UNIVERSITY OF NAIROBI

CENTRE FOR BIOTECHNOLOGY AND BIOINFORMATICS

IN SILICO EXPLORATION OF 3D STRUCTURES OF *PfMSP3* AND *PfMSP6* INVASION GENES AND THEIR ALLELIC DIFFERENCES

BY

PAULINE WAMBUI WANGUNYU

I56/68370/11

SUPERVISORS

PROFESSOR JAMES OCHANDA

PROFESSOR PETER WAGACHA

DR. ISABELLA OYIER

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

DECLARATION

This thesis as presented is my original work and to the best of my knowledge has not been presented in any other institution.

Signature: ……..

Date ..…….

Pauline Wambui Wangunyu

I56/68370/11

This thesis has been submitted with my approval as university supervisor

Signature: ……..

Date ..…….

Professor James Ochanda

Centre for Biotechnology and Bioinformatics

Signature: ……..

Date ..…….

Professor Peter Wagacha

School of Computing and Informatics

Signature: ……

Date ..……...

Dr. Isabella Oyier

Centre for Biotechnology and Bioinformatics

ABSTRACT

Malaria is one of the leading causes of child mortality in Africa. According to the *World Health Organization report 2011,* there was an estimated 655,000 deaths in 2010 in Africa most of which have been reported among children. This implies that malaria is a threat to the human population. One of the ways to deal with this problem is to develop effective vaccines against malaria. Antibody Dependent Cellular Inhibition (ADCI) is a phenomenon where individuals who have been exposed to malaria develop a form of immunity also known as premunition. The ADCI effect is effective against some *Plasmodium falciparum* invasion proteins such as Merozoite Surface Protein (MSP) 3 and 6 making them potential candidates for a malaria vaccine.

This study was conducted to determine the 3D structures of MSP3 and MSP6 as well as determine the allelic differences between 3D7 and K1 strains. Due to the lack of good templates we resorted to use *ab initio* modeling which is only able to model very small proteins or peptides from larger ones. Smaller regions of interest such as high activity binding peptides, regions that have been shown to elicit inhibition in *in vitro* assays as well as epitopes were identified and modeled. For the regions modeled and verified to be correct, we were able to identify pockets that are potential protein-ligand interaction as well as antibody binding sites. This research will be useful to researchers focusing on malaria drugs and vaccines who need to know what type of ligands to design and which areas to target.

ACKNOWLEDGEMENTS

I am very grateful to God for enabling me to carry out this work successfully. Special appreciation goes to the University of Nairobi that gave me a scholarship that enabled me to undertake my Masters degree. I would also like to appreciate all my supervisors for their guidance and support. I particularly would like to thank Dr. Oyier who proposed the idea and has been there throughout providing input and positive criticism which has led to the positive culmination of this work. I also thank the School of computing and Informatics for allowing me to use their resources. I also appreciate all my friends and family that have given me great support in more ways than I can count. Thank you all.

Contents

LIST OF TABLES

LIST OF FIGURES

Figure B(i): Rosetta's HABP2 decoy

Figure B(ii): QUARK's HABP2 model

Figure B(iii): I-TASSER's HABP2 model

Figure C: Knowledge based energy plot for Rosetta's model S_00004882

Figure D(i): Rosetta's MSP3b Decoy

Figure D(ii): QUARK's MSP3b model

Figure D(iii): I-TASSER's MSP3b model

Figure E(i): Rosetta's MSP6BC S_00000190 decoy

Figure E(ii): Rosetta's MSP6BC S_00004658decoy

Figure E(iii) : QUARK's MSP6BC model

Figure (iv): I-TASSER's MSP6BC model

Figure F(i): Rosetta's MSP6D decoy

Figure F(ii): QUARK's MSP6D's model

Figure F(iii): I-TASSER's MSP6D model

Figure G(i): MSP6BC

Figure G(ii): MSP6D

H(i): Knowledge based energy plots for Rosetta's MSP3 K1 model S_00000467

H(ii):S_00006919

Figure I(i): Rosetta's MSP3 K1 indel Fragment decoy

Figure I(ii): QUARK's MSP3 K1indel fragment model

Figure I(iii): I-TASSER's MSP3 K1indel Fragment model

Figure J: Knowledge based energy plot of Rosetta's MSP6 3D7 model S_00009724_0001.

Figure:K(i) Rosetta's MSP6 3D7 indel fragment decoy

Figure K(ii): QUARK's MSP6 3D7 indel fragment model

Figure K(iii): I-TASSER's MSP6 3D7indel fragment model

Figure L(i):Rosetta's MSP6 K1 indel fragment decoy

Figure L(ii):QUARK's MSP6 K1 indel fragment model

Figure L(iii): I-TASSER's MSP6 K1 indel fragment model

Figure M: Residue coverage in Pocket 1 in 70aa fragment

Figure N: 3D structure of MSP6D showing the pocket identified (spheres)

Figure O: Residues forming the MSP6D pocket highlighted with green

Figure P: 3D structure of MSP3 K1 indel fragment showing pocket 2 (blue spheres)

Figure Q: Residues forming the MSP3 K1indel fragment pocket 2 highlighted in blue

Figure R: 3D structure of MSP6 K1 indel fragment showing pockets (green spheres)

Figure S(i): I-TASSER's MSP3 3D7 model

Figure S(ii):I-TASSER's MSP3 K1 model

Figure T(i): I-TASSER's MSP6 3D7 model

Figure T(ii): I-TASSER's MSP6 K1 model

LIST OF ABBREVIATIONS

MSP - Merozoite Surface Protein

RON – Rhoptry Neck Proteins

EBA – Erythrocyte Binding Antigen

PATH - Program for Appropriate Technology in Health

GLURP - Glutamate Rich Protein

AMA1 - Apical Membrane Antigen 1

ADCI - Antibody Dependent Cellular Inhibition

RAMA -Rhoptry-Associated Membrane Antigen

ACTs -Artemisinin-based Combination Therapies

NCBI - National Center for Biotechnology Information

MMV - Medicines for Malaria Venture

PDB – Protein Data Bank

CASP -Critical Assessment of Techniques for Protein Structure Prediction

SCWRL - Side Chain With Rotamer Library

HMM – Hidden Markov Model

HABP – High Activity Binding Peptide

NMR – Nuclear Magnetic Resonance

RH - reticulocyte-binding-like protein homologue

RMSD – Root Mean Square Deviation

MSAs – Multiple Sequence Alignments

PSI-BLAST – Position Specific Iterative - Basic Local Alignment Search Tool

1. INTRODUCTION

Malaria is one of the leading causes of child mortality in Africa. According to the *World Health Organization report 2011,* there was an estimated 655,000 deaths in 2010 in Africa most of which have been reported among children [1]. This statistic implies that malaria is a threat to human populations. It is therefore important that a lasting solution is found to this killer disease. Previous studies have shown that *Plasmodium falciparum* is the most prevalent and is responsible for malaria associated mortality [2]. To this end, drugs have been designed that target different stages of the *P. falciparum* life cycle. However, the biggest challenge remains the development of resistance to these drugs leading to the need to develop more efficient ways of dealing with the parasite.

There has been increased funding towards malaria vaccine development. A report by Program for Appropriate Technology in Health (PATH, 2011) indicates that funding towards malaria Research & Development has increased from $121 million in 1993 to $612 million in 2009 with an accelerated increase since 2004 [3]. The increase in funding creates great opportunities for researchers to explore different avenues and techniques through which to curb the malaria menace. One of the ways to achieve this would be to develop drugs that can interfere with the invasive stage of the malaria parasite to ensure that it does not successfully enter the host.

2. LITERATURE REVIEW

2.1 Malaria Parasite Invasion of Human Erythrocyte

The malaria parasite is a member of the phylum Apicomplexa. Members of the apicomplexa are characterized by the apical complex [4] found on the parasite during the merozoite stage of the parasite's lifecycle. Human malaria infection happens when *Plasmodium* sporozoites are transmitted from a female anopheles mosquito during a bloodmeal. Once a female anopheles mosquito is infected with the malaria parasite, the sporozoites travel to the salivary gland. During a bloodmeal, the sporozoites migrate into the host entering the bloodstream and then move to invade the hepatocytes. The sporozoites differentiate and divide mitotically into thousands of merozoites [5], which eventually invade the erythrocytes initiating the blood asexual stage of the *Plasmodium* parasite.

The erythrocyte entry stage begins by an initial weak attachment of the merozoite to the erythrocyte through some interaction between the parasite's receptors and the red blood cell ligands [6]. This is then followed by the merozoite reorienting itself in such a way that the apical complex is attached to the host membrane. The merozoite reorientation is facilitated by two proteins families, the Erythrocyte Binding Like and the Reticulocyte Binding Like proteins which bring the parasite to a close apposition to the red blood cell membrane [6]. A tight junction is formed between the merozoite and the erythrocyte. The junction, referred to as the moving junction, moves inwards as the merozoite invaginates the erythrocyte [7]. During the invagination process, the merozoite forms a vacuole around itself as a protection mechanism from the host's cytoplasm. Through a complex series of events, the junction moves from the apical to the posterior end of the merozoite [7]. The events are powered by the actin-myosin motor [8]. The moving junction consists of components that are secreted from the secretory

organelles known as the rhoptries [9][10]. Studies conducted on the parasite *Toxoplasma gondii*, a member of the apicomplexa, have shown that four rhoptry neck proteins (RON2, RON4, RON5 and RON8) are secreted and targeted towards the erythrocyte membrane. The moving junction complex was also found to consist of the apical membrane antigen (AMA) 1 which was shown to be secreted by other secretory organelles known as the micronemes [11]. Lamarque et al. (2011), were able to demonstrate that there is an association between the rhoptry neck proteins and the microneme secreted AMA1 protein during the invasion stage of both *P. falciparum* and *Toxoplasma gondii* [12].

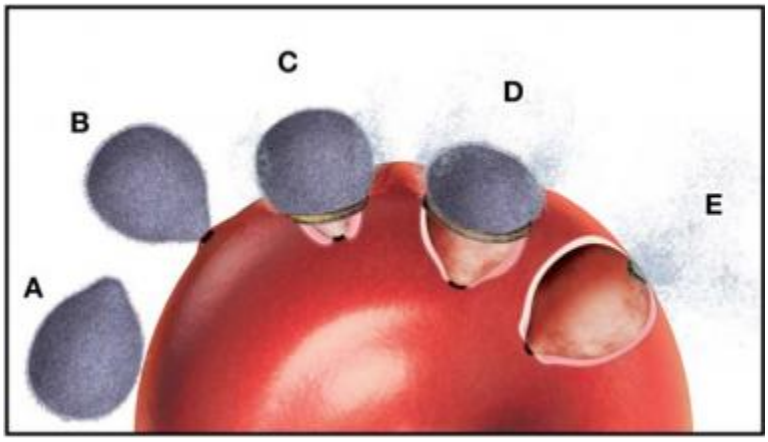


Figure 1.1 Merozoite invasion of Erythrocytes (Cowman et al., 2006)

Erythrocyte invasion is a process that is characterized by a series of molecular interactions between the merozoite, the invading bloodstage parasite, and the host membrane [13] (Figure 1). These are ligand-receptor interactions. A review by Cowman et al. (2006), classifies merozoite proteins in *P. falciparum* into five categories. They include the GPI anchored surface proteins such as Merozoite Surface Proteins (MSP) e.g. MSP-1, MSP-2, MSP-10, microneme proteins e.g. AMA1, Erythrocyte Binding Antigen (EBA) 140, EBA 175, EBA 181, peripheral surface proteins e.g. Glutamate Rich Protein (GLURP), MSP3, MSP6, rhoptry e.g. Rhoptry-Associated

Membrane Antigen (RAMA) and the rhoptry neck proteins e.g. reticulocyte-binding-like protein homologues Rh1, Rh4 [5]. The discovery of the functions of some of these proteins has given scientists insight on the possible target areas in the effort to curb malaria.

From the peripheral surface proteins category, MSP3 and MSP6, which belong to the same family, have been chosen as candidates for malaria vaccines. Both of these antigens have been identified as targets of Antibody Dependent Cellular Inhibition (ADCI) [14]. This is a phenomenon whereby individuals who have been exposed to the malaria parasite over time have developed some form of immunity known as premunition.

The immune system works in such a way that once the body recognizes an antigen, it produces B cells, which are part of adaptive immunity, to produce antibodies that can neutralize the antigens [15]. The body also has immune memory that allows it to recognize an antigen it has previously encountered. When it does, the memory cell elicits an immune response that destroys the antigen before it can cause an infection [15]. Premunition is therefore part of the human body's adaptive immunity that works in cooperation with innate immunity, and particularly the monocyte cells, to protect an individual who has previously been infected by the malaria parasite from getting reinfected.

The ADCI effect is effective against some *Plasmodium falciparum* invasion proteins such as MSP3 and MSP6 making them potential candidates for the malaria vaccine. A vaccine against these two antigens would stimulate the immune system so as to protect individuals who have not acquired premunition. Moreover, a vaccine against MSP3 and MSP6 would also enhance the inhibitory effect different from what is observed with having premunition only. An individual who is exposed to the malaria parasite may have better inhibition towards it if they have been given a vaccine. Several studies conducted in Asian and African settings showed that there was

a strong association between anti-MSP3 antibodies and acquired clinical protection against malaria [16][17][18][18]. For example, parasite killing was observed in an experiment where human recombinant anti MSP3 antibodies in cooperation with monocytes were used [15]. This is a good indication that there is some great potential to develop an effective malaria vaccine.

An antigenic analysis performed by Singh et al., (2009), demonstrates that irrespective of the differences found in the MSP3 family members, they observed that the antigenic properties are conserved to generate cross reactive antibodies [16]. As an example, a study conducted by Singh et al., 2005, established that MSP6 has some epitopes that are cross reactive with MSP3. This means that an antibody designed for one member of the MSP3 family could have the same anti-parasite inhibition effect observed in a different member of the same family. This cross reactive property would mean higher efficacy of vaccines developed against the malaria parasite.

Studying the properties of these two proteins would give us a better understanding of how they function and in effect lead to the design of effective malaria vaccines. One of the ways to accomplish this is by deciphering the structure of the protein.

2.2 Protein Structure Prediction

Protein structure prediction is the process of finding the three dimensional (3D) structure of a protein. Proteins are very important molecules in any given organism. Virtually every cell function is controlled by a protein. There are proteins for structural support, bodily movement, defense mechanism among other functions. A protein is a properly folded peptide sequence. It is the folding of a protein into its 3D structure that determines its function. The amino acid sequence forms what is known as the protein primary structure whereas the 3D structure is referred to as the tertiary structure. Studying the structure of proteins gives insight into how they function, how different processes in an organism are controlled, how the body fights against

foreign bodies among other things. One could also study the structure of proteins in parasites, bacteria among other organisms that cause disease in humans, plants and animals to understand their interaction with their hosts and how best to prevent or cure the diseases they cause. There are different ways of obtaining a protein structure, the experimental techniques such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and electron microscopy. Experimental techniques have a common disadvantage of being expensive and time consuming [20]. Other methods include comparative modeling, fold recognition or threading, and *ab initio* prediction. Unlike the experimental methods, these are computational methods that are faster and less labor intensive. The sheer number of unknown structure of proteins has led to a lot of work being put into improving the computational techniques.

2.2.1 Threading (Fold Recognition)

This technique searches through a protein templates database using a query sequence of the protein with unknown structure, trying to find a match from which the structure can be predicted [20]. This method was developed after it was found that two proteins had the same structure despite having minimal similarity in their protein sequences [21]. This method has two weaknesses. One, the impreciseness of the energy functions leading to an inability to determine any given conformation. Secondly, recognizing the conformation still remains a challenge since so far there is no direct computational method to do it [20]. A scoring function is used to rate the results of a search and the best template is used to build the model.

2.2.2 Ab initio *Modeling*

This is a de novo structure prediction method meaning it starts with no prior structures of the protein under study. The goal of this technique is to find the native structure of a protein. This is done by performing a conformational search by generating a pool of structure decoys from which

the final model is selected [22]. Selection of the model that is closest to the native state is done by identifying the decoy with the energy function that corresponds to the most stable thermodynamic state [22]. The principle behind use of thermodynamic stability is that proteins in their native state are at their lowest free energy. There are two major types of energy functions: physics and statistical based. Physics based or the true effective energy function, analyses the forces between particles [23]. The statistical or knowledge based energy function utilizes statistical observations [24] or statistical information of solved structures that have been stored in the Protein Data Bank (PDB) [22]. Simulations to the biological folding process and conformational changes are done while observing the energy changes. Although *ab initio* has the strength of not being dependent on the number of available templates like homology modeling and threading methods, it has two main disadvantages. One is the inaccuracy of energy functions and two, the immense number of possible conformations [21]. Practice has also shown that this method is not able to handle large proteins and currently has an upper limit of approximately 200-250 residues [25].

2.2.3 Homology Modeling

Comparative (homology) modeling is a modeling method that searches through a database to find a known structure of a close homolog to the sequence in question [21]. The idea behind this technique is the fact that proteins with similar function, and especially the ones that have some evolutionary relationship, have similar sequences which tend to adopt similar structural conformations [26]. Sequence similarity is therefore a key aspect of this method. It is important to note that the higher the level of similarity between the sequence with a known template and the sequence with unknown structure, the higher the level of accuracy of the results found. Homology modeling process involves doing a fold assignment template selection, target-

template alignment, model building and model evaluation. Of the three computational methods, comparative modeling has been found to be the most accurate prediction method [27]. There are many important applications of protein structures including drug design in medicine. For example, traditionally, drug discovery process was a multi-step time consuming process that required *in vivo* biological screens and additional investigation of the pharmacokinetic properties, metabolism and potential toxicity [1]. However, there has been a remarkable improvement in technology which has seen the introduction of *in silico* drug design.

In silico structure prediction methods have made it possible to understand protein function and conduct experiments on proteins in a much simpler way using less time. With the available structure prediction tools we can now model proteins such as the MSP 3 and 6 to help define important structural regions in the proteins that could be potential vaccine targets to inhibit the protein's function.

3. RESEARCH QUESTION

Can we determine the 3D protein structure of MSP3 and MSP6 to examine structural differences in the two main allelic variants?

4. RESEARCH OBJECTIVES

- To computationally determine the 3D structures of the Merozoite Surface Protein 3 and 6.
- To map MSP3 and MSP6 allelic variants from a malaria endemic population to determine whether amino acid changes alter protein conformation
- To identify possible binding pockets on the modeled structures

5. JUSTIFICATION

Malaria is a real threat to the human population. According to the World Health Organization, about 3.3 billion people, which is approximately half of the world's population, are at risk of malaria. It was estimated that in the year 2010, there were 216 million cases of malaria [28]. In 2009, 765 million people in sub Saharan Africa alone were estimated to be at risk of malaria infection [3]. These statistics are a clear indication that a lasting solution needs to be found.

There has been an accelerated increase in funding towards malaria Research & Development since 2004 which allows for scientists to research further and find more effective ways to prevent and treat malaria. For example, we have Artemisinin-based Combination Therapies (ACTs) which became available in the 1990's have been used for malaria treatment replacing some of the drugs that had become ineffective as a result of resistance [29]. Partnerships such as Medicines for Malaria Venture (MMV) have also been put in place to design malaria treatments [29]. Despite the existing breakthroughs in malaria research, there is room for more work to be done and better solutions to be found.

One of the ways of creating better vaccines and treatment for any disease is to understand the function of the proteins of the parasites causing the disease. Today, it has become very important to study the 3D structure of proteins in order to understand their functions. There are several proteins involved during the invasion of erythrocytes by the malaria parasite some of which their 3D structures have been found and led to better understanding of the invasion process. This project proposes to add to the knowledge of the invasion process of the malaria parasite by modeling the 3D structure of the MSP3 and MSP6 proteins that are also involved in the invasion process. The 3D structures of these two proteins will go a long way to inform the process of malaria vaccine design. Discovering the similarities or differences of these two antigens that

belong to the same family will enable researchers to understand their function and in turn target them appropriately.

6. METHODOLOGY

6.1 Target Sequence Retrieval

Before any modeling work can be done, one has to verify that the structures have not been modeled before. This is accomplished by doing a search through the PDB, a repository for all existing protein structures. Development of the 3D structure of the protein in question then begins with identifying the sequence of that particular protein. This can be done by searching for it in the existing biological databases for example the National Center for Biotechnology Information (NCBI). The protein sequence can then be used to do a template search to find the most similar template to the protein in question.

6.2 Template Search

For homology modeling, a template with high levels of similarity is preferred for good results to be obtained. Similarity is measured in terms of percentage identity between the target and the templates found. The protein sequences of our two antigens were therefore used to find similar templates. Sequence alignments are then performed to find the templates with high sequence identity and similarity. A sequence identity of greater than 50% has been shown to produce accurate results of up to ~1Angstrom at alpha carbon root mean square deviation (rmsd) from the experimental structure [30]. Alignments with 30% sequence identity have been shown to give near optimal results for their targets. Below this threshold, alignment quality decreases sharply which could even result in misalignment when the sequence identity is less than 20% [30]. If good templates are not found, other methods such as threading and *ab initio* modeling should be employed.

Searching for proteins with evolutionary relationships requires the use of advanced techniques that can look beyond the plain protein sequence. This has led to the introduction of search methods that use sequence profiles for both the query and the databases. Sequence profiles are created from Multiple Sequence Alignments (MSAs) of related sequences where position specific substitution scores [31] are computed to indicate the probabilities of observing each of the 20 amino acids at any given position in the sequence [32]. Searches that use HMM-HMM and Profile-Profile have been found to be the most sensitive [31]. Although sequence profile based searches e.g. PSI-BLAST are sensitive, they are also very slow.

6.3 Analysis of genes under study

6.3.1 MSP3

Various studies have been done on MSP3, also known as secreted polymorphic antigen associated with the merozoite (SPAM) [33], in an effort to identify regions that would be key in informing the design of vaccines as well as possible drug sites. These are the regions that were identified and modeled. A study conducted by Quevray et al. (1994) identified a 27 amino acid region, referred to as MSP3b, which they found to be a target for naturally occurring antibodies and inhibited *in vitro* growth in cooperation with monocytes [34]. Another study conducted by Singh et al. (2004) that used 6 overlapping MSP3 peptides found that 3 out of the 6 peptides had a major inhibitory effect on the growth of the malaria parasite. They found a 70 amino acid region that was a target of antibodies and which they suggested should be part of a malaria vaccine construct [18]. Other regions found on MPS3 include three segments identified by Rodriguez et al. (2005) which they characterized as High Activity Binding Peptides (HABPs). These were found to inhibit *in vitro* invasion of the erythrocyte by the merozoite. One of the

regions was in the N terminal, another in the middle while the other was found to be located in the C terminal end.

Table1: MSP3 Sequence indicating fragments modeled

>gi|23495212|gb|AAN35542.1| merozoite surface protein 3 [Plasmodium falciparum 3D7]
MKSFINITLSLFLLHLYIYINNVASKEIVKKYNLNLRNAILNNNSQIENEENVNTTITGNDFSGGEFLWP
GYTEELKAKKASEDAEKAANDAENASKEAEEAAKEAVNLKESDKSYTKAKEACTAASKAKKAVETALKAK
DDAEKSSKADSISTKTKEYAEKAKNAYEKAKNAYQKANQAVLK***A*KEASSYDYI***LGWEFGGGVPEHKKEE
MLSHLYVSSKDKENISKENDDVLDEKEEEAEETEEEELEEKNEEETESEISEDEEEEEEEEEKEEENDKK
KEQEKEQSNENNDQKKDMEAQNLISKNQNNNEKNVKEAAESIMKTLAGLIKGNNQIDSTLKDLVEELSKY
FKNH

***Key**: Red – MSP3b target of antibodies (identified by Quevray et al., 1994)*

A++Red+Green - 70aa target of antibodies (identified by Singh et al., 2004)*

Blue + Italicized Red – 3HABPs (identified by Rodriguez et al., 2005) ordered according to appearance on the table.

6.3.2 MSP6

MSP6 has been shown to be homologous to MSP3 by having 50% identity and 85% similarity [33]. Through a study conducted to establish whether the two antigens also share the property of being targets of naturally occurring antibodies, it was found that MSP6 has some epitopes that are cross reactive with MSP3 [14]. These are peptides B, D, E and F as indicated in Table 2 below. Peptides B and F were found to be fully cross reactive in MSP3 and MSP6 whereas D and E were partially cross-reactive.

Table2: MSP6 Sequence indicating peptides identified by Singh et al., 2005

>gi|23495213|gb|AAN35543.1| merozoite surface protein 6 [Plasmodium falciparum 3D7]
MNKIYNITFLFILLNLYINENNFIRNELINEKNHNLRNGSMYNNDKILSKNEVDTNIESNENSIHESGHK
IDGEEVLKANVDDITYKKKNVDDSEIPFSGYDIQATYQFPSTSGGNNVIPLPIKQSGENQYTVTSISGIQ
KGANGLTGATENITQVVQANSETNKNPTSHSNSTTTSLNNNILGWEFGGGAPQNGAAEDKKTEYLLEQIK
IPSWDRNNIPDENEQVIEDPQEDNKDEDEDEETETENLETEDDNNEEIEENEEDDIDEESVEEKEEEEEK
KEEEEKKEEKKEEKKPDNEITNEVKEEQKYSSPSDINAQNLISNKNKKNDETKKTAENIVKTLVGLFNEK
NEIDSTINNLVQEMIHLFSNN

Key: *Red – Peptide B*
Light Blue – Peptide E
Blue + Orange – Peptide D *Green - Peptide F*
Red with Yellow Highlight + Blue – Peptide C

It is important to note is that our HABP2 from MSP3 was extended on the right side to make it 30 residues long since the fragment server could only generate fragments of proteins longer than 30aa. Also, in the case of MSP6 peptide C was included due to the fact that peptide B was much shorter and therefore had to be extended to have a fragment that the fragment server would use to generate the fragment files.

6.3.3 Exploring Allelic Differences

In this project, we sought to explore the differences between the two alleles, chloroquine-sensitive, 3D7 and chloroquine resistant, K1 strains. One main characteristic difference between the two is the high number of inserts in the K1 strain. The MSP3 K1 variant has a length of 379aa (Accession No.AAC47831.1), while the 3D7 has 354aa (AAN35542.1). In the case of MSP6, K1 has 414aa (ACR10029.1) whereas 3D7 has 371 (AAN35543.1) residues. The structural differences were therefore investigated. Due to the fact that *ab initio* modeling was being used, the antigens were divided into shorter manageable segments whose structures were then determined and their differences highlighted. The regions chosen for exploration of allelic differences in both MSP3&6 were the areas dominated by deletions/insertions. The diagrams below (*figure 2&3*) show the alignments of the two alleles for both MSP3 and MSP6. The alignment was performed using ClustalW2 alignment tool.

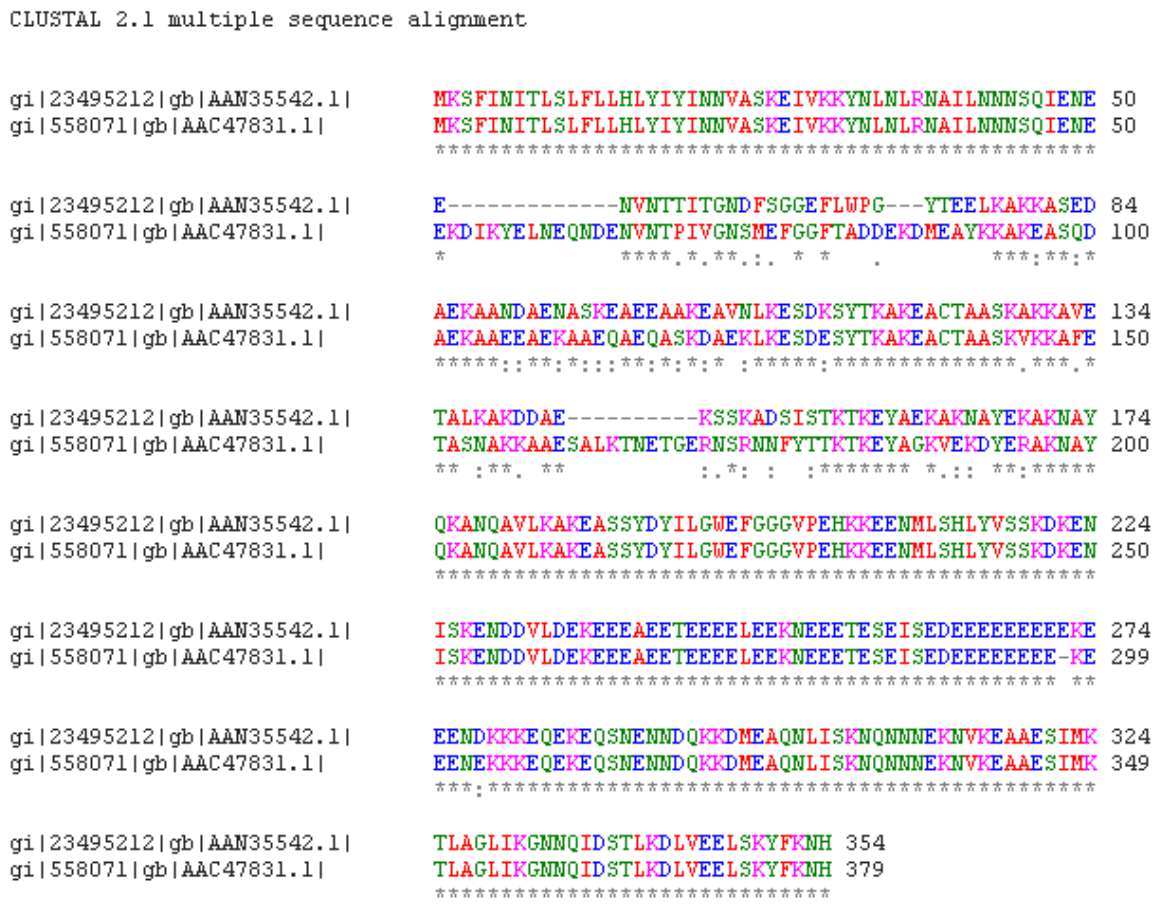


Figure 2 – MSP3, 3D7(AAN35542.1) vs K1 (AAC47831.1) Alignment

In the case of MSP3 the fragment selected for exploration of allelic differences began from residues **NNVTTI**, the residues after the first gap in the 3D7 strain, and ran all the way to residues **TALKA**, the residue right before the last gap. The K1 strain ran from residues **KDIKYE** to **SALKT**. The reasoning behind choosing this section is that it displayed the major differences between the two strains as demonstrated in the figure 5 above.

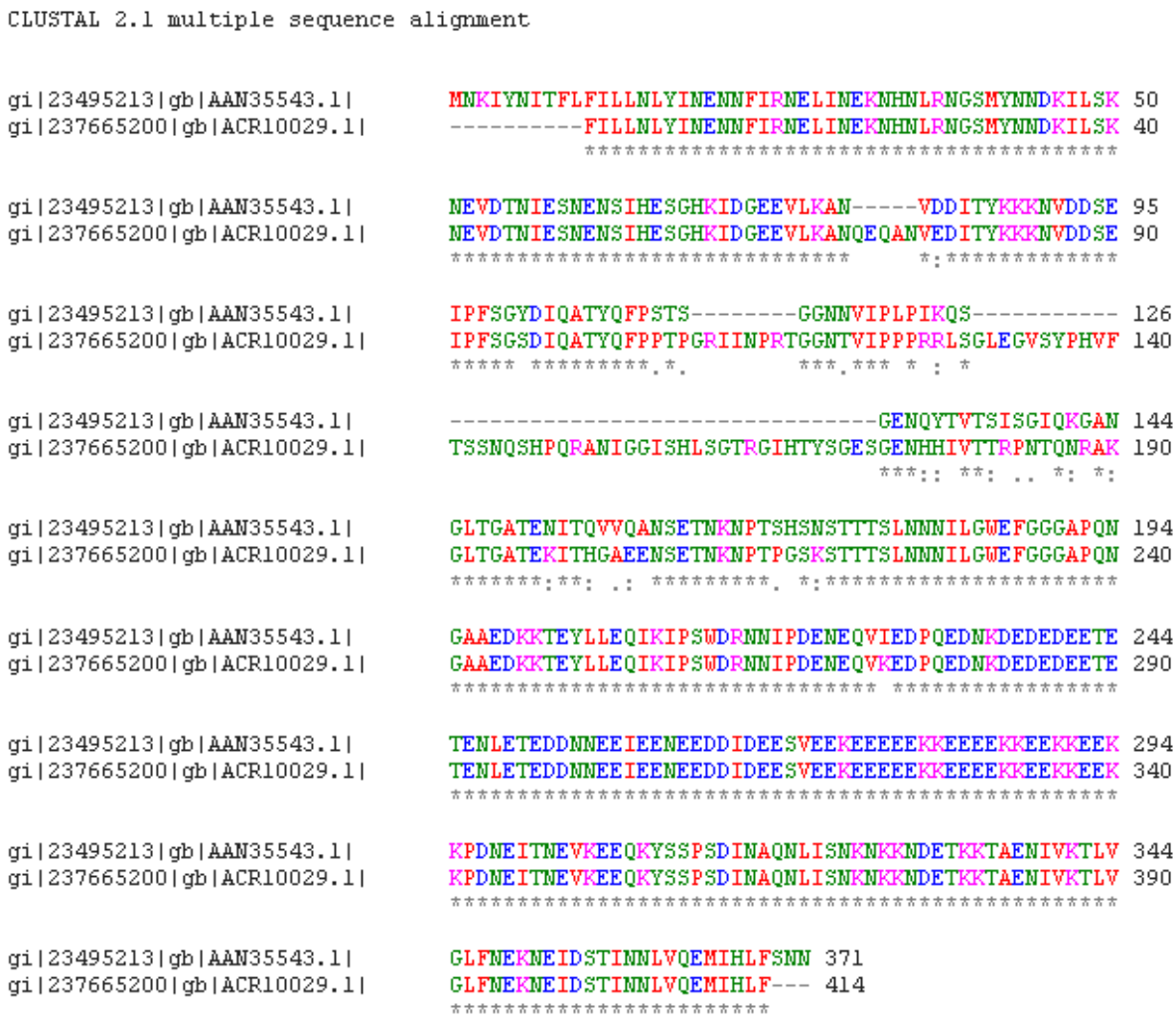


Figure 3: MSP6, 3D7(AAN35543.1) vs K1(ACR10029.1) Alignment

The section chosen in MSP6's 3D7 strain for the purpose of exploration of differences ran from **NVDDI** to **SGEN**, the residues before and after the insertion/deletion. As for the K1 strain, our fragment began from **NNQEA** to **ESGEN**. There was also fragment to cover the insert/deletion at the beginning of the protein from **MNKI** to **NIESEN** in the case of 3D7 as well as from **FILL** to **NIESEN** in the case of K1 strain.

6.4 Modeling of the Protein Fragments

Three tools, QUARK, ROSETTA and I-TASSER, were used so as to provide some comparison of the models. QUARK and ROSETTA have been reported [33][34] to perform well in the CASP experiments under the category of *ab initio* modeling. I-TASSER is a server based on the principle of threading. It was chosen for this study since it has also been reported [37] to be one of the best servers in its category. Since *ab initio* has only been known to perform well on small protein fragments [25], a decision was made to fragment the proteins into smaller peptides. The criterion used to do this was that some research was done to find out any interesting regions such as those that have shown to be targets of antibodies.

To model the protein fragments using Rosetta, the fasta sequences were first prepared by extracting from the longer MSP3 and MSP6 sequences. From there, the fragment library files for each of the fragments were generated. Rosetta uses 3mer and 9mer fragment files to begin the protein assembly process. A secondary structure prediction file was also generated to be used by Rosetta. Before the job could be run, several parameters were set including the number of decoys (the term used to refer to Rosetta models), that one needs to be generated. For this study, we generated 10,000 decoys. This was a compromise between the number of recommended models and the available resources which included the computing power and the time available. It is recommended to generate 20,000 models for proteins of sizes of up to 150 residues. Therefore, the 10,000 decoys generated for the fragments in this project was found to be a good number given that most of the fragments were less than 100 residues long. Running one job would take a minimum of 20 days and this was because the fragment sizes were much smaller than the full sequence. The decision to model the smaller fragments also meant that it would be possible to get more accurate results since *ab initio* modeling works best for small proteins.

From here, the generated decoys are run through a clustering algorithm. The goal for performing clustering on the large number of decoys is that we would like to find the decoy with the least amount of energy in the largest cluster. The principle behind selecting the decoy in the largest cluster is that a conformation that is sampled many times over others tends to be the correct one. Since the goal was to establish the decoy that is as close as possible to the native, the model with the least amount of free energy was chosen. Rosetta generates a silent file where the coordinates of the generated models are written. The extract algorithm was then executed to retrieve the models of choice. From here, the relax algorithm was ran refine the selected models as well as pack the side-chains.

Visual inspections as well as quality assessment tests were conducted on the best performing decoys in the top 5-10 clusters in order to pick the one that was most favorable and had minimal steric clashes.

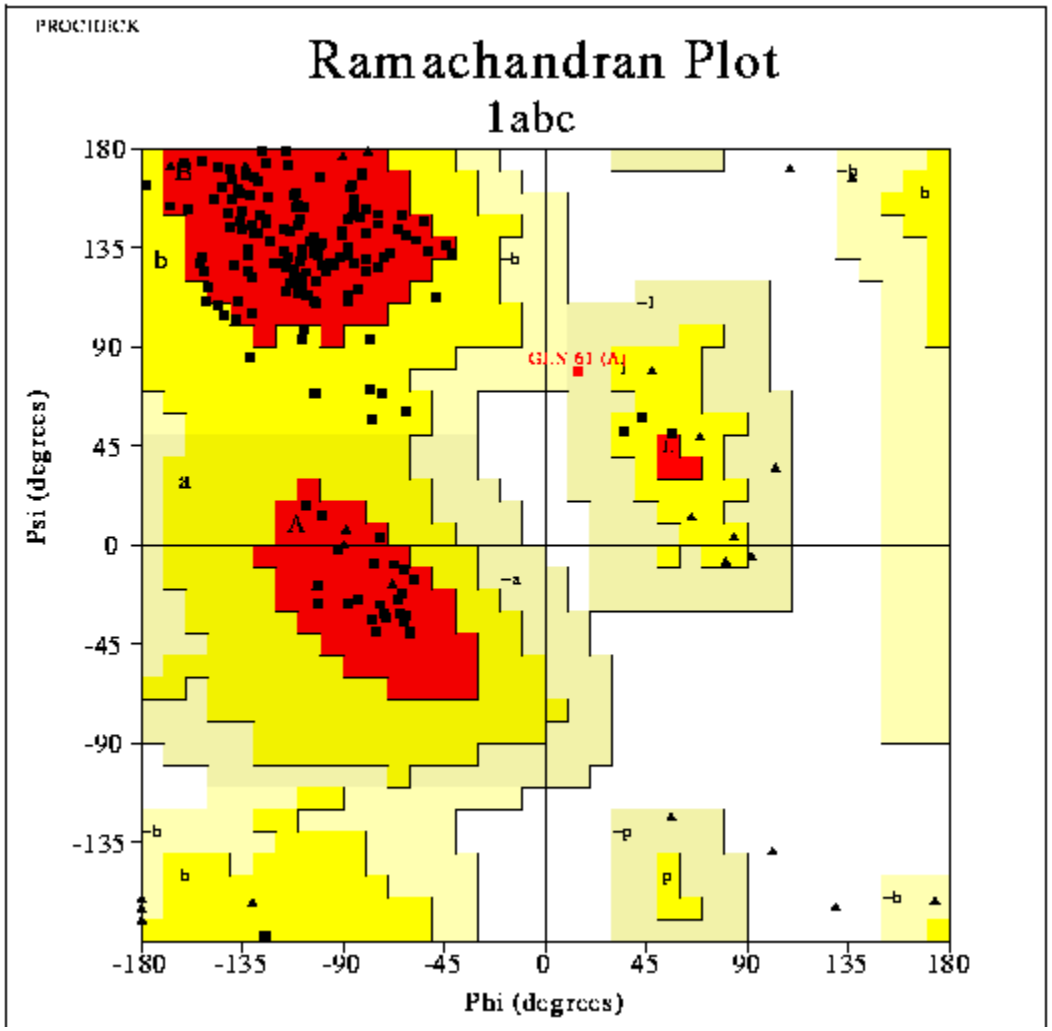
Since *ab initio* does not use templates which can tell how accurate the models found are, the results were compared with those models generated by two highly rated automated webservers, QUARK and I-TASSER.

6.6 Model Refinement and Visualization

After all the above steps have been implemented, it is always important to go back and do some structural refinement to ensure that the model we have obtained is accurate. Methods employing molecular dynamics have been found to be more promising as they use conformational space of the sequence to find the most favored arrangement. However, these methods require high computational power. Calculating the position of atoms as a probability density function has been found to be a faster method [26].

Rosetta uses the relax algorithm which attempts to wiggle the atoms in to a low energy state so as to get the conformation with the minimum energy as part of the refinement process.

Validation of the models was done using tools such as PROCHECK, ERRAT and ProSA-web. PROCHECK performs a check on the stereochemical properties of a protein categorizing the residues to either falling in the allowed or disallowed regions on the ramachandran plot.



Key:

A	- Core alpha	**L**	- Core left-handed alpha
a	- Allowed alpha	**l**	- Allowed left-handed alpha
~a	- Generous alpha	**~l**	- Generous left-handed alpha
B	- Core beta	**p**	- Allowed epsilon
b	- Allowed beta	**~p**	- Generous epsilon
~b	- Generous beta		

Figure 4 : Sample Ramachandran Plot borrowed from www.ebi.ac.uk

The diagram above, (figure 4), shows the different regions on the plot. The ramachandran plot shows the phi-psi torsion angles for all residues in the protein structure. Models were qualified as good if over 90% of the residues fell in the allowed regions. These allowed regions represented the most favorable combinations of phi-psi values [38]. Glycine residues were represented using triangles to distinguish them from the other residues as it is the only amino acid that does not have a side-chain. For this reason, it is not restricted to the regions in our plot. The generous areas have been reported to be in the disallowed region [38] and should therefore be investigated further. PROCHECK was used to assess the unusualness of properties such as phi-psi distribution, main chain covalent forces among others, using a measurement known as G-Factor. G-Factor values less than -0.5 were considered unusual whereas those less than -1.0 were considered highly unusual [38].

PROSA is a tool that evaluates correctness of protein models. However, Pawlowski et al., 2008, recognized that PROSA can be very strict since it detects very minor errors [39] and would therefore be more suitable to validate homology models that have a higher accuracy. ProSA-web was however used to determine the Z-score of the models in relation to the existing database of experimental and NMR structures.

ERRAT is another tool used in the verification of protein structures. It is based on the premise that geometric and energetic effects lead to the nonrandom distribution of different atom types with respect to each other. It therefore uses statistics of pairwise atomic interactions to identify erroneous regions in protein structures [40]. High resolution structures have been found to attain 95% and above in the overall quality factor whereas those with lower resolution (2.5-3A) have an overall quality factor of 91% [41]. This attribute and the fact that it also gives a graph

indicating confidence levels with which to reject any given section of a structure made it a useful tool in this study.

Visual inspection is also important since it is during this that some important differences in the models may be recognized and further research conducted to establish the significance. The visualization tools used here were RASMOL and CHIMERA.

6.6 Pocket Identification

The final stage was pocket identification. This was done using Computed Atlas of Surface Topography of proteins (CASTp), a webserver that finds possible pockets and voids. It also gives a description of the atoms participating in the formation of the pockets [42]. CASTp uses solvent accessibility surface model as well as molecular surface model to calculate the area and volume of the pockets found. Besides doing pocket identification for analysis of our structures, we also used ProFunc to find special features that may be realized after folding of the protein.

7. RESULTS

7.1 Template Search

To find homologous templates, a search was done on the HHPRED and Position Specific Iterative – Basic Local Alignment Search (PSI-BLAST) servers. HHPRED server uses the HHpred search algorithm. The reason for choosing this server was because it uses hidden markov models (HMMs) which perform sensitive searches. It is important to note that the database used by the search algorithm for HHpred was pdb70, an alignment database built around sequences of known structures.

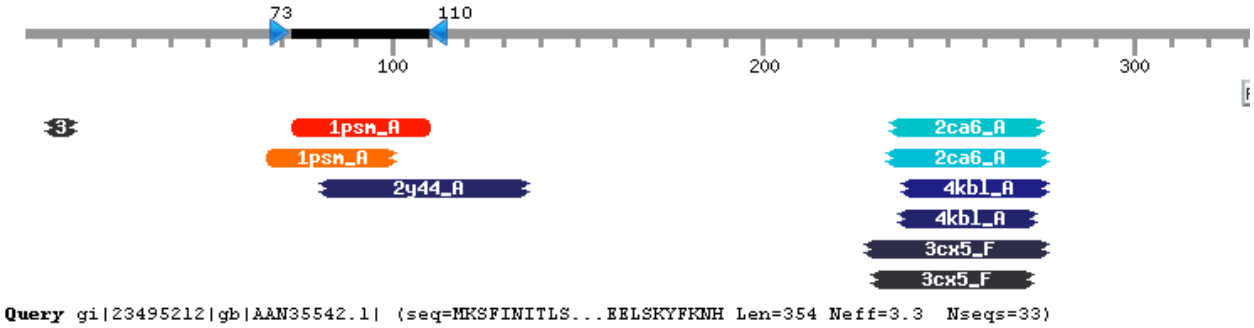


Figure 5: Results of HHpred search using MSP3. Red bars signify close hits. The further the color of the bar is from red, the poorer the hit

The proteins under study had very poor hits in the template search as shown in the *figure 2* above. The region with red gave the highest identity score of 79%. It was later discovered that the 3D structure of a 38 residue region (73- 110) had already been determined using NMR [43] and hence the high score. The other hits had very low identity scores as well as high E-values which meant that the templates found were not so good. Moreover, as it can be observed above, a very large proportion of the protein did not have any hits at all. The results of PSI-BLAST were not any better.

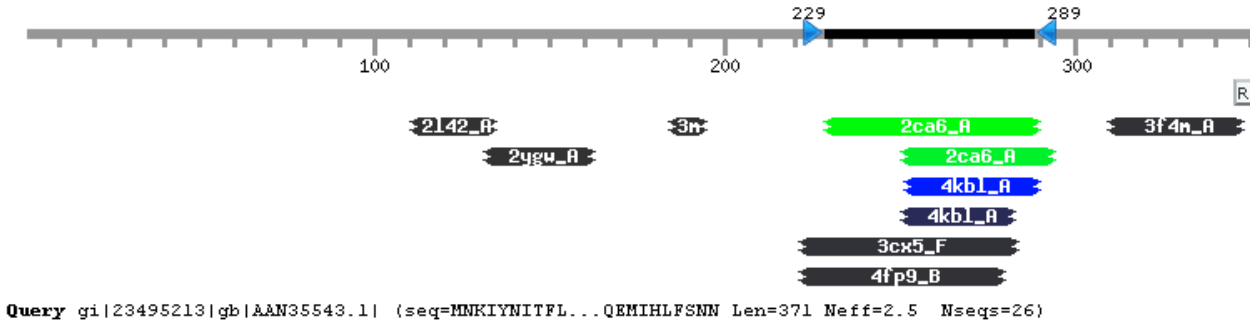


Figure 6: Results of HHpred search using MSP6. Red bars signify close hits. The further the color of the bar is from red, the poorer the hit

The search for homologous proteins for MSP6 did not yield good results either as illustrated by figure 3 above. The top hit, Ran Gtpase-Activating Protein 1(PDB code: 2c6_A), had an identity of 27% and an E-value (an E-value close to 1 signifies close hits) of 0.7. This can be confirmed with the alignment figure 4 shown below where there are hardly any matching regions.

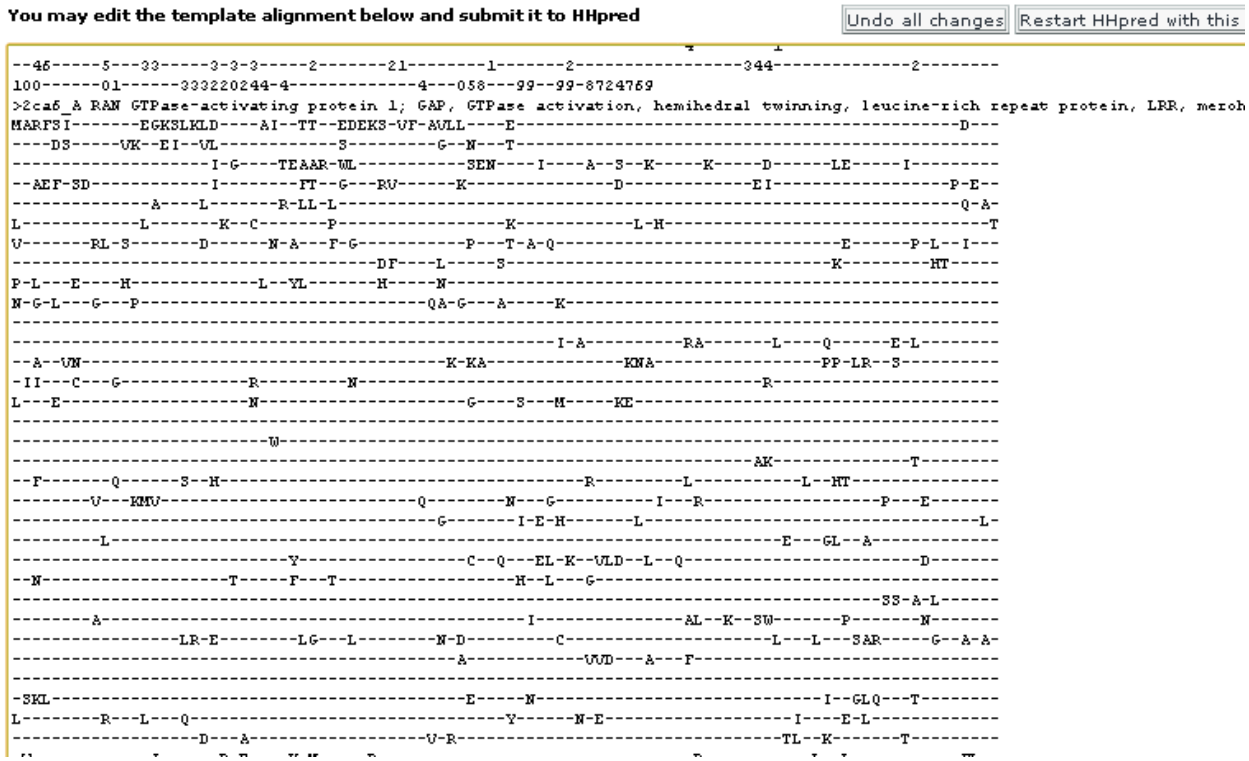


Figure 7: Top hit - Ran Gtpase-Activating Protein 1(PDB code: 2c6_A) alignment to MSP6

The poor results led us to using *ab initio* modeling as well as threading.

7.2 Benchmark Structures

Our benchmark structures were EBA175 (583aa) and EBA140 (595aa). These two proteins have been categorized as Duffy Binding Like proteins and have been found to be involved in the activation of the invasion process after merozoite reorientation [5]. The fact that these two proteins are also involved in the invasion process and have had their structures experimentally determined, with EBA175 having a resolution of 2.4Angstroms and EBA140 at 2.3Angstroms, made them good candidates for benchmark structures.

We conducted our evaluation using our test tools to see how the experimentally determined structures performed. This helped inform on the quality of the models found.

7.2.1 EBA175 and EBA140 PROCHECK Results

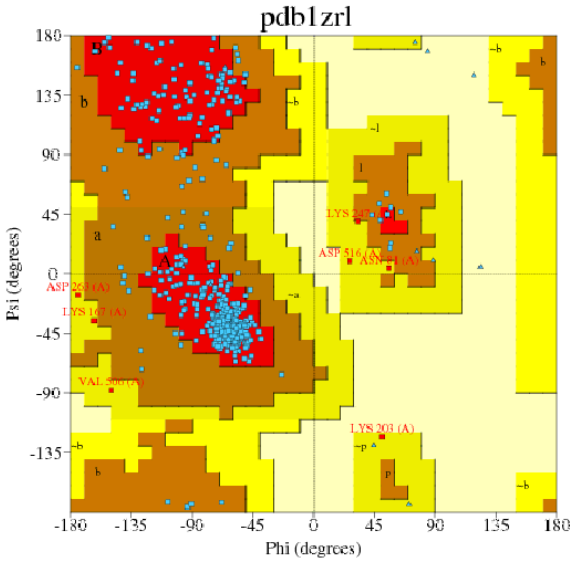


Figure 8: EBA175 ramachandran plot

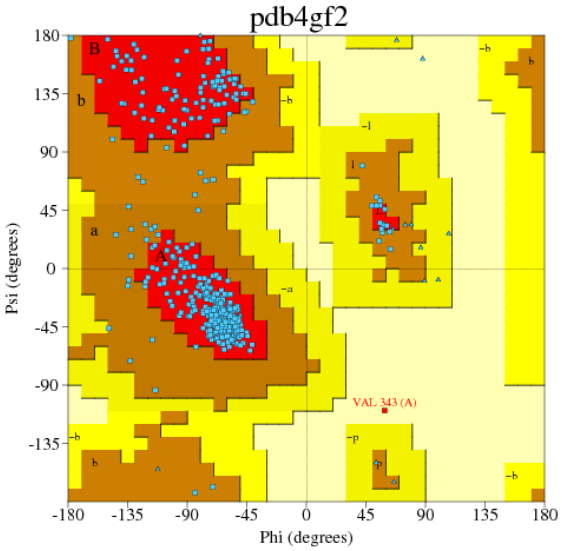


Figure 9: EBA140 ramachandran plot

The plots above show the distribution of EBA175 and EBA140 residues on the ramachandran plot. The red regions represent the most favored regions where most of the residues of a good structure should fall under.

Table 3: Ramachandran plot statistics for EBA175 and EBA140

Aspect	Protein	No. of Residues	Percentage(%)
Most favored regions	EBA175	**489**	**88.7**
	EBA140	**513**	**91.8**
Additionally Allowed regions	EBA175	**55**	**10**
	EBA140	**45**	**8.1**
Generously allowed regions	EBA175	**7**	**1.3**
	EBA140	**0**	**0**
Disallowed Region	EBA175	**0**	**0**
	EBA140	**1**	**0.2**
G-Factor	**Overall Average**		
EBA175	0.31		
EBA140	0.66		

Good structures are those that have over 90% of their residues in the favored regions

Table 3 above shows the scores obtained by EBA175 and EBA140 when run through PROCHECK. The table shows that EBA140 met the quality mark of 90% whereas EBA175 fell a little short scoring 88.7%. However, EBA140 had one of its residues falling in the disallowed regions.

PROCHECK was also used to check whether a model's residue had any unusual properties. This was graded using GFactor. Values below -0.5 were rated as unusual whereas those below -1.0 were said to be highly unusual. EBA 175 and EBA140 had GFactors of 0.31 and 0.34 respectively.

7.2.2 EBA175 and EBA140 ProSA-web Results

We submitted our benchmark structures EBA175 and EBA140 to ProSA-web and got the following plots. Both proteins had low z-scores of -9.33 and -9.38 values, respectively as shown

in figures 10 and 11 below, and depict the region of z scores for experimentally determined structures.

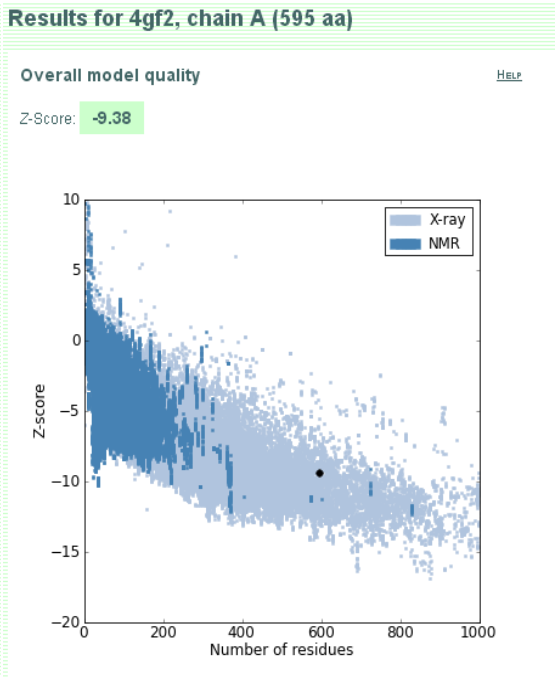


Figure 10 EBA175ProSA-web Output Figure 11: EBA140 ProSA-web Output
The black dots on the diagrams show the position of the protein in relation to all the structures in the PDB

Figures 12 and 13 below show knowledge based energy plots for EBA175 and EBA140. From the plots we see that good structures are characterized by having most of their knowledge based energies below 0. This was helpful in judging the quality of our models. However, transmembrane regions tend to create energy spikes as can be seen towards the end of the plot.

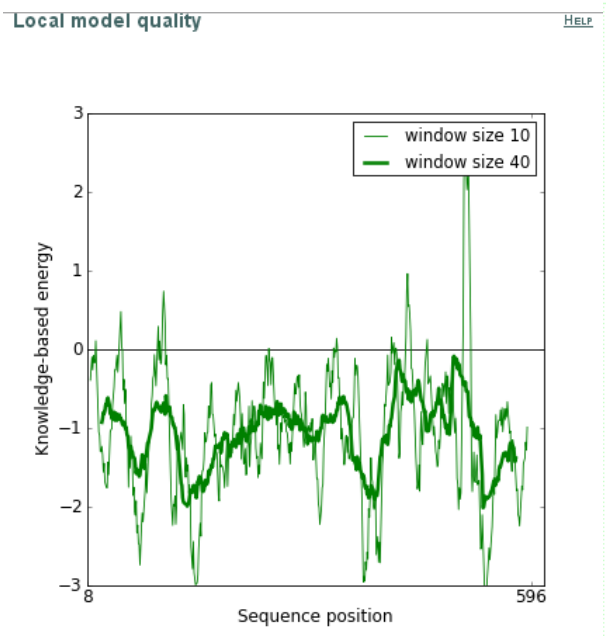


Figure 12:EBA175 Knowledge Based Energy Plot

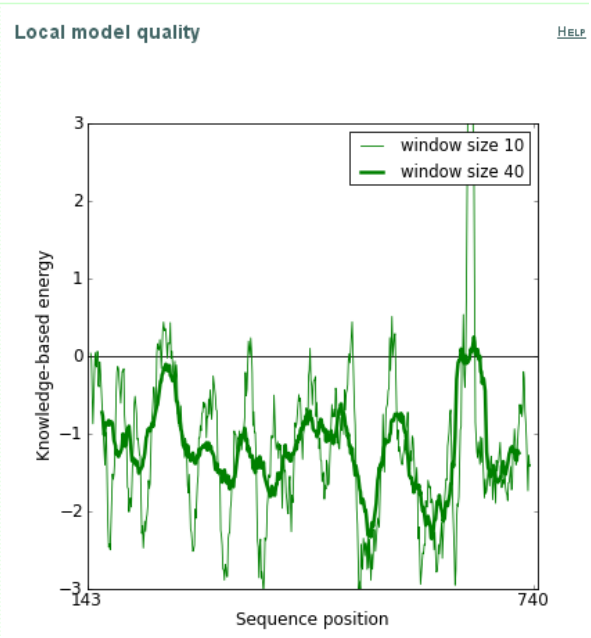


Figure 13:EBA140 Knowledge Based *Energy Plot*

The green lines, thin and thick, represent plots based on a window of size10 and 40, respectively.

We also submitted the two proteins to ERRAT server and the overall quality was graded at 95.812 for EBA175 and 96.724 for EBA140. Figures 14 and 15 below show the overall performance of the different sections of EBA175 and EBA140 X-ray structures on ERRAT.

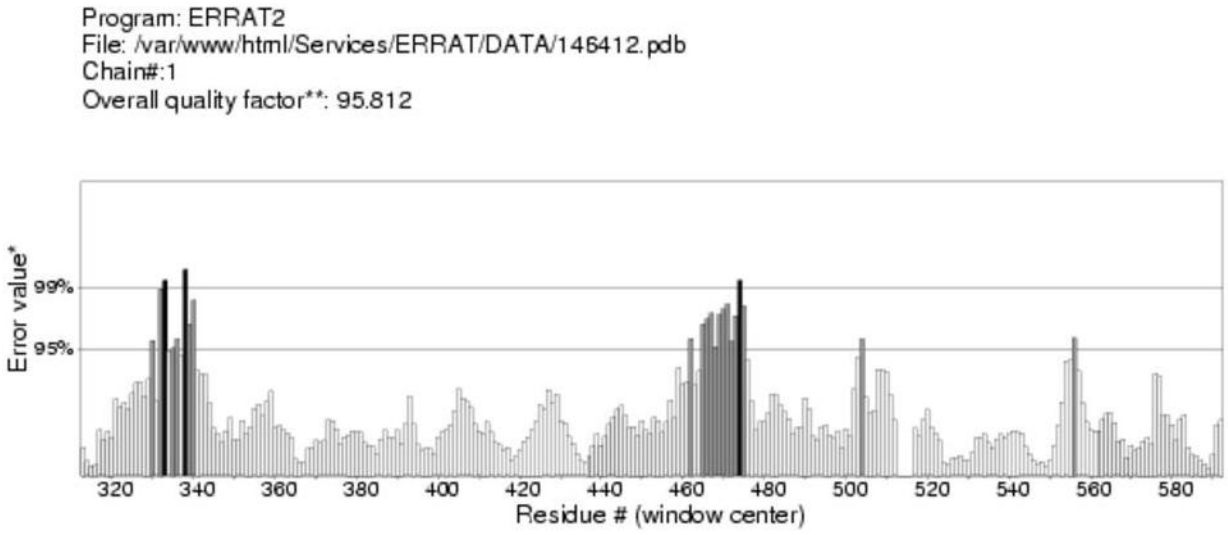


Figure 14: EBA175 ERRAT Output.

Regions that can be rejected at the 95% and 99% confidence levels are those that go beyond the 95% and 99% marks as seen in the diagram above.

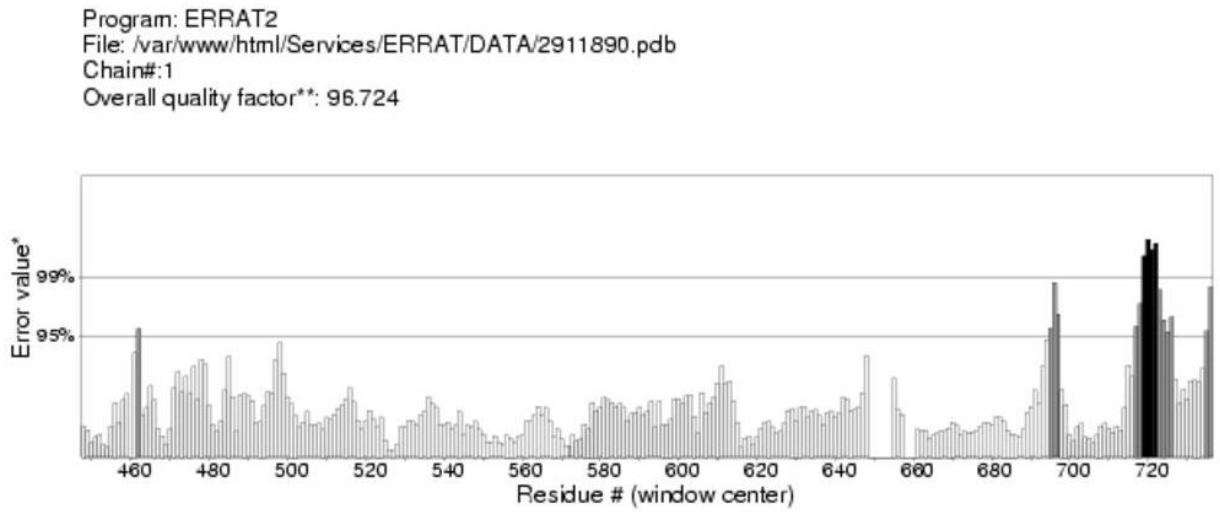


Figure 15: EBA140 ERRAT Output
Regions that can be rejected at the 95% and 99% confidence levels are those that go beyond the 95% and 99% marks as seen in the diagram above.

Figures 14 and 15 showed that both EBA175 and EBA140 had good structures as their overall quality factors were above the threshold given by ERRAT of 90%. Most of the residues fall within the acceptance region.

7.3 HABP1

The first protein fragment under study was the HABP1 [33] from MSP3. This was a 30 residue long fragment. After running the job, 10,000 decoys were generated. Clustering was then performed which generated 17 clusters which were then sorted according to their energy scores. From the 17 clusters, the top three which had 5036, 2004 and 1129 members respectively were inspected. The top three decoys were then selected from each of the three top clusters for further analysis.

The next step was to discriminate between the top scoring decoys by Rosetta, take the best, and compare it with the output from I-TASSER and QUARK servers. Table 4 in appendix A shows

the nine top scoring decoys from Rosetta and their resulting scores from the verification servers: ProSA-web, PROCHECK and ERRAT. ProSA-web gave a z-score as well as an energy plot. The column showing PROCHECK percentage in the most favored region was chosen since models with greater than 90% are considered to be of good quality as far as the stereochemical properties are concerned. The overall quality given by ERRAT is also useful in identifying regions that are problematic since it gives a plot indicating the confidence with which a given region should be rejected. The last two rows of the table also show the scores obtained by the models generated by I-TASSER and QUARK servers. This was to help judge which of the models from the three tools was better.

For a model to be classified as having a good structure, it had to perform well in all the three tests. This was because each of these verification tools scrutinizes different aspects of protein structures. It found that in some cases, a model would score very highly in one verification tools and yet so poorly in another. To ensure that the structure satisfies as many structural rules as possible, we chose to take those that had over 90% of residues in the favored region, a quality mark by PROCHECK, over 91% overall quality factor, the threshold by ERRAT, and had low energy plots, below the zero level, when tested by ProSA- web.

From the table 4 (appendix A), S_00009435, a model by Rosetta, had the lowest energy and z-score. It also performed well in the other two tests. Due to its overall performance in the three tests in comparison with the other decoys, as well as the general outlook of its knowledge based energy plot, it was qualified as the best in its category. After comparing it with the models generated by I-TASSER and QUARK, the model was found to have better scores and was therefore a better choice for the HABP1 fragment model. On observing the PROCHECK column of the QUARK model, we found that it had 96.4% of its residues in the favored region.

However, its GFactor was low meaning that some of the properties of the model were unusual which led to its elimination. Since PROCHECK does not define a range of good GFactor values, a decision was made to only indicate the models that had unusual GFactors. These were the ones that had a GFactor of less than -0.5.

It was also noted that despite the models generated by I-TASSER and QUARK having lower scores, they also played a key role in giving confidence in the overall conformation of the final model due to the similarities observed after visualization, as shown in the diagrams below. From the diagrams below, figure 16-19, a general consensus can be observed on the fragment starting off as a right handed helix.

The knowledge based plot generated by S_00009435(figure A in appendix B) showed that the overall energy was still a bit high especially at the beginning of the fragment and therefore the model can still be improved further to minimize the energy score.

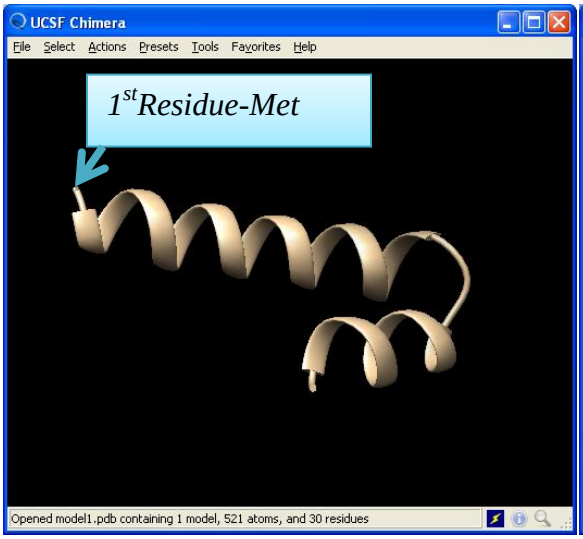


Figure 16: QUARK's HABP1 model

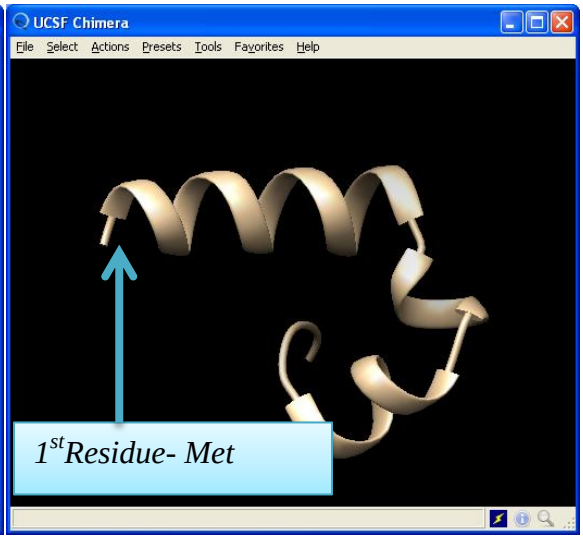


Figure 17: Rosetta's HABP1 model

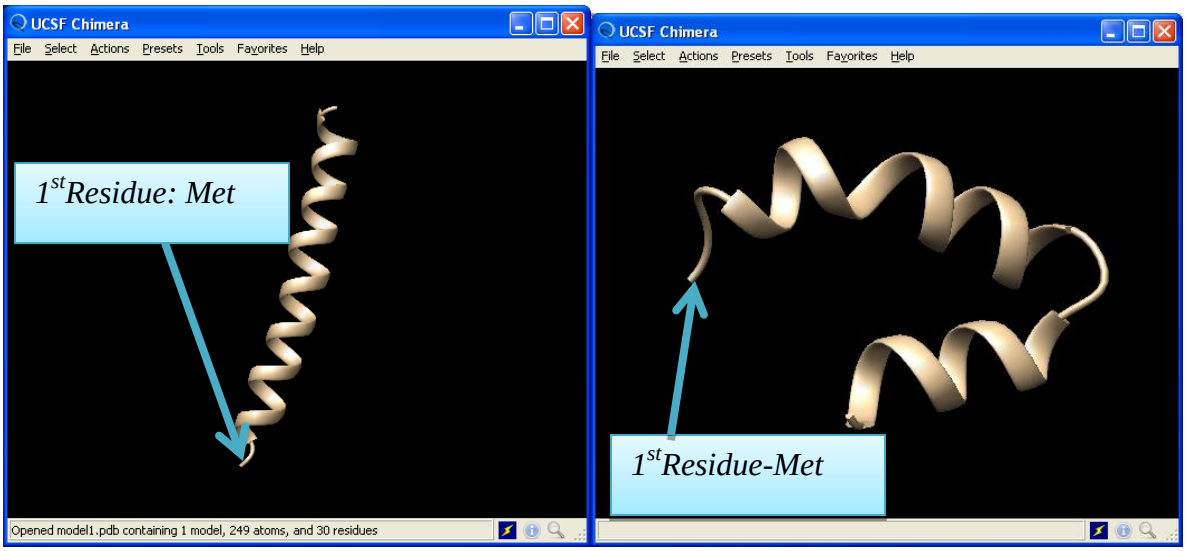


Figure 18&19 : I-TASSER's HABP1 Model1 & 2

7.4 HABP2

The second high activity binding peptide was 30 residues long. In the case of Rosetta, we generated 5000 models. After clustering, 10 clusters were formed and the top three decoys with the lowest energy scores selected for further investigation. The largest cluster had 2745 decoys followed by 841 and 603 in the second and third clusters respectively. The models obtained from I-TASSER and QUARK were then compared with the top scoring decoy by Rosetta. Table 5 in appendix A shows the scores obtained by the selected models in the different verification tests.

As can be observed in table 5, the top scoring model by Rosetta falls short of the 90% quality mark put by ERRAT. For this fragment of MSP3, S_00004882, the third model in the first cluster was chosen as the better model under the Rosetta decoys due to its better scores in the three tests. The latter part of the fragment showed higher energy levels as shown in figure C in appendix B.

On comparing with QUARK and I-TASSER models, the Rosetta decoy was found as the better model and therefore selected as the model of choice. Both I-TASSER and QUARK models

returned an error when submitted to the ProSA-web server. In addition, it was found that the two do not fare well in the ERRAT assessment. However, we observed that the QUARK model had a high PROCHECK score. Figures B (i-iii) in appendix B show the structures of the top scoring models by the three tools.

On visualizing the top scoring models, it was established that there was a disagreement with regard to the latter part of the fragment after the turn. This was because Rosetta had a right handed helix whereas QUARK generated beta sheets and finally I-TASSER had no secondary structure after the turn. The three tools however seemed to agree on the structure of the first section.

7.5 HABP3

The third High Activity Binding Peptide, HABP3, was 40 amino acids long. 10, 000 decoys were generated using Rosetta and the resulting models clustered. There were 93 clusters formed with the top three clusters having 3621, 1253 and 873 members respectively. The top three models in each of the three clusters were taken through refinement using Rosetta's relax algorithm. Table 6 in appendix A shows the different scores for these decoys as well as QUARK's and I-TASSER's top models.

A closer look at table 6 showed that the energy scores by Rosetta were lower than what those obtained for the other two HABPs. Upon inspection of the top scoring decoys by Rosetta, it was found that both top models,S_00005426 and S_00007390, in the first and second clusters had scored fairly well. They both scored above the 90% mark required by PROCHECK and 91% mark by ERRAT. In addition, their knowledge based energy plots by ProSA (figure 20 for S_*00005426_0001*) also showed that the entire fragment fell below the zero level hence the conclusion that they were good models.

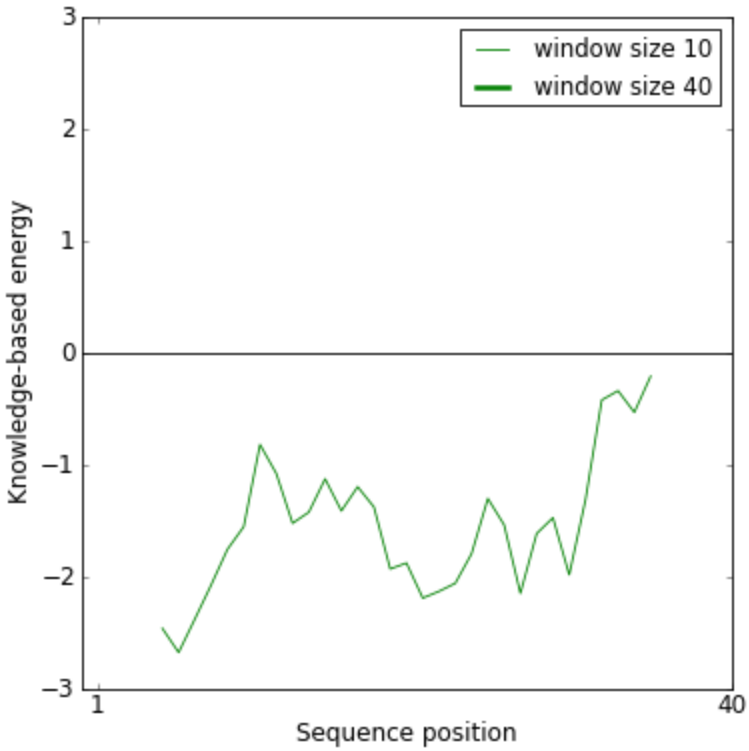


Figure 20: Knowledge-based energy plot of Rosetta's HABP3 model S_00005426_0001

The thin green line represents the energy plot based on a window of size 10.

I-TASSER and QUARK models also had good scores in the three tests. However, it was observed that the GFactor for both of these models was a bit low meaning that there was something unusual about their stereochemical properties. Therefore, once again, the models generated by Rosetta were qualified as the more accurate ones.

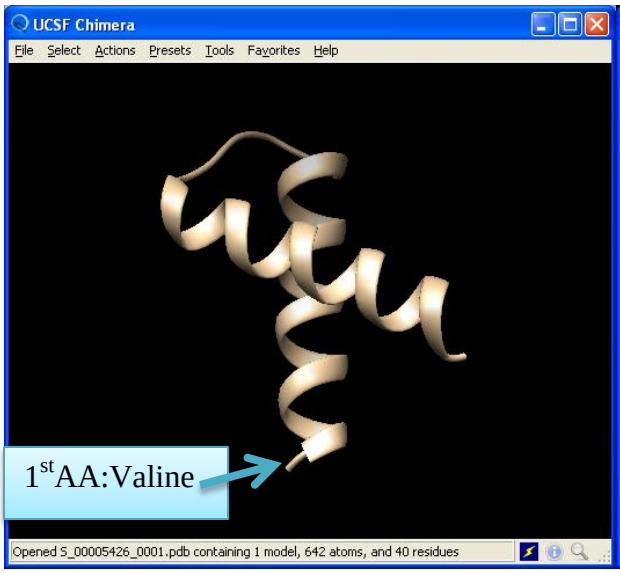


Figure 21:Rosetta's HABP3 Decoy S_00005426

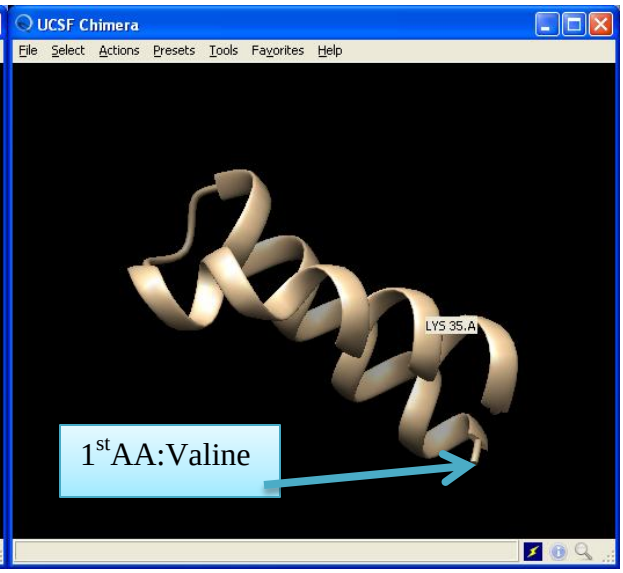


Figure 22:Rosetta's HABP3 Decoy S_00007390

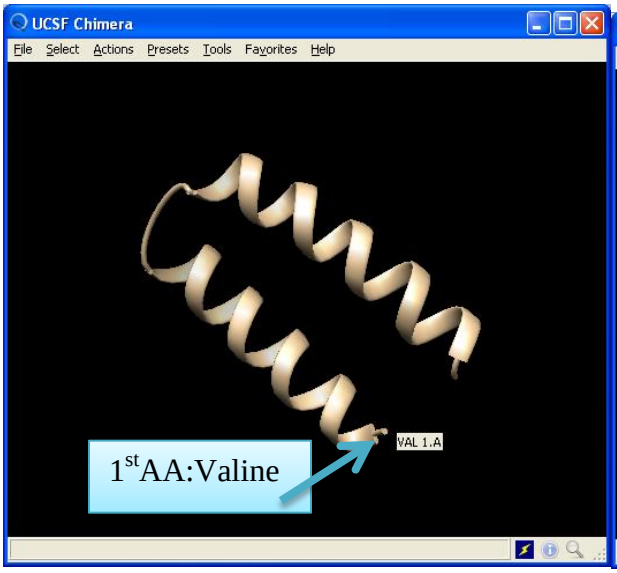


Figure 23: I-TASSER's HABP3 model

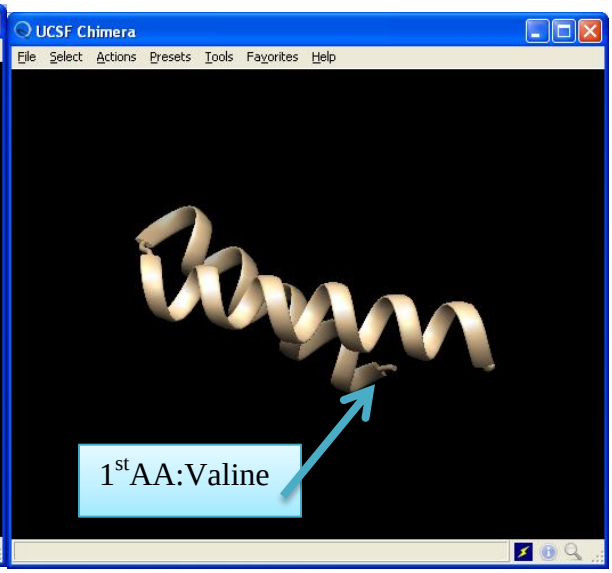


Figure 24: QUARK's HABP3 model

A lot of similarities were observed on visualization of the top models as shown on figures 21-24 above.

7.6 MSP3b Fragment

The other fragment that was modeled was the one that was identified by Quevray et al., 1994. Due to the fact that the minimum number of amino acids required by the fragment library is 30; the MSP3b fragment was extended to 30. 10,000 decoys were generated using Rosetta and then clustered. Five clusters were formed with the top three having 4103, 3873 and 1730 members respectively. For each of the clusters, the three top decoys were selected and taken through relax, the refinement algorithm. The sequence was also submitted to QUARK and I-TASSER servers and the models' performance with regard to the various tests was as shown in table 7 in appendix A.

From table 7, it was observed that the energy level as scored by Rosetta was very low compared to that of the fragments analyzed previously e.g. HABPs. This helped to explain the scores obtained from the various tests which is not so good overall. On assessment of the decoys by Rosetta, it was found that the top model in the second cluster, S_00004297, was a better alternative although it had not scored very well according to PROCHECK.

On taking a closer look at the scores obtained by the models by QUARK and I-TASSER, it was found that they did not score well either. I-TASSER's model passed the stereochemical test but did very poorly in the ERRAT test. QUARK's model had a very low GFactor.

From the models shown in figures D (i-iii) in appendix B, it was observed that Rosetta and QUARK showed some similarities in the majority of the fragment whereas I-TASSER's latter part also agreed with Rosetta's latter section. However, the overall assessment, from the results in table 7, is that this was a problematic segment to model.

7.7 SINGH 70aa Fragment

This 70aa long fragment was identified by Singh et al., 2004 as a target for antibodies which they suggested should be part of a malaria vaccine construct [18]. For this fragment, 4766 decoys were generated using Rosetta. The number was lower than for the other fragments since the processing was interrupted and due to time constraints, entire job could not be rerun. The silent file was therefore edited to remove the 4767th record which was being modeled at the time of interruption. After performing clustering, 8 clusters were formed with the top 3 clusters having 1966, 1322 and 367 members respectively. The sequence was also submitted to QUARK and I-TASSER servers. The verification results are as shown below in table

On assessing the decoys by Rosetta (table 8 in appendix A), it was found that S_00002298, the top model in the first cluster was a better model compared to the rest of them. The choice was guided by the fact that this decoy scored well in the three tests scoring 95.3% in PROCHECK's test and 98.413% as the overall quality factor by ERRAT. Its knowledge based energy plot by ProSA, figure 25, also showed that it conformed more to the pattern observed in our bench mark structures. On taking a closer look at the models generated by the automatic servers, it was found that both fell below the quality marks set by ERRAT and PROCHECK.

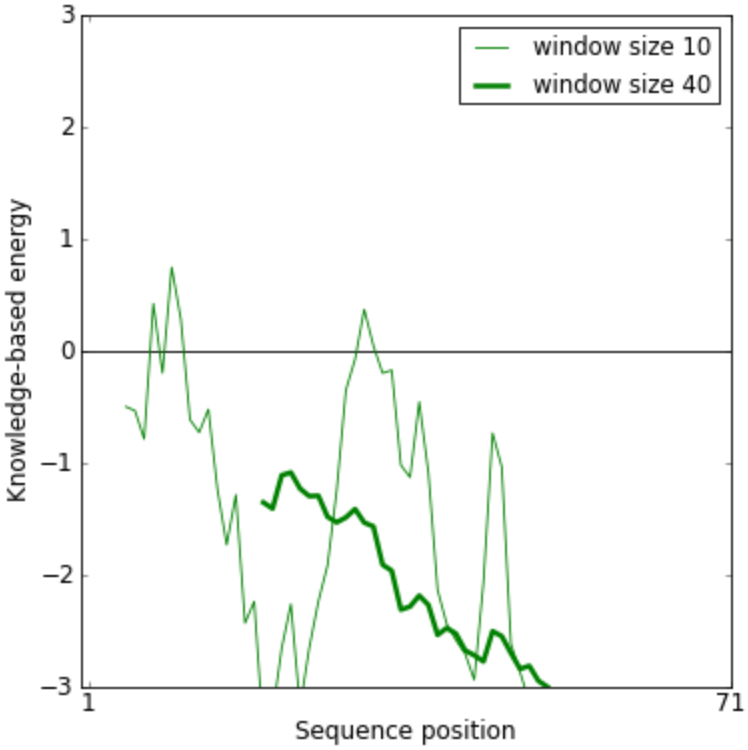


Figure 25: Knowledge based energy plot by Rosetta's 70aa fragment S_00002298_0001.
We consider the thin green line which coincides with the plot with window of size of 10. The dark green line is the plot formed from sliding window of size 40.

An important note was that this fragment shared a common region with the MSP3b long fragment that was identified by Quevray et al., 1994. As mentioned earlier, the MSP3b fragment was somewhat problematic to model illustrated by the scores obtained in the various tests in table 7 in appendix A. This was confirmed further on visualization of the 70aa fragment as shown in the diagrams below (figures 26-28).

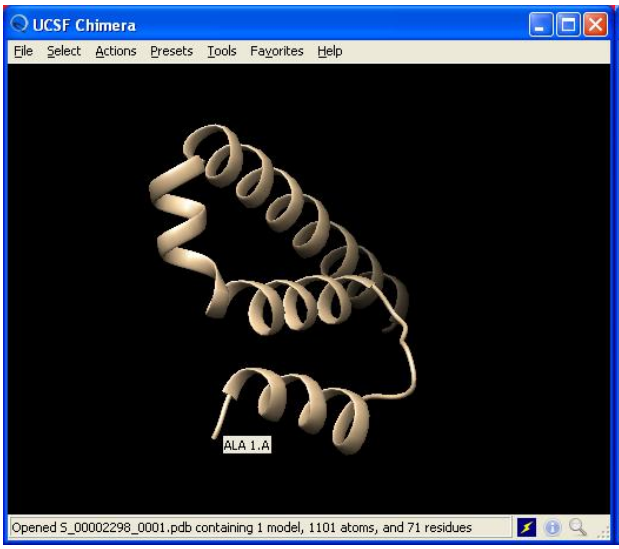


Figure 26: Rosetta's 70aa decoy

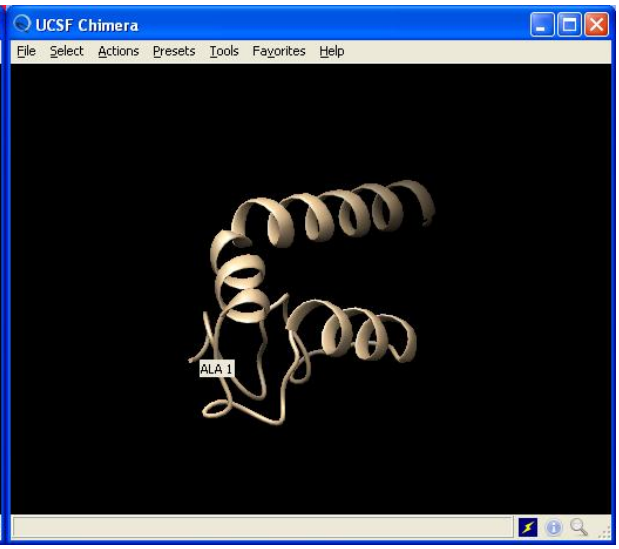


Figure 27: QUARK'S 70aa model

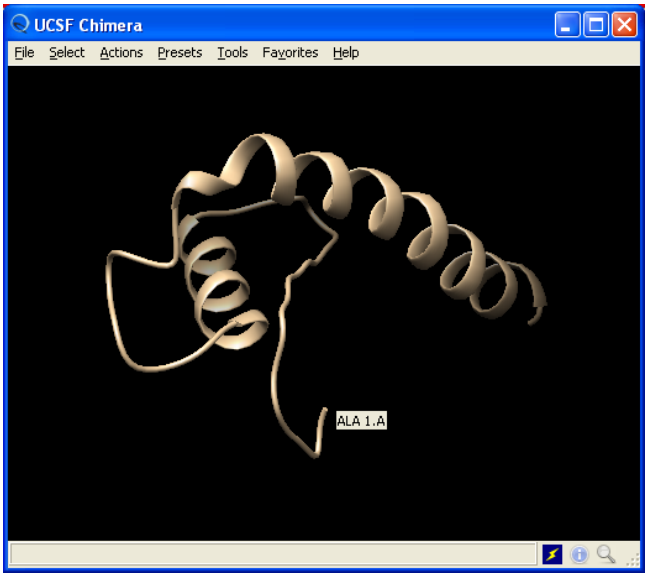


Figure 28: I-TASSER's 70aa model

As demonstrated in the diagrams above, the three tools did not agree on the structure of the first part which was the region shared with the MSP3b fragment in section 7.5 above.

7.8 MSP6BC

Merozoite Surface Protein 6 was the other protein under study of which we also identified some fragments and modeled them. The first fragment was MSP6BC that contained two of six overlapping peptides, B and C, selected for a study on cross-reactive antigenicity between MSP3 and MSP6 [14]. B was found to be cross-reactive but C was not. However, our fragment contained C for the purpose of having a longer fragment that the fragment library could use to generate fragment files required by Rosetta. This fragment was 47aa long. Ten thousand decoys were generated using Rosetta. Clustering was then performed leading to formation of 7 clusters. The top three clusters had 5567, 3050 and 667 members respectively. The top three decoys from each of the top three clusters were selected and run through the refinement algorithm. This was then followed by verification tests and the results are as shown in table 9 in appendix A.

Decoy S_00000190 and S_00004658 which are the first and second in the top cluster seemed to perform fairly well. Specifically, S_00000190 had 91% of its residues falling in the favored region and scored 100% in ERRAT test. The knowledge based energy plot for S_00000190, figure 29, shows that most of the fragment has the line falling below the zero level.

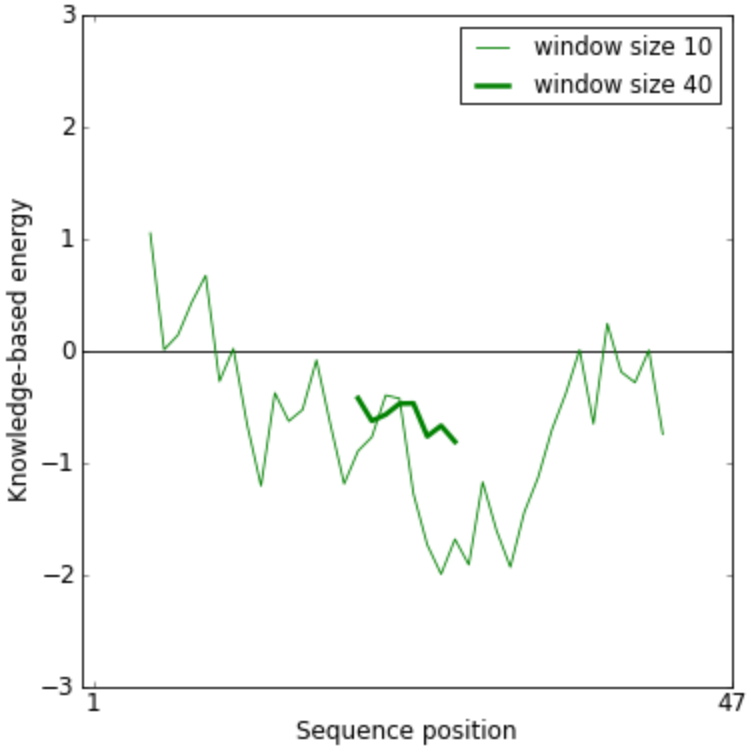


Figure 29: Knowledge based energy plot of Rosetta's MSP6BC model S_00000190.
The dark green line is the plot formed from sliding window of size 40

A closer observation of the models generated by QUARK and I-TASSER showed that they both failed PROCHECK and ERRAT tests. The top decoys are shown in figures E (i-iv) in appendix B for visualization.

7.9 MSP6D

The other peptide under study, reported by Singh et al. (2005), was peptide D which was also found to be cross-reactive. This fragment was 53aa long. Ten thousand decoys were generated and clustering performed on them. Nine clusters were formed with the top three clusters having 4467, 2215 and 1266 members respectively. For each of these clusters, the top three decoys were

chosen and taken through refinement. The scores for the refined decoys as well as those of the other two servers are as shown in table 10 in appendix A.

On further investigation of Rosetta decoys, it was found that S_00000562 was a better model due to the fact that its knowledge based energy plot, figure 30, had most of its points below the zero mark. Test scores obtained by the models generated by QUARK and I-TASSER, showed that they both scored poorly in PROCHECK but ERRAT appeared to score QUARK's model highly and I-TASSER's extremely poorly. The diagrams F (i-iii) in appendix B show the visual aspect of the models.

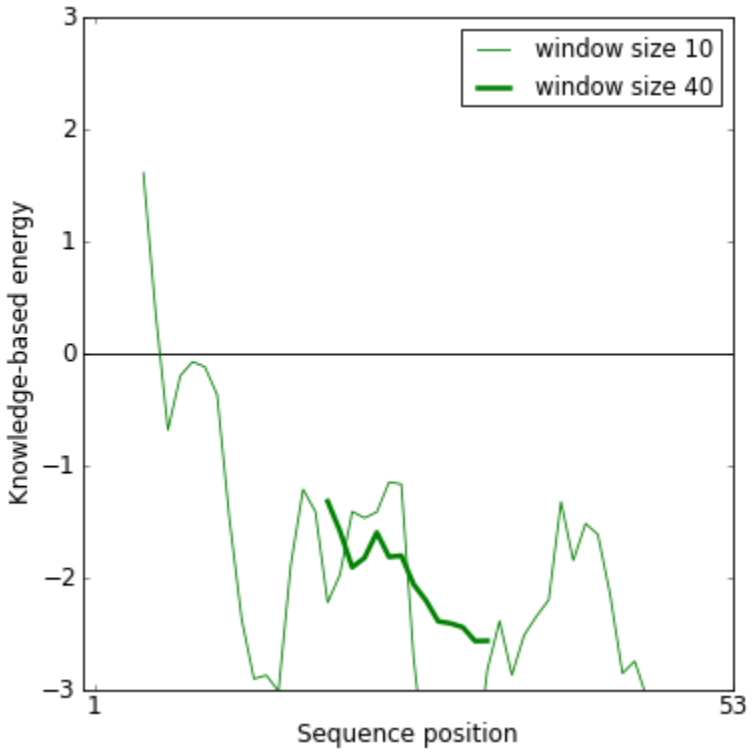


Figure 30: Knowledge based energy plot for Rosetta's MSP6D fragment S_00000562_0001.
The thin green line created from window of size ten shows the energy for this fragment is low. The dark green line is the plot formed from sliding window of size 40.

Visualization of the models helped one understand why ERRAT would score I-TASSER's model so poorly. This is because it was evident that I-TASSER was not able to fold a big percentage of the fragment.

Also of importance is the fact that peptide C and D shared a region where the two peptides overlapped. There were striking similarities observed in the two common regions as illustrated in the figures G (i-ii) in appendix B.

7.10 MSP6F

The last peptide modeled was MSP6F which was also found to be cross-reactive. This fragment was 52aa long. For this region, 6295 decoys were generated using Rosetta. After clustering, 15 clusters were formed and the top three clusters had 4152, 605 and 434 members, respectively. From each of the selected clusters, the top three decoys were selected for further analysis. The sequence was also submitted to QUARK and I-TASSER servers. The results of the tests are as shown in table 11 in appendix A.

On further assessment of the Rosetta decoys, it was found that S_00004550, the top model in the first cluster, was a better decoy as it performed well on all the tests as well as the knowledge based energy plot, figure 31, shows that the entire fragment is below the zero level.

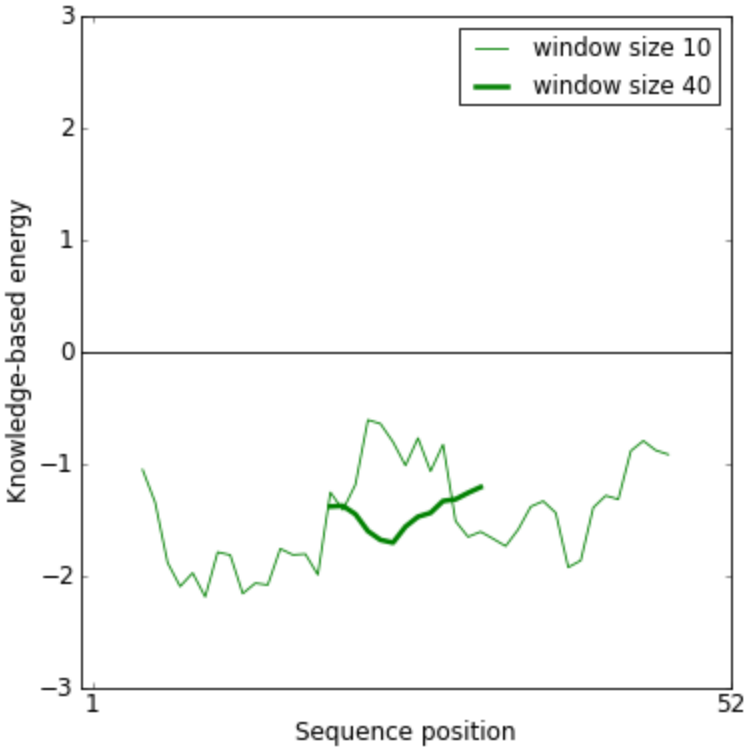


Figure 31: Knowledge based energy plot of Rosetta's MSP6F model S_00004550_0001.
The thin green line shows that the energy level of this model falls below the zero level. The dark green line is the plot formed from sliding window of size 40.

The models generated by I-TASSER and QUARK scored above 90%, the quality mark by PROCHECK but still had relatively low GFactors implying that there was something unusual about their stereochemistry. QUARK's model however also did score well on ERRAT which was not the case with I-TASSER's model.

The diagrams below show the models by the three tools.

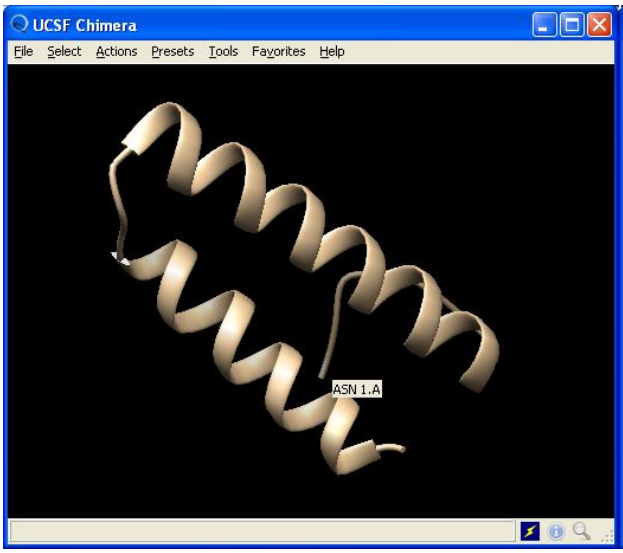


Figure 32: Rosetta's MSP6F decoy

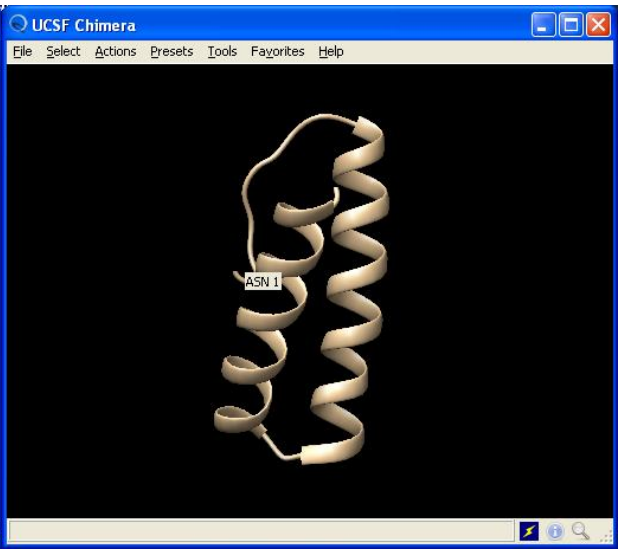


Figure 33: QUARK's MSP6F model

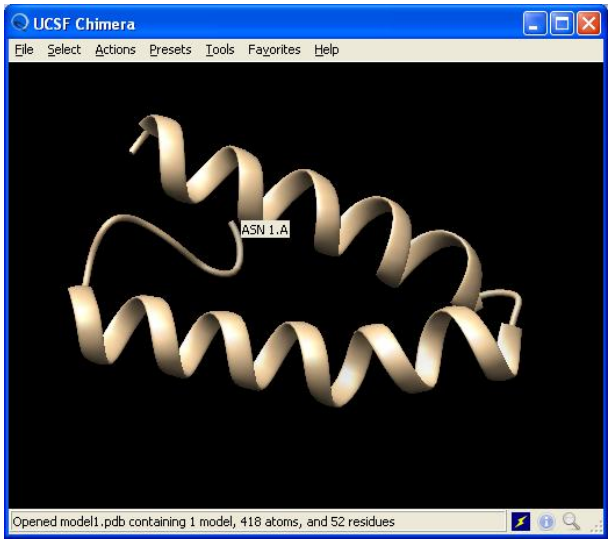


Figure 34: I-TASSER's MSP6Fmodel

From the diagrams above, one can see that the three tools seemed to have a consensus on the overall structure of this fragment.

7.11 MSP3 3D7 Indel Segment

One of the objectives of this study was to find out the implications of the allelic differences on the 3D structure of MSP3 and MSP6. Since *ab initio* modeling does not allow one to model big proteins, it was decided that it would be better to narrow down to the regions with great

differences in the two proteins under study and see the effect of the insertions/deletions in the two alleles.

From alignment figure 5 above, one can observe that the region with most disparities between the two strains lies in the first half of the sequence. This is the part we describe below in the case of MSP3.

This fragment was 88 residues long. Ten thousand decoys were generated and later clustered. 12 clusters were formed with the top three clusters having 4777, 1886, and 980 members respectively. The top three decoys for each of the top three clusters were then taken through a validation step and the test results are as shown in table 12 in appendix A. The results of QUARK and I_TASSER are also shown in the same table.

From the inspection conducted on the top Rosetta decoys, S_00005572, the second model in the first cluster was chosen as the better one in the set. The knowledge based energy plot for this model, figure 35, shows that most of the fragment had the line falling below the zero level which agreed with our benchmark structures.

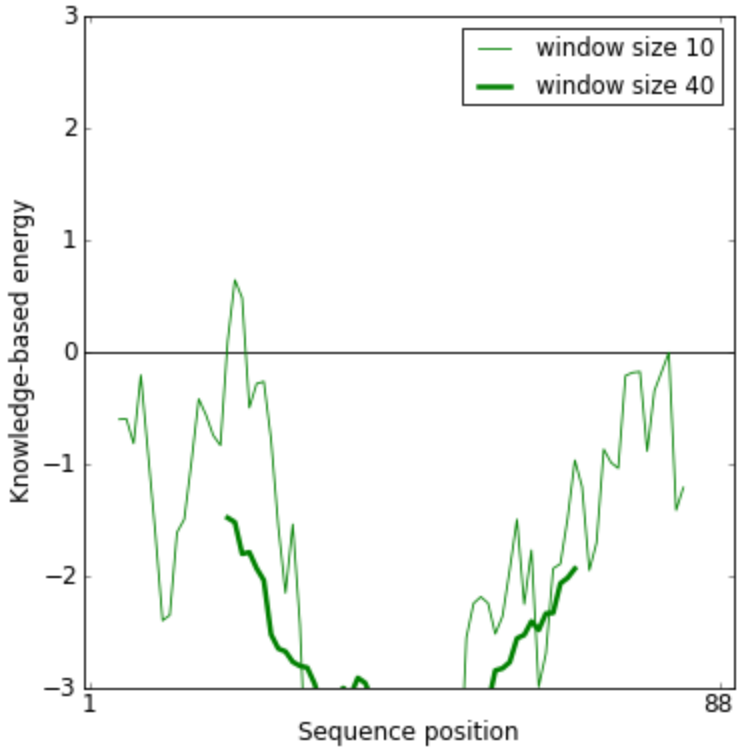


Figure 35: Knowledge based energy plot of Rosetta's MSP3 3D7 indel model. Most of the thin green (window of size 10) line is below the zero level. The dark green line is the plot formed from sliding window of size 40.

On comparing Rosetta's model with those by I-TASSER and QUARK, it was found that this model scored better than the two. I-TASSER's model had a GFactor of -0.08 and 2.5% of the residues were falling in the disallowed region. QUARK's model had a high percentage of its residues in the favored regions but its GFactor was a bit low at -0.18. However, QUARK and I-TASSER models still served the purpose of giving confidence on the general structure of the fragment. The figures below show the visual aspect of the structures by the three tools.

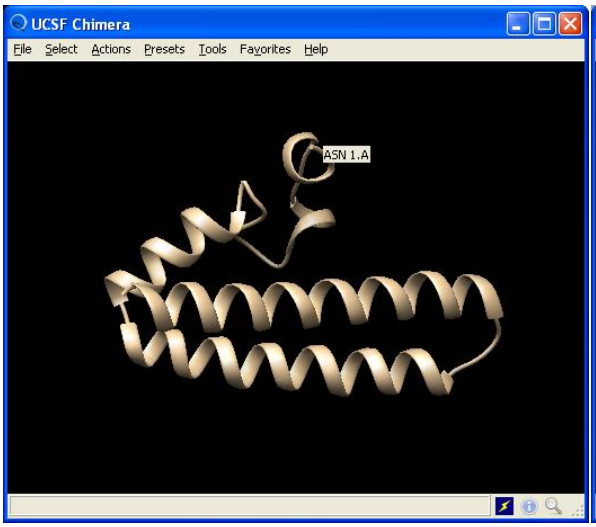


Figure 36: Rosetta's MSP3 3D7 indel model

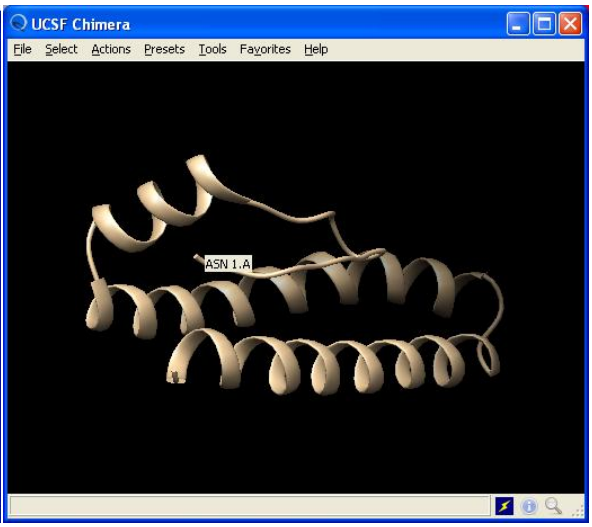


Figure 37: I-TASSER's MSP3 3D7 indel fragment model

Figure 38: QUARK's MSP3 3D7 indel fragment model

On visual inspection of the 3 models by the three tools, it was found that Rosetta and I-TASSER had more similarities. The two models were therefore considered in the comparison with the models for the K1 strain to identify the differences introduced by the indels.

7.12 MSP3 K1 Indel Segment

The fragment in our other strain under study, K1, was 114 residues long. For this, ten thousand decoys were generated and clustered. Clustering led to formation of 12 clusters. The top three clusters had 5285, 2213 and 544 members, respectively. The three clusters were a good representation of this dataset of decoys as the total number of the members was coming to 8042. The top members in the three clusters were therefore investigated further to look at their correctness and quality. The table 13 in appendix A has the score for the Rosetta decoys as well as QUARK and I-TASSER's models.

From the table 13, S_00000467, the top model in the first cluster appeared to score well and therefore chosen as the better decoy. Despite S_00006919 having similar scores as S_00000467, the latter was chosen because the energy plot, as shown in figure H(i) (appendix B), showed that it had lower energy than S_00006919 H(ii). I-TASSER and QUARK models had good scores by PROCHECK although the GFactors were still a bit low. ERRAT however seemed to score the model by QUARK better than the one by I-TASSER. Figures I (i-iii) in appendix B show the visual aspect of the models.

The diagrams show some agreement between QUARK and Rosetta for the better part of the fragment. However, Rosetta's model has a turn which is then followed by an alpha helix which is different from QUARK's that maintains a helix all the way to the end. The 3D structures of MSP3 3D7 and K1 were then studied side by side to highlight the key differences. Since visual inspection may not be very useful in highlighting the major differences in the arrangement of atoms in a protein, it was resolved that a better approach would be to observe the changes in the possible ligand interaction and antibody binding sites in the two strains which are discussed in the pocket identification section below.

7.13 MSP6 3D7 Indel Segment

Regions of main differences between the K1 and 3D7 strains of MSP6 were also selected and modeled. In the case of the 3D7 strain, the fragment size was 50aa long. Using Rosetta, 10,000 decoys were generated and then clustered. Five clusters were generated with the top three clusters having 5156, 2662 and 1780 members respectively. The top three members in each of the top three clusters were then taken through validation tests. The details of the results obtained for the structures generated as well as their test results are shown in table 14 in appendix A.

From the results shown in table 14, it was found that the top decoy by Rosetta did not score well in the ERRAT test. On further assessment of the decoys, it was found that S_00009724, the third decoy in the third cluster, was a better choice among the Rosetta decoys. Its knowledge based energy plot is shown in figure J in appendix B. I-TASSER and QUARK decoys did not perform very well since as can be observed in the table above; both had some residues, 5% and 2.5% respectively, falling within the disallowed regions. Figures K (i-iii) in appendix B help to visualize the results in table 14.

Figures K (i-iii) (appendix B) showed that Rosetta and QUARK had some similarities in their final models. However, it appeared that I-TASSER was not able to fold the protein fragment which would explain the scores the model after carrying out the verification tests.

7.14 MSP6 K1 Indel Segment

The last fragment, 106 residues long, was the indel fragment from MSP K1 strain. This was modeled for the purpose of finding the allelic differences between 3D7 and K1 strains. Ten thousand decoys were generated using Rosetta and submitted the same sequence to I-TASSER and QUARK servers. For Rosetta decoys, clustering was performed which led to 9 clusters being

created. The top three clusters had 4178, 1833 and 1395 members respectively. From these clusters, we took the top three members and did some tests to verify their structures. The table 15 in appendix A shows the scores obtained by each of these models including those by QUARK and I-TASSER.

From the scores in table 15, it was found that none of the decoys by Rosetta obtained the 91% quality mark set by ERRAT. The best decoy by Rosetta scored 86%. This decoy, S_00005452, was the one chosen as a better choice in its category. Figure 39 below shows the knowledge based energy plot by this model. The region with a spike into the positive side of the graph may show that that region needs improvement.

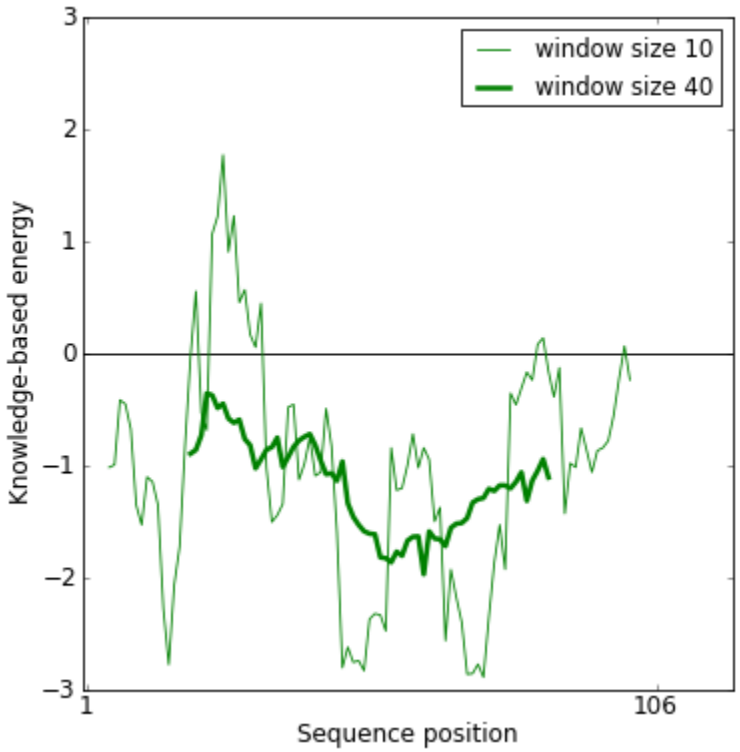


Figure 39: Knowledge based energy plot by Rosetta's MSP6 K1 model S_00005452.
The thin line shows the plot formed from a sliding window of size 10 whereas the dark one is from a window of size 40.

Yet again it was found that I-TASSER's model got a zero in ERRAT's test which shows that it did not manage to fold it correctly. QUARK's model did not do very well either. The diagrams below illustrate how the three tools folded this segment.

On visualization of the models by the three tools, figures L (i-iii) in appendix B, it was found that Rosetta and QUARK had some similarities. I-TASSER's model yet again did not have any well-defined secondary structures which would explain the low test scores.

Overall Positioning of our models in relation to other structures in the PDB is shown below (figure 40).

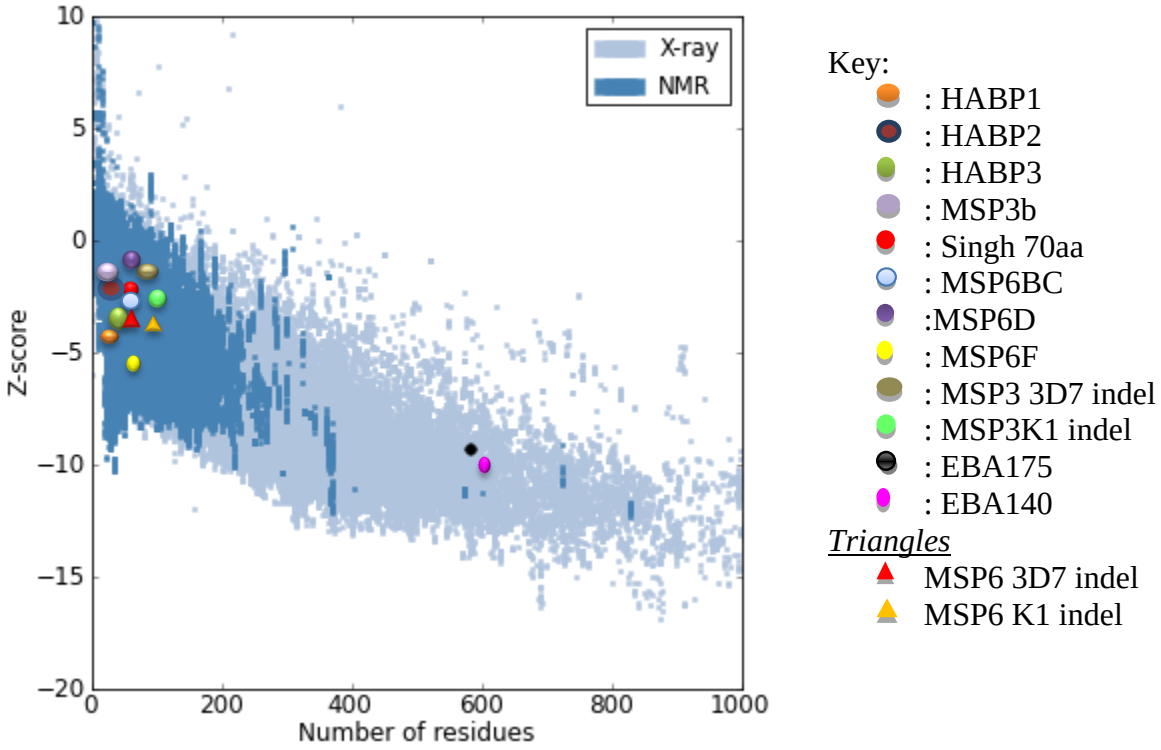


Figure 40: Overall positioning (estimates) of fragment structures in relation to PDB structures by ProSA-web

7.15 Pocket Identification

The final task in this study was to identify possible pockets for the models which had good scores in the verification tests performed. In this section we highlight the pockets found and their properties. This was done using CASTp webserver.

From the results outlined in table 4, the scores obtained gave confidence that the HABP1 models were good; and especially given the similarities observed from the three test tools used. This prompted for further investigation with the aim of identifying pockets on HABP1. The first high activity binding peptide was found to have one pocket as is illustrated by figure 40 below. This pocket was 20.4 cubic angstroms.

HABP1 structure was also submitted to ProFunc which found a nest at the residues Lys26, Glu27, Ile28, and Val29. A nest is a region characterized by having an anion or cation and has been shown to be associated with functional sites of proteins [44]. However, this nest fell within the extension region of our fragment and not the initial HABP1 that was identified.

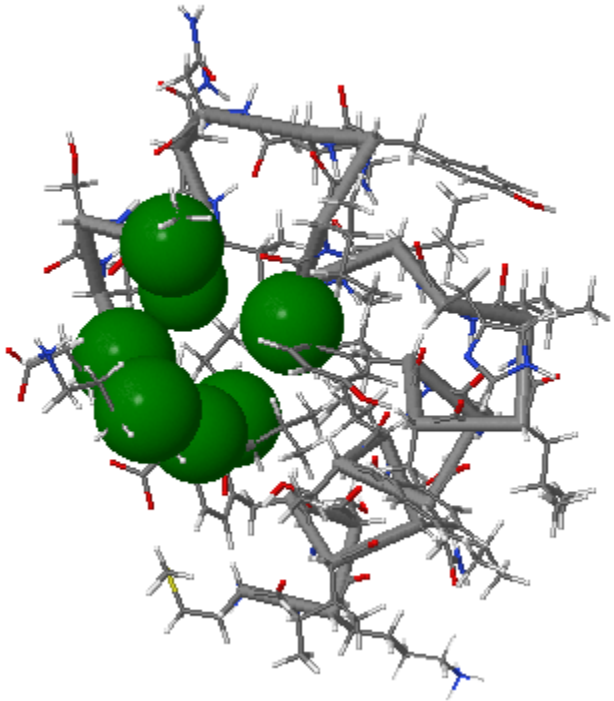


Figure 41: 3D structure of HABP1 showing the pocket identified in green

In the case of the second peptide, HABP2 there was a disagreement by the three tools on the latter part of the fragment. This is because as can be seen in the figures B (i-iii) in appendix B, Rosetta modeled the region as a helix whereas QUARK found beta sheets. I-TASSER did not fold the segment after the turn. For this reason, the structure of this fragment gave less confidence with regard to its correctness and therefore no further analysis was carried out.

The third peptide which was 40 amino acids long had fairly good scores from the tests that were performed. The models were visualized and inspected to see whether there were any identifiable pockets in the structure. From figure 41 below, there were three pockets that were identified by CASTp. The largest pocket, in terms of volume (25.3 cubic angstroms), was represented by the green colored spheres. It was bordered by Ala(5), Iso(8), Tyr(36), Phe(37), and His(40) residues in the fragment most of which are hydrophobic in nature. This was followed by the pocket marked by the cyan colored balls having a volume of 13.1cubic angstroms and which was bordered by four Leu residues. The smallest pocket, represented by the blue balls, had 7.5 angstroms^{3}and was marked by Leu(12),Ile(16 &22) and Val(30) which are hydrophobic in nature.

A nest was also found by ProFunc at residues 18, 19 and 20 which were Gly, Asn and Asn which implied that that region could have an important functional site. This could confirm the results of the study that identified this region as part of the high activity binding peptides.

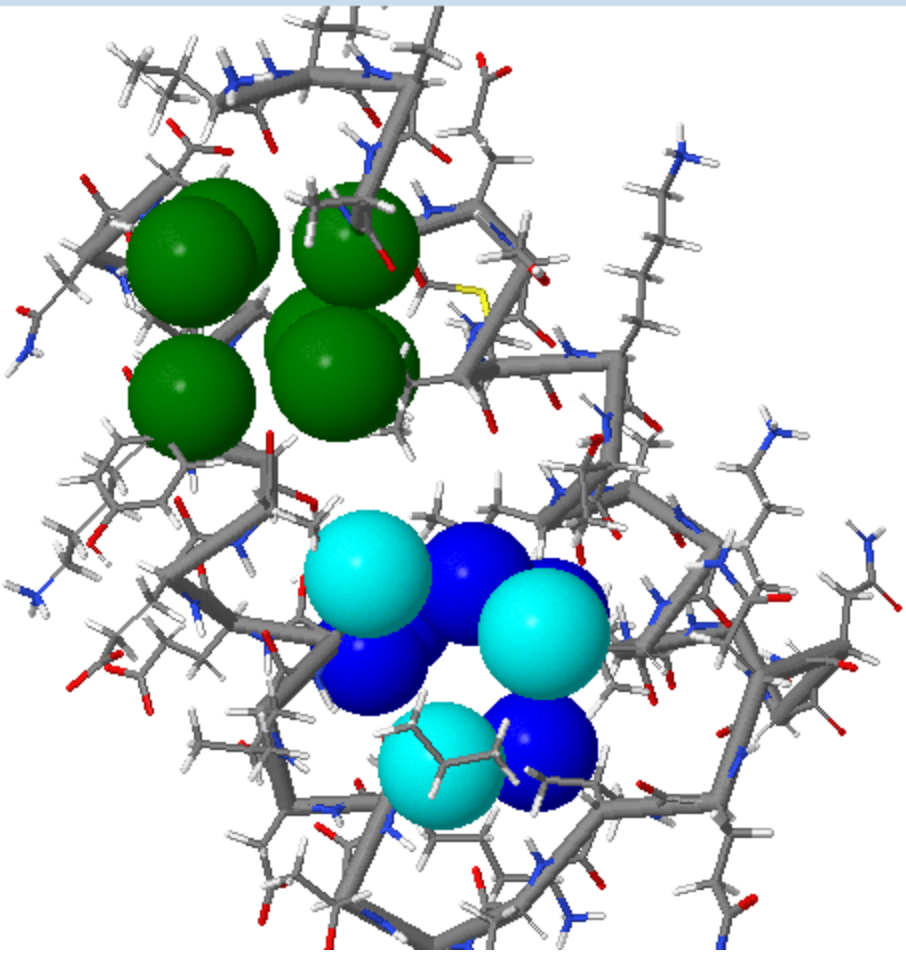


Figure 42: 3D structure of HABP3 showing the pockets identified (spheres). Green, Cyan and blue is the order of the sizes of the pockets found, from largest to smallest

The next fragment we modeled was the MSP3b long region identified by Quevray et al., (1994). From the test scores obtained by the models, as shown in table 7, in appendix A, it was evident that the models were not entirely correct. This was confirmed by the results obtained for the 70aa region was modeled and the results found demonstrated that the MSP3b fragment is a problematic region to model. The three tools did not have a consensus on the structure of the first segment of the 70aa fragment as can be observed in the figures 26-28 on page 37 above. However, it was found that despite the low scores by the three tools, Rosetta and QUARK had beta sheets and turns. Given that two tools out of three had similar output when modeling the

MSP3b region, it was concluded that this could be how the region folds. However, this should be further improved to reduce the energy as well as improve the stereochemical properties and the overall quality of the fragment.

In the case of the latter part of the 70aa fragment, it was found that all the tools had a general consensus on the structure and therefore the model was used to perform further analysis. The first task was to identify possible pockets in the fragment. CASTp was able to identify several pockets and therefore we discuss the deepest pockets which are most likely to be used in ligand interaction [45].

The first pocket, represented with green spheres, had a surface area of 142A^{2}and a volume of 151.8A^3. This pocket was found to cover the first part of the fragment with residues Lys, Glu, Ala, Tyr. Like mentioned earlier, poor models were obtained with regard to the structure of the first few residues which coincided with the MSP3b region that was also modeled but gave poor results. The fact that this pocket was made of residues that are less hydrophobic confirmed that this was likely to be a false hit. Figure M in appendix B shows the residue coverage of this pocket.

The second pocket, represented with cyan colored spheres, had a surface area of 97.6A^2 and a volume of 148.9A^3. This pocket was dominated by Val residues which are hydrophobic as well as Ser and His which are less hydrophobic. Few Asn residues were also found. These are known to be neutral in that they do not carry any charge. The diagram below shows the coverage of this second pocket.

Figure 43: Residue coverage in Pocket 2 in 70aa fragment

The third largest pocket had an area of 94.0A^2 and a volume of 111.2A^3. However, since the pocket was covered by the first few residues in the fragment (figure 44 below) which fell under the region that was poorly modeled, it would be prudent to conduct further investigation to confirm the existence of this pocket. The other property that led to decreased confidence in this pocket is the fact that was dominated by Glutamic acid residues which have acidic polarity.

Figure 44: Residue coverage in Pocket 3 in 70aa fragment

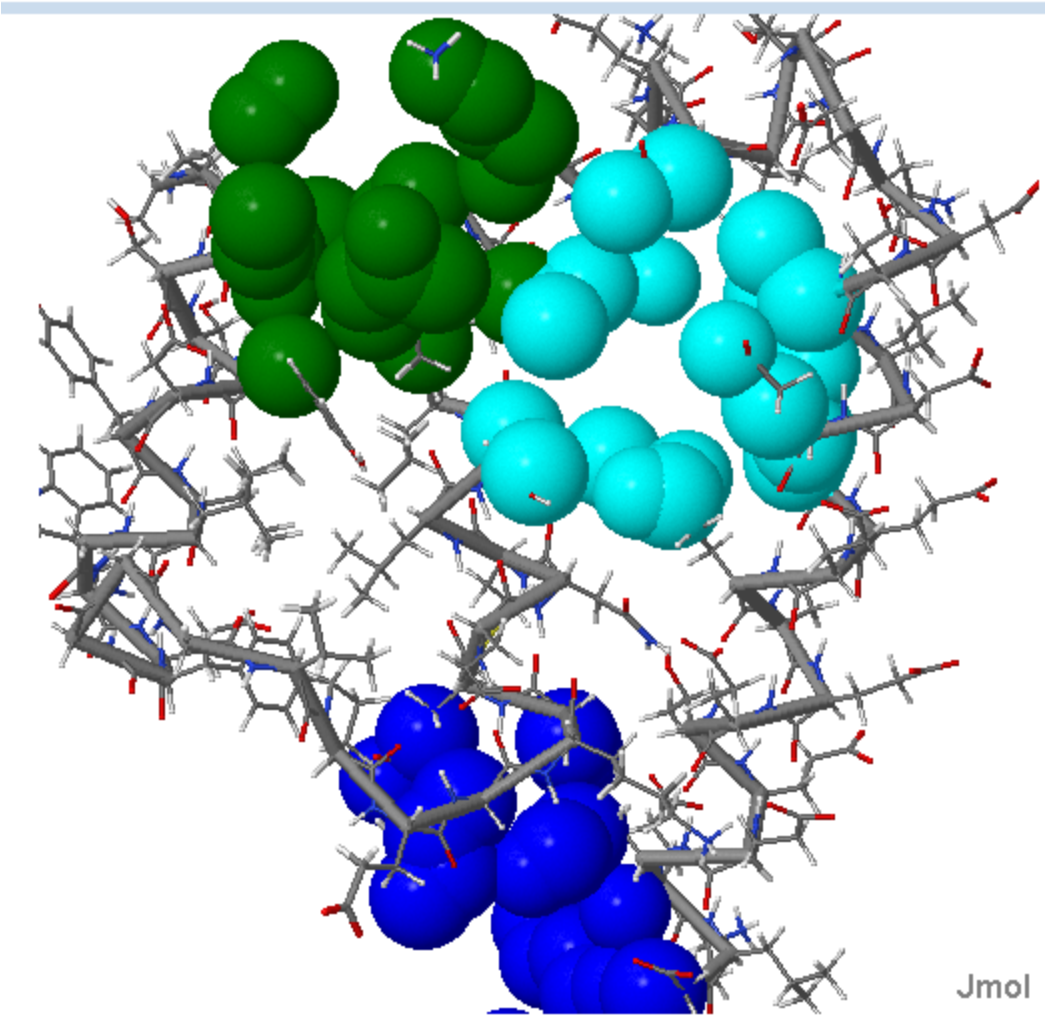


Figure 45: 3D structure of 70aa fragment showing the pockets identified (spheres) Green, Cyan and blue is the order of the sizes of the pockets found, from largest to smallest

In the case of MSP6, three fragments were modeled. These were identified by Singh et al., (2005) to be epitopes which shared cross reactivity with MSP3. The models obtained for the first fragment MSP6BC had little agreement when a comparison was done on the output of the three tools. This was especially in the overall structure of the models and therefore led to decreased confidence in the structure of this region. For this reason no further analysis was carried out.

The second peptide in MSP6 was MSP6D which, from figures F (i-iii) in appendix B, it was found that the three tools did not converge on the same result. Both Rosetta and QUARK showed some similarities. PROCHECK output indicated that there was something unusual about the model generated by QUARK. Rosetta's model was the model of choice and on further inspection of the model, one deep pocket with a volume of $129.5A^3$ was found as shown in figure N in appendix B. However, due to the lack of consensus on the general structure and the presence of non-hydrophobic residues in the pocket such as glutamic acid, figure O in appendix B, it would be advisable that further improvement be done so as to make a conclusive decision about the structure of this region.

The last fragment of MSP6 was MSP6F which was 52aa long. From table 11 (appendix A), the test results showed that majority of the models performed well. On looking at the overall structure depicted from the three tools, it was established that there was some consensus. Further inspection to identify potential binding pockets was then done. CASTp was able to identify several pockets but only the first two were selected since they had larger volume sizes. The first pocket, shown in figure 46, had a surface area of $132.2A^2$ and an area of $118.5A^3$.

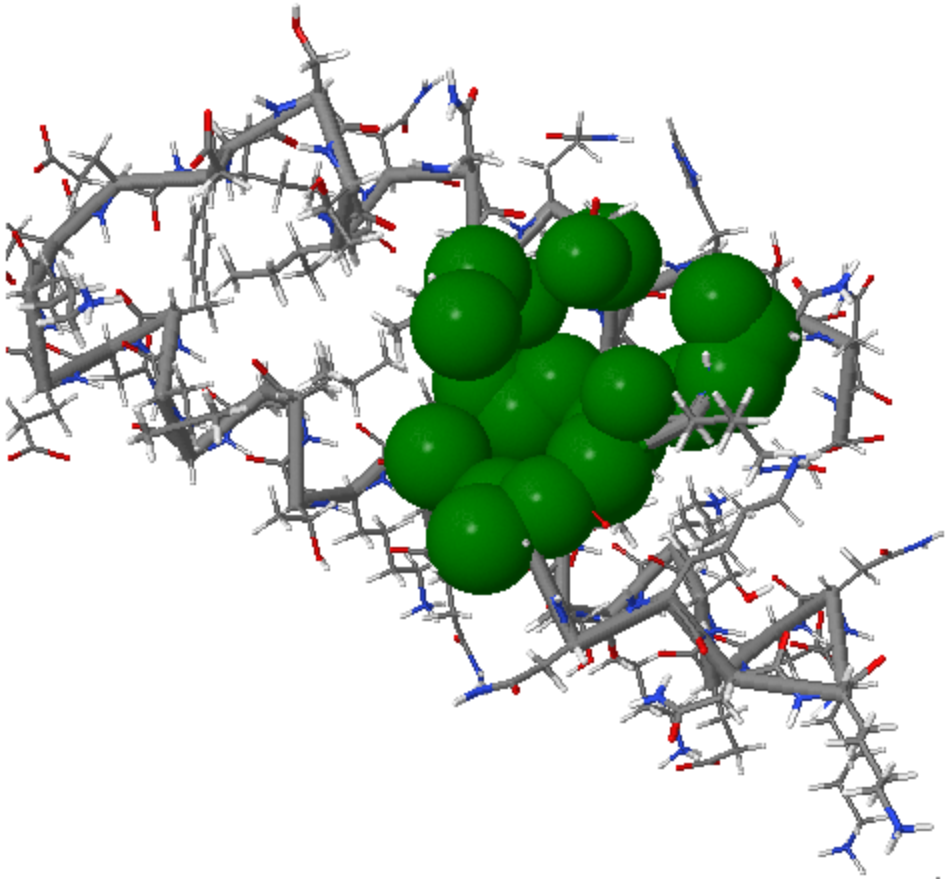


Figure 46:3D structure of MSP6F showing pocket 1 (Green spheres)

The residues that bordered the region, shown in figure 47, were mostly hydrophobic which gave inspired confidence that this was a region that could be used for ligand interaction.

Figure 47: Residues forming the MSP6F pocket highlighted in green

The second largest pocket had a surface area of 106.1A^2 and a volume of 99.3A^3 and was characterized by having hydrophobic residues bordering it. Figures 48 and 49 below depict the location and the specific residues.

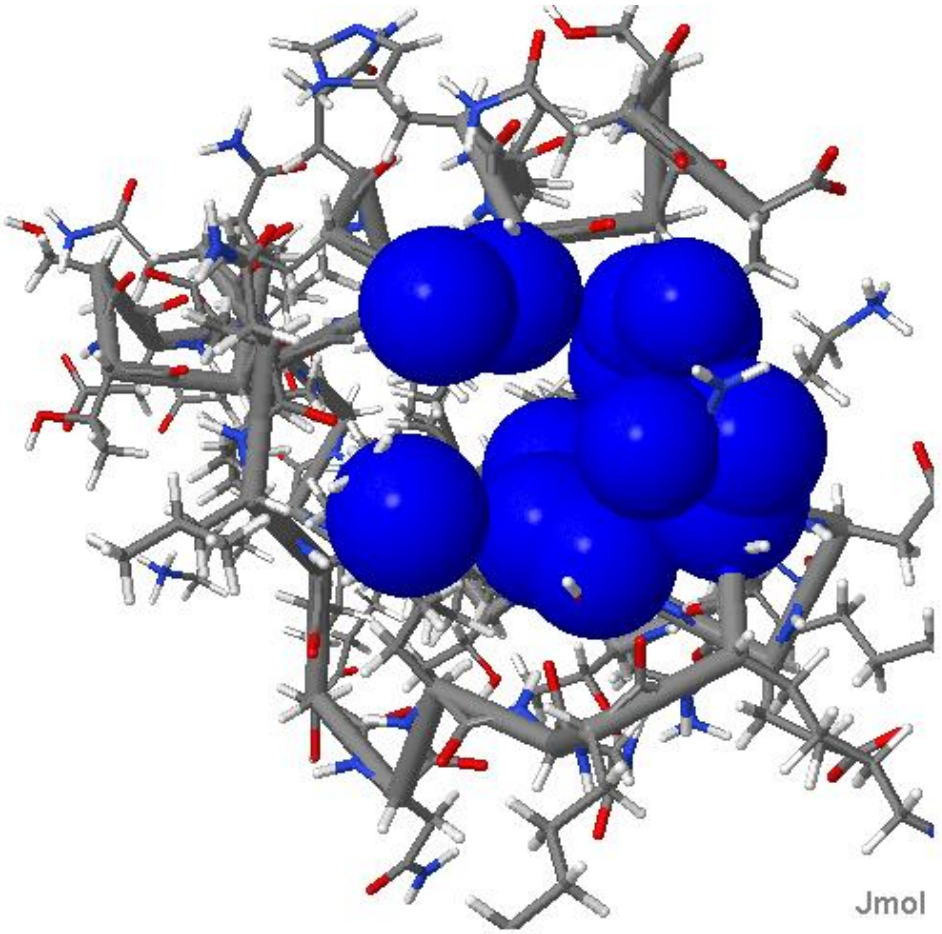


Figure 48: 3D structure of MSP6F showing pocket 2 (Blue spheres)

Chain A

1- NLISNKNKKN DETKKTAENI VKTLVGLFNE KNEIDSTINN LVQEMIHLFS

51- NN

Figure 49: Residues forming the MSP6F pocket 2 highlighted in blue

Indel fragments of MSP3 and 6 were also modeled so as to find out the overall implication of the insertions/deletions on the 3D structure of the two strains. First, a superimposition was done

using the 3Dimensional Structure Superposition (3d-SS) server which resulted in the following diagrams.

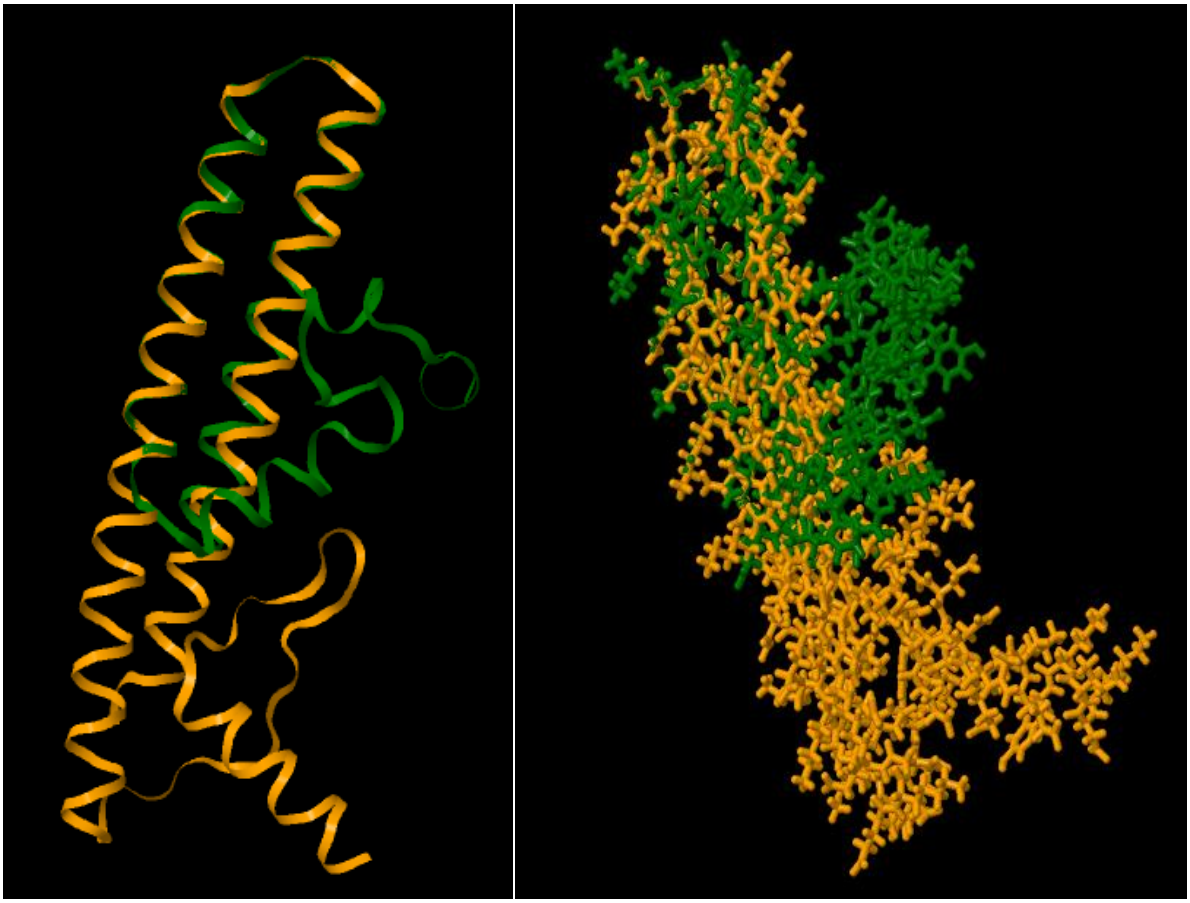


Figure 50: Superimposition of MSP3 3D7(green) and K1(orange) strains indel fragments(left –ribbons, right- wireframe)

From the diagrams above, we observe that the K1 strain forms longer alpha helices. To analyze and see the effect on the 3D structure, we tried to find pockets in both in order to establish whether the polymorphisms interfered with the possible ligand interaction sites. The first pocket, in MSP3 3D7, was very deep with a volume of 765.3A^3 and a surface are of 393.5A^2. The diagram depicting this pocket is shown below, figure 51, as well as the one showing the residues that border it, figure 52.

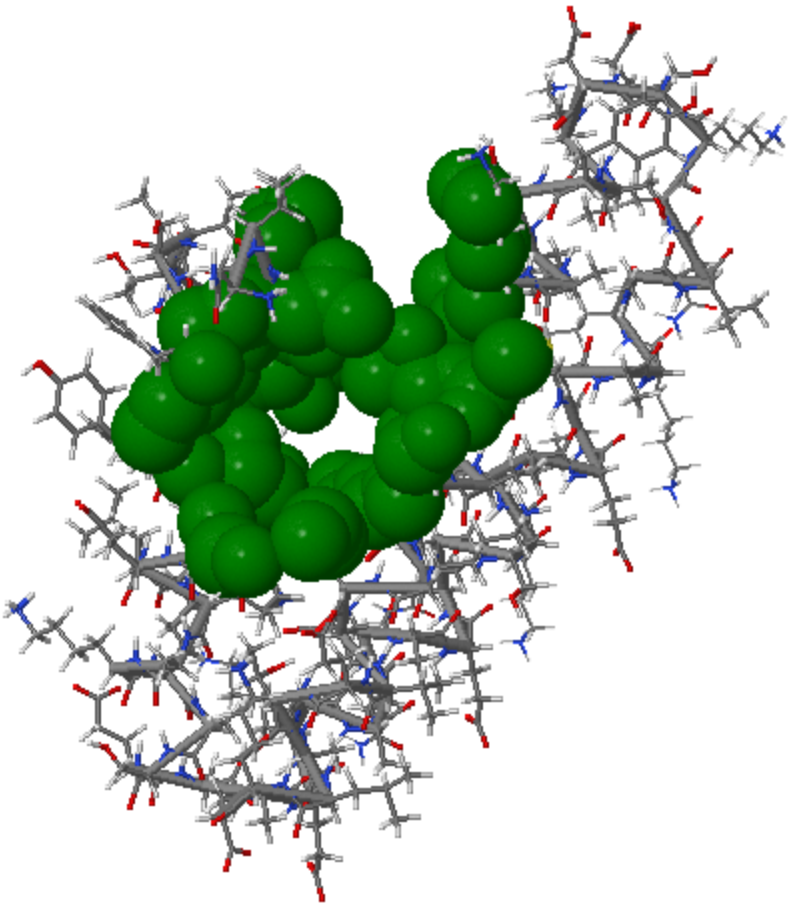


Figure 51:3D structure of MSP3 3D7 indel showing pocket 1 (Green spheres)

Figure 52: Residues forming the MSP3 3D7 indel fragment pocket 1 highlighted in green

The second largest pocket had a surface area of 58.6A^2 and a volume of 49.2A^3. The diagrams below show the location of the pocket.

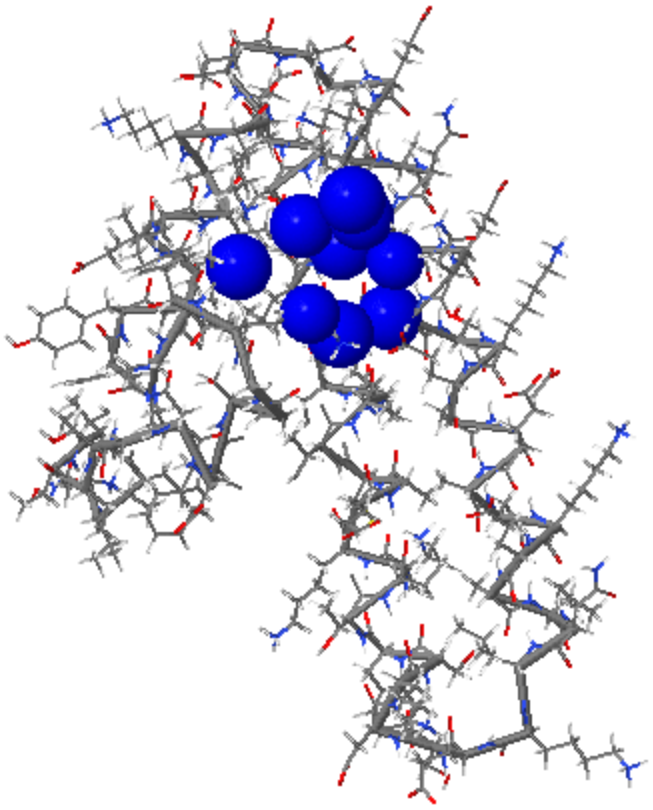


Figure 53: 3D structure of MSP3 3D7 indel fragment showing pocket 2 (blue spheres)

Figure 54: Residues forming the MSP3 3D7 indel fragment pocket 2 highlighted in blue

ProFunc was also able to find a nest in this indel fragment consisting of Gly(20), Tyr(21) and Thr(22) residues.

K1 strain is much longer compared to the 3D7 strain since there are insertions in K1 sequence. The diagram below, figure 55, depicts that there is a possible pocket. This is the largest pocket identified by CASTp. It has a surface area of 187.7A^2 and a volume of 258.7A^3. However, from

figure 55 below, we find this pocket bordered by hydrophilic residues Asp and Glu. This tells us that this could be a false hit on possible pockets.

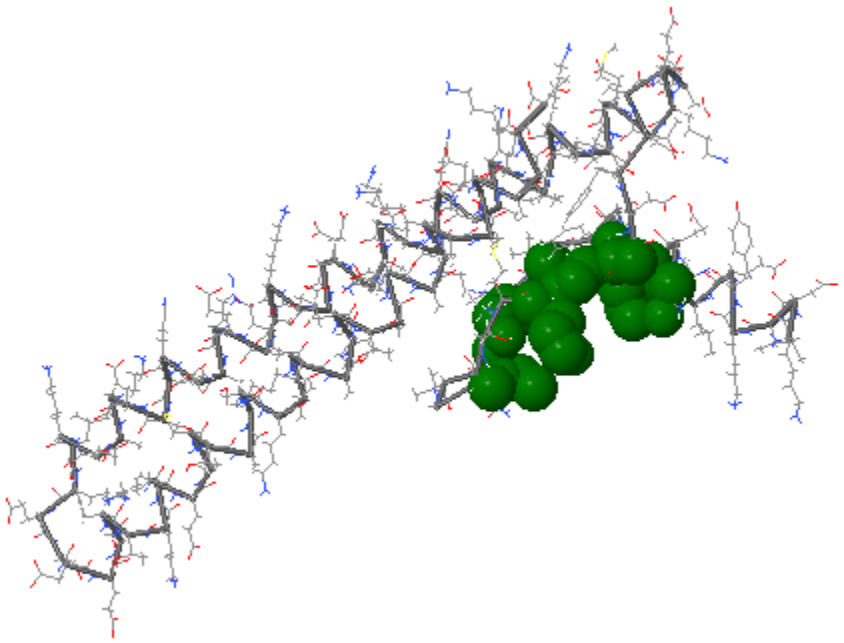


Figure 55: 3D structure of MSP3 K1 indel fragment showing pocket 1 (green spheres)

Figure 56: Residues forming the MSP3 K1indel fragment pocket 1 highlighted in green

The second largest pocket had a surface area of 44.3A^2 and a volume of 73.2A^3. This pocket was characterized by having Glu residues that are non-hydrophobic which led to the conclusion that it may be a false positive pocket hit. Figures P and Q in appendix B show the position of this second pocket and the residues involved, respectively.

Given the differences observed in MSP3 3D7 and K1 strains above, it was observed that the insertions into K1 seemed to have interfered with the pockets.

The study also looked at the differences brought about by the insertions into the K1 strain. Something that stood out was that the new residues led to the second turn of the structure occur much later at residue 75 as opposed to that in 3D7 which was found from residues 58. There is therefore a possible interference of the interaction of side chains of residues by the insertions as shown in the diagrams below.

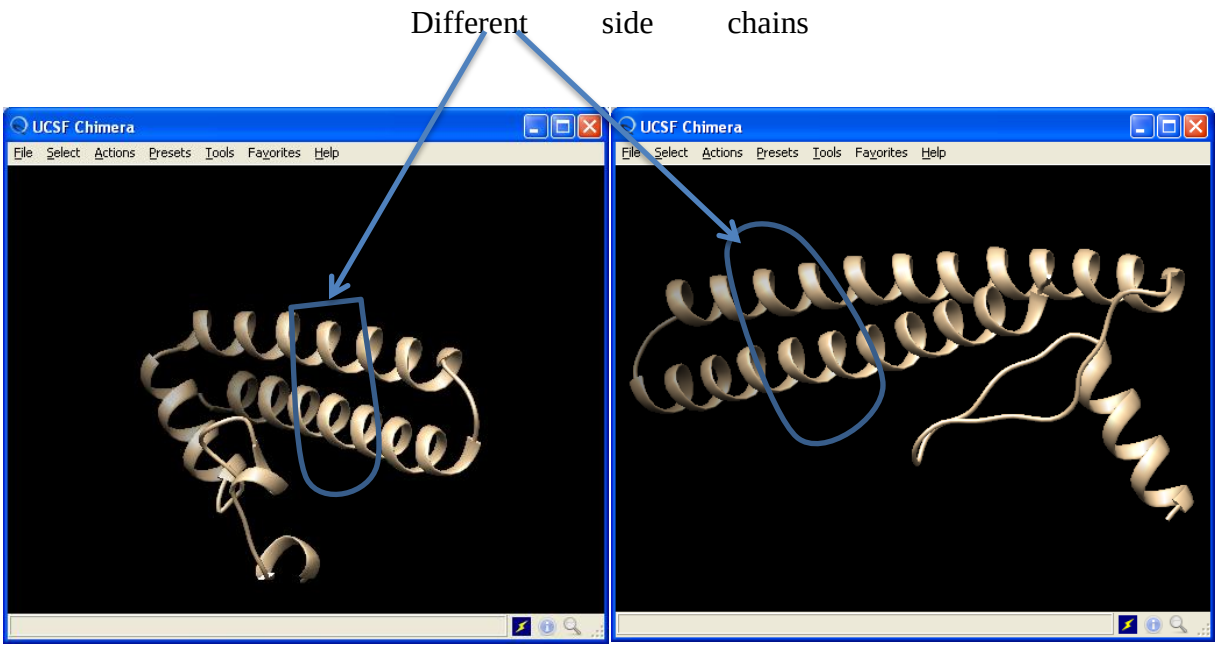


Figure 57: MSP3 3D7 indel fragment secondturn

Figure 58: MSP3 K1 indel fragment showing the occuring in latter residues

The interaction of a different set of side chains could explain the absence of a similar pocket as that identified in the 3D7 strain.

From the results obtained on testing the models generated for MSP6, it was clear that MSP6 was difficult to model and especially the K1 variant. However, due to the similarities observed in the models generated by Rosetta and QUARK, a decision was made to explore them further to see

whether there were outstanding differences between the two strains even though they did not meet the quality marks set out by PROCHECK and ERRAT.

In the case of MSP6 3D7 strain, several pockets were identified. Here, we illustrate the two largest pockets in this strain which are the ones most likely to form the binding pocket [45]. The first pocket had a surface area of 83.4A2 and a volume of 84.6A3 whereas the second had 57.4A^2 and 80.1A^3 in area and volume respectively. The diagram below, figure 59, shows the two pockets.

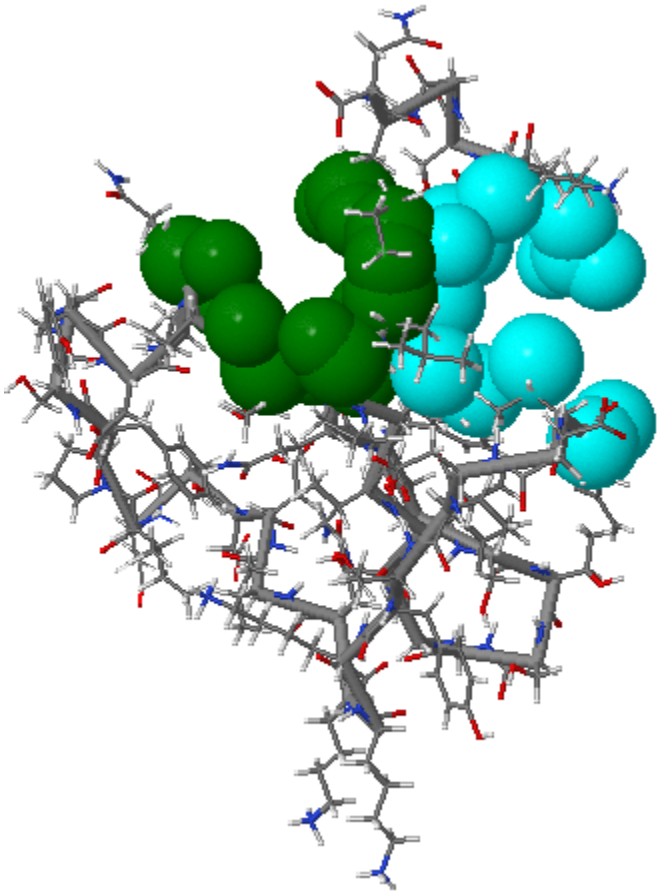


Figure 59: 3D structure of MSP6 3D7 indel fragment showing pockets 1(green) & 2 (cyan spheres) Green (largest) then Cyan (2nd largest) is the order of the sizes of the pockets found

Figure 60: Residues covering the two pockets in MSP6 3D7 indel fragment strain, green for pocket 1 and cyan for pocket 2

Looking at the properties of the residues making up the two pockets, the first pocket consisted of more hydrophobic residues than the second one and therefore more probable.

The largest pocket in the K1 strain had a surface area of 1473.3A^2 and a volume of 3356.7A^3. This is larger than the other pockets we found although it was smaller than one of the largest ligand site found by Liang et al., (1998) that had a volume of 10,048A^3. The second largest had an area of 45.4A^2 and a volume of 26.0A^3. The diagram below, figure R in appendix B, shows that the first pocket is so large when compared with the overall size of the molecule which may also mean that the structure is not correct.

Due to the poor models obtained for MSP6 K1 strain, it was not possible to clearly outline the implication of the insertions in this strain compared to 3D7.

The full sequences were also submitted to I-TASSER which is able to handle longer sequences. Figures S (i-ii) and T (i-ii) in appendix B show the models generated for both strains. MSP3 full models by I-TASSER showed a lot of similarities whereas MSP6 had major differences when it came to the overall conformation.

8. DISCUSSION

The aim of this study was to determine the 3D structures of Merozoite Surface Protein 3 and 6 which have been found to be targets of naturally occurring antibodies [34]. Moreover, the study went further and investigated the differences between the two main allelic variants, 3D7 and K1. Determining the 3D structure of the entire fragments was to be performed *in silico* using comparative modeling. However, due to the absence of good templates, it was found that *ab initio* and threading techniques would have to be used for this purpose. Since *ab initio* only works best for small proteins, the target proteins were divided into fragments.

Three tools, Rosetta, QUARK and I-TASSER were chosen for this task. The first two perform *ab initio* modeling while the last one uses folding to determine the 3D structures of proteins. The three tools were chosen as they have been reported to perform well in CASP experiments [33][34][37]. CASP experiments are experiments performed biannually to test the ability of structure prediction tools to determine the 3D structures of proteins correctly.

Of the three tools, Rosetta performed the best giving models that scored well in the three validation tests performed. I-TASSER gave poor results with some of the models not having any secondary structures. QUARK gave average models in that they did not achieve the thresholds set by the validation tools but were somewhat close. The use of the three tools was done so as to give confidence in the models found and especially where consensus was found. Despite Rosetta giving the best models, its computational time was long with the jobs for longer fragments taking over sixty days to complete. The reason for this is that the conformational search space increases considerably as the number of residues increase.

Two bench mark structures of EBA175 and EBA140 genes, whose structures have already been determined experimentally, were used to help in assessing the quality of the models. This was

through looking at the overall positioning of the X-ray structures with respect to all other structures that exist in the PDB. Their performance in the ERRAT and PROCHECK tests also helped judge how good the models obtained were. Both EBA175 and EBA140 got overall quality factors above 91%, the threshold set by ERRAT and had most of their residues falling in the allowed regions in the PROCHECK test. Therefore, all the models which scored values equal or above these thresholds, were categorized as good models as they tallied with the results of the X-ray structures.

Several fragments were identified for both MSP3 and MSP6. Of the fragments modeled in MSP3, HABP1, HABP3 and the 70aa fragment had good models as well as MSP6F fragment in MSP6. This was established by the fact that most of them achieved the thresholds set out by the validation tools. Their correctness was measured in terms of proper atomic interactions, stereochemical properties as well as their knowledge based energies; by ERRAT, PROCHECK and ProSA-web respectively. Besides that, ProSA-web was able to map the models on to a plot showing their positions in relation with the already existing structures in the PDB. All of the models of the fragments modeled fell in the NMR range in the ProSA-web plot (figure 40) that maps all existing 3D structures that have been determined through either X-ray crystallography or NMR. Falling in the NMR range was a positive achievement since it meant that *in silico* modeling can help answer structure related questions without spending too many resources as well as in a much shorter time frame. The structures obtained using computational methods are also valid despite them being of medium resolution.

One of the benefits of such medium resolution structures is that one is able to identify possible binding pockets that can be used for ligands. This is what was done in this study where the good models were taken through further analysis to establish whether there were any probable ligands

or antibody binding sites. The larger pockets found for the good models were highlighted since research has shown that these are ones that are the better candidates for binding ligands [45]. The confidence of such regions being binding pockets was increased in the cases where the pockets were bordered by hydrophobic residues which are most likely to be found on the inner regions away from the surface of proteins.

This study also set out to establish whether the insertions and deletions in the K1 variant had any effect on the 3D structures of the proteins. After modeling the fragments of the regions with the most differences for both strains, the possible binding regions were identified using CASTp. This was found to be a better way of establishing the differences since visual inspection may not be as informative. In the case of MSP3 where good models were obtained, it was found that the large pocket that was identified in the 3D7 strain had been interfered with and was way smaller in the K1 strain. This could explain the effect of such polymorphisms as those found in the K1 strain. A smaller binding region would mean that a previously designed drug meant to lodge in a given pocket would not be effective as the pocket would already be interfered with.

Some of the fragments were problematic to model. These include MSP3b, HABP2, MSP6BC as well as MSP6 indel fragments. The length of MSP6 indel fragment could explain the poor results. MSP3b and MSP6BC were glycine rich regions. Glycine is a unique amino acid that has a single hydrogen atom as its side chain. This gives it great conformational flexibility meaning that it can fall in regions that are not allowed for other residues [46]. However, more research should be performed to establish the effect of glycine in these structures. MSP6 indel fragment model should also be improved. This could be done by generating more decoys to cater for the increased conformational search space for this larger protein.

This study recommends that the pockets identified for the good models should be subjected to docking experiments to establish their viability.

9. CONCLUSION

In silico modeling has given researchers a cheaper and faster alternative to determining the 3D structures of proteins. Despite the fact that comparative modeling is not possible for proteins with poor templates and some proteins are very large, it is still possible to determine the structures of fragments of such proteins using *ab initio* modeling. This study has shown that *ab initio* is also possible to obtain structures with the accuracy of NMR structures which is a positive finding and especially since NMR has not been fully automated.

MSP3 and MSP6 have regions that have probable ligand and antibody binding pockets that should be subjected to further research including designing probes and taking the experiments further to test their viability *in vivo* and establish which ones are the most effective sites that can be targeted by peptide vaccines and drugs.

One of the possible effects of insertions/deletions observed in the K1 variant could be altering the regions which ligands can bind to, as was observed in the reduction of a probable pocket. The reduced probable pocket identified in K1 could be tested further to find out whether alternative ligands can be designed to target it.

10. RECOMMENDATIONS

Since time as well as computational capacity constraints did not allow generation of more than 10,000 decoys, this study recommends that future studies should consider using clusters to harness the computational capacity. This would highly improve the structures found as well as cover longer regions of the protein.

This study also recommends that structure prediction research should consider using more than one tool so as to verify the validity and correctness of the models obtained. It would also be

beneficial for the developers of the I-TASSER tool to refine their algorithm. This will give the user some kind of consistency in the results obtained and avoid having some cases where no structures are found for a small fragment and yet the same tool finds a structure when given a longer protein containing the same fragment.

11. REFERENCES

[1] World Health Organization, “World Malaria Report 2011,” Geneva, Switzerland, 2011.

[2] N. H. Tolia, E. J. Enemark, B. K. L. Sim, and L. Joshua-Tor, “Structural Basis for the EBA-175 Erythrocyte Invasion Pathway of the Malaria Parasite Plasmodium falciparum,” *Cell*, vol. 122, no. 2, pp. 183–193, Jul. 2005.

[3] Program for Appropriate Technology in Health, “Staying the Course? Malaria Research and Development in a Time of Economic Uncertainty,” PATH, Seattle, 2011.

[4] “Tree of Life Web Project,” *Tree of Life Web Project*, 2012. [Online]. Available: http://tolweb.org/Apicomplexa. [Accessed: 15-Aug-2012].

[5] A. F. Cowman and B. S. Crabb, “Invasion of Red Blood Cells by Malaria Parasites,” *Cell*, vol. 124, no. 4, pp. 755–766, Feb. 2006.

[6] P. Srinivasan, W. L. Beatty, A. Diouf, R. Herrera, X. Ambroggio, J. K. Moch, J. S. Tyler, D. L. Narum, S. K. Pierce, J. C. Boothroyd, J. D. Haynes, and L. H. Miller, “Binding of Plasmodium merozoite proteins RON2 and AMA1 triggers commitment to invasion,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13275–13280, Jul. 2011.

[7] M. Aikawa, L. H. Miller, J. Johnson, and J. Rabbege, “Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite,” *The Journal of cell biology*, vol. 77, no. 1, pp. 72–82, 1978.

[8] D. Keeley, A., and Soldati, “The glideosome: a molecular machine poweringmotility and host-cell invasion by Apicomplexa.,” *Trends Cell Biology*, no. 14, pp. 528–532, 2004.

[9] B. P. Lebrun M, Michelin A, El Hajj H, Poncet J, “The rhoptry neck protein RON4 relocalizes at the moving junction during Toxoplasma gondii invasion,” *Cell Microbiology*, no. 7, pp. 1823–1833, 2005.

[10] B. J. Alexander DL, Mital J, Ward GE, Bradley P, “Identification of the Moving Junction Complex of Toxoplasma gondii: A Collaboration between Distinct Secretory Organelles.,” *PLoS Pathog*, no. 17, 2005.

[11] S. L. Carruthers VB, “Sequential protein secretion from three distinct organelles of Toxoplasma gondii accompanies invasion of human fibroblasts,” *Eur J Cell Biology*, vol. 73, pp. 114–123, 1997.

[12] M. Lamarque, S. Besteiro, J. Papoin, M. Roques, B. Vulliez-Le Normand, J. Morlon-Guyot, J.-F. Dubremetz, S. Fauquenoy, S. Tomavo, B. W. Faber, C. H. Kocken, A. W. Thomas, M. J. Boulanger, G. A. Bentley, and M. Lebrun, “The RON2-AMA1 Interaction

is a Critical Step in Moving Junction-Dependent Invasion by Apicomplexan Parasites," *PLoS Pathogens*, vol. 7, no. 2, p. e1001276, Feb. 2011.

[13] J. Baum, A. G. Maier, R. T. Good, K. M. Simpson, and A. F. Cowman, "Invasion by P. falciparum Merozoites Suggests a Hierarchy of Molecular Interactions," *PLoS Pathogens*, vol. 1, no. 4, p. e37, 2005.

[14] S. Singh, S. Soe, C. Roussilhon, G. Corradin, and and P. Druilhe1, "Plasmodium falciparum merozoite surface protein 6 displays multiple targets for naturally occurring antibodies that mediate monocyte-dependent parasite killing," *Infection and Immunity*, 2005.

[15] "National Institute of Allegy and Infectious Diseases," 2014. [Online]. Available: http://www.niaid.nih.gov/topics/immuneSystem/Pages/features.aspx. [Accessed: 25-Feb-2014].

[16] S. Singh, S. Soe, S. Weisman, J. W. Barnwell, J. L. Pérignon, and P. Druilhe, "A Conserved Multi-Gene Family Induces Cross-Reactive Antibodies Effective in Defense against Plasmodium falciparum," *PLoS ONE*, vol. 4, no. 4, p. e5410, Apr. 2009.

[17] C. Roussilhon, C. Oeuvray, C. Müller-Graf, A. Tall, C. Rogier, J. F. Trape, M. Theisen, A. Balde, J. L. Pérignon, and P. Druilhe, "Long-term clinical protection from falciparum malaria is strongly associated with IgG3 antibodies to merozoite surface protein 3," *PLoS medicine*, vol. 4, no. 11, p. e320, 2007.

[18] S. Singh, S. Soe, J. Mejia, C. Roussilhon, M. Theisen, G. Corradin, and P. Druilhe, "Identification of a Conserved Region of Plasmodium falciparum MSP3 Targeted by Biologically Active Antibodies to Improve Vaccine Design," *The Journal of infectious diseases*, vol. 190, no. 5, pp. 1010–1018, 2004.

[19] S. Soe and E. Al, "Association between protection against clinical malaria and antibodies to merozoite surface antigens in an area of hyperendemicity in Myanmar: complementarity between responses to merozoite surface protein 3 and the 220-kilodalton glutamate-rich protein," *Infection and immunity*, 2004.

[20] M. S. Abual-Rub and R. Abdullah, "A Survey of Protein Fold Recognition Algorithms," *Journal of Computer Science*, vol. 4, no. 9, pp. 768–776, 2008.

[21] K. Ginalski, N. V. Grishin, A. Godzik, and L. Rychlewski, "Practical lessons from protein structure prediction," *Nucleic acids research*, vol. 33, no. 6, pp. 1874–1891, 2005.

[22] L. et al Jooyoung, "Ab Initio Protein Structure Prediction," *Springer Science + Business Media*, 2009.

[23] T. Lazaridis and M. Karplus, "Effective energy functions for protein structure prediction," *Elsevier Science Ltd*, pp. 139–145, 2000.

[24] G. Helles, “A comparative study of the reported performance of ab initio protein structure prediction algorithms.,” *Journal of the Royal Society, Interface / the Royal Society*, vol. 5, no. 21, pp. 387–96, Apr. 2008.

[25] D. Kihara, H. Chen, and Y. D. Yang, “Quality assessment of protein structure models,” *Current Protein and Peptide Science*, vol. 10, no. 3, pp. 216–228, 2009.

[26] D. Constant, “Template-Based Modeling of Protein Structure,” 2011.

[27] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Šali, “Comparative protein structure modeling of genes and genomes,” *Annual review of biophysics and biomolecular structure*, vol. 29, no. 1, pp. 291–325, 2000.

[28] World Health Organization, “10 Facts on malaria,” 2012. [Online]. Available: http://www.who.int/features/factfiles/malaria/en/index.html.

[29] Global Health Group, “Progress Against Malaria,” 2009. [Online]. Available: www.livingproofproject.org. [Accessed: 29-Sep-2012].

[30] K. Ginalski, “Comparative modeling for protein structure prediction,” *Current Opinion in Structural Biology*, vol. 16, no. 2, pp. 172–177, Apr. 2006.

[31] J. Söding and M. Remmert, “Protein sequence comparison and fold recognition: progress and good-practice benchmarking.,” *Current opinion in structural biology*, vol. 21, no. 3, pp. 404–11, Jun. 2011.

[32] M. Remmert, A. Biegert, A. Hauser, and S. Johannes, “HHblits : Lightning-fast iterative protein sequence searching by HMM-HMM alignment,” *Elsevier Science Ltd*, 2011.

[33] L. E. Rodríguez, H. Curtidor, M. Ocampo, J. Garcia, A. Puentes, J. Valbuena, R. Vera, R. López, and M. E. Patarroyo, “Identifying Plasmodium falciparum merozoite surface antigen 3 (MSP3) protein peptides that bind specifically to erythrocytes and inhibit merozoite invasion.,” *Protein science : a publication of the Protein Society*, vol. 14, no. 7, pp. 1778–86, Jul. 2005.

[34] H. B.-T. Claude Quevray, “A novel merozoite antigen of MSP3 identified by cellular antibody cooperative mechanism.pdf,” 1994.

[35] D. Xu and Y. Zhang, “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.,” *Proteins*, vol. 80, no. 7, pp. 1715–35, Jul. 2012.

[36] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V Grishin, and D. Baker, “Structure prediction for CASP8 with all-atom refinement using Rosetta.,” *Proteins*, vol. 77 Suppl 9, pp. 89–99, Jan. 2009.

[37] A. Roy, A. Kucukural, and Y. Zhang, “I-TASSER: a unified platform for automated protein structure and function prediction.,” *Nature protocols*, vol. 5, no. 4, pp. 725–38, Apr. 2010.

[38] R. Laskowski, “European Bioinformatics Institute: Ramachandran Plot,” 1995. [Online]. Available: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?doc=TRUE&pdbcode=n/a&template=doc_procheck01.html. [Accessed: 01-Jul-2013].

[39] M. Pawlowski, M. J. Gajda, R. Matlak, and J. M. Bujnicki, “MetaMQAP: a meta-server for the quality assessment of protein models.,” *BMC bioinformatics*, vol. 9, p. 403, Jan. 2008.

[40] C. Colovos and T. O. Yeates, “Verification of protein structures: patterns of nonbonded atomic interactions.,” *Protein science : a publication of the Protein Society*, vol. 2, no. 9, pp. 1511–9, Sep. 1993.

[41] T. Yeates, “Institute for Genomics and Proteomics.” [Online]. Available: http://www.doe-mbi.ucla.edu/people/yeates/errat. [Accessed: 09-Sep-2013].

[42] T. a. Binkowski, “CASTp: Computed Atlas of Surface Topography of proteins,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3352–3355, Jul. 2003.

[43] T. D. Mulhern, G. J. Howlett, G. E. Reid, R. J. Simpson, D. J. McColl, R. F. Anders, and R. S. Norton, “Solution structure of a polypeptide containing four heptad repeat units from a merozoite surface antigen of Plasmodium falciparum.,” *Biochemistry*, vol. 34, no. 11, pp. 3479–91, Mar. 1995.

[44] R. a Laskowski, J. D. Watson, and J. M. Thornton, “ProFunc: a server for predicting protein function from 3D structure.,” *Nucleic acids research*, vol. 33, no. Web Server issue, pp. W89–93, Jul. 2005.

[45] J. Liang, H. Edelsbrunner, and C. Woodward, “Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.,” *Protein science : a publication of the Protein Society*, vol. 7, no. 9, pp. 1884–97, Sep. 1998.

[46] R. B. Betts, M.j., Russell, *Bioinformatics for geneticists*. 2003.

Server References

[46] PROCHECK: a program to check the stereochemical quality of protein structures J. Appl. Cryst., Vol. 26 (1993), pp. 283-291 by R. A. Laskowski, M. W. Macarthur, D. S. Moss, J. M. Thornton

[47Wiederstein & SipplProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407-W410 (2007).

[48] Y. Zhang. I-TASSER server for protein 3D structure prediction.BMC Bioinformatics, vol 9, 40 (2008).

[49] The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research, Vol. 33, No. Web Server issue. (1 July 2005), pp. W244-W248, doi:10.1093/nar/gki408 by Johannes Söding, Andreas Biegert, Andrei N. Lupas

[50] CASTp: Computed Atlas of Surface Topography of proteins Nucl. Acids Res., Vol. 31, No. 13. (1 July 2003), pp. 3352-3355, doi:10.1093/nar/gkg512by Andrew T. Binkowski, ShaporNaghibzadeh, Jie Liang

[51] Free modeling with Rosetta in CASP6. Proteins, Vol. 61 Suppl 7 (26 September 2005), pp. 128-134, doi:10.1002/prot.20729 by Philip Bradley, Lars Malmström, Bin Qian, et al.

[52] Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega Molecular Systems Biology, Vol. 7, No. 1. (11 October 2011), doi:10.1038/msb.2011.75 by Fabian Sievers, Andreas Wilm, David Dineen, et al.

[53] ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology, Vol. 487 (2011), pp. 545-574, doi:10.1016/b978-0-12-381270-4.00019-6 by Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, et al.

[54] K. Sumathi, P. Ananthalakshmi, M. N. A. M. Roshan, and K. Sekar, “3dSS: 3D structural superposition,” Nucleic Acids Research , vol. 34 , no. suppl 2 , pp. W128–W132, Jul. 2006.

12. APPENDIX A

Table 4: Test Results for HABP1

13. Rosetta					
Cluster No.	**Top Decoys HABP1**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00009435	-60.31	-4.58	100	100
	S_00001806	-56.933	-2.33	100	100
	S_00001116	-55.561	-2.36	96.4	100
2	S_00002126	-55.675	-2.12	96.4	100
	S_00000096	-55.147	-3.35	89.3	100
	S_00006231	-54.896	-2.49	89.3	100
3	S_00009171	-58	-2.41	92.9	72.727
	S_00009798	-53.002	-1.39	100	100
	S_00001997	-52.908	-2.88	96.4	100
I-TASSER			-0.04	96.4	100
QUARK			-2.95	96.4(GFactor: unusual)	100

Table 5: Test Results for HABP2

Rosetta					
Cluster No.	**Top Decoys HABP2**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00001218	-56.731	-2.52	95.8	85.714
	S_00002847	-55.439	-3.05	91.7	95.455
	S_00004882	-55.109	-2.79	100	100
2	S_00000279	-55.465	-2.79	95.8	100
	S_00004831	54.964	-2.86	95.8	95.455

	S_00001688	-54.85	-2.4	95.8	95.455
3	S_00004967	-54.965	-2.45	95.8	100
	S_00002337	-54.49	-1.43	91.7	100
	S_00000128	-53.321	-2.89	100	100
I-TASSER			Error	83.3	54.545
QUARK			Error	95.8	63.636

Table 6: Test results for HABP3

Rosetta					
Cluster No.	**Top Decoys HABP3**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00005426	-86.718	-4.11	100	100
	S_00009685	-86.701	-4.07	100	100
	S_00005285	-84.011	-4.17	97.2	100
2	S_00007390	-86.712	-5.29	100	100
	S_00005678	-85.826	-5.49	88.9	100
	S_00004225	-85.538	-4.5	97.2	100
3	S_00007005	-84.454	-5.42	100	100
	S_00007467	-82.438	-5.22	97.2	100
	S_00008707	-82.086	-5.26	100	93.75
I-TASSER			-4.59	91.7(GFactor:0.1-Low)	96.875
QUARK			-3.2	97.2(GFactor:-0.44)	100

Table 7: MSP3b *Fragment Test Results*

Rosetta					
Cluster No.	**Top Decoys MSP3b**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region**	**Overall Quality(ERRAT)**

				PROCHECK	
1	S_00006397	-38.8	-1.47	95	33.33
	S_00000834	-38.352	-2.21	90	64.706
	S_00007914	-35.74	-1.18	85	47.368
2	S_00004297	-38.133	-2.12	85	100(T)
	S_00005819	-37.674	-2.26	100	89.474
	S_00005403	-37.604	-2.31	90	100
3	S_00004413	-39.048	-1.78	85	57.895
	S_00007176	-38.948	-1.58	90	100
	S_00007552	-37.203	-1.09	90	100
I-TASSER			-2.49	100	37.5
QUARK			-0.93	70(GFactor:-2.10)	52.632

Table 8: 70aa Fragment Test Results

Rosetta					
Cluster No.	**Top Decoys SINGH70**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00002298	-142.673	-3.49	95.3	98.413
	S_00001285	-141.131	-3.1	92.2	93.651
	S_00002102	-140.83	-3.87	95.3	100
2	S_00004128	-142.037	-2.98	93.8	90.476
	S_00001392	-140.753	-3.78	90.6	87.302
	S_00001952	-140.676	-3.06	93.8	98.413
3	S_00000870	-136.607	-2.52	89.1	93.651
	S_00001514	-134.859	-2.54	95.3	93.651
	S_00002534	-134.787	-2.61	92.2	98.413
I-TASSER			-4.74	85.9(GFactor:-0.26)	74.603
QUARK			-4.21	85.9(GFactor:-	77.778

		1.05)Disallowed-6.2	

Table 9: MSP6BC Fragment Test results

Rosetta					
Cluster No.	**Top Decoys MSP6BC**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00000190	-85.438	-2.08	91.9	100
	S_00004658	-85.011	-3.11	100	100
	S_00001185	-84.511	-1.96	86.5	84.211
2	S_00003086	-81.495	-2.25	100	94.872
	S_00001323	-81.055	-3.63	91.9	87.179
	S_00007097	-81.051	-3.49	91.9	76.923
3	S_00001979	-78.349	-3.74	100	94.872
	S_00003869	-78.316	-1.19	91.9	100
	S_00002357	-77.376	-1.98	94.6	76.923
I-TASSER			-5.36	59.5(GFactor:-0.83)	33.333
QUARK			-5.05	81.1(GFactor:-1.9)	53.846

Table 10:MSP6D Test results

Rosetta					
Cluster No.	**Top Decoys MSP6D**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00000562	-86.652	-1.63	91.7	97.778
	S_00008065	-86.58	0.22	91.7	100
	S_00001848	-86.21	-1.49	91.7	95.556
2	S_00001103	-91.083	-1.89	93.8	95.455
	S_00004776	-88.359	-0.96	87.5	100

	S_00007117	-86.663	-2.55	89.6	97.778
3	S_00009610	-88.154	-1.15	93.8	84.444
	S_00004241	-87.97	-1.98	89.6	100
	S_00001344	-86.825	-1.37	87.5	100
I-TASSER			-2.52	68.8(GFactor: -0.5)	0.0
QUARK			-3.71	72.9(GFactor: -1.24)	93.333

Table 11:MSP6F Test results

Rosetta					
Cluster No.	**Top Decoys MSP6F**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00004550	-103.723	-5.39	100	100
	S_00003564	-101.904	-4.57	95.9	100
	S_00005289	-101.062	-5.9	93.9	100
2	S_00001722	-98.972	-5.86	98	100
	S_00000474	-95.857	-5.42	91.8	97.727
	S_00005153	-95.002	-5.8	95.9	100
3	S_00001191	-101.177	-4.91	98	100
	S_00005789	-98.883	-4.34	98	100
	S_00002361	-98.455	-3.46	93.9	93.182
I-TASSER			-4.96	91.8(GFactor:-0.4)	56.818
QUARK			-6.31	93.9(GFactor:-0.74)	93.023

Table 12: MSP3 indel Fragment Test Results

Rosetta					
Cluster No.	**Top Decoys MSP3 3D7**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00001996	-190.571	-3.36	96.3	93.750

	S_00005572	-188.704	-2.3	97.5	100
	S_00008330	-188.6	-3.8	95.1	98.734
2	S_00009691	-185.099	-3.76	93.8	98.750
	S_00004678	-183.805	-3.71	95.1	92.405
	S_00006799	-183.623	-2.28	96.3	100
3	S_00008241	-179.321	-3.06	95.1	97.5
	S_00000068	-177.678	-3.27	97.5	91.139
	S_00001867	-177.534	-2.7	91.4	94.872
I-TASSER			-3.86	90.1(GFactor:-0.08) Disallowed:2.5	86.076
QUARK			-3.06	93.8(GFactor:-0.18)	82.5

Table 13: MSP3 K1 indel Fragment Test Results

Rosetta					
Cluster No.	**Top Decoys MSP3 K1**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00000467	-244.909	-2.48	98.1	100
	S_00007748	-244.89	-2.99	95.4	92.308
	S_00008623	-244.471	-2.42	95.4	90.476
2	S_00008387	-236.684	-2.75	94.4	100
	S_00000171	-236.313	-3.16	95.4	97.170
	S_00003510	-235.86	-3.17	95.4	95.283
3	S_00006979	-238.128	-2.95	97.2	100
	S_00005267	-237.099	-3	95.4	88.679
	S_00006919	-236.078	-3.07	98.1	100
I-TASSER			-2.7	98.1(GFactor:0.15-Low)	83.962
QUARK			-3.19	93.5(GFactor:-0.31)	93.396

Table 14: MSP6 3D7 indel Fragment Test Results

Rosetta					
Cluster No.	**Top Decoys MSP6 3D7**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00001622	-83.209	-3.09	92.5	59.459
	S_00009465	-83.137	-3.12	87.5	94.286
	S_00000439	-82.892	-3.39	95.0	97.5
2	S_00004651	-86.067	-2.95	87.5	88.571
	S_00007879	-83.073	-3.05	82.5	97.5
	S_00000082	-81.764	-4.86	95.0	100
3	S_00008864	-83.424	-3.88	90	100
	S_00008720	-83.39	-5.64	90	100
	S_00009724	-83.297	-3.19	100	97.436
I-TASSER			-4.24	62.5(GFactor:-0.77)Disallowed: 5.0	0.0
QUARK			-3.4	85.0(GFactor:-1.62)Disallowed:2.5	40.476

Table 15: MSP6 K1 indel Fragment Test Results

Rosetta					
Cluster No.	**Top Decoys MSP6 K1**	**Energy Level**	**Z-Score (ProSA)**	**%Most Favored Region PROCHECK**	**Overall Quality(ERRAT)**
1	S_00001436	-175.639	-3.12	82.9	81.053
	S_00005452	-172.714	-3.27	93.9	86.170

	S_00005566	-172.689	-3.95	91.5	78.261
2	S_00000994	-176.446	-2.15	90.2	77.66
	S_00007925	-175.692	-3.32	84.1	63.158
	S_00004910	-173.57	-4.24	90.2	83.516
3	S_00000388	-167.392	-3.85	86.6	61.957
	S_00002215	-166.749	-3.78	90.2	75.556
	S_00006900	-165.898	-2.46	89.0	85.714
I-TASSER			-3.98	63.4(GFactor:-0.98) Disallowed: 3.7	0.0
QUARK			-3.65	72(GFactor:-1.72) Disallowed: 4.9	36.735

14. APPENDIX B

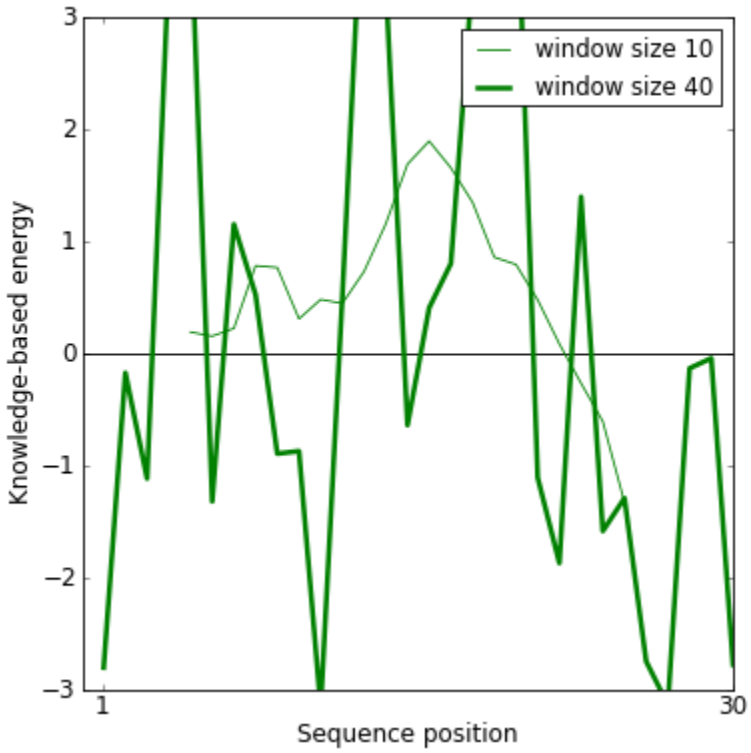


Figure A: Knowledge-based energy plot for Rosetta's HABP1(MSP3) model S_00009435. Due to the short length of the sequence, the thin light green line was used to judge the knowledge-based energy of this fragment. The dark green line is the plot formed from sliding window of size 40.

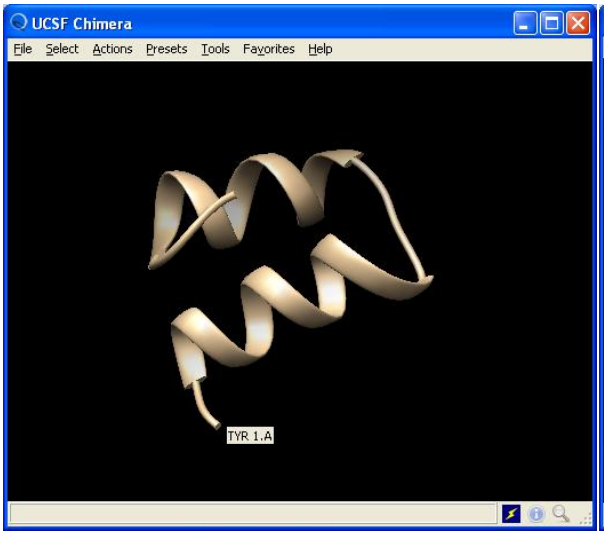


Figure B(i):Rosetta's HABP2 decoy

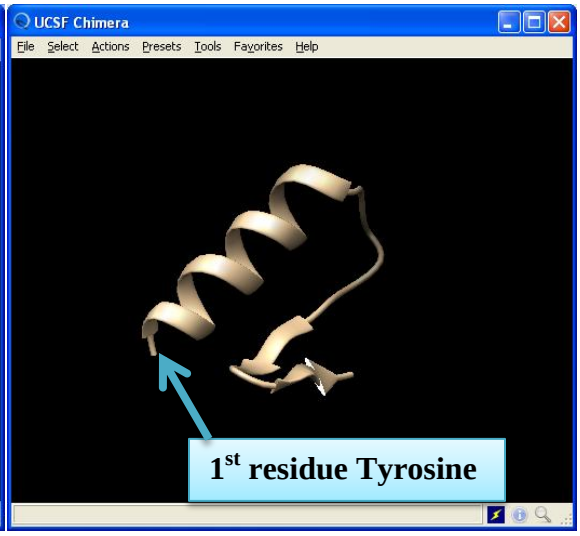


Figure B(ii):QUARK's HABP2 model

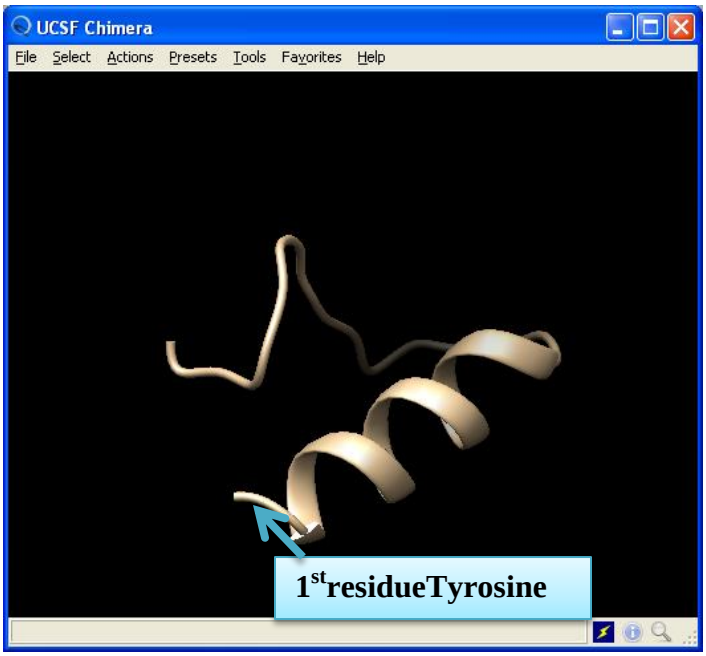


Figure B(iii): I-TASSER's HABP2 model

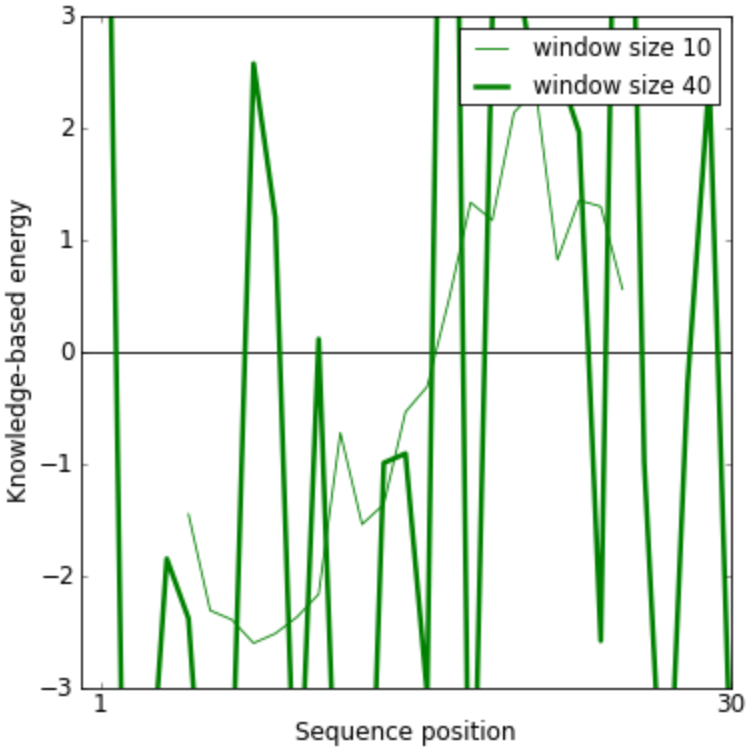


Figure C: Knowledge based energy plot for Rosetta's model S_00004882. We focus on the thinner green line. The region with high energy levels coincides with the part which the three tools did have a consensus on the structure. The dark green line is the plot formed from sliding window of size 40.

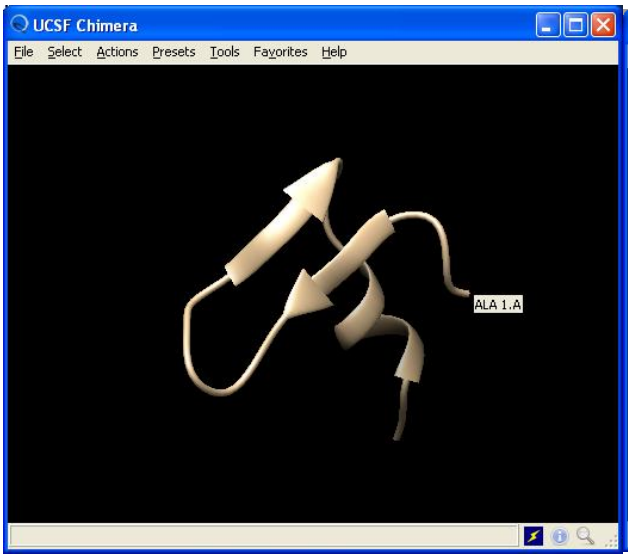


Figure D(i): Rosetta's MSP3b Decoy

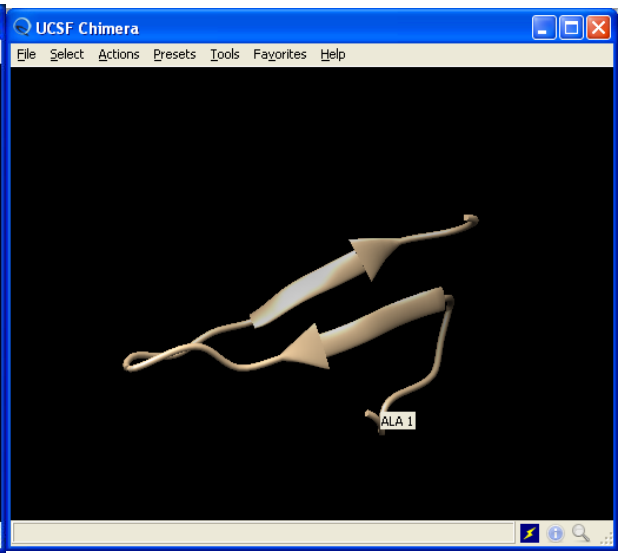


Figure D(ii): QUARK's MSP3b model

Figure D(iii): I-TASSER's MSP3b model

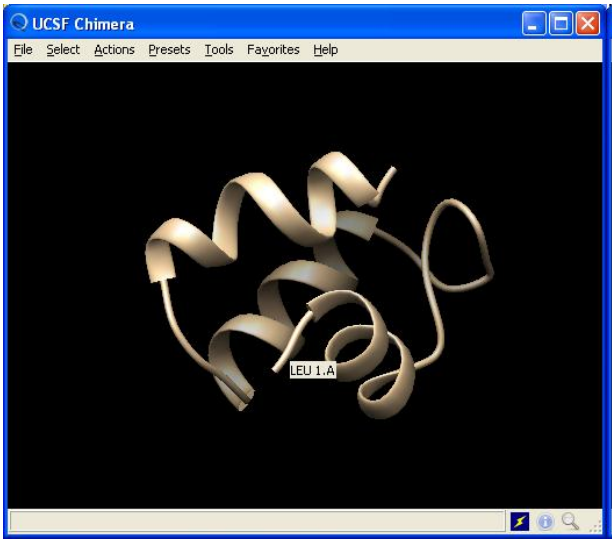


Figure E (i): Rosetta's MSP6BC S_00000190 decoy

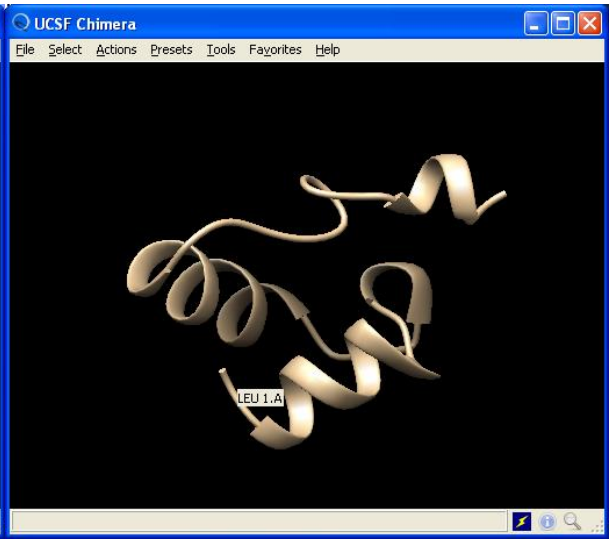


Figure E(ii): Rosetta's MSP6BC S_00004658decoy

Figure E(iii) : QUARK's MSP6BC model

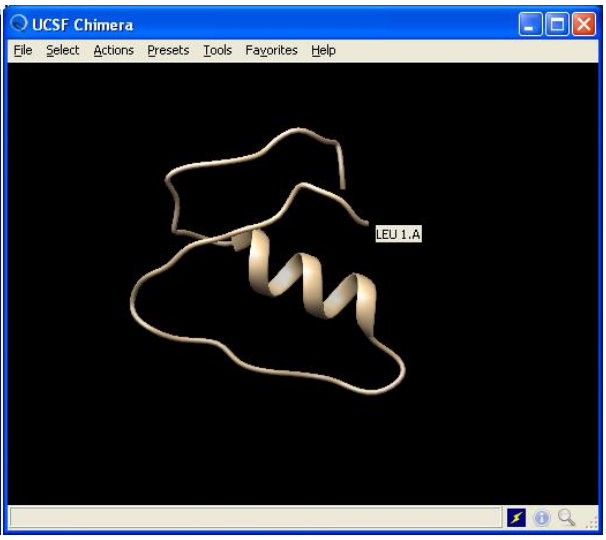


Figure E(iv): I-TASSER's MSP6BC model

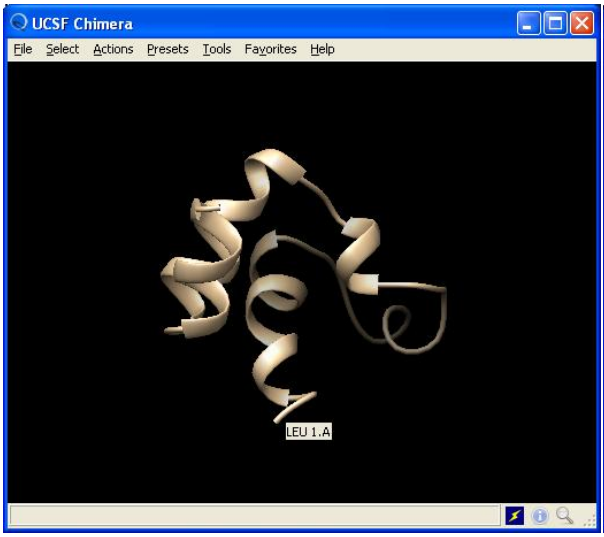


Figure F(i): Rosetta's MSP6D decoy

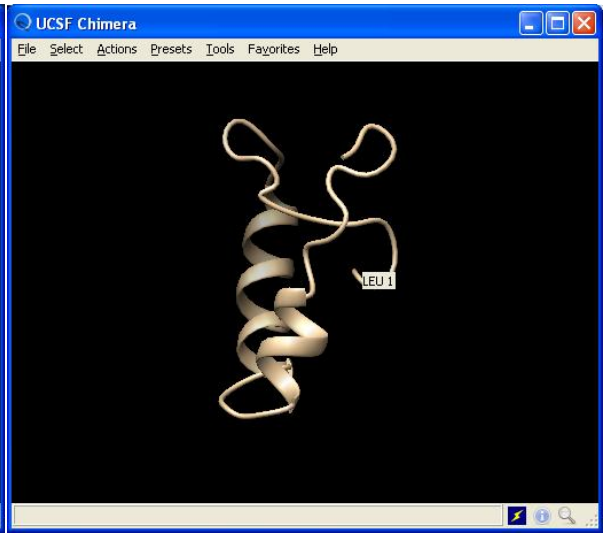


Figure F(ii): QUARK's MSP6D's model

Figure F(iii): I-TASSER's MSP6D model

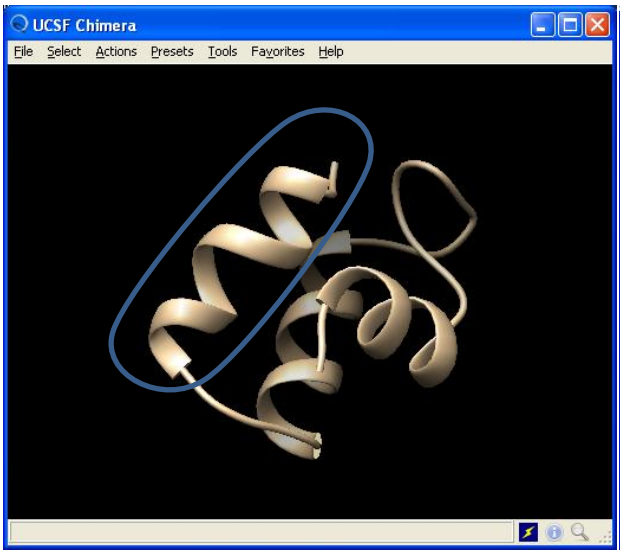


Figure G(i): MSP6BC

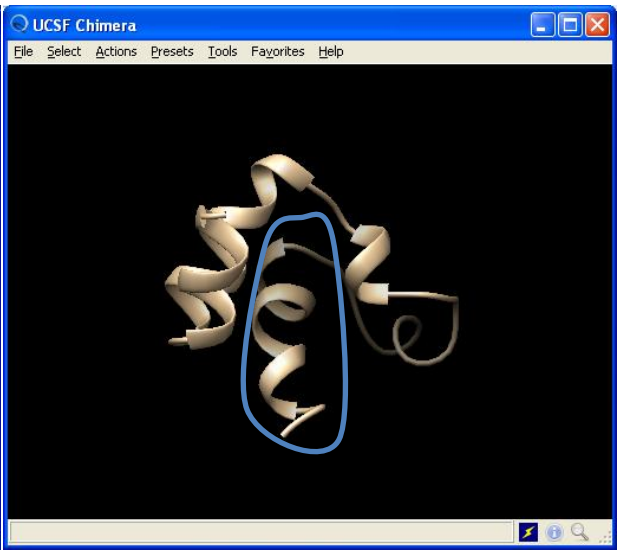


Figure G(ii): MSP6D

The circled regions show the parts of the fragments that overlap in the sequence. The models have been tilted to expose the two points of focus.

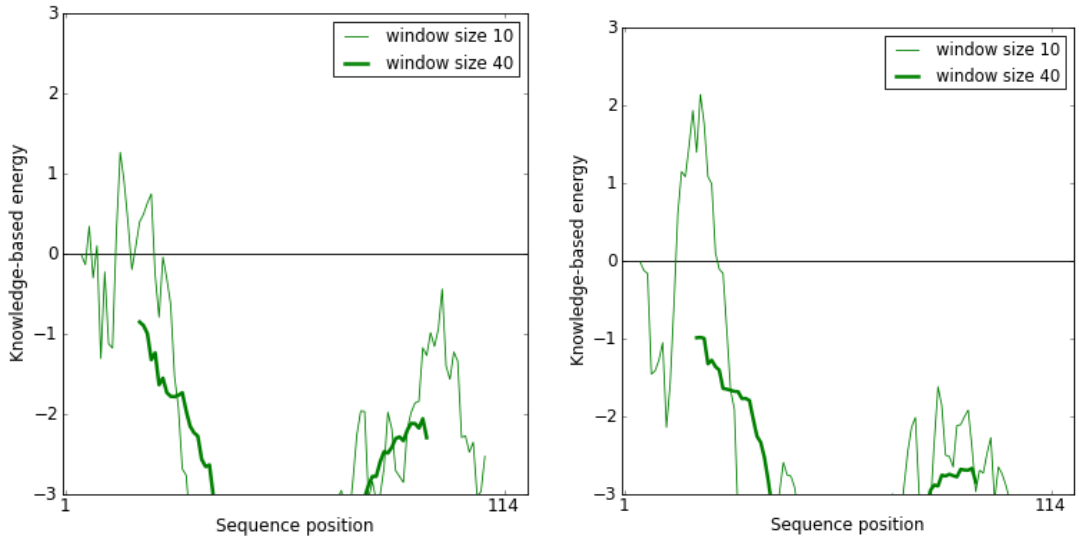


Figure H(i): Knowledge based energy plots for Rosetta's MSP3 K1 model S_00000467(left) and Figure H(ii):S_00006919(Right). The right shows a higher energy spike than the left one(thin green line which represents window of size 10). The dark green line is the plot formed from sliding window of size 40.

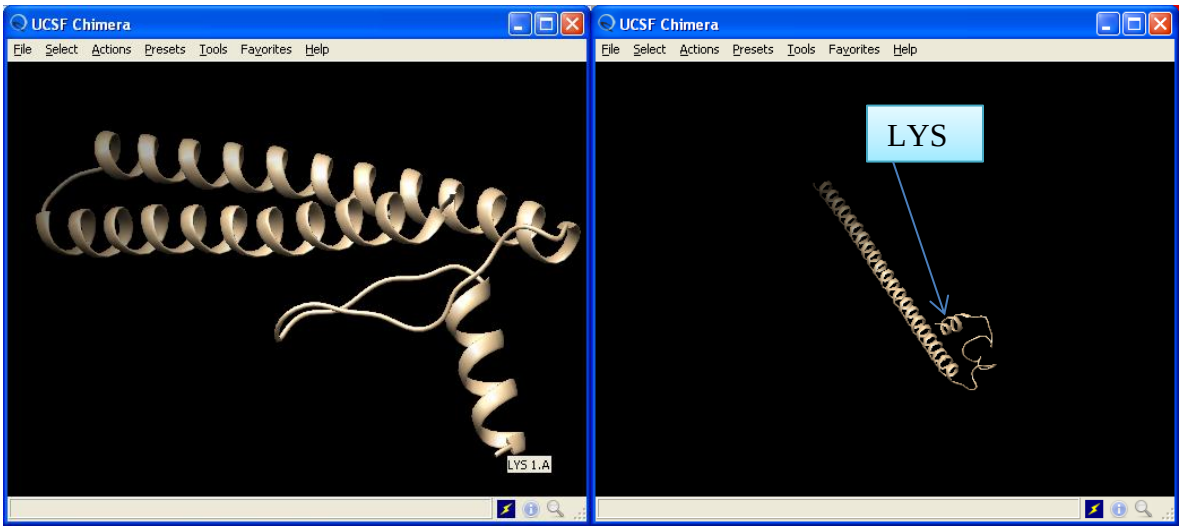


Figure I(i): Rosetta's MSP3 K1 indel Fragment decoy

Figure I(ii): QUARK'sMSP3 K1indel fragment model

Figure I (iii): I-TASSER's MSP3 K1 indel fragment model

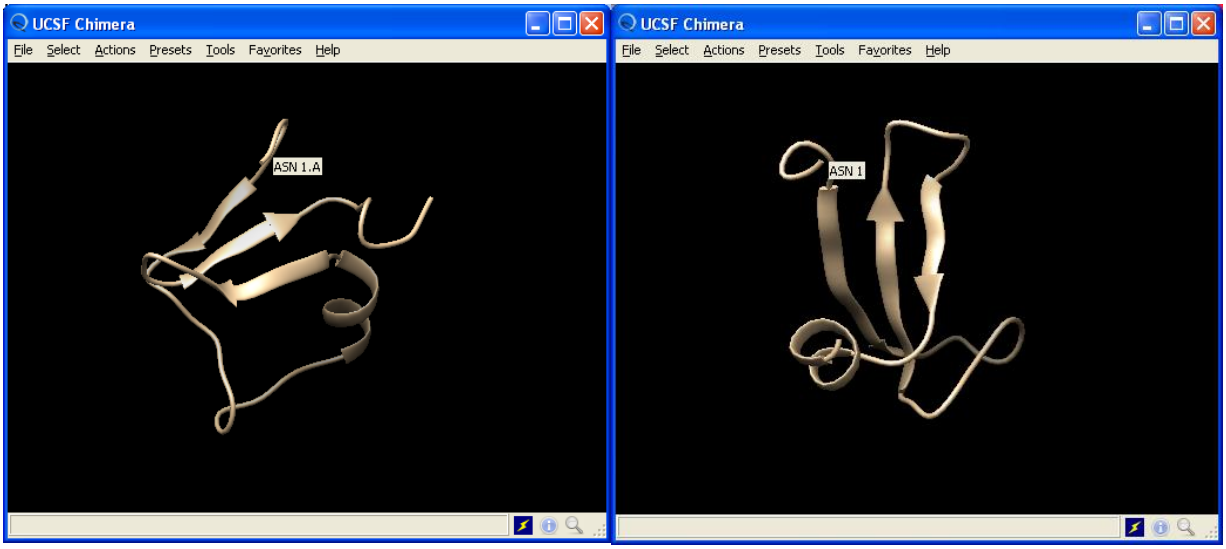


Figure:J(i) Rosetta's MSP6 3D7 indel fragment decoy Figure J(ii): QUARK's MSP6 3D7 indel fragment model

Figure J(iii): I-TASSER's MSP6 3D7 indel fragment model

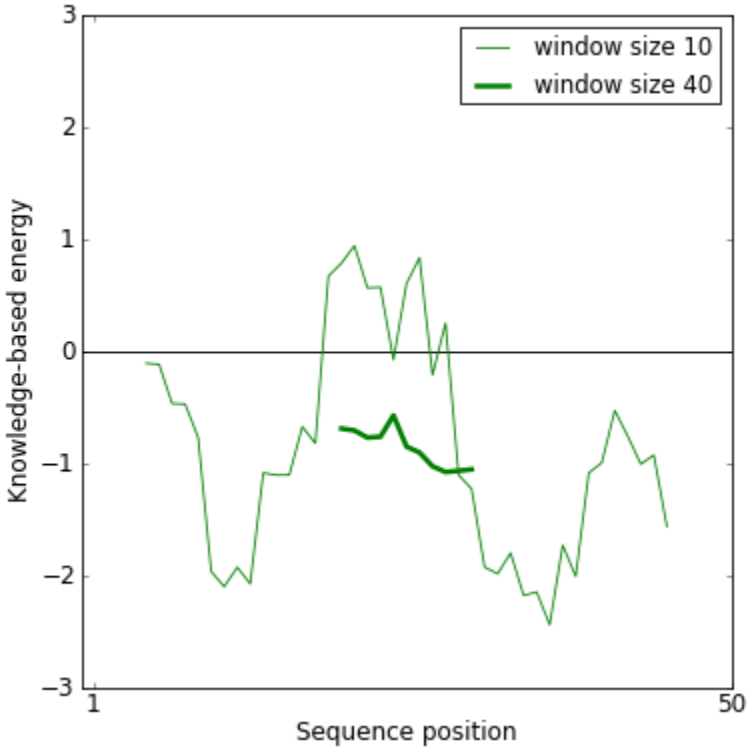


Figure K: Knowledge based energy plot of Rosetta's MSP6 3D7 model S_00009724_0001. We consider the thin green line. Regions above the zero line show regions that should be improved to reduce their energy levels.

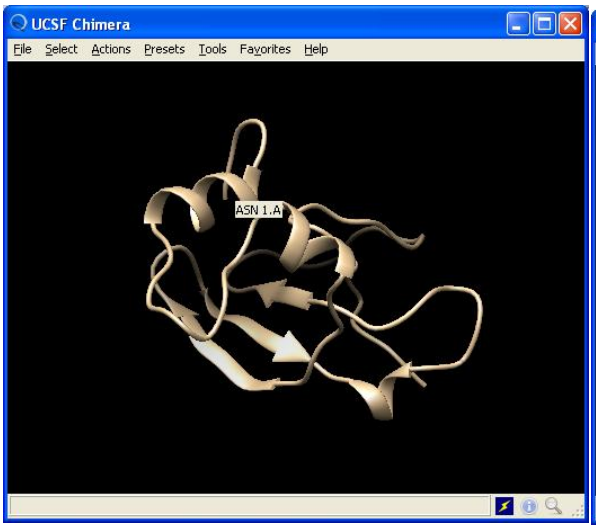


Figure L(i):Rosetta's MSP6 K1 indel fragment decoy

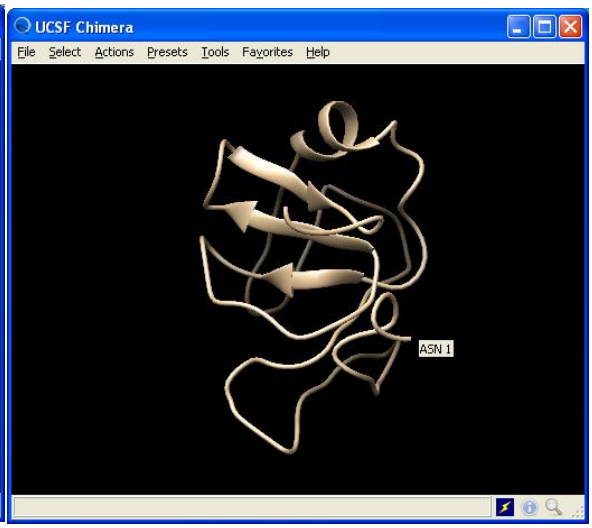


Figure L(ii):QUARK's MSP6 K1 indel fragment model

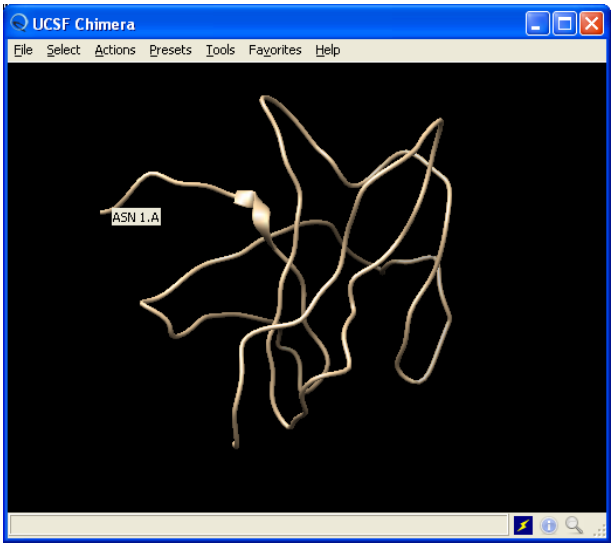


Figure L(iii): I-TASSER's MSP6 K1 indel fragment model

Figure M: Residue coverage in Pocket 1 in 70aa fragment

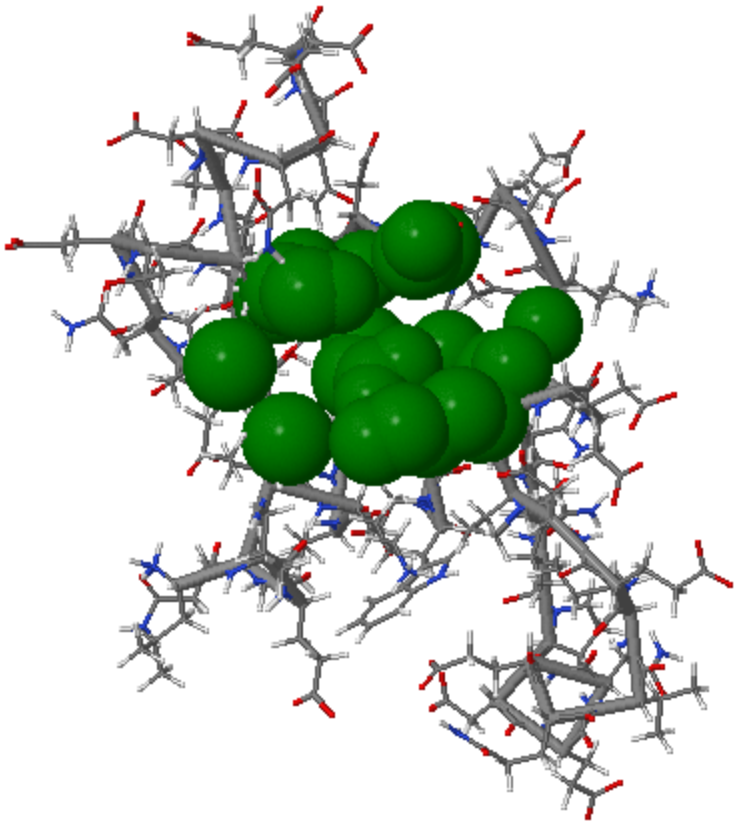


Figure N: 3D structure of MSP6D showing the pocket identified (spheres)

Chain A

1- L L E Q I K I P S W D R N N I P D E N E Q V I E D P Q E D N K D E D E D E E T E T E N L E T E D D N

51- N E E

Figure O: Residues forming the MSP6D pocket highlighted with green

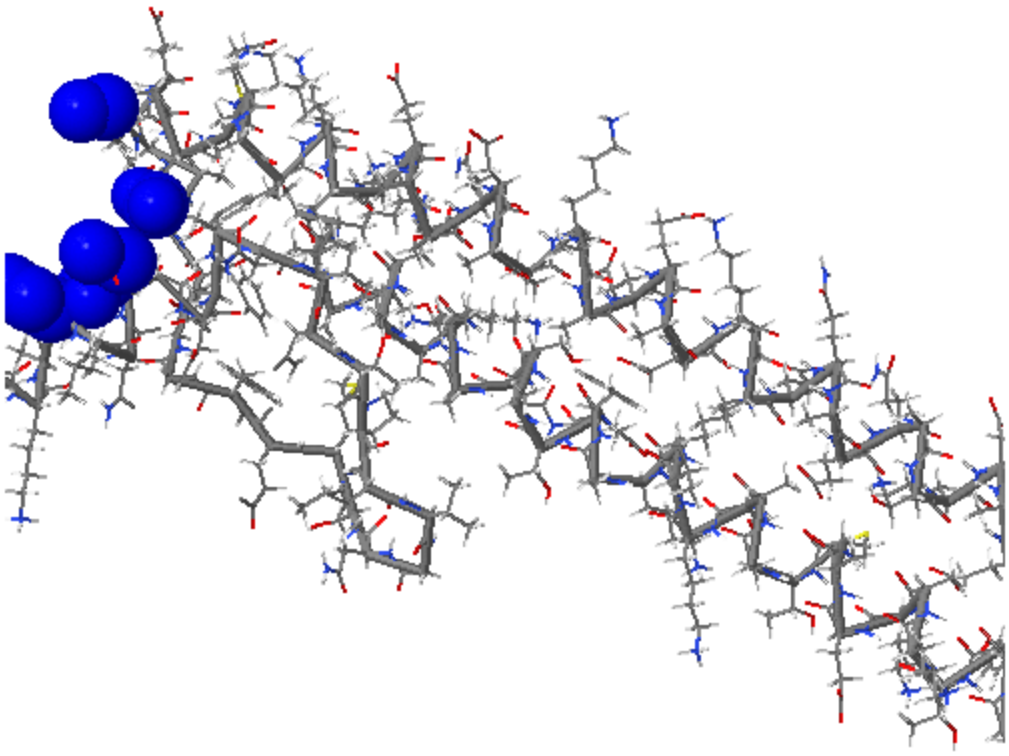


Figure P: 3D structure of MSP3 K1 indel fragment showing pocket 2 (blue spheres)

Chain A

1- K D I K Y E L N E Q N D E N V N T P I V G N S M E F G G F T A D D E K D M E A Y K K A K E A S Q D A

51- E K A A E E A E K A A E Q A E Q A S K D A E K L K E S D E S Y T K A K E A C T A A S K V K K A F E T

101- A S N A K K A A E S A L K T

Figure Q: Residues forming the MSP3 K1indel fragment pocket 2 highlighted in blue

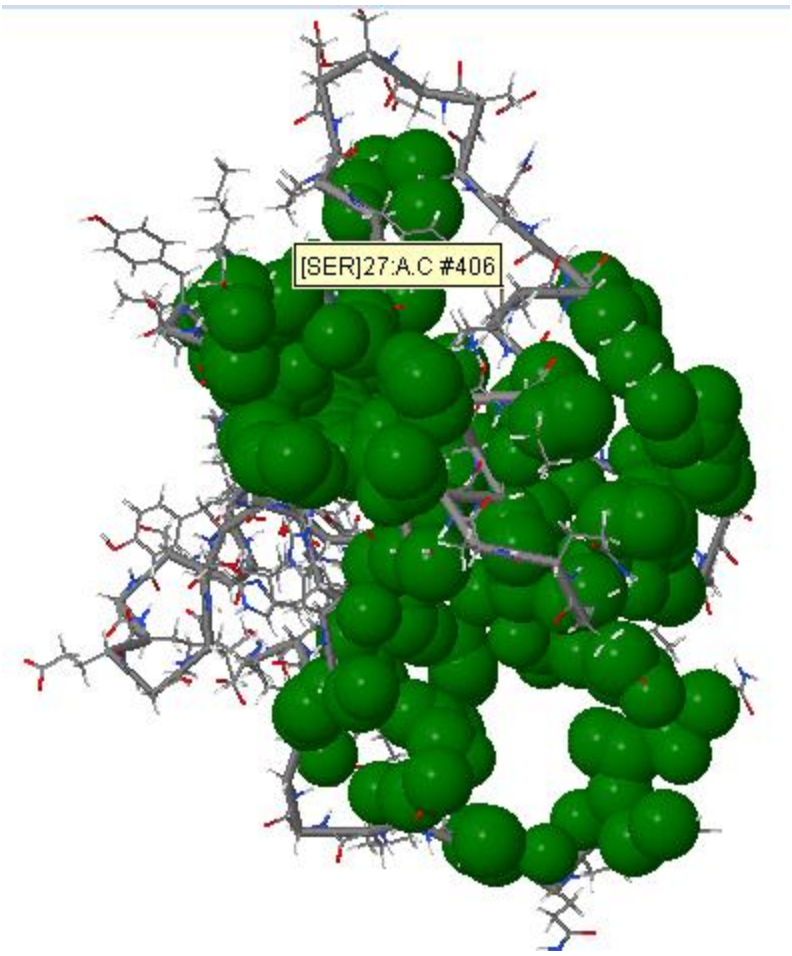


Figure R: 3D structure of MSP6 K1 indel fragment showing pockets (green spheres)

Figure S(i): I-TASSER's MSP3 3D7 model

Figure S(ii):I-TASSER's MSP3 K1 model

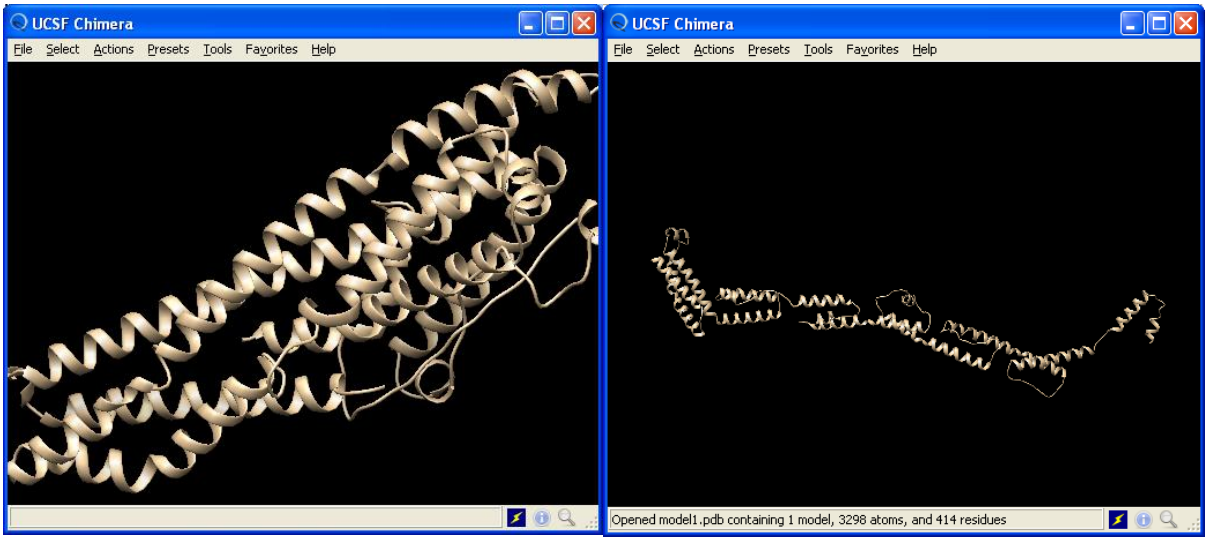


Figure T(i): I-TASSER's MSP6 3D7 model

Figure T(ii): I-TASSER's MSP6 K1 model