cause hemorrhagic fever in both human and nonhuman primates (Serena A. Carroll, 2013). According to a proposal for revised taxonomy of the family *Filoviridae* (Kuhn JH, 2010), there are three genera in the family of *Filoviridae*, including *Ebolavirus*, *Marburgvirus*, and a tentative genus *Cuevavirus*. The detailed classification is shown in Table 1. In the genus of *Ebolavirus*, Ebola virus has the highest case fatality rate, up to 89%. The species of *Sudan Ebolavirus* was also discovered in 1976 in Nzara, Maridi and the surrounding area, and 284 people were infected and 151 of them died, causing a fatality rate of 53% (CDC, 2015). Reston virus was discovered in 1989 in monkeys imported from Philippines. This virus was also found in four people in USA and three people in Philippines, but all of them were asymptomatic. Tai Forest virus was found in 1994, and only one case has been found so far. The infected patient was treated in Switzerland. Bundibugyo virus is the most recently found species. It was found in Uganda, in 2007, with a fatality rate of 25%. In the genus of *Marburgvirus*, there is only one species, consisting of two viruses, Marburg virus and Ravn virus. Marburg virus was first found in 1967 in Germany and Yugoslavia at the same time. In large outbreaks in 1998-2000 and in 2004-2005, 83% and 90% infected people dies respectively, according to the chronology of Marburg Hemorrhagic Fever Outbreaks summarized by Centers for Disease Control and Prevention (Centers for Disease Control and Prevention, 2014). The last genus, *Cuevavirus*, was found in bat carcasses from Spain in 2002 (Ana Negredo, 2011). No fatal case in human has been found.

Table 1: Taxonomy of Family *Filoviridae* Proposed in 2010 (Kuhn JH, 2010)

Genus	Species	Virus
<i>Marburgvirus</i>	<i>Marburg marburgvirus</i>	Marburg virus (MARV)
		Ravn virus (RAVV)
<i>Ebolavirus</i>	<i>Tai Forest ebolavirus</i>	Tai Forest virus (TAFV)
	<i>Reston ebolavirus</i>	Reston virus (RESTV)
	<i>Sudan ebolavirus</i>	Sudan virus (SUDV)
	<i>Zaire ebolavirus</i>	Ebola virus (EBOV)
	<i>Bundibugyo ebolavirus</i>	Bundibugyo virus (BDBV)
<i>Cuevavirus</i> (tentative)	<i>Lloviu cuevavirus</i> (tentative)	Lloviu virus (LLOV)

Ever since the discovery of Ebola and Marburg, people have been studying the pathogenesis, genome, and genome products of them. On the other hand, the genus *Cuevavirus* is the least studied, simply because it was discovered only a couple of years ago.

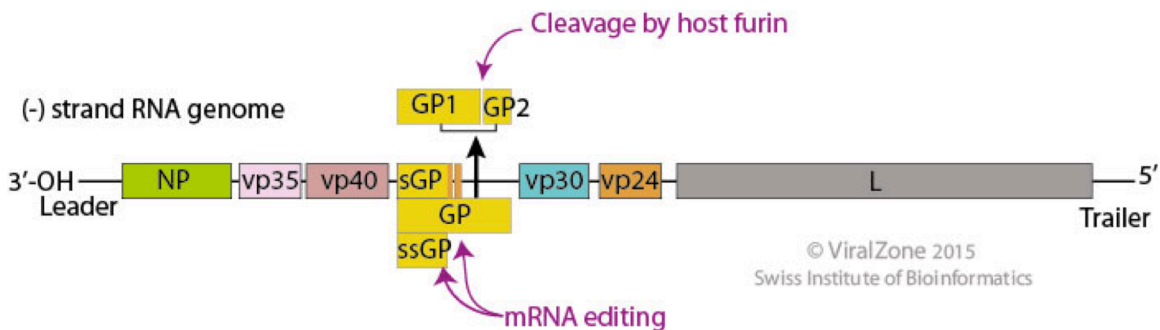


Figure 1: Genome of Ebolavirus (Extracted from [http://education.expasy.org/images/Filovirus\\_genome.jpg](http://education.expasy.org/images/Filovirus_genome.jpg))

Triggered by the outbreak in 2014, we wonder whether we have enough information about Ebola and its family on the molecular level, how Ebola is related to other species within the family in terms of molecular evolution, and what useful information we can find out based on the structural information. Unlike the previous studies on genome of filovirus, which mainly focus on the two dimensional sequences of proteins, this project focuses on the

Table 2: Proteome of Zaire Ebolavirus. The organism is Zaire Ebolavirus (Strian Mayinga-76) (ZEBOV) (Zaire Ebola Virus). The entry ID can be used in UniProt.

Entry	Entry name	Protein names	Gene names	Length
Q05318	L_EBOZM	RNA-directed RNA polymerase L (Protein L) (Large structural protein) (Replicase) (Transcriptase) [Includes: RNA-directed RNA polymerase (EC 2.7.7.48); mRNA (guanine-N(7)-)-methyltransferase (EC 2.1.1.56); mRNA guanylyltransferase (EC 2.7.7.-)]	L	2212
P18272	NCAP_EBOZM	Nucleoprotein (Nucleocapsid protein) (Protein N)	NP	739
Q05320	VGP_EBOZM	Envelope glycoprotein (GP1,2) (GP) [Cleaved into: GP1; GP2; GP2-delta]	GP	676
Q05322	VP24_EBOZM	Membrane-associated protein VP24	VP24	251
Q05323	VP30_EBOZM	Minor nucleoprotein VP30 (Transcription activator VP30)	VP30	288
Q05128	VP40_EBOZM	Matrix protein VP40 (Membrane-associated protein VP40)	VP40	326
Q05127	VP35_EBOZM	Polymerase cofactor VP35	VP35	340
P60170	VSGP_EBOZM	Pre-small/secreted glycoprotein (pre-sGP) [Cleaved into: Small/secreted glycoprotein (sGP); Delta-peptide]	GP	364
Q9YMG2	VSSGP_EBOZM	Super small secreted glycoprotein (SsGP)	GP	298

three dimensional structures of proteins. The advantage of studying the three dimensional structures over the two dimensional sequences lies in that some distantly related amino acids from the view of sequence can be involved in active interactions due to the folding of protein sequences. Intuitively, if a region on a protein is very consistent across species, it is likely to be the essential region for the virus to survive. On the contrast, if a region on a protein changes over time, it is likely to have evolved to avoid antibodies.

## 1.2 Previous Studies

### 1.2.1 Sequence Information

In order to solve the above questions, we studied the viruses in the family *Filoviridae*, which are nonsegmented, negative-strand RNA viruses (Serena A. Carroll, 2013). Among the family, the species *Zaire Ebola virus* is the one that causes the 2014 outbreak. There are seven genes in the proteome of Zaire Ebola, encoding the following proteins: nucleoprotein (NP), polymerase cofactor VP35 (VP35), matrix protein VP40 (VP40), glycoprotein (GP), minor nucleoprotein VP30 (VP30), membrane-associated protein VP24 (VP24), and RNA-directed RNA polymerase L (L). Furthermore, glycoprotein can form two variants, pre-small/secreted glycoprotein (sGP), and super small secreted glycoprotein (ssGP) (Swiss-Prot, 2002-2015). Figure 1 shows the genome of Ebola virus (Bioinformatics, 2015). Table 2 summarizes the proteome information described above, along with the length of each protein in terms of amino acids.

### 1.2.2 3D Structures of the Proteome

In terms of three dimensional (3D) structures of the proteins, there is still a significant gap in the knowledge in this area. According to the video that PDB (Protein Data Bank) published in 2014 (PDB, 2014), the crystal structures of VP24 and VP40 have been fully resolved, while those of glycoprotein, VP30, nucleoprotein, and VP35 have not been fully resolved, and the structure of large structural protein is still a myth yet.

### 1.2.3 Previous Study on Sequence Evolution

The conventional way to study the evolution is to align sequences of interest and then construct the phylogenetic tree. A research paper published in 2013 analyzed the molecular evolution of viruses in the family *Filoviridae* based on 97 whole-genome sequences (Serena A. Carroll, 2013). Using the “Bayesian coalescent” analyses, they have constructed phylogenetic trees for major genera and the whole family. The result also predicts the emergence of the common ancestor between two genera. However, this does not provide us with any insight on the structural details.

### 1.2.4 Previous Study on Structural Bioinformatics

In terms of protein structures in this family, many studies have examined the crystal structures of the proteins; one study examined the structure of the same protein VP35 in Zaire Ebola virus in various forms, and it also suggests VP35 is a potential target for drugs (Craig S. Brown, 2014); some studies found the structures for interactions between a certain protein and other proteins, such as the interaction between NP and VP35. However, none of them has studied the overall evolution of this family based on the structural information.

## 1.3 Our Approach

Given the fact that the family-wise evolution based on the protein structures has not been studied, we aimed to tackle this task from the point of view of structural bioinformatics. Inspired by a research about influenza published in 2013, we decided to follow a similar approach. From the paper about influenza, the researchers filtered out 75 non-redundant strains from 1100 complete genomes (Samantha Warren, 2013). However, in order to reduce computational cost, they only built models for all ten proteins in one randomly selected strain. Then they identified the exterior residues and interior residues using CalcSurface. Next they checked the evolution for the exterior and interior residues

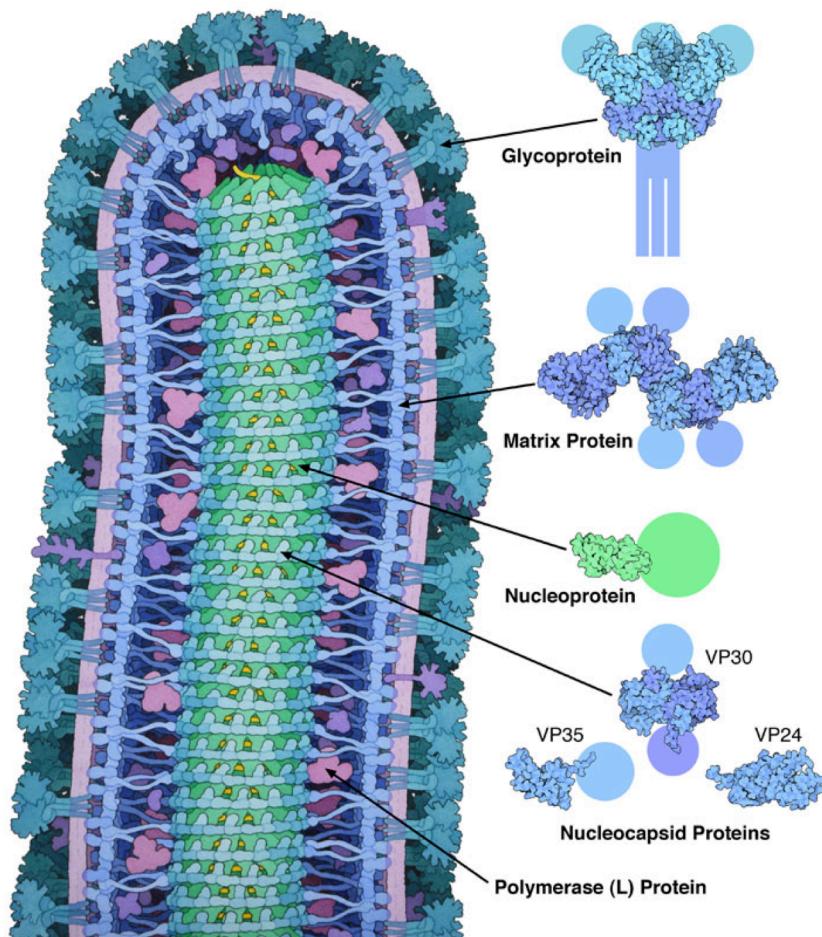


Figure 2: Overview of proteins in Ebolavirus. Extracted from Protein Data Bank, Molecule of the Month: [http://www.rcsb.org/pdb/education\\_discussion/molecule\\_of\\_the\\_month/download/Ebola.pdf](http://www.rcsb.org/pdb/education_discussion/molecule_of_the_month/download/Ebola.pdf)

separately, first by aligning the sequences and finding the 100% conserved sequences, and then by calculating the Euclidean distance between the all pairs of 100% conserved residues. They also constructed phylogenetic trees to help analyze. As a result, they found that the highly conserved regions on the surface do not overlap with the regions involved in host-viral interactions. However, such regions are associated with the intra-viral interactions.

Similarly, we are interested in the spatial locations of the conserved regions as well as the diverse regions in the family of *Filovirus*. Unlike the paper we referred to, where most concepts are defined by a certain scientifically rigorous algorithm, such as the definitions of exterior residues (surface residues), interior residues (core residues), conserved residues, and diverse residues, we decide most of the cutoffs by intuition. To start with, since not all the structures in the proteome of Ebola have been solved, we constructed the three-dimensional models for all the proteins based on homology modeling method. Then we aligned representative sequences of all species within this family and tried to find the extremely conserved residues as well as the extremely diverse residues, which are then mapped back to the three-dimensional structures we built. Finally, we predicted the functional regions in the proteins and mapped back to the models, too. This way, we can compare the highly conserved and diverse regions with the predicted functional regions.



Our hypothesis is similar to the conclusion in the reference paper about influenza (Samantha Warren, 2013): The highly conserved regions across species should be involved in intra-viral interactions, while the highly diverse regions across years should be involved in host-viral interactions. The logic behind this hypothesis is intuitive: In order to reach minimal requirement for survival, virus has to keep certain essential regions unchanged so that the proteins within this virus can recognize itself and collaborate with other proteins or RNA. On the other hand, in order for the species of virus to survive over years, the virus has to “evolve” to avoid being detected by the host antibodies. Therefore, the interactions between the host and virus are unlikely to stay the same over years. So, the residues involved in host-viral interactions in an outbreak several decades ago should be found extremely diverse throughout the years.

## Chapter 2: Methods

The basic methodology is similar to the one regarding H1N1 (Warren S, 2013). Overall, to discover the conserved and diverse regions about the proteins in Ebolavirus, the models for the 7 proteins were built using homology modeling method. Then, a representative from each species in the *Filoviridae* family was chosen and highly conserved and highly diverse residues were found and mapped back to the models built in the previous steps. Careful observations were made and the potential functional annotations were created through searching SCOP family of the related sequences. After studying the conservation and diversity across species, the similar approach is applied to studying the conservation and diversity across years. Detailed procedures will be discussed in each section below.

### 2.1 Data Collection

To start with, we tried to collect as many sequences in the *Filoviridae* family as possible. The searching databases used to find the sequences included UniProt, founded by European Bioinformatics Institute et al, (European Bioinformatics Institute, the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource, 2015) and NCBI founded by National Center for Biotechnology Information in US (National Center for Biotechnology Information, U.S. National Library of Medicine, 2015). For each species found in the Introduction section, we searched the protein names combined with the species in UniProt. Then the summary is downloaded with the information about the entry, entry name, status, protein names, gene name, organism, length, date of creation, place, strain, and year. The place, strain, and year were parsed from the strain information, and the date of creation was the date the entry was created in UniProt, which does not reflect the actual date when the sequence was found. Sequences that are segments of a protein were ignored. The summaries can be found in the appendices.

Similar searching was done in NCBI. However, most of the results overlapped with those from UniProt, since both search engines share the data. At the time of searching, NCBI provided 99 genomes of Ebolavirus from 2014 outbreak, which were not available in UniProt, and the protein ID for each sequence was also recorded in the appendices.

### 2.2 Comparative Modeling

After finding the sequences, models for each protein are made using MODELLER (Andrej Šali, 1993) in order to further study the 3D structure in the future steps. As described in its website, MODELLER uses existing structures as template and build the target model by satisfaction spatial restraints (Sali Lab, 2015). This method is also called homology modeling. There are five types of modeling in the tutorial (Sali Lab, 2010). In this project, only the Basic Modeling method and the Advanced Modeling method were used to construct the models. The procedures used were very similar to the ones described in the tutorial, with only a minor modification in the first two steps. Here is a basic flow of how to build models in MODELLER.

#### 2.2.1 Pre-Processing Steps

To start with, the target sequence for each protein was to be determined. Here, the first sequence for each protein in the 99 sequences found in the 2014 outbreak was chosen as our targets. Their sequence IDs in NCBI are as follows: [AIE11806.1](#) (NP), [AIE11807.1](#) (VP35), [AIE11808.1](#) (VP40), [AIE11809.1](#) (GP), [AIE11812.1](#) (VP30), [AIE11813.1](#) (VP24), and [AIE11814.1](#) (L). In the tutorial, the first step is to find structures related to the target sequence using a

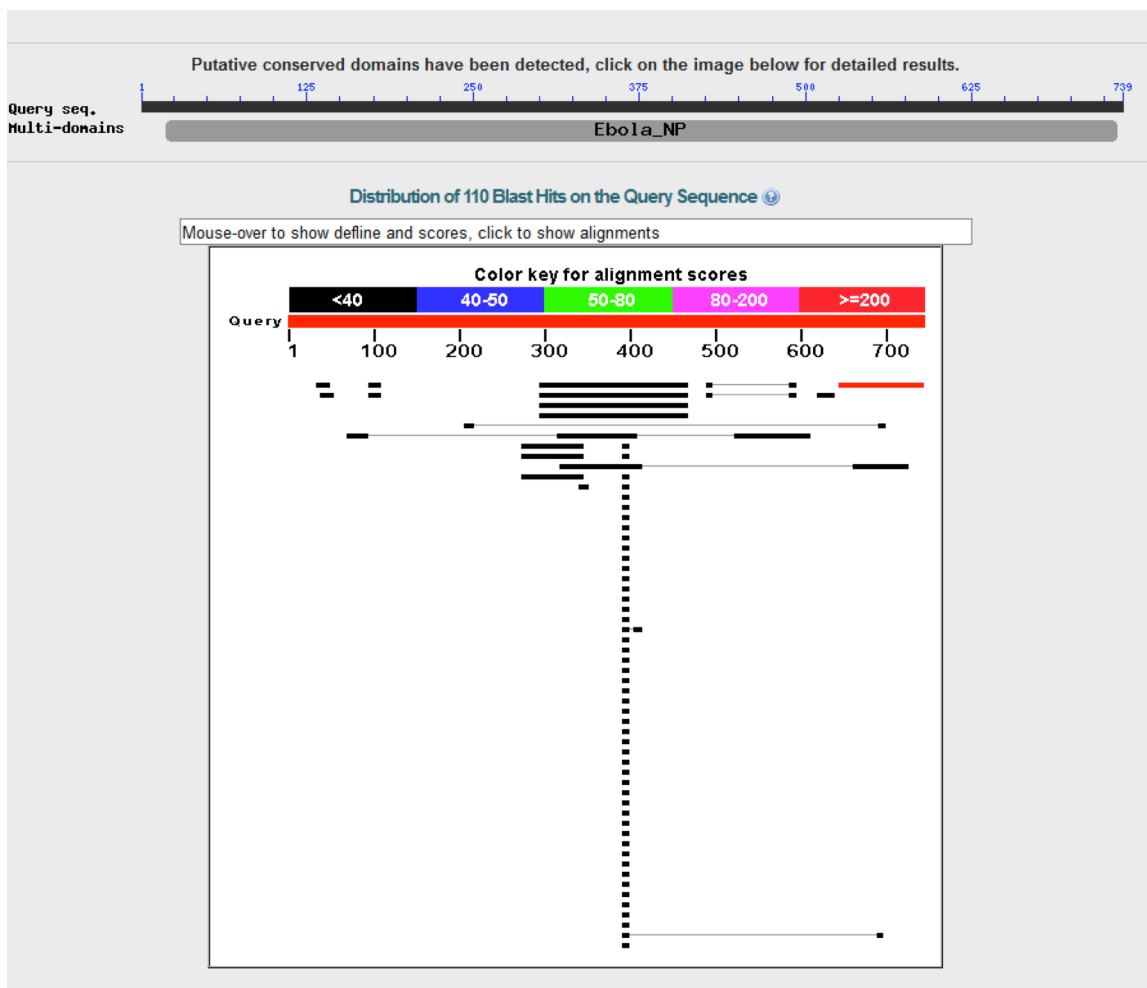


Figure 3: NP PSI-BLAST Result

binary file called “pdb\_95.bin”, which contains structures in PDB that are less than 95% similar. However, the pdb\_95.bin file can be outdated. So, in this project, we chose an alternative approach. For each sequence, PSI-BLAST was performed against Protein Data Bank. The tool we used is the internal link in NCBI website, which will give us a very useful graphical summary.

For example, Figure 3 is a screenshot of the PSI-BLAST result of NP. The color key for alignment scores shows that a red sequence has the highest score, while a black sequence has a relatively low score. The red bar below the color key is the query sequence, and the coordinates roughly show the positions where the two sequences are aligned. Specifically, in Figure 3, a small red sequence is found to align with our query near the end.

While in Figure 4, more than one red sequence is aligned to the same portion of our query at the same time. In this case, we can build a model using multiple templates, corresponding to the advanced modeling method. Usually, we first build a basic model before building an advanced one, even when more than one template can be chosen. Then, we would build an advanced model and compare it with the basic one using DOPE score, which will be further discussed in later sections.

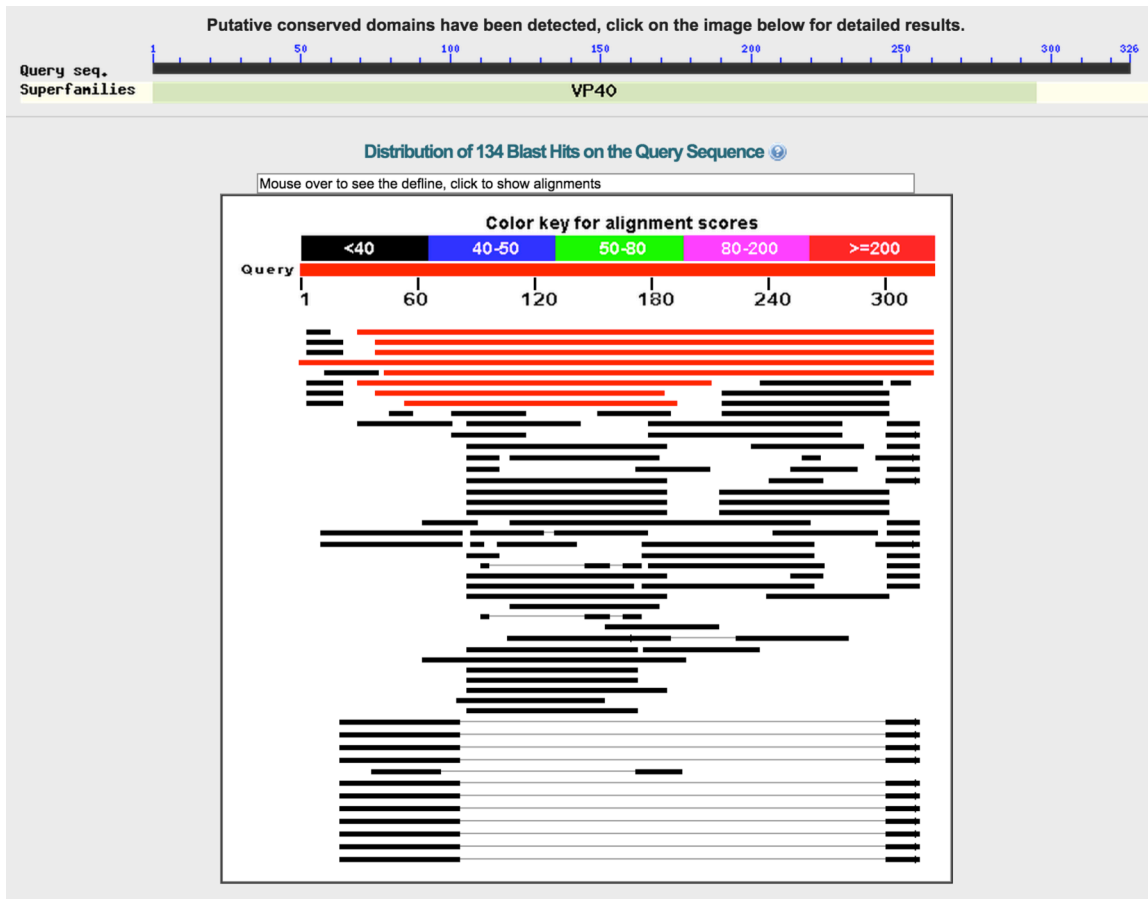


Figure 4: VP40 PSI-BLAST Result

If several sequences can be chosen as the template, but they vary in coverage and identity, the sequence with the best combination of coverage and identity will be chosen finally.

After selecting a template (or multiple templates) for the target sequence, we then can build models using the template(s).

### 2.2.2 Basic Modeling Method

If a single template is selected, we can continue to build the model using the basic modeling method, which is comprised of the following steps.

1. Align the target sequence with the template.
2. Build models based on the template using python script.
3. Evaluate the models.

#### 1. Aligning the Target Sequence with the Template

After selecting a template, we can align the template with the sequence using MODELLER library (See align2d.py in the appendices). Before the alignment, small adjustment has to be done to convert the ".fasta" file to ".ali" file, according to the PIR format instructions (Webb, 2007).

#### 2. Model Building

Then we can build the model using the alignment file, our target sequence and our template (See model-single.py in the appendices). 5 models will be generated.

### 3. Model Evaluation

As introduced in the tutorial, several measurements can be used to evaluate models and therefore, to select the best one among the models built in the previous step. One of them is DOPE (Discrete Optimized Protein Energy) score, which calculates the potential in a certain model. The lower the DOPE score is, the better the model is. To evaluate whether the models have been built successfully, we can open the five models together in Chimera at the same time and superimpose them. If they align pretty well, the models are very consistent, and thus the models are generally acceptable. However, if the models differ from each other greatly, or it contains a long free segment, the modeling process has not been very successful. In such cases, advanced modeling method might be applied if more than one template can be found.

#### 2.2.3 Example of Basic Modeling Method

Here is an example of how to construct a basic model for VP40 using one template. The corresponding NCBI protein ID is AIE11808.

First, a PSI-BLAST against the known structures in PDB was done. Figure 4 and Table 3 summarize the results. (Note that the order of sequences in the graphical summary is different from that of the table. When you read the graphical summary in NCBI, the sequence information will show up when mouse-over.) As we can see, 1H2D and 1H2C have the highest identity (100%); 1ES6, 4LDB, 4LDD, 4LDI, and 4LDM are almost identical to our target (99%), and among them 1ES6 has the longest coverage. On the other hand, 3TCQ has the overall longest coverage, but it is only 75% identical to the target. So, in order to have the best model possible, we need to balance the identity and the coverage. If we consider building a basic model, we can either use 3TCQ or 1ES6. Since 3TCQ has the longest coverage and 75% is a decent identity. However, when opening 3TCQ, it shows that the structure corresponding to the first 44 residues and the last 18 residues is missing. Overall, 1ES6: A should be the best template, with a 99% identity and coverage of 296 residues in length. After selecting the template, we aligned the template with our target sequence using Modeller script “align2d.py” (See Appendix B. 1). The alignment result is shown below in pap format in Figure 5.

Then we could construct the models based on the alignment by running the script “model-single.py” (See Appendix B. 2). Five models were generated after running the script. Among them, the best is the fifth one, with the lowest total DOPE score. Below is the summary of

**Table 3: PDB Advanced Sequence Search Result for VP40**

Code	Chains	E-value	Identity	Start	End	Length
1ES6	A	6.44E-170	294/296 (99%)	31	326	296
4LDB	A,B,C,D	9.16E-163	282/284 (99%)	43	326	284
4LDD	A,B,C	9.16E-163	282/284 (99%)	43	326	284
4LDI	A,B	5.79E-162	281/284 (99%)	43	326	284
3TCQ	A	3.59E-141	248/329 (75%)	1	326	329
4LD8	A	1.16E-136	232/282 (82%)	45	326	282
1H2D	A,B	7.13E-103	182/182 (100%)	31	212	182
4LDM	A	6.72E-80	145/146 (99%)	43	188	146
1H2C	A	1.61E-77	140/140 (100%)	55	194	140

```

|_aln.pos      10      20      30      40      50      60
1es6A  -----GDTPSNPLRPIADDTIDHASHTPGS
AIE11808 MRRVILPTAPPEYMEAIYPARSNSTIARGGNSNTGFLTPEVNGDTPSNPLRPIADDTIDHASHTPGS
  _consrvd  *****

|_aln.p      70      80      90      100     110     120     130
1es6A  VSSAFILEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSFDSTTAAIMLASYTITHFGKATNPLVRVN
AIE11808 VSSAFILEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSFDSTTAAIMLASYTITHFGKATNPLVRVN
  _consrvd  *****

|_aln.pos     140     150     160     170     180     190     200
1es6A  RLGPGIPDHPLRLLRIGNQAFLEFVLPVQLPQYFTFDLTALKLITQPLPAATWTD-----DGALR
AIE11808 RLGPGIPDHPLRLLRIGNQAFLEFVLPVQLPQYFTFDLTALKLITQPLPAATWTDPTGNSGALR
  _consrvd  *****

|_aln.pos     210     220     230     240     250     260     270
1es6A  PGISFHPKLRPILLPN-----KNSADLTSPEKIQAIMTSLQDFKIVPIDPTKNIMGIEVPETLVKLKLT
AIE11808 PGISFHPKLRPILLPNKSGKGNKNSADLTSPEKIQAIMTSLQDFKIVPIDPTKNIMGIEVPETLVHKLKLT
  _consrvd  *****

|_aln.pos     280     290     300     310     320
1es6A  -----GGQPIIPVLLPKYIGLDPVAPGDLTMVITQDCDTCHSPASLP-----
AIE11808 GKKVTSKNGQPIIPVLLPKYIGLDPVAPGDLTMVITQDCDTCHSPASLPVAVVEK
  _consrvd  *****

```

Figure 5: Alignment between 1ES6 and VP40 (AIE11808-13s6A.pap)

scores for each model produced by Modeller:

```

>> Summary of successfully produced models:
Filename          molpdf  DOPE score  GA341 score
-----
AIE11808.B99990001.pdb  2246.65991 -32501.79883  1.00000
AIE11808.B99990002.pdb  2261.87061 -32414.62695  1.00000
AIE11808.B99990003.pdb  2250.31812 -32688.30273  1.00000
AIE11808.B99990004.pdb  2314.44580 -32276.85547  1.00000
AIE11808.B99990005.pdb  2123.74365 -32741.74609  1.00000

```

When opening the five models in Chimera at the same time and superimpose all of them, most of them can be aligned well, except for the ends, as shown in Figure 6. The consistency of models indicates the models are reliable to some degree. Also, since the GA341 scores for all models equal 1, all models can be considered as “good” models.

Now, we could further evaluate our best model based on the DOPE per-residue score, and plot it against that for the template structure. Before plotting the DOPE scores, we need “profile” files for both the template and our model. Using “evaluate\_model.py” (See Appendix B. 3) we could get the corresponding “profile” file. Then we could generate the plot using “plot\_profiles.py” (See Appendix B. 4). The plot generated is shown in Figure 7. As we can see, there is space for improvement around residue 280, where the chosen template does not have a corresponding structure. We could try to construct an advanced model using multiple templates.

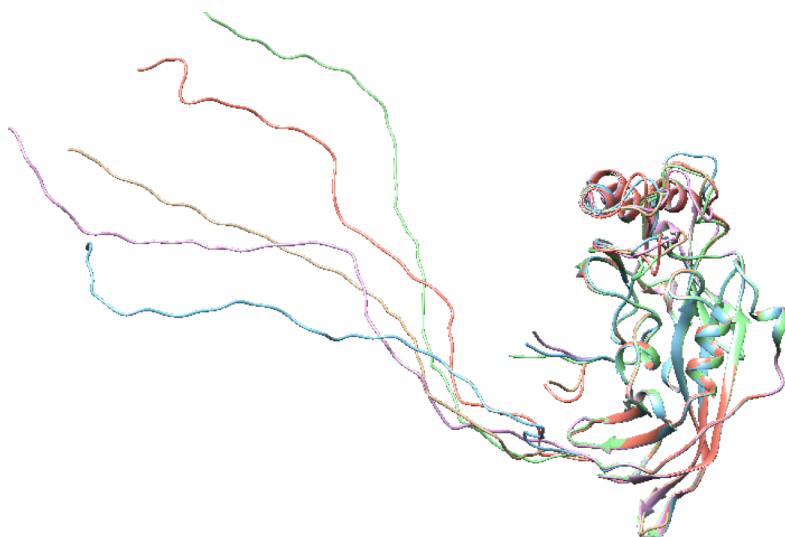


Figure 6: Basic Models for VP40 Aligned in Chimera

### 2.2.4 Advanced Modeling Method

In the advanced tutorial, there are three techniques to improve the models: multiple templates, loop refining, and modeling ligands in the binding site. In this project, only the first method is used to construct better models because of time constraint. As illustrated in the pre-processing section, sometimes, we can find more than one sequence aligned with a certain part of our query. Such information is useful for constructing a better model because the information from other templates may compliment the information from our basic template. The steps to construct an advanced model is as follows:

#### 1. Selecting Templates

First, we need to determine which templates we should use. In the tutorial, they find the related structures by searching the family members for the template used in the basic modeling method. But in this project, we simply continued using the PSI-BLAST result.

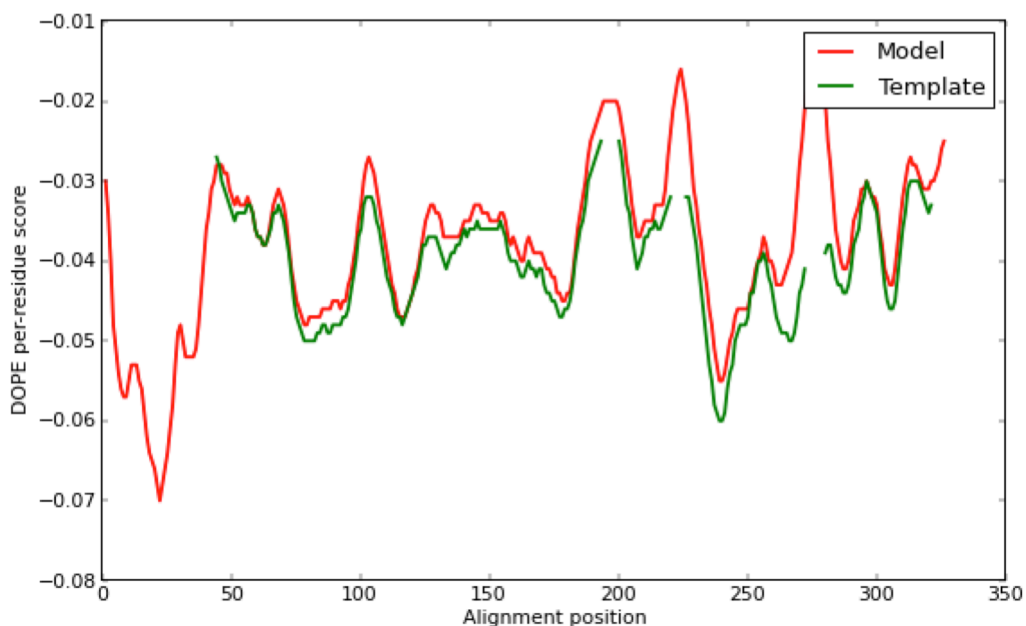


Figure 7: Model Assessment Using DOPE Score Per Residue measure

```

|_aln.pos      10      20      30      40      50      60
1es6A      GDTPSNPLRPIADDTIDHASHTPGSVSSAFIEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSDST
4ldbB      GDTPSNPLRPIADDTIDHASHTPGSVSSAFIEAMVNVISGPKVLMKQIPIWLPLGVADQKTYSDST
3tcqA      MDTPSNMRPVADDNIDHTSHTPNGVASAFIEATVNVISGPKVLMKQIPIWLPLGIADQKTYSDST
_consrvd    ***** ** *** ** * ***** *****

|_aln.p       70      80      90      100     110     120     130
1es6A      TAAIMLASYTITHFGKATNPLVRVNRLLGPGIPDHPLRLLRIGNQAFLLQEFVLPVPPVQLPQYFTFDLTAL
4ldbB      TAAIMLASYTITHFGKATNPLVRVNRLLGPGIPDHPLRLLRIGNQAFLLQEFVLPVPPVQLPQYFTFDLTAL
3tcqA      TAAIMLASYTITHFGKANNPLVRVNRLLGQIPDHPLRLLRMGNQAFLLQEFVLPVPPVQLPQYFTFDLTAL
_consrvd    ***** ***** ***** *****

|_aln.pos     140     150     160     170     180     190     200
1es6A      KLITQPLPAATWTDGALRPGISFHPKLRPILLPNKNSADLTSPEKIQAIMTSLQDFKIVPIDPTKNI
4ldbB      KLITQPLPAATWTD-GALRPGISFHPKLRPILLPN-----LTSPEKIQAIMTSLQDFKIVPIDPTKNI
3tcqA      KLVTQPLPAATWTD----RPGLSFHPKLRPVLLPG-----TAPDKIQTIVNLMQDFKIVPIDPAKSI
_consrvd    ** ***** ** ***** ** * * * * * ***** * *

|_aln.pos     210     220     230     240     250     260
1es6A      MGIEVPETLVLLKLTG-----GQPIIPVLLPKYIGLDPVAPGDLTMVITQDCDTCCHSPASLP
4ldbB      MGIEVPETLVHKLTKGKVTSKNGQPIIPVLLPKYIG-----GD-LTMVITQDCDTC-----CH
3tcqA      IGIEPELLVHKLTKGKMSQKNGQPIIPVLLPKYIG-----PGDLTMVIT-----
_consrvd    ***** ** **** ***** *****

```

Figure 8: Alignment Among Templates (multitemp.pap)

2. Aligning the Target with Multiple Templates

After determining the templates, we would align the templates using the command “salgn()” (See the complete script in Appendix B.1). After checking the alignment file, we can align our query sequence to the templates (See Appendix B. 6).

3. Model Building

Similar to what we have done in the basic modeling method, we would then construct the model based on the alignment.

4. Model Evaluation

Finally, we would evaluate the model using DOPE score. To compare the advanced model with the basic one, we can plot the DOPE per residue score for both models, and see whether there exists any difference. The scripts used for the last two steps are almost identical to those used in the basic modeling method. Please check Appendix B. 3 and 4. An additional step can be to compare the best model we constructed using the basic method with that using the advanced method.

2.2.5 Example of Advanced Modeling Method

As shown in Figure 4, the first four sequences (colored red) have a good coverage of our query, but after checking the sequences, the second and the third sequence are almost identical. So, three sequences were chosen to build the model: 1ES6:A, 4LDB:B, and 3TCQ:A. Then, we aligned the templates using “salgn.py”, which can be found in Appendix B. 5. Figure 8 is the alignment result among the templates. Next, we aligned the query with the aligned templates using “align2d\_mult.py”, which can be found in Appendix B. 6. Figure 8 is the alignment result between the templates and the query. Finally, we constructed the model using the script “model\_mult.py”, which can be found in Appendix B. 7. As in the basic method, MODELLER generated 5 models. The DOPE scores for the models are listed below:



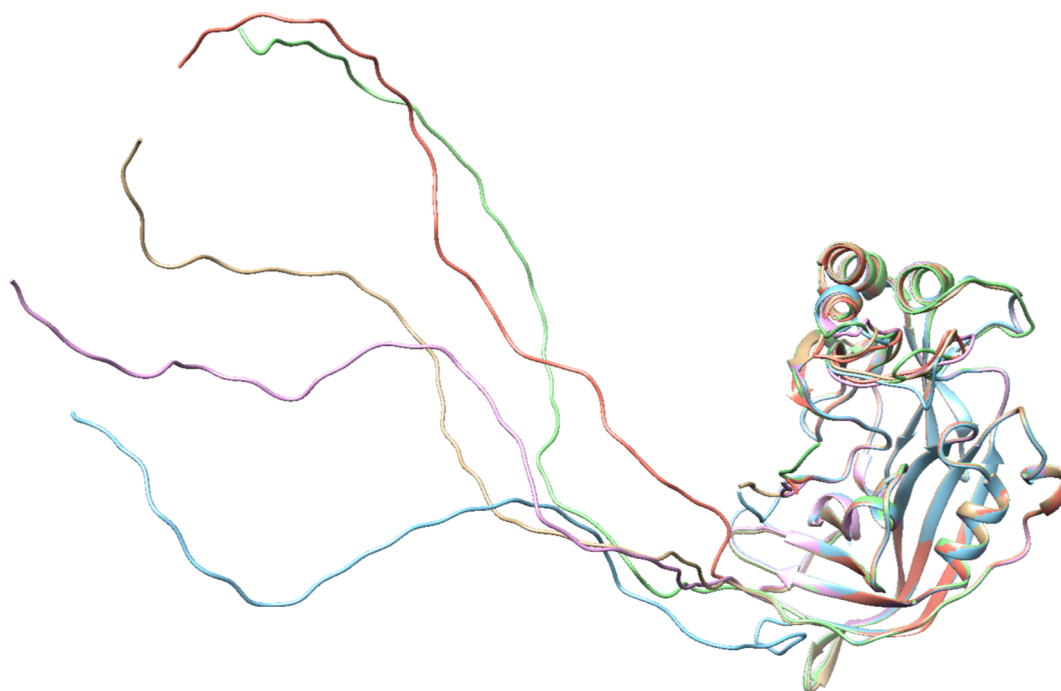


Figure 9: Advanced Models for VP40 Aligned in Chimera

>> Summary of successfully produced models:

Filename	molpdf	DOPE score	GA341 score
AIE11808.B99990001.pdb	10951.37500	-32580.52734	1.00000
AIE11808.B99990002.pdb	10891.88184	-32887.29297	1.00000
AIE11808.B99990003.pdb	10914.78125	-32747.57422	1.00000
AIE11808.B99990004.pdb	10882.16211	-32652.26562	1.00000
AIE11808.B99990005.pdb	10948.46484	-32930.75781	1.00000

So, the best model according to the DOPE score is the fifth one. When opening the five models in Chimera at the same time, they can superimpose with each other very well, which indicates the models are very consistent.

Finally, we can compare the best advanced model with the best basic model to see whether the model has been improved. To compare them, we could plot out the DOPE score per residue for both models. As we can see from Figure 10, the advanced model generally has a lower DOPE score per residue. Especially in the region around residue 270, the advanced model has a lower DOPE score. Therefore, the advanced model is better.

### 2.3 Family-Wise Sequence Alignment

After building the models, we want to find out how much similar the Ebola proteome is to those from other species. So, we continued to align the Ebola protein sequence one by one with the corresponding sequence from other species in the same family, and then score the alignment per residue based on trident score, which will be further discussed in this chapter.

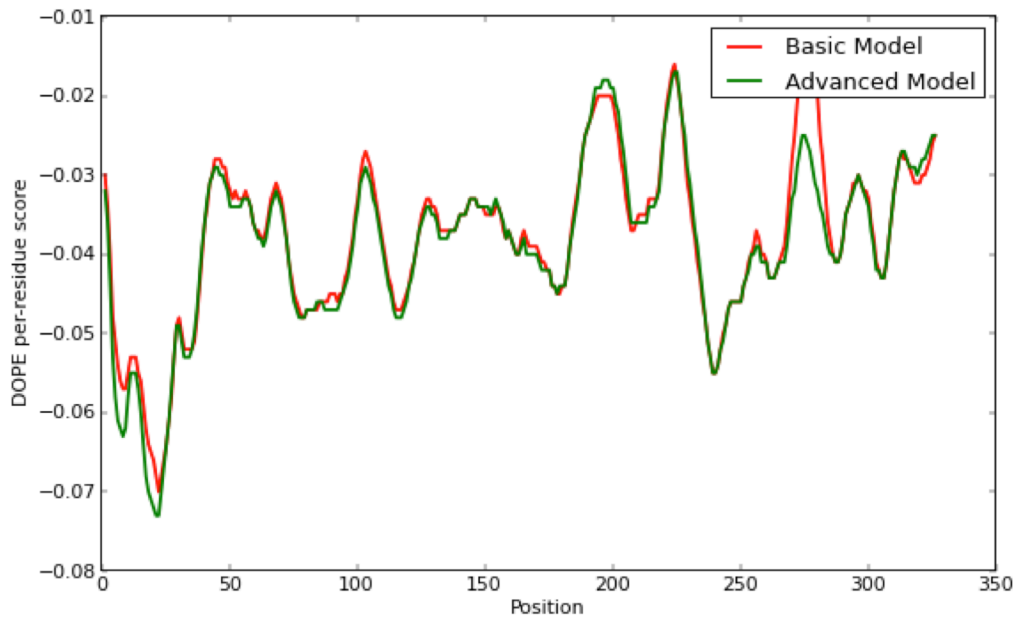


Figure 10: DOPE Score Per Residue for Both Models

### 2.3.1 Sequence Selection

As mentioned in Section 2.1, we collected data for 99 Zaire Ebolavirus genomes from 2014 outbreak, and all sequences available in UniProt for all species under the *Filoviridae* family. However, the numbers of sequences from different species differ greatly. For example, there are 43 Marburg sequences, while there is only 1 Lloviu cuevavirus sequence. Besides, if we use the 99 sequences found in the year 2014, we would have more than 100 Zaire

- [Filoviridae](#) *Click on organism name to get more information.*
- [Cuevavirus](#)
  - [Lloviu cuevavirus](#)
- [Ebolavirus](#)
  - [Bundibugyo virus](#)
  - [Reston ebolavirus](#)
    - [Reston ebolavirus - Reston](#)
    - [Reston ebolavirus - Reston \(1989\)](#)
    - [Reston ebolavirus - Siena/Philippine-92](#)
  - [Sudan ebolavirus](#)
    - [Sudan ebolavirus - Maleo \(1979\)](#)
    - [Sudan ebolavirus - Nakisamata](#)
    - [Sudan ebolavirus - Uganda \(2000\)](#)
    - [Sudan virus - Boniface, Sudan, 1976](#)
  - [Tai Forest ebolavirus](#)
    - [Tai Forest virus - Cote d'Ivoire, Cote d'Ivoire, 1994](#)
  - [Zaire ebolavirus](#)
    - [Ebola virus](#)
  - [unclassified Ebolavirus](#)
    - [Ebola virus Yambio0401](#)
    - [Ebola virus Yambio0402](#)
    - [Ebola virus Yambio0403](#)
    - [Ebola virus sp.](#)
- [Marburgvirus](#)
  - [Marburg marburgvirus](#)
    - [Lake Victoria marburgvirus - Angola2005](#)
    - [Lake Victoria marburgvirus - Ci67](#)
    - [Lake Victoria marburgvirus - DRC1999](#)
    - [Lake Victoria marburgvirus - Leiden](#)
    - [Lake Victoria marburgvirus - Ozolin](#)
    - [Lake Victoria marburgvirus - Popp](#)
    - [Marburg virus - Musoke, Kenya, 1980](#)
    - [Ravn virus - Ravn, Kenya, 1987](#)
  - [unclassified Filoviridae](#)

Figure 11: Filoviridae Taxonomy from NCBI

The screenshot shows the 'The Scorecons Server' interface. At the top, it says 'A program to score residue conservation in a multiple sequence alignment'. Below this are sections for 'Alignment input', 'Output options', 'Scoring options', and 'Action'. In the 'Scoring options' section, the 'Scoring method' is set to 'trident'. Under 'General score parameters', 'substitution matrix' is 'BLOSUM62', 'matrix transformation' is 'karlinlike', and 'gapphilia' is '0'. Under 'Parameters for trident score', 'diversity' is '1', 'chemistry' is '0.5', and 'gaps' is '3'. There are also checkboxes for 'replace raw scores with their fractional rank' and 'Calculate scores for...' with radio buttons for 'all positions' (selected) and 'only positions relating to sequence 0'.

Figure 12: Parameters to Compute the Trident Score

Ebolavirus genomes in the end. Since we wanted to compare the sequences in a quantitative way, which will be introduced later in this chapter, the abundance of each species would introduce bias for the analysis step. Therefore, we altered our method. Instead of using all the sequences, we selected one genome for each species identified in NCBI Taxonomy, and aligned them using MAFFT automatic alignment. Below is a screenshot of the *Filoviridae* taxonomy from NCBI.

### 2.3.2 Trident Score

Then, to measure the degree to which Ebola is different from other species within the *Filoviridae* family, we need to find a way to assess the alignment quantitatively. The measurement we used is called “trident” score (Valdar, 2002), which considers symbol diversity, stereochemical diversity and gaps. The formula to compute the trident score is shown below:

$$C_{trident}(x) = (1 - t(x))^{\alpha} (1 - r(x))^{\beta} (1 - g(x))^{\gamma}$$

In this formula,  $t(x)$  is Shannon’s entropy, or symbol diversity, while  $r(x)$  is the stereochemical diversity, and  $g(x)$  is the fraction of symbols in a certain column that are gaps. Each of them is continuous and normalized from 0 to 1. And  $\alpha$ ,  $\beta$ , and  $\gamma$  are the weights of each parameter. For example, if  $\alpha$ ,  $\beta$ , and  $\gamma$  all equal 1, then they are equally weighted, but in this way too much weight is on stereochemistry but too little weight is on gaps. Since all three parameters range from 0 to 1, the trident score will range from 0 to 1 as well. The advantage of using trident score lies in that not only diversity is considered, but also the chemical property is taken into account.

Since the calculation of the trident score is too complicated, we used an online tool, called Scorecons (Valdar, 2002). The parameters we used to compute the trident score are shown in the screenshot in Figure 12. We chose BLOSUM62 for two reasons. First, Scorecons provides three categories of substitution matrices, PAM, BLOSUM, and modified PET91, and among the three categories, the BLOSUM matrices match our purpose best, since they were designed to find the conserved regions of proteins (Marketa Zvelebil, 2008). Second, when we aligned the sequences through MAFFT, an automatic computational method was chosen, and through the output log, BLOSUM62 was chosen as the substitution matrix.

Therefore, in order to keep consistent, we continued using this matrix. In terms of the parameters for  $\alpha$ ,  $\beta$ , and  $\gamma$ , we just used the default, 1 for diversity, 0.5 for chemistry, and 3 for gaps.

After getting the trident scores, we can then analyze the distribution of the scores. First, since only seven species are chosen, if more than two gaps occur at the same position, this position is not very informative. In other words, the trident score for a certain position will be omitted, if nearly half of the residues, or even more are gaps (at least 3/7). Then, we need to find the distribution of the trident scores. A simple and common way to analyze the distribution of a variable over a continuous range is to discretize the range into bins, and regard each bin as a group. Then, find the distribution for the bins. Here, bins with equal width are used. We made two types of analysis, one with 20 bins and the other with 10 bins. Specifically, the bin width is set to 0.05 in order to have 20 bins. The ranges are (0, 0.05], (0.05, 0.10], ... , (0.95, 1]. The bin width is set to 0.1 in order to have 10 bins. The ranges are (0, 0.1], (0.1, 0.2], ... , (0.9, 1]. Value 0 is not included in either of the two ways, since no such value is observed in our cases.

Theoretically, we expect there to be three types of trident score distributions, left-skewed, normal bell-curved, and right-skewed, which corresponds to situations where there are too many scores near 0, where the scores are normally distributed, and where there are too many scores near 1, respectively. (However, none of these is true when we get the results, which will be further discussed in the Results Section.)

### 2.3.3 Mapping Conserved and Diverse Regions

After computing the trident scores and discretizing the scores, we plan to find the extremely conserved regions and diverse regions. Since we planned to fit the distribution into a normal distribution, we computed the mean and standard deviation. Then, for each remaining position, Z-score is computed. We defined the residues with trident score=1 as the extremely conserved residues, while those with Z-scores symmetric to those with trident score=1 will be considered as extremely diverse residues. (However, based on the distributions we found, we changed our methodology. Since some proteins are really consistent within the family, very few diverse residues can be found in this way. As a consequence, some meaningful diverse regions may be ignored. Therefore, we lowered the threshold by considering positions with trident scores less than or equal to 0.3 as highly diverse.) Then we map the residues back to the models we have built in Section 2.2, and try to find the relationship between the highly conserved regions or diverse regions and functional regions.

### 2.4 Clustering and Function Mapping

At the same time, we mapped the functional clusters to the models we built. For each protein, we searched the ID of the template used to build the protein in DOMMINO, which is a “database of macromolecular interactions” (Kuang X, 2012). For example, to find the interaction pairs for protein VP30, we searched the template used to build the homology model for this protein, 2I8B. DOMMINO gave us an undefined interaction. Then we downloaded the interaction pairs and compared the alignment with the PDB residue coordinate to calculate the corresponding coordinate in the model. Then, we search for related structures in PDB, and see whether we can find functional annotation from other related sequences. To make sure that the related structure has something in common with our target sequence, we always superimpose the structures together in Chimera. If they

overlap with each other to a great portion, we can use this as a reference. If not, we disregard such sequences.

Finally, after we finish collecting all the related sequences and map the interacting residues to the model we have built, we can then use 1 to mark the residues that are involved in interactions, and count the number of interactions for each residue. We intended to put different weights for residues involved in intra-viral interactions and host-viral interactions. For example, put 1 for intra-viral interactions, and 2 for host-viral interactions. However, in order to avoid ambiguity, intra-viral interactions are separated from host-virus interactions.

## Chapter 3: Results

### 3.1 Structural modeling of viral proteins

Homology models have been built for each protein in Ebola. As a result, in terms of coverage, the models we have built are very consistent with the known structures for Ebola proteins mentioned earlier in Chapter 1, mainly because the models are constructed based on known structures. Below is a table summarizing the models we have constructed. For each protein, the names we gave for the models are recorded, as well as the corresponding templates used to construct the models, the start and end of the model in terms of the loci of our query, and the identity to the template. For each protein we will give a paragraph of evaluation for the model. As described in Chapter 2, five models are generated for each protein. However, for simplicity, only the models with the lowest DOPE are described.

Note that the sequence identity was recorded from the result of PSI-BLAST. For AIE11807 (the first model for VP35), the sequence identity should be even poorer than 19%, because we could not force MODELLER to get the same result as that from NCBI PSI-BLAST. For AIE11811 (ssGP) (the first model for GP), the sequence identity should be higher than 67%, since we chopped off a part of our target sequence and the final model for GP should be more reliable, which will be further discussed in the later parts.

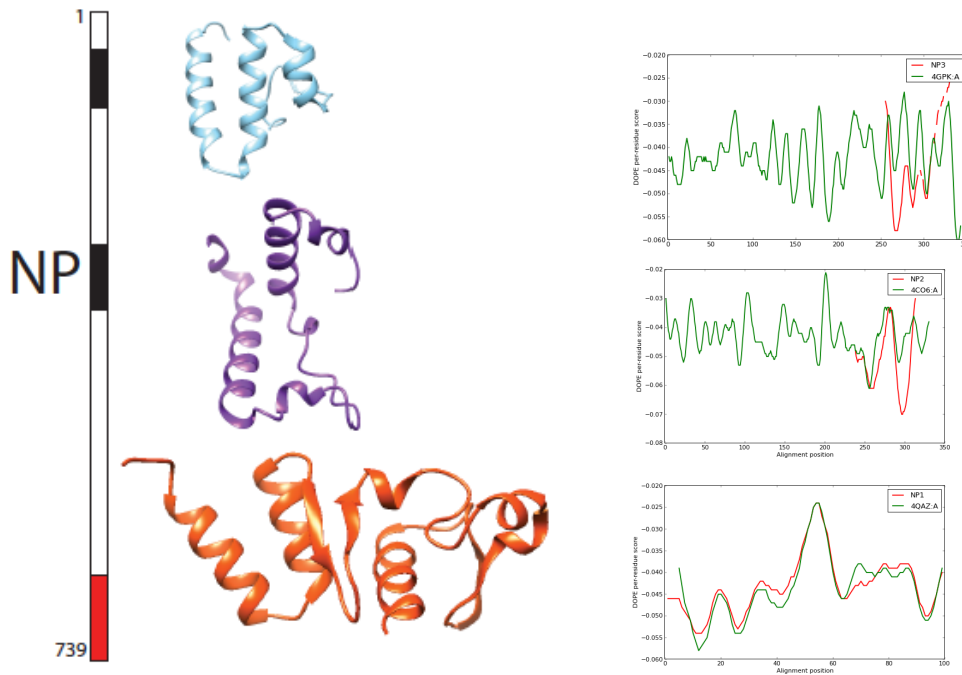
Table 4: Summary of Models

Protein	Name	Start	End	Target-template Seq. ID	Template
NP	NP3	45	110	30%	4GPK:A
	NP2	265	340	25%	4CO6:A
	NP1	641	739	97%	4QAZ:A
VP35	AIE11807	35	262	19%*	4L78:A
	VP35_1	213	340	98%	3FKE:A
VP40	AIE11808	1	326	99%	1ES6:A
				99%	4LDB:B
				75%	3TCQ:A
GP	AIE11811 (ssGP)	1	297	67%*	3CSY:I
	GP2	502	599	99%	3CSY:J
VP30	AIE11812	1	288	98%	2I8B:A
VP24	AIE11813	1	251	99%	4MOQ:A
L	L2	384	558	23%	2FX0:A
	L1	2043	2154	24%	1YXA:A

#### 3.1.1 NP

Three models were constructed that correspond to the following three protein segments: from 641 to 739 (NP1), from 265 to 340 (NP2), and from 45 to 110 (NP3). Among them, the most informative and reliable structure is NP1, since the template for NP1 is a solved structure of the C-terminal domain of NP. According to this paper, experiments show that the C-terminal fragment of NP is a major antigenic determinant, indicating it could be used

in diagnosis (Paulina J. Dziuban´ska, 2014). Figure 13 contains the three models built for NP, as well as the DOPE score per residue for the three models versus the corresponding templates. Overall, the coverage is not very good. Specifically, large portion of N-terminal structure is unsolved.



**Figure 13: NP Models and DOPE Scores**

In Figure 13, we can see a bar on the left side, which represents the length of the amino acid sequence for NP. Next to the bar, there are three DOPE score per residue plots of the protein models we constructed, each corresponding to the black/red portion of sequence on the left. Here, and throughout the paper, the following coloring notation of model coverage is used. Black portion represents the model constructed with low identity template, while red represents the model constructed with high identity template. The white parts are the regions that do not have structural templates. The red line in the DOPE score plots is the DOPE score of the models we constructed, while the green line is the DOPE score of the template. The way to interpret other graphs in the following sections is the same.

### 3.1.2 VP35

Two models were constructed for VP35, one corresponding to the portion spanning from residue 35 to residue 262, and the other from 213 to 340. Furthermore, there is an overlap between the two ranges. If we can superimpose the overlapping part, we can get a hypothetical model for the whole protein. However, given that the identity between the first model and the template (4L78) is only 19%, the first model is not very successful, and therefore will be omitted in later steps regarding conserved region mapping. Figure 14 shows the two models and the corresponding DOPE scores.

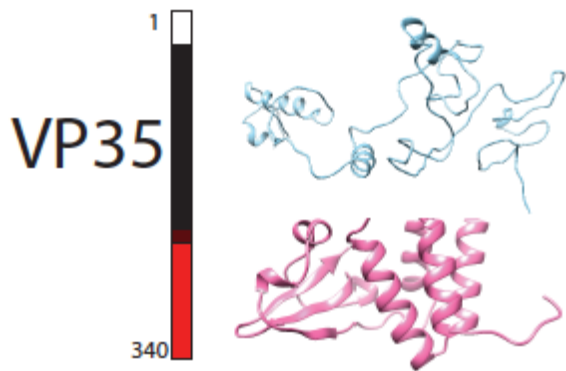
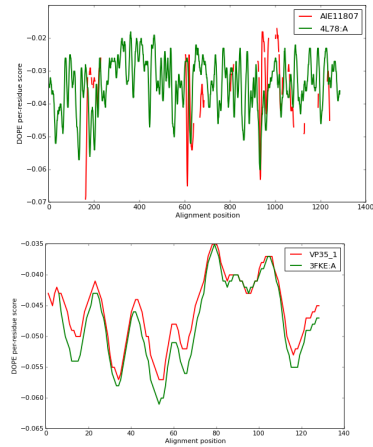


Figure 14: VP35 Models and DOPE Scores



### 3.1.3 VP40

As mentioned in Chapter 2, one model using multiple templates was constructed for VP40. This is the only model that has been built using the multiple template method. According to previous studies, the structure of VP40 has been fully resolved (Dessen A, 2000). Note that the DOPE score here shows the comparison between the basic model and advanced model.

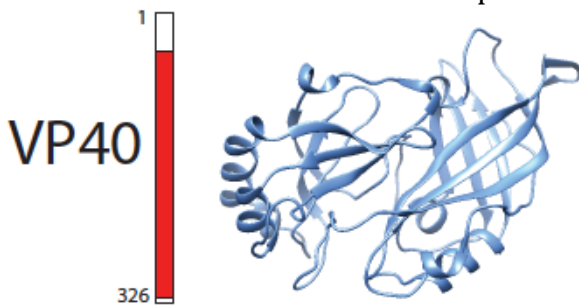
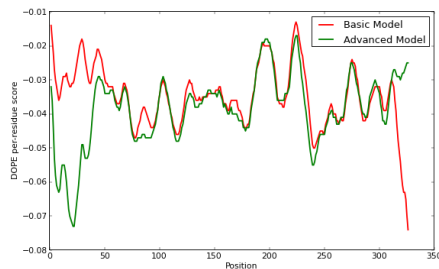


Figure 15: VP40 Model and DOPE Score



### 3.1.4 GP

At first two models were constructed for GP separately based on PDB structure 3CSY, Chains I and J. As shown in Figure 16, there is a gap between the two structures. The

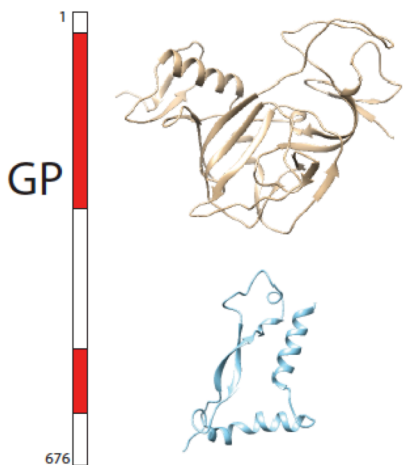
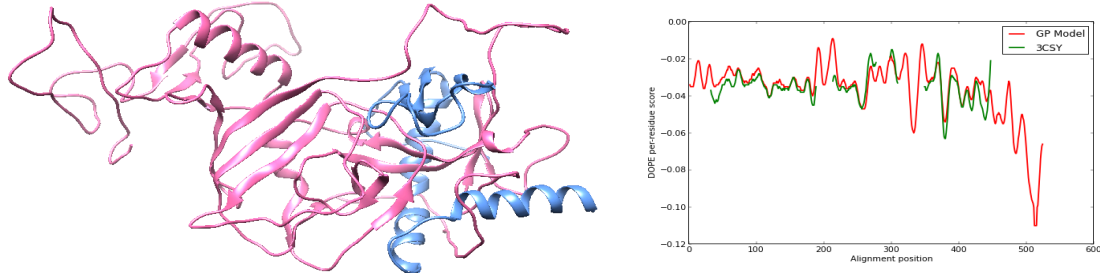


Figure 16: GP Models (Original)



disadvantage of building model in this way lies in that the interaction between the two parts cannot be shown in the structure. After discovering on the structure of the chosen template 3CSY, we find that a composite model with two chains can be constructed using 3CSY as the template, which will better reveal the intra-viral interactions. The newly constructed model is shown in Figure 17. The chain in hot pink is Chain A, while that in cornflower blue is Chain B.

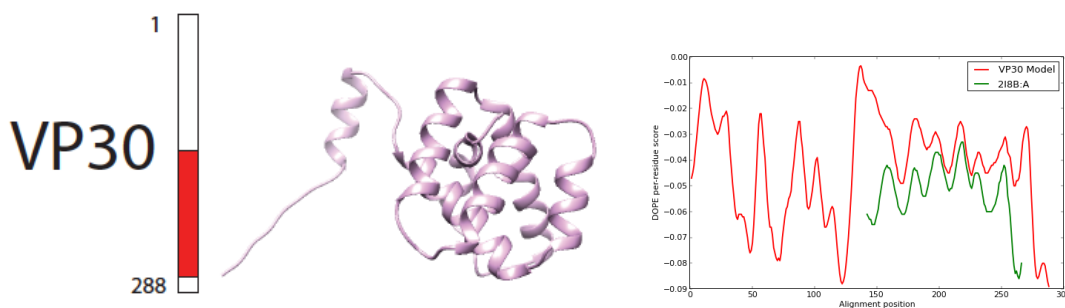


**Figure 17: GP Model (New) and DOPE Score**

Gene products of GP here can be referred as GP<sub>1,2</sub>, and two other gene products can be produced through RNA editing of GP gene: sGP and ssGP. The protein product of ssGP is almost identical to the first of the original models for GP, AIE11811, with only three residues in difference. And this is why the first model of GP is also called ssGP. On the other hand, sGP differs from GP in around 30 residues (Masfique Mehedi, 2011). However, since we did not focus on the mechanism of RNA editing or the byproducts, the models for these two products are not shown here.

### 3.1.5 VP30

One model was constructed for VP30, and large part of the model (the first 144 residues and the last 24 residues) has to be chopped off, since there is no structural information available for that part. Note that although ignored in the model, the DOPE scores for the first 144 residues were plotted in Figure 18.



**Figure 18: VP30 Model and DOPE Score**

### 3.1.6 VP24

One model was constructed for VP24, and this is the model with the highest coverage and identity among all the models that have been constructed. (Weirdly, the name for the template in PDB is called “Ebola Virus VP24 Structure”, while the primary citation used for this protein is called “The Marburg virus VP24 protein interacts with Keap1 to activate the cytoprotective antioxidant response pathway”.)

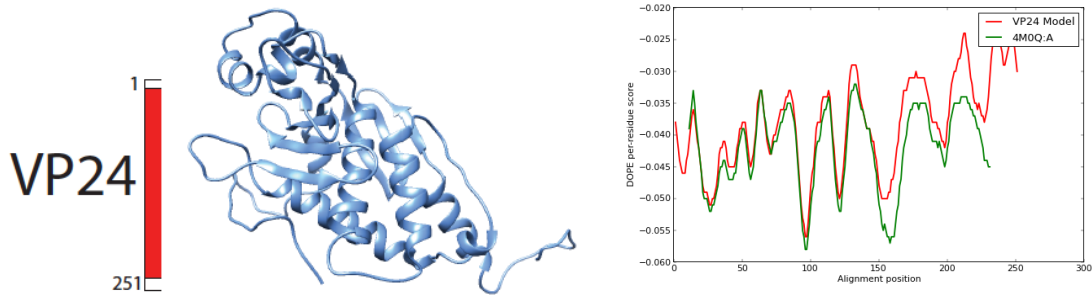


Figure 19: VP24 Model and DOPE Score

### 3.1.7 L

Two models were constructed for L, but unfortunately, the coverage is only around  $(174+111)/2122=13\%$ , and the identity between the templates and the query is very low as well. The result is very consistent with what most resources comment on L protein: Many viruses have the large polymerase protein but the structure has not been solved yet. This can be proven by the fact that many sequences can be found with high coverage and identity if a PSI-BLAST is performed against all sequences, while only a few sequences can be found with low coverage and identity if a PSI-BLAST is performed against PDB structures.

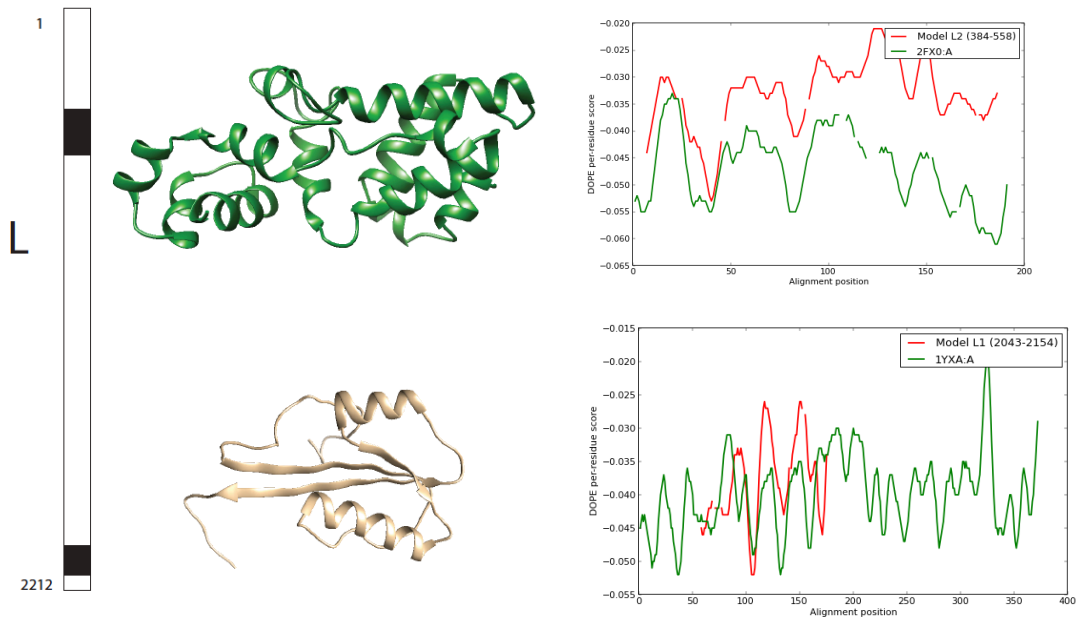


Figure 20: L Models and DOPE Score

### 3.1.8 Summary of Models

In general, by constructing the models, we did not expand our knowledge about the protein structures very much, since we did not have a good amount of templates for each protein. So far, the most reliable model should be VP24, since it has the longest coverage and highest identity. Then comes VP40. The coverage and sequence-template identity for VP30 is very similar to GP, both with half of the protein structure unknown. Although it seems that we have a good coverage of VP35, because the first model of VP35 was not very successful due to low identity to the template, we still need more information about VP35. For NP, although we have three separate models for different parts, how the three parts connect with each other is still a myth. For L, if we compare the known portion with the

unknown portion, it seems that only a tenth of the whole protein structure can be predicted with a high uncertainty.

## 3.2 Family-Wise Sequence Alignment of *Filoviridae* Proteomes

### 3.2.1 Data Selection

As described in the methodology part, to discover the evolutionary pattern within the family of *Filoviridae*, we decided to perform a family-wise sequence alignment and then evaluate the result. One representative genome was selected from each species within the family *Filoviridae*. The genomes selected are the reference genomes from NCBI, which are listed in Appendix C. Genome Selection for Family-Wise Sequence Alignment, together with the sequence ID for each protein.

### 3.2.2 Trident Score Distribution

Then, for each protein among the seven main proteins in the Ebola genome, sequence alignment is performed. Most *Ebolavirus* sequences align well with each other, while Marburg virus sequences and *Lloviu cuevavirus* seem to have difficulties in aligning perfectly with all other sequences. To quantitatively evaluate the alignment, trident scores were computed, and then the data were discretized into bins to find out the distributions. Figure 23 is composed of distribution charts for all seven proteins. Since the pattern in the 20-bin method is even more obscure than that in the 10-bin method, here only the 10-bin method result is shown. From the charts, we can see that most distributions follow a bimodal or even trimodal trend, if we do not take GP (8) Trident Score Distribution into account.

Note that there are two charts for one protein, GP, one labeled as GP (7) and the other GP (8). This is due to the genome of *Lloviu cuevavirus* having two protein sequences affiliated with GP, YP\_004928138.1 and YP\_004928139.1, both of which fall into the normal loci for gene GP in other species in *Filoviridae* family. GP (8) consider the two sequences separately. Therefore, there are only a few residues with trident score=1, which should result from false alignment, since there is no overlap in the two sequences from *Lloviu cuevavirus*. Figure 21 and Figure 22 are the three possible Open Reading Frames (ORFs) for GP1 and GP2 in *Lloviu cuevavirus*. In both cases, Frame 1 represents the final protein sequence. However, the end of 5'3' Frame 3 in Figure 21 is the same as the first several residues in 5'3' Frame 1 in Figure 22, which means there is a frame shift during translation, and therefore the two sequences would not be combined into one. A naïve way to deal with this problem is to simply concatenate the two sequences together into one, and align it with other sequences. And the result for the trident score distribution is shown in GP (7) Trident Score Distribution.

5'3' Frame 1

Met VPTYYPYSSLLDWRPPNPNTLPWILNLVVFYTIAWLPGGVSGIPLGLLGNNISITQTVVD  
NVVCKEHLATTDQLQAIGLGLGLEGLGEHADLPTATKRWGFSDVIPKIVGYTAGGEWVEN  
CYNLEITKKDGHPCLPSPPTGLLGYPRCRYVHRAKAGGPCPGGNAFHKKHGSFFLYHG  
Met ASTVIYHGVTFTTEGTIAFLIVPKDAPRLKAGLGTGFSHQAENQNPNNQFRTTLDY  
DV Met SPW Met DNATFFFRAREDTS Met LIQTRYPPANLELVQERLANLTGDQADPSK Met  
EEIVAEVLTLELGDWSGWTTKKTAVQTIIRLNPSASGSTKDKTGQKP Met TDHQEFIL  
QPHSAVGQPCLWNILRTPGRNPARRHRRETPT Met SITAAPGSGYKPYIAIPLVKFR  
CHWEGLRHVCRRYPSWVQ Stop

5'3' Frame 2

WCPPTRTAAAY Stop IGDHHQTPYHGSSTLWSFIP Stop PGCPGESQEFHVCWETTASP  
LSWT Met Stop CARNTLPQQISYRLLDWD Stop RGLVN Met LTSRLPPSDGVFDL Met SSQKS  
WDTP LGNGWKTATILKSPRK Met VILASPARQLAYLAIPDAA Met STEPKEQALAQVG Met  
LSTN Met VLSFCTTVWLLQ Stop FI Met V Stop PLRKAQLLS Stop LSRR Met HPVSRQGLEQDS  
VIKQRTKTQTTNFEQQL Stop I Met Met Stop Stop VLGWT Met LPSSFERKTHQC Stop SKQGT  
LQI Stop SLFKKDWLILPEIKLIHQRWKRLSLRF Stop HWSSVIGPVGQLKPKQYKPYG  
Stop ETLHQHLVQPRTRLARSP Stop RIIRSSSSNLILLDNPASGTFEFELRGGTLHEGTG  
GKHHQQCP SLLLLGQDTSRTSRQYLW Stop NFDAIGRVFG Met CVVDTPPGFS

5'3' Frame 3

GAHLPVQQPIRLETTTKHPT Met DPQPCGLLYHSLAARGSLRNSTRFVKGQQHHPNCR  
GQCSVQGT PCHNRSATGYWIGTRGAW Stop TC Stop PPDCHQA Met GFSI Stop CHPKNRGI  
HRWG Met GGKLLQS Stop NHQERWSSLPPQANWLTWLS Met PLCPQSQRSRPLRWE  
CFPQTWFFLSVPRYGFYSNLSWCNLYGRHNCFPNCEGCTPSQGRAWNRIQSSSRE  
PKPKQPI SNNNFRL Stop CNESLDGQCYLLSSEGRHINANPNKVPSSKSRACSRKIG  
Stop SYRRSS Stop SIKDGRDCR Stop GFDIGAR Stop LVRLDN Stop KNRSTNHTAKKPTFSIW  
FNQGDWPEAHDGSSGVHPPTS FCCWTTLPLEHSSNSGAEPCTKAPAGNTTNNVH  
CCSWVRIQAVHPGNTSGEIS Met PLGGSSACVSSIPLLSV

Figure 21: 3 Possible ORFs of GP1 in Lloviu Cuevavirus

5'3' Frame 1

Met PLGGSSACVSSIPLLSV SNNSSIQELETSSKSATELTTPINHSQSLQLASVTNTP  
PTTQSKSWTVDYNNTTPT Met DPTTILTPDTATIPPNNSSDHNATTTSKTRRRRQVNP  
VPPTITQQTSTSINTSHHPN Met TTQLARHPSVQTR Met QNPSCNPNLRYWTSRE Met SN  
AGGLAWIPWIGPGIEGGITDGI Met EHQNITVQQLRELANTTTKALQLFLRATTELRTYS  
LNRHAIDFLLQRWGGTCRILGPNCCIEPHDWSANITAEINHIREDLNHHEIQPSQDPS  
FWTGWQQWIPTGASALGILAILALICLCRITR Stop

5'3' Frame 2

CHWEGLRHVCRRYPSWVQ Stop ATIVQYRSLRLHLKVQQN Stop QLPSITPNHYSSHPSQ  
TPPHRQHSPSPGQLTTTTQRQPWIPQQY Stop RHPTPQPFLTTHLITTPQQQAKQDEG  
DRSTQCPQRSPNKPLQASIPPTPI Stop QHS Stop QDIRVCKQGCKTPAVIPTLDTGQAG  
R Stop V Met LGGLHGFHGLDQGLREGSQTG Stop WSIRTQLSVSYGSSRTPLLKPYSFSSG  
LPLSSEPTLSSTA Met RLTFYYSVGVVPAESLAQTAVSNL Met IGLPTLRLR Stop IILEKIS  
Stop TI Met RSNLLKTPPFGLDGNSSGSQQEPVLSESSWQY Stop P Stop FVCAE Stop HD

5'3' Frame 3

AIGRVFG Met CVVDTPPGFSEQQ Stop FNTGA Stop DFI Stop KCNRIIDNSHQSLPITTARIRH  
KHPHTDNTVQVLDS Stop LQQHNANHGSHNNTDDTRHRNHSP Stop QLI Stop SQRHNNKQ  
NKTKETGQPSAPNDHPTNLYKHQYLPQPQYDNTVSKTSECANKDAKPQL Stop SQP Stop  
ILDKPGDE Stop CWGAC Met DS Met DWTRD Stop GRDHRRDNGASEHNCLSVTGAREHHY  
Stop SPTAFPPGYH Stop APNLLYPQPPCD Stop LSTTALGWYLQNPWPKLLYRTS Stop LVC  
QHYG Stop DKSY Stop RRYPEPS Stop DPTFSRPLLLDW Met ATVDPNRSQC SRNHGPNISL  
DLSVQNTT Met

Figure 22: 3 Possible ORFs of GP2 in Lloviu Cuevavirus

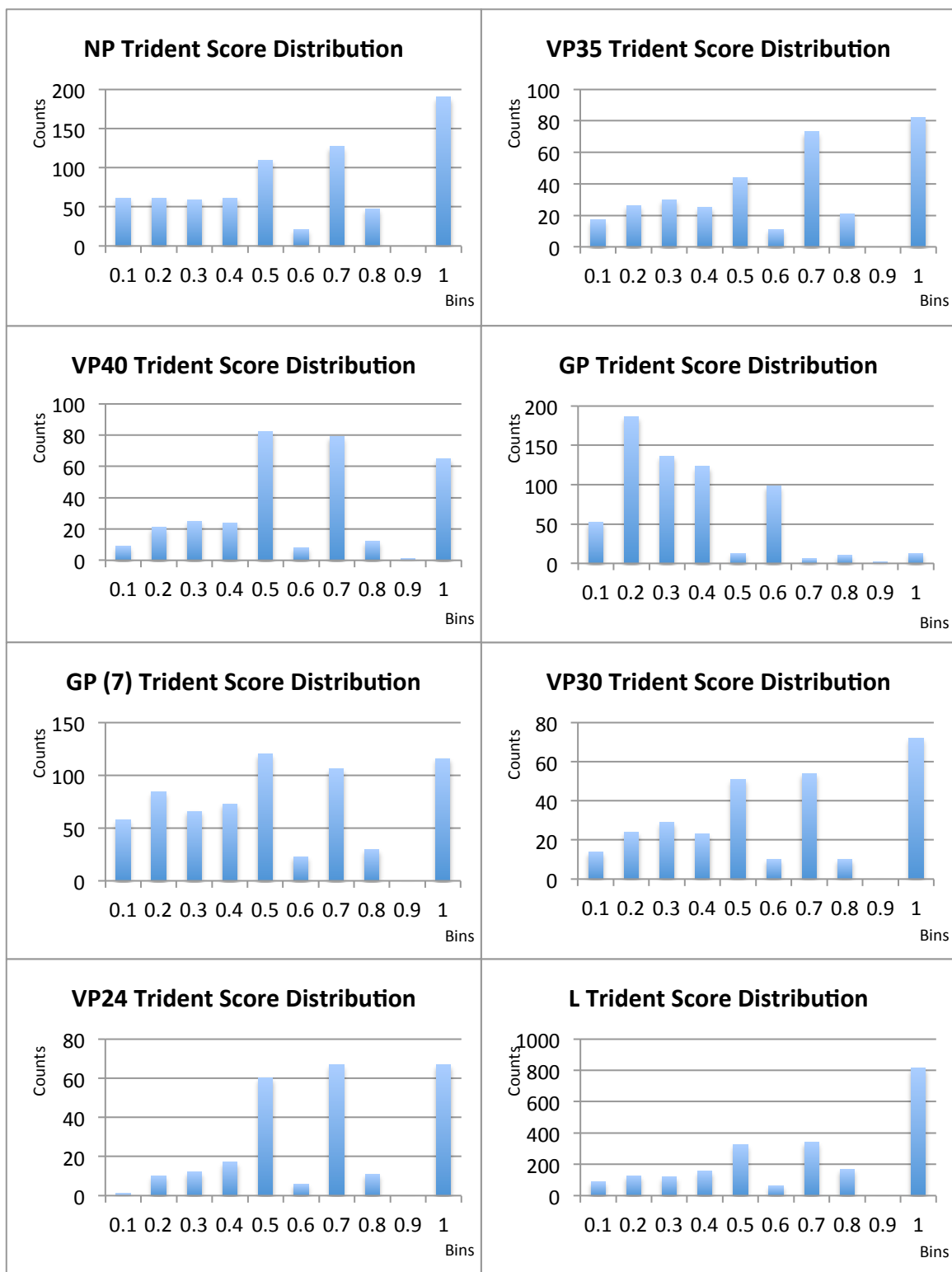


Figure 23: Trident Score Distributions of 7 Proteins

### 3.2.3 Mapping Conserved and Diverse Regions

Then we mapped the extremely conserved residues and extremely diverse residues back to the models we constructed in previous step. We rendered the models using the following color code: The model itself is shown in light gray surface mode; Residues with trident score equal to 1 are considered as extremely conserved ones, which are colored in blue;

Residues with trident score less than or equal to 0.4 are considered as extremely diverse ones. Among them, if no gaps are found at a certain locus in the alignment, the corresponding residue is colored in red; if no more than two gaps are found at a certain locus in the alignment, the corresponding residue is colored in magenta. The residues that fall into each category (conserved, diverse without gaps, or diverse with up to 2 gaps) in each model are recorded in Appendix D. Conserved and Diverse Residues Mapping.

### 3.2.3.1 NP

As mentioned before, for NP, three models were constructed separately. The first two models were not very reliable based on the identity with the template sequences. However, we can see a large amount of conserved regions in these models. On the other hand, the last model was considered more reliable since it had a higher identity to the template, but it contains more diverse residues.

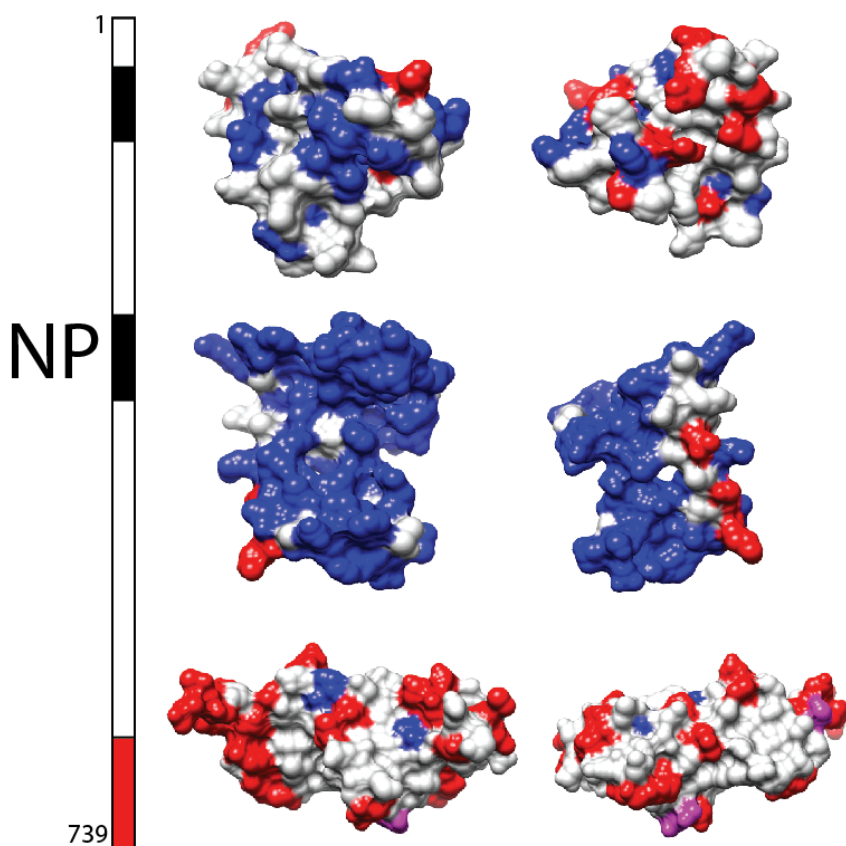


Figure 24: NP Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

### 3.2.3.2 VP35

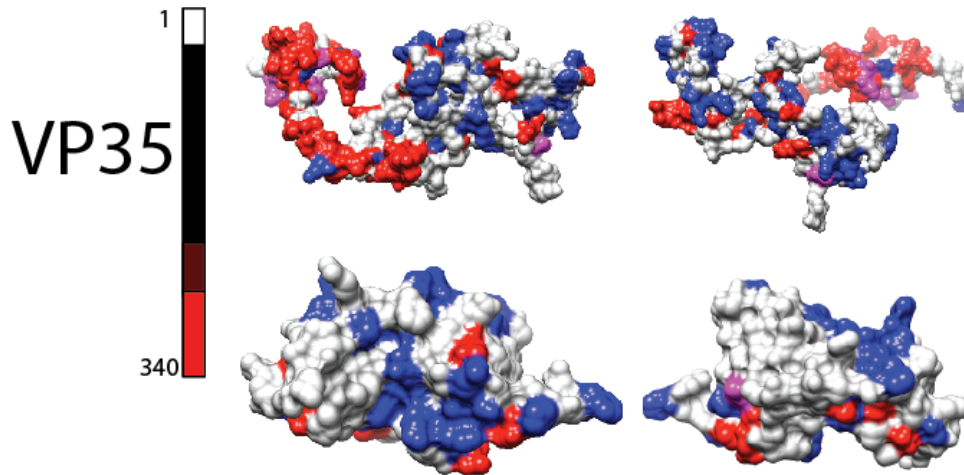


Figure 25: VP35 Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

For VP35, two models were constructed and there was an overlap between the models. However, the first model was not considered successful since it had a low identity to its template. But from the color code, we can see a large number of diverse residues. As for the second model, we can see clusters of conserved residues on the surface of the molecule.

### 3.2.3.3 VP40

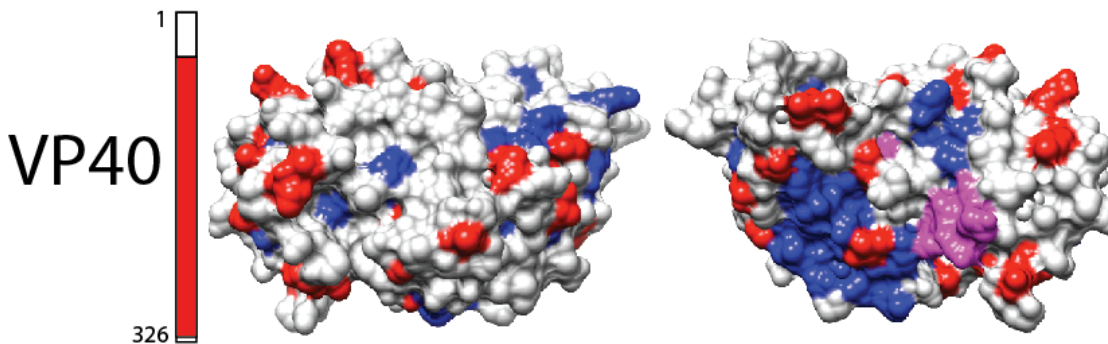


Figure 26: VP40 Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

For VP40, with the long free end being chopped off, we can see a cluster of conserved residues, and several diverse residues around the surface.

### 3.2.3.4 GP

On the surface of GP, more diverse residues can be found than conserved residues. This is mainly because Lloviu cuevavirus does not align to all other species very well, and since we only defined residues with trident score strictly equal to 1 as “extremely conserved” ones, therefore, there are not many extremely conserved ones. We can also get this conclusion from the distribution charts. As we can see, the distribution of trident scores for GP (7) is different from those for other proteins, in that for proteins like L and NP, more residues fall into the bin of (0, 1], while for GP, more residues fall into the bin of (0.4, 0.5].

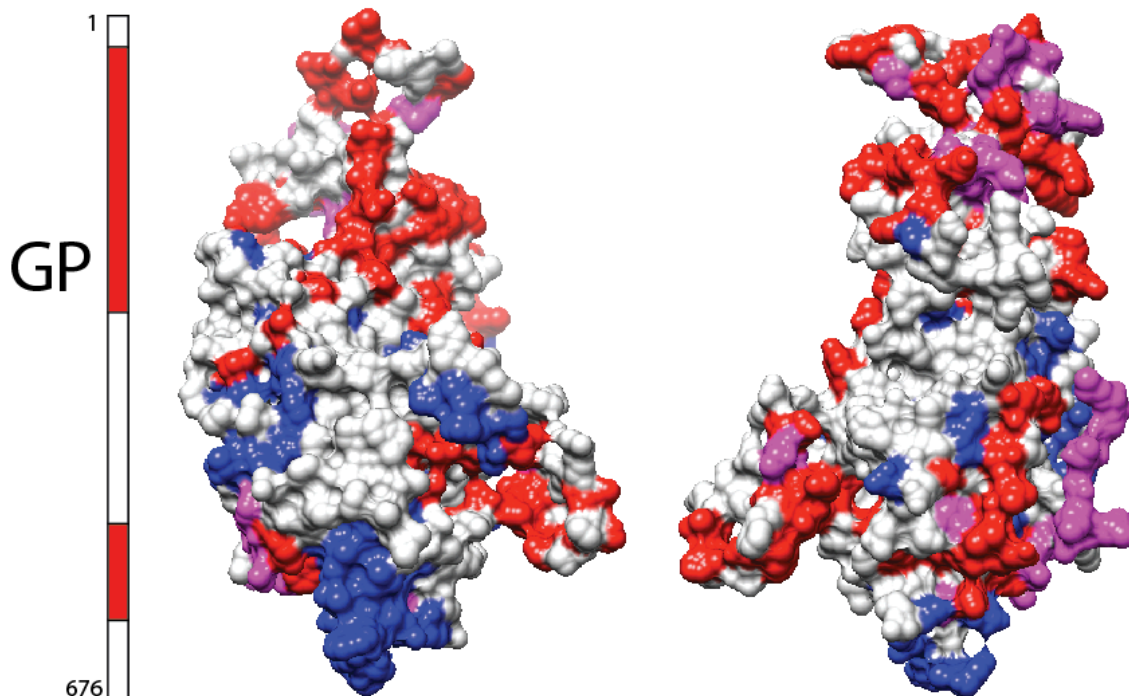


Figure 27: GP Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

### 3.2.3.5 VP30

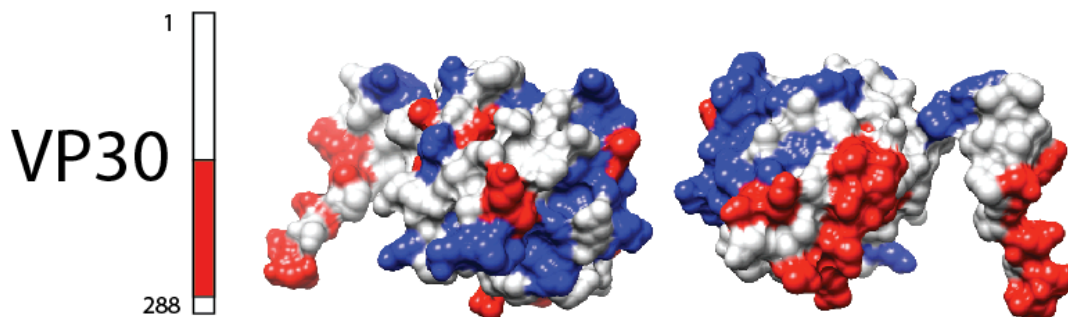


Figure 28: VP30 Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

For VP30, only half of the whole protein was modeled. As shown in Figure 28, one region can be considered as highly conserved while two clusters can be considered highly diverse.

### 3.2.3.6 VP24

For VP24, the model is almost complete. However, the pattern of conserved and diverse regions seems to be random.



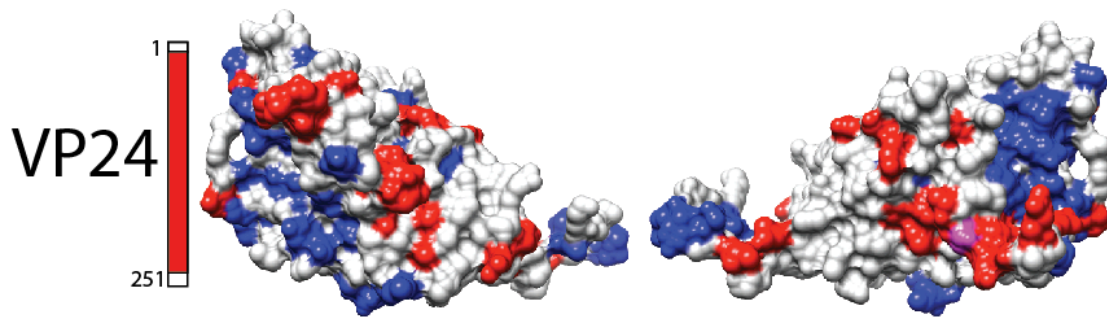


Figure 29: VP24 Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

### 3.2.3.7 L

For L, as shown in Figure 30, more residues are found to be highly conserved in the first model, while more are found to be diverse in the second model. It is note worthy that there are two putative conserved domain in the protein L, Mononeg\_RNA\_pol superfamily and Paramyx\_RNAcap superfamily, as shown in Figure 31.

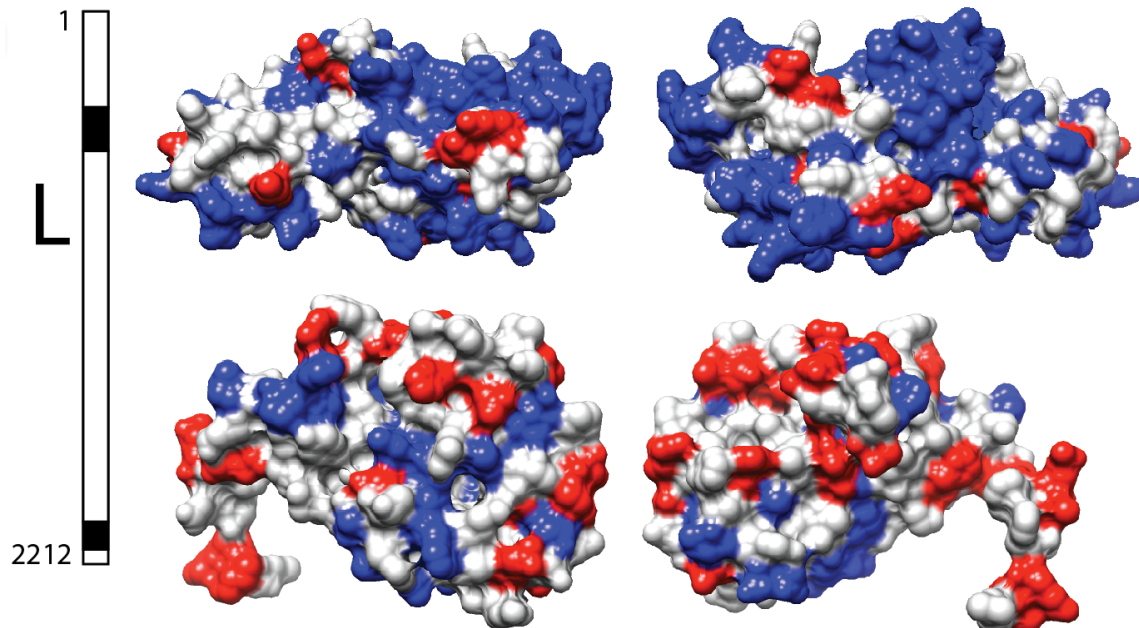


Figure 30: L Conserved and Diverse Regions Mapping. Colored in red are extremely diverse residues (trident score of 0.4 and lower with no gap); colored in magenta are diverse residues (trident score of 0.4 and lower with up to 2 gaps); colored in blue are extremely conserved residues (trident score of 1).

## 3.3 Result for Clustering and Function Mapping

In order to test our hypothesis that the evolutionary pattern within the family *Filoviridae* is potentially related to the functional clusters, we then went on and mapped all the potential binding sites by finding related structures in DOMMINO and aligning the functional sites

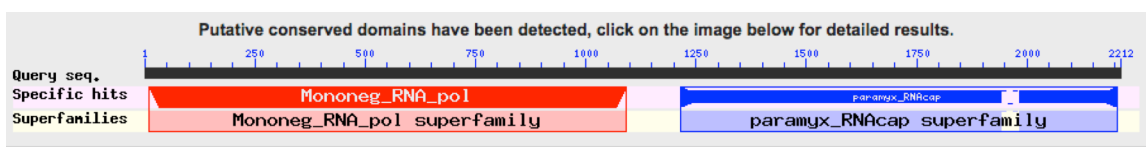


Figure 31: Screen Shot of Putative Conserved Domains in L

**Table 5: Color Code for Figures in Section 3.3**

Color	Conserved	Diverse	Intra-viral	Host-viral
Blue	√			
Red		√		
Yellow			√	
Orange				√
Cyan	√		√	
Cyan	√			√
Cyan	√		√	√
Magenta		√	√	
Magenta		√		√
Magenta		√	√	√

with our models. Detailed study notes can be found in Appendix E. Model Check and Functional Annotation, which records the related structures we found for each protein. Then, based on the alignment, we recorded the score for each residue to be involved in intra-viral interactions or host-viral interactions separately, and mapped the binding sites to the models we built using the following color code: The base model is rendered in light gray; the residues involved in intra-viral interactions are colored in yellow, while those involved in human-viral interactions are colored in light green. Now, we will show the binding site mapping result for each protein. And here is a table for the color code in general. Basically, we will render the models in the following way. If a residue is only classified as conserved, it will be colored in blue; if a residue is only classified as diverse, it will be colored in red; if a residue is only involved in intra-viral functional sites, it will be colored in yellow; if a residue is only involved in host-viral functional sites, it will be colored in orange. If a residue is both conserved and involved in functional sites (either intra-viral or host-viral or both), it will be colored in cyan. If a residue is both diverse and involved in functional sites (either intra-viral or host-viral or both), it will be colored in magenta. Here, we do not distinguish within the diverse residues based on the number of gaps as we did in Section 3.2, since we altered our focus here.

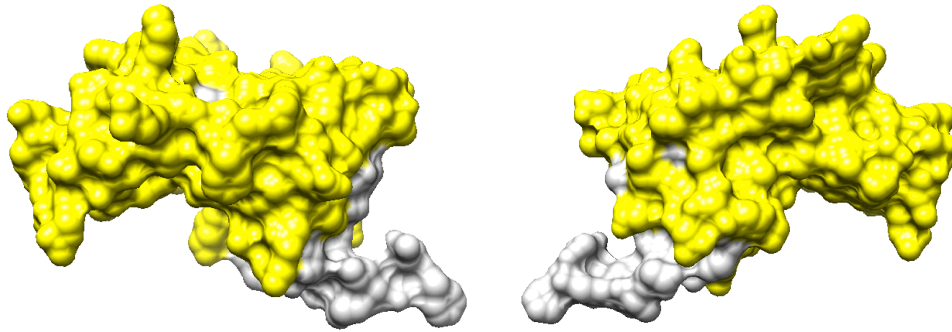
### 3.3.1 NP

For NP, it shows that two solved crystal structures can be found in PDB, 4QAZ and 4QB0. We tried DOMMINO, SUPERFAMILY, and PDB to find its related structures. Unfortunately, no functional domains have been found.

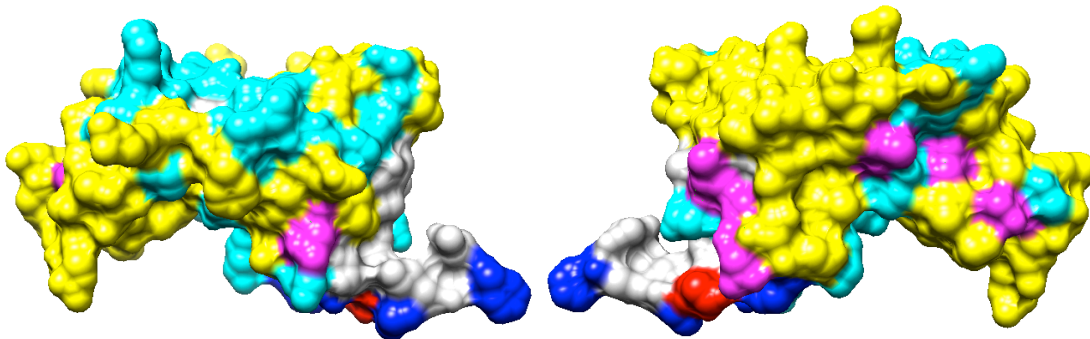
### 3.3.2 VP35

For VP35, although two models were built, since the first model had a really poor identity to the template, here only the second model is considered (what we called “basic”, or “VP35\_1”). For this model, we chose 3FKE as its template, which contains 1 intra-viral interaction. Also, 3KS8, 4LG2, 3L25, 3L28, and 4GHL are considered as related structures. Only intra-viral interactions and RNA-RNA interactions are found in these structures. Figure 32 shows the model in surface mode with all the residues involved in intra-viral interactions highlighted in yellow. Furthermore, among all 128 residues in the model, there are 49 residues that are categorized as extremely conserved ones, and 34 of them overlap with those that are potentially involved in intra-viral interactions. On the other hand, 9 residues are categorized as extremely diverse ones, and 8 of them overlap with those

involved in intra-viral interactions. Figure 33 is the model with conserved, diverse regions mapped together with the intra-viral functional sites.



**Figure 32: Residues in VP35 that Are Potentially Involved in Intra-viral Interactions.** In this figure, only the residues involved in intra-viral interactions are highlighted in yellow.



**Figure 33: Conserved and Diverse Regions in VP35 Mapped with Intra-viral Interactions.** Colored in red are extremely diverse residues (trident score of 0.4 and lower), and only 1 residue is colored in red; colored in blue are extremely conserved residues (trident score of 1), and 15 residues belong to this category; colored in magenta are the common residues between the diverse and intra-viral functional sites, and 8 residues belong to this category; colored in cyan are the common residues between the conserved and intra-viral functional sites, and 34 residues belong to this category.

### 3.3.3 VP40

For VP40, since we constructed the model by the advanced method, we already have three templates, which are 3TCQ, 4LDB, and 4LDI/4LDD. The three templates share the SCOP family 50013. In fact, there are 7 proteins under this SCOP family, but they are highly redundant, and only the templates are studied, since they can represent monomer, tetramer, and dimer respectively. Figure 34 shows the residues potentially involved in intra-viral interactions. Among all 326 residues in the model, 65 residues are categorized as extremely conserved ones, and 41 of them overlap with those that are potentially involved in intra-viral interactions. (But since some residues are not in the model, only 61 residues from the conserved are shown. Among them, 20 are colored in blue, and the rest 41 residues are colored in cyan because they overlap with intra-viral functional sites.) On the other hand, 79 residues are categorized as extremely diverse ones, and 25 of them overlap with those involved in intra-viral interactions. (But 32 out of the 79 residues are missing in the model. So, 47 diverse residues are shown. Among them, 22 residues are colored in red, and the rest 25 residues are colored in cyan because they overlap with the intra-viral functional sites.)

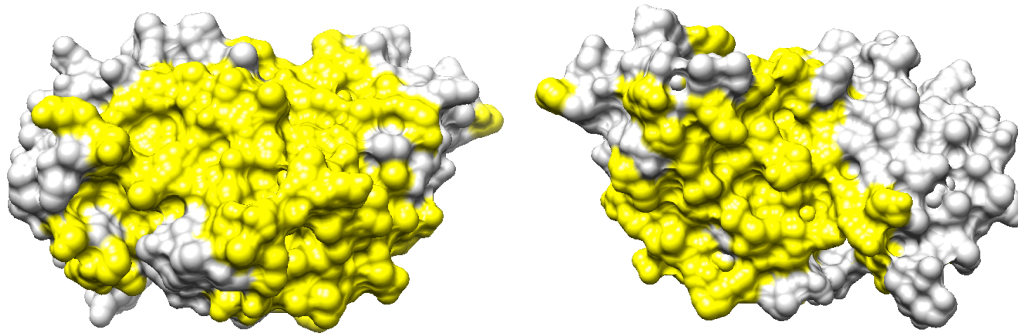


Figure 34: Residues in VP40 that Are Potentially Involved in Intra-viral Interactions. In this figure, only the residues involved in intra-viral interactions are highlighted in yellow.

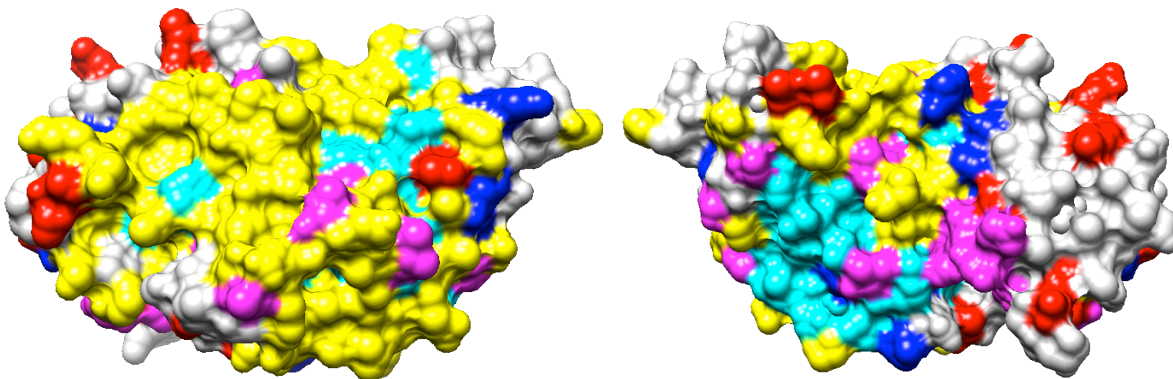
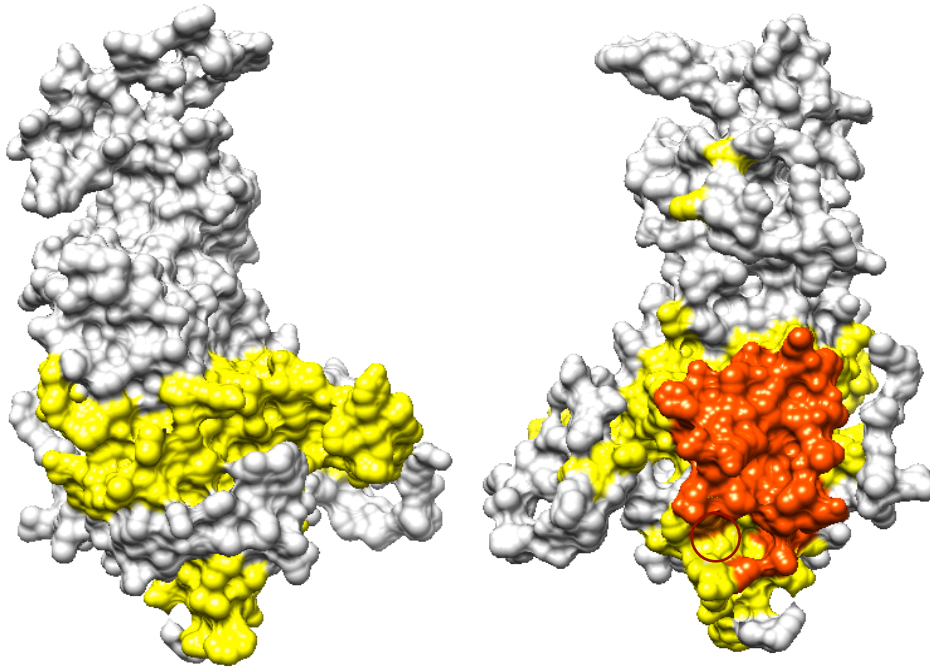


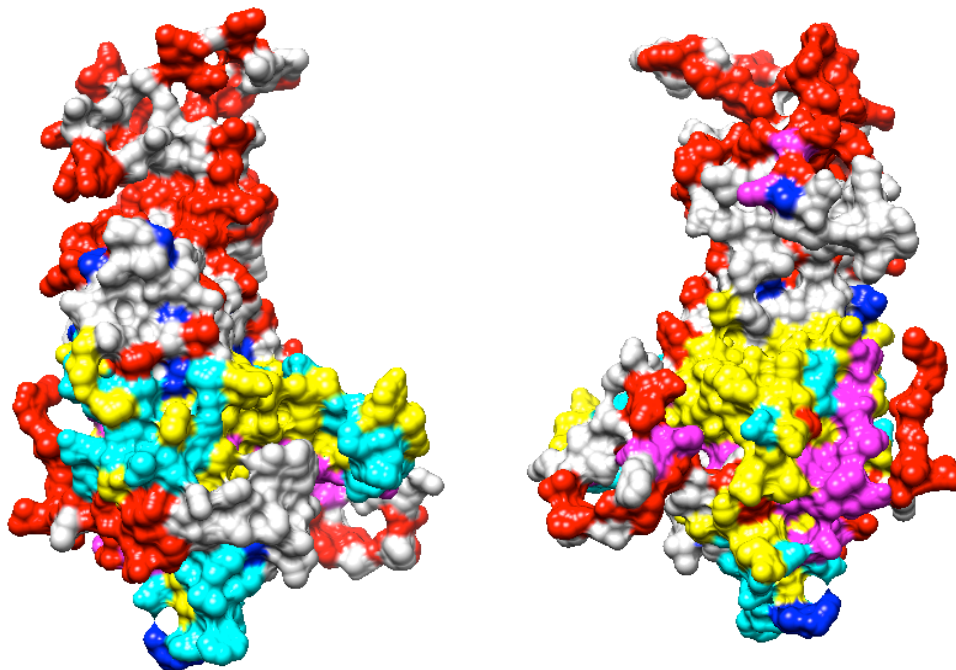
Figure 35: Conserved and Diverse Regions in VP40 Mapped with Intra-viral Interactions. Colored in red are extremely diverse residues (trident score of 0.4 and lower), and 22 residues are colored in red; colored in blue are extremely conserved residues (trident score of 1), and 20 residues belong to this category; colored in magenta are the common residues between the diverse and intra-viral functional sites, and 8 residues belong to this category; colored in cyan are the common residues between the conserved and intra-viral functional sites, and 34 residues belong to this category.

### 3.3.4 GP

For GP, as mentioned previously, a model based on 3CSY has been built. According to the cited paper, 3CSY is the crystal structure of glycoprotein in Ebola in the form of trimeric prefusion with an antibody from a human survivor. This reveals that we have found host-viral interactions in GP. We also found another three related structures, 3S88, 3VE0, and 1EBO. However, after reading the paper about 1EBO, we found that 1EBO is a hybrid trimer between GP2 and modified GCN4 (Winfried Weissenhorn, 1998), so as a result, we did not consider this structure any more. As shown in Figure 36, the residues involved in intra-viral interactions are in yellow. Those involved in host-viral interactions should be colored in orange. However, since all residues belonging to this category are also involved in intra-viral interactions expect for Residue 39. Therefore, to emphasize the overlap, residues involved in both interactions are colored in orange red. However, residue 39, which is only involved in host-viral interaction, is colored in orange, and circled in dark red. This overlap between the two types of interactions does not give us a significant difference between the intra-viral interactions and host-viral interactions as expected.



**Figure 36: Residues in GP that Are Potentially Involved in Functional Sites.** In this figure, the residues involved in intra-viral interactions are highlighted in yellow, while those only involved in host-viral interactions are supposed to be colored in orange according to the color code mentioned before. Residue 39 is the only one involved in host-viral interaction. So, we colored it in orange and circled it with dark red. The overlap between the two kinds of functional sites are colored in orange red. In GP, all residues involved in host-viral interactions are involved in intra-viral interactions as well, with the exception of Residue 39 mentioned before.



**Figure 37: Conserved and Diverse Regions in GP Mapped with Intra-viral Interactions.** Colored in red are extremely diverse residues (trident score of 0.4 and lower); colored in blue are extremely conserved residues (trident score of 1); colored in yellow are residues involved in intra-viral interactions; colored in magenta are the common residues between the diverse and intra-viral functional sites; colored in cyan are the common residues between the conserved and intra-viral functional sites.

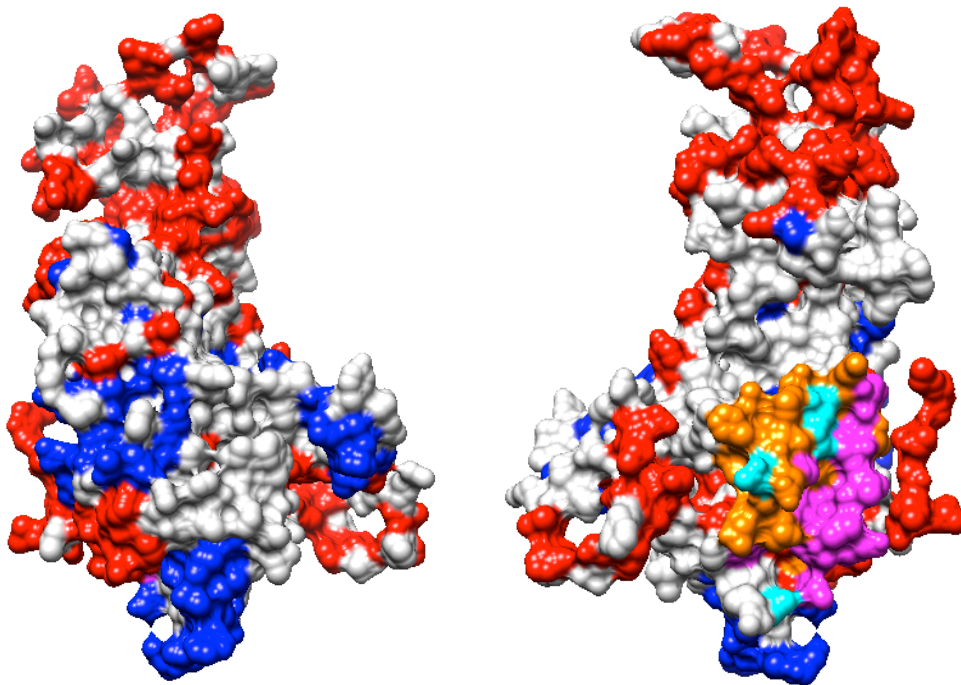


Figure 38: Conserved and Diverse Regions in GP Mapped with Host-viral Interactions. Colored in red are extremely diverse residues (trident score of 0.4 and lower); colored in blue are extremely conserved residues (trident score of 1); colored in orange are residues involved in host-viral interactions; colored in magenta are the common residues between the diverse and host-viral functional sites; colored in cyan are the common residues between the conserved and host-viral functional sites.

### 3.3.5 VP30

We built the model for VP30 using the template 2I8B. Another structure considered related to VP30 is 3V70. We found a SCOP family 54237 in 3V70, but none of the organisms inside this family is closely related to Ebola virus except 3V70. 72 out of 288 residues are classified as highly conserved residues. However, only 22 of them overlap with the residues involved in intra-viral interactions. On the other hand, 90 residues are classified as diverse, but 18 of them overlap with those involved in intra-viral interactions.

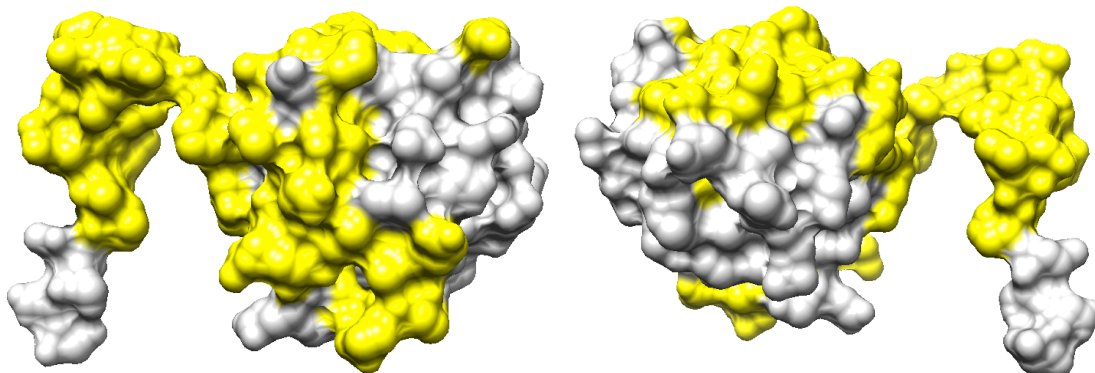


Figure 39: Residues in VP30 that Are Potentially Involved in Intra-viral Interactions. In this figure, only the residues involved in intra-viral interactions are highlighted in yellow.

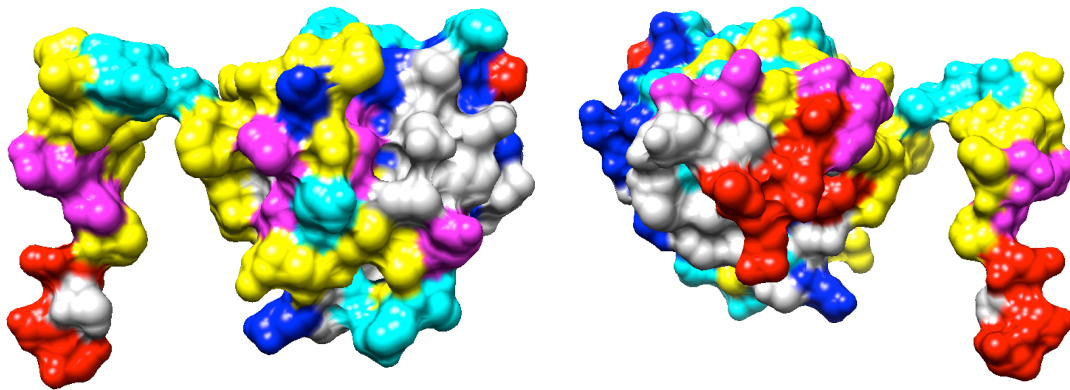


Figure 40: Conserved and Diverse Regions in VP30 Mapped with Intra-viral Interactions. Colored in red are extremely diverse residues (trident score of 0.4 and lower); colored in blue are extremely conserved residues (trident score of 1); colored in yellow are the residues involved in intra-viral interactions; colored in magenta are the common residues between the diverse and intra-viral functional sites; colored in cyan are the common residues between the conserved and intra-viral functional sites.

### 3.3.6 VP24

Besides our template 4M0Q, which contains one intra-viral interaction, we used 4U2X as a related structure to VP24 as well, which reveals both intra-viral interactions and host-viral interactions. As shown in Figure 41, unlike the situation in GP, residues involved in intra-viral interactions are almost mutually exclusive to those involved in host-viral interactions, with the exception of 5 residues highlighted in orange red.

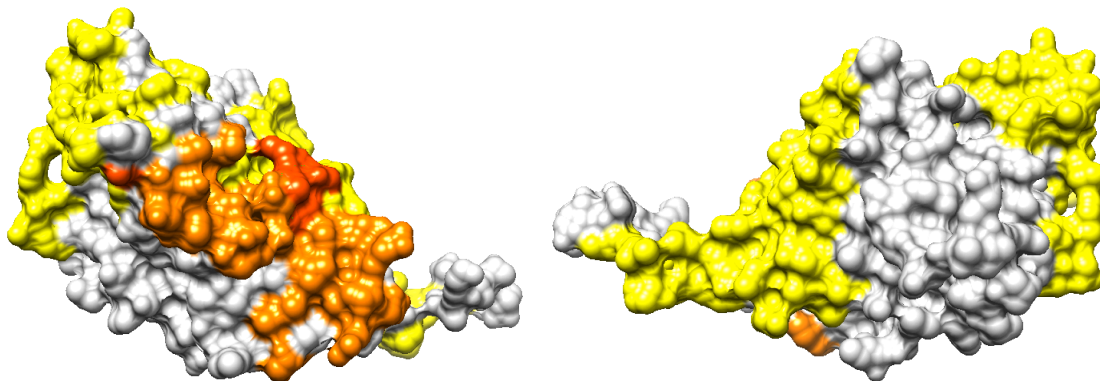


Figure 41: Residues in VP24 that Are Potentially Involved in Functional Sites. In this figure, the residues involved in intra-viral interactions are highlighted in yellow, while those involved in host-viral interactions are colored in orange. Residues involved in both host-viral interaction and intra-viral interactions will be colored in orange red.

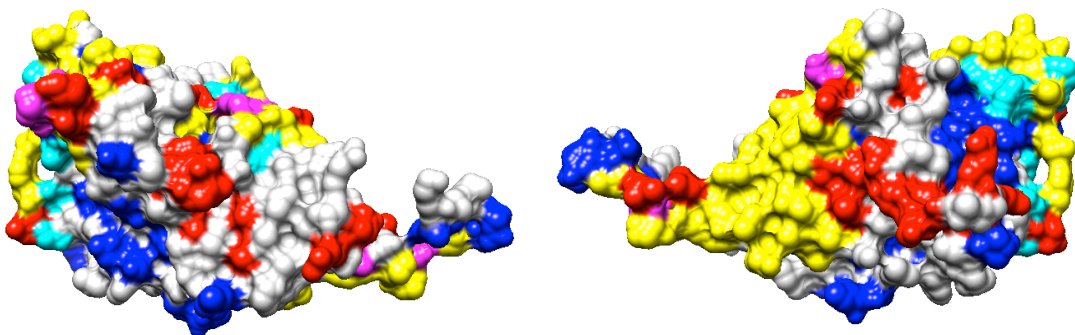


Figure 42: Conserved and Diverse Regions in VP24 Mapped with Intra-viral Interactions. Colored in red are extremely diverse residues (trident score of 0.4 and lower); colored in blue are extremely conserved residues (trident score of 1); colored in yellow are residues involved in intra-viral interactions; colored in magenta are the

common residues between the diverse and intra-viral functional sites; colored in cyan are the common residues between the conserved and intra-viral functional sites.

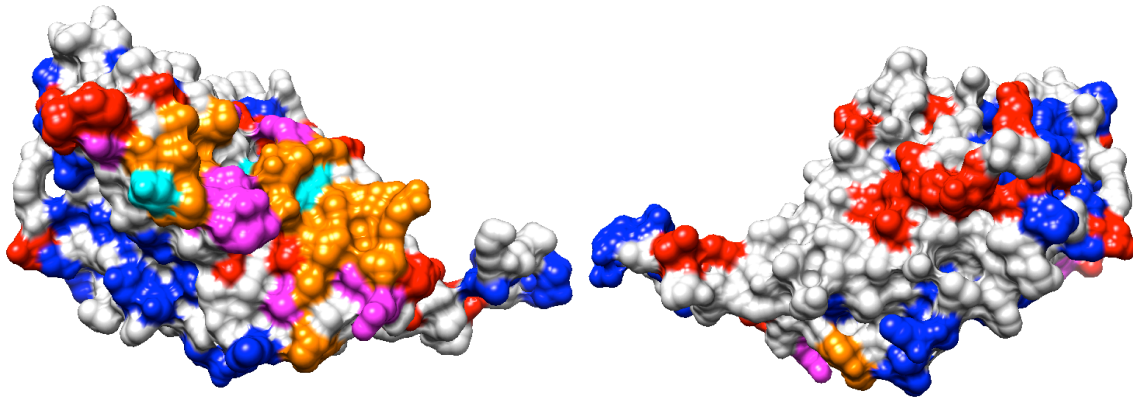


Figure 43: Conserved and Diverse Regions in VP24 Mapped with Host-viral Interactions. Colored in red are extremely diverse residues (trident score of 0.4 and lower); colored in blue are extremely conserved residues (trident score of 1); colored in orange are residues involved in host-viral interactions; colored in magenta are the common residues between the diverse and host-viral functional sites; colored in cyan are the common residues between the conserved and host-viral functional sites.

### 3.3.7 L

Since we did not get good templates for L, we could not perform a legitimate search in DOMMINO. Instead, we chose PDB advanced searching method for L. However, no significant result has shown.

(The residues that are potentially involved in functional sites can be found in Appendix F. Functional Residues Mapping.)



## Chapter 4: Conclusions

In this project motivated by a recent 2014 Ebola outbreak, we aimed to find evolutionary patterns within the family of *Filoviridae*. Inspired by a study on influenza (Warren S, 2013), we posed a hypothesis similar to the one based on the conclusions from the influenza study: the highly conserved residues across the species of *Filoviridae* family should be associated with the intra-viral interaction regions. Indeed, intuitively, in order for a virus to survive, it has to maintain minimal set of critical functions, such as replication of RNA, where the viral proteins interact with each other and are supposed to work together synergistically. Thus, genetic variations in the RNA sequence that lead to the change in the protein sequence might affect those interactions, potentially resulting in the reduced functional efficiency of the proteins. In contrast, the extremely diverse residues from a single viral species sampled over multiple years are likely to be associated with the host-viral interactions binding sites, specifically the antibody binding sites. To reach our goal, we determined and analyzed conserved and diverse regions both across species and across years. Also, in order to compare the conserved/diverse regions with the functional sites and ultimately to test our hypothesis, we also needed to annotate the viral proteins with the protein binding sites.

Overall, the proposed research could be split into three major steps. First, we would build three-dimensional models for the seven proteins in Ebola virus using the homology modeling method. Second, we would find the conserved and diverse regions across family, and map them to the models. At the same time, we would find the potential binding sites on each model by checking the binding sites on related structures for each model, and map them to the models. Residues involved in intra-viral interactions and in host-viral interactions will be recorded separately. Then, we would compare the conserved and diverse regions with the residues involved in the two types of functional sites, to test the first part of our hypothesis. Third, we would do the same for each species across years, and compare the conserved and diverse regions with the residues involved in the two types of interactions to test the second half of our hypothesis. However, due to the time limitation, in this project we have successfully achieved the first two major steps.

We first made models for each of the seven main proteins in the Ebola proteome using the homology modeling approach. Specifically, we built 3 separate models for NP, 2 for VP35, 1 for VP40, 2 for GP, 1 for VP30, 1 for VP24, and 2 for L. Due to the limitation in the number of structural templates and in their coverage of the target sequences, for most proteins, we could not build a complete model covering the whole protein. For instance, the structural coverage for NP and L is very low. Then, we selected one representative from each species within the *Filoviridae* family, aligned the sequences for each protein, and calculated the trident score for each position. We then found the distributions of trident scores for each protein and determined the cutoff for highly conserved residues to be trident score strictly equal to 1 and that for diverse residues to be trident score less than or equal to 0.4. Then, we mapped the highly conserved and highly diverse residues back to the models we built in the first step. In most cases, one can clearly see clusters of the conserved residues (*e.g.*, see the conserved and diverse regions mapping for NP2 and L2 – the upper model for L.) However, the diverse residues are somewhat randomly scattered on the protein surfaces for most proteins (*e.g.*, see the conserved and diverse regions mapping for VP40 and VP24). Finally, we annotated each model with its potential functional sites by searching for

existing related structures in the DOMMINO database [ref]. For NP and L, no related structures have been found. For VP35, VP40, and VP30, only intra-viral interactions have been found. For GP and VP24, both intra-viral interactions and host-viral interactions have been found.

Finally, in order to test our hypothesis that intra-viral interactions would be highly associated with the conserved regions, while the host-viral interactions would be associated with the diverse regions, we analyze the binding site annotation results. From the results, intra-viral interactions are only predicted on five models, which are VP35, VP40, GP, VP30, and VP24. Among those proteins, we also found those ones, GP and VP24, involved in the inter-viral binding sites. The results show that intra-viral binding sites have indeed a higher proportion of conserved residues than the diverse ones, while the reverse is also true for the host-virus binding sites. For some of the proteins it is harder to draw this conclusion than for the others. This is likely due to the lack of structural information and contrasts with the annotation of influenza proteome whose proteins are mostly structurally resolved. These results also suggest that an accurate measure is needed that estimates the proportion of the conserved and diverse residues and compares it to the background distribution of such residues on the entire protein surface.

## Chapter 5: Potential Problems and Future Studies

1. As the research goes, more and more data come out. However, due to time limitation, we did not include the newest data, *e.g.*, the structure of the newly discovered NP structure from Bundibugyo and Tai, and the VP35 structure from Ebola virus.
2. In this project, there are many steps involving protein sequence alignment. However, because different software packages are involved, we did not keep it consistent on which alignment method we used. If we had done an alignment for all the procedures, what would the results be? Will it be highly affected?
3. We assumed that after concatenating the two chunks of proteins we would get the corresponding structure of GP in Lloviu cuevavirus. However, from the perspective of mutation, we know that mutations started from RNA. Even one insertion could completely alter the open reading frame sequence. Hence, a better approach needs to be designed.

Finally, although the scores are calculated when we find the related structures for functional sites, the scores were not used in the visualizations, instead simpler binary annotation was implemented that identifies a residue as a binding residue or not. This avoids the problem of how much weight we should put on each interaction, since for some solved structures, there are redundant interactions, and we have not defined a proper way to take care of this. This will bring in many residues that may not be actively involved in interactions and should be resolved as one of the future steps.

## Reference

- Ana Negredo, G. P.-M. (2011). Discovery of an Ebolavirus-like Filovirus in Europe. *Plos Pathogens*, 7 (10), e1002304.
- Andrej Šali, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234 (3), 779-815.
- Bioinformatics, S. I. (2015). Retrieved May 12, 2015, from ViralZone: [http://viralzone.expasy.org/complete\\_by\\_protein/207.html](http://viralzone.expasy.org/complete_by_protein/207.html)
- CDC. (2015, May 9). *2014 Ebola Outbreak in West Africa - Case Counts*. Retrieved May 11, 2015, from Centers for Disease Control and Prevention: <http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html>
- CDC. (2015, May 22). *Outbreaks Chronology: Ebola Virus Disease*. Retrieved May 22, 2015, from Centers for Disease Control and Prevention: <http://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html>
- Centers for Disease Control and Prevention. (2014, October 9). Retrieved August 19, 2015, from Marburg Hemorrhagic Fever (Marburg HF): <http://www.cdc.gov/vhf/marburg/resources/outbreak-table.html>
- Craig S. Brown, M. S. (2014). In Silico Derived Small Molecules Bind the Filovirus Vp35 Protein and Inhibit Its Polymerase Cofactor Activity. *Journal of Molecular Biology*, 426 (10), 2045-2058.
- Dessen A, V. V. (2000). Crystal Structure of the Matrix Protein VP40 From Ebola Virus. *The EMBO Journal*, 19 (16), 4228-4236.
- European Bioinformatics Institute, the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource. (2015, January). Retrieved January 8, 2015, from UniProt: <http://www.uniprot.org/>
- Kiley, M. P. (1988). Filoviridae: Marburg and Ebola Viruses. In *Laboratory Diagnosis of Infectious Diseases Principles and Practice* (pp. 595-601). Springer New York.
- Kuang X, H. J. (2012). DOMMINO: A database of macromolecular interactions. *Nucleic Acids Res.* (40), D501-D506.
- Kuhn JH, B. S. (2010). Proposal for a Revised Taxonomy of the Family Filoviridae: Classification, Names, of Taxa and Viruses, and Virus Abbreviations. *Archives of Virology*, 155 (12), 2083-21-3.
- Marketa Zvelebil, J. O. (2008). Substitution Matrices. In J. O. Marketa Zvelebil, *Understanding Bioinformatics* (pp. 84-85). New York, NY, USA: Garland Science, Taylor & Francis Group, LLC.
- Masfique Mehedi, D. F.-J. (2011). A New Ebola Virus Nonstructural Glycoprotein Expressed through RNA Editing. *Journal of Virology*, 85 (11), 5406-5414.
- National Center for Biotechnology Information, U.S. National Library of Medicine. (2015, January). Retrieved January 8, 2015, from NCBI Protein: <http://www.ncbi.nlm.nih.gov/protein>
- National Science Foundation, the National Institutes of Health, and the US Department of Energy. (2015, January). Retrieved January 8, 2015, from Protein Data Bank: <http://www.rcsb.org/pdb/home/home.do>
- Paulina J. Dziuban'ska, U. D. (2014). The Structure of the C-terminal Domain of the Zaire Ebolavirus Nucleoprotein. *Acta Crystallographica Section D, Biological Crystallography*, 70, 2420-2429.
- PDB. (2014, December 10). Retrieved May 25, 2015, from Ebola Virus Proteins: <https://www.youtube.com/watch?v=9SptF9bCyd0>

PDB, D. S. (2014, October). *Ebola Virus Proteins - October 2014 Molecule of the Month*. Retrieved August 24, 2015, from RCSB PDB:  
<http://www.rcsb.org/pdb/101/motm.do?momID=178>

Sali Lab. (2015, May 19). *About Modeller*. Retrieved June 5, 2015, from Modeller:  
[https://salilab.org/modeller/about\\_modeller.html](https://salilab.org/modeller/about_modeller.html)

Sali Lab. (2010, January 12). *Tutorial*. Retrieved June 5, 2015, from Modeller:  
<https://salilab.org/modeller/tutorial/>

Samantha Warren, X.-F. W. (2013). Extreme Evolutionary Conservation of Functionally Important Regions in H1N1 Influenza Proteome. *PLoS ONE*, 8 (11), 1-14.

Serena A. Carroll, J. S. (2013). Molecular Evolution of Viruses of the Family Filoviridae Based on 97 Whole-Genome Sequences. *Journal of Virology*, 87 (5), 2608-2616.

Swiss-Prot. (2002-2015). Retrieved May 12, 2015, from UniProt:  
<http://www.uniprot.org/uniprot/?query=database%3A%28type%3Aembl+AF086833%29&sort=score>

Valdar, W. S. (2002). Scoring Residue Conservation. *Proteins*, 48 (2), 227-241.

Warren S, W. X.-F. (2013). Extreme Evolutionary Conservation of Functionally Important Regions in H1N1 Influenza Proteome. *PLoS ONE*, 8 (11), 1-14.

Webb, B. (2007, August 3). *Alignment File (PIR)*. Retrieved January 8, 2015, from Modeller:  
<http://salilab.org/modeller/9v2/manual/node412.html>

# Appendices

## Appendix A. Protein Sequences Found in UniProt

### Appendix A1. Protein Sequences Found in UniProt (NP)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCM7	B8XCM7_9MO NO	U	Bundibugyo ebolavirus	739	3/3/09	Bundibugyo		
R4QJ68	R4QJ68_9MO NO	R	Bundibugyo ebolavirus	739	7/24/13	Bundibugyo		
A9Q1H2	A9Q1H2_9MO NO	U	Lake Victoria marburgvirus - Ci67	695	2/5/08			
Q1PDB6	Q1PDB6_9MO NO	U	Lake Victoria marburgvirus - DRC1999	695	5/16/06	Democratic Republic of the Congo		1999
Q1PDC3	Q1PDC3_9MO NO	U	Lake Victoria marburgvirus - DRC1999	695	5/16/06	Democratic Republic of the Congo		1999
Q1PDA9	Q1PDA9_9MO NO	U	Lake Victoria marburgvirus - DRC1999	695	5/16/06	Democratic Republic of the Congo		1999
G9HWE2	G9HWE2_9MO NO	U	Lake Victoria marburgvirus - Leiden	695	2/22/12	Leiden		
Q1PD74	Q1PD74_MAB VA	U	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	695	5/16/06		Angola	2005
Q1PD53	NCAP_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	695	1/15/08		Angola	2005
P27588	NCAP_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	695	8/1/92	Musoke, Kenya	Musok e-80	1980
Q6UY69	NCAP_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	695	1/15/08		Ozolin -75	
P35263	NCAP_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	695	2/1/94	Popp, West Germany	Popp- 67	1967
C0IT83	C0IT83_MABV R	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	695	5/5/09	Ravn, Kenya	Ravn- 87	1987
C0IT76	C0IT76_MABV R	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	695	5/5/09	Ravn, Kenya	Ravn- 88	1987
A9Q1G5	A9Q1G5_MAB VR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	695	2/5/08	Ravn, Kenya	Ravn- 89	1987
Q1PDD0	NCAP_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	695	1/15/08	Ravn, Kenya	Ravn- 90	1987
G8EF11	G8EF11_LLOV A	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias- Bat86/2003) (LLOV)	749	1/25/12	Asturias, Spain	Asturia s- Bat86	2003
K4MSQ7	K4MSQ7_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4N3C7	K4N3C7_9MO NO	U	Marburg marburgvirus	695	1/9/13			
C7B248	C7B248_9MON O	U	Marburg marburgvirus	695	9/22/09			
E0X104	E0X104_9MON	U	Marburg marburgvirus	695	11/2/10			

	O							
E0X111	E0X111_9MON O	U	Marburg marburgvirus	695	11/2/10			
E0X0Z7	E0X0Z7_9MON O	U	Marburg marburgvirus	695	11/2/10			
K4MT16	K4MT16_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MRA8	K4MRA8_9MO NO	U	Marburg marburgvirus	695	1/9/13			
C7B283	C7B283_9MON O	U	Marburg marburgvirus	695	9/22/09			
C7B276	C7B276_9MON O	U	Marburg marburgvirus	695	9/22/09			
K4N3V7	K4N3V7_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MQC5	K4MQC5_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MSZ0	K4MSZ0_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MQ25	K4MQ25_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4N3K2	K4N3K2_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MR07	K4MR07_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MR89	K4MR89_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MQY0	K4MQY0_9MO NO	U	Marburg marburgvirus	695	1/9/13			
C7B269	C7B269_9MON O	U	Marburg marburgvirus	695	9/22/09			
K4MQF4	K4MQF4_9MO NO	U	Marburg marburgvirus	695	1/9/13			
A0A077D 355	A0A077D355_9 MONO	U	Marburg marburgvirus	695	10/29/14			
K4N3V7	K4N3V7_9MO NO	U	Marburg marburgvirus	695	1/9/13			
E0X111	E0X111_9MON O	U	Marburg marburgvirus	695	11/2/10			
K4MQC5	K4MQC5_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MSZ0	K4MSZ0_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MQ25	K4MQ25_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4N3K2	K4N3K2_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MR07	K4MR07_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MR89	K4MR89_9MO NO	U	Marburg marburgvirus	695	1/9/13			
E0X0Z7	E0X0Z7_9MON O	U	Marburg marburgvirus	695	11/2/10			
K4MQY0	K4MQY0_9MO NO	U	Marburg marburgvirus	695	1/9/13			
K4MQF4	K4MQF4_9MO NO	U	Marburg marburgvirus	695	1/9/13			
C7B269	C7B269_9MON O	U	Marburg marburgvirus	695	9/22/09			
L7RI74	L7RI74_9MON O	U	Reston ebolavirus	739	4/3/13			
L7REU1	L7REU1_9MO NO	U	Reston ebolavirus	739	4/3/13			
C6G8D3	C6G8D3_EBO RE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	739	9/1/09			Philipp ines- 96
C6G8E1	C6G8E1_EBO RE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	739	9/1/09			Philipp ines- 96
C6G8E9	C6G8E9_EBO RE	U	Reston ebolavirus (strain Philippines-96) (REBOV)	739	9/1/09			Philipp ines-

			(Reston Ebola virus)				96	
Q91DE1	NCAP_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	739	7/11/06		Philippines-96	
Q8JPY1	NCAP_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	739	7/11/06		Reston-89	
B0LPL4	B0LPL4_9MONO	U	Sudan ebolavirus	738	3/18/08			
C4PK55	C4PK55_9MONO	U	Sudan ebolavirus	738	7/7/09			
M4J9H3	M4J9H3_9MONO	U	Sudan ebolavirus	738	5/29/13			
R4QTJ0	R4QTJ0_9MONO	U	Sudan ebolavirus	738	7/24/13			
R4NSX3	R4NSX3_9MONO	U	Sudan ebolavirus	738	7/24/13			
R4QTJ0	R4QTJ0_9MONO	U	Sudan ebolavirus	738	7/24/13			
I7FUR1	I7FUR1_9MONO	U	Sudan ebolavirus - Nakisamata	738	10/3/12	Nakisamata		
Q9QP77	NCAP_EBOSB	R	Sudan ebolavirus (strain Boniface-76) (SEBOV) (Sudan Ebola virus)	738	5/30/00		Boniface-76	
Q5XX08	NCAP_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	738	7/11/06		Uganda-00	
B8XCN6	B8XCN6_9MONO	U	Tai Forest ebolavirus	739	3/3/09			
L7QI51	L7QI51_9MONO	U	Zaire ebolavirus	739	4/3/13			
L7QHT6	L7QHT6_9MONO	U	Zaire ebolavirus	739	4/3/13			
I0DHA6	I0DHA6_9MONO	U	Zaire ebolavirus	739	6/13/12			
L7QIA8	L7QIA8_9MONO	U	Zaire ebolavirus	739	4/3/13			
L7QI42	L7QI42_9MONO	U	Zaire ebolavirus	739	4/3/13			
L7QHU5	L7QHU5_9MONO	U	Zaire ebolavirus	739	4/3/13			
G8DB37	G8DB37_9MONO	U	Zaire ebolavirus	739	1/25/12			
A0A068J945	A0A068J945_9MONO	U	Zaire ebolavirus	739	10/1/14			
X5GXS8	X5GXS8_9MONO	U	Zaire ebolavirus	739	6/11/14			
A0A075WG99	A0A075WG99_9MONO	U	Zaire ebolavirus	739	10/29/14			
A9QPL6	A9QPL6_9MONO	U	Zaire ebolavirus	739	2/5/08			
X5HMY5	X5HMY5_9MONO	U	Zaire ebolavirus	739	6/11/14			
A0A075W933	A0A075W933_9MONO	U	Zaire ebolavirus	739	10/29/14			
A0A068J9B9	A0A068J9B9_9MONO	U	Zaire ebolavirus	739	10/1/14			
Q9QCE9	NCAP_EBOG4	R	Zaire ebolavirus (strain Gabon-94) (ZEBOV) (Zaire Ebola virus)	739	5/30/00		Gabon-94	
O72142	NCAP_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	739	5/30/00		Kikwit-95	
P18272	NCAP_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	739	11/1/90		Mayinga-76	

## Appendix A2. Protein Sequences Found in UniProt (VP35)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCM8	B8XCM8_9MONO	U	Bundibugyo ebolavirus	341	3/3/09			
R4QGV8	R4QGV8_9MONO	U	Bundibugyo ebolavirus	341	7/24/13			
A9Q1H3	A9Q1H3_9MONO	U	Lake Victoria marburgvirus - Ci67	329	2/5/08			
Q1PDB5	Q1PDB5_9MONO	U	Lake Victoria marburgvirus - DRC1999	329	5/16/06	Democratic Republic of the Congo		1999
Q1PDA8	Q1PDA8_9MONO	U	Lake Victoria marburgvirus - DRC1999	329	5/16/06	Democratic Republic of the Congo		1999
G9HWE3	G9HWE3_9MONO	U	Lake Victoria marburgvirus - Leiden	329	2/22/12	Leiden		
Q1PD52	VP35_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	329	1/15/08	Angola		2005
P35259	VP35_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	329	2/1/94	Musoke, Kenya	Musoke-80	1980
Q6UY68	VP35_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	329	1/15/08	Ozolin, South Africa	Ozolin-75	1975
Q03039	VP35_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	329	2/1/94	Popp, West Germany	Popp-67	1976
Q1PDC9	VP35_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	329	1/15/08	Ravn, Kenya	Ravn-87	1987
A9Q1G6	A9Q1G6_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	329	2/5/08	Ravn, Kenya	Ravn-87	1987
C0IT77	C0IT77_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	329	5/5/09	Ravn, Kenya	Ravn-87	1987
C0IT91	C0IT91_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	329	5/5/09	Ravn, Kenya	Ravn-87	1987
G8EFI2	G8EFI2_LLOVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	320	1/25/12	Asturias, Spain	Asturias-Bat86	2003
C7B270	C7B270_9MONO	U	Marburg marburgvirus	329	9/22/09			
K4MSU6	K4MSU6_9MONO	U	Marburg marburgvirus	329	1/9/13			
E0X0Z8	E0X0Z8_9MONO	U	Marburg marburgvirus	329	11/2/10			
C7B249	C7B249_9MONO	U	Marburg marburgvirus	329	9/22/09			
C7B277	C7B277_9MONO	U	Marburg marburgvirus	329	9/22/09			
C7B256	C7B256_9MONO	U	Marburg marburgvirus	329	9/22/09			
K4MQB0	K4MQB0_9MONO	U	Marburg marburgvirus	329	1/9/13			
K4MST7	K4MST7_9MONO	U	Marburg marburgvirus	329	1/9/13			
K4MTB0	K4MTB0_9MONO	U	Marburg marburgvirus	329	1/9/13			



K4MQU1	K4MQU1_9MON O	U	Marburg marburgvirus	329	1/9/13			
K4MTE0	K4MTE0_9MON O	U	Marburg marburgvirus	329	1/9/13			
A0A077 CY87	A0A077CY87_9 MONO	U	Marburg marburgvirus	329	10/29/14			
O36427	O36427_9MON O	U	Marburg marburgvirus	329	1/1/98			
L7RFC8	L7RFC8_9MON O	U	Reston ebolavirus	329	4/3/13			
L7RF08	L7RF08_9MON O	U	Reston ebolavirus	329	4/3/13			
Q91DE0	VP35_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	329	7/11/06		Philipp ines- 96	
C6G8F0	C6G8F0_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	329	9/1/09		Philipp ines- 96	
C6G8E2	C6G8E2_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	329	9/1/09		Philipp ines- 96	
C6G8D4	C6G8D4_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	329	9/1/09		Philipp ines- 96	
Q8JPY0	VP35_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	329	7/11/06		Resto n-89	
B0LPL5	B0LPL5_9MON O	U	Sudan ebolavirus	329	3/18/08			
C4PK56	C4PK56_9MON O	U	Sudan ebolavirus	329	7/7/09			
R4NUK7	R4NUK7_9MON O	U	Sudan ebolavirus	329	7/24/13			
R4QGR7	R4QGR7_9MON O	U	Sudan ebolavirus	329	7/24/13			
I7FLL9	I7FLL9_9MONO	U	Sudan ebolavirus - Nakisamata	329	10/3/12	Nakisamata		
Q5XX07	VP35_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	329	7/11/06		Ugand a-00	
B8XCN7	B8XCN7_9MON O	U	Tai Forest ebolavirus	341	3/3/09			
A9QPL7	A9QPL7_9MON O	U	Zaire ebolavirus	340	2/5/08			
G8DB47	G8DB47_9MON O	U	Zaire ebolavirus	340	1/25/12			
X5GXU1	X5GXU1_9MON O	U	Zaire ebolavirus	340	6/11/14			
A0A068J 6A9	A0A068J6A9_9 MONO	U	Zaire ebolavirus	340	10/1/14			
G8DB38	G8DB38_9MON O	U	Zaire ebolavirus	340	1/25/12			
Q6V1Q9	VP35_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	340	7/11/06		Kikwit- 95	
Q05127	VP35_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	340	2/1/94		Mayin ga-76	

### Appendix A3. Protein Sequences Found in UniProt (VP40)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCM9	B8XCM9_9MONO	U	Bundibugyo ebolavirus	326	3/3/09			
R4QGW8	R4QGW8_9MONO	U	Bundibugyo ebolavirus	326	7/24/13			
A9Q1H4	A9Q1H4_9MONO	U	Lake Victoria marburgvirus - Ci67	303	2/5/08			
Q1PDB4	Q1PDB4_9MONO	U	Lake Victoria marburgvirus - DRC1999	303	5/16/06	Democratic Republic of the Congo		1999
Q1PDA7	Q1PDA7_9MONO	U	Lake Victoria marburgvirus - DRC1999	303	5/16/06	Democratic Republic of the Congo		1999
G9HWE4	G9HWE4_9MONO	U	Lake Victoria marburgvirus - Leiden	303	2/22/12	Leiden		
Q1PD51	VP40_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	303	1/15/08	Angola		2005
P35260	VP40_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	303	2/1/94	Musoke, Kenya	Musoke-80	1980
Q6UY67	VP40_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	303	1/15/08	Ozolin, South Africa	Ozolin-75	1975
Q03040	VP40_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	303	2/1/94	Popp, West Germany	Popp-67	1976
Q1PDC8	VP40_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	303	1/15/08	Ravn, Kenya	Ravn-87	1987
C0IT92	C0IT92_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	303	5/5/09	Ravn, Kenya	Ravn-87	1987
A9Q1G7	A9Q1G7_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	303	2/5/08	Ravn, Kenya	Ravn-87	1987
C0IT78	C0IT78_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	303	5/5/09	Ravn, Kenya	Ravn-87	1987
C0IT85	C0IT85_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	303	5/5/09	Ravn, Kenya	Ravn-87	1987
G8EFI3	G8EFI3_LLOVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	329	1/25/12	Asturias, Spain	Asturias-Bat86	2003
C7B271	C7B271_9MONO	U	Marburg marburgvirus	303	9/22/09			
C7B278	C7B278_9MONO	U	Marburg marburgvirus	303	9/22/09			
C7B250	C7B250_9MONO	U	Marburg marburgvirus	303	9/22/09			
E0X113	E0X113_9MONO	U	Marburg marburgvirus	303	11/2/10			
K4MQT4	K4MQT4_9MONO	U	Marburg marburgvirus	303	1/9/13			
C7B285	C7B285_9MONO	U	Marburg marburgvirus	303	9/22/09			
E0X0Z9	E0X0Z9_9MONO	U	Marburg marburgvirus	303	11/2/10			

K4MQ95	K4MQ95_9MON O	U	Marburg marburgvirus	303	1/9/13			
E0X106	E0X106_9MON O	U	Marburg marburgvirus	303	11/2/10			
K4MR86	K4MR86_9MON O	U	Marburg marburgvirus	303	1/9/13			
A0A077D 0Z4	A0A077D0Z4_9 MONO	U	Marburg marburgvirus	303	10/29/14			
L7RDZ6	L7RDZ6_9MON O	U	Reston ebolavirus	331	4/3/13			
L7REV2	L7REV2_9MON O	U	Reston ebolavirus	331	4/3/13			
Q91DD9	VP40_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	331	7/11/06		Philipp ines- 96	
C6G8D5	C6G8D5_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	331	9/1/09		Philipp ines- 96	
C6G8F1	C6G8F1_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	331	9/1/09		Philipp ines- 96	
C6G8E3	C6G8E3_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	330	9/1/09		Philipp ines- 96	
Q8JPX9	VP40_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	331	7/11/06		Resto n-89	
B0LPL6	B0LPL6_9MON O	U	Sudan ebolavirus	326	3/18/08			
R4QRB1	R4QRB1_9MON O	U	Sudan ebolavirus	326	7/24/13			
I7F2J6	I7F2J6_9MONO	U	Sudan ebolavirus - Nakisamata	326	10/3/12	Nakisamata		
Q5XX06	VP40_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	326	7/11/06		Ugand a-00	
B8XCN8	B8XCN8_9MON O	U	Tai Forest ebolavirus	326	3/3/09			
A9QPL8	A9QPL8_9MON O	U	Zaire ebolavirus	326	2/5/08			
X5H596	X5H596_9MON O	U	Zaire ebolavirus	326	6/11/14			
G8DB48	G8DB48_9MON O	U	Zaire ebolavirus	326	1/25/12			
A0A075 WBZ0	A0A075WBZ0_9 MONO	U	Zaire ebolavirus	326	10/29/14			
Q2PDK5	VP40_EBOG4	R	Zaire ebolavirus (strain Gabon-94) (ZEBOV) (Zaire Ebola virus)	326	7/11/06		Gabon -94	
Q77DJ6	VP40_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	326	7/11/06		Kikwit- 95	
Q05128	VP40_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	326	2/1/94		Mayin ga-76	

## Appendix A4. Protein Sequences Found in UniProt (GP)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCN2	B8XCN2_9MONO	U	Bundibugyo ebolavirus	302	3/3/09			
B8XCN0	B8XCN0_9MONO	U	Bundibugyo ebolavirus	676	3/3/09			
B8XCN1	B8XCN1_9MONO	U	Bundibugyo ebolavirus	373	3/3/09			
R4QGV6	R4QGV6_9MONO	U	Bundibugyo ebolavirus	676	7/24/13			
R4QRC0	R4QRC0_9MONO	U	Bundibugyo ebolavirus	676	7/24/13			
A9Q1H5	A9Q1H5_9MONO	U	Lake Victoria marburgvirus - Ci67	681	2/5/08			
Q1PDA6	Q1PDA6_9MONO	U	Lake Victoria marburgvirus - DRC1999	681	5/16/06	Democratic Republic of the Congo		1999
Q1PDB3	Q1PDB3_9MONO	U	Lake Victoria marburgvirus - DRC1999	681	5/16/06	Democratic Republic of the Congo		1999
Q1PDC0	Q1PDC0_9MONO	U	Lake Victoria marburgvirus - DRC1999	681	5/16/06	Democratic Republic of the Congo		1999
G9HWE5	G9HWE5_9MONO	U	Lake Victoria marburgvirus - Leiden	681	2/22/12	Leiden		
Q1PD50	VGP_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	681	1/15/08	Angola		2005
P35253	VGP_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	681	2/1/94	Musoke, Kenya	Musoke-80	1980
Q6UY66	VGP_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	681	1/15/08	Ozolin, South Africa	Ozolin-75	1975
P35254	VGP_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	681	2/1/94	Popp, West Germany	Popp-67	1976
Q1PDC7	VGP_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	681	1/15/08	Ravn, Kenya	Ravn-87	1987
C0IT79	C0IT79_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	681	5/5/09	Ravn, Kenya	Ravn-87	1987
A9Q1G8	A9Q1G8_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	681	2/5/08	Ravn, Kenya	Ravn-87	1987
G8EFI4	G8EFI4_LLOVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	415	1/25/12	Asturias, Spain	Asturias-Bat86	2003
G8EFI5	G8EFI5_LLOVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	320	1/25/12	Asturias, Spain	Asturias-Bat86	2003
K4MQZ0	K4MQZ0_9MONO	U	Marburg marburgvirus	681	1/9/13			
K4MR14	K4MR14_9MONO	U	Marburg marburgvirus	681	1/9/13			
K4N3N9	K4N3N9_9MONO	U	Marburg marburgvirus	681	1/9/13			
C7B251	C7B251_9MONO	U	Marburg marburgvirus	681	9/22/09			
Q71VM1	Q71VM1_9MONO	U	Marburg marburgvirus	681	7/5/04			

K4MSG5	K4MSG5_9MON O	U	Marburg marburgvirus	681	1/9/13			
C7B286	C7B286_9MON O	U	Marburg marburgvirus	681	9/22/09			
K4N3G2	K4N3G2_9MON O	U	Marburg marburgvirus	681	1/9/13			
A0A077C Y85	A0A077CY85_9 MONO	U	Marburg marburgvirus	681	10/29/14			
K4MQA7	K4MQA7_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MT22	K4MT22_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MQ81	K4MQ81_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MR75	K4MR75_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MSZ7	K4MSZ7_9MON O	U	Marburg marburgvirus	681	1/9/13			
C7B272	C7B272_9MON O	U	Marburg marburgvirus	681	9/22/09			
K4MSY0	K4MSY0_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MST4	K4MST4_9MON O	U	Marburg marburgvirus	681	1/9/13			
C7B279	C7B279_9MON O	U	Marburg marburgvirus	681	9/22/09			
K4N3W4	K4N3W4_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MSV6	K4MSV6_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MTD6	K4MTD6_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4N3L1	K4N3L1_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MR55	K4MR55_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MT76	K4MT76_9MON O	U	Marburg marburgvirus	681	1/9/13			
C7B293	C7B293_9MON O	U	Marburg marburgvirus	681	9/22/09			
E0X100	E0X100_9MON O	U	Marburg marburgvirus	681	11/2/10			
K4MSU2	K4MSU2_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MR93	K4MR93_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4MTA6	K4MTA6_9MON O	U	Marburg marburgvirus	681	1/9/13			
K4N3Z7	K4N3Z7_9MON O	U	Marburg marburgvirus	681	1/9/13			
R4QGR2	R4QGR2_9MON O	U	Marburg marburgvirus	681	7/24/13			
L7REV7	L7REV7_9MON O	U	Reston ebolavirus	677	4/3/13			
L7RFD3	L7RFD3_9MON O	U	Reston ebolavirus	381	4/3/13			
L7RF11	L7RF11_9MON O	U	Reston ebolavirus	367	4/3/13			
Q91DD8	VGP_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	677	7/11/06		Philipp ines- 96	
Q91DD7	VSGP_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	367	7/11/06		Philipp ines- 96	
P0C770	VSSGP_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	332	2/9/10		Philipp ines- 96	
C6G8F2	C6G8F2_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	677	9/1/09		Philipp ines- 96	
C6G8E5	C6G8E5_EBOR	U	Reston ebolavirus (strain	367	9/1/09		Philipp	

	E		Philippines-96) (REBOV) (Reston Ebola virus)				ines-96	
C6G8D7	C6G8D7_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	381	9/1/09		Philipp ines-96	
C6G8E4	C6G8E4_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	677	9/1/09		Philipp ines-96	
C6G8F3	C6G8F3_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	367	9/1/09		Philipp ines-96	
C6G8D6	C6G8D6_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	677	9/1/09		Philipp ines-96	
C6G8D7	C6G8D7_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	381	9/1/09		Philipp ines-96	
C6G8E5	C6G8E5_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	367	9/1/09		Philipp ines-96	
C6G8F3	C6G8F3_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	367	9/1/09		Philipp ines-96	
Q66799	VGP_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	677	5/30/00		Resto n-89	
Q66800	VSGP_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	367	12/1/00		Resto n-89	
P0C771	VSSGP_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	332	2/9/10		Resto n-89	
Q89853	VGP_EBORS	R	Reston ebolavirus (strain Siena/Philippine-92) (REBOV) (Reston Ebola virus)	677	5/30/00		Philipp ine-92	
Q89569	VSGP_EBORS	R	Reston ebolavirus (strain Siena/Philippine-92) (REBOV) (Reston Ebola virus)	367	12/1/00		Philipp ine-92	
B0LPL8	B0LPL8_9MON O	U	Sudan ebolavirus	372	3/18/08			
B0LPL7	B0LPL7_9MON O	U	Sudan ebolavirus	676	3/18/08			
M4JBD3	M4JBD3_9MON O	U	Sudan ebolavirus	676	5/29/13			
C4PK58	C4PK58_9MON O	U	Sudan ebolavirus	676	7/7/09			
C4PK59	C4PK59_9MON O	U	Sudan ebolavirus	372	7/7/09			
R4P4N7	R4P4N7_9MON O	U	Sudan ebolavirus	676	7/24/13			
R4QJ45	R4QJ45_9MON O	U	Sudan ebolavirus	676	7/24/13			
I7FLM2	I7FLM2_9MONO	U	Sudan ebolavirus - Nakisamata	372	10/3/12	Nakisamata		
I7F2J9	I7F2J9_9MONO	U	Sudan ebolavirus - Nakisamata	676	10/3/12	Nakisamata		
Q66814	VGP_EBOSB	R	Sudan ebolavirus (strain Boniface-76) (SEBOV) (Sudan Ebola virus)	676	5/30/00		Bonifa ce-76	
P60172	VSGP_EBOSB	R	Sudan ebolavirus (strain Boniface-76) (SEBOV) (Sudan Ebola virus)	372	12/15/03		Bonifa ce-76	
Q66798	VGP_EBOSM	R	Sudan ebolavirus (strain Maleo-79) (SEBOV) (Sudan Ebola virus)	676	5/30/00		Maleo-79	
P60173	VSGP_EBOSM	R	Sudan ebolavirus (strain Maleo-79) (SEBOV) (Sudan Ebola virus)	372	12/15/03		Maleo-79	
Q7T9D9	VGP_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV)	676	2/5/08		Ugand a-00	

			(Sudan Ebola virus)				
P0C772	VSSGP_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	319	2/9/10		Uganda-00
Q7T9E0	VSGP_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	372	7/11/06		Uganda-00
B8XCP0	B8XCP0_9MONO	U	Tai Forest ebolavirus	365	3/3/09		
B8XCN9	B8XCN9_9MONO	U	Tai Forest ebolavirus	676	3/3/09		
B8XCP1	B8XCP1_9MONO	U	Tai Forest ebolavirus	302	3/3/09		
Q66810	VGP_TAFVC	R	Tai Forest ebolavirus (strain Cote d'Ivoire-94) (TAFV) (Cote d'Ivoire Ebola virus)	676	5/30/00		Cote d'Ivoire-94
Q66811	VSGP_TAFVC	R	Tai Forest ebolavirus (strain Cote d'Ivoire-94) (TAFV) (Cote d'Ivoire Ebola virus)	365	12/1/00		Cote d'Ivoire-94
L7QHW1	L7QHW1_9MONO	U	Zaire ebolavirus	297	4/3/13		
A9QPM0	A9QPM0_9MONO	U	Zaire ebolavirus	364	2/5/08		
A9QPL9	A9QPL9_9MONO	U	Zaire ebolavirus	676	2/5/08		
G8DB50	G8DB50_9MONO	U	Zaire ebolavirus	364	1/25/12		
G8DB49	G8DB49_9MONO	U	Zaire ebolavirus	676	1/25/12		
X5HMX4	X5HMX4_9MONO	U	Zaire ebolavirus	676	6/11/14		
A0A068J419	A0A068J419_9MONO	U	Zaire ebolavirus	676	10/1/14		
A0A068J949	A0A068J949_9MONO	U	Zaire ebolavirus	297	10/1/14		
A0A068J941	A0A068J941_9MONO	U	Zaire ebolavirus	364	10/1/14		
G8DB42	G8DB42_9MONO	U	Zaire ebolavirus	297	1/25/12		
X5H5I7	X5H5I7_9MONO	U	Zaire ebolavirus	364	6/11/14		
X5HMZ0	X5HMZ0_9MONO	U	Zaire ebolavirus	297	6/11/14		
A0A068J4A4	A0A068J4A4_9MONO	U	Zaire ebolavirus	676	10/1/14		
X5HL92	X5HL92_9MONO	U	Zaire ebolavirus	364	6/11/14		
X5H5A9	X5H5A9_9MONO	U	Zaire ebolavirus	676	6/11/14		
A0A068J471	A0A068J471_9MONO	U	Zaire ebolavirus	364	10/1/14		
A0A068J9C3	A0A068J9C3_9MONO	U	Zaire ebolavirus	297	10/1/14		
V9TJ34	V9TJ34_9MONO	U	Zaire ebolavirus	410	3/19/14		
V9TJ39	V9TJ39_9MONO	U	Zaire ebolavirus	298	3/19/14		
P87671	VGP_EBOEC	R	Zaire ebolavirus (strain Eckron-76) (ZEBOV) (Zaire Ebola virus)	676	5/30/00		Eckron-76
P87670	VSGP_EBOEC	R	Zaire ebolavirus (strain Eckron-76) (ZEBOV) (Zaire Ebola virus)	364	12/1/00		Eckron-76
O11457	VGP_EBOG4	R	Zaire ebolavirus (strain Gabon-94) (ZEBOV) (Zaire Ebola virus)	676	5/30/00		Gabon-94
O11458	VSGP_EBOG4	R	Zaire ebolavirus (strain Gabon-94) (ZEBOV) (Zaire Ebola virus)	364	12/1/00		Gabon-94

P87666	VGP_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	676	5/30/00		Kikwit-95	
P60171	VSGP_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	364	12/15/03		Kikwit-95	
P0C773	VSSGP_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	298	2/9/10		Kikwit-95	
Q6V1Q5	Q6V1Q5_EBOZ5	U	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	297	7/5/04		Kikwit-95	
Q05320	VGP_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	676	2/1/94		Mayinga-76	
P60170	VSGP_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	364	12/15/03		Mayinga-76	
Q9YMG2	VSSGP_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	298	2/9/10		Mayinga-76	



## Appendix A5. Protein Sequences Found in UniProt (VP30)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCN3	B8XCN3_9MONO	U	Bundibugyo ebolavirus	289	3/3/09			
R4QUH9	R4QUH9_9MONO	U	Bundibugyo ebolavirus	289	7/24/13			
A9Q1H6	A9Q1H6_9MONO	U	Lake Victoria marburgvirus - Ci67	281	2/5/08			
Q1PDB9	Q1PDB9_9MONO	U	Lake Victoria marburgvirus - DRC1999	281	5/16/06	Democratic Republic of the Congo		1999
Q1PDB2	Q1PDB2_9MONO	U	Lake Victoria marburgvirus - DRC1999	281	5/16/06	Democratic Republic of the Congo		1999
Q1PDA5	Q1PDA5_9MONO	U	Lake Victoria marburgvirus - DRC1999	281	5/16/06	Democratic Republic of the Congo		1999
G9HWE6	G9HWE6_9MONO	U	Lake Victoria marburgvirus - Leiden	281	2/22/12	Leiden		
Q1PD56	VP30_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	281	1/15/08		Angola	2005
Q1PD49	Q1PD49_MABVA	U	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	281	5/16/06		Angola	2005
P35258	VP30_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	281	2/1/94	Musoke, Kenya	Musoke-80	1980
Q6UY65	VP30_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	281	1/15/08		Ozolin-75	
P41326	VP30_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	281	2/1/95	Popp, West Germany	Popp-67	1967
Q1PDC6	VP30_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	281	1/15/08	Ravn, Kenya	Ravn-87	1987
A9Q1G9	A9Q1G9_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	281	2/5/08	Ravn, Kenya	Ravn-88	1987
C0IT80	C0IT80_MABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	281	5/5/09	Ravn, Kenya	Ravn-89	1987
G8EFI6	G8EFI6_LLOVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	328	1/25/12	Asturias, Spain	Asturias-Bat86	2003
K4MSY4	K4MSY4_9MONO	U	Marburg marburgvirus	281	1/9/13			
C7B280	C7B280_9MONO	U	Marburg marburgvirus	281	9/22/09			
C7B273	C7B273_9MONO	U	Marburg marburgvirus	281	9/22/09			
K4MSZ1	K4MSZ1_9MONO	U	Marburg marburgvirus	281	1/9/13			
K4MQ93	K4MQ93_9MONO	U	Marburg marburgvirus	281	1/9/13			
K4N3M2	K4N3M2_9MONO	U	Marburg marburgvirus	281	1/9/13			
C7B252	C7B252_9MONO	U	Marburg marburgvirus	281	9/22/09			
E0X101	E0X101_9MONO	U	Marburg marburgvirus	281	11/2/10			

K4N3J5	K4N3J5_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MT90	K4MT90_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MQE5	K4MQE5_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MSM3	K4MSM3_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MR00	K4MR00_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MQB7	K4MQB7_9MON O	U	Marburg marburgvirus	281	1/9/13			
K4MSU4	K4MSU4_9MON O	U	Marburg marburgvirus	281	1/9/13			
R4QIZ8	R4QIZ8_9MON O	U	Marburg marburgvirus	281	7/24/13			
A0A077C ZD4	A0A077CZD4_9 MONO	U	Marburg marburgvirus	281	10/29/14			
L7RI71	L7RI71_9MONO	U	Reston ebolavirus	281	4/3/13			
L7RFD4	L7RFD4_9MON O	U	Reston ebolavirus	287	4/3/13			
Q91DD6	VP30_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	287	7/11/06		Philipp ines- 96	
C6G8F4	C6G8F4_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	287	9/1/09		Philipp ines- 96	
C6G8E6	C6G8E6_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	287	9/1/09		Philipp ines- 96	
C6G8D8	C6G8D8_EBOR E	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	281	9/1/09		Philipp ines- 96	
Q8JPX6	VP30_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	287	7/11/06		Resto n-89	
B0LPL9	B0LPL9_9MON O	U	Sudan ebolavirus	288	3/18/08			
M4J9H6	M4J9H6_9MON O	U	Sudan ebolavirus	288	5/29/13			
R4QUF0	R4QUF0_9MON O	U	Sudan ebolavirus	288	7/24/13			
R4P2H0	R4P2H0_9MON O	U	Sudan ebolavirus	288	7/24/13			
I7FWC9	I7FWC9_9MON O	U	Sudan ebolavirus - Nakisamata	288	10/3/12	Nakisamata		
Q5XX03	VP30_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	288	7/11/06		Ugand a-00	
B8XCP2	B8XCP2_9MON O	U	Tai Forest ebolavirus	289	3/3/09			
A9QPM2	A9QPM2_9MON O	U	Zaire ebolavirus	288	2/5/08			
Q77DJ5	VP30_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	288	7/11/06		Kikwit- 95	
Q05323	VP30_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	288	2/1/94		Mayin ga-76	

## Appendix A6. Protein Sequences Found in UniProt (VP24)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XCN4	B8XCN4_9 MONO	U	Bundibugyo ebolavirus	251	3/3/09			
R4QRB9	R4QRB9_9 MONO	U	Bundibugyo ebolavirus	251	7/24/13			
A9Q1H7	A9Q1H7_9M ONO	U	Lake Victoria marburgvirus - Ci67	253	2/5/08			
Q1PDB1	Q1PDB1_9 MONO	U	Lake Victoria marburgvirus - DRC1999	253	5/16/06	Democratic Republic of the Congo		1999
Q1PDA4	Q1PDA4_9 MONO	U	Lake Victoria marburgvirus - DRC1999	253	5/16/06	Democratic Republic of the Congo		1999
G9HWE7	G9HWE7_9 MONO	U	Lake Victoria marburgvirus - Leiden	253	2/22/12	Leiden		
Q1PD62	VP24_MAB VA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	253	1/15/08		Angola	2005
Q1PD48	Q1PD48_M ABVA	U	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	253	5/16/06		Angola	2005
P35256	VP24_MAB VM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	253	2/1/94	Musoke, Kenya	Musoke-80	1980
Q6UY64	VP24_MAB VO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	253	1/15/08		Ozolin-75	
P41325	VP24_MAB VP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	253	2/1/95	Popp, West Germany	Popp-67	1967
Q1PDC5	VP24_MAB VR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	253	1/15/08	Ravn, Kenya	Ravn-87	1987
A9Q1H0	A9Q1H0_M ABVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	253	2/5/08	Ravn, Kenya	Ravn-88	1987
G8EF18	G8EF18_LL OVA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias-Bat86/2003) (LLOV)	250	1/25/12	Asturias, Spain	Asturias-Bat86	2003
E0X102	E0X102_9M ONO	U	Marburg marburgvirus	253	11/2/10			
C7B288	C7B288_9M ONO	U	Marburg marburgvirus	253	9/22/09			
C7B253	C7B253_9M ONO	U	Marburg marburgvirus	253	9/22/09			
C7B281	C7B281_9M ONO	U	Marburg marburgvirus	253	9/22/09			
C7B295	C7B295_9M ONO	U	Marburg marburgvirus	253	9/22/09			
K4MT73	K4MT73_9M ONO	U	Marburg marburgvirus	253	1/9/13			
K4MQD0	K4MQD0_9 MONO	U	Marburg marburgvirus	253	1/9/13			
R4QGQ6	R4QGQ6_9 MONO	U	Marburg marburgvirus	253	7/24/13			
A0A077D3Q8	A0A077D3Q8_9MONO	U	Marburg marburgvirus	253	10/29/14			
L7RF04	L7RF04_9M ONO	U	Reston ebolavirus	251	4/3/13			
L7RE03	L7RE03_9M ONO	U	Reston ebolavirus	251	4/3/13			

Q91DD5	VP24_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	251	7/11/06		Philippines-96	
C6G8D9	C6G8D9_EBORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	251	9/1/09		Philippines-96	
C6G8F5	C6G8F5_EBORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	251	9/1/09		Philippines-96	
C6G8E7	C6G8E7_EBORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	251	9/1/09		Philippines-96	
Q77DB4	VP24_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	251	7/11/06		Reston-89	
B0LPM0	B0LPM0_9MONO	U	Sudan ebolavirus	251	3/18/08			
R4QUG5	R4QUG5_9MONO	U	Sudan ebolavirus	251	7/24/13			
I7ETY5	I7ETY5_9MONO	U	Sudan ebolavirus - Nakisamata	251	10/3/12	Nakisamata		
Q5XX02	VP24_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	251	7/11/06		Uganda-00	
B8XCP3	B8XCP3_9MONO	U	Tai Forest ebolavirus	251	3/3/09			
L7QHW5	L7QHW5_9MONO	U	Zaire ebolavirus	251	4/3/13			
G8DB44	G8DB44_9MONO	U	Zaire ebolavirus	251	1/25/12			
L7QI07	L7QI07_9MONO	U	Zaire ebolavirus	251	4/3/13			
A9QPM3	A9QPM3_9MONO	U	Zaire ebolavirus	251	2/5/08			
A0A089WWJ4	A0A089WWJ4_9MONO	U	Zaire ebolavirus	251	11/26/14			
A0A075WG23	A0A075WG23_9MONO	U	Zaire ebolavirus	251	10/29/14			
O11459	VP24_EBOG4	R	Zaire ebolavirus (strain Gabon-94) (ZEBOV) (Zaire Ebola virus)	251	5/30/00		Gabon-94	
Q6V1Q3	VP24_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	251	7/11/06		Kikwit-95	
Q05322	VP24_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	251	2/1/94		Mayinga-76	

## Appendix A7. Protein Sequences Found in UniProt (L)

Entry	Entry name	Status	Organism	Length	Date of creation	Place	Strain	Year
B8XC�5	B8XC�5_9M ONO	U	Bundibugyo ebolavirus	2210	3/3/09			
R4QUH5	R4QUH5_9M ONO	U	Bundibugyo ebolavirus	2210	7/24/13			
A9Q1H8	A9Q1H8_9M ONO	U	Lake Victoria marburgvirus - Ci67	2331	2/5/08			
Q1PDA3	Q1PDA3_9M ONO	U	Lake Victoria marburgvirus - DRC1999	2327	5/16/06	Democratic Republic of the Congo		1999
Q1PDB0	Q1PDB0_9M ONO	U	Lake Victoria marburgvirus - DRC1999	2331	5/16/06	Democratic Republic of the Congo		1999
Q1PDB7	Q1PDB7_9M ONO	U	Lake Victoria marburgvirus - DRC1999	2331	5/16/06	Democratic Republic of the Congo		1999
G9HWE8	G9HWE8_9 MONO	U	Lake Victoria marburgvirus - Leiden	2331	2/22/12	Leiden		
Q1PD54	L_MABVA	R	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	2331	1/15/08		Angola	2005
Q1PD47	Q1PD47_MA BVA	U	Lake Victoria marburgvirus (strain Angola/2005) (MARV)	2331	5/16/06		Angola	2005
P31352	L_MABVM	R	Lake Victoria marburgvirus (strain Musoke-80) (MARV) (Marburg virus (strain Kenya/Musoke/1980))	2331	7/1/93	Musoke, Kenya	Musok e-80	1980
Q6UY63	L_MABVO	R	Lake Victoria marburgvirus (strain Ozolin-75) (MARV) (Marburg virus (strain South Africa/Ozolin/1975))	2331	1/15/08		Ozolin- 75	
P35262	L_MABVP	R	Lake Victoria marburgvirus (strain Popp-67) (MARV) (Marburg virus (strain West Germany/Popp/1967))	2331	2/1/94	Popp, West Germany	Popp- 67	1967
Q1PDC4	L_MABVR	R	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	2327	1/15/08	Ravn, Kenya	Ravn- 87	1987
A9Q1H1	A9Q1H1_MA BVR	U	Lake Victoria marburgvirus (strain Ravn-87) (MARV) (Marburg virus (strain Kenya/Ravn/1987))	2327	2/5/08	Ravn, Kenya	Ravn- 88	1987
G8EF19	G8EF19_LLO VA	U	Lloviu cuevavirus (isolate Bat/Spain/Asturias- Bat86/2003) (LLOV)	2196	1/25/12	Asturias, Spain	Asturia s- Bat86	2003
K4MR40	K4MR40_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MST1	K4MST1_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MQ40	K4MQ40_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
C7B254	C7B254_9M ONO	U	Marburg marburgvirus	2327	9/22/09			
K4MQ89	K4MQ89_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MQ15	K4MQ15_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MR20	K4MR20_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MR98	K4MR98_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4N3X4	K4N3X4_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MSY7	K4MSY7_9M ONO	U	Marburg marburgvirus	2327	1/9/13			
K4MSU1	K4MSU1_9M ONO	U	Marburg marburgvirus	2331	1/9/13			

K4MQE0	K4MQE0_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4N3H7	K4N3H7_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
C7B296	C7B296_9M ONO	U	Marburg marburgvirus	2331	9/22/09			
K4MT06	K4MT06_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MSW9	K4MSW9_9 MONO	U	Marburg marburgvirus	2331	1/9/13			
K4MSM0	K4MSM0_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
C7B275	C7B275_9M ONO	U	Marburg marburgvirus	2327	9/22/09			
K4MSW2	K4MSW2_9 MONO	U	Marburg marburgvirus	2331	1/9/13			
K4N3U7	K4N3U7_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MSH4	K4MSH4_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MTE3	K4MTE3_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MQB3	K4MQB3_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
K4MR82	K4MR82_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
C7B261	C7B261_9M ONO	U	Marburg marburgvirus	2327	9/22/09			
K4MTB4	K4MTB4_9M ONO	U	Marburg marburgvirus	2331	1/9/13			
C7B289	C7B289_9M ONO	U	Marburg marburgvirus	2331	9/22/09			
C7B282	C7B282_9M ONO	U	Marburg marburgvirus	2331	9/22/09			
A0A077D0Y9	A0A077D0Y9 9MONO	U	Marburg marburgvirus	2331	10/29/14			
R4QIZ2	R4QIZ2_9M ONO	U	Marburg marburgvirus	2331	7/24/13			
E0X103	E0X103_9M ONO	U	Marburg marburgvirus	2331	11/2/10			
L7REU7	L7REU7_9M ONO	U	Reston ebolavirus	2212	4/3/13			
L7RI78	L7RI78_9MO NO	U	Reston ebolavirus	2212	4/3/13			
C6G8E8	C6G8E8_EB ORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	2212	9/1/09			Philippi nes-96
C6G8F6	C6G8F6_EB ORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	2212	9/1/09			Philippi nes-96
C6G8E0	C6G8E0_EB ORE	U	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	2212	9/1/09			Philippi nes-96
Q91DD4	L_EBORE	R	Reston ebolavirus (strain Philippines-96) (REBOV) (Reston Ebola virus)	2212	7/11/06			Philippi nes-96
Q8JPX5	L_EBORR	R	Reston ebolavirus (strain Reston-89) (REBOV) (Reston Ebola virus)	2212	7/11/06			Reston -89
B0LPM1	B0LPM1_9M ONO	U	Sudan ebolavirus	2210	3/18/08			
M4JA89	M4JA89_9M ONO	U	Sudan ebolavirus	2210	5/29/13			
C4PK62	C4PK62_9M ONO	U	Sudan ebolavirus	2210	7/7/09			
R4NUL2	R4NUL2_9M ONO	U	Sudan ebolavirus	2210	7/24/13			
R4QGU1	R4QGU1_9M ONO	U	Sudan ebolavirus	2210	7/24/13			
I7FUR5	I7FUR5_9M ONO	U	Sudan ebolavirus - Nakisamata	2210	10/3/12	Nakisamata		

Q66802	L_EBOSM	R	Sudan ebolavirus (strain Maleo-79) (SEBOV) (Sudan Ebola virus)	2210	5/30/00		Maleo-79	
Q5XX01	L_EBOSU	R	Sudan ebolavirus (strain Uganda-00) (SEBOV) (Sudan Ebola virus)	2210	7/11/06		Uganda-00	
B8XCP4	B8XCP4_9MONO	U	Tai Forest ebolavirus	2210	3/3/09			
G8DB45	G8DB45_9MONO	U	Zaire ebolavirus	2212	1/25/12			
A0A068J9B1	A0A068J9B1_9MONO	U	Zaire ebolavirus	2212	10/1/14			
A0A075W7N6	A0A075W7N6_9MONO	U	Zaire ebolavirus	2212	10/29/14			
A0A075W9P7	A0A075W9P7_9MONO	U	Zaire ebolavirus	2212	10/29/14			
A0A068J465	A0A068J465_9MONO	U	Zaire ebolavirus	2212	10/1/14			
A0A075WAW2	A0A075WAW2_9MONO	U	Zaire ebolavirus	2212	10/29/14			
L7QHR8	L7QHR8_9MONO	U	Zaire ebolavirus	2212	4/3/13			
X5H5B6	X5H5B6_9MONO	U	Zaire ebolavirus	2212	6/11/14			
G8DB54	G8DB54_9MONO	U	Zaire ebolavirus	2212	1/25/12			
A9QPM4	A9QPM4_9MONO	U	Zaire ebolavirus	2212	2/5/08			
Q6V1Q2	L_EBOZ5	R	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	2212	7/11/06		Kikwit-95	
Q05318	L_EBOZM	R	Zaire ebolavirus (strain Mayinga-76) (ZEBOV) (Zaire Ebola virus)	2212	2/1/94		Mayinga-76	

## Appendix B. Modeller Scripts

Note: all the scripts are adapted from the Modeller tutorial.

### 1. align2d.py

```
/* This is the script to align two sequences together. The template is in pdb format, while
 * the target is in PIR (ali) format.
 * All the text in green should be substituted by proper names corresponding to either the
 * template or the target sequence.
 */
```

```
from modeller import *
```

```
env = environ()
aln = alignment(env)
# If a chain other than A is selected as template, please change the letter correspondingly.
mdl = model(env, file='1es6', model_segment=('FIRST:A','LAST:A'))
aln.append_model(mdl, align_codes='1es6A', atom_files='1es6.pdb')
aln.append(file='AIE11808.ali', align_codes='AIE11808')
aln.align2d()
aln.write(file='AIE11808-1es6A.ali', alignment_format='PIR')
aln.write(file='AIE11808-1es6A.pap', alignment_format='PAP')
```

### 2. model-single.py

```
/* This script will build 5 models based on the alignment file in ali format.
 * All the text in green should be substituted by proper names corresponding to either the
 * template or the target sequence.
 */
```

```
from modeller import *
from modeller.automodel import *
#from modeller import soap_protein_od
```

```
env = environ()
a = automodel(env, alnfile='AIE11808-1es6A.ali',
              knowns='1es6A', sequence='AIE11808',
              assess_methods=(assess.DOPE,
                              #soap_protein_od.Scorer(),
                              assess.GA341))
a.starting_model = 1
a.ending_model = 5
a.make()
```



### 3. evaluate\_model.py

```
/* This script generates the "profile" file for the corresponding sequence using the pdb file.
 * All text in green should be substituted with proper file names.
 */

from modeller import *
from modeller.scripts import complete_pdb

log.verbose() # request verbose output
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters

# read model file
mdl = complete_pdb(env, 'AIE11808.B99990005.pdb')

# Assess with DOPE:
s = selection(mdl) # all atom selection
s.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='AIE11808.profile',
              normalize_profile=True, smoothing_window=15)
```

### 4. plot\_profiles.py

```
/* This script plots the DOPE score per residue for both the template and the model, so
 * that we can see the places that need improvement in terms of model-constructing.
 * All the text in green should be substituted by proper names corresponding to either the
 * template or the target sequence.
 */

import pylab
import modeller

def r_enumerate(seq):
    """Enumerate a sequence in reverse order"""
    # Note that we do not use reversed() since Python 2.3 does not have it
    num = len(seq) - 1
    while num >= 0:
        yield num, seq[num]
        num -= 1

def get_profile(profile_file, seq):
    """Read `profile_file` into a Python array, and add gaps corresponding to
    the alignment sequence `seq`."""
    # Read all non-comment and non-blank lines from the file:
```

```

f = file(profile_file)
vals = []
for line in f:
    if not line.startswith('#') and len(line) > 10:
        spl = line.split()
        vals.append(float(spl[-1]))
# Insert gaps into the profile corresponding to those in seq:
for n, res in r_enumerate(seq.residues):
    for gap in range(res.get_leading_gaps()):
        vals.insert(n, None)
# Add a gap at position '0', so that we effectively count from 1:
vals.insert(0, None)
return vals

e = modeller.environ()
a = modeller.alignment(e, file='AIE11808-1es6A.ali')

template = get_profile('1ES6.profile', a['1es6A'])
model = get_profile('AIE11808.profile', a['AIE11808'])

# Plot the template and model profiles in the same plot for comparison:
pylab.figure(1, figsize=(10,6))
pylab.xlabel('Alignment position')
pylab.ylabel('DOPE per-residue score')
pylab.plot(model, color='red', linewidth=2, label='Model')
pylab.plot(template, color='green', linewidth=2, label='Template')
pylab.legend()
pylab.savefig('dope_profile.png', dpi=65)

```

## 5. `salign.py`

# Illustrates the SALIGN multiple structure/sequence alignment

/\* This script aligns the templates together.

\* All the text in green should be substituted by proper names corresponding to either the

\* template or the target sequence.

\*/

```
from modeller import *
```

```
log.verbose()
```

```
env = environ()
```

```
env.io.atom_files_directory = './:../atom_files/'
```

```
aln = alignment(env)
```

```
for (code, chain) in (('1es6', 'A'), ('4ldb', 'B'), ('3tcq', 'A')):
```

```
    mdl = model(env, file=code, model_segment=('FIRST:'+chain, 'LAST:'+chain))
```

```
    aln.append_model(mdl, atom_files=code, align_codes=code+chain)
```

```

for (weights, write_fit, whole) in (((1., 0., 0., 0., 1., 0.), False, True),
                                   ((1., 0.5, 1., 1., 1., 0.), False, True),
                                   ((1., 1., 1., 1., 1., 0.), True, False)):
    aln.salign(rms_cutoff=3.5, normalize_pp_scores=False,
              rr_file='$(LIB)/as1.sim.mat', overhang=30,
              gap_penalties_1d=(-450, -50),
              gap_penalties_3d=(0, 3), gap_gap_score=0, gap_residue_score=0,
              dendrogram_file='multitemp.tree',
              alignment_type='tree', # If 'progresive', the tree is not
                                   # computed and all structues will be
                                   # aligned sequentially to the first
              feature_weights=weights, # For a multiple sequence alignment only
                                   # the first feature needs to be non-zero
              improve_alignment=True, fit=True, write_fit=write_fit,
              write_whole_pdb=whole, output='ALIGNMENT QUALITY')

aln.write(file='multitemp.pap', alignment_format='PAP')
aln.write(file='multitemp.ali', alignment_format='PIR')

aln.salign(rms_cutoff=1.0, normalize_pp_scores=False,
          rr_file='$(LIB)/as1.sim.mat', overhang=30,
          gap_penalties_1d=(-450, -50), gap_penalties_3d=(0, 3),
          gap_gap_score=0, gap_residue_score=0, dendrogram_file='1is3A.tree',
          alignment_type='progressive', feature_weights=[0]*6,
          improve_alignment=False, fit=False, write_fit=True,
          write_whole_pdb=False, output='QUALITY')

```

## 6. align2d\_mult.py

```

/* This script aligns our query with the templates.
 * All the text in green should be substituted by proper names corresponding to either the
 * template or the target sequence.
 */

```

```

from modeller import *

log.verbose()
env = environ()

env.libs.topology.read(file='$(LIB)/top_heav.lib')

# Read aligned structure(s):
aln = alignment(env)
aln.append(file='multitemp.ali', align_codes='all')
aln_block = len(aln)

```

```
# Read aligned sequence(s):
aln.append(file='AIE11808.ali', align_codes='AIE11808')

# Structure sensitive variable gap penalty sequence-sequence alignment:
aln.salign(output='', max_gap_length=20,
           gap_function=True, # to use structure-dependent gap penalty
           alignment_type='PAIRWISE', align_block=aln_block,
           feature_weights=(1., 0., 0., 0., 0., 0.), overhang=0,
           gap_penalties_1d=(-450, 0),
           gap_penalties_2d=(0.35, 1.2, 0.9, 1.2, 0.6, 8.6, 1.2, 0., 0.),
           similarity_flag=True)

aln.write(file='VP40-mult.ali', alignment_format='PIR')
aln.write(file='VP40-mult.pap', alignment_format='PAP')
```

## 7. model\_mult.py

```
/* This script builds the model using the alignment information.
 * All the text in green should be substituted by proper names corresponding to either the
 * template or the target sequence.
 */

from modeller import *
from modeller.automodel import *

env = environ()
a = automodel(env, alnfile='VP40-mult.ali',
             knowns=('1es6A', '4ldbB', '3tcqA'), sequence='AIE11808')
a.starting_model = 1
a.ending_model = 5
a.make()
```

## Appendix C. Genome Selection for Family-Wise Sequence Alignment

Genomes:

Lloviu cuevavirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/355469071?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/355469071?report=graph</a>
Marburg Marburgvirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/158539108?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/158539108?report=graph</a>
Sudan Ebolavirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/55770807?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/55770807?report=graph</a>
Reston Ebolavirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/22789222?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/22789222?report=graph</a>
Zaire Ebolavirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/10313991?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/10313991?report=graph</a>
Tai Forest Ebolavirus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/302315369?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/302315369?report=graph</a>
Bundibugyo virus	<a href="http://www.ncbi.nlm.nih.gov/nuccore/302371213?report=graph">http://www.ncbi.nlm.nih.gov/nuccore/302371213?report=graph</a>

### 1. NP

Lloviu cuevavirus	YP_004928135.1
Marburg Marburgvirus	YP_001531153.1
Sudan Ebolavirus	YP_138520.1
Reston Ebolavirus	NP_690580.1
Zaire Ebolavirus	NP_066243.1
Tai Forest Ebolavirus	YP_003815423.1
Bundibugyo virus	YP_003815432.1

### 2. VP35

Lloviu cuevavirus	YP_004928136.1
Marburg Marburgvirus	YP_001531154.1
Sudan Ebolavirus	YP_138521.1
Reston Ebolavirus	NP_690581.1
Zaire Ebolavirus	NP_066244.1
Tai Forest Ebolavirus	YP_003815424.1
Bundibugyo virus	YP_003815433.1

### 3. VP40

Lloviu cuevavirus	YP_004928137.1
Marburg Marburgvirus	YP_001531155.1
Sudan Ebolavirus	YP_138522.1
Reston Ebolavirus	NP_690582.1
Zaire Ebolavirus	NP_066245.1
Tai Forest Ebolavirus	YP_003815425.1
Bundibugyo virus	YP_003815434.1

### 4. GP

Lloviu cuevavirus	YP_004928138.1 & YP_004928139.1
Marburg Marburgvirus	YP_001531156.1

Sudan Ebolavirus	YP_138523.1
Reston Ebolavirus	NP_690583.1
Zaire Ebolavirus	NP_066246.1
Tai Forest Ebolavirus	YP_003815426.1
Bundibugyo virus	YP_003815435.1

**5. VP30**

Lloviu cuevavirus	YP_004928140.1
Marburg Marburgvirus	YP_001531157.1
Sudan Ebolavirus	YP_138525.1
Reston Ebolavirus	NP_690585.1
Zaire Ebolavirus	NP_066249.1
Tai Forest Ebolavirus	YP_003815429.1
Bundibugyo virus	YP_003815438.1

**6. VP24**

Lloviu cuevavirus	YP_004928142.1
Marburg Marburgvirus	YP_001531158.1
Sudan Ebolavirus	YP_138526.1
Reston Ebolavirus	NP_690586.1
Zaire Ebolavirus	NP_066250.1
Tai Forest Ebolavirus	YP_003815430.1
Bundibugyo virus	YP_003815439.1

**7. L**

Lloviu cuevavirus	YP_004928143.1
Marburg Marburgvirus	YP_001531159.1
Sudan Ebolavirus	YP_138527.1
Reston Ebolavirus	NP_690587.1
Zaire Ebolavirus	NP_066251.1
Tai Forest Ebolavirus	YP_003815431.1
Bundibugyo virus	YP_003815440.1

## Appendix D. Conserved and Diverse Residues Mapping

### Color Code:

Basic color -> Light grey

Conserved residues -> Blue

Diverse residues with no gaps -> Red

Diverse residues with up to 2 gaps -> Magenta

### 1. GP

Conserved:

26,34,49,52,53,57,60,67,70,72,82,84,85,86,88,89,92,94,95,97,102,103,108,109,110,118,122,126,133,135,139,143,146,152,154,157,159,160,161,162,166,167,168,171,176,178,179,182,185,250,286,359,363,366,374,376,379,380,381,384,385,388,389,399,404,406,407,409,410,411,413,417,419,421,422,424,425,426,428,429,434,435,437,438,439,440,441,442,444,445,446,447,448,449,452,453,454,456,457,458,462,466,467,471,474,477,493,494,496,503,510,515,516,518,520

Diverse (No Gaps):

15,31,32,37,39,44,45,47,50,61,64,73,79,98,112,116,124,151,156,187,189,190,191,192,194,195,197,199,203,207,212,214,216,219,221,222,223,224,227,228,229,231,232,235,236,238,239,241,243,245,246,248,254,261,262,263,264,265,266,269,282,283,284,285,297,299,302,310,311,312,316,325,326,327,329,330,334,335,338,339,340,343,344,345,347,348,350,351,352,353,354,356,357,368,369,371,372,375,401,479,482,485,486,487,488,490,496,506,517,523,524

Diverse (Up to 2 Gaps):

1,2,3,4,5,6,7,8,10,11,12,13,14,43,51,205,206,272,273,274,275,276,277,278,279,280,281,308,309,313,314,315,317,318,319,320,321,337,478

Note:

1. 1 residue classified as conserved, 80 residues classified as diverse ones with no gaps, and 51 residues classified as diverse ones with up to 2 gaps cannot be mapped in the model, because when constructing the models, we deleted the 150 residues in the middle in order to have it aligned better with our template.
2. The **red numbers** at the end are not shown in the visualization, since the file GP\_No\_free\_end.pdb only contains residues up to 449. (The original pdb file GP.B99990003.pdb has up to 524 residues but contains large amount of free ends, which are not significantly useful.)
3. There is one mutation in our model in the conserved residues, 82. (All the representatives from different species give us A, but our model is V.)

### 2. L

*L2 (In Query: 384-558; In Aln: 394-568)*

Conserved:

2,5,6,7,9,14,16,18,22,24,26,27,28,29,34,41,43,44,49,52,53,54,56,60,63,65,66,75,76,77,79,83,86,87,88,89,90,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,110,111,112,114,1

15,119,120,121,122,124,125,128,131,132,133,134,135,136,137,138,140,143,144,152,155,  
157,158,160,163,164,165,166,167,168,169,170,171,172,173,174,175

Diverse (No Gaps):

1,12,19,31,57,61,73,82,108,126,129,147,150,156,161

Diverse (Up to 2 Gaps):

None

### ***L1 (In Query: 2043-2154; In Aln: 2187-2299)***

Conserved:

1,10,11,18,19,20,22,29,30,40,57,60,61,73,74,75,80,82,83,84,87,88,90,94,98

Diverse (No Gaps):

3,7,8,9,12,13,16,36,37,41,44,45,48,49,51,59,64,66,78,93,97,101,103,104,106,107,110,111

Diverse (Up to 2 Gaps):

None

## **3. NP**

### ***NP3 (In Query: 45-110; In Aln: 46-111)***

Conserved:

9,13,19,21,30,33,34,37,39,42,46,49,54,55,59,62

Diverse (No Gaps):

1,3,7,11,25,43,44,50,61,64,65

Diverse (Up to 2 Gaps):

None

### ***NP2 (In Query: 265-340; In Aln: 266-341)***

Conserved:

1,2,3,5,6,8,9,12,16,17,19,20,23,24,26,28,29,30,31,32,33,34,36,37,38,39,40,42,43,44,45,46,47,48,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,74,75,76

Diverse (No Gaps):

14,21,25

Diverse (Up to 2 Gaps):

None

### ***NP1 (In Query: 641-739; In Aln: 695-793)***

Conserved:

35,49,57,58,82,85



Diverse (No Gaps):

1,2,3,5,6,8,9,13,17,30,32,33,37,41,42,47,54,55,56,61,65,66,67,72,73,74,75,76,77,81,87,88,94

Diverse (Up to 2 Gaps):

7,99

#### 4. VP24

Conserved:

1,2,7,8,9,37,38,40,42,47,53,56,57,60,61,67,69,72,74,75,76,78,79,80,82,85,90,94,95,98,104,105,110,111,125,126,127,149,160,167,168,169,171,174,177,178,180,181,188,189,190,192,194,195,196,198,199,200,205,219,221,224,241,246,247,250,251

Diverse (No Gaps):

11,12,17,19,32,39,45,50,70,84,88,89,91,103,116,120,130,132,139,140,158,165,173,184,185,186,210,211,212,213,229,230,232,233,234,235,237,239,240

Diverse (Up to 2 Gaps):

231

#### 5. VP30

Conserved:

1,7,8,9,26,29,31,51,65,66,68,69,70,72,78,81,86,89,90,92,95,99,100,101,102,103,105,106,123,126,133,143,144,147,167,168,169,171,173,179,180,184,191,193,197,199,201,202,203,211,215,217,218,219,220,222,223,224,225,226,227,228,230,231,234,238,241,243,251,252,253,254

Diverse (No Gaps):

2,5,12,13,16,17,18,19,21,22,23,24,25,30,33,34,35,36,37,39,40,41,42,45,48,52,53,56,84,112,114,115,116,117,124,130,132,137,138,139,145,146,148,149,153,156,157,158,160,161,165,166,188,192,205,206,221,240,244,262,263,265,267,269,270,271,272,273

Diverse (Up to 2 Gaps):

3,4,14,60,61,62,63,119,274,275,276,277,278,279,280,281,282,283,284,285,286,287

Note:

The **red numbers** at the beginning and the end are not shown in the visualization, since in the file VP30\_nf.pdb residues before 145 and after 270 are chopped off. (The original pdb file AIE11812.B99990005.pdb has 288 residues but contains large amount of free ends, which are not significantly useful.)

#### 6. VP35

*Black/AIE11807 (In Query: 35-262; In Aln: 36-291)*

Conserved:

1,2,41,52,66,84,103,104,106,107,108,109,111,112,115,116,117,119,122,123,124,127,128,  
131,132,136,137,139,158,165,177,180,183,186,187,188,190,197,198,199,200,203,205,206  
,208,210,211,214

Diverse (No Gaps):

7,8,21,22,23,24,25,26,27,28,29,31,33,34,35,36,37,43,44,47,48,49,51,53,54,55,57,58,60,61,6  
2,63,64,65,67,68,69,70,75,78,79,82,88,89,93,100,130,134,145,149,152,153,154,155,156,15  
7,160,161,163,167,170,173,174,184,192,220,221

Diverse (Up to 2 Gaps):

11,12,13,14,15,16,18,19,20,45,46,50,224

### **Basic/VP35\_1 (In Query: 213-340; In Aln: 238-369)**

Conserved:

2,5,8,9,10,12,19,20,21,22,25,27,28,30,32,33,36,38,39,44,45,51,55,58,62,64,65,75,80,81,92,  
95,96,97,98,99,100,103,104,106,108,111,112,113,114,121,126,127,128

Diverse (No Gaps):

6,14,42,43,79,115,117,118

Diverse (Up to 2 Gaps):

46

### **7. VP40**

Conserved:

**1,10,11,13**,53,72,73,75,76,77,90,93,95,96,97,98,99,106,111,114,119,120,121,123,125,133,  
134,136,137,138,139,141,142,143,145,146,147,150,151,153,154,155,156,157,159,162,164  
,172,173,176,177,179,181,187,191,205,208,211,213,215,218,219,264,273,290

Diverse (No Gaps):

**4,5,14,34,35,36,37,38,40,41,42,43**,44,46,50,54,58,62,67,70,78,105,128,129,152,182,197,19  
8,199,200,202,209,220,222,227,228,241,243,245,259,266,269,276,277,278,293,298,299,3  
00,320,**322,323,324,325**

Diverse (Up to 2 Gaps):

**9,16,17,18,20,21,22,23,24,25,26,27,28,29,30,31**,308,309,310,311,312,313,315,316,318

Note:

The **red numbers** at the beginning and the end are not shown in the visualization, since in the file VP40\_nf.pdb residues before 44 and after 321 are chopped off. (The original pdb file AIE11808.B99990005.pdb has 326 residues but contains large amount of free ends, which are not significantly useful.)

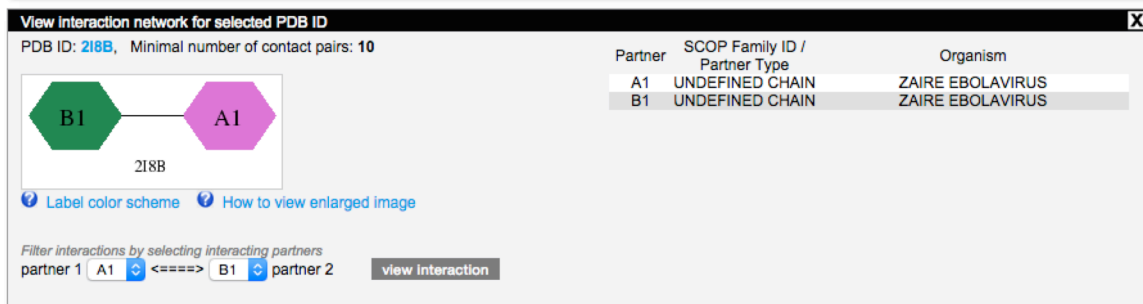


Figure 44: 2I8B B1 A1 Interaction

## Appendix E. Model Check and Functional Annotation

Based on the meeting with Professor Korkin on the following two dates:

05/19/2015: VP30, VP40, VP35, VP24, GP

05/21/2015: GP, NP, L

### 1. VP30

#### **1 Model, 1 Template**

##### **Model:**

Average quality

##### **Annotation:**

1). Template: 2I8B

Interactions: 1 intra-virus interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_interaction/2I8B](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_interaction/2I8B)

2I8B: A interacts with 2I8B: B, which should be identical to A.

In the pdb file, the two chains start from different residues, A from 140, B from 135, but they both end at 266.

The interacting residues in Chain A are exactly the same as those in Chain B.

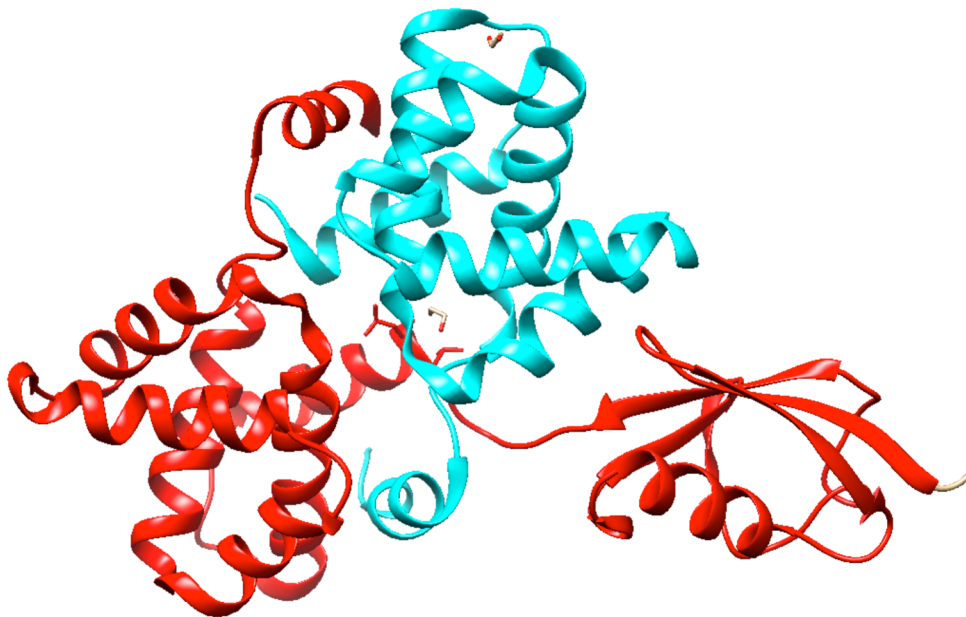


Figure 45: 3V70 Chain A (Cyan) and Chain B (Red)

## 2). Template: 3V70

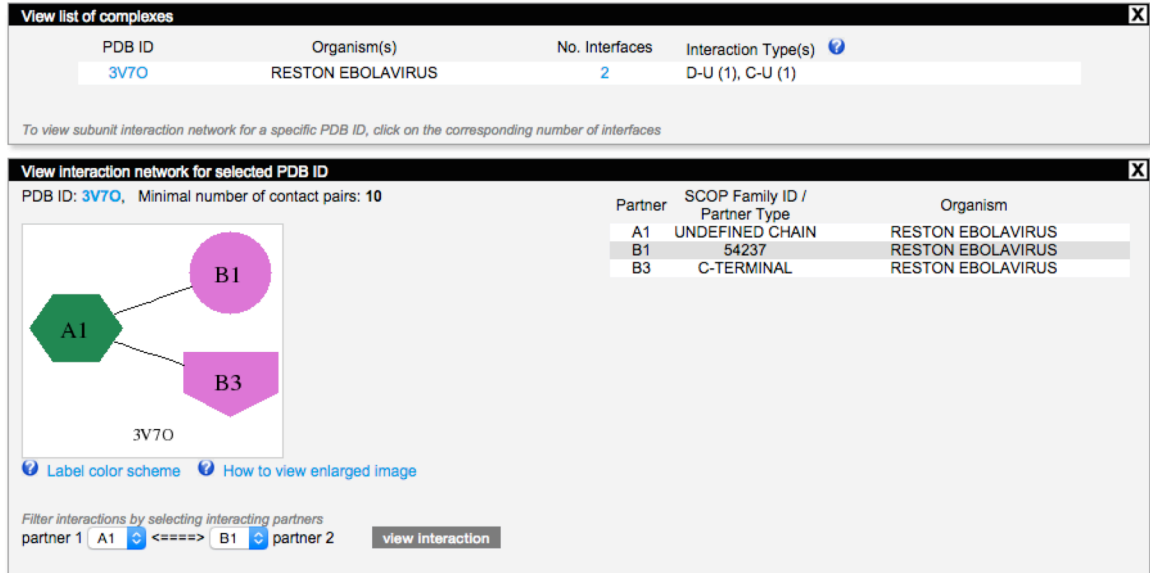


Figure 46: 3V70 Interaction

Interactions: 2 intra-virus interactions

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_interaction/3V70](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_interaction/3V70)

Chain A is identical to Chain B.

The two chains both start from -95. In the interaction, Chain A starts from 0 and ends at 125, and Chain B starts from -76 and ends at 0. The structures available for Chain A and Chain B are different.

SCOP Family: 54237

However, none of the organisms except *Reston Ebolavirus* is closely related to our target.

## 2. VP40

### 1 Model, 3 Templates

#### Model:

Cleave residues before locus 44 and after 321.

#### Annotation:

SCOP family: 50013

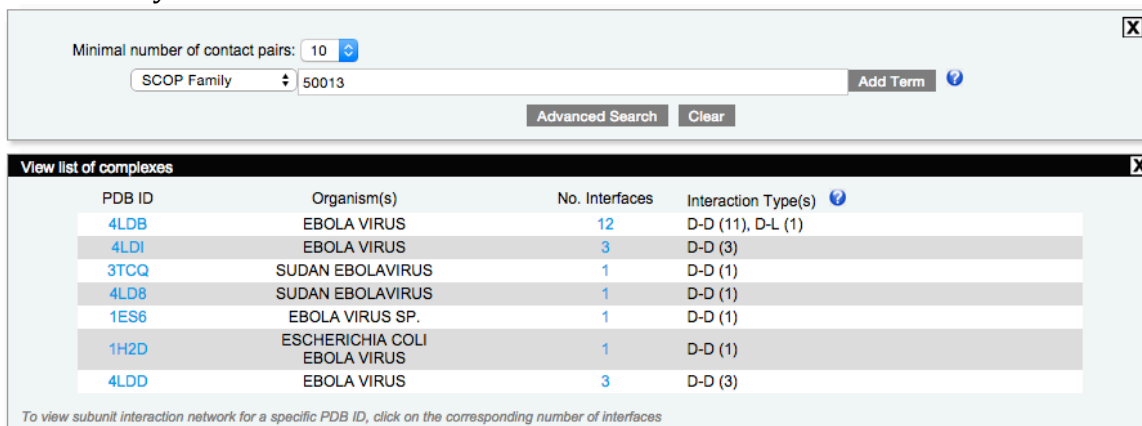


Figure 47: SCOP Family 50013

## 1). Template: 3TCQ (monomer)


PDB ID	Organism(s)	No. Interfaces	Interaction Type(s) 
4N9F	HUMAN IMMUNODEFICIENCY VIRUS 1 HOMO SAPIENS	132	U-U (4), D-U (47), D-D (51), C-U (12), D-C (7), D-N (11)
2J7Q	MURINE CYTOMEGALOVIRUS HOMO SAPIENS	5	D-D (5)
3PSE	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	6	C-N (1), D-C (1), D-N (2), D-D (2)
3PRM	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	5	U-U (1), D-U (4)
3RT3	INFLUENZA B VIRUS HOMO SAPIENS	3	D-N (1), D-D (2)
3R66	INFLUENZA B VIRUS HOMO SAPIENS	10	D-D (8), D-N (2)
3PHW	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS STRAIN HOMO SAPIENS	24	D-D (12), D-N (8), D-C (4)
3DCG	HUMAN IMMUNODEFICIENCY VIRUS TYPE 1 (NEW YORK-5) HOMO SAPIENS	6	D-P (2), D-U (3), D-D (1)
3SDL	INFLUENZA B VIRUS HOMO SAPIENS	10	D-D (8), D-N (2)
4HXD	HOMO SAPIENS DUGBE VIRUS (ISOLATE ARD44313)	10	D-N (5), D-C (2), D-D (3)
4WUR	HOMO SAPIENS HUMAN CORONAVIRUS EMC	1	D-U (1)
3PHX	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	5	D-C (1), D-N (2), D-D (1), C-N (1)
3ZNH	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	4	D-C (1), D-N (2), D-D (1)
4MM3	SARS CORONAVIRUS HOMO SAPIENS	1	D-U (1)
3V7O	RESTON EBOLAVIRUS HOMO SAPIENS	2	D-U (1), C-U (1)
2MA9	HUMAN IMMUNODEFICIENCY VIRUS TYPE 1 HOMO SAPIENS	4	D-D (1), D-C (2), D-U (1)
4IUM	HOMO SAPIENS EQUINE ARTERITIS VIRUS	1	D-U (1)
4RF1	HOMO SAPIENS HUMAN BETACORONAVIRUS 2C JORDAN- N3/2012	1	D-U (1)
3PRP	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	13	D-D (5), D-N (4), C-N (2), D-C (2)
3PT2	CRIMEAN-CONGO HEMORRHAGIC FEVER VIRUS HOMO SAPIENS	1	D-U (1)
4RF0	HOMO SAPIENS HUMAN BETACORONAVIRUS 2C JORDAN- N3/2012	1	D-U (1)
4M0W	SARS CORONAVIRUS BOS TAURUS	1	D-U (1)

Figure 48: SCOP Family 54237

Interaction: 1 intra-virus interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3TCQ](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3TCQ)

3TCQ only has chain A.

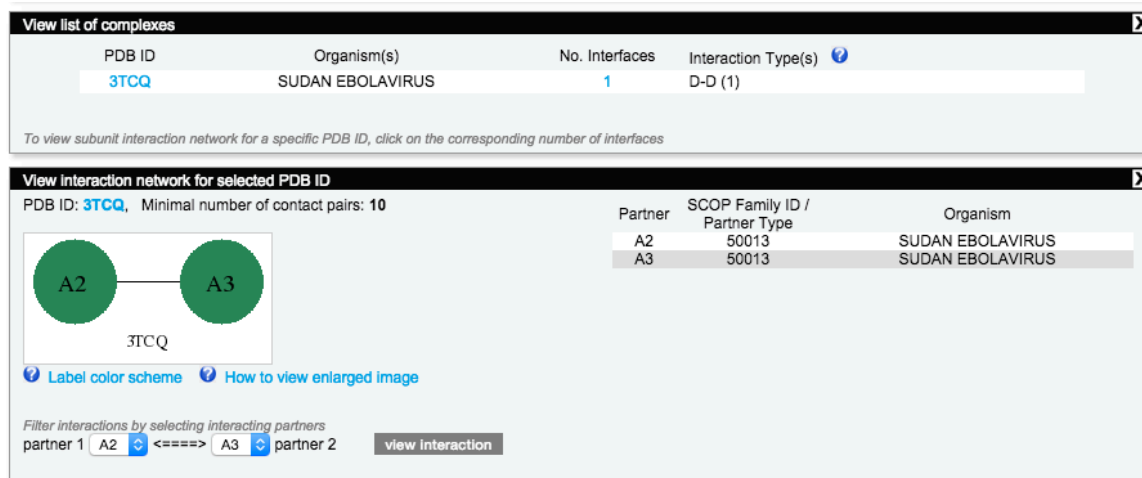


Figure 49: 3TCQ Interaction

2). Template: 4LDB (tetramer), 12 intra-virus interactions

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/4LDB](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/4LDB)

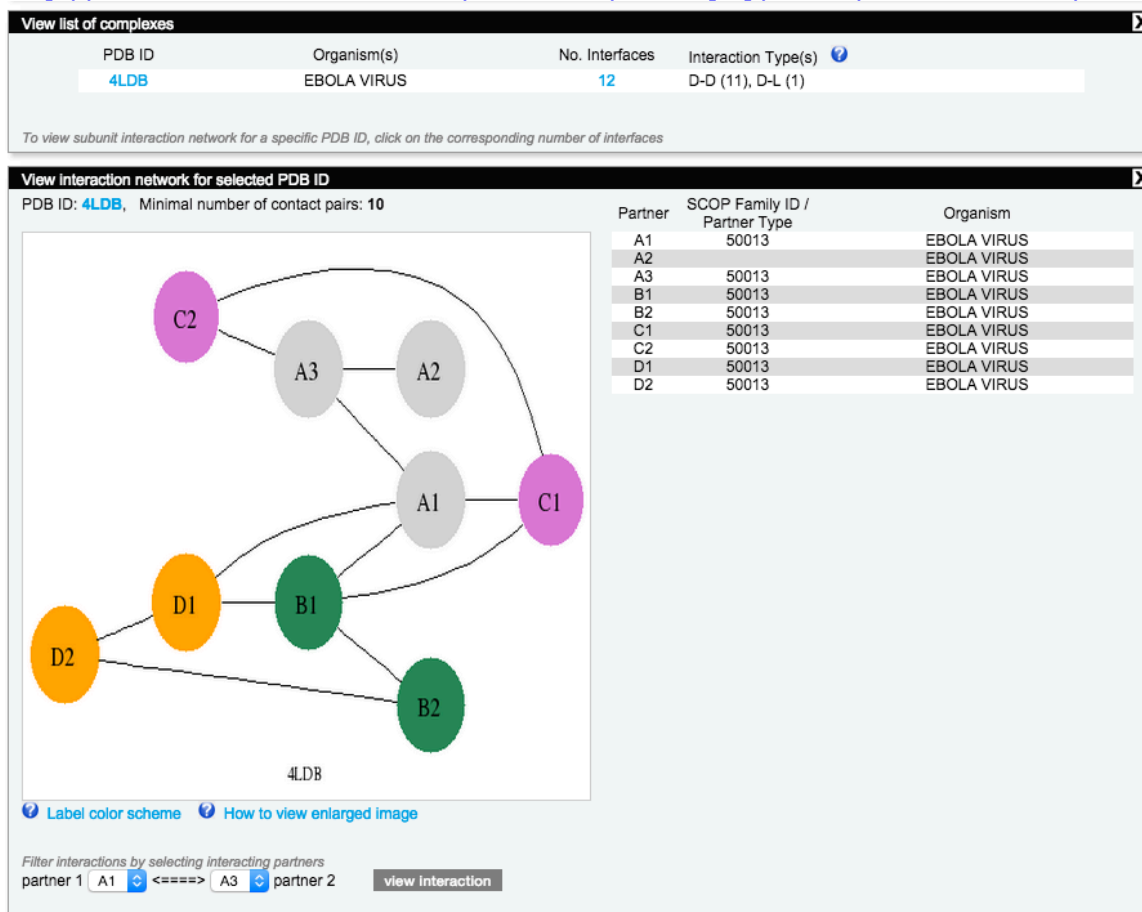


Figure 50: 4LDB Interaction

A, B, C, and D chains are identical in terms of sequence, but they are slightly different in terms of structure. Since the sequences are identical, there is no need to align them one by one with our target. A single alignment with the template would suffice. The alignment shows that the untruncated sequence has the exactly same residue at each position after position 44.

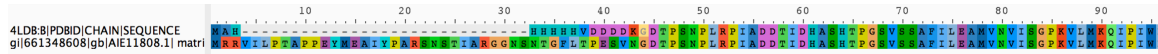


Figure 51: Alignment Between 4LDB and VP40

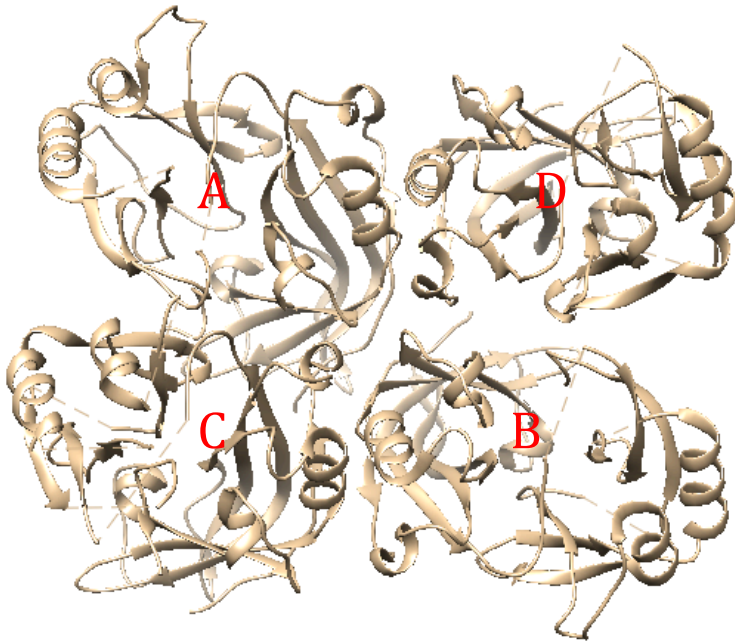


Figure 52: Structure of 4LDB

3). 4LDI/4LDD (dimer) -> 3 interactions

Interaction: 3 intra-virus interactions

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_interaction/4LDI](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_interaction/4LDI)

4LDI has two identical chains.

The alignment result shows that 4LDI is exactly the same as 4LDB except at one position (212).

### 3. VP35

**2 Models, 1 Template for Each -> 1 Model, 1 Template** (See Model section)

**Model:**

Remove the first model, since the template is too bad.

The main model is of average quality.

**Annotation:**

1). Template: 3FKE

Interaction: 1 undefined intra-virus interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3FKE](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3FKE)

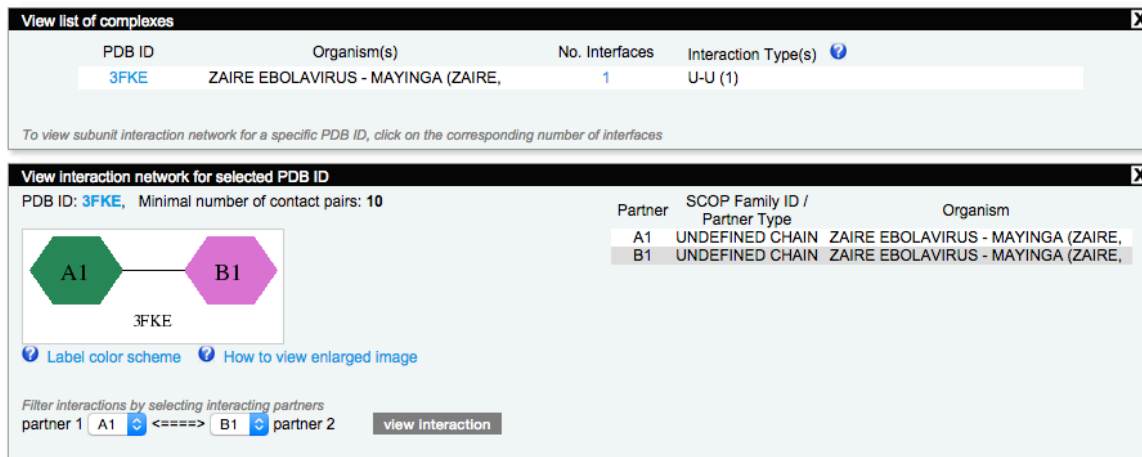


Figure 53: 3FKE Interaction

Chain A and Chain B are identical to each other.

2). Other candidate template: 3KS8

Interactions: 2 undefined intra-virus interactions, 1 RNA-RNA interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3KS8](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3KS8)

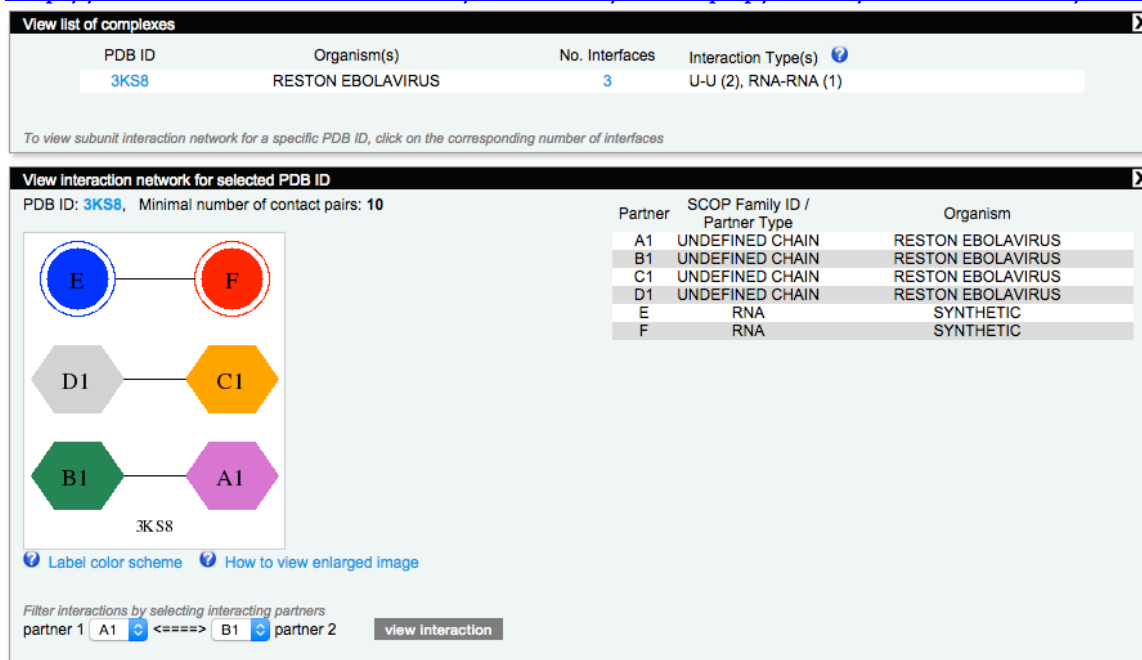


Figure 54: 3KS8 Interaction

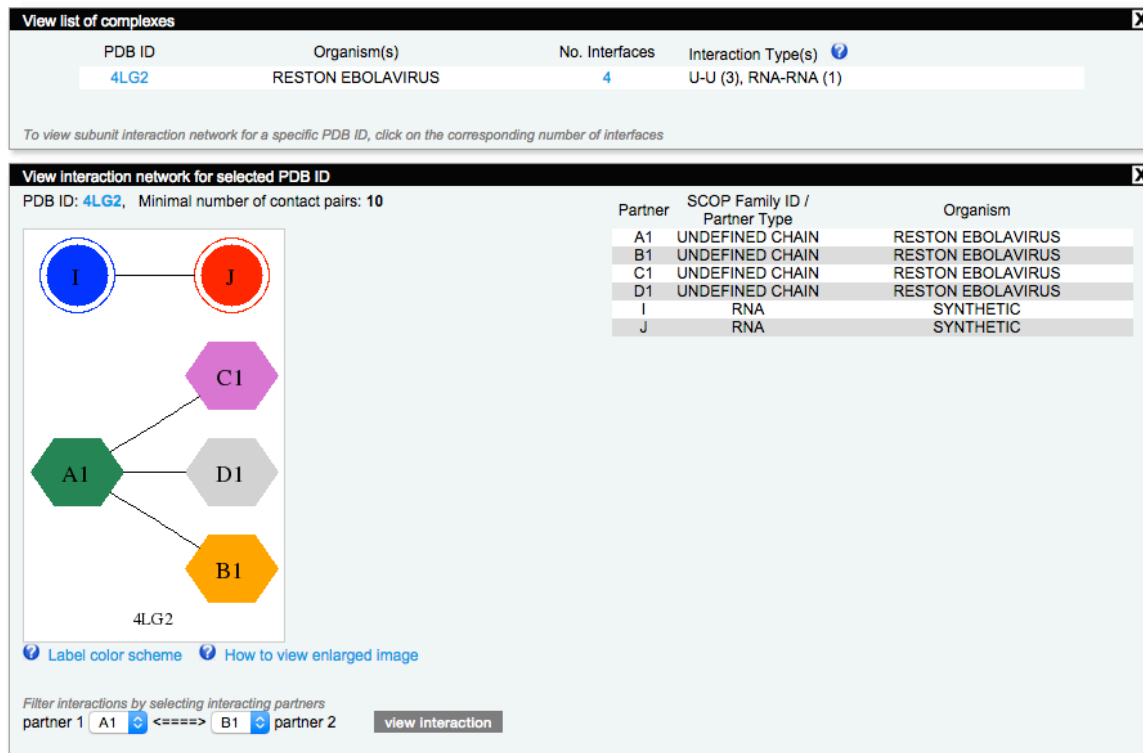
A, B, C, and D chains are identical. The four chains are almost identical in terms of structure. There are also two chains of RNA. As a result, the interaction pairs between Chain A and Chain B are exactly the same as those between Chain C and Chain D.

3). Other candidate template: 4LG2

Interactions: 3 undefined intra-virus interactions, 1 RNA-RNA interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/4LG2](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/4LG2)





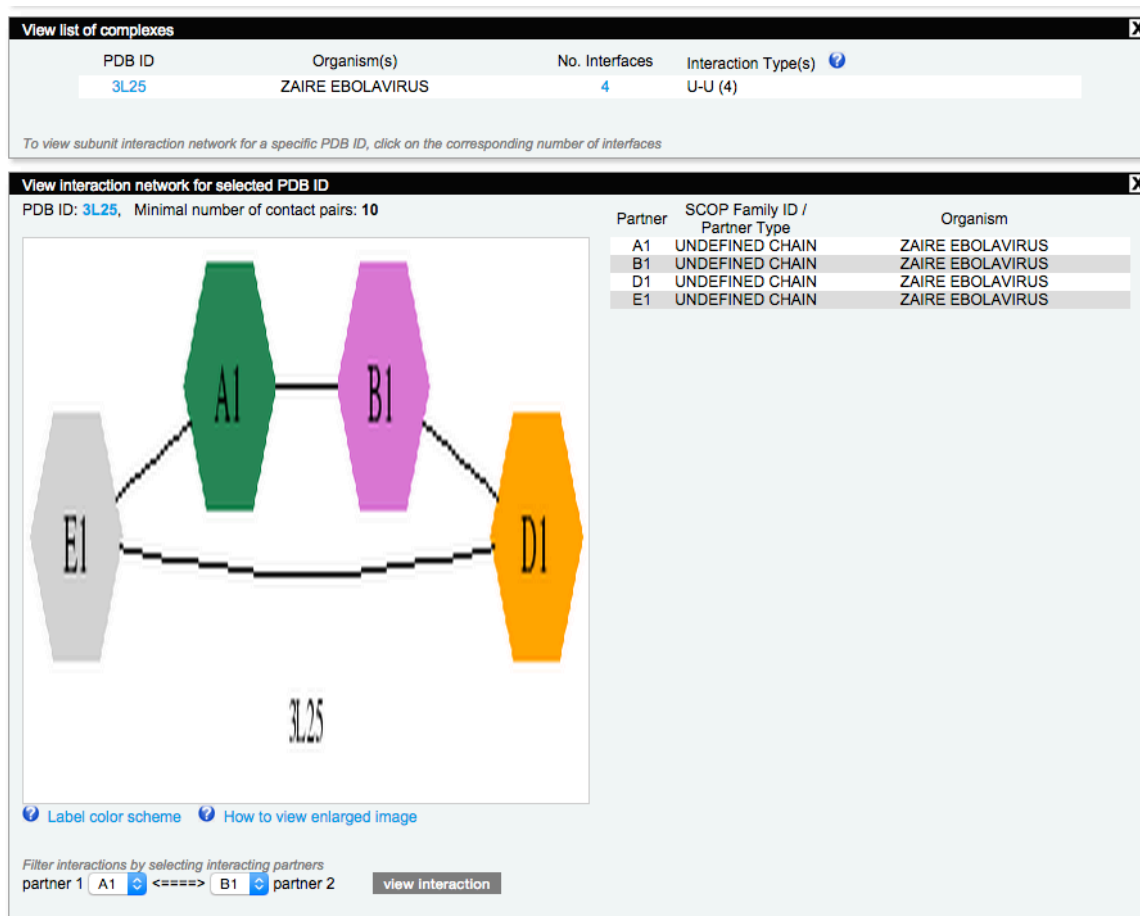
**Figure 55: 4LG2 Interaction**

A, B, C, and D Chains are identical. The four chains are almost identical in terms of structure. There are also four chains of RNA.

4). Other candidate template: 3L25

Interactions: 4 undefined intra-virus interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3L25](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3L25)



**Figure 56: 3L25 Interaction**

A, B, C, and D Chains are identical. The four chains are almost identical in terms of structure. There are also two chains of RNA.

The residues involved in the interaction between Chain A and Chain E are the same as those between Chain B and Chain D.

5). Other candidate template: 3L28

Interactions: 6 undefined intra-virus interactions

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3L28](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3L28)

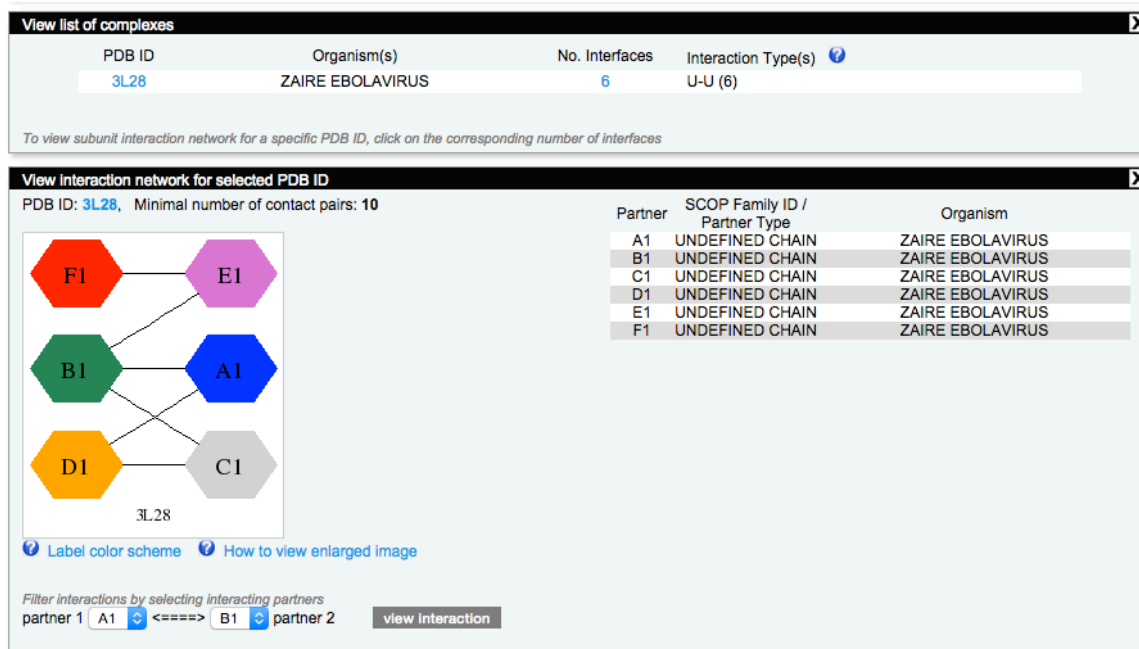


Figure 57: 3L28 Interaction

There are 6 identical chains in this protein. The sequence in each chain is also identical to that in 3L25, differing only at position 339 (3L25: K; 3L28: A). So, the alignment is similar to that with 3L25.

6). Other candidate template: 4GHL

Interactions: 3 undefined intra-virus interactions, 1 RNA-RNA interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/4GHL](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/4GHL)

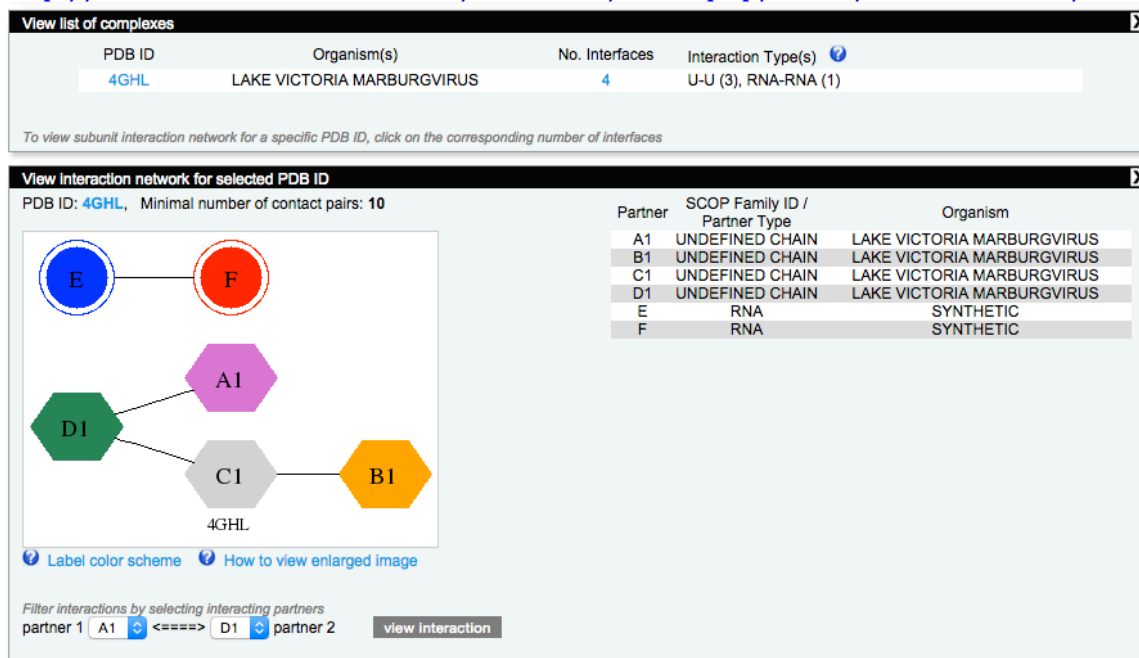


Figure 58: 4GHL Interaction

There are four identical chains and two RNA chains in this protein.

Note, that since the template we used to build this model only contains 1 undefined interaction, we continued to search for related structures in PDB using advanced search.

#### 4. VP24

##### **Model:**

Average quality. All termini were cleaved.

##### **Annotation:**

1). Template: 4M0Q

Interactions: 1 undefined intra-virus interaction

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/4M0Q](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/4M0Q)

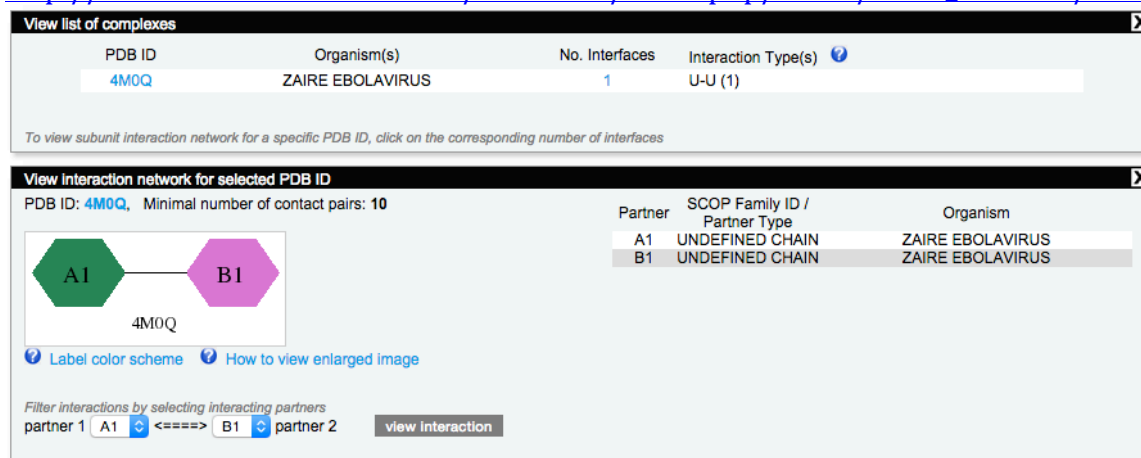


Figure 59: 4M0Q Interaction

The two chains are identical. The interaction pairs are almost symmetric between Chain A and Chain B.

2). Other candidate template: 4U2X (human-viral interaction)

Interactions: 2 U-U intra-virus interactions, 3 D-U host-viral interactions

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/4U2X](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/4U2X)

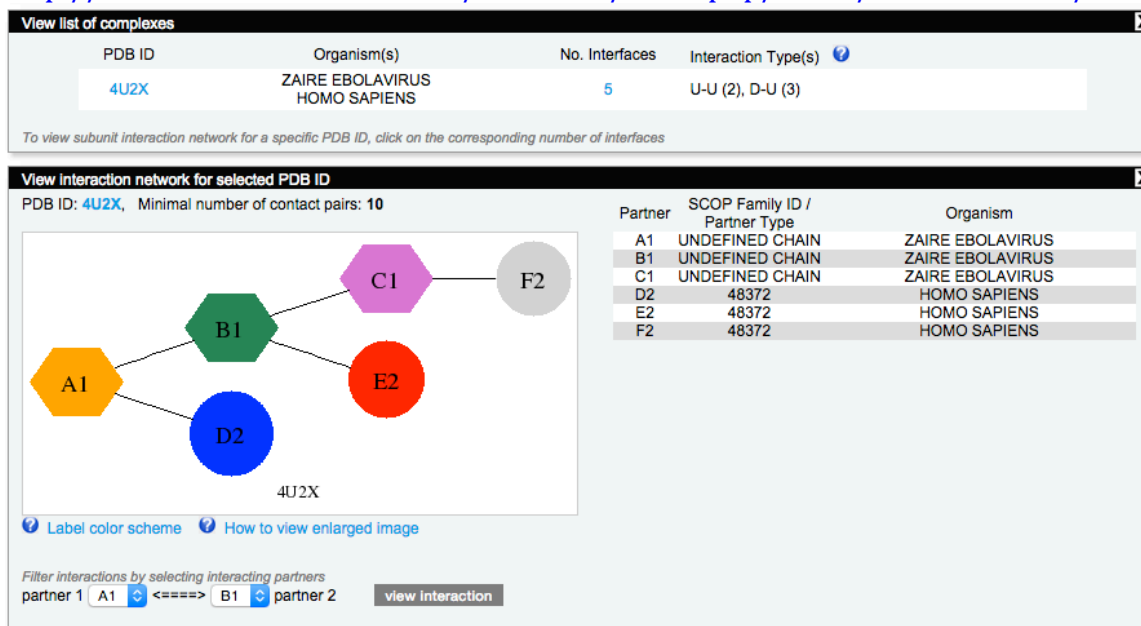


Figure 60: 4U2X Interaction

Chains A, B, and C are identical and are from *Zaire Ebolavirus*, while Chains D, E, and F are identical and are from *homo sapiens*.

## 5. GP

### **Model:**

Cleaved the middle part and rebuilt the model based on 3CSY template.

### **Annotation:**

1). Template: 3CSY

Interactions: D-N (16), D-D (23), D-L (7), D-U (11), N-U (4)

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3CSY](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3CSY)

PDB ID	Organism(s)	No. Interfaces	Interaction Type(s)
3CSY	ZAIRE EBOLA VIRUS HOMO SAPIENS	61	D-N (16), D-D (23), D-L (7), D-U (11), N-U (4)

To view subunit interaction network for a specific PDB ID, click on the corresponding number of interfaces

**Figure 61: 3CSY Interaction**

The structure **3CSY** has in total **16** chains. These are represented by **4** sequence-unique entities. [Crystal structure of the trimeric prefusion Ebola virus glycoprotein (GP1+GP2) in complex with a neutralizing antibody from a human survivor]

Fab KZ52 heavy chain: ACEG  
Fab KZ52 light chain: BDFH  
Envelop Glycoprotein GP2: JLNP  
Envelop Glycoprotein GP1: IKMO

There are 21 intra-virus interactions and 13 host-virus interactions.

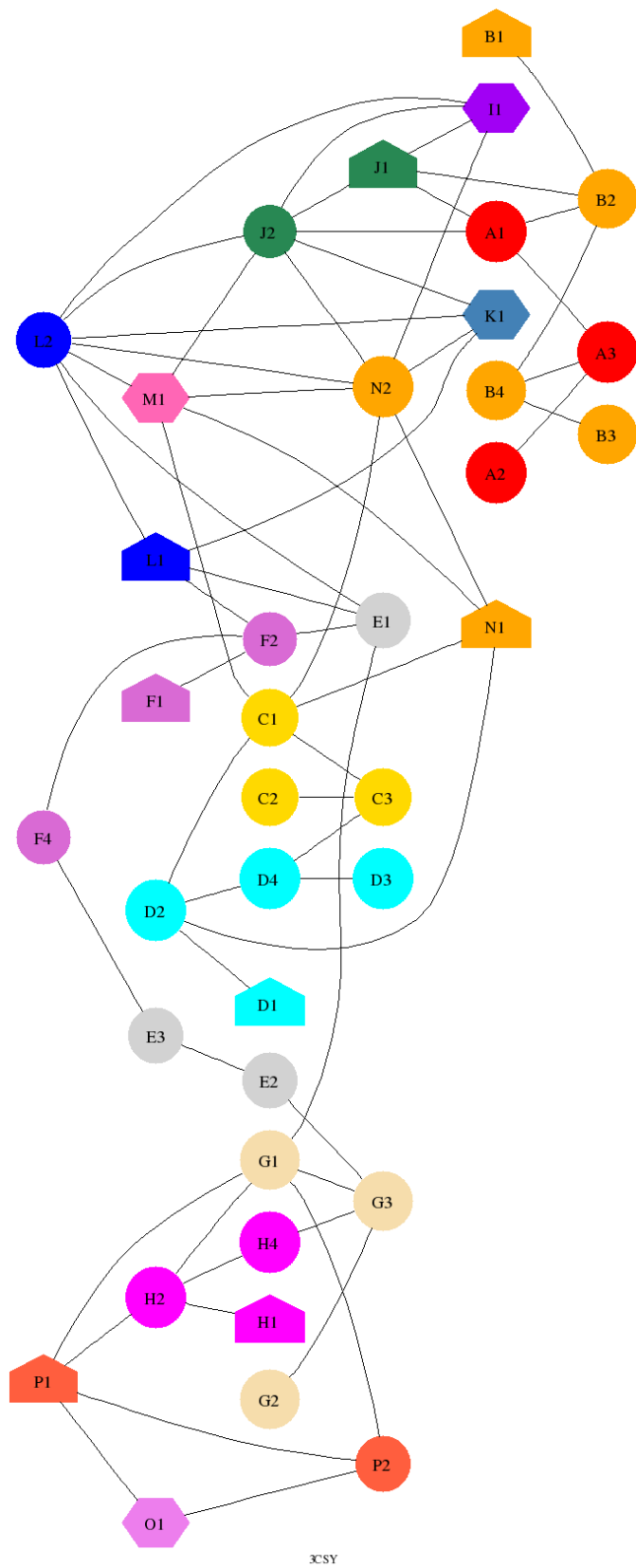


Figure 62: 3CSY Interaction Map Enlarged

Partner	SCOP Family ID / Partner Type	Organism
A1	48727	HOMO SAPIENS
A2		HOMO SAPIENS
A3	48942	HOMO SAPIENS
B1	N-TERMINAL	HOMO SAPIENS
B2	48727	HOMO SAPIENS
B3		HOMO SAPIENS
B4	48942	HOMO SAPIENS
C1	48727	HOMO SAPIENS
C2		HOMO SAPIENS
C3	48942	HOMO SAPIENS
D1	N-TERMINAL	HOMO SAPIENS
D2	48727	HOMO SAPIENS
D3		HOMO SAPIENS
D4	48942	HOMO SAPIENS
E1	48727	HOMO SAPIENS
E2		HOMO SAPIENS
E3	48942	HOMO SAPIENS
F1	N-TERMINAL	HOMO SAPIENS
F2	48727	HOMO SAPIENS
F4	48942	HOMO SAPIENS
G1	48727	HOMO SAPIENS
G2		HOMO SAPIENS
G3	48942	HOMO SAPIENS
H1	N-TERMINAL	HOMO SAPIENS
H2	48727	HOMO SAPIENS
H4	48942	HOMO SAPIENS
I1	UNDEFINED CHAIN	ZAIRE EBOLA VIRUS
J1	N-TERMINAL	ZAIRE EBOLA VIRUS
J2	58070	ZAIRE EBOLA VIRUS
K1	UNDEFINED CHAIN	ZAIRE EBOLA VIRUS
L1	N-TERMINAL	ZAIRE EBOLA VIRUS
L2	58070	ZAIRE EBOLA VIRUS
M1	UNDEFINED CHAIN	ZAIRE EBOLA VIRUS
N1	N-TERMINAL	ZAIRE EBOLA VIRUS
N2	58070	ZAIRE EBOLA VIRUS
O1	UNDEFINED CHAIN	ZAIRE EBOLA VIRUS
P1	N-TERMINAL	ZAIRE EBOLA VIRUS
P2	58070	ZAIRE EBOLA VIRUS

Figure 63: 3CSY Interaction Partners Enlarged

## 2). 3S88 (Sudan Ebolavirus)

Interactions: D-U (2), D-L (2), D-D (6), N-U (1), D-C (1), D-N (1)

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3S88](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3S88)

The structure **3S88** has in total **4** chains. These are represented by **4** sequence-unique entities.

16F6 heavy chain: H

16F6 light chain; L

Envelop glycoprotein: I, J

There are 3 intra-viral interactions and 3 host-virus interactions.

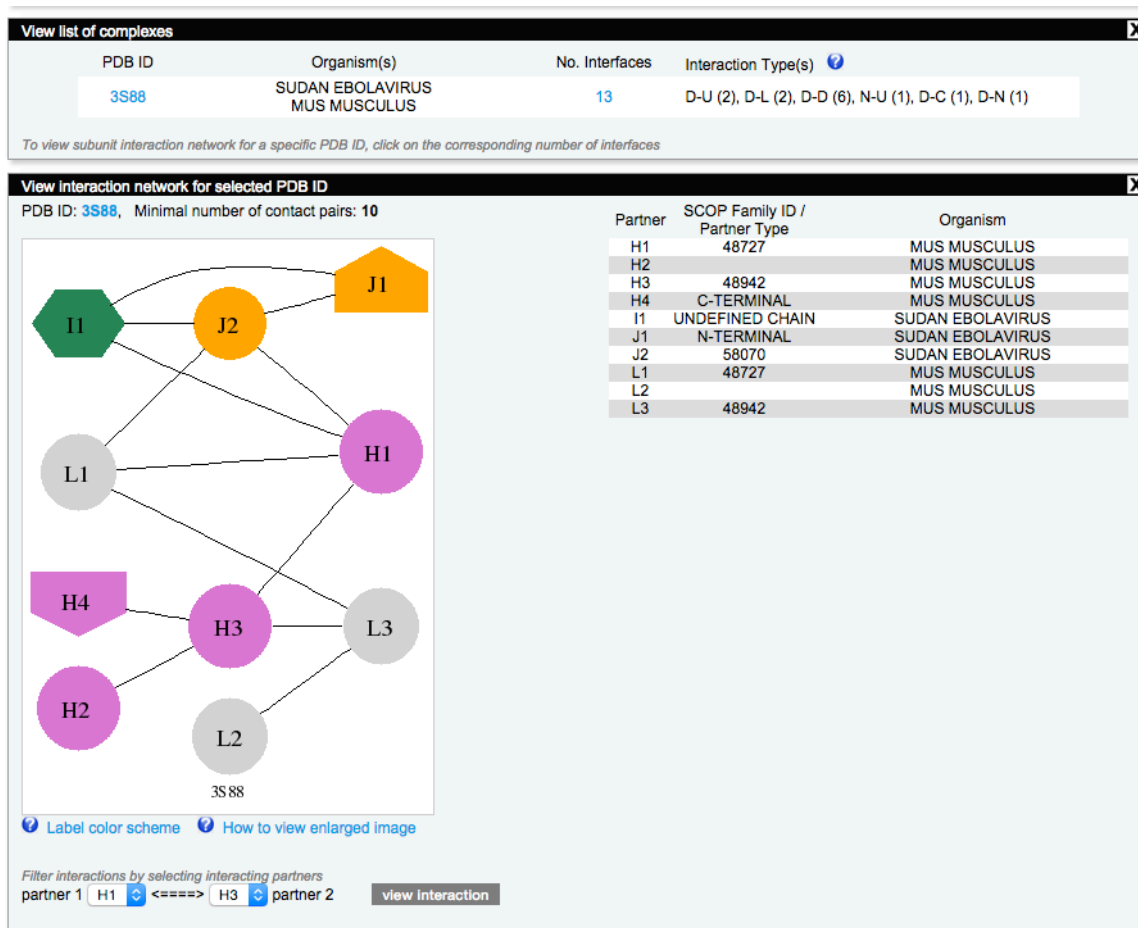
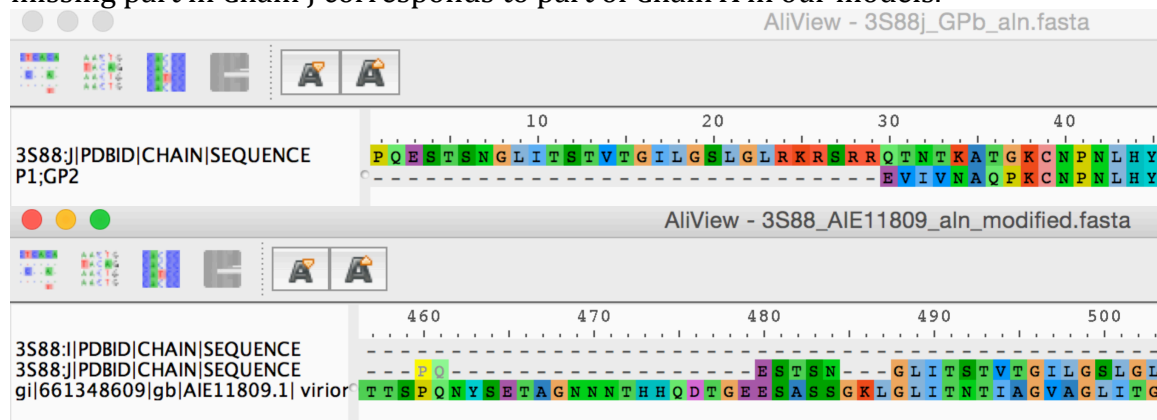


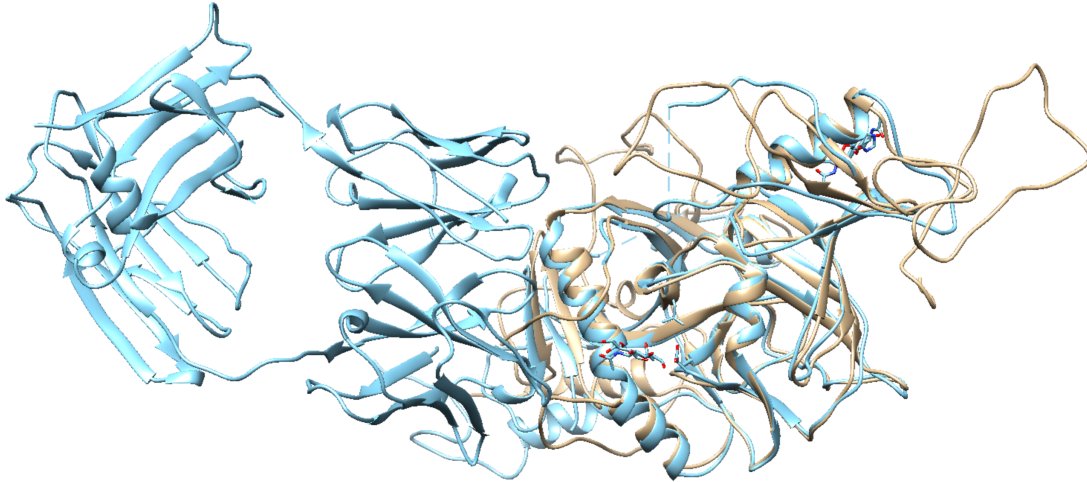
Figure 64: 3S88 Interaction

It seems that the cutoff between the two chains in 3S88 is different from the one in 3CSY. The alignment therefore is a little different. As shown in the two pictures below, the missing part in Chain J corresponds to part of Chain A in our models.



Below is a screenshot of the superimposed GP model and 3S88. But the misaligned part does not have structural information in 3S88. So, it will not affect the calculation of the scores.



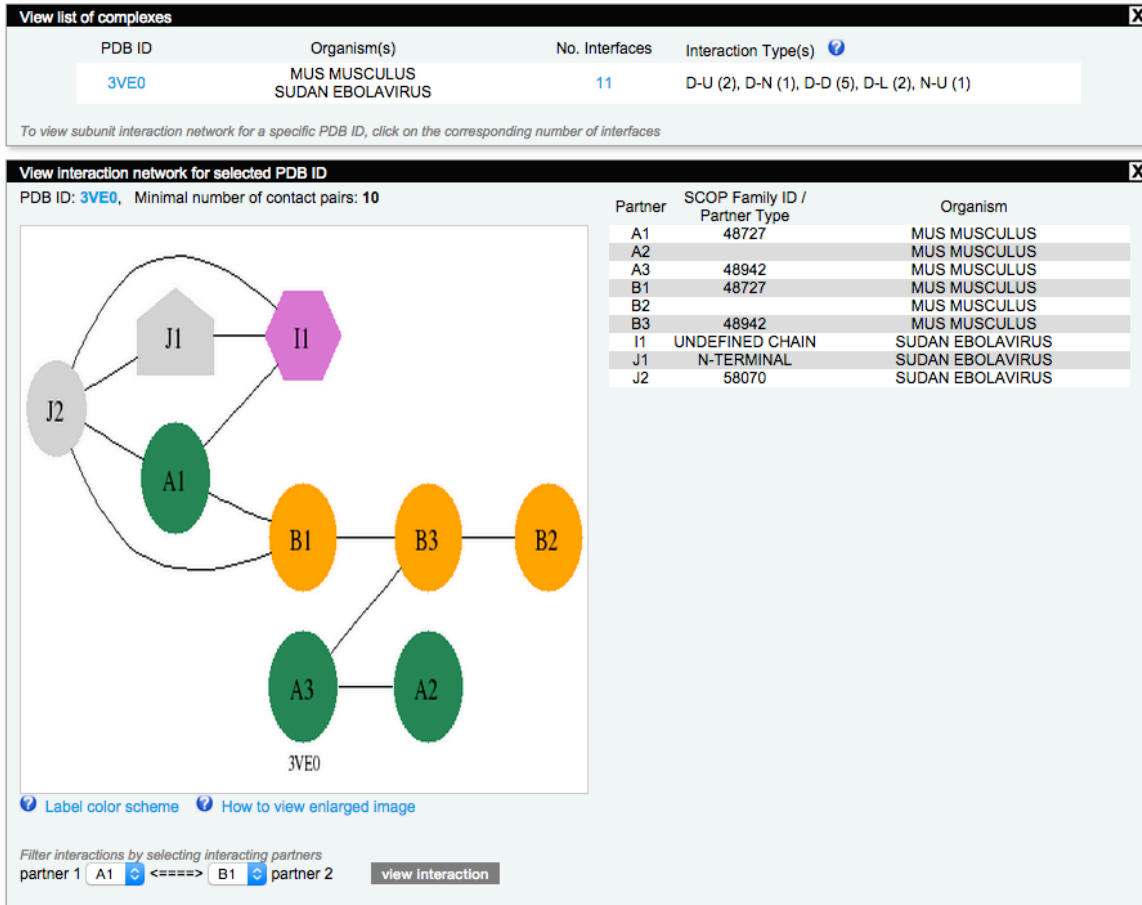


### 3). 3VE0 (Sudan Ebolavirus)

Interactions: D-U (2), D-N (1), D-D (5), D-L (2), N-U (1)

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/3VE0](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/3VE0)

The structure **3VE0** has in total **4** chains. These are represented by **4** sequence-unique entities.



**Figure 65: 3VE0 Interaction**

16F6 Antibody	A
16F6 Antibody	B
Envelope glycoprotein	I

## Envelope glycoprotein J

Similar to 3S88, the beginning of Chain J in 3EV0 can probably be aligned to the portion excluded in the model, but since the structural information for that part is missing in PDB, this will not affect our result. So, there is no need to reconstruct the model.

### 4). 1EBO

Interactions: D-D (8)

[http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show\\_network/1EBO](http://kestrel2.rnet.missouri.edu/dommino/index.php/result/show_network/1EBO)

The structure **1EBO** has in total **6** chains. These are represented by **1** sequence-unique entity.

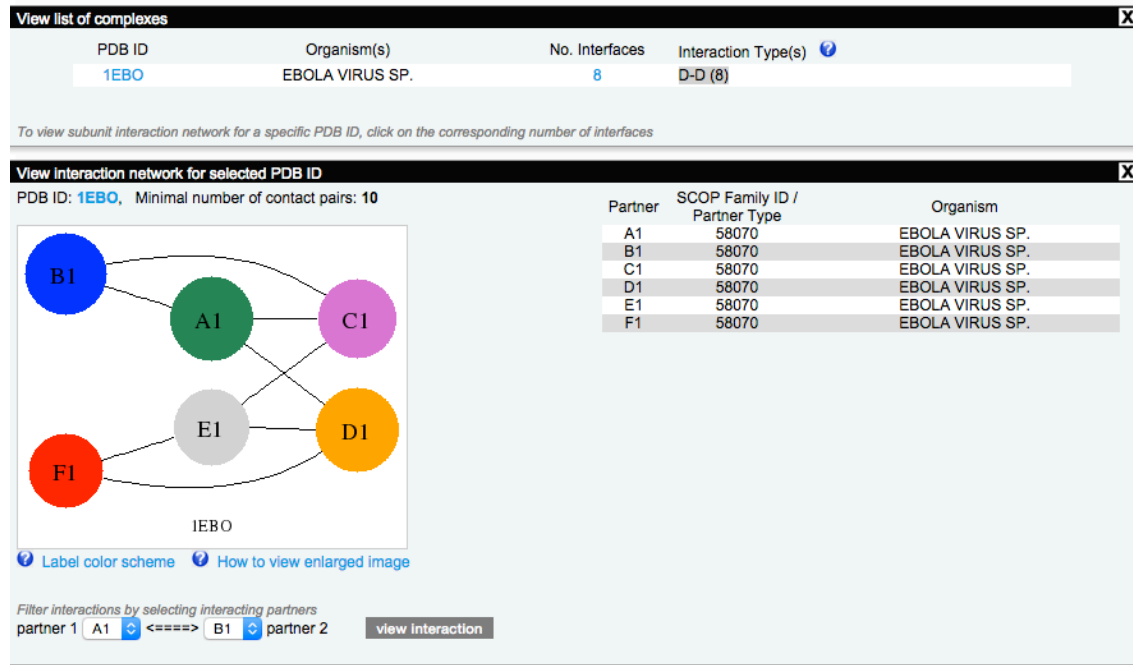


Figure 66: 1EBO Interactions

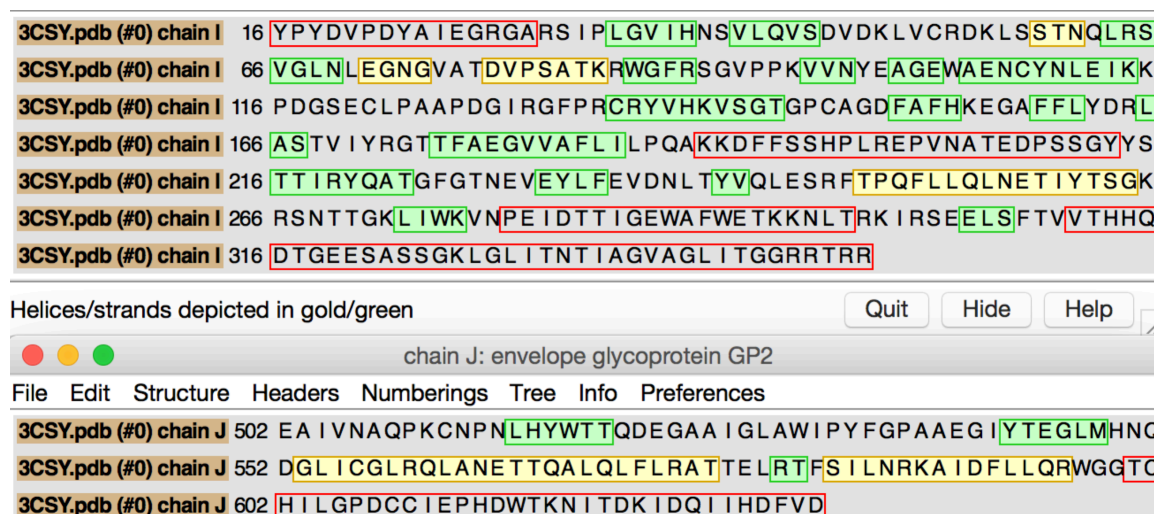


Figure 67: Template 3CSY

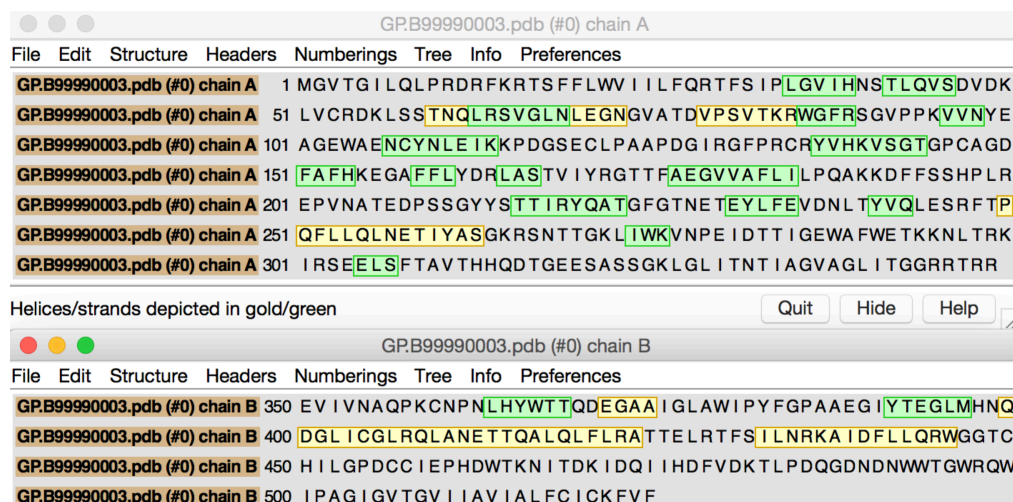


Figure 68: Model

## 6. NP

### **Model:**

Average quality.

### **Annotation:**

- 1). 4QAZ -> 0 interactions
- 2). 4QB0 -> 0 interactions

## 7. L

### **Model:**

Poor quality due to the lack of structural templates with high target-template sequence identity.

### **Annotation:**

PDB advanced search did not provide any significant result. This protein was discarded from the further analysis.

## 8. RNA-Protein Interactions

Search in Dommino using advanced search:

Minimal number of contact pairs: 0

Interaction Type: U-RNA

Organism: virus

### **Annotation:**

- 1). 3KS8 (Reston Ebolavirus) -> 11 interactions (8 U-RNA)
- 2). 4GHL (Lake Victoria Marburgvirus) -> 8 interactions (2 U-RNA)
- 3). 3L26 (Zaire Ebolavirus) -> 3 interactions (2 U-RNA)
- 4). 4GHA (Marburg virus) -> 13 interactions (8 U-RNA)
- 5). 4LG2 (Reston Ebolavirus) -> 10 interactions (4 U-RNA)
- 6). 3L25 (Zaire Ebolavirus) -> 13 interactions (8 U-RNA)

## Appendix F. Functional Residues Mapping

### VP35

Intra-viral interactions:

10,13-15,17-25,27-29,32-33,36-37,40-43,45-46,50,53-64,66-67,70-71,73-74,76-94,96-112,114-124,127-128

### VP40

Intra-viral interactions:

44,46-64,70,72-74,76,85,91-100,107-119,121,123,125,127,130,132,134,136-142,146-148,151-152,154-165,171,178-194,196-197,199-202,210,212-214,217-221,247,249-259,262-263,267-268,271-272,274,279-294,309-321

### GP

Host-viral interactions:

32-35,39-47,50,52,350-363,396-410,412

Intra-viral interactions:

31-38,40-61,63-75,86-105,126-134,136,153-159,162-166,168,179-185,191-192,212-213,283,285,350-447,454-458

Overlap between HV and IV interactions:

32-35,40-47,50,52,350-363,396-410,412

### VP30

Intra-viral interactions:

46-51,95,133-150,152-154,156,165-170,172-174,176-177,179-183,185,187-188,190-192,194-195,199-202,204-205,218,223,246-266

### VP24

Host-viral interactions:

88,113-114,116-118,121-122,124,125,127-131,134-141,184-186,201-205,207,210,216-218

Intra-viral interactions:

10-15,17-22,24-25,28-34,38-41,43-46,54-56,58-59,63-71,73-79,81-83,91-92,95,125,129-133,147-148,150-152,154-155,171-174,208-212,214

Overlap between HV and IV interactions:

125,129-131,210