

**PROTEIN STRUCTURE PREDICTION ON A GENOMIC SCALE: APPLICATION TO  
THE AFRICAN SWINE FEVER VIRUS GENOME.**

**Dickson Bennet Kinyanyi.**

**REG NO: I56/9274/05**

**COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES  
Centre for Biotechnology and Bioinformatics  
University of Nairobi**

**A Thesis submitted to the Board of Postgraduate Studies,  
University of Nairobi, in partial fulfillment for the award  
of  
Master of Science in Bioinformatics**

**© 2014**

**DECLARATION AND APPROVAL**

The work reported herein is original and has not been presented for a degree program in any academic institution or university.

Mr. Dickson Bennet Kinyanyi

Registration Number: I56/9274/05

Signature; .....  ..... Date; ..... 18/08/2014 .....

**Approval**

This thesis has been submitted for examination with our approval as the University Supervisors:

Professor James Ochanda, PhD.

Director - CEBIB Department

University of Nairobi

Signature; .....  ..... Date; ..... 18/08/2014 .....

Dr. Mark Wamalwa, PhD

Lecturer - CEBIB Department


University of Nairobi

Signature; .....  ..... Date; ..... 18/08/2014 .....

Dr. Elisha Opiyo, PhD.

Lecturer – School of Computing & Informatics Department.

University of Nairobi

Signature; .....  ..... Date; ..... 18/08/2014 .....

## Acknowledgements

I give thanks to almighty God Yahweh and Jesus Christ for the strength bestowed unto me while undertaking this research study. I am highly grateful to my research supervisors: Dr. Mark Wamalwa for the excellent opportunity to work in ILRI. Dr Mark Wamalwa and Dr Elisha Opiyo, you have been wonderful supervisors and admirable mentors, for your invaluable guidance in my research work, I am also grateful to ILRI for giving me the opportunity to work for them, I acknowledge the constant support from Prof. James Ochanda and Dr Isabella Oyier through whom I got the opportunity of studying in one of the best Bioinformatics, academic and research programs in the country. Thanks to Dr. Elisha Opiyo for playing a vital role in enhancing my curiosity in computational sciences. I am sincerely grateful to my family for constant support throughout my whole life and most importantly during my research work.

## Dedication

I dedicate this thesis to my parents, William Kinyanyi and Agnes Kinyanyi for facilitating my education up to this level.

# CONTENTS

CONTENTS.....	v
LIST OF FIGURES .....	viii
LIST OF TABLES.....	viii
LIST OF ABBREVIATIONS.....	ix
ABSTRACT.....	x
1.0 INTRODUCTION.....	1
1.1 PROBLEM STATEMENT .....	3
1.2 RESEARCH QUESTION:.....	3
1.3 MAIN OBJECTIVE:.....	3
1.3.1 SPECIFIC OBJECTIVES.....	3
1.4 JUSTIFICATION.....	4
1.5 CURRENT STATUS.....	4
1.6 SCOPE .....	4
1.7 CONTRIBUTIONS.....	4
2.0 LITERATURE REVIEW:.....	5
2.1 AFRICAN SWINE FEVER VIRUS .....	5
2.2 AFRICAN SWINE FEVER DISEASE .....	5
2.3 ASFV INFECTION.....	6
2.4 ASFV HOST EVASION MECHANISMS .....	6
2.5 UNCHARACTERISED OPEN READING FRAMES OF BA71V ISOLATE .....	8
2.6 STRUCTURAL BIOINFORMATICS.....	9
2.6.1 COMPARATIVE MODELING .....	9
2.6.2 FOLD RECOGNITION.....	9
2.6.3 ABINITIO .....	9
2.6.4 I-TASSER .....	10
2.6.5 GENOME-SCALE STRUCTURE PREDICTIONS .....	10
2.6.6 STRUCTURAL SIMILARITY .....	11
2.7 FUNCTIONAL ANNOTATION.....	13
2.7.1 THE GENE ONTOLOGY .....	13
2.7.2 LIGAND BINDING SITES.....	13

2.8	MODEL EVALUATION.....	14
2.9	PROTEIN STRUCTURE CLASSIFICATIONS .....	14
2.10	BIOLOGICAL USEFULNESS OF PREDICTED MODELS .....	15
2.11	PEPTIDE VACCINES AND IMMUNOINFORMATICS .....	16
3.0	METHODOLOGY: .....	17
3.2	DATA SOURCES MATERIALS:.....	17
3.3	DOMAIN BOUNDARY PREDICTION USING THREADOM: .....	18
3.4	I-TASSER .....	19
3.4.1	TEMPLATE IDENTIFICATION.....	19
3.4.2	STRUCTURE ASSEMBLY.....	19
3.4.2.1	ENERGY FORCE FIELD .....	20
3.4.3	ATOMIC MODEL CONSTRUCTION.....	20
3.4.4	MODEL SELECTION AND EVALUATION.....	20
3.5.	FUNCTION PREDICTION OF THE PREDICTED STRUCTURE.....	21
3.6	TIMING.....	21
3.7	STATISTICAL ANALYSES AND VISUALIZATIONS .....	21
4.0	RESULTS .....	23
4.1	THREADOM RESULTS:.....	23
4.2	I TASSER RESULTS .....	24
4.3	STRUCTURAL CLASSIFICATION .....	24
4.4	H171R THREADOM RESULTS .....	28
4.5	H171R TM FOLD AND I-TASSER RESULTS .....	28
4.5.1	H171R BINDING SITE RESULTS .....	31
4.5.2	H171R PROCHECK RESULTS .....	31
4.6	IMMUNODOMINANT PEPTIDES RESULTS: .....	33
5.0	DISCUSSION AND CONCLUSION .....	34
5.1	DISCUSSION .....	34
5.1.1	PROTEIN STRUCTURE PREDICTION AND CLASSIFICATION .....	34
5.1.2	ASFV H171R.....	34
5.1.3.	EPITOPE PREDICTION DISCUSSION .....	36
5.2	CONCLUSION .....	37

5.3 RECOMMENDATIONS: .....	37
BIBLIOGRAPHY .....	38
APPENDICES .....	43
APPENDIX I : PREDICTED PROTEINS.....	43
APPENDIX II : MULTIGENE FAMILIES.....	45
APPENDIX III: H171R.....	46
APPENDIX 1V: SLA DQ AND DR PROTEINS.....	46
APPENDIX V: PERL EXCERPT FOR STATISTICAL CALCULATIONS .....	47
APPENDIX VI: I TASSER STRUCTURAL PREDICTIONS .....	48
APPENDIX VII: COFACTOR BINDING SITE PREDICTION .....	48
APPENDIX VIII: GENE ONTOLOGY PREDICTIONS .....	49
APPENDIX IX : PREDICTED ENZYME FUNCTION:.....	49
APPENDIX X: UNCHARACTERIZED PROTEINS OF H171R .....	50

## LIST OF FIGURES

Figure 1:A schematic representation of the protein structure prediction protocol .....	17
Figure 2:Domain Distribution of single domain and multiple domained of ASFV .....	23
Figure 3:TM-scores of 83 I- Tasser Modeled 3D structures.....	24
Figure 4:Illustration of domain decision by ThreaDom for ORF ASFV 108.....	28
Figure 5:H171R I-Tasser modeled subunit.....	30
Figure 6:Structural superimposition of H171R and the native structure .....	30
Figure 7: Predicted binding site of the model with FEO ligand .....	31
Figure 8: A ramachandran plot showing the distribution of H171R model amino acids.....	32
Figure 9:Iron Oxidation and OH radical formation .....	35

## LIST OF TABLES

Table 1: Summary of 83 uncharacterized proteins in BA71V ASFV isolate .....	8
Table 2: Structural classification of uncharacterised BA71 V ORFs .....	25
Table 3: MHC class II binding peptides ranked on the basis of strong binding to SLA-DQ.....	33
Table 4:MHC class II binding peptides ranked on the basis of strong binding to SLA-DR .....	33



# LIST OF ABBREVIATIONS

3D	Three Dimensional
ASFV	African Swine Fever Virus
ASF	African Swine Fever
CASP	Critical Assessment of Structure Prediction
CATH	Class Architecture Topology and Homology Database
BLAST	Basic Local Alignment Search Tool
DAG	Directed Acyclic Graph
DCS	Domain Conservation Score
DPS	DNA-binding proteins from starved cells
EC	Enzyme Commission
GO	Gene Ontology
I TASSER	<b>I</b> terative <b>T</b> hreading <b>ASSEMBLY</b> <b>R</b> efinement
LOMETS	Local MetaThreading Server
ORF	Open Reading Frame
PDB	Protein DataBase
RMSD	Root Mean Square Deviation
SCOP	Structural Classification of Protein
TM-Score	Template Modelling Score
NFAT	Nuclear Factor of Activated T-cells
NF-KB	Nuclear Factor kappa-light-chain-enhancer of activated B cells
RBCs	Red Blood Cells

# ABSTRACT

Pork is widely consumed worldwide. One potential threat to food security is ASFV (African Swine Fever Virus), the etiological agent of ASF (Africa Swine Fever) in pigs. To date there is no known vaccine for this disease that is characterized by 95-100% mortality rate and frequent outbreaks. The study aimed at predicting protein structures of the 83 unpredicted ORFs in ASFV because of the low sequence identity. I-TASSER an integrated platform that uses low sequence identity is used for protein structure prediction. It combines *abinitio* folding and template-based modeling for genome-wide structure prediction. The study managed to identify an essential protein for the survival of ASFV in stressful environments. The identified protein, H171R, is predicted to belong to the ferritin group and is essential for survival of the virus in the macrophages. Of the 83 uncharacterized proteins in ASFV ORFs, I-TASSER generated models with an average TM-score of 0.7185. TM-Score is a quantitative criterion for structure classification, if it is greater than 0.5 protein structures are classified to belong to the same class. SCOP (Structural Classification of Proteins), a manual based fold classification system for proteins was unable to classify 56% of PDB101 experimentally solved structures we used as templates. TM fold classified 99%. NetMHCIIpan3.0 identified viral peptides that bind to Swine Leukocyte Antigen. This may be useful for vaccine development. In conclusion, the study of ASFV proteins, represent promising progress towards genome-wide structure modeling and fold family assignment when the sequence homology is less than 20%. We identified an immunogenic protein H171R, a protein previously uncharacterised, and found it is a strong binder to Swine Leukocyte Antigen-DQ, hence it should be considered as a vaccine candidate.

# Chapter 1

## 1.0 INTRODUCTION.

Africa swine fever virus from Asfarviridae family infects domesticated pigs causing a fatal hemorrhagic fever. Their natural hosts are warthogs, bush pigs and the soft tick vector. It was described in Kenya in the 1920`s, however, regular outbreaks of ASFV (Dixon et al. 2004; Atuhaire et al. 2013) have emerged and reemerged on a regular basis. Lack of a vaccine contributes to the difficulties in controlling this contagious disease that is characterised by 95-100% mortality rate. In the absence of vaccines, the only available option for ASFV eradication is by slaughter and disposal of all infected and potentially infected pigs causing huge loses. In Africa the significance of pigs in assuring food security is being recognized especially by the rural poor. An extensive free-range pig breeding is of a growing importance for the subsistence of village farming in sub-Saharan African countries. During the last decade, small and sometimes bigger commercial pig farms have been developed in urbanized areas. Pig owners are mostly women farmers who play a major role in the feeding and the management. Availability of several sequenced viral genomes (<http://www.uniprot.org/taxonomy/10497>) have been made, these will help in the understanding the function of ASFV encoded proteins and of the host response to infection. Methods of manipulation of ASFV genome and new virus vectors make it suitable for delivery of foreign genes in pigs. These advances make the development of new diagnostic reagents and a vaccine for control of ASFV a realistic possibility. Currently genome sequencing projects end up producing linear amino acid sequences, but full understanding of the biological role of these proteins will require knowledge of their structure and function. Although experimental structure determination methods are providing high-resolution structures, they are often expensive and a certain subset of proteins is not crystallizable. Computational structure prediction methods used here provide valuable information for the large fraction of sequences whose structures have not been determined experimentally. Computational structure prediction methods have produced structures that are as good as native and therefore are of considerable accuracy. Computational structure prediction methods generally fall into three classes. The first class of protein structure prediction methods is knowledge based and includes comparative modeling. In comparative modeling the protein structure is predicted by aligning the unknown

sequence to an evolutionary related solved template structure sequence. The second class is threading, where similarity at sequence level is low having less than 30 percent, the structure within the amino acid is calculated by a suitable scoring function that is devised to help match the sequence and the existing finite fold numbers. The third class of fold prediction methods uses lattices and simulations to build full length models extracted from already established databases. These methods are also known as de novo or *Abinitio* methods, in predicting the structure from sequence alone searches can be done to establish the protein functions by comparing the generated models to an existent protein database, in this project, threading and *abinitio* method have been applied to the ORFs sequences of the African Swine Fever Virus with the aim of annotating all uncharacterized hypothetical proteins of the ASFV. This has helped to bridge the gap between known sequences and existing structures.

## **1.1 PROBLEM STATEMENT**

A significantly low number of Open Reading Frames in ASFV viral genome have been predicted and methods of ASFV survival are therefore not fully understood. By predicting ASFV protein structures and assigning structures to function, we hope to come up with new viral proteins involved in evading host defenses and therefore have more understanding in developing better vaccines against ASFV.

## **1.2 RESEARCH QUESTION**

The biological question of genome scale protein structure prediction and its application to ASFV when sequence identity is below 30% has not been exploited fully. The study attempts answer this question using I-TASSER method.

## **1.3 MAIN OBJECTIVE**

To use I-TASSER to predict 83 unknown proteins of the ASFV genome.

### **1.3.1 SPECIFIC OBJECTIVES**

1. To structurally classify predicted proteins.
2. To identify a possible vaccine candidate in ASFV.
3. To understand mechanisms of viral survival.

## **1.4 JUSTIFICATION**

ASF is a threat to food security and the disease is endemic in half the African continent with frequent emergence and reemergence. Many ORFS in the ASFV genome have no known functions, with a large portion of Open Reading Frames annotated as hypothetical (83 out of 149), therefore its of much use to predict the proteins to enable us understand their functions and design potentially better effective vaccines, experimental protein structure prediction methods are expensive and take quite long. Therefore computational protein structure prediction offer a reasonable means for this undertaking.

## **1.5 CURRENT STATUS**

A large proportion of Open Reading Frames are annotated as hypothetical with no known function based on homology searches.

There is no known effective vaccine against the African Swine Fever Virus with frequent emergence and reemergence.

## **1.6 SCOPE**

This project involves protein structure prediction on a genomic scale. Understanding the protein structure and function during the prediction is a useful step in effective vaccine development.

## **1.7 CONTRIBUTIONS**

The project has identified an immunogenic protein H171R that is essential for survival in macrophages. The protein was previously uncharacterized. This has added new knowledge to the existent knowledge base, and therefore contributed towards understanding how the virus survives in the host. The project has also classified 82 of the 83 proteins this provide a base for function prediction all the objectives set out to be achieved have been realized.

# Chapter 2

## 2.0 LITERATURE REVIEW:

### 2.1 AFRICAN SWINE FEVER VIRUS

African Swine Fever Virus (ASFV), the etiological agent of ASF, is a large, enveloped, double-stranded DNA virus which replicates predominantly in the cytoplasm of Macrophages. The virus is morphologically similar to the *Iridoviridae* virus family but has a similar genome structure and replication strategy to the *Poxviridae* and therefore has been placed in a separate virus family the *Asfarviridae*, genus *Asfivirus* in which it is the only member. ASFV is currently the only known DNA arbovirus. The *Iridoviridae*, *Poxviridae* and *Asfarviridae* virus families all belong to a group of Nucleocytoplasmic Large DNA Viruses (NCLDV) which also includes the *Phycodnaviridae*, and *Mimivirus* (Iyer et al. 2001).

### 2.2 AFRICAN SWINE FEVER DISEASE

African swine fever (ASF) is a devastating disease of domestic swine. The disease is characterized as a severe hemorrhagic fever with up to 100% mortality in infected herds. The disease was first described by Montgomery in 1920. Contact of domestic pigs with warthogs was identified as the source of infection. In Africa, ASF remains endemic, affecting almost half the continent. Outbreaks of ASF have occurred on a regular basis in many African countries since the mid 90s to date causing devastating losses to the rural poor as well as commercial farms. In Kenya the disease re-emerged in 2013 in Mahimahi killing 80 Pig. In Uganda and Tanzania, frequent outbreaks of African Swine Fever (Atuhaire et al. 2013) have occurred.

The global pig industry is worth around \$150 billion. In absence of a vaccine, the only available option for ASF eradication is stamping out by slaughter and disposal of all infected and potentially infected pigs. However this method is not practical in the poor African farmer context. ASFV is thus a huge threat to food security for the rural poor whose livelihood depends on the pig industry.

### **2.3 ASFV INFECTION.**

Macrophages are key cells involved in activating and co-coordinating the innate and adaptive immune response to infection. The ability of ASFV to infect and replicate in macrophages (Basta et al. 2010) is thought to play a critical role in ASFV disease pathology. ASFV has been reported to infect other cell types, including endothelial cells, fibroblasts and reticular cells (Abrams et al. 2008) . However, infection of these cells is limited to the late stages of infection after the characteristic symptoms of ASFV and therefore does not play a central role in disease pathology. ASFV exhibits a predominantly cytoplasmic replication cycle. Entry of the virus occurs *via* receptor-mediated endocytosis, and is energy and temperature dependent. Following internalization, the viral envelope fuses with that of endosomes at an acidic pH releasing the viral core to the cytoplasm of the host cell (Alonso et al. 2013). The virus initiates gene expression immediately following entry into the cytoplasm, using enzymes and factors packaged into the virus core. Viral gene transcription does not require the host RNA polymerase and is dependent on the viral RNA polymerase and specific virus-encoded transcription factors.

### **2.4 ASFV HOST EVASION MECHANISMS**

Various authors have made an effort in understanding viral proteins involved in evading host defense mechanisms (Dixon et al. 2004; Yáñez et al. 1995). ASFV uses various mechanisms to evade the host immune system: The virus encodes A238L protein that inhibits activity of NF-KB directly and NFAT indirectly via inhibition of calcineurin activity (Abrams et al. 2008; Miskin et al. 2000). Hence this protein can inhibit the expression of a wide range of immunomodulatory proteins in macrophages whose expression depends on these factors.

C-type lectin prevents the presentation of MHC class 1 antigens, ASFV CD2v protein encoded by the virus is expressed on the surface of infected cells causes red blood cells to adhere to the surface of infected cells, camouflaging them from the immune system (Goatley & Dixon 2011). ASFV CD2v is also present on extracellular virus particles, which adhere to red blood cells, facilitating dissemination of ASFV in infected animals. Expression of CD2v in infected macrophage also interferes with the ability of T cells to divide (Goatley & Dixon 2011; Dixon et al. 2004). Virus infection causes apoptosis in neighboring T and B cells (Vallée et al. 2001), thus reducing populations of these important immune cells. The virus also encodes proteins which inhibits apoptosis (Revilla et al. 1997) of the host cell, thereby promoting virus



replication. In this study as one of the objectives was finding any protein the virus uses for survival and evading host defenses, previous studies show that half to two-thirds of the approximately 149 genes encoded by ASFV are not essential for replication in cells but have an important role for virus survival and transmission in its hosts (Dixon et al. 2004). These genes provided an untapped repository, and will be valuable tools for deciphering not only how the virus manipulates the host response to infection to avoid elimination, but also provided a useful understanding important in host anti-viral mechanisms and discovery of novel vaccines.

## 2.5 UNCHARACTERISED OPEN READING FRAMES OF BA71V ISOLATE

A comprehensive list of 83 uncharacterized proteins that genomic scale protein structure prediction will be applied to using I-TASSER is shown in *Table 1*.

**Table 1: Summary of 83 uncharacterized proteins in BA71V ASFV isolate**

	ORF	SEQ No		ORF	SEQ No		ORF	SEQ No		ORF	SEQ No		ORF	SEQ No
1	KP86R*	1	21	A280R	22	41	C122R	60	59	H124R	109	79	DP148R	144
2	KP93L	2	22	A505R	23	42	C275L	61	60	H233R	112	80	DP96R	146
3	KP360L	3	23	A498R	24	43	C62L	65*	61	H240R	113	81	DP363R	147
4	KP362L	4	24	A528R	25	44	B169L	71	62	E184L	118	82	DP42R	148
5	L83L	6	25	A506R	26*	45	B475L	72	63	E423R	120*	83	DP60R	149
6	L356L	7*	26	A542R	27	46	B354L	73	64	E146L	122			
7	L270L*	8	27	A118R	31	47	B125R	77*	65	E111R	128			
8	U104L	9	28	A151R	32	48	B117L	78	66	E66L	129			
9	XP124L	10*	29	A276R	33	49	B407L	79	67	I267L	130			
10	V82L	11	30	F317L	38	50	B263R	81	68	I226R	131			
11	Y118L	12	31	F165R	41	51	B66L	82	69	I73R	133			
12	UP60L	13	32	K205R	43	52	CP123L	85	70	I329L	134			
13	X69R	14	33	K145R	46	53	CP312R	90	71	I177L	136			
14	J268L	15	34	K421R	*47	54	D129L	97	72	I196L	137			
15	J154R	16	35	EP84R	49	55	D79L	98	73	DP238L	138			
16	J104L	17	36	EP152R	51	56	D339L	99	74	DP311R	139			
17	J182L	18	37	M1249L	55	57	S183L	104	75	DP63R	140			
18	J319L	19	38	M448R	56	58	H171R	108*	76	DP542L	141*			
19	A125L	20	39	C84L	58				77	DP141L	142			
20	A489R	21	40	C717R	59*				78	DP146L	*143			

## **2.6 STRUCTURAL BIOINFORMATICS**

Structural bioinformatics and modeling of protein tertiary structure is a potentially attractive route to identify unknown proteins and understand the biological role of these proteins. This will improve our understanding of an organism's biology. Although experimental structure determination methods provide high-resolution structure information, computational structure prediction methods provide valuable information for the large fraction of sequences whose structures will not be determined experimentally. There are 3 main methods of structure prediction. Comparative modeling, fold recognition and *abinitio* methods.

### **2.6.1 COMPARATIVE MODELING**

If the sequence for which the structure to be predicted has a close homolog of greater than 50% sequence identity with an experimental structure solved, (Schwede 2003) it is possible to use the solved experimental structure and the sequence template to accurately build a full length structure. Comparative modeling does this and it's possible to assign function.

### **2.6.2 FOLD RECOGNITION**

Fold recognition involves fitting a protein sequence into a protein structure. subsequently an overall score is then assigned and its structure predicted. The general belief is that different proteins fold into similar 3D shapes because at some level, they share similar interaction patterns among their residues and between the residues and the environments. It has been shown that these interaction patterns could possibly be captured using simple statistics-based energy models (Bowie et al. 1991; Fischer et al. 1996). These simple statistics-based energy functions have been used in many cases, to distinguish the correct structural folds from the incorrect ones and to distinguish the correct placements of the residues in a query protein into the structural positions of a correct structural template.

### **2.6.3 ABINITIO**

In many cases neither comparative modelling nor threading can provide a useful model for a sequence under study if local conformations are nonexistent. If protein templates are not available, we have to build the 3D models from scratch. This procedure is called *abinitio* modelling (Klepeis et al. 2005).

#### **2.6.4 I-TASSER**

I-TASSER ( Roy et al. 2010a; Zhang 2008; Wu et al. 2007; Zhou et al. 2007) is an integrated package for protein structure prediction employing both threading and *Abinitio* methods, it runs as a standalone algorithm or may be run from the main server Zhang lab. I-TASSER algorithm is based on a knowledge based potential that reduces the search space by using local conformations of short amino acid fragments from the PDB ( Dong Xu and Yang Zhang 2013; Zhang & Skolnick 2005a). The initial search is done in centroid mode and an optional subsequent model refinement is done in a full atom mode. The algorithm was developed under the assumption that a short sequence of amino acids have a finite number of conformations (Bystroff & Baker 1998) and that these conformations are represented in the PDB (Berman 2008). For a given sequence, I-TASSER first identifies template proteins from the Protein Data Bank PDB by multiple threading techniques LOMETS (Local Metathreading Server) (Wu & Zhang 2007). The continuous fragments excised from the template alignments are used to assemble full-length models by iterative Monte Carlo simulations. The best models are then selected from the Monte Carlo trajectories by decoy clustering. The final atomic models are rebuilt from the structure clusters by atomic-level structural refinements. The I-TASSER algorithm ("Zhang-Server") participated in the Server Section of 7<sup>th</sup> (2006), 8<sup>th</sup> (2008), 9<sup>th</sup> (2010) 10<sup>th</sup> (2012) and was ranked as the No1 server in CASP7 (Critical Assessment of Structure Prediction) and CASP8 (Zhou et al. 2007). In CASP9, I-TASSER server and QUARK (another server from the same lab) were ranked as number one and number two servers, respectively in CASP (2010) while in 2012 it was ranked first. Based on these performances, we will implement the algorithm in our genomic scale Protein Structure Prediction procedure.

#### **2.6.5 GENOME-SCALE STRUCTURE PREDICTIONS**

Genome scale structure predictions have been performed by many authors (Zhang & Skolnick 2004a; Malmström et al. 2007; Wu et al. 2007; Fischer & Eisenberg 1997; Sánchez & Sali 1998; Kihara et al. 2002; Kim et al. 2003) on different organisms. The methods used have been based on threading and refined forms of threading (Roy et al. 2010a; Zhang 2008; Wu et al. 2007; Zhou et al. 2007) these methods have been performing considerably well in CASP on medium-size proteins particularly (Zhang & Skolnick 2004e; Zhang 2008; Zhou et al. 2007). This could be attributed to the completeness of PDB for single domain proteins majority of which are

medium sized proteins ( Skolnick et al. 2012; Skolnick et al. 2009 ; Zhang, Hubner, et al. 2006; Zhang & Skolnick 2005a; Kihara & Skolnick 2003;), it was therefore necessary to check the number of domains in an ORF (Xue et al. 2013) then apply genome scale prediction to it, if the prediction target is multidomained and hard to recognize then splitting it into individual domains can help in protein structure prediction. Studies have been applied to various organisms ( Idrees & Ashfaq 2013; Idrees et al. 2013; Xu & Zhang 2013; Jethra et al. 2012; Kemege et al. 2011; Franceschini et al. 2006; Kihara et al. 2002; Fischer & Eisenberg 1997) and have demonstrated genome scale predictions as being possible, recently (Xu & Zhang 2013) demonstrated with a 65% success rate on hard proteins having less than 12% identity, using Quark an *abinitio* protein structure prediction software from Zhang lab (Zhang 2014) that 321 of 465 hard targets proteins were predicted as having a TM-Score of greater than 0.5. These have a high reliability for fold family classification based on TM align. TASSER has also been benchmarked on automated prediction of weakly homologous protein on a genomic scale and the accuracy was approximately 67%, Our study therefore finds it appropriate to apply the protein structure prediction on a genomic scale to the ASFV genome consisting of 83 Open Reading Frames, we expect a similar success rate of fold family classification with Template Modeling Score (TM-Score ) values of ~65% of the Open Reading Frames being greater than 0.5 for accurate protein structure prediction and classification.

### 2.6.6 STRUCTURAL SIMILARITY

Structures are traditionally compared using a metric called Root Mean Square Deviation (RMSD). That is, the root of the sum of the squared distance between alpha carbons of equivalent amino acids.

$$RMSD = \sqrt{\frac{\sum (r_{ai} - r_{bi})^2}{n}}$$

*Equation 1*

$r_{ai}$  and  $r_{bi}$  are the coordinates of atom  $i$  of structure  $a$  and structure  $b$   $n$  is the number of residues. See equation 1. When structures differ by a mean deviation less than 2Å, they are considered structurally equivalent. RMSD has two fundamental flaws of being dependent on length and sensitive to outliers, A unified statistical framework developed by (Levitt & Gerstein 1998) made improvements on sensitivity to outliers during structural comparisons of protein structures, later

improvements led to the development of the TM Score. TM Score is length dependent and independent on outliers. TM-align is a structural alignment algorithm that uses TM-score to identify the best structural alignment between protein pairs. It combines the TM-score, a rotation matrix and Dynamic Programming (DP). TM align finds close structural analogs and ranks them based on a TM-score see **Equation 2**. The TM-score defined assess the topological similarity of two protein structures where  $L$  is the length of the target protein, and  $L_{ali}$  is the number of the equivalent residues in two proteins.  $d_i$  is the distance of the  $i$ -th pair of the equivalent residues between the two structures, which depends on the superposition matrix; the ‘max’ means the procedure to identify the optimal superposition matrix that maximizes the sum in **Equation 2**. The scale  $d_0 = \sqrt[3]{(L-15)-1.8}$  is defined to normalize the TM-score in a way that the magnitude of the average TM-score for random protein pairs is independent on the size of the proteins. TM-score ranges between (0, 1] with a higher value indicating a stronger similarity.

$$\text{TM-score} = \frac{1}{L} \left[ \sum_{i=1}^{L_{ali}} \frac{1}{1 + d_i^2 / d_0^2} \right]_{\max} \quad \text{Equation 2}$$

The TM-score of protein pairs sharing the same fold in *abinitio* and template-based protein structure prediction helps in judging whether a predicted model has the same fold or topology as the experimental solved structure (Zhang & Skolnick 2004b; Xu & Zhang 2010; Zhang 2010). For TM-score <0.5, proteins are mostly not in the same fold while for TM-score >0.5, proteins are generally in the same fold, there exists a significant correlation between the correctness of the predicted structure and the structural similarity of the model to the other proteins in the PDB. The correlation could be used to assist in model selection in blind protein structure predictions.

## **2.7 FUNCTIONAL ANNOTATION**

Involves the use of an online server Cofactor.( Roy et al. 2012).Cofactor employs Enzyme Commissioning, Gene Ontology and ligand binding site prediction.

### **2.7.1 THE GENE ONTOLOGY**

The Gene Ontology (GO) is a hierarchical ontology developed by Ashburner and colleagues (Gene & Consortium 2000).GO describes a proteins function from three perspectives or branches, its localization (cellular component), its biochemical function (molecular function) and the proteins context in the cell (biological process). Each branch is organized in a tree like structure with a single root with one or more children. A function is a node in this tree like structure and a relation between functions is called an edge. The lower down in the tree like structure, the more specific the term, and terms with no children are the most specific functions called leafs. Each branch is a directed acyclic graph (DAG), in our study GO score values range between zero to one where a higher value indicates a better confidence in predicting the function using an existent template.

### **2.7.2 LIGAND BINDING SITES**

Protein–ligand binding sites and ligand-interacting residues in the query protein are identified based on both global and local structural similarities to a comprehensive binding site template library BioLiP (Yang et al. 2013) BioLiP is a semi-manually curated database for biologically relevant ligand–protein interactions, which contains greater than 204 223 entries, including information on protein–protein, protein–nucleic acid, protein–lipid and protein–small molecule interactions. Cofactor uses this database for functional site assignment. The binding position of the template ligand in the query structure is predicted based on the superposition matrix acquired from the local alignment of query and template binding site residues. Binding site score greater than 1.1, indicates a better confidence in predicting the function of the modeled structure using the template.

## 2.8 MODEL EVALUATION

Procheck (Laskowski et al. 1993) evaluates the quality of the model, this is summarized via a Ramachandran plot showing the phi-psi torsion angles for all residues in the structure (except those at the chain termini). Glycine residues are separately identified by triangles as these are not restricted to the regions of the plot appropriate to the other side chain types. The coloring/shading on the plot represents the different regions; the darkest red areas correspond to the "core" regions, representing the most favorable combinations of phi-psi values. Ideally, it's suitable to have over 90% of the residues in these "core" regions. The percentage of residues in the "core" regions is one of the better guides to stereochemical quality. We used Procheck to evaluate our predicted models and establish its accuracy.

## 2.9 PROTEIN STRUCTURE CLASSIFICATIONS

The number of potential protein structures is enormous but many of these potential protein structures will resemble each other. By grouping proteins with structures that resembles each other in a tree structure, where groups closer to the leafs are more similar, and closer to the root are less similar, it is possible to reduce the complexity. This is difficult because it is complicated to define a single metric that describes structural similarity. Murzin and colleagues (Murzin et al. 1995; Fox et al. 2014) have developed a classification system named SCOP, (SCOP; Structural classification of Proteins). The SCOP classification classifies protein structures according to a hierarchical 4 level tree. The levels are Class, Fold, Superfamily and Family. The current SCOP database, version 1.75C or 2.03e, has 59514 protein structures, 167547 domains divided into 1194 folds, 1961 superfamilies and 4493 protein families. There are a number of classes in SCOP, 4 of which are more prominent than the others. All alpha proteins consist of mostly alpha helices and beta proteins contain only beta sheets. The alpha+beta proteins contain both alpha helices and beta sheets but the different elements are spatially grouped with secondary structure elements of similar kind and beta sheets being mostly antiparallel. The last group, the alpha/beta group contains alpha helices and beta sheets mixed together and beta strands are parallel. Other minor groups exist, for example multidomain proteins, membrane and cell surface proteins, the study attempted to classify proteins based on SCOP classification. We used the SCOP database for fold assignment during the study to help us assign structure to function.



## 2.10 BIOLOGICAL USEFULNESS OF PREDICTED MODELS

The biological usefulness of the predicted protein models relies on the accuracy of the structure prediction (Zhang 2010). For example, high-resolution models with RMSD in 1–2 Armstrong's, typically generated by Comparative modeling using close homologous templates, usually meet the highest structural requirements and are suitable for computational ligand-binding studies and virtual compound screening (Ekins et al. 2007). Medium-resolution models, roughly in the RMSD range of 2~5 Armstrong's and typically generated by threading and CM from distantly homologous templates, can be used for identifying the spatial locations of functionally important residues, such as active sites and the sites of disease-associated mutations (Arakaki et al. 2004). However, many of the functionally important sites are located on loop regions show large structural variability although the scaffold of the protein structures is conserved. Thus, accurately modeling of the loop regions is still an important yet unsolved problem in template-based modeling, models with the lowest resolution, from an otherwise meaningful prediction, models with an approximately correct topology, predicted using either *abinitio* approaches or based on weak hits from threading, have a number of uses including protein domain boundary identification, topology recognition and family/superfamily assignment (Zhang, Devries, et al. 2006).the biological function of protein molecules is determined by their 3D shape (which dictates how the protein interacts with ligands or other protein molecules), one of the most common motivations for predicting the protein structure is to use the structural information to gain insight into the protein's biological function. A convenient approach to the structure-based functional assignment involves global structural comparison of protein pairs for fold recognition and family assignment which in many cases can be directly used to infer function. However, it is increasingly recognized that the relationship between structure and function is not always straightforward, as many protein folds/families are known to be functionally promiscuous, and different folds can perform the same function (Bork et al. 1993). When the global structures are not similar, functional similarity may arise due to the conserved local structural motifs which perform the same biochemical function, Cofactor helped us in binding site prediction for functionally promiscuous structures that cannot be accurately determined but local binding sites are of consensus. One main challenge encountered was predicting small domain proteins that are below 80 residues, ThreaDom assumes they are single domain proteins and ITASSER predictions are not very reliable because they lack templates in the PDB.

## 2.11 PEPTIDE VACCINES AND IMMUNOINFORMATICS

The general idea behind peptide vaccines is based on the chemical approach to synthesize the identified T-cell epitopes that are immunodominant and can induce specific immune responses. Because ASFV infects macrophages which interact with T cells the study will confine itself to T-Cell epitopes. The T-cell epitopes are typically peptide fragments, an attempt to predict T-cell epitopes that are immunogenic is therefore necessary. Peptides have become desirable vaccine candidates owing to their comparatively easy production and construction. The methodology of analyzing the pathogen genome to identify potential antigenic proteins is known as 'reverse vaccinology'. Normally, the investigation of the binding affinity of antigenic peptides to the MHC Class II molecules is the main goal when predicting epitopes. Using NetMHCIIpan3.0 (Nielsen et al. 2010) the study intended to predict epitope binders that potentially elicited a strong immune response,

NetMHCIIpan3.0 (Karosiene et al. 2013) server (<http://www.cbs.dtu.dk/services/NetMHCIIpan/>) predicted binding of peptides to MHC class II molecules, in swine macrophages SLA-DR and SLA-DQ membrane markers present peptides (Piriou-Guzylack & Salmon 2008), for T-Cell recognition, NetMHCIIpan 3.0 server produced predictions for peptides of 9 - 19 amino acids in length. The server also provides a possibility for the user to upload the Specific SLA protein sequence of interest that present the cleaved peptides to the T-cells, The prediction values are given in nM IC50 values and as % Rank to a set of 200000 random natural peptides. Strong and weak binding peptides were indicated in the output.

In summary genomic scale sequencing projects end up producing linear amino acid sequences, but full understanding of the biological role of these proteins require knowledge of their structure and conserved functional sites. There is a high demand of the for protein structures to help us understand how an organism functions. Computer-based protein structure prediction, provides a means to alleviate the problem that is growing at an unprecedentedly critical position is, to bridge the gap between sequence and functional structure annotation, we used I TASSER the best ranked algorithm in protein structure prediction in an attempt to annotate proteins of low sequence identity. NetMHCIIpan3.0 was used to check if a protein is immunogenic, this information is useful in vaccine and diagnostic kit development.

# Chapter 3

## 3.0 METHODOLOGY:

Five main processes that were involved in the prediction are; domain prediction, structure modeling, model evaluation, function prediction and immunodominance determination a similar methods have been successfully used (Roy et al. 2010b; Wu et al. 2007) in protein structure prediction . A schematic representation for the modeling of tertiary structures for the proteins under study and function assignment of the entire ASFV genome is illustrated below see *figure 1*.

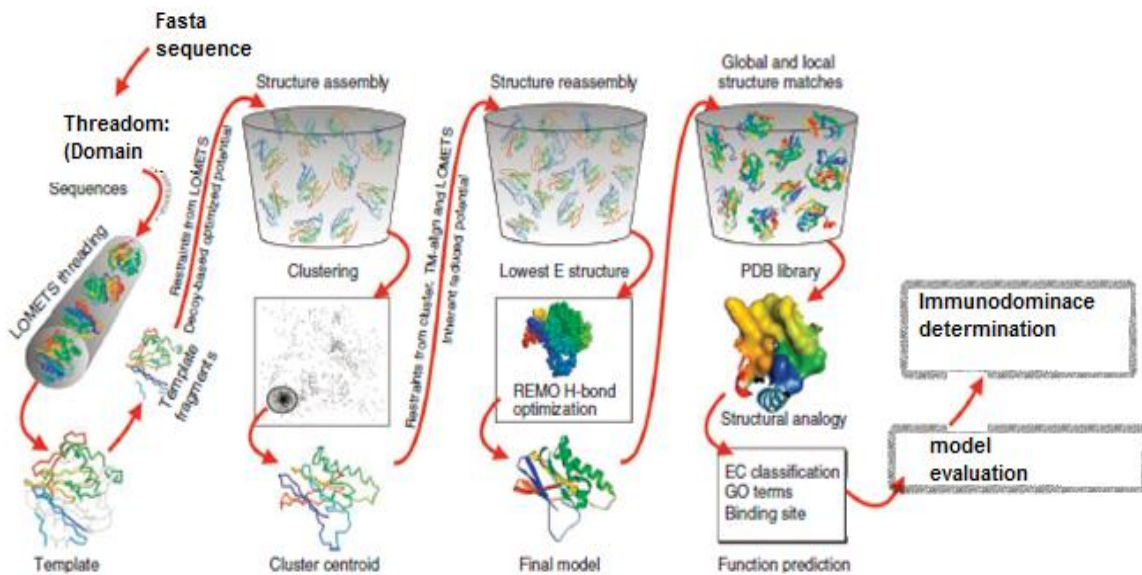


Figure 1: A schematic representation of the protein structure prediction and immunodominance determination. Diagram courtesy of Roy et al. 2010.

### 3.2 DATA SOURCES AND MATERIALS:

ASFV isolate BA71V avirulent strain sequenced ORF was used, the ORFs were downloaded in Fasta format available online at (<http://www.uniprot.org/taxonomy/10497>).The SLA-DR and SLA-DQ Porcine peptides that bind immunogenic peptides were also downloaded from Uniprot <http://www.uniprot.org/uniprot/Q31072> and <http://www.uniprot.org/uniprot/?query=sla+dq&sort=score>. The study used the following software; ThreaDom for domain prediction, I-TASSER for structural modeling, TM Align for structural comparison, TM fold for model classification, Cofactor for function prediction and Procheck for stereochemistry checking these softwares are

available as online servers for academic users, we opted for the online version because of time constraints and powerful processing technologies.

### **3.3 DOMAIN BOUNDARY PREDICTION USING THREADOM:**

The sequenced ORFs of the ASFV genome BA71V were fed into ThreaDom, a domain boundary prediction algorithm, <http://zhanglab.ccmb.med.umich.edu/ThreaDom/>

Domain predictions in ThreaDom are based on two assumptions: first homologous proteins have similar domain structures; second residues in the core regions of domain structures are evolutionally more conserved than those in the boundary (or linker) regions between domains. Following these assumptions, the ThreaDom procedure contains three steps, Target sequences were threaded through the PDB by eight, LOMETS programs, and a multiple sequence alignment is constructed based on the target sequence. A Domain Conservation Score (DCS) is calculated for each residue position based on the LOMETS multiple sequence alignments, which counts for the balance of conservation and gap penalty scores. Domain boundaries were assigned based on the DCS profile using a target-specific scoring cut-off. The target open reading frame was then judged as being a multiple domain protein or a single domain.

We proceeded with I TASSER modeling for all ORFs but if the TM-SCORE was less than 0.5 for multidomained proteins then it was desirable to split the long multi-domain proteins and model each domain separately, if the single domain had a TM Score less than 0.5 then we assumed it belonged to a new fold, Modeling domains individually has been shown to speed up the prediction process, it also increases the quality of query-template alignment resulting in more reliable structure, the PDB library is complete for single domain proteins (Skolnick, Zhou, & Brylinski, 2012; Zhang, Hubner, Arakaki, Shakhnovich, & Skolnick, 2006; Zhang & Skolnick, 2005; Zhang & Skolnick, 2005; Kihara & Skolnick, 2003), at low to moderate resolution, therefore there is a higher expectation in modeling single domain proteins in that one already solved structure exists in the PDB that has a RMSD from native  $< 4$  Angstroms for 90% of its residues (Zhang & Skolnick 2004a). Therefore the likelihood of success in protein prediction is maximized. Domains having less than 80 residues are considered to be single domains. (Xue et al. 2013) if a segment of the target sequence with  $>80$  residues has no aligned residues in the top two threading templates.

### 3.4 I-TASSER PREDICTION PROCESS

The I-TASSER prediction process consisted of four major steps: Template identification, Structure assembly/reassembly, Atomic model construction, and Final model selection. I-TASSER has been benchmarked (Zhang 2008) with good performance on medium sized proteins.

#### 3.4.1 TEMPLATE IDENTIFICATION

The Open Reading Frames of the ASFV were threaded through a non-redundant PDB structure library (<http://zhanglab.ccmb.med.umich.edu/library/PDB.tar.bz2>) within the I-TASSER server for identifying appropriate global-structure templates. Threading was done by LOMETS (Wu & Zhang 2007) a Local Metathreading Server consisting of 10 individual servers, LOMETS was used to detect homologous templates from the PDB library, for each threading program the significance of target template alignment was measured by a Z-score, consensus amongst the individual servers via a 3D jury (Ginalski et al. 2003) system, 3D-Jury is an algorithm that aggregates and compares models from various protein structure prediction servers. It takes in groups of predictions made by a collection of threading software and assigns each pair a 3D-Jury score, based on structural similarity. The score is generated by counting the number of C $\alpha$  atoms in the two predictions within 3.5 Å of each other after being super positioned; the approach was used to identify the 10 best protein scoring templates from LOMETS.

#### 3.4.2 STRUCTURE ASSEMBLY

Continuous fragments excised from the LOMETS threading templates were used to assemble full-length models (Roy et al. 2010b) the unaligned loop regions were built by *abinitio* modeling in a lattice system (Zhang et al. 2003). Structure assembly process consisted of two sets of simulations. The first set uses the threading templates as initial structures. In the second set, the simulations start from the cluster centroids generated by SPICKER a simple and efficient strategy to identify near native folds by clustering protein structures generated during computer simulations (Zhang & Skolnick 2004d), Spatial restraints collected from the PDB structures hit by TM-align (Zhang & Skolnick 2005b) using the cluster centroids as query structures are also incorporated in the I-TASSER simulations. The purpose of the second stage is to refine the local geometry as well as the global topology of the SPICKER centroids.

### **3.4.2.1 ENERGY FORCE FIELD**

The structure assembly simulations (for both the threading-aligned and the *Abinitio* modeled regions) were guided by a unified knowledge-based force field, which includes four components: (1) general knowledge-based statistics terms from the PDB (C-alpha/side-chain Correlations) (2) H-bonds and hydrophobicity. (3) spatial restraints from threading templates; (4) sequence-based contact predictions from SVMSEQ(Wu & Zhang 2008).SVMSEQ is a Support-Vector-Machine (SVM) based residue-residue contact predictor that only uses sequence information. It was trained using local window features (position-specific scoring matrices, secondary structure and solvent accessibility predictions) and in-between segment features (residue separations, secondary structure of the contacting residues, and state distributions of the contacting residues).

### **3.4.3 ATOMIC MODEL CONSTRUCTION**

The SPICKER cluster centroids from I-TASSER are reduced models, with each residue represented by its C $\alpha$  and side-chain center. The full-atomic models were built by REMO (Li & Zhang 2009), a protocol developed for constructing full-atomic models from C-alpha traces by optimizing the H-bond networks. The basic backbone fragments (C $\alpha$ , C, N, O) were matched from a secondary structure specific backbone isomer library which consist of a total of 68,206 non-redundant isomers from high-resolution PDB structures. The driving force in the REMO refinement protocol includes H-bonding, clash/break-amendment, I-TASSER restraints, and the CHARMM22 potential. Based on a test set of 230 non-homologous proteins, REMO has the ability of removing steric clashes while retaining a topology score (TM-score).

### **3.4.4 MODEL SELECTION AND EVALUATION**

Structural analogs of the top-scoring I-TASSER model in the PDB library as identified by the structural alignment program TM-align(Zhang & Skolnick 2005b) were ranked based on the TM-score(Zhang & Skolnick 2004c) this was done by comparison of the I-TASSER model and the experimentally solved PDB proteins. Structural analogs with a TM-score >0.5 were used for classifying the modeled structure (Xu & Zhang 2010)TM fold server/software calculated the posterior probability of the modeled structure and an experimentally solved PDB protein in belonging to the same class. High scoring relevant models were evaluated using Procheck to ascertain if greater than 90 % of residues lie in the favourable region, if it met this criteria then

the model was classified as being good enough.

### **3.5. FUNCTION PREDICTION OF THE PREDICTED STRUCTURE**

. The function prediction result was divided into four subsections: TM Align, Enzyme Classification, (EC) numbers, Gene Ontology (GO) terms, and Ligand binding. Structural homology was used to assign if a model and an experimentally solved PDB native perform the same function, though there is also overlapping functions amongst various microorganisms having similar structures hence binding sites may assist in assignment of functions since they are the basic level of conservation. Enzyme commissioning was also used to predict enzyme-catalyzed reactions .TM-score (template modeling score) was defined to assess the topological similarity of protein structure pairs with a value in the range of (0, 1], a higher score indicated better structural match both in Binding site assignment and fold function assignment ,binding site score was incorporated to assess function prediction, a threshold greater than 1.1 is assigned as good binding sites and can be used to infer functions(Roy et al. 2010b).for the gene ontology Go score of 0.5 are desirable for functional inferences

### **3.6 TIMING**

On average the procedure of structure and function prediction by the I-TASSER server is 72 hours for a typical medium-size protein (~100–300 residues), although larger proteins required a longer Monte Carlo simulation and hence longer waiting time the actual processing time also depended on the number of jobs in the queue. I TASSER online server is run on a cluster of 2000 HP DL1000h (Nehalem) processors.

### **3.7 STATISTICAL ANALYSES AND VISUALIZATIONS**

Posterior probabilities and p-Value calculations were calculated by TM-Fold (CSU & Zhang 2010) the software is available online and downloadable as a free academic software at <http://zhanglab.ccmb.med.umich.edu/TM-fold>, 3D modeled structures for the generated models were visualized using chimera version 1.8.

### 3.8 IMMUNODOMINANCE

Investigation of the binding affinity of antigenic peptides to the MHC Class II was done using NetMHCIIpan3.0 (Nielsen et al. 2010; Lundegaard et al. 2010) this was done to predict epitope binders which potentially elicited immune responses for the ASFV Viral proteins. NetMHCIIpan 3.0 server (<http://www.cbs.dtu.dk/services/NetMHCIIpan/>) outputted 15mer peptides that bind to MHC class II porcine peptides, the porcine isotypes SLA-DR (Gustafsson et al. 1990) and SLA-DQ (Kim et al. 2012) were used because these two membrane markers are the ones that present cleaved peptides (Piriou-Guzylack & Salmon 2008) on infected swine macrophages, NetMHCIIpan 3.0 server can produce predictions for peptides of 9 - 19 amino acids in length. The MHC pseudo sequence generated by the servers for optimal output based on the Servers trained neural network algorithms, are as follows.

QEFFIASGAAVDAILHLFLEQYDLQRETYHILFL for SLA-DRalpha\beta and  
YLFHETSGARTLHIVYFGHTYFDFQTETVHIETT for SLA-DQalpha\beta.

*(see appendix page 46 for SLA sequences).*

NetMHCpanII3.0 server provided a possibility for the user to upload the specific MHC protein sequence of interest, for the pig (*sus-scrofa*) we have the SLA-DQ and SLA-DR experimental solved proteins available, we use the SLA-DQ and SLA-DR to maintain the accuracy and avoid cross species based predictions. The prediction values were given in nM and to a set of 200.000 random natural peptides. Strong and weak binding peptides were ranked in the output and the top 10 peptide sequences of ASFV used.

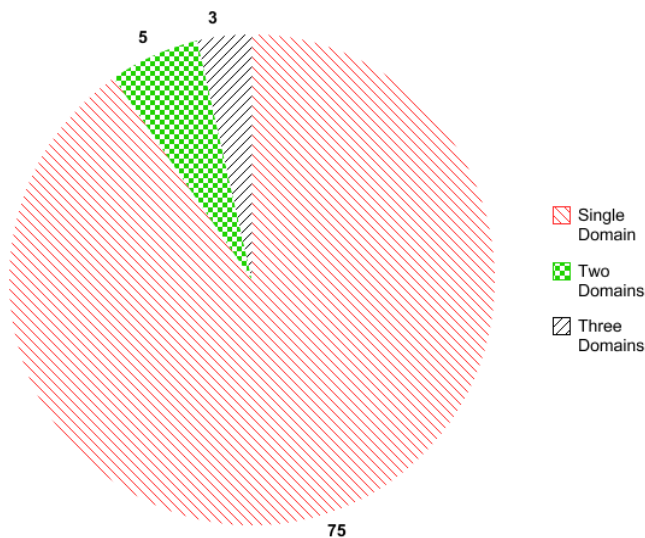


# Chapter 4

## 4.0 RESULTS

### 4.1 THREADOM RESULTS:

ThreaDom (Xue et al. 2013) was used for domain boundary identification the reason being the force field of I TASSER performs much better on single domain proteins and the domain space for single domain proteins is nearly complete in the PDB (Zhang, Hubner, et al. 2006; Zhang & Skolnick 2005a; Skolnick et al. 2012). ThreaDom defines a total of 94 domains from the 83 Open Reading Frames.

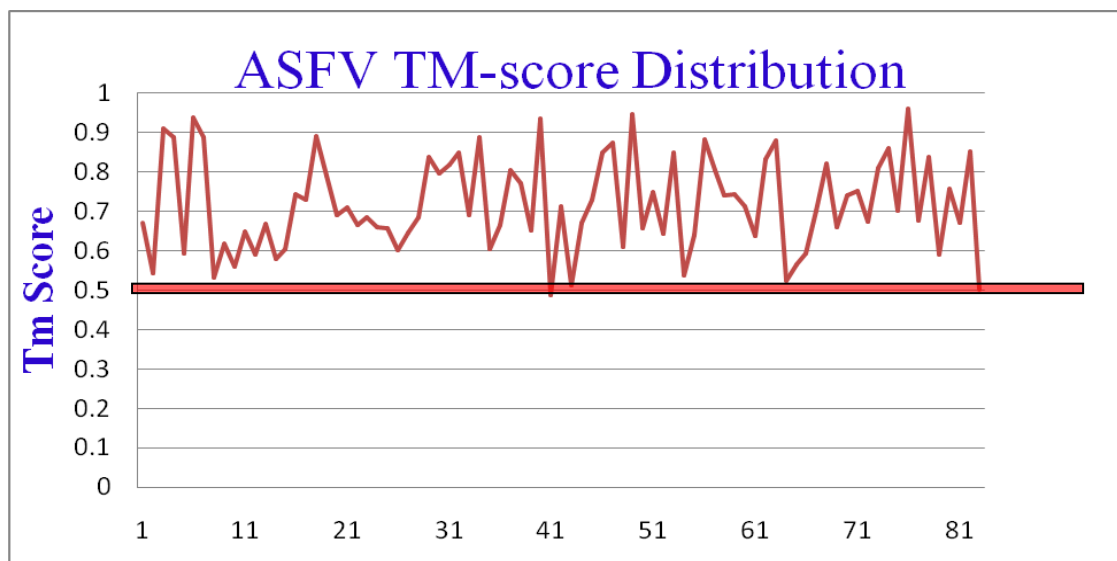


**Figure 2: Showing the Domain distribution of single domain and multiple domained ORFs of ASFV a total of 83 ORFs have been classified to fall into the 3 groups. red are 1 domained , green 2 domained and blue 3 domained ,a total of 94 domains are classified**

the 83 uncharacterized Open Reading Frames have their domain boundaries determined,75 of them are single domained, Meaning they have one functional domain,5 of them have 2 domains , while 3 of them have 3 domains making a total of 94 domains (75+10+9). For the predicted 75 single domain open reading frames we are certain of structurally predicting each one of them after I-TASSER modelling this is because of the completeness of the PDB for single domain proteins (Skolnick et al. 2012; Zhang, Hubner, et al. 2006; Zhang & Skolnick 2005a).

## 4.2 I-TASSER RESULTS

To analyze the ASFV uncharacterized ORFs, the sequenced BA71V genome and the sequences were pipelined to I-TASSER for protein structure prediction. I-TASSER pulled the modeled templates closer to PDB structures, as observed in previous benchmark tests (Wu et al. 2007) and CASP (Zhang 2009; Kopp et al. 2007; Zhou et al. 2007), the fragment assembly I-TASSER procedure consistently drives the local excised fragment structures closer to the experimentally solved PDB natives. As anticipated nearly all single domain (74 out of 75) modeled structures have a structurally related fold in the PDB (Zhang, Hubner, et al. 2006; Kihara & Skolnick 2003), from the I-TASSER results 82 of the 83 TM-Scores lie above 0.5 *see figure3*, meaning 82 models are viable for fold classification because they have a TM score greater than 0.5 (Xu & Zhang 2010). The average TM score is 0.7185. There was no need for splitting multiple domain proteins into single domains because the I-TASSER predicted models had TM Scores greater than 0.5. A summary of the TM-Scores is as shown. See *figure3*.



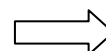
**Figure 3: TM-scores of 83 I-TASSER Modelled 3D structures. TM-Scores lie above the 0.5 minimal threshold for fold family assignment in SCOP. Y-axis represents TM-Score and the X axis 83 proteins that will be classified by TM-Fold.**

## 4.3 STRUCTURAL CLASSIFICATION

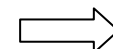
Structural classification was carried out by TM-Fold software that incorporates TM-Align and a Perl script that calculates the posterior probabilities and p-Values, Structure similarity was measured by Template Modeling score (TM-score). TM score was used because it provided a quantitative relation between scores and conventional fold classifications (Xu & Zhang 2010; Zhang & Skolnick 2005b; Zhang & Skolnick 2004b).

**Table 2: A summary of the structural classification of ASFVs' 83 I-TASSER predicted protein structures done using TM-Fold a fold classification software, I-TASSER modelled the structures, TM Align calculated the structural similarity , TM-Fold calculated the fold assignment posterior probabilities and p-Values for all the 83 Predicted proteins.**

ORF	Sequence number of ORF	NO OF DOMAINS	HIGHEST TM SCORE	POSTERIOR PROBABILITY TO SCOP	P-Value (to 4dp)	SCOP-Assignment
<b>KP86R*</b>	1	1	0.6700	0.8636	<0.0000	a.45.1.1
<b>KP93L</b>	2	1	0.5420	0.4332	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>KP360L</b>	3	1	0.9100	0.9969	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>KP362L</b>	4	1	0.8890	0.9946	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>L83L</b>	6	1	0.5920	0.7120	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>L356L</b>	7	1	0.9390	0.9974	<0.0000	<a href="#">c.37.1.8</a>
<b>L270L*</b>	8	2	0.8880	0.9962	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>U104L</b>	9	1	0.5330	0.5199	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>XP124L</b>	10	1	0.6170	0.8107	<0.0000	<a href="#">a.29.2.0</a>
<b>V82L</b>	11	1	0.5590	0.8778	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>Y118L</b>	12	1	0.6490	0.8214	<0.0000	<a href="#">a.45.1.0</a>
<b>UP60L</b>	13	1	0.5900	0.5371	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>X69R</b>	14	1	0.6670	0.7755	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>J268L</b>	15	1	0.5800	0.5012	<0.0000	<a href="#">a.118.1.1</a>
<b>J154R</b>	16	1	0.6040	0.6246	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>J104L</b>	17	1	0.7430	0.9638	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>J182L</b>	18	1	0.7300	0.9589	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>J319L</b>	19	1	0.8900	0.9963	<0.0000	<a href="#">d.211.1.1</a>
<b>A125L</b>	20	1	0.7900	0.9847	<0.0000	<a href="#">d.211.1.1</a>
<b>A489R</b>	21	1	0.6910	0.9931	<0.0000	<a href="#">d.211.1.1</a>
<b>A280R</b>	22	1	0.7110	0.9405	<0.0000	<a href="#">d.211.1.1</a>
<b>A505R</b>	23	1	0.6660	0.9946	<0.0000	<a href="#">d.211.1.1</a>
<b>A498R</b>	24	1	0.6850	0.9954	<0.0000	<a href="#">d.211.1.1</a>
<b>A528R</b>	25	1	0.6600	0.9947	<0.0000	<a href="#">d.211.1.1</a>
<b>A506R</b>	26	1	0.6570	0.9921	<0.0000	<a href="#">d.211.1.1</a>
<b>A542R</b>	27	1	0.6010	0.9873	<0.0000	<a href="#">d.211.1.1</a>
<b>A118R</b>	31	1	0.6450	0.6650	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>A151R</b>	32	1	0.6860	0.9832	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>A276R</b>	33	1	0.8380	0.993	<0.0000	d.211.1.1
<b>F317L</b>	38	1	0.7950	0.986	<0.0000	<a href="#">d.185.1.1</a>
<b>F165R</b>	41	2	0.8190	0.9927	<0.0000	PDB entry is not classified in SCOPe 2.03



ORF	Sequence	Domain	TM-Score	Posterior prob	P-VALUE	SCOP assignment
<b>K205R</b>	43	1	0.8500	0.9941	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>K145R</b>	46	1	0.6890	0.9696	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>K421R</b>	47	3	0.8880	0.9765	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>EP84R</b>	49	1	0.6050	0.6379	<0.0000	d.68.4.1
<b>EP152R</b>	51	1	0.6650	0.9893	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>M1249L</b>	55	1	0.8050	0.997	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>M448R</b>	56	1	0.7710	0.9956	<0.0000	<a href="#">a.4.5.11</a>
<b>C84L</b>	58	1	0.6510	0.8101	<0.0000	d.58.7.1
<b>C717R</b>	59	1	0.9340	0.9974	<0.0000	a.118.1.1
<b>C122R</b>	60	1	0.4860	0.1863	<0.0000	Probability of classification low
<b>C275L</b>	61	1	0.7120	0.9509	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>C62L</b>	65	3	0.5130	0.525	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>B169L</b>	71	1	0.6720	0.8648	<0.0000	<a href="#">c.37.1.3</a>
<b>B475L</b>	72	1	0.7290	0.9524	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>B354L</b>	73	3	0.8500	0.9964	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>B125R</b>	77	1	0.8750	0.9953	<0.0000	<a href="#">b.91.1.1</a>
<b>B117L</b>	78	1	0.6100	0.692	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>B407L</b>	79	1	0.9450	0.9975	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>B263R</b>	81	2	0.6560	0.9965	<0.0000	<a href="#">d.129.1.1</a>
<b>B66L</b>	82	1	0.7490	0.9791	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>CP123L</b>	85	1	0.6440	0.8801	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>CP312R</b>	90	2	0.8480	0.9967	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>D129L</b>	97	1	0.5360	0.7424	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>D79L</b>	98	1	0.6390	0.7719	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>D339L</b>	99	1	0.8820	0.996	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>S183L</b>	104	1	0.8050	0.9912	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>H171R</b>	108	1	0.7400	0.9929	<0.0000	<a href="#">a.25.1.0</a>
<b>H124R</b>	109	1	0.7430	0.9867	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>H233R</b>	112	2	0.7130	0.9372	<0.0000	<a href="#">c.94.1.1</a>
<b>H240R</b>	113	1	0.6380	0.7665	<0.0000	b.1.18.15
<b>E184L</b>	118	1	0.8320	0.996	<0.0000	<a href="#">a.4.1.9</a>
<b>E423R</b>	120	1	0.8790	0.9975	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>E146L</b>	122	1	0.5220	0.6791	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>E111R</b>	128	1	0.5650	0.709	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>E66L</b>	129	1	0.5940	0.5631	<0.0000	PDB entry is not classified in SCOPe 2.03
<b>I267L</b>	130	1	0.7010	0.978	<0.0000	<a href="#">c.108.1.10</a>
<b>I226R</b>	131	1	0.8200	0.9923	<0.0000	<a href="#">d.159.1.3</a>
<b>I73R</b>	133	1	0.6590	0.8583	<0.0000	<a href="#">a.4.5.34</a>
<b>I329L</b>	134	1	0.7410	0.9861	<0.0000	<a href="#">c.72.1.5</a>

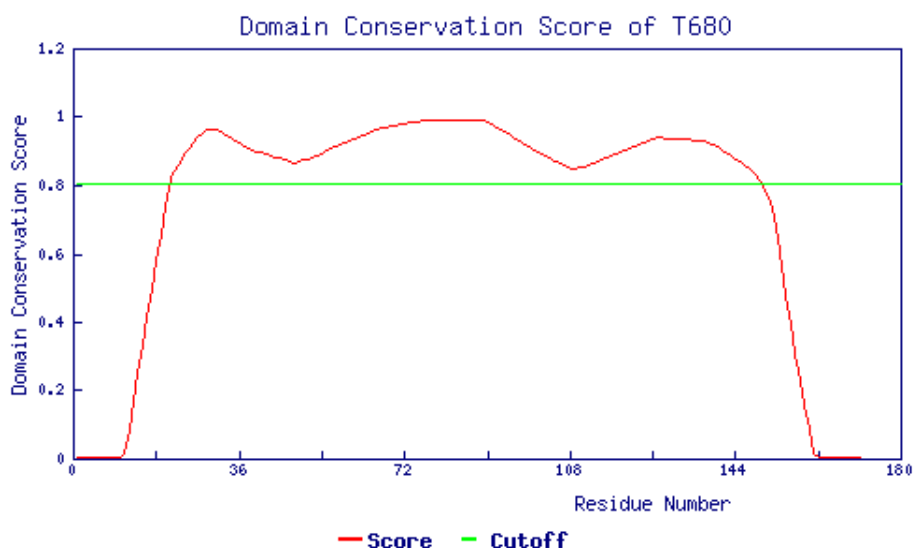


ORF	SEQ.No	Domain	TM-Score	Posterior prob	P-value	SCOP assignment
I177L	136	1	0.7510	0.9825	<0.0000	PDB entry is not classified in SCOPe 2.03
I196L	137	1	0.6750	0.9089	<0.0000	b.18.1.3
DP238L	138	1	0.8110	0.9909	<0.0000	c.1.8.0
DP311R	139	1	0.8610	0.9949	<0.0000	<a href="#">d.211.1.1</a>
DP63R	140	1	0.7020	0.9229	<0.0000	<a href="#">a.47.4.0</a>
DP542L	141	1	0.9590	0.9976	<0.0000	<a href="#">a.118.1.2</a>
DP141L	142	1	0.6770	0.6559	< 0.0000	PDB entry is not classified in SCOPe 2.03
DP146L	*143	1	0.8370	0.9929	<0.0000	<a href="#">g.3.11.1</a>
DP148R	144	1	0.5900	0.567	<0.0000	<a href="#">a.118.8.1</a>
DP96R	146	1	0.7560	0.9857	<0.0000	PDB entry is not classified in SCOPe 2.03
DP363R	147	1	0.6700	0.9919	<0.0000	<a href="#">d.211.1.1</a>
DP42R	148	1	0.8510	0.9954	<0.0000	PDB entry is not classified in SCOPe 2.03
DP60R	149	1	0.5010	0.7787	<0.0000	PDB entry is not classified in SCOPe 2.03
		94	0.7185			

TM-Score was based on a previously benchmarked set (Xu & Zhang 2010), with an all-to-all gapless structural match on 6684 non-homologous single-domain proteins in the PDB. TM-score of the non homologous set followed an extreme value distribution. The dataset used was from SCOP 1.75. Ankyrin group d.21.1.1 formed the largest group of families their functions are very diverse.

#### 4.4 H171R THREADOM RESULTS

The 171 residue H171R (ASFV108) ORF has not been characterized and the protein function has not been assigned. (<http://www.uniprot.org/uniprot/P0CA73>), *see appendix page 50&51*. Based on our results this protein may be of importance to the virus in macrophage survival and may be immunogenic. (Papinutto et al. 2002). Prediction of the domain boundary was done by ThreaDom and found to be a single domain protein. *see figure4*. Meaning we expect to find a PDB hit due to the likely completeness of the PDB for all single domain proteins (Zhang, Hubner, et al. 2006).



**Figure 4: Illustration of domain decision by ThreaDom for H171R ORF based on domain conservation score profile one domain has been defined by the ThreaDom algorithm**

#### 4.5 H171R TM FOLD AND I-TASSER RESULTS

From the prediction study of H171R, I-TASSER pipeline generated 5 models, the first model which represents the best template was considered, *Figure 5* was used (Roy et al. 2010a; Zhang 2008). Structural comparison of this model was done to the preexisting experimentally solved proteins in PDB 101 database by the structural alignment program TM align, a ferritin like subunit (PDB ID: 2C41C) (Franceschini et al. 2006) was found to be the structural analog. *Figures 6*, *TM align* the structural alignment programs' TM-score was 0.740 with an RMSD of 2.25 to the top scoring PDB 2C41C experimentally solved native. *See appendix page 48*. The significance of this TM-score was calculated to prove it was not a random score, TM fold was

used to calculate posterior probabilities and p-Values, we used an Extreme Value Distribution, the best fitting parameters being  $\mu=0.1512$  and  $\sigma=0.0242$  which were estimated by maximum likelihood from a benchmarked set of 71583085 random protein pairs. Equation 4 was used

$$P\text{-value}(x) = \int_x^1 f(x|\mu, \sigma) dx = 1 - \exp\left[-\exp\left(\frac{\mu-x}{\sigma}\right)\right] \quad \text{Equation 4}$$

$$P\text{value} = 1 - \exp(0 - \exp((\mu - x) / \sigma)); \quad \mu = 0.1512 \text{ and } \sigma = 0.0242.$$

The p-Value calculated by our TM-Fold Perl script was found to be 0.000000000002354 see excerpt page 47 showing the significance of the TM-score against randomly paired TM-Score.

A SCOP assignment was done to the modeled protein to show if it belonged to the same group as the 2C41C native, TM-Fold a structural classification software was used to ascertain the SCOP family probability of classification, the posterior probabilities of SCOP family have been benchmarked (Xu & Zhang 2010) and calculated using Bayes formula in Equation 3,

$$P(F|TM) = \frac{P(TM|F)P(F)}{P(TM|F)P(F) + P(TM|F^-)P(F^-)} \quad \text{Equation 3 (Bayes Equation)}$$

Here in Equation 3, TM stands for the TM-score of the compared proteins as calculated by the structural alignment program TM-align, (F) and (F<sup>-</sup>) represent the events that the proteins belong to the same and different Fold in SCOP, respectively P(F) and P(F<sup>-</sup>) are the prior probabilities of proteins in same and different folds; P(TM|F) and P(TM|F<sup>-</sup>) are the conditional probabilities of TM-score when the two proteins are in the same or different Fold families respectively. P(F|TM) is the posterior probability of the fold sharing the same group conditioned at a particular TM. All the data and parameters have been extracted from SCOP database. The posterior probabilities were curve fitted into a sigmoidal Boltzmann model and used to calculate the posterior probabilities. See Equation 5

$$\text{Boltzmann Model. } Y = \max + (\min - \max) / (1 + \exp((TM - Tm_{scop}) / dx_{scop})) \quad \text{Equation 5}$$

The parameters for the Boltzmann model incorporated into a Perl script to calculate the posterior probability are:  $a1_{scop}(\min) = -0.0071735$ ;  $a2_{scop}(\max) = 0.99803$ ;  $x0_{scop}$

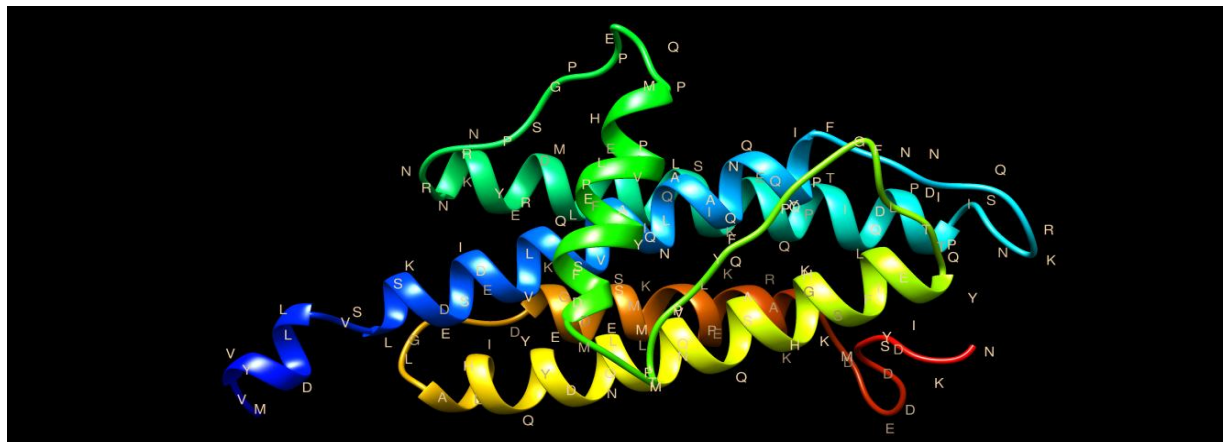
$$Tm_{scop} = 0.57917 \quad dx_{scop} = 0.048934; \quad (\text{see Perl excerpt pg 46})$$

*min* is the initial value (left horizontal asymptote) *max* final value (right horizontal asymptote) *x0* center (point of inflection), *dx* is the width (the change in X corresponding to the most

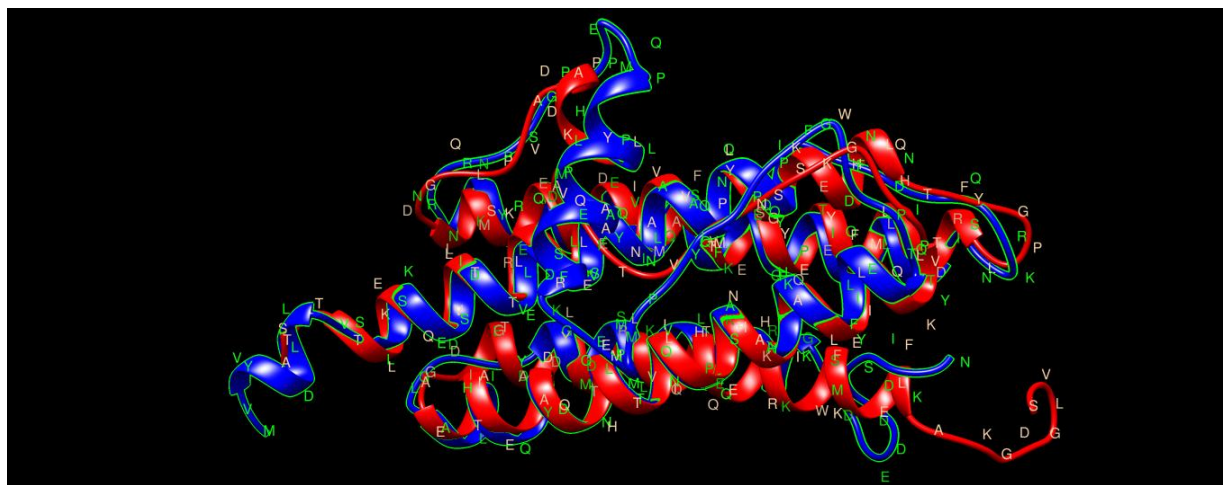


significant change in Y values).

The posterior probability from the Boltzmann fit equation was found to be 0.9929 for the SCOP family classification of H171R (ASFV108). Which proved with a 99.29% certainty, the I-TASSER modeled protein belonged to the same family as the experimentally solved PDB 2C41 Native, SCOP classification code for the family was a.25.1.0. A ferritin like protein. This close analogy protein (PDB ID: 2C41C) resembles a subunit of DNA-binding protein from starved cells in bacteria.



**Figure 5:**H171R I-TASSER modelled subunit resembling Ferritin like domain by I-TASSER the TM score .740 and the RMSD 2.25 to 2C41C,The N terminus starts from M (BLUE ) to (N red) .

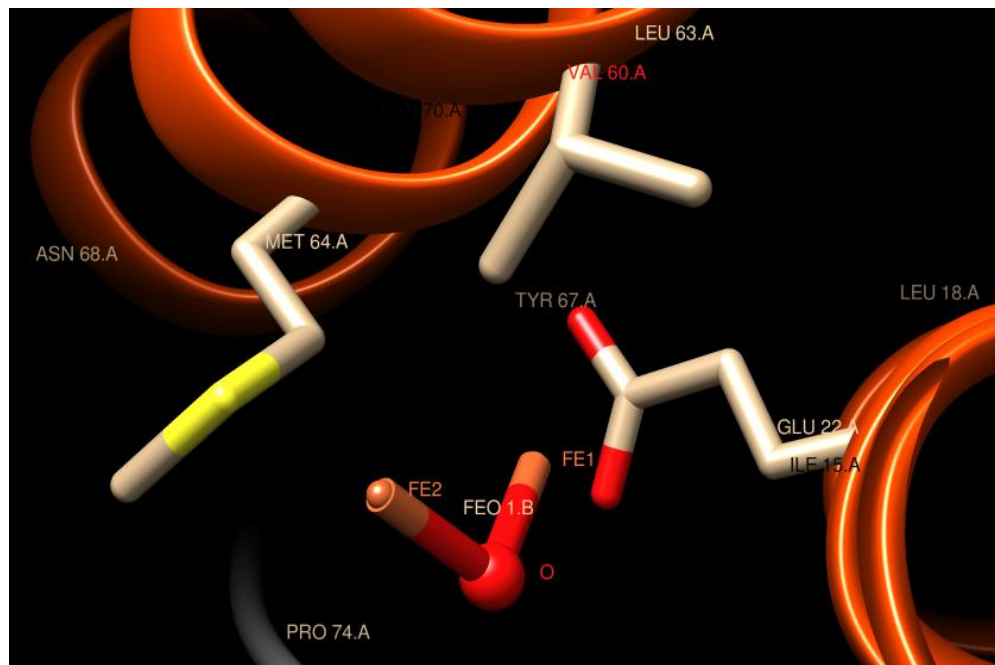


**Figure 6:** Structural superimposition of H171R I-TASSER model (Blue)and the native structure 2C41C ,The red model 2C41C is a representative of the PDB.



#### 4.5.1 H171R BINDING SITE RESULTS

Further analysis of the binding site and ligand screening see *Figure 7 and appendix page 48* confirmed that the binding site was similar to that of 1n1qA which is a subunit of DPS with ferroxidase activity in the

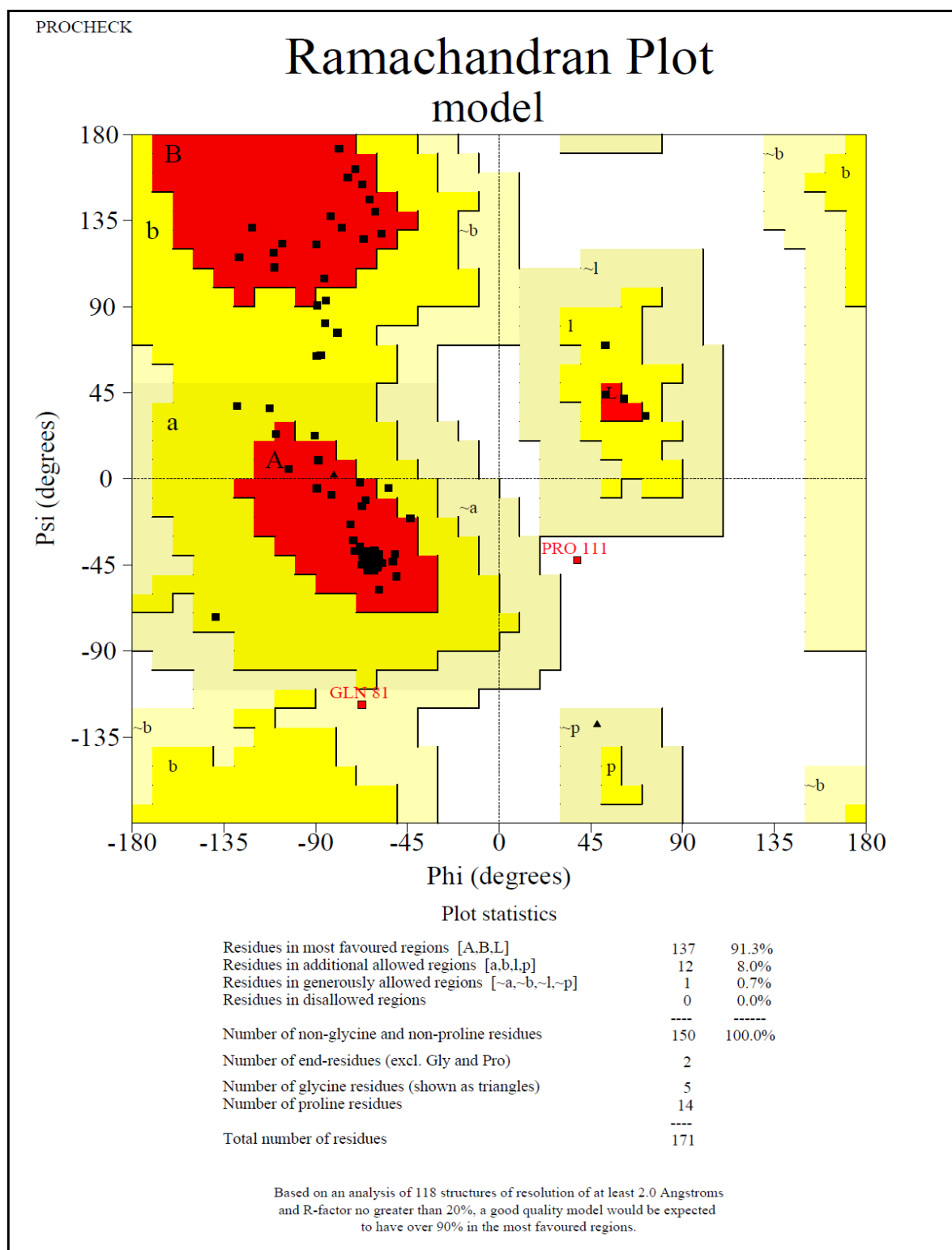


**Figure 7: Predicted binding site of the model with FEO ligand the model binds to residues Glu22, Val60 and Tyr 67.**

same family with a TM-Score of 0.707 at the ligand binding site, the ligand that binds was found to be FEO and it binds to residues 22 Glu, 60 Val and 67 Tyr, the binding site score was found to be 1.17 which is greater than 1.1. BS-score is a measure of local similarity between template binding site and predicted binding site in the query structure, It is based on large scale benchmarking analysis, it has been observed that a BS-score >1.1 reflects a significant local match between the predicted and template binding site.

#### 4.5.2 H171R PROCHECK RESULTS

A Ramachandran plot was done using Procheck (Laskowski et al. 1993) for the structure validation *Figure 8*. The plot developed validates the structure as all the empirically distributed data-points present in the structure are observed to lie in the allowed region. This indicates, the H171R model 1 predicted with the help of I-TASSER has approx. 90% of the residues in the allowed region conformation. The I-TASSER modeled and refined structure had approximately. 91.3% of the residues in the allowed region conformation by Procheck.



**Figure 8: A Ramachandran plot showing angular distribution in H171R , 91.3% of the residues lie in the allowed region using Procheck if a model has 90% of its residues in the allowed region it is good enough.**

#### 4.6 IMMUNODOMINANT PEPTIDES RESULTS:

NetMHCIIpan 3.0 an online web tool for the prediction of peptide binding was used, we used Swine Leukocyte Antigens DQ and Swine Leukocyte Antigens DRB proteins. The top ten epitopes in the tabular form are listed below *see Table 3 and Table 4.*

**Table 3: MHC class I1 binding peptides ranked on the basis of strong binding to MHC pseudo sequence YLFHETSGARTLHIVYFGHTYFDFQTETVHIETT: Swine Leukocyte Antigen –DQ Alpha and Beta protein sequences were used.**

PROTEIN	POSITION	PEPTIDE	1-log50k	AFFINITY
ASFVgp069	146	WARFGVAKAQMSALA	0.79	9.71
ASFVgp145	13	KHVRFAAAVEVWEAD	0.782	10.53
ASFVgp047	4	VDVVGIAEASAALYV	0.755	14.12
ASFVgp083	284	LINFTYARAQQVIK	0.745	15.72
ASFVgp067	135	GLIYATAGVLLAQLH	0.742	16.29
*ASFVgp108	14	IDVLRVFEANLAAFN	0.738	17.01
ASFVgp095	37	LFKTVYEALVAQEIP	0.726	19.46
ASFVgp064	73	FEATRLVAVRAQQLA	0.726	19.47
ASFVgp030	71	HHSEFSAEIAALLKL	0.724	19.81
*ASFVgp010	2	VIFLGILGLLANQVL	0.723	19.96

**Table 4: MHC class I1 binding peptides ranked on the basis of strong binding strength to MHC pseudo sequence QEFFIASGAAVDAILHLFLEQYDLQRETYHILFL: Swine Leukocyte Antigen –DR Alpha and Beta protein sequences were used.**

Protein	Position	Peptide	1-log50k	affinity
ASFVgp038	236	ILDIFMMLTSRRSLV	0.931	2.12
ASFVgp127	159	KELFLRIRNTRLKQI	0.922	2.33
ASFVgp093	1380	NVLLRMALSSPVQVL	0.911	2.61
ASFVgp026	370	PERVVKMAARLMRVD	0.896	3.09
ASFVgp091	119	VSYLIRIRAALKKKN	0.89	3.3
*ASFVgp010	4	FLGILGLLANQVLGL	0.887	3.39
ASFVgp059	26	RKWLTLQPSLLRYSG	0.887	3.41
ASFVgp037	74	ILNFLRLISGHRVVL	0.885	3.48
ASFVgp115	190	YNNIMQAKNIRILFL	0.885	3.48
ASFVgp053	3	FISIISVLSIRKRK	0.884	3.51

# Chapter 5

## 5.0 DISCUSSION AND CONCLUSION

### 5.1 DISCUSSION

#### 5.1.1 PROTEIN STRUCTURE PREDICTION AND CLASSIFICATION

As anticipated in our study, nearly all the 75 single domain Open Reading Frames assigned domain boundaries by ThreaDom and modeled by I TASSER, have structurally related experimentally solved folds in the PDB. This is shown by I TASSER having 74 out of the 75 Modeled structures, with a TM score greater than 0.5. Eight of the multidomained proteins predicted, had using had TM scores greater than 0.5. Therefore, there was no need for splitting the domains further. Structures were classified to belong to the same fold group if the TM. Score is greater than 0.5 (Xu & Zhang 2010).The likelihood of correctly assigning an accurately modeled single domain protein structure to a fold family was observed, and the near completeness of the PDB, for single domain proteins, as previously seen in genome scale applications is verifiable, making the protein structure prediction problem solvable (Zhang, Hubner, et al. 2006; Kihara & Skolnick 2003). The average TM score was 0.7185 for all the 83 Open Reading Frames. Only the ORF; C122R, had a TM-Score of 0.486, most probably it belonged to the category of new folds which do not have any local conformations in the PDB. SCOPe 2.03e, the manual classification system, assigns only 37 of the 83 modeled structures which have experimentally solved natives as templates and TM Scores greater than 0.5, to known protein families highlighting the slow nature and shortcomings of a manual structural classification system.This represents only 44% of the modeled structures.

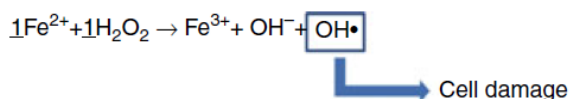
#### 5.1.2 ASFV H171R

ASFV has a host of survival mechanism that facilitate its persistence in macrophages, as one of the objectives the study has identified a subunit of the self assembling DPS (DNA binding Protein during Stress.) like protein (Zhang & Orner 2011). DPS was first characterized as a DNA-binding protein and was not at that time known to exhibit any ferroxidase activity. Extensive studies have shown it possesses, ferroxidase activity, the protein is classified to belong

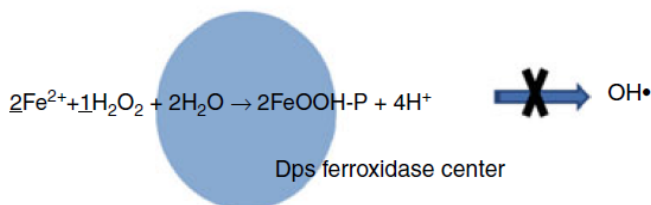
to the ferritin like family. DPS maybe crucial for viral survival in macrophages. Previous studies have shown the usefulness of DPS in surviving the oxidative stressful environment of macrophages (Theoret et al. 2011). Macrophages are phagocytic cells involved in immune coordination and RBC destruction ,90% of RBCs are removed from the circulation by the phagocytic activities of macrophages in the liver, spleen and lymph nodes, the Macrophage oxidative environment has free Fe(II) and Hydrogen peroxide which are potentially dangerous to ASFV, damaging effects arises when Fe<sup>2+</sup> and H<sub>2</sub>O<sub>2</sub> through Fentons reaction, *see figure 9*. leads to the production of OH· radicals .The OH· radicals can damage virtually all types of macromolecules (Tu et al. 2012): nucleic acids damage is severe as OH· radicals cause strand breaks, depyrimidation, depurinations and oxidation of bases. Lipids are damaged by peroxidation that decreases membrane fluidity. Proteins are damaged by oxidation of amino acids leading to fragmentation.

The diffusion of OH· radicals is limited, in that it is likely to react with an oxidizable substrate before travelling a long distance. However, OH· radicals can start a radical reaction which can result in injury far away from the site of OH· radical formation. From an enzymatic point of view, OH· radical, has a half life of 10<sup>-9</sup>s .Therefore its detoxification is not enzymatically feasible. The study results therefore hypothesize, based on Insilco protein structure prediction, the existence of a DPS like protein, that self assembles, and ASFV uses to protect itself from peroxide stress by inhibiting the iron-catalyzed production of OH· radicals. *See figure 9*.

(a) Iron oxidation via the Fenton reaction



(b) Iron oxidation in Dps ferroxidase center



**Figure 9: Iron Oxidation: Oxidation of ferrous ions in may occur via the Fenton Reaction. In this instance, the stoichiometric ratio of iron to hydrogen peroxide is 1: 1 and thus results in the production of hydroxyl radicals that can lead to damage of vital macromolecules and cell death. (b) Oxidation of ferrous ions occurs in the ferroxidase centre of Dps and proceeds via a reaction whereby the stoichiometric ratio of iron to hydrogen peroxide is 2 : 1. This reaction does not produce hydroxyl radicals; thus, the potential damage to cellular components is alleviated. Diagram courtesy of (Calhoun & Kwon 2011)**

### 5.1.3. EPITOPE PREDICTION

For proteins binding to SLA-DQ, Epitope **WARFGVAKAQMSALA** of ASFV 069 was found to have the highest antigenicity among all epitopes assuring maximum binding affinity after proteasome digestion. It is a prenyl transferase, gene deletion of this protein has been shown to have an effect virulence replication in macrophages and its conserved in isolates, (Alejo et al. 1999; Dixon et al. 2004).

H171R predicted peptide sequence **IDVLRFVEANLAAFN** ranked 6<sup>th</sup> Of the 149 peptides in binding to Swine Leukocyte Antigen DQ. confirming it was good enough as a potential candidate in peptide vaccine development, the protein has an effect on replication in macrophages (Carrillo et al. 1994; Moore et al. 1998),the primary cells ASFV targets. The peptide is conserved in all ASFV isolates.

Another peptide sequence which may be of value is ASFV 010 which binds both to SLA DQ and SLA DR strongly using the same peptide core **ILGLLANQV** underlying its importance as an immunogenic peptide, The protein for this sequence has only been assigned a family there is no general consensus on its function, it may be a Bromodomain, Bromodomains (BRDs) are protein interaction modules that specifically recognize  $\epsilon$ -N-lysine acetylation motifs but further investigation should be done.

## 5.2 CONCLUSION

In conclusion, this study found that the library of solved PDB structures is likely complete and clearly demonstrates that the majority of secondary structural elements and global chain contour similarity are retained for structures with a TM-score to native greater than 0.5. The key issue of identifying templates for the 83 of 149 targets where contemporary sequence identity and structure prediction methods fail is improved and applied on a genomic scale.

TM-score thresholds may be suitable for automated protein structure classifications, because of the rapid increase of protein structures accelerated by various proteomic projects, it is becoming increasingly infeasible as exemplified by SCOP, for the manual human visualization to conduct large-scale protein structure classifications. SCOP only managed to classify 37 out of the 83 possible targets. Therefore the usage of these quantitative scoring functions that corresponds to specific structural similarity levels should be adopted. The severity of this is seen with more than half the experimentally solved PDB natives not being able to be classified by SCOP.

The study identified and modeled ASFV108 ORF H171R, accurately, I-TASSER managed to build a full length model for this BA71V Isolate ORF that shares less than 13% primary sequence homology with any experimental solved protein structure an accurate functional assignment of ferroxidase activity was inferred ,the protein is crucial for ASFV survival in macrophages oxidatively stressful conditions, this protein has been considered as a vaccine candidate in various studies (Papinutto et al. 2002) and is a strong binder to SLA-DQ.

## 5.3 RECOMMENDATIONS:

There are still questions that remain to be addressed in future studies, one is the function of Ankyrin repeats in ASFV. Ankyrin repeats form the largest family and their functions are diverse. An extensive study should be done to see if they are implicated in virulence enhancement. Investigation of ASFV 010 should also be done it uses the same peptide binding core in binding to SLA DQ and SLA DR and is also immunogenic. However, there is no general consensus in its function. Investigations involving gene deletion of H171R as an attenuation mechanism and observation of the virus in an oxidatively stressful environment having H<sub>2</sub>O<sub>2</sub> and Fe<sup>2+</sup> can add more insight.

# BIBLIOGRAPHY

Abrams, C.C. et al., 2008. Domains involved in calcineurin phosphatase inhibition and nuclear localisation in the African swine fever virus A238L protein. *Virology*, 374(2), pp.477–86.

Alonso, C. et al., 2013. African swine fever virus-cell interactions: from virus entry to cell survival. *Virus research*, 173(1), pp.42–57.

Arakaki, A.K., Zhang, Y. & Skolnick, J., 2004. Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics (Oxford, England)*, 20(7), pp.1087–96.

Atuhaire, D.K. et al., 2013. Molecular characterization and phylogenetic study of African swine fever virus isolates from recent outbreaks in Uganda (2010-2013). *Virology journal*, 10(1), p.247.

Basta, S. et al., 2010. Cellular processes essential for African swine fever virus to infect and replicate in primary macrophages. *Veterinary microbiology*, 140(1-2), pp.9–17

Berman, H.M., 2008. The Protein Data Bank: a historical perspective. *Acta crystallographica. Section A, Foundations of crystallography*, 64(Pt 1), pp.88–95

Bork, P., Sander, C. & Valencia, A., 1993. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein science : a publication of the Protein Society*, 2(1), pp.31–40.

Bowie, J.U., Lüthy, R. & Eisenberg, D., 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (New York, N.Y.)*, 253(5016), pp.164–70.

Bystroff, C. & Baker, D., 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of molecular biology*, 281(3), pp.565–77

Calhoun, L.N. & Kwon, Y.M., 2011. Structure, function and regulation of the DNA-binding protein Dps and its role in acid and oxidative stress resistance in *Escherichia coli*: a review. *Journal of applied microbiology*, 110(2), pp.375–86.

Dixon, L.K. et al., 2004. African swine fever virus proteins involved in evading host defence systems. *Veterinary immunology and immunopathology*, 100(3-4), pp.117–34.



Dong Xu and Yang Zhang, 2013. Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-based Force Field. *Proteins*, 80(7), pp.1715–1735.

Ekins, S., Mestres, J. & Testa, B., 2007. In silico pharmacology for drug discovery: applications to targets and beyond. *British journal of pharmacology*, 152(1), pp.21–37.

Fischer, D. et al., 1996. Assigning amino acid sequences to 3-dimensional protein folds. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 10(1), pp.126–36

Fischer, D. & Eisenberg, D., 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proceedings of the National Academy of Sciences of the United States of America*, 94(22), pp.11929–34.

Fox, N.K., Brenner, S.E. & Chandonia, J.-M., 2014. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic acids research*, 42(Database issue), pp.D304–9.

Franceschini, S. et al., 2006. Antioxidant Dps protein from the thermophilic cyanobacterium *Thermosynechococcus elongatus*. *The FEBS journal*, 273(21), pp.4913–28

Gene, T. & Consortium, O., 2000. Gene Ontology : tool for the. *nature genetics*, 25(may), pp.25–29.

Goatley, L.C. & Dixon, L.K., 2011. Processing and localization of the african swine fever virus CD2v transmembrane protein. *Journal of virology*, 85(7), pp.3294–305

Idrees, S. & Ashfaq, U. a, 2013. Structural analysis and epitope prediction of HCV E1 protein isolated in Pakistan: an in-silico approach. *Virology journal*, 10, p.113

Idrees, S., Ashfaq, U. a & Khaliq, S., 2013. HCV Envelope protein 2 sequence comparison of Pakistani isolate and In-silico prediction of conserved epitopes for vaccine development. *Journal of translational medicine*, 11(1), p.105.

Iyer, L.M., Aravind, L. & Koonin, E. V, 2001. Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of virology*, 75(23), pp.11720–34

Jethra, G. et al., 2012. Structure and function prediction of unknown wheat protein using LOMETS and I-TASSER. *The Indian Journal of Agricultural Sciences*, 82(October), pp.867–874.

Karosiene, E. et al., 2013. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10), pp.711–24.

- Kemege, K.E. et al., 2011. Ab initio structural modeling of and experimental validation for Chlamydia trachomatis protein CT296 reveal structural similarity to Fe(II) 2-oxoglutarate-dependent enzymes. *Journal of bacteriology*, 193(23), pp.6517–28.
- Kihara, D. et al., 2002. Ab initio protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9), pp.5993–8
- Kihara, D. & Skolnick, J., 2003. The PDB is a covering set of small protein structures. *Journal of molecular biology*, 334(4), pp.793–802.
- Kim, D. et al., 2003. PROSPECT II: protein structure prediction program for genome-scale applications. *Protein engineering*, 16(9), pp.641–50.
- Kopp, J. et al., 2007. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, 69 Suppl 8(April), pp.38–56
- Laskowski, R.A. et al., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), pp.283–291
- Levitt, M. & Gerstein, M., 1998. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.5913–20.
- Malmström, L. et al., 2007. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS biology*, 5(4), p.e76.
- Miskin, J.E., Abrams, C.C. & Dixon, L.K., 2000. African swine fever virus protein A238L interacts with the cellular phosphatase calcineurin via a binding domain similar to that of NFAT. *Journal of virology*, 74(20), pp.9412–20.
- Murzin, A.G. et al., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4), pp.536–40.
- Nielsen, M. et al., 2010. NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome research*, 6(1), p.9.
- Papinutto, E. et al., 2002. Structure of two iron-binding proteins from Bacillus anthracis. *The Journal of biological chemistry*, 277(17), pp.15093–8
- Piriou-Guzylack, L. & Salmon, H., 2008. Membrane markers of the immune cells in swine: an update. *Veterinary research*, 39(6), p.54.
- Revilla, Y. et al., 1997. Inhibition of apoptosis by the African swine fever virus Bcl-2 homologue: role of the BH1 domain. *Virology*, 228(2), pp.400–4.

- Roy, A., Kucukural, A. & Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4), pp.725–38.
- Roy, A., Yang, J. & Zhang, Y., 2012. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40(Web Server issue), pp.W471–7.
- Sánchez, R. & Sali, A., 1998. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23), pp.13597–602
- Skolnick, J. et al., 2009. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(37), pp.15690–5.
- Skolnick, J., Zhou, H. & Brylinski, M., 2012. Further evidence for the likely completeness of the library of solved single domain protein structures. *The journal of physical chemistry. B*, 116(23), pp.6654–64.
- Theoret, J.R. et al., 2011. A *Campylobacter jejuni* Dps homolog has a role in intracellular survival and in the development of campylobacteriosis in neonate piglets. *Foodborne pathogens and disease*, 8(12), pp.1263–8
- Tonello, F. et al., 1999. The *Helicobacter pylori* neutrophil-activating protein is an iron-binding protein with dodecameric structure. , 34, pp.238–246
- Vallée, I., Tait, S.W. & Powell, P.P., 2001. African swine fever virus infection of porcine aortic endothelial cells leads to inhibition of inflammatory responses, activation of the thrombotic state, and apoptosis. *Journal of virology*, 75(21), pp.10372–82
- Wu, S., Skolnick, J. & Zhang, Y., 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC biology*, 5, p.17.
- Wu, S. & Zhang, Y., 2008. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics (Oxford, England)*, 24(7), pp.924–31
- Wu, S. & Zhang, Y., 2007. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*, 35(10), pp.3375–82
- Xu, D. & Zhang, Y., 2013. Ab Initio structure prediction for *Escherichia coli*: towards genome-wide protein structure modeling and fold assignment. *Scientific reports*, 3, p.1895.
- Xu, J. & Zhang, Y., 2010. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics (Oxford, England)*, 26(7), pp.889–95

- Xue, Z. et al., 2013. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* (Oxford, England), 29(13), pp.i247–56.
- Yang, J., Roy, A. & Zhang, Y., 2013. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research*, 41(Database issue), pp.D1096–103.
- Yáñez, R.J. et al., 1995. Analysis of the complete nucleotide sequence of African swine fever virus. *Virology*, 208(1), pp.249–78.
- Zhang, Y., 2014. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins*, 82 Suppl 2(April), pp.175–87.
- Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, 9, p.40
- Zhang, Y., 2009. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, 77 Suppl 9(Suppl 9), pp.100–13.
- Zhang, Y., Hubner, I. a, et al., 2006. On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp.2605–10
- Zhang, Y., 2010. Protein Structure Prediction: Is It Useful? *Current Opinion in Structural Biology*, 19(2), pp.145–155.
- Zhang, Y., Devries, M.E. & Skolnick, J., 2006. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS computational biology*, 2(2), p.e13.
- Zhang, Y. & Skolnick, J., 2004a. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20), pp.7594–9.
- Zhang, Y. & Skolnick, J., 2004b. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4), pp.702–10.
- Zhang, Y. & Skolnick, J., 2004c. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophysical journal*, 87(4), pp.2647–55.
- Zhang, Y. & Skolnick, J., 2005a. The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4), pp.1029–34.
- Zhang, Y. & Skolnick, J., 2005b. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research*, 33(7), pp.2302–9

## APPENDICES

### APPENDIX I : PREDICTED PROTEINS

	<b>PROTEINS BA71V</b>	<b>GENE NAME</b>	<b>ORF NUMBER</b>
1	Thymidylate kinase	A240L	30
2	Thymidine kinase	K196R	45
3	dUTPase*	E165R	124
4	Ribonucleotide reductase (small subunit)	F334L	39
5	Ribonucleotide reductase (large subunit)	F778R	40
6	DNA polymerase $\beta$	G1211R	84
7	DNA topoisomerase type II*	P1192R	106
8	Proliferating cell nuclear antigen (PCNA) like	E301R	121
9	DNA polymerase family X *	O174L	91
10	DNA ligase*	NP419L	94
11	Putative DNA primase	C962R	66
12	AP endonuclease class II*	E296R	127
13	RNA polymerase subunit 2	EP1242L	48
14	RNA polymerase subunit 6	C147L	64
15	RNA polymerase subunit 1	NP1450L	93
16	RNA polymerase subunit 3	H359L	107
17	RNA polymerase subunit 5	D205R	102
18	RNA polymerase subunit 10	CP80R	89
19	TFIIB like	C315R	63
20	Helicase superfamily II	A859L	35
21	Helicase superfamily II similar to origin binding protein	F1055L	42
22	Helicase superfamily II	B962L	67
23	Helicase superfamily II VV D6/D11 like involved in transcription termination	D1133L	100
24	Helicase superfamily II VV D5 like	Q706L	115
25	Helicase superfamily II VV A18 like	QP509L	116
26	Transcription factor SII	I243L	132
27	Guanyl transferase*	NP868R	95
28	Poly A polymerase large subunit	C475L	62
29	FTS J like methyl transferase domain	EP424R	50
30	ERCC4 nuclease domain	EP364R	54
31	Lambda-like exonuclease	D345L	103
32	VV A2L like transcription factor	B385R	75
33	VV A7L like transcription factor	G1340L	83
34	VV VLTF2 like late transcription factor, FCS like finger	B175L	80
35	VV D5 like ATPase involved in replication	C962R	66

## APPENDIX I

	<b>OTHER ENZYMES WITH UNKNOWN ROLES</b>		
1	Prenyl transferase*	B318L	69
2	Serine protein kinase*	R298L	114
3	Ubiquitin conjugating enzyme*	I215L	135
4	Nudix hydrolase*	D250R	96
	<b>HOST CELL INTERACTIONS</b>		
1	IAP apoptosis inhibitor*	A224L	28
2	Bcl 2 apoptosis inhibitor*	A179L	36
3	IkB homolog and inhibitor of calcineurin phosphatase*	A238L	34
4	C type lectin like*	EP153R	52
5	CD2 like. Causes haemadsorption to infected cells	EP402R	53
6	Similar to HSV ICP34.5 neurovirulence factor	DP71L	145
7	Nif S like PLP dependent transferase	QP383R	117
8	Mn dependent superoxide dismutase	C129R	57
	<b>STRUCTURAL PROTEINS AND PROTEINS INVOLVED IN</b>		
1	P22 Transmembrane domain.	KP177R	5
2	A104R Histone-like structural protein.HF-like DNA binding protein	A104R	29
3	P11.5	A137R	37
4	P10	A78R	44
5	P72 Major capsid protein. Involved in virus entry	B646L	76
6	B438L Required for formation of vertices in icosahedral capsid	B438L	70
7	B602L Chaperone. Involved in folding of p72 capsid protein. .	B602L	74
8	B119L ERV 1 like. Involved in redox metabolism* Not incorporated into	B119L	68
9	Sumo 1 like protease. Involved in polyprotein cleavage	S273R	105
10	P220 Polyprotein precursor of p150, p37, p14, p34 coreshell components.	CP2475L	86
11	P30 Phosphoprotein. Involved in virus entry	CP204L	87
12	P60 Polyprotein precursor of p35 and p15. Present in core shell	CP530R	88
13	P12 Attachment protein involved in virus entry.Transmembrane domain	O61R	92
14	P17 Transmembrane domain.	D117L	101
15	J5R Transmembrane domain.	H108R	111
16	P54 Binds to LC8 chain of dynein, involved in virus entry.Transmembrane	E183L	119
17	J18L Transmembrane domain. VV J5 like membrane protein	E199L	123
18	E248R Possible component of redox pathway required disulfide bond	E248R	125
19	A151R Contains CXXC motif similar to that in thioredoxins. Binds to	A151R	32
20	P14.5 DNA binding. Required for movement of virions to plasma	E120R	126
21	XP124L Multigene family 110 member. Contains KDEL ER retrieval	XP124L	110

## APPENDIX II : MULTIGENE FAMILIES

	<b>MULTIGENE FAMILIES UNANNOTATED</b>		
1	<b>Multi gene family 360</b>	KP360L	003
2		KP362L	004
3		UP60L	013
4		L356L	007
5		J319L	019
	<b>Multi gene family 110</b>		
1		U104L	009
2		XP124L	010
3		V82L	011
4		Y118L	012
	<b>Multi gene family 300</b>		
1		J154R	016
2		J104R	017
3		J182L	018
4		J318L	019
	<b>Multi gene family 505/530</b>		
1		A125L	020
2		A489R	021
3		A280R	022
4		A505R	023
5		A498R	024
6		A528R	025
7		A506R	026
8		A542R	027
	<b>Multi gene family 142(DP141L)</b>	DP141L	142

### APPENDIX III: H171R

#### Avirulent isolate pig BA71V ASFV genome uncharacterised

>ASFVgp108 H171R pH171R 135445:135960 forward MW:19952

MVVYDLLVSLSKESIDVLRVFEANLAAFNQQYIFFNIQRKNSITTPLLITPQQEKISQIVEF  
LMDEYNKNNRRPSGPPREQPMHPLLPYQQSSDEQPMMPYQQPPGNDDQPYEQIYHKKH  
ASQQVNTTELNDYYQHILALGDEDKGMDSMLKLPEKAKRGSDDDEDDMFSIKN

### APPENDIX 1V: SLA DQ AND DR PROTEINS.

>tr|Q31065|Q31065\_PIG MHC ClassII OS=Sus scrofa GN=SLA-DR-alpha PE=2 SV=1

MTILGVPVVLGFVITILNLQKSWAIVENHVIIQAEFYLSPKSGEFMFDFDGDGEIFHVDMEKRETVW  
RLEEFGHFASFEAQGALANIAVDKANLEILIKRSNNTPTNTNVPPEVTVLSDKPVELGEPNIIICFID  
KFSPPVNVNVTWLRNGSPVTRGVSETVFLPREDHLFRKFHYLPFMPSTEDVYDCQVEHWGLDKPL  
LKHWEFEAQTPLETTENTVCALGLIVALVGIIVGTVLIKGVVRKGNATERRGPL

>tr|Q31072|Q31072\_PIG MHC class II antigen OS=Sus scrofa GN=LA-DRB-d PE=2 SV=1

MLHLCFSRGFWMAALTVMLVVLSPPLALARDTPPHFLHLLKFECHFFNGTERVRLLERQYYNGE  
EFLRFSDSDVGEYRAVTELGRPDAKDWNSQKDLLEQRRAEVDTYCRHNYRILDFTFLVPRRAEPTV  
TVYPAKTQPLQHHNLLVCSVTGFYPGHVEVRWFRNGQEEAAGVVSTGLIPNGDWTFTQTMVML  
ETVPQSGEVYSCRVEHPSLTSPVTVEWRARSESAQGMMSGIGGFVLGLLFVAVGLFIYFKNQK  
GRPALQPTGLLS

>tr|Q4W5W7|Q4W5W7\_PIG MHC class II antigen OS=Sus scrofa GN=SLADQA PE=2 SV=1

MVPGRVLMWGALALTTVMSACGGEDIAADHVASYGLNVYQSYGPSGYFTHEFDGDEEFYVDL  
EKKETVWRLPLFSEFTSFDPPQALRNIA TLKHNLNIVTKRSNNTAAVNQVPEVTVFSKSPVILGQP  
NTLICHVDSIFPPVINITWLKNGHSVKGFSETSFLSKNDHSFLKISYLTFLPSDDDFYDCKVEHWGL  
DKPLLKHWEPEIPAPMSELTETVVCALGLIVGLVGIIVGTVFIIQGLRSGGPSRHQGS

>tr|O98263|O98263\_PIG MHC class II OS=Sus scrofa GN=SLADQB PE=2 SV=1

MSGMVALRLPRGLWTAALTVMLVVLGAPVAEGRDSPQDFVYQFKGECYFFNGTQVRVHVTRY  
IYNQEEHVRFSDSDVGEFRAVTPLRPDADYWNGQKDFLEQTRAELDTVCKHNYQIEEGTTLQRR  
VQPTVTISPSKAEALNHHNLLVCAVTDYFYSQVKVQWFRNGQEETAGVVSTPLIRNGDWTYQVL  
VMLEMNLQRGDVYTCRVEHSSLQSPILVEWRAQSESAQSKMLSGVGGFVLGLIFLGLGLFIRHRS  
QKGLVR



## APPENDIX V: PERL EXCERPT FOR STATISTICAL CALCULATIONS

```
#!/usr/bin/perl
use Math::Trig;
#that line indicates the path to the language interpreter when the script is run from the command
line. In the case of perl /usr/bin/perl is the path to the perl interpreter.
#defines many trigonometric functions not defined by the core Perl
##### statistical analysis of tmscore #####
# set parameters of posterior probability and EVD
$a1_scop=-0.0071735;
$a2_scop=0.99803;
$x0_scop=0.57917;
$dx_scop=0.048934;
$u=0.1512;
$sigma=0.0242;
# pvalue and posterior probability calculation
$pvalue=1-exp(0-exp(($u-$tms_new)/$sigma));
$prob_scop=$a2_scop+($a1_scop-$a2_scop)/(1+exp((($tms_new-$x0_scop)/$dx_scop));
if($prob_scop<0){$prob_scop=0;}
if($prob_scop>1){$prob_scop=1;}
if($pvalue<0){$pvalue=0;}
if($pvalue>1){$pvalue=1;}
$prob_scop=sprintf("%.4f",$prob_scop);
$pvalue=sprintf("%.15f",$pvalue);

#####outputresult#####
print "Statistical Scores:\n";
print "Probability of sharing same fold (SCOP):           $prob_scop\n";
print "P-value (Significance of the alignment):           $pvalue\n\n";
```

## APPENDIX VI: I TASSER STRUCTURAL PREDICTIONS

Proteins with highly similar structure in PDB (as identified by [TM-align](#))

### Top 10 Identified structural analogs in PDB

Click to view	Rank	PDB Hit	TM-score	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov.	Download Alignment
	1	<a href="#">2c41C</a>	0.740	2.25	0.124	0.848	<a href="#">Download</a>
	2	<a href="#">2vxxD</a>	0.735	2.84	0.131	0.860	<a href="#">Download</a>
	3	<a href="#">1tjoD</a>	0.733	2.75	0.105	0.854	<a href="#">Download</a>
	4	<a href="#">2fjcA</a>	0.725	2.71	0.110	0.854	<a href="#">Download</a>
	5	<a href="#">2c2uA</a>	0.715	2.42	0.107	0.842	<a href="#">Download</a>
	6	<a href="#">2d5kD</a>	0.715	2.38	0.134	0.830	<a href="#">Download</a>
	7	<a href="#">2chpA</a>	0.711	2.48	0.098	0.836	<a href="#">Download</a>
	8	<a href="#">2c2jA</a>	0.710	3.03	0.128	0.866	<a href="#">Download</a>
	9	<a href="#">1o9rF</a>	0.709	2.64	0.081	0.842	<a href="#">Download</a>
	10	<a href="#">2vzbA</a>	0.708	2.88	0.085	0.854	<a href="#">Download</a>

ALL  
ARE  
DPS

- (a) Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.  
 (b) Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library.  
 (c) RMSD<sup>a</sup> is the RMSD between residues that are structurally aligned by TM-align.  
 (d) IDEN<sup>a</sup> is the percentage sequence identity in the structurally aligned region.  
 (e) Cov. represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

All identified structural analogs Similar to the H171R model belong to the Dps (ferritin group)

## APPENDIX VII: COFACTOR BINDING SITE PREDICTION

### Predicted Binding Site

#### Template proteins with similar binding site:

Click to view	Rank	Cscore <sup>LB</sup>	PDB Hit	TM-score	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov.	BS-score	Lig. Name	Download Complex	Predicted binding site residues
	1	0.11	<a href="#">1n1qA</a>	0.707	2.58	0.124	0.842	1.17	FEO	<a href="#">Download</a>	22,60,67
	2	0.09	<a href="#">1hrsA</a>	0.629	3.34	0.125	0.807	1.07	PP9	<a href="#">Download</a>	26,29,30,33,53,54,57,61,85
	3	0.09	<a href="#">3ravA</a>	0.632	3.33	0.126	0.807	0.97	RAV	<a href="#">Download</a>	26,29,30,53,57
	4	0.01	<a href="#">1csmB</a>	0.374	5.73	0.058	0.655	0.62	TRP	<a href="#">Download</a>	108,110,113

Click on the radio buttons to visualize predicted binding site and residues.

- (a) Cscore<sup>LB</sup> is the confidence score of predicted binding site. Cscore<sup>LB</sup> values range in between [0-1]; where a higher score indicates a more re  
 (b) BS-score is a measure of local similarity (sequence & structure) between template binding site and predicted binding site in the query structu  
 significant local match between the predicted and template binding site.  
 (c) TM-score is a measure of global structural similarity between query and template protein.  
 (d) RMSD<sup>a</sup> the RMSD between residues that are structurally aligned by TM-align.  
 (e) IDEN<sup>a</sup> is the percentage sequence identity in the structurally aligned region.  
 (f) Cov. represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the

## APPENDIX VIII: GENE ONTOLOGY PREDICTIONS

Predicted GO terms															
Rank	C <sub>score</sub> <sup>GO</sup>	TMscore	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov.	PDB Hit	Associated GO Terms								
1	0.26	0.6943	2.89	0.10	0.84	<a href="#">2bk6C</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0055114</a>	<a href="#">GO:0005737</a>	<a href="#">GO:0046872</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0009405</a>
2	0.24	0.6873	2.63	0.09	0.81	<a href="#">1j14A</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0005737</a>	<a href="#">GO:0055114</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0009405</a>	<a href="#">GO:0046872</a>
3	0.24	0.7066	2.58	0.12	0.84	<a href="#">1n1qA</a>	<a href="#">GO:0003677</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0005737</a>	<a href="#">GO:0009295</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0046872</a>	<a href="#">GO:0030261</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0046914</a>
4	0.24	0.7244	2.70	0.11	0.85	<a href="#">2fjzB</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>			
5	0.23	0.7107	2.48	0.10	0.84	<a href="#">2chpA</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0055114</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0003677</a>		
6	0.23	0.7078	2.61	0.09	0.84	<a href="#">3ge4A</a>	<a href="#">GO:0046872</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>		
7	0.23	0.7081	2.48	0.13	0.83	<a href="#">2d5kB</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>			
8	0.23	0.6548	2.30	0.10	0.75	<a href="#">2qqvA</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>				
9	0.22	0.7084	2.88	0.09	0.85	<a href="#">2vzbA</a>	<a href="#">GO:0046872</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0008199</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>			
10	0.22	0.7403	2.25	0.12	0.85	<a href="#">2c41C</a>	<a href="#">GO:0046914</a>	<a href="#">GO:0055114</a>	<a href="#">GO:0006950</a>	<a href="#">GO:0006879</a>	<a href="#">GO:0016491</a>	<a href="#">GO:0003677</a>	<a href="#">GO:0008199</a>		

**ALL ARE FERRITIN**

Consensus Prediction of Gene Ontology terms					
Molecular Function		Biological Process		Cellular Location	
GO term	GO-Score	GO term	GO-Score	GO term	GO-Score
<a href="#">GO:0008199</a>	0.75	<a href="#">GO:0006879</a>	0.75	<a href="#">GO:0005737</a>	0.57
<a href="#">GO:0016491</a>	0.75	<a href="#">GO:0006950</a>	0.75	<a href="#">GO:0043232</a>	0.48
<a href="#">GO:0003677</a>	0.41	<a href="#">GO:0055114</a>	0.75		
		<a href="#">GO:0006323</a>	0.48		
		<a href="#">GO:0051276</a>	0.48		
		<a href="#">GO:0009405</a>	0.44		

**ALL ARE DPS (FERRITIN)**

## APPENDIX IX : PREDICTED ENZYME FUNCTION:

### Function Prediction using COFACTOR

#### Predicted EC Numbers

#### Top 5 enzyme homologs in PDB

Click to view	Rank	C <sub>score</sub> <sup>EC</sup>	PDB Hit	TM-score	RMSD <sup>a</sup>	IDEN <sup>a</sup>	Cov.	EC Number	Predicted Active Site Residues
	1	0.203	<a href="#">2htnG</a>	0.658	3.23	0.101	0.801	<a href="#">1.16.3.1</a>	25
	2	0.183	<a href="#">1uzrB</a>	0.665	3.48	0.045	0.871	<a href="#">1.17.4.1</a>	NA
	3	0.182	<a href="#">2wlaA</a>	0.708	2.76	0.065	0.842	<a href="#">1.16.3.1</a>	30
	4	0.177	<a href="#">3ee4A</a>	0.695	3.32	0.045	0.883	<a href="#">1.17.4.1</a>	NA
	5	0.173	<a href="#">1h0nA</a>	0.659	3.39	0.086	0.866	<a href="#">1.17.4.1</a>	NA

The enzymes 1.16.3.1 are classified as having ferroxidase activity:

## APPENDIX X: UNIPROT UNCHARACTERISED PROTEINS OF H171R

Firefox | Hydrogen-peroxide-induced heme d... | http://www.ebi.ac.../2004\_9/Page2.htm | h171r in UniProtKB | Analysis of the Complete Nucleotide ...

www.uniprot.org/uniprot/?query=h171r&sort=score

UniProtKB

Search | Blast | Align | Retrieve | ID Mapping \*

Search in: Protein Knowledgebase (UniProtKB) | Query: h171r | Search | Advanced Search | Clear

9 results for h171r in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Page 1 of 1

Results Customize

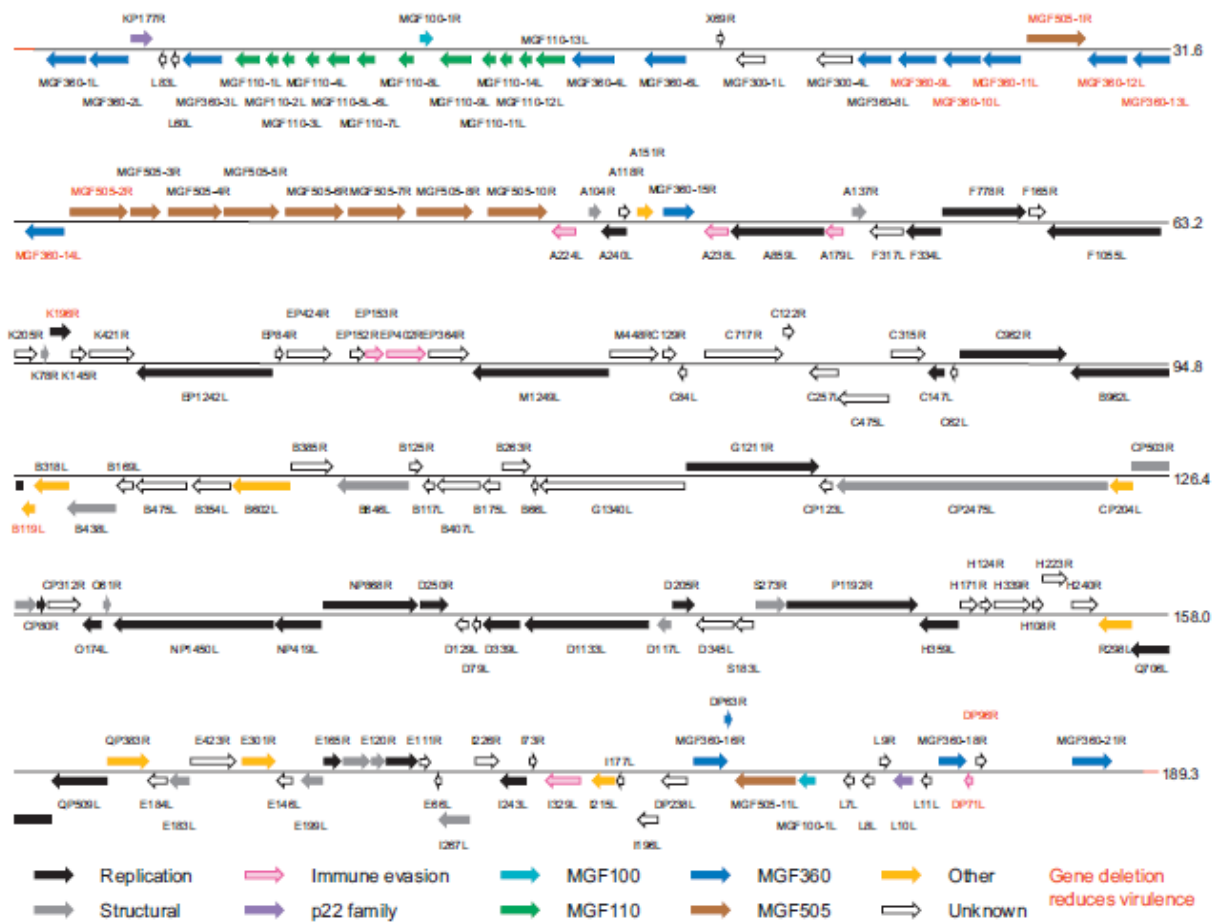
Show only reviewed (5) (UniProtKB/Swiss-Prot) or unreviewed (4) (UniProtKB/TrEMBL) entries

Did you mean 171r (35)?

Restrict term "h171r" to protein family (5), gene name (5), protein name (7)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q65185	VF171_ASF7	★	Uncharacterized protein H171R	Ba71V-114 H171R	African swine fever virus (strain Badajoz 1971 Vero-adapted) (Ba71V) (ASFV)	171
<input type="checkbox"/> Q65230	VF171_ASF2	★	Uncharacterized protein H171R	Mal-122 j2R	African swine fever virus (isolate Tick/Malawi/Lil 20-1/1983) (ASFV)	175
<input type="checkbox"/> P0CA75	VF171_ASF4	★	Uncharacterized protein H171R	War-124	African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)	171
<input type="checkbox"/> P0CA73	VF171_ASF5	★	Uncharacterized protein H171R	Ken-126	African swine fever virus (isolate Pig/Kenya/KEN-50/1950) (ASFV)	175
<input type="checkbox"/> P0CA74	VF171_ASF6	★	Uncharacterized protein H171R	Pret-126	African swine fever virus (isolate Tick/South Africa/Pretoriuskop Pr4/1996) (ASFV)	171
<input type="checkbox"/> E0WM64	E0WM64_ASF	★	H171R	H171R ASFV-Georgia_4-138	African swine fever virus Georgia 2007/1	171
<input type="checkbox"/> D4I5V1	D4I5V1_ASF	★	BA71V-H171R (J2R) protein	BA71V-H171R (j2R) E75_121	African swine fever virus E75	171
<input type="checkbox"/> A9JM28	A9JM28_ASFPP	★	PH171R	H171R	African swine fever virus (isolate Pig/Portugal/OURT88/1988) (ASFV)	171
<input type="checkbox"/> A9JL93	A9JL93_ASFPP	★	PH171R	H171R	African swine fever virus (isolate Pig/Benin/Ben 97-1/1997) (ASFV)	171

Page 1 of 1



Genome organisation map showing uncharacterised proteins of H171R. Diagram courtesy of (Dixon 2013).