Employing Computational Approach to Predict Critical Residues in PFEMP1 (Information Theoretic Measures)

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology

In

Biomedical engineering

Submitted by

Kalpana

(DTU/13/M.Tech/395)
Delhi Technological University, Delhi, India

Under the guidance of

Professor Dr. Bansi D. Malhotra

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

DECLARATION

I hereby declare that the work entitled “**Employing Computational Approach to Predict Critical Residues in PFEMP1 (Information Theoretic Measures)**” submitted in partial fulfillment of the requirement for the award of the degree of **Master of Technology** in Biomedical Engineering from Delhi Technological University (formerly DCE), is an authentic record of my work carried out under the guidance of **Professor Dr. Bansi D Malhotra.**

The information and data enclosed in this dissertation is original and has not been submitted anywhere for honoring of any other degree.

Date: Signature:

Place:

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **"Employing Computational Approach to Predict Critical Residues in PFEMP1 (Information Theoretic Measures)"**, submitted by KALPANA **(DTU/13/M.Tech/395)** in partial fulfilment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of my work carried out under the guidance of **Professor Dr. Bansi D Malhotra.**.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Professor D. Kumar
Head of the Department
Department of Bio-Technology
Delhi Technological University
(Formerly DCE, University of Delhi)

Professor Dr. Bansi D. Malhotra
(Project Mentor)
Department of Bio-Technology
Delhi Technological University
(Formerly DCE, University of Delhi

ACKNOWLEDGEMENT

The completion of any project requires a lot of guidance and support from many people and I am extremely fortunate to have got this all along during the course of my project work.

I take this opportunity to express my profound gratitude and deep regards to my guide **Mr. Sayan Chatterjee,** Department of Biotechnology, University school of Biotechnology, Guru Gobind Singh Intraprastha University, Delhi**,** for his exemplary guidance, monitoring and constant encouragement throughout the course of this work. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

I am highly indebted to **Professor Dr. Bansi D Malhotra,** Department of Biotechnology, Delhi Technological University, and Delhi, for his support, guidance and constant supervision as well as for allowing me to carry out my work from University school of Biotechnology, Guru Gobind Singh Intraprastha University, Delhi.

/I am also thankful to **Prof. D. Kumar**, HOD, Department of Biotechnology, Delhi Technological University, for his constant co-operation.

I also take this opportunity to express my heartiest thanks to lab members for their able guidance and help which they gave me in every step of my project period.

Finally I thank God and my parents for being always beside me and their unflagging blessing and support without which this training would not have been possible.

Date - With Regards,

Place - Kalpana

CONTENTS

List of Figure

List of Tables

List of Abbreviations

RE	Relative Entropy
CRE	‘ Cumulative Relative Entropy
PFEMP	Plasmodium falciparum Erythrocyte Membrane Protein
DBL	Duffy Binding – Like
CIDR	Cysteine Rich Inter Domain Region
HMM	Hidden Markov Model
CD36	Cluster Of Differentiation 36
MAFFT	Multiple Alignment using Fast Fourier Transform
DOPE	Discrete Optimized Protein Energy
YASARA	Yet Another Scientific Artificial Reality Application
HADDOCK	High Ambiguity Driven protein-protein DOCKing

Employing Computational Approach to Predict Critical Residues in PFEMP1 (Information Theoretic Measures)

Kalpana

Delhi Technological University, Delhi, India

ABSTRACT

PFEMP1 (*Plasmodium falciparum erythrocyte membrane protein*) is an important target for defensive immunity and is involve in the pathology of malaria through its ability to adhere to host endothelial receptors. PFEMP1 has specific domains which are essential in its cytoadherence function. PFEMP1 binds to CD36, an 88kDa glycoprotein found in several cell types including platelets, monocytes, dendritic cells, and micro vascular endothelial cells. This cythoadherence of PFEMP1 to CD36 receptor is due to a particular domain called CIDR1α domain. This cytoadherence function of CIDR1α to CD36 receptor is support by various conserved motifs which may be targeted to disrupt the parasite cytoadherence system. The knowledge of these critical residues can lead to the better considerate of the molecular basis of diseases which arise due to modified protein functions. This knowledge also would play a vital role in rational protein engineering and drug designing. So, in-depth knowledge of structure and function of various severely important residues of CIDR1α is necessary for effective drug design and vaccine designing.

Herein, we will be employing computational approaches to predict fold and functionally vital residues of CIDR1α domain. Traditional use of conservation scores are enhanced with Information Theoretic scores – the Relative Entropy (RE) and Cumulative Relative Entropy (CRE) calculated from Multiple Sequence Alignment (MSA) have been shown to adversely identify residues important for the fold and specificity of the protein.

These methods were enforced to predict residues of CIDR1α with high RE and CRE to be fold and functionally important respectively.

CHAPTER 1

INTRODUCTION

Plasmodium *falciparum* is the most noxious of all other species of microorganisms and accountable for maximum human deaths (Warrell DA *et.al.,* 1990). The different pathological characteristic of Plasmodium *falciparum* contagion is that the parasite infected erythrocytes attach to host endothelium and are eventually isolated from the blood circulation. This allows the parasite to avoid spleen-dependent killing and persist for further transmittance. However, this may produce lethal difficulty in case isolation of infected erythrocyte takes place in the vital organs.

Another adaption of parasite to keep away from immune response is by increase variability by consistently replacing the antigens expressed on the surface that are disclose to the host immune system. Malaria parasites contain a wide family of genes for mutant antigens called *var* genes that play a critical role in the differential expression of these antigens. *Var* gene family are grouped into three subgroups UpsA, UpsB and UpsC this grouping is done according to chromosomal localization of their 5' transcribed region (Lavstsen T *et.al,* 2003; Yvonne K *et.al,* 2010). These genes code for two exons: the extracellular region and putative transmembrane domain; and the second encode the acidic terminal segment or ATS that is theorise to anchor PFEMP1 at knobs (Su XZ *et.al.,* 1995).

PFEMP1 contain two different adhesive modules: the Cysteine-rich Inter Domain Region (CIDR1α) (Baruch DI et.al., 1997; Smith JD et.al., 1998) and Duffy binding-like (DBL) domain which is illustrate as adhesive region in different Plasmodium proteins which are involved in erythrocyte invasion (Adams JH et.al. 1990; Sim BK et.al., 1994). DBL domains bind to distinct molecules like intercellular adhesion molecule 1 (ICAM-1) (Smith JD et.al., 1994), chondroitin sulphate A (CSA) (Rowe JA et.al., 1997) and undefined heparin sulphate molecule on erythrocyte surface (Chen Q et.al., 1998). The CIDR1α domain binds to CD36 receptors. Difference in the PFEMP1 primary sequence is such that this function of the protein-binding to CD35 – remains the same, while the epitopes associated to antigen are changed. Most of the parasites isolated have the capacity, to bind to the CD36 receptor. This gives us a hint that there must be some main residues which remain conserved in each variant, which would enable the application of methods to extract the structural and functional residues that exists in the protein family. These residues can be largely classified as single site residues which include (a) residues that are conserved throughout a protein family thereby responsible for the fold of the protein termed 'fold specific' and (b) residues that are differentially conserved along many subfamilies within a protein family which are responsible for substrate or functional specificity in the protein subfamily.

From a structural point of view, fold specific residues are those that are responsible for the general scaffold common across a specific protein family and random mutations of residues on these scaffold results in paralogous proteins with a distinct functional or substrate specificity. The knowledge of these crucial residues can lead to the well understanding of the molecular basis of diseases which appear due to altered protein functions. This knowledge also would play a critical role in logical protein engineering (Baker D et.al. 2010) and drug designing (Tramonotano A. Et.al., 2005). Moreover, the direct involvement of these critical residues with substrates/ligands and their involvement in maintaining the stability of protein can be efficiently clarified. These structural and functional constraints implant in a particular protein family are efficiently reflected by their Multiple Sequence Alignments (MSA). A multiple sequence alignment perform as a historical record of amino acids variability that has been assemble at each sequences positions of a protein family throughout the course of evolution. Once a protein has evolved to a useful level of functionality, a most of the mutations are selectively neutral at the molecular level and do not influence the function and fold of the proteins, since those mutations which are harmful provide selection pressure for residue conservation (Kimura et.al., 1983). Thus, the residue conservation in a multiple sequence alignment of a protein and its homolog's specify the importance of the residues for maintaining the structure and function of proteins. Traditionally used conservation scores can recognize the fold specific residues that are conserved throughout the alignment (Valdar WS et.al. 2002). However, these are not efficient in identifying differentially conserved and co-evolving residues in the alignments. Further, using wide sets of sequences permit for the efficient distinction of functionally crucial residues from those related to phylogenetic conservation, which is a standard error from conservation patterns derived from smaller collections of sequences from closely associated organisms. Therefore in order to solve the drawbacks of this traditional scoring techniques we have made use of information theory (Christoph A et.al., 2004) and consider measures that can accurately differentiate these crucial signals with that of the background noises. We have also carried out Hidden Markov Models to estimate the probabilities (Sriastava P.K et.al. 2007) of amino acids which in turn will be used as predictors in different information theoretic measures.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Malaria

Malaria is the problem especially in developing and tropical country; there is 300 – 500 million case of malaria in each year and 2-3 million deaths annually, mainly in children. In human malaria is caused by the protozoan genus Plasmodium, these are four types of Plasmodium species, namely, P. *falciparum,* P. *vivax,* P. *ovale,* and P. *malariae.* The complex life cycle of malaria parasites is completed by passing through both the anopheline mosquito and human, with asexual reproduction occurring in the mammalian host and sexual reproduction in the anopheles mosquito vectors (Fig.1) (White et.al., 1998). Infection in humans begins when anopheles mosquito (female) which carry plasmodium parasite, bites the human, with which the sporozoite stage of parasite gets transmitted. Sporozoites get injected through the bloodstream these sporozoites travel to the liver via blood and take up residues in hepatocytes (liver cell). In the liver the sporozoites multiply asexually and become many merozoites collectively known as schizonts, the hepatocytes then burst releasing it into the blood and thus occur in 7-10 days later. In the blood all the merozoites invade erythrocytes and multiply again until the cell burst and releasing into bloodstream. After several asexual cycles merozoites can invade RBC and instead of replicating they can develop into sexual form of the parasites, which are plasmodium gametocytes, therefore, if another unaffected anopheles mosquito (female) comes and bite to particular infected human it was suck of these gametocytes. It digest gametocytes which will they allow it to develop into many sex cell called gametes, then they fused together to form zygotes which forming oocytes where we have sporozoites begins to developed, they can multiply then they cause oocytes ruptured releasing the sporozoites, these moved into the salivary gland of the mosquito leads to injected into the another human so that male and female gametes enter sporozenic cycle producing more pathogenic sporozoites. So that female mosquito can infect human causing malaria.

2.2 Cytoadherence

The pathogenicity of P. *falciparum* increases due to its unique ability to adhere to the capillary and post capillary venular endothelium, this process is known as cytoadherence (Lus S. A et.al; 1971, MacPherson G. G et.al; 1985). Cytoadherence gives the survival advantage to the parasite; the major advantage is the escape from the clearance by the spleen. This safeguards the parasite from the immune response.

Cytoadherence resulting the sequestration of infected erythrocytes (IRBC) leads to the alteration in the microcirculatory blood flow, the metabolic dysfunction, and as a consequence, many of the manifestations of severe falciparum malaria components (Ho M et.al; 1990).

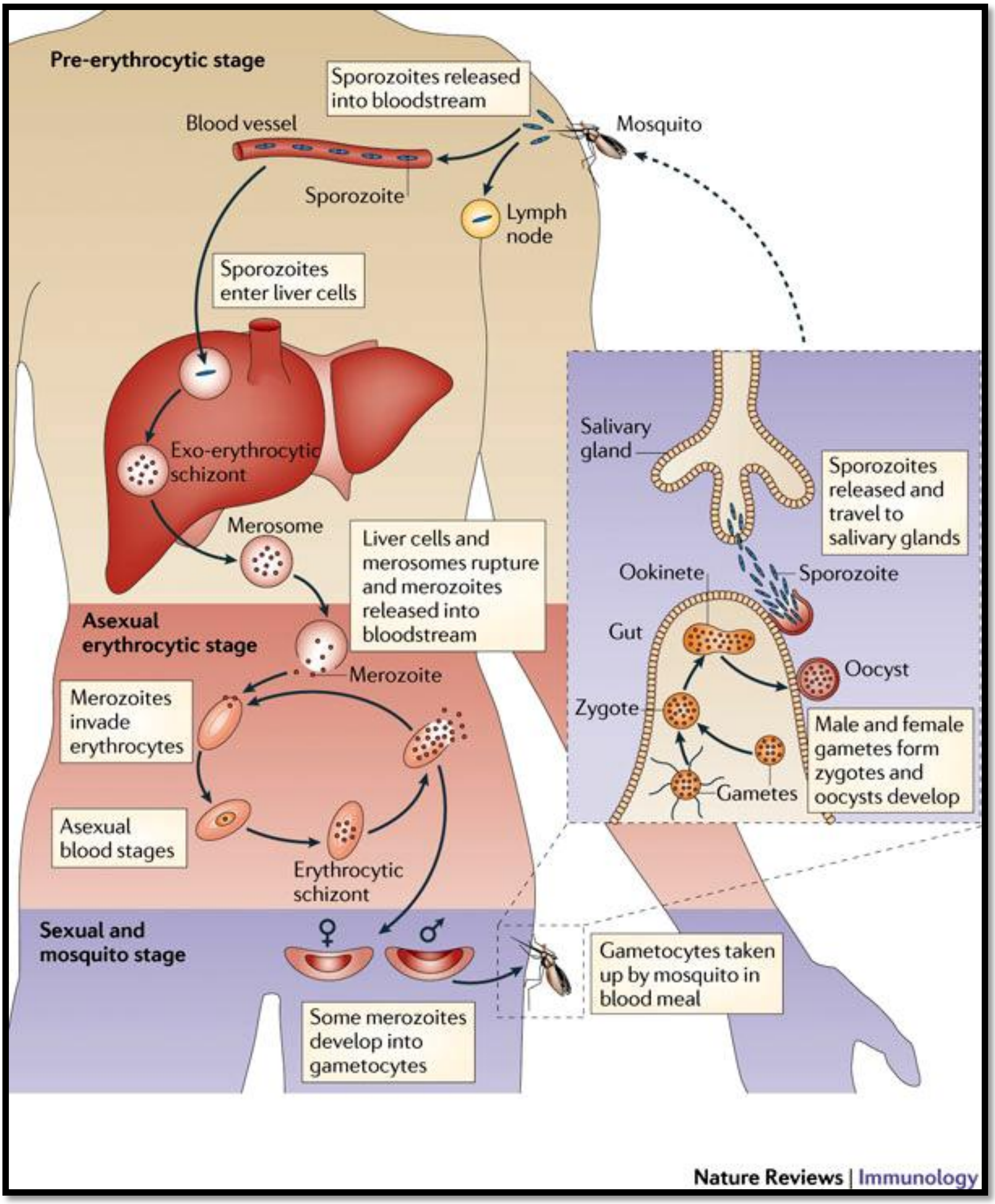

Pre-erythrocytic stage
Sporozoites released into bloodstream
Blood vessel
Mosquito
Sporozoite
Lymph node
Sporozoites enter liver cells
Exo-erythrocytic schizont
Merosome
Liver cells and merosomes rupture and merozoites released into bloodstream
Asexual erythrocytic stage
Merozoite
Merozoites invade erythrocytes
Asexual blood stages
Erythrocytic schizont
Sexual and mosquito stage
♀
♂
Some merozoites develop into gametocytes
Gametocytes taken up by mosquito in blood meal
Salivary gland
Sporozoites released and travel to salivary glands
Ookinete
Sporozoite
Gut
Oocyst
Zygote
Gametes
Male and female gametes form zygotes and oocysts develop
Nature Reviews | Immunology

Figure 1: Malaria Life Cycle (Robert W.S *et.al;* 2011)

2.3 Plasmodium falciparum erythrocyte membrane protein 1

During the merozoite stage of Plasmodium *falciparum*, the Plasmodium *falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) is expressed on the surface of infected RBC and mediates adhesion of infected erythrocytes (IE) to the various host cells on the vascular lining. (Baruch DI et.al; 1995, Su XZ et.al; 1995). PfEMP1 is encoded by ~60 var genes, a majority of which are situated in the sub telomeric regions while remaining ~40% are found centrally in the chromosomes (Lavstsen T. Et.al; 2003, Kraemer SM et.al; 2003). To large extent hyper-variable *var* gene repertoire generated by frequent meiotic ectopic recombination in the mosquito abdomen, this is possible by alignment of *var* genes in the nuclear periphery (Taylor HM et.al; 2000, Freitas-Junior LH et.al; 2000). Most of the PfEMP1 (even proteins with the same domain architecture) display less than 50% amino acid sequence identity between individual domains (Kraemer SM et.al; 2007). Several human cell receptors involved in adhesion of PfEMP1 are CD36 and intercellular adhesion molecule 1 (ICAM-1), although no consensus on association between receptor binding and severe malaria has been reached (reviewed in Rowe JA et.al; 2009). PfEMP1 has previously been described as composed of several domains N-terminal segments (NTS), Duffy binding-like (DBL) domains, Cystine rich inter-domain regions (CIDR1α), C2 domains, one transmembrane region (TM) and the acidic terminal segment (ATS). The CIDR1 domains have been divided into three classes: CIDR1α, β and γ (MacPherson G. G et.al; 1985).

Among these only CIDR1α binds to the CD36 receptor (Baruch DI et.al; 1995). CIDR1α domain consisting of three regions, which are minimal CD36 binding region denoted M2, flanked by less conserved M1 and M3 regions (Smith JD et.al; 2000). Several CIDR1α class domains have been found to mediate binding to the human CD36 receptor. Furthermore, CIDR1α domains have been found to bind immunoglobulin M and PECAM-1 (Chen Q et.al; 2000).

2.4 Information Theoretic Measures

In information theory, entropy is a measure of the uncertainty associated with a random variable. In this context, the term usually refers to the Shannon entropy (H) that is one of the simplest and most common information theoretic scores which estimate the diversity of a system and sequence variability at a position in the alignment (Sander S et.al;1991, Kullback S et.al;1991). It is defined for a column i as:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Where M = 20, the number of possible amino acids.

Pi = amino acid frequency distribution in column i of the alignment.

The unit of entropy is ‘bit’. However, it is immaterial which logarithm is used as all logarithms are proportional. Shannon entropy would be highest for a completely variable column where every amino acid is equally likely whereas it would be zero for a completely conserved column.

2.4.1 Relative Entropy

Relative Entropy (RE) also called the Kullback-Leibler divergence (KL divergence) as the direct divergence between two distributions (Kullback SK et.al; 1951). It is used to compare two probability distributions (Cover T et.al; 2009) and is to measure the difference of an amino acid distribution P from some background distribution Q. The relative entropy was calculated according to the formula:

$$\textbf{Relative Entropy} = \sum_{\mathbf{n=1}}^{\mathbf{20}} \mathbf{P(n) log} \frac{\mathbf{P(n)}}{\mathbf{Q(n)}}$$

Where the summation is over all amino-acid types n in the alignment; P(n) is the probability of the amino acid n in the column; Q(n) is the background probability of the amino acid n in all columns of the multiple sequence alignment, which is calculated as the probability of finding an amino acid n in all available protein sequences ie, protein sequences in Swiss-Prot database. It is always greater than or equal to zero. The relative entropy reached its maximum value if the amino acid alone is observed which is the least probable according to the background distribution.

2.4.2 Cumulative Relative Entropy (CRE)

The cumulative relative entropy of an alignment is simply the sum of the information / relative entropy of all of the positions. Hannenhelli and Russel represented the CRE method for identification of Specificity Determining Residues (SDRs) given an alignment and its classification into subfamilies (Hannenhali S et.al;2000). For alignment position i, the CRE is calculated as:

$$\mathbf{RE_i(y_1 - y_2) = \sum_{x=1}^{20} p_i\,(x,y_1) log \frac{p_i(x,y_1)}{p_i(x,y_2)}}$$

Where $p_i(x,y_1)$ and $p_i(x,y_2)$ denote the probabilities of amino acid x in the subfamily y and the rest of the subfamilies at position i of the alignment respectively.

The method was implemented using HMM and further HMM profiles were also used to predict the subfamilies of the unclassified proteins. Authors performed a large scale assessment of their method by applying PFAM collection of multiple sequence alignment partitioned into subfamilies by using Swiss-Port functional assignment. The good performance of the method has been shown by the fact the predicted SDRs were in close agreement with the experiment.

CHAPTER 3

AIM AND OBJECTIVE

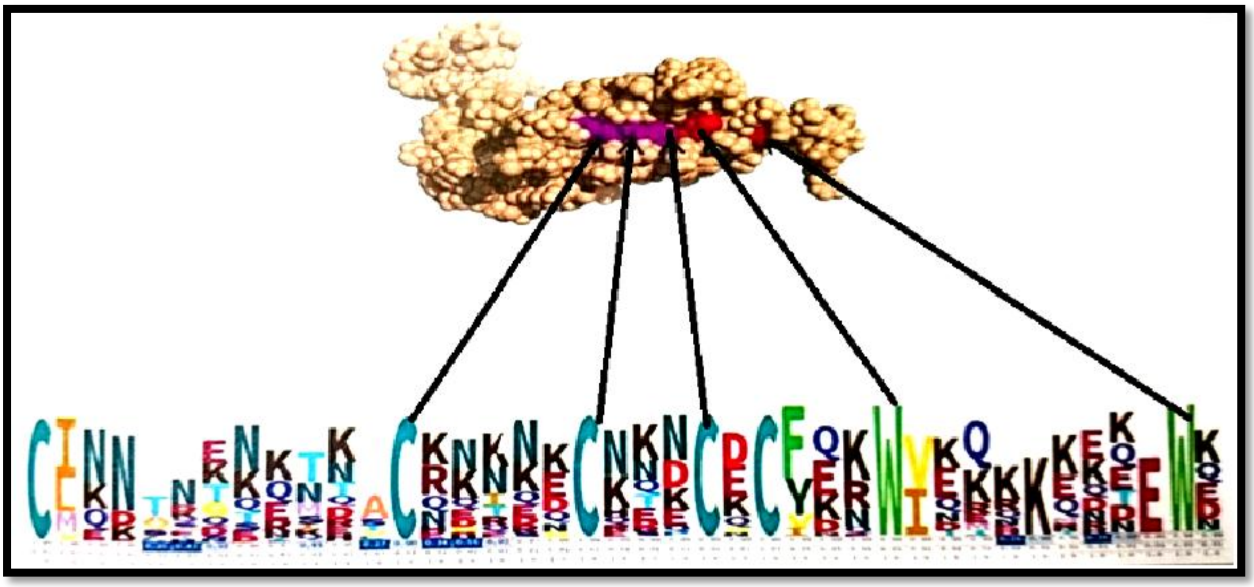
This study is aimed at the identification, modification and implementation of information theoretic measures for the prediction of critical residues from a sequence analysis perspective. To maintain the fold and function of a protein family, various groups of residues follow different conservation patterns across the different subfamilies. The residues responsible for maintaining the fold and for conferring specificity are collectively termed “critical” residues. These conservation patterns identified from the multiple sequence alignment of the proteins, organised into various subfamilies.

From a sequence analysis standpoint, fold determining residues are conserved throughout the family while specificity determining residues can be interpreted as differentially conserved residues of different subfamilies. In order to predict fold determining residues Kullback – Leibler distance (Relative Entropy) is used.

Protein function can be studied hierarchically, e.g., the broader function of a GPCR family is single transduction, but at a finer level the binding sites of these single transducing molecules tend to vary across subfamilies giving rise to different signal transduction pathway activation.

Specificity determining residues can be interpreted as differentially conserved residues of different subfamilies. In order to predict these functionally relevant conservations of each subfamily distinctively from the conservation associated universal across all the subfamilies we have developed a Cumulative Relative Entropy approach to identify residues responsible for a specific function by not only considering the differentially conserved residues but also those residues that are conserved only in the concerned subfamily.

a)



b)

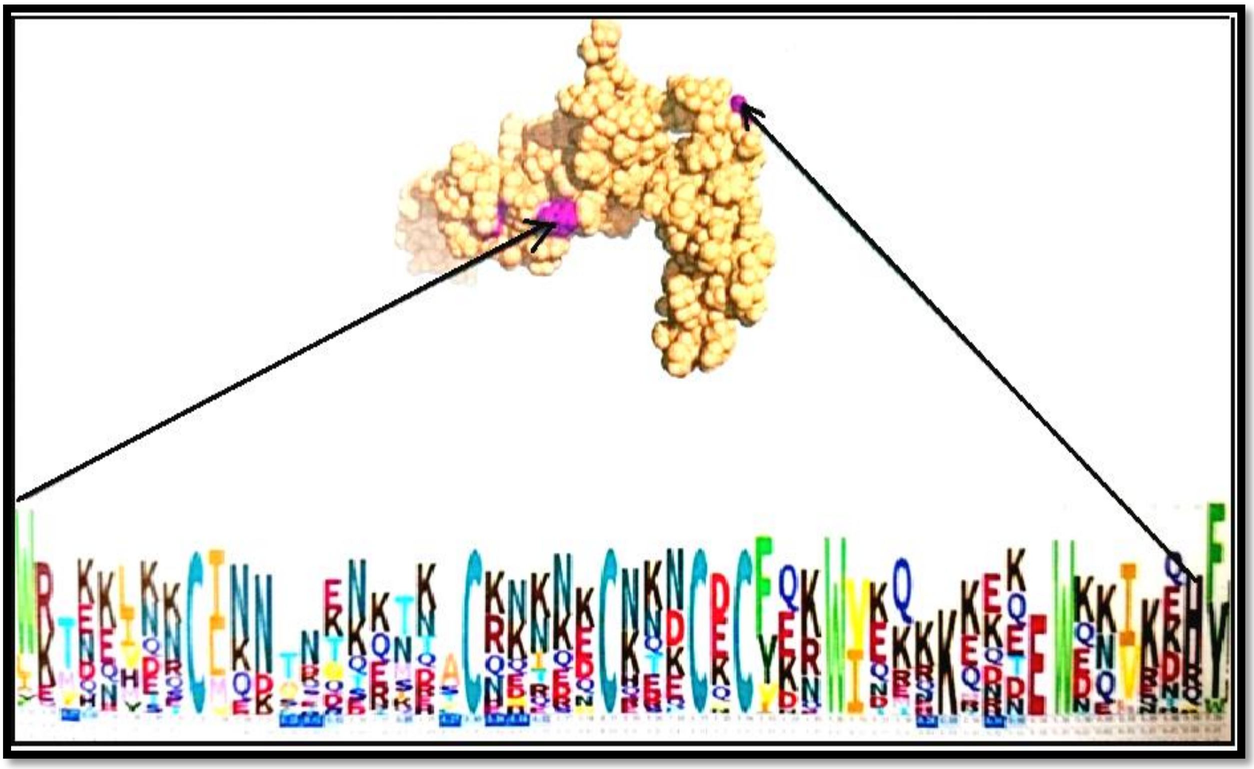


Figure 2: Schematic description of sequence conservation and its implication on protein function. (a) Show residues conserved across the alignment. These are responsible for the broad function or thermodynamic integration of the protein. (b) Patterns of differential conservation are seen in the case of residues conserved only within the subfamily and are presumed to be responsible for its specific function.

CHAPTER 4

MATERIAL AND METHODS

4.1 CIDR1α domain sequences

The protein sequences of PFEMP1 were obtained from the CIDR1α. These sequences were trimmed to get CIDR1α and CIDR1β domain using local pair alignment using Mafft. This was done by performing local pair alignment MC179 sequence and trimming.

CIDR1α domain was obtained by performing local pair alignment with MC179.

4.2 Building Multiple Sequence Alignment

Mafft is used for the local pair option in multiple sequence alignment of all CIDR1α sequences, using the option providing an iterative refinement method incorporating local pairwise alignment. We used standalone version of Mafft.

4.3 Prediction of Fold and Function specific residues

4.3.1 Calculation of Relative Entropy (RE) Scores

As explained in Section 4.1, Relative Entropy (RE) scores are calculated by comparing the amino acid probability distribution for each column of the multiple sequence alignment with that of the background distribution. The background probability distributions for all the 20 amino acids were calculated directly from the alignment.

Where alignment.fasta is the input alignment file from Section 2 and background.txt is the output file with the background frequencies for the 20 amino acids specific for the alignment in alphabetic order. The Relative Entropy scores for all columns in the alignment are calculated.

This script makes use of the HMMER package 2.3.2 and module hmmer.pm to calculate the position specific information for all columns of the multiple sequence alignment, alignment.fasta and compare it with the background probabilities present in background.txt. The Relative Entropy scores are written in the output file alignment_RE. This file has two fields: alignment column positions and RE scores. All the columns of the alignment were accounted for irrespective of the number of gaps present. Therefore it was necessary to weight the columns based on the number of gaps present in each columns which was incorporated by a scaling factor given as:

S_i = sum (Non gap sites in column i) / No of sites in column i

$RE_i = RE_i \text{ X } S_i$

Where S_i is the scaling factor for each column i in the MSA.

Mapping of the RE scores to a specific protein sequence in the alignment is carried out using mapping_protein.pl, where alignment_Res is the scaled RE scores, alignment.fasta is the alignment file and id is the sequence id of the protin sequence to be mapped. The output file alignment_Remapped contains 4 fields: alignment column positions, scaled RE scores, amino acid of the protein sequence id corresponding to each column position and sequence positions.

4.3.2 Calculation of Cumulative Relative Entropy (CRE) Scores

The alignment sequences can be grouped separate subfamilies by preparing a list of sequence id that belong to a particular subfamily of interest as one list (subfamily.fasta) and the rest of sequence ids for all the other subfamilies as another (rest.fasta) list.

Where alignment.fasta is th alignment file, list contains the ids separated into two groups that are to be studied (subfamily, rest). The outputs are the alignment of sequences for each group under study.(subfamily.fa, rest.fa).

RE_subfamily.pl builds hmm profiles and extracts out the probabilities from HMM profiles using hmmer.pm. Similar to that of RE calculation where the comparison is done with the background frequencies but here, RE_subfamily.pl compares the probability distribution of the subfamily under study (subfamily1.fa) with the rest of subfamilies (rest.fa). The output file is subfamily12_RE.

Similarly,

As mentioned earlier in Section 3 scaling.pl makes correction for gaps in the scores obtained.

Similarly RE_family.pl builds and extracts probabilities from HMM profiles using hmmer.pm for calculation of Relative Entropy Scores. Here RE is subjected for the concerned subfamily. The output file is subfamily_RE which is later scaled.

Differentially conserved residues for each subfamilies, and those residues that are present in one subfamily but absent in others can be efficiently extracted by CRE calculations. The intuitive procedure is to weight more for these residues from the others, thereby giving this formula for CRE calculation:

$\mathrm{CRE}_i = (\mathrm{RE12}_i + \mathrm{RE21}_i) \text{ x } (\mathrm{RE1}_i)$

Where i = each columns of the multiple sequence alignment. These CRE scores are later normalized resulting in CREs scores. It requires the module REcontext.pm.

The output file alignment_mapped CRE contains four fields: alignment column positions, CREs scores, amino acids of the protein sequence id corresponding to each column position and sequence positions.

4.3.3 Generating Null Models

To assess the significance of the results obtained through RE, REcontext and DCA it was necessary to compare the results with that obtained from the Null model. The Null models were generated by randomizing the data sets, which in our case is the sequence alignment files.

Randomizing was done keeping in mind the following criteria's (Rost B et.al;1993):

- The randomize data should nullify the property established in the native alignment.
- The gap integrity of the alignment should be maintained as it was necessary to maintain the topological stacking of various compartments of the protein sequences.

4.3.4 Null Models for RE calculation

The native alignment for RE calculations establishes the property of residue conservations across certain columns of the Multiple Sequence Alignment. These conservations as explained in Sections (4.1, 4.2) are necessary to reflect the fold and function specific residues in the protein family under study. So a random alignment intuitively should reside in the residue columns that night be conserved by chance. More importantly the properties retained in the native alignment are single site constraints. Therefore randomizing was done by shuffling each rows/sequences of the multiple sequence alignment keeping the gaps of the alignment undisturbed as they are placed such that an optimal alignment of the sequence is produced.

Where alignment.fa is the input alignment file and output.fa is the row shuffled randomized alignment.

The above procedures for RE and CRE calculations (Section 3, 4, 5 and 7) were later implemented on the randomized datasets, which apart from predicting random fold specific, function specific and co-evolving residues would also identify the threshold values that are to be set to obtain significant predictions.

Where input files are RE_results were that obtained from the native alignment, rand_RE_results were that obtained from the random dataset, z_results are those residues that are significant greater than the threshold value obtained from the null model.

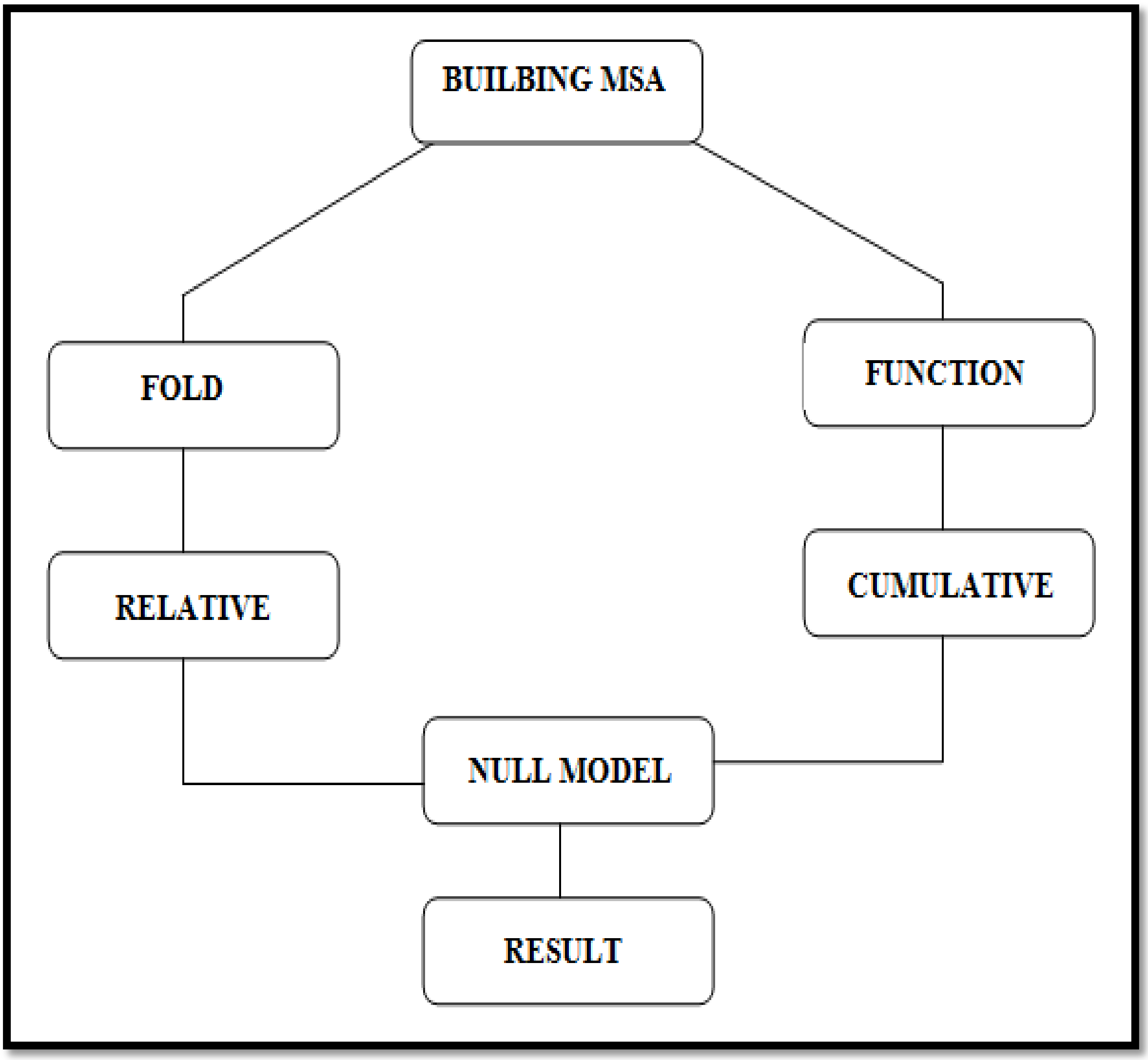


Figure 3: The work follows of various methodologies that were implementing in this thesis.

4.4 Modelling CD36

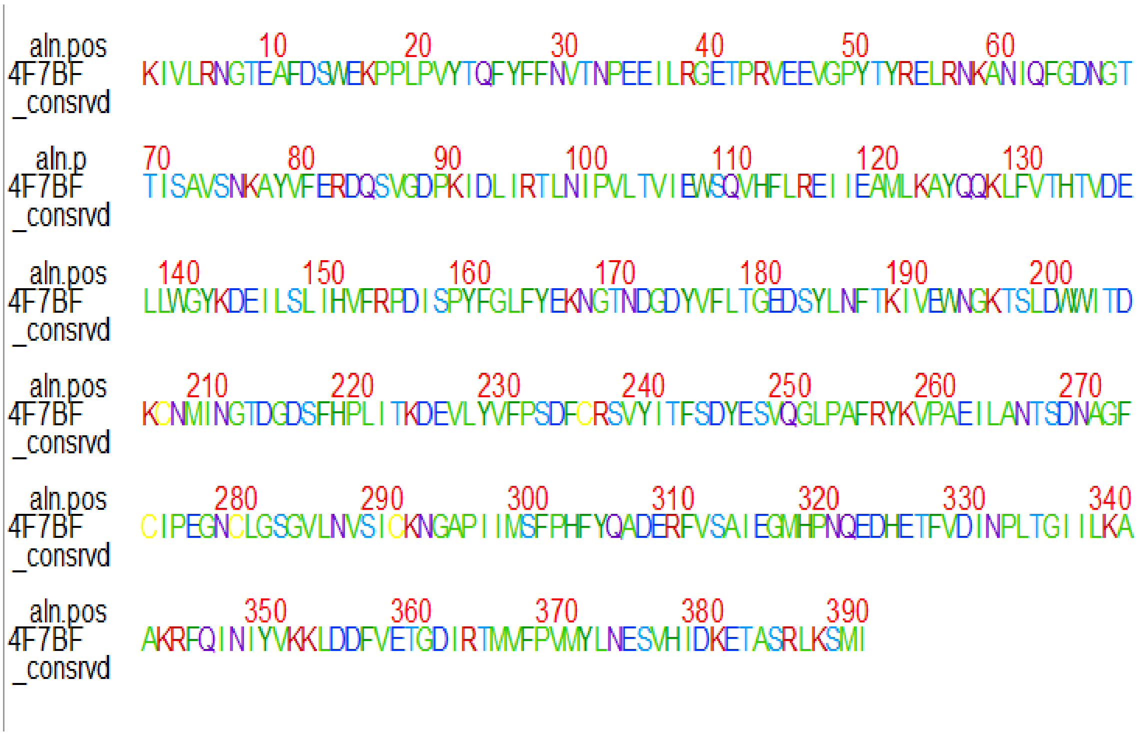
Homology model of CD36 sequence was generated using Modeller9v7Package. All the scripts used for modelling are available at the website.

https://salilab.org/modeller/tutorial/basic.html. The steps taken for building the structure model of CD36 sequence are as follows:

4.4.1 Template Identification

The template used for modelling is 4F7B.pdb is from *Homo sapiens* (Neculai D et.al; 2013).

```
_aln.pos         10        20        30        40        50        60
4F7BF
KIVLRNGTEAFDSWEKPPLPVYTQFYFFNVTNPEEILRGETPRVEEVGPYTYRELRNKANIQFGDNGT
 _consrvd
 _aln.p   70        80        90       100       110       120
130
4F7BF
TISAVSNKAYVFERDQSVGDPKIDLIRTLNIPVLTVIEWSQVHFLREIIEAMLKAYQQKLFVTHTVDE
 _consrvd

 _aln.pos  140       150       160       170       180       190
200
4F7BF
LLWGYKDEILSLIHVFRPDISPYFGLFYEKNGTNDGDYVFLTGEDSYLNFTKIVEWNGKTSLDWWITD
 _consrvd

 _aln.pos    210       220       230       240       250       260
270
4F7BF
KCNMINGTDGDSFHPLITKDEVLYVFPSDFCRSVYITFSDYESVQGLPAFRYKVPAEILANTSDNAGF
 _consrvd

 _aln.pos      280       290       300       310       320       330
340
4F7BF
CIPEGNCLGSGVLNVSICKNGAPIIMSFPHFYQADERFVSAIEGMHPNQEDHETFVDINPLTGIILKA
 _consrvd

 _aln.pos        350       360       370       380       390
4F7BF     AKRFQINIYVKKLDDFVETGDIRTMVFPVMYLNESVHIDKETASRLKSMI
 _consrvd
```

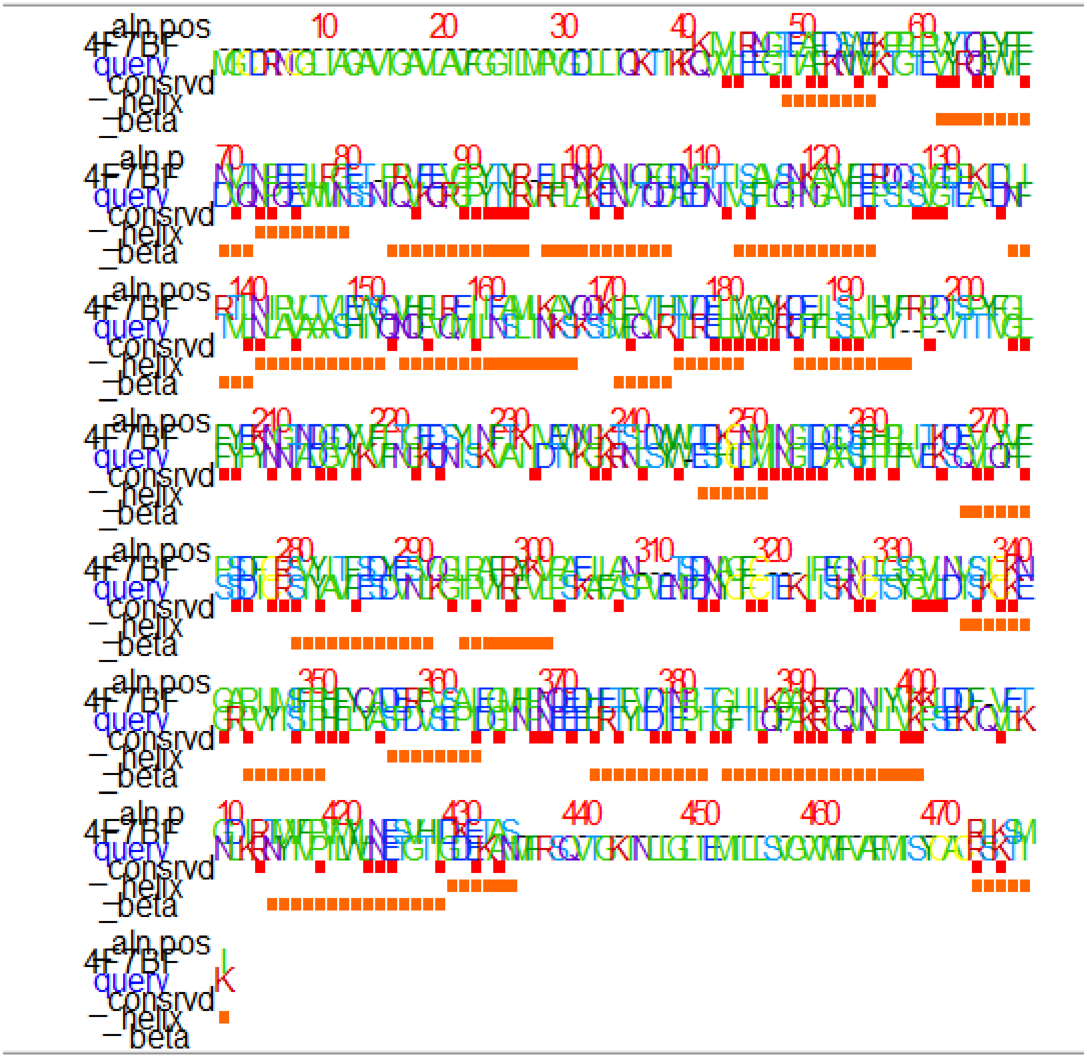


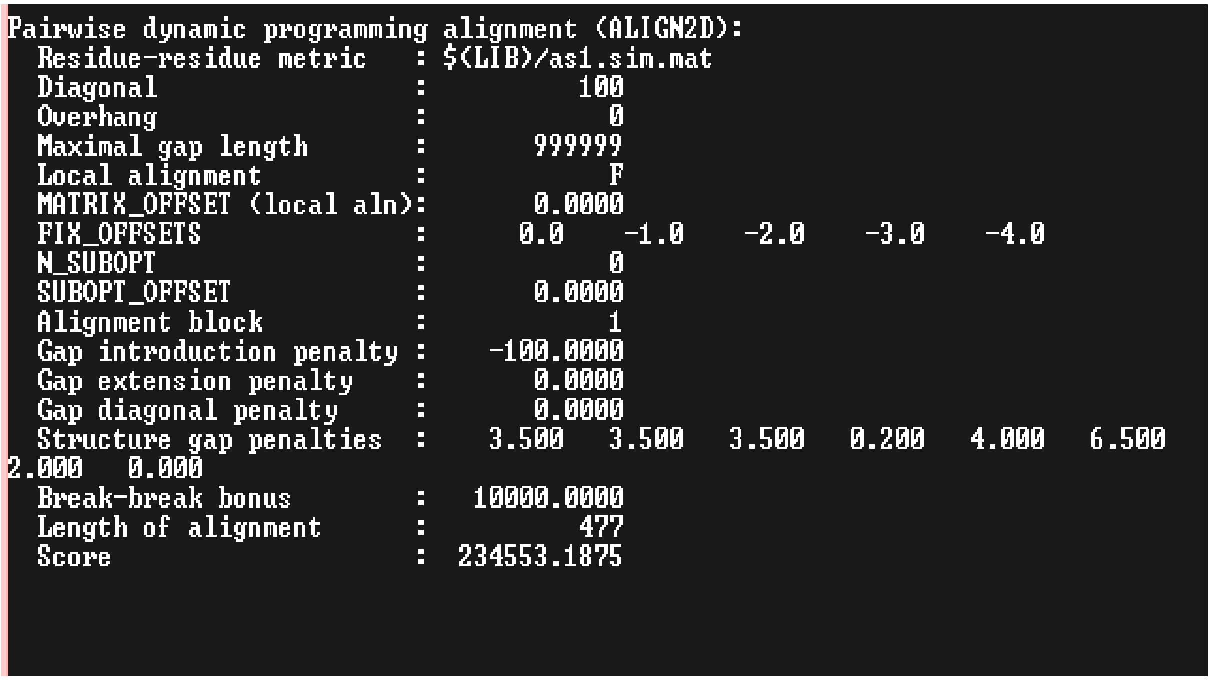
```
 _aln.pos            10        20        30        40        50        60
4F7BF      KIVLRNGTEAFDSWEKPPLPVYTQFYFFNVTNPEEILRGETPRVEEVGPYTYRELRNKANIQFGDNGT
 _consrvd

  aln.p    70        80        90       100       110       120       130
4F7BF      TISAVSNKAYVFERDQSVGDPKIDLIRTLNIPVLTVIEWSQVHFLREIIEAMLKAYQQKLFVTHTVDE
 _consrvd

 _aln.pos    140       150       160       170       180       190       200
4F7BF      LLWGYKDEILSLIHVFRPDISPYFGLFYEKNGTNDGDYVFLTGEDSYLNFTKIVEWNGKTSLDWWITD
 _consrvd

 _aln.pos      210       220       230       240       250       260       270
4F7BF      KCNMINGTDGDSFHPLITKDEVLYVFPSDFCRSVYITFSDYESVQGLPAFRYKVPAEILANTSDNAGF
 _consrvd

 _aln.pos        280       290       300       310       320       330       340
4F7BF      CIPEGNCLGSGVLNVSICKNGAPIIMSFPHFYQADERFVSAIEGMHPNQEDHETFVDINPLTGIILKA
 _consrvd

 _aln.pos          350       360       370       380       390
4F7BF      AKRFQINIYVKKLDDFVETGDIRTMVFPVMYLNESVHIDKETASRLKSMI
 _consrvd
```

4.4.2 Aligning CD36 with the template

Python script align2d.py is used to align the CD36 with the template structure. It uses align2d command which is based on a dynamic programming algorithm and is different from general sequence alignment methods as it takes into consideration the structural information from the template while constructing an alignment. By the appropriate insertion of gaps using variable gap penalty function which tends to place gaps in solvent exposed and curved regions, outside secondary structure segments, and between two positions that are in close proximity, the alignment errors are reduced significantly in comparison to general sequence alignment methods. The above improvement becomes more critical for the sequences exhibiting less similarity and harbouring more number of gaps in the alignment.

```
>P1;4F7BF
structureX:4F7B:  40 :F:+390 :F:MOL_ID  1; MOLECULE  LYSOSOME MEMBRANE
PROTEIN 2; CHAIN  A, B, C, D, E, F; FRAGMENT  LYSOSOME MEMBRANE PROTEIN
2 (UNP REISDUES 34-429) SYNONYM  85 KDA LYSOSOMAL MEMBRANE
SIALOGLYCOPROTEIN, LGP85 ANTIGEN-LIKE 2, LYSOSOME MEMBRANE PROTEIN II,
LIMP II, SCAV RECEPTOR CLASS B MEMBER 2; ENGINEERED  YES:MOL_ID  1;
ORGANISM_SCIENTIFIC  HOMO SAPIENS; ORGANISM_COMMON  HUMAN;
ORGANISM_TAXID  9606; GENE  CD36L2, LIMPII, SCARB2: 3.00:-1.00
KIVLRNGTEAFDSWEKPPLPVYTQFYFFNVTNPEEILRGETPRVEEVGPYTYRELRNKANIQFGDNGTTIS

sAVSN

KAYVFERDQSVGDPKIDLIRTLNIPVLTVIEWSQVHFLREIIEAMLKAYQQKLFVTHTVDELLWGYKDEIL

SLIH

VFRPDISPYFGLFYEKNGTNDGDYVFLTGEDSYLNFTKIVEWNGKTSLDWWITDKCNMINGTDGDSFHPLI

TKDE

VLYVFPSDFCRSVYITFSDYESVQGLPAFRYKVPAEILANTSDNAGFCIPEGNCLGSGVLNVSICKNGAPI

IMSF

PHFYQADERFVSAIEGMHPNQEDHETFVDINPLTGIILKAAKRFQINIYVKKLDDFVETGDIRTMVFPVMY

LNES

VHIDKETASRLKSMI*
```

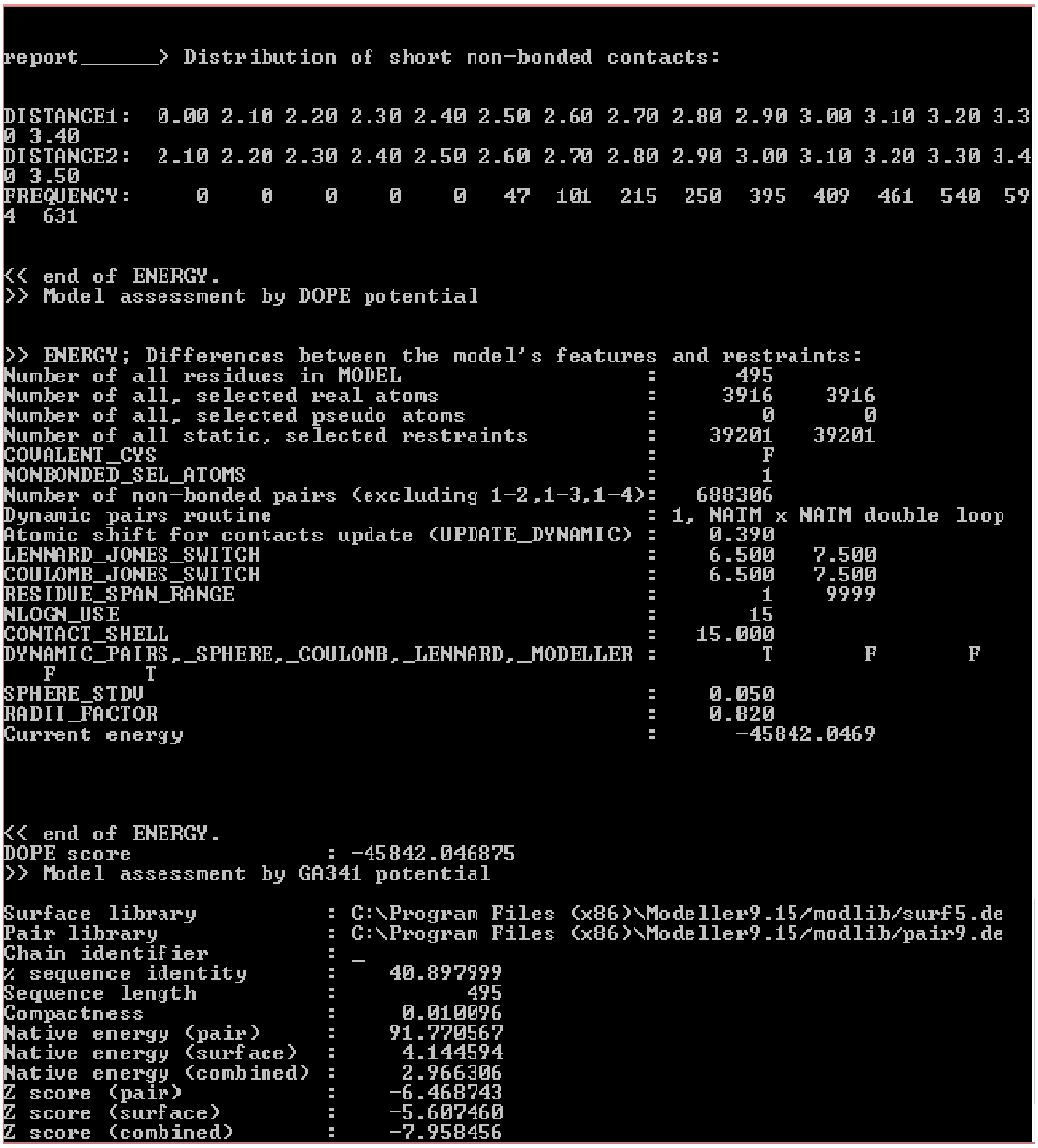




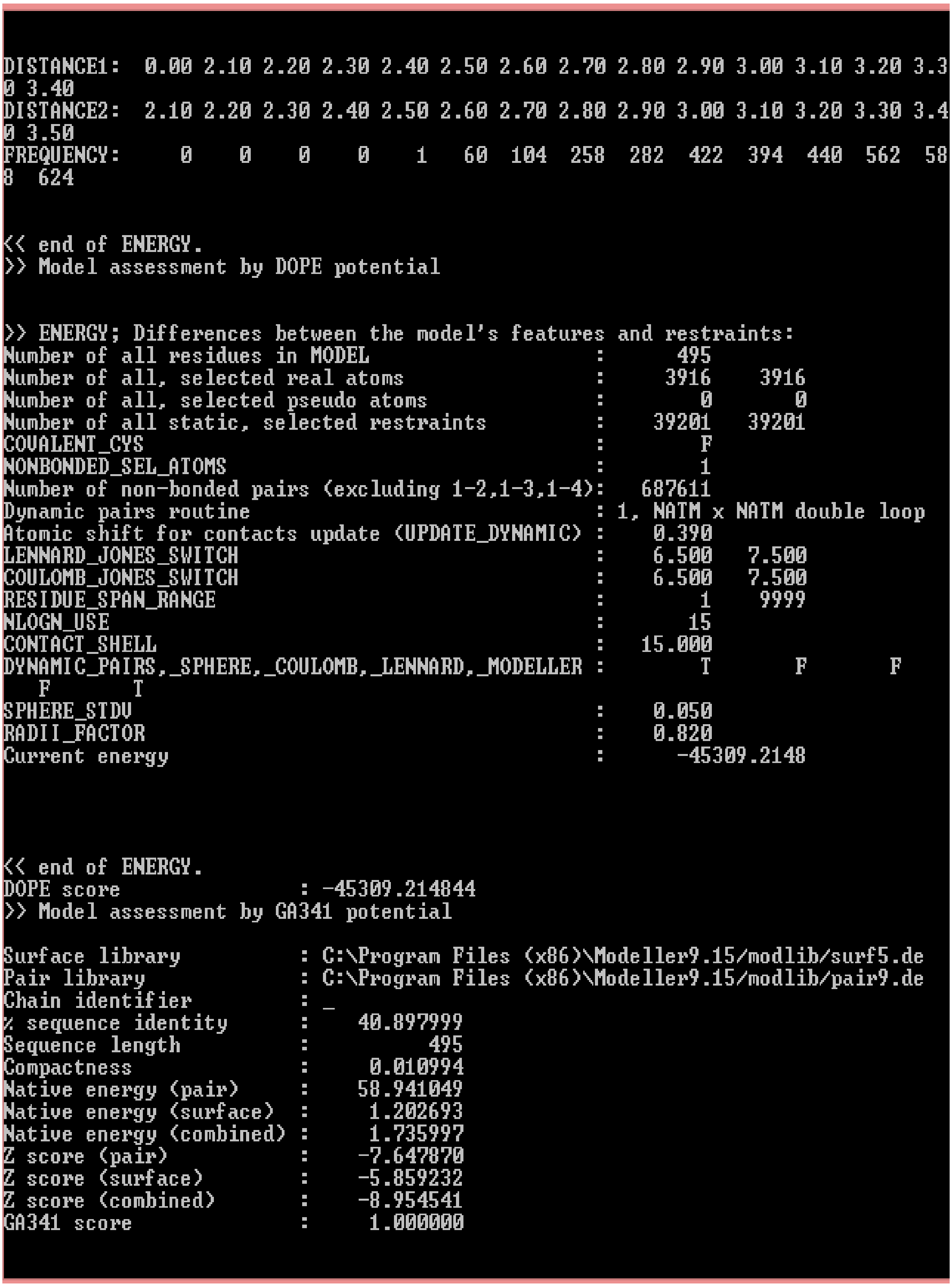
```
Pairwise dynamic programming alignment (ALIGN2D):
  Residue-residue metric    : $(LIB)/as1.sim.mat
  Diagonal                  :          100
  Overhang                  :            0
  Maximal gap length        :       999999
  Local alignment           :            F
  MATRIX_OFFSET (local aln):       0.0000
  FIX_OFFSETS               :      0.0     -1.0     -2.0     -3.0     -4.0
  N_SUBOPT                  :            0
  SUBOPT_OFFSET             :       0.0000
  Alignment block           :            1
  Gap introduction penalty :    -100.0000
  Gap extension penalty     :       0.0000
  Gap diagonal penalty      :       0.0000
  Structure gap penalties   :    3.500    3.500    3.500    0.200    4.000    6.500
2.000    0.000
  Break-break bonus         :   10000.0000
  Length of alignment       :          477
  Score                     :  234553.1875
```

4.4.3 Model Building

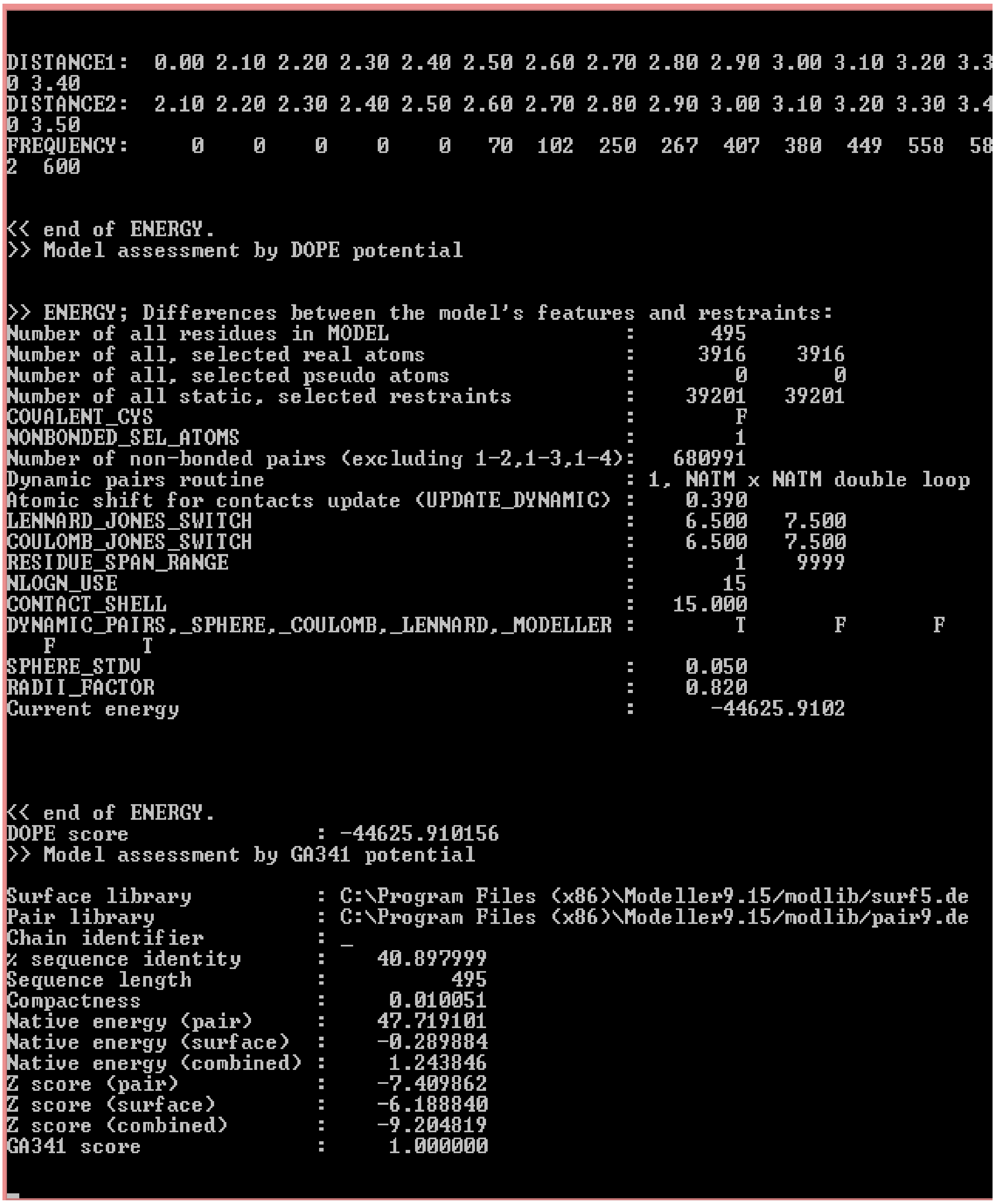
Another python script named as model-single.py is use for building 3D model of CD36 from the sequence template alignment. The objective is achieved by auto-model class of Modeller. A total of 100 3D models of CD36 were prepared as result of “model_single.py” and an output file “model-single.log” summarizing all the models built.



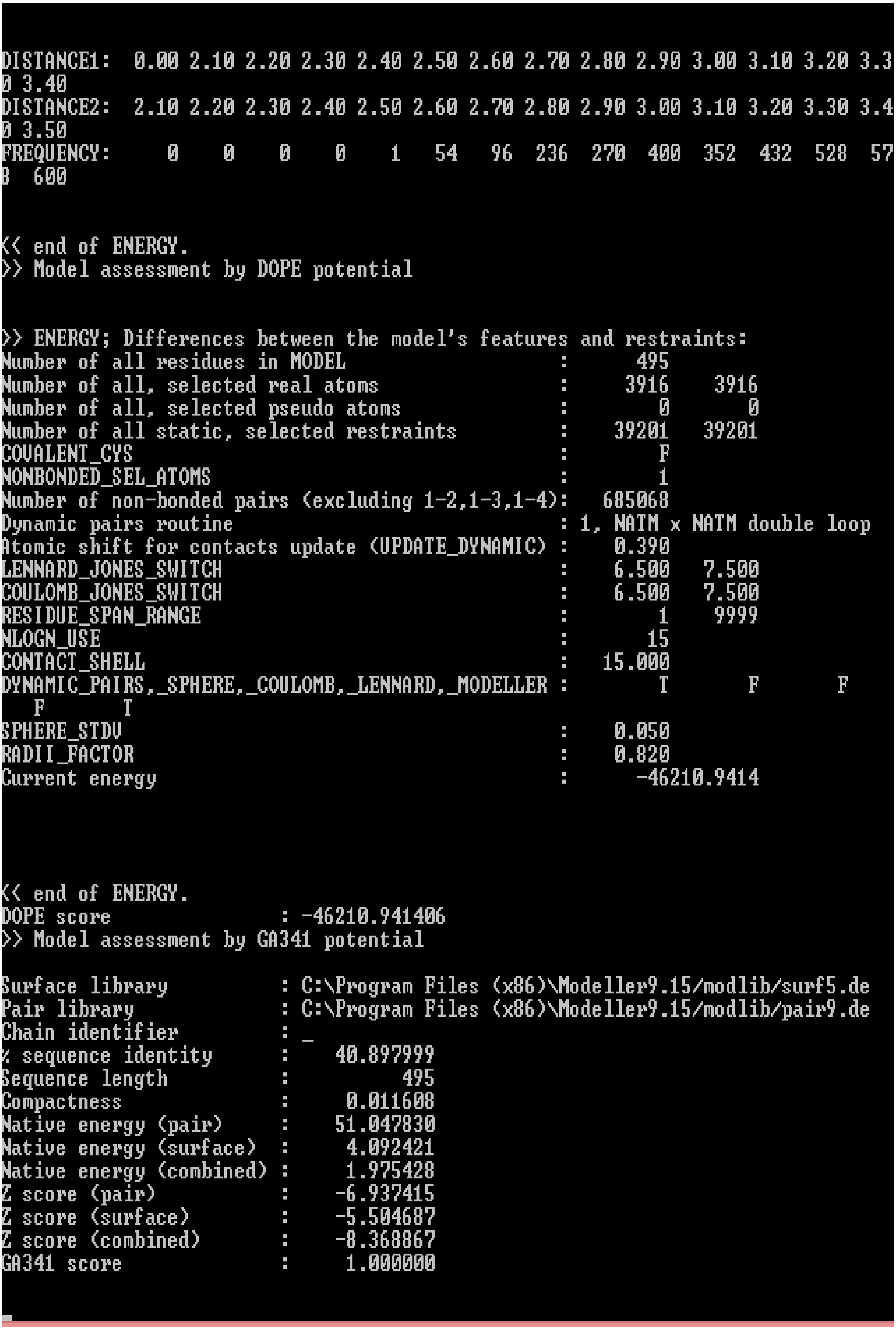
```
report______> Distribution of short non-bonded contacts:


DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:     0    0    0    0    0   47  101  215  250  395  409  461  540  59
4  631


<< end of ENERGY.
>> Model assessment by DOPE potential


>> ENERGY; Differences between the model's features and restraints:
Number of all residues in MODEL                   :      495
Number of all, selected real atoms                :     3916     3916
Number of all, selected pseudo atoms              :        0        0
Number of all static, selected restraints         :    39201    39201
COVALENT_CYS                                      :        F
NONBONDED_SEL_ATOMS                               :        1
Number of non-bonded pairs (excluding 1-2,1-3,1-4):   688306
Dynamic pairs routine                             : 1, NATM x NATM double loop
Atomic shift for contacts update (UPDATE_DYNAMIC) :    0.390
LENNARD_JONES_SWITCH                              :    6.500    7.500
COULOMB_JONES_SWITCH                              :    6.500    7.500
RESIDUE_SPAN_RANGE                                :        1     9999
NLOGN_USE                                         :       15
CONTACT_SHELL                                     :   15.000
DYNAMIC_PAIRS,_SPHERE,_COULOMB,_LENNARD,_MODELLER :        T        F        F
       F        T
SPHERE_STDV                                       :    0.050
RADII_FACTOR                                      :    0.820
Current energy                                    :      -45842.0469




<< end of ENERGY.
DOPE score               : -45842.046875
>> Model assessment by GA341 potential

Surface library          : C:\Program Files (x86)\Modeller9.15/modlib/surf5.de
Pair library             : C:\Program Files (x86)\Modeller9.15/modlib/pair9.de
Chain identifier         : _
% sequence identity      :   40.897999
Sequence length          :         495
Compactness              :    0.010096
Native energy (pair)     :   91.770567
Native energy (surface)  :    4.144594
Native energy (combined) :    2.966306
Z score (pair)           :   -6.468743
Z score (surface)        :   -5.607460
Z score (combined)       :   -7.958456
```



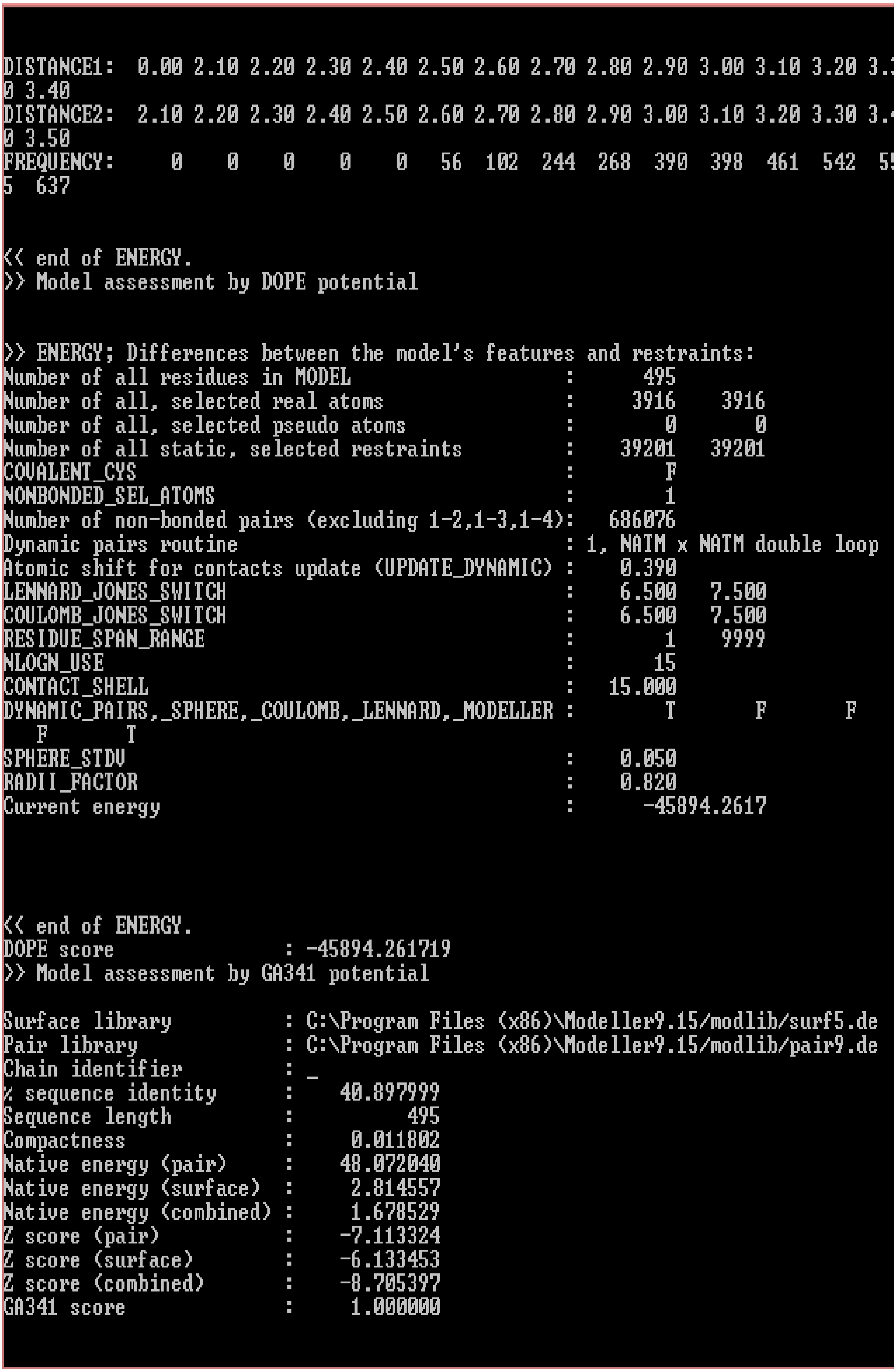
```
DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:     0    0    0    0    1   60  104  258  282  422  394  440  562  58
8  624


<< end of ENERGY.
>> Model assessment by DOPE potential


>> ENERGY; Differences between the model's features and restraints:
Number of all residues in MODEL                   :       495
Number of all, selected real atoms                :      3916     3916
Number of all, selected pseudo atoms              :         0        0
Number of all static, selected restraints         :     39201    39201
COVALENT_CYS                                      :         F
NONBONDED_SEL_ATOMS                               :         1
Number of non-bonded pairs (excluding 1-2,1-3,1-4):    687611
Dynamic pairs routine                             : 1, NATM x NATM double loop
Atomic shift for contacts update (UPDATE_DYNAMIC) :     0.390
LENNARD_JONES_SWITCH                              :     6.500    7.500
COULOMB_JONES_SWITCH                              :     6.500    7.500
RESIDUE_SPAN_RANGE                                :         1     9999
NLOGN_USE                                         :        15
CONTACT_SHELL                                     :    15.000
DYNAMIC_PAIRS,_SPHERE,_COULOMB,_LENNARD,_MODELLER :         T        F        F
   F        T
SPHERE_STDV                                       :     0.050
RADII_FACTOR                                      :     0.820
Current energy                                    :      -45309.2148




<< end of ENERGY.
DOPE score               : -45309.214844
>> Model assessment by GA341 potential

Surface library          : C:\Program Files (x86)\Modeller9.15/modlib/surf5.de
Pair library             : C:\Program Files (x86)\Modeller9.15/modlib/pair9.de
Chain identifier         : _
% sequence identity      :   40.897999
Sequence length          :         495
Compactness              :    0.010994
Native energy (pair)     :   58.941049
Native energy (surface)  :    1.202693
Native energy (combined) :    1.735997
Z score (pair)           :   -7.647870
Z score (surface)        :   -5.859232
Z score (combined)       :   -8.954541
GA341 score              :    1.000000
```



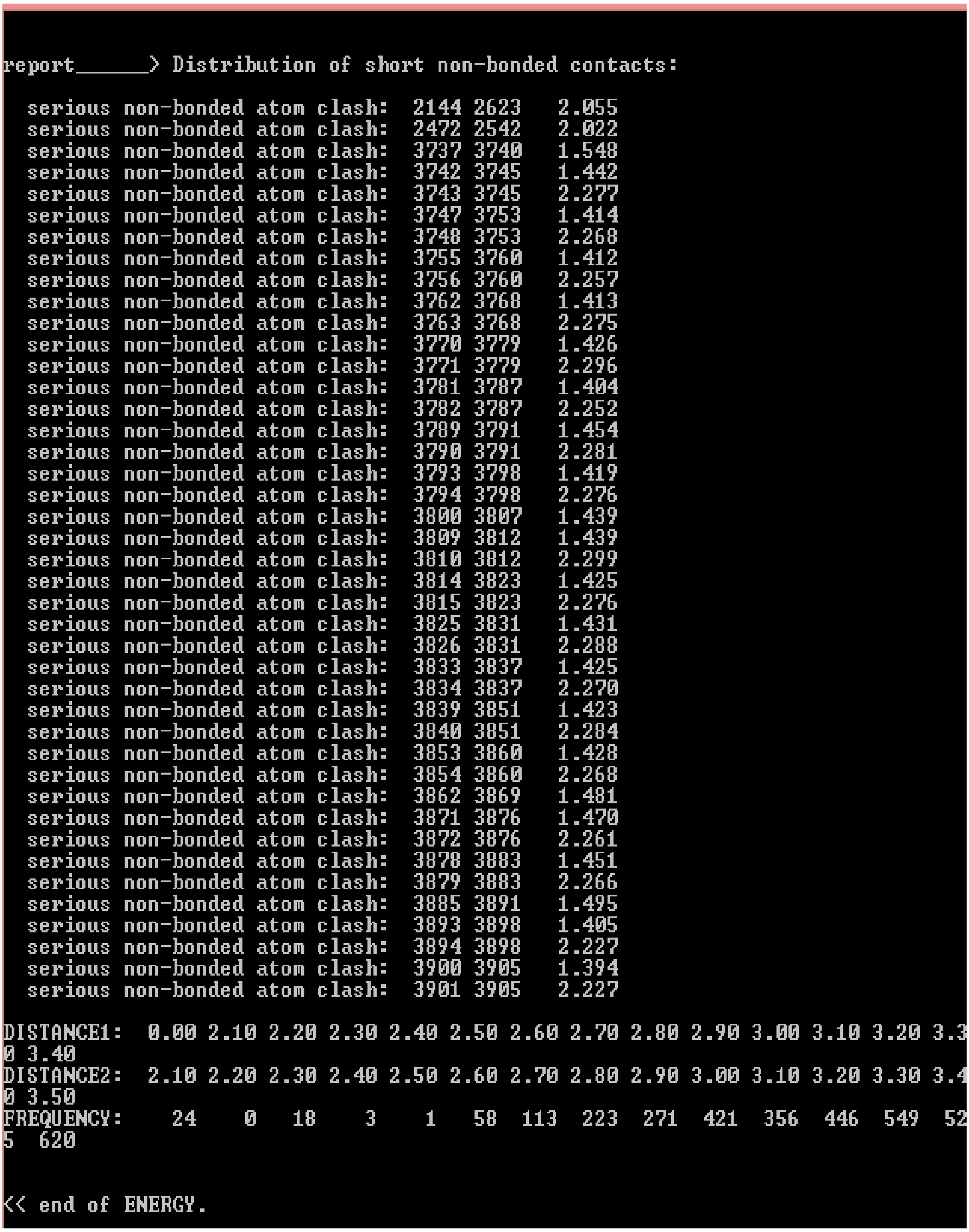
```
DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:     0    0    0    0    0   70  102  250  267  407  380  449  558  58
2  600


<< end of ENERGY.
>> Model assessment by DOPE potential


>> ENERGY; Differences between the model's features and restraints:
Number of all residues in MODEL                   :      495
Number of all, selected real atoms                :     3916    3916
Number of all, selected pseudo atoms              :        0       0
Number of all static, selected restraints         :    39201   39201
COVALENT_CYS                                      :        F
NONBONDED_SEL_ATOMS                               :        1
Number of non-bonded pairs (excluding 1-2,1-3,1-4):   680991
Dynamic pairs routine                             : 1, NATM x NATM double loop
Atomic shift for contacts update (UPDATE_DYNAMIC) :    0.390
LENNARD_JONES_SWITCH                              :    6.500   7.500
COULOMB_JONES_SWITCH                              :    6.500   7.500
RESIDUE_SPAN_RANGE                                :        1    9999
NLOGN_USE                                         :       15
CONTACT_SHELL                                     :   15.000
DYNAMIC_PAIRS,_SPHERE,_COULOMB,_LENNARD,_MODELLER :        T       F       F
   F       T
SPHERE_STDV                                       :    0.050
RADII_FACTOR                                      :    0.820
Current energy                                    :      -44625.9102



<< end of ENERGY.
DOPE score               : -44625.910156
>> Model assessment by GA341 potential

Surface library          : C:\Program Files (x86)\Modeller9.15/modlib/surf5.de
Pair library             : C:\Program Files (x86)\Modeller9.15/modlib/pair9.de
Chain identifier         : _
% sequence identity      :   40.897999
Sequence length          :         495
Compactness              :    0.010051
Native energy (pair)     :   47.719101
Native energy (surface)  :   -0.289884
Native energy (combined) :    1.243846
Z score (pair)           :   -7.409862
Z score (surface)        :   -6.188840
Z score (combined)       :   -9.204819
GA341 score              :    1.000000
```



```
DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:     0    0    0    0    1   54   96  236  270  400  352  432  528  57
8  600


<< end of ENERGY.
>> Model assessment by DOPE potential


>> ENERGY; Differences between the model's features and restraints:
Number of all residues in MODEL                   :      495
Number of all, selected real atoms                :     3916     3916
Number of all, selected pseudo atoms              :        0        0
Number of all static, selected restraints         :    39201    39201
COVALENT_CYS                                      :        F
NONBONDED_SEL_ATOMS                               :        1
Number of non-bonded pairs (excluding 1-2,1-3,1-4):   685068
Dynamic pairs routine                             : 1, NATM x NATM double loop
Atomic shift for contacts update (UPDATE_DYNAMIC) :    0.390
LENNARD_JONES_SWITCH                              :    6.500    7.500
COULOMB_JONES_SWITCH                              :    6.500    7.500
RESIDUE_SPAN_RANGE                                :        1     9999
NLOGN_USE                                         :       15
CONTACT_SHELL                                     :   15.000
DYNAMIC_PAIRS,_SPHERE,_COULOMB,_LENNARD,_MODELLER :        T        F        F
   F        T
SPHERE_STDV                                       :    0.050
RADII_FACTOR                                      :    0.820
Current energy                                    :      -46210.9414




<< end of ENERGY.
DOPE score               : -46210.941406
>> Model assessment by GA341 potential

Surface library          : C:\Program Files (x86)\Modeller9.15/modlib/surf5.de
Pair library             : C:\Program Files (x86)\Modeller9.15/modlib/pair9.de
Chain identifier         : _
% sequence identity      :    40.897999
Sequence length          :          495
Compactness              :     0.011608
Native energy (pair)     :    51.047830
Native energy (surface)  :     4.092421
Native energy (combined) :     1.975428
Z score (pair)           :    -6.937415
Z score (surface)        :    -5.504687
Z score (combined)       :    -8.368867
GA341 score              :     1.000000
```



```
DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:     0    0    0    0    0   56  102  244  268  390  398  461  542  55
5  637


<< end of ENERGY.
>> Model assessment by DOPE potential


>> ENERGY; Differences between the model's features and restraints:
Number of all residues in MODEL                   :      495
Number of all, selected real atoms                :     3916    3916
Number of all, selected pseudo atoms              :        0       0
Number of all static, selected restraints         :    39201   39201
COVALENT_CYS                                      :        F
NONBONDED_SEL_ATOMS                               :        1
Number of non-bonded pairs (excluding 1-2,1-3,1-4):   686076
Dynamic pairs routine                             : 1, NATM x NATM double loop
Atomic shift for contacts update (UPDATE_DYNAMIC) :    0.390
LENNARD_JONES_SWITCH                              :    6.500   7.500
COULOMB_JONES_SWITCH                              :    6.500   7.500
RESIDUE_SPAN_RANGE                                :        1    9999
NLOGN_USE                                         :       15
CONTACT_SHELL                                     :   15.000
DYNAMIC_PAIRS,_SPHERE,_COULOMB,_LENNARD,_MODELLER :        T       F       F
   F       T
SPHERE_STDV                                       :    0.050
RADII_FACTOR                                      :    0.820
Current energy                                    :      -45894.2617




<< end of ENERGY.
DOPE score               : -45894.261719
>> Model assessment by GA341 potential

Surface library          : C:\Program Files (x86)\Modeller9.15/modlib/surf5.de
Pair library             : C:\Program Files (x86)\Modeller9.15/modlib/pair9.de
Chain identifier         : _
% sequence identity      :    40.897999
Sequence length          :          495
Compactness              :     0.011802
Native energy (pair)     :    48.072040
Native energy (surface)  :     2.814557
Native energy (combined) :     1.678529
Z score (pair)           :    -7.113324
Z score (surface)        :    -6.133453
Z score (combined)       :    -8.705397
GA341 score              :     1.000000
```



```
report______> Distribution of short non-bonded contacts:

  serious non-bonded atom clash:  2144 2623   2.055
  serious non-bonded atom clash:  2472 2542   2.022
  serious non-bonded atom clash:  3737 3740   1.548
  serious non-bonded atom clash:  3742 3745   1.442
  serious non-bonded atom clash:  3743 3745   2.277
  serious non-bonded atom clash:  3747 3753   1.414
  serious non-bonded atom clash:  3748 3753   2.268
  serious non-bonded atom clash:  3755 3760   1.412
  serious non-bonded atom clash:  3756 3760   2.257
  serious non-bonded atom clash:  3762 3768   1.413
  serious non-bonded atom clash:  3763 3768   2.275
  serious non-bonded atom clash:  3770 3779   1.426
  serious non-bonded atom clash:  3771 3779   2.296
  serious non-bonded atom clash:  3781 3787   1.404
  serious non-bonded atom clash:  3782 3787   2.252
  serious non-bonded atom clash:  3789 3791   1.454
  serious non-bonded atom clash:  3790 3791   2.281
  serious non-bonded atom clash:  3793 3798   1.419
  serious non-bonded atom clash:  3794 3798   2.276
  serious non-bonded atom clash:  3800 3807   1.439
  serious non-bonded atom clash:  3809 3812   1.439
  serious non-bonded atom clash:  3810 3812   2.299
  serious non-bonded atom clash:  3814 3823   1.425
  serious non-bonded atom clash:  3815 3823   2.276
  serious non-bonded atom clash:  3825 3831   1.431
  serious non-bonded atom clash:  3826 3831   2.288
  serious non-bonded atom clash:  3833 3837   1.425
  serious non-bonded atom clash:  3834 3837   2.270
  serious non-bonded atom clash:  3839 3851   1.423
  serious non-bonded atom clash:  3840 3851   2.284
  serious non-bonded atom clash:  3853 3860   1.428
  serious non-bonded atom clash:  3854 3860   2.268
  serious non-bonded atom clash:  3862 3869   1.481
  serious non-bonded atom clash:  3871 3876   1.470
  serious non-bonded atom clash:  3872 3876   2.261
  serious non-bonded atom clash:  3878 3883   1.451
  serious non-bonded atom clash:  3879 3883   2.266
  serious non-bonded atom clash:  3885 3891   1.495
  serious non-bonded atom clash:  3893 3898   1.405
  serious non-bonded atom clash:  3894 3898   2.227
  serious non-bonded atom clash:  3900 3905   1.394
  serious non-bonded atom clash:  3901 3905   2.227

DISTANCE1:  0.00 2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.3
0 3.40
DISTANCE2:  2.10 2.20 2.30 2.40 2.50 2.60 2.70 2.80 2.90 3.00 3.10 3.20 3.30 3.4
0 3.50
FREQUENCY:    24    0   18    3    1   58  113  223  271  421  356  446  549  52
5  620


<< end of ENERGY.
```

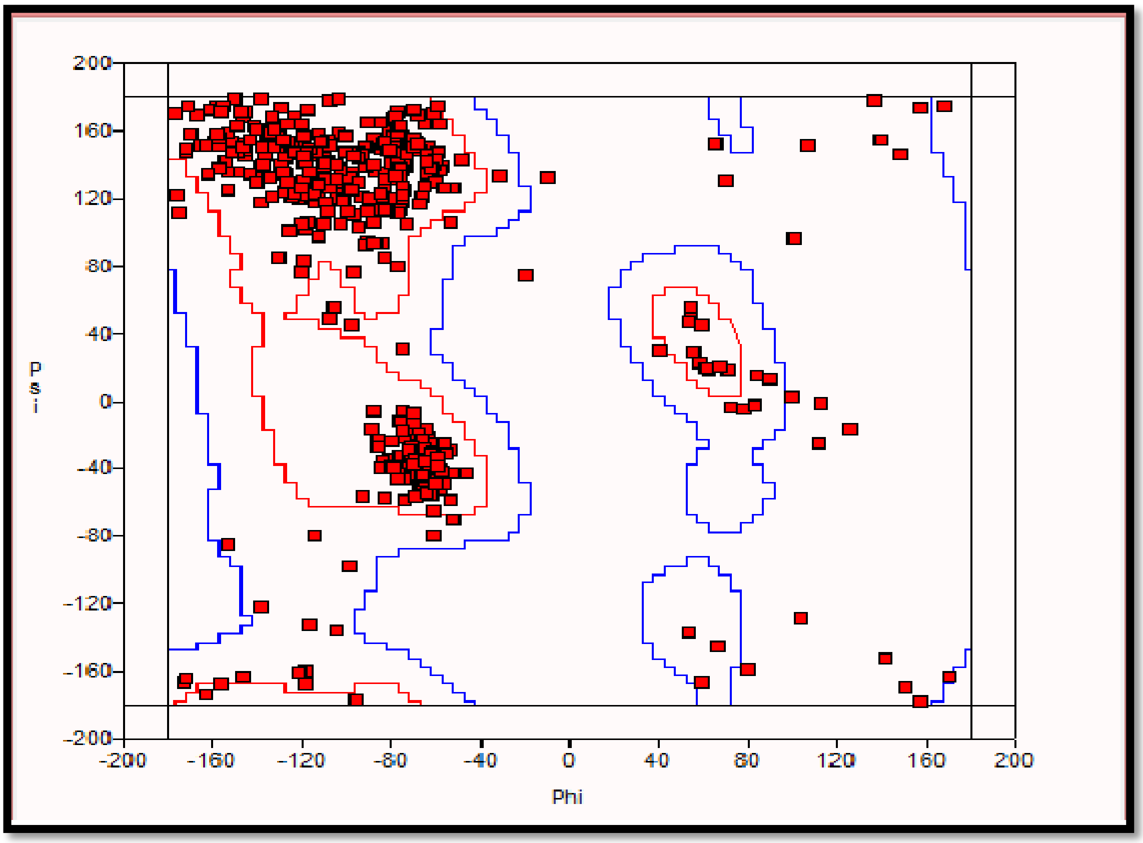
4.4.4 Selection of the best Model

The model can be picked with the lowest value of the modeller objective function or the DOPE assessment score and with the highest GA341 assessment score reported in log file “model-single.log”. GA341 score ranges from 0.0 (worst) to 1.0 (native-like). The DOPE and GA341 score are not absolute measure, they are only used to rank models calculated from the sequence alignment. However, DOPE score is better than GA341 in distinguishing ‘good’ models from ‘bad’ models. The GA341 was highest, 1 for all the hundred models, so, 5 lowest DOPE scoring models are picked out of 100 further evaluation.

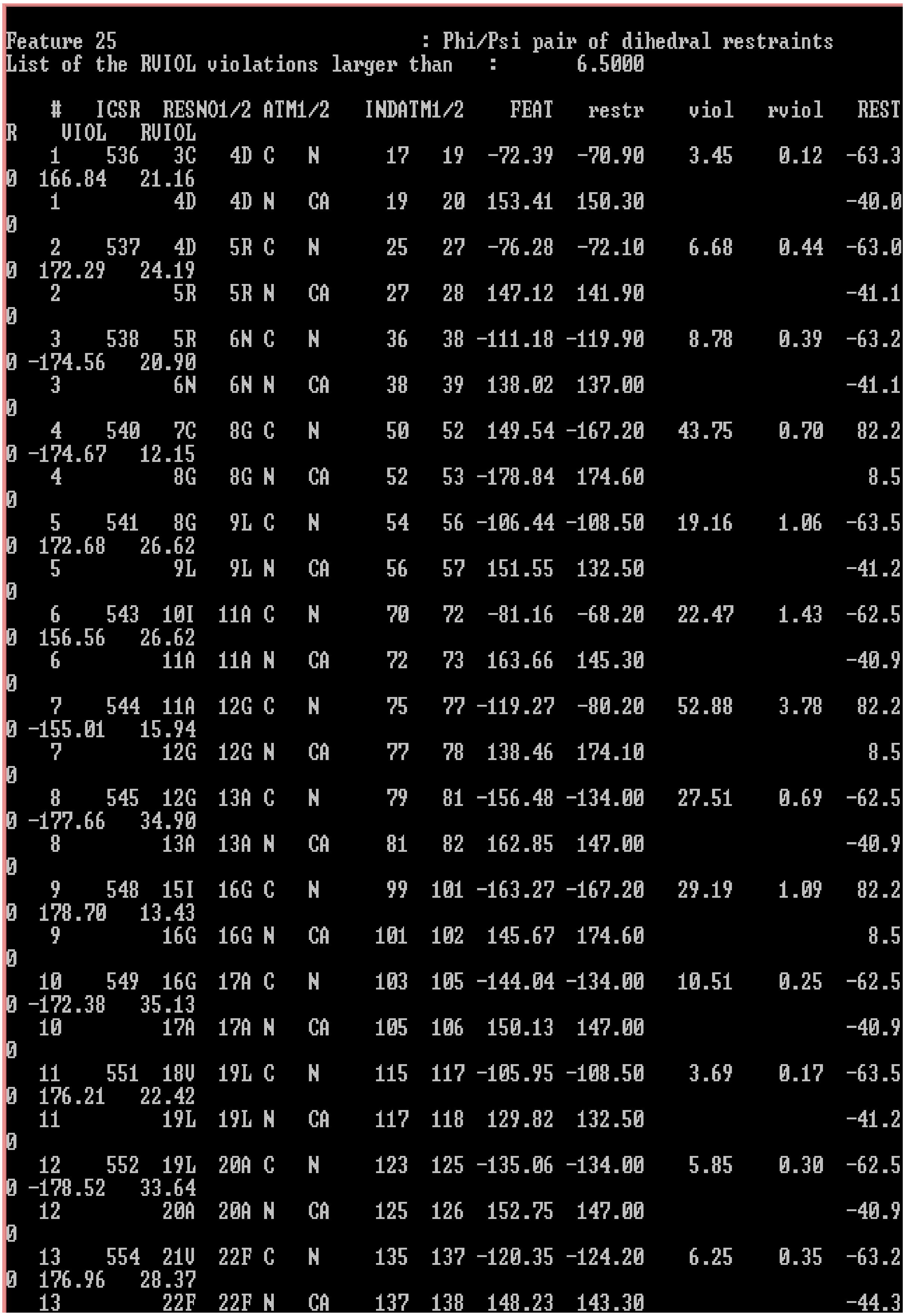
Python script evaluate_model.py uses complete_pdb to residues in a PDB file and prepares it for energy calculation. The DOPE energy is calculated with the assess_dope command and the energy profile is smooth ednedoevr a 15 residue window and normalised by the number of restrain acting on each residue is written to the output file “4F7B.profile(for template)”, which can be plotted using python script plot_profiles.py.

4.4.5 Model Evaluation

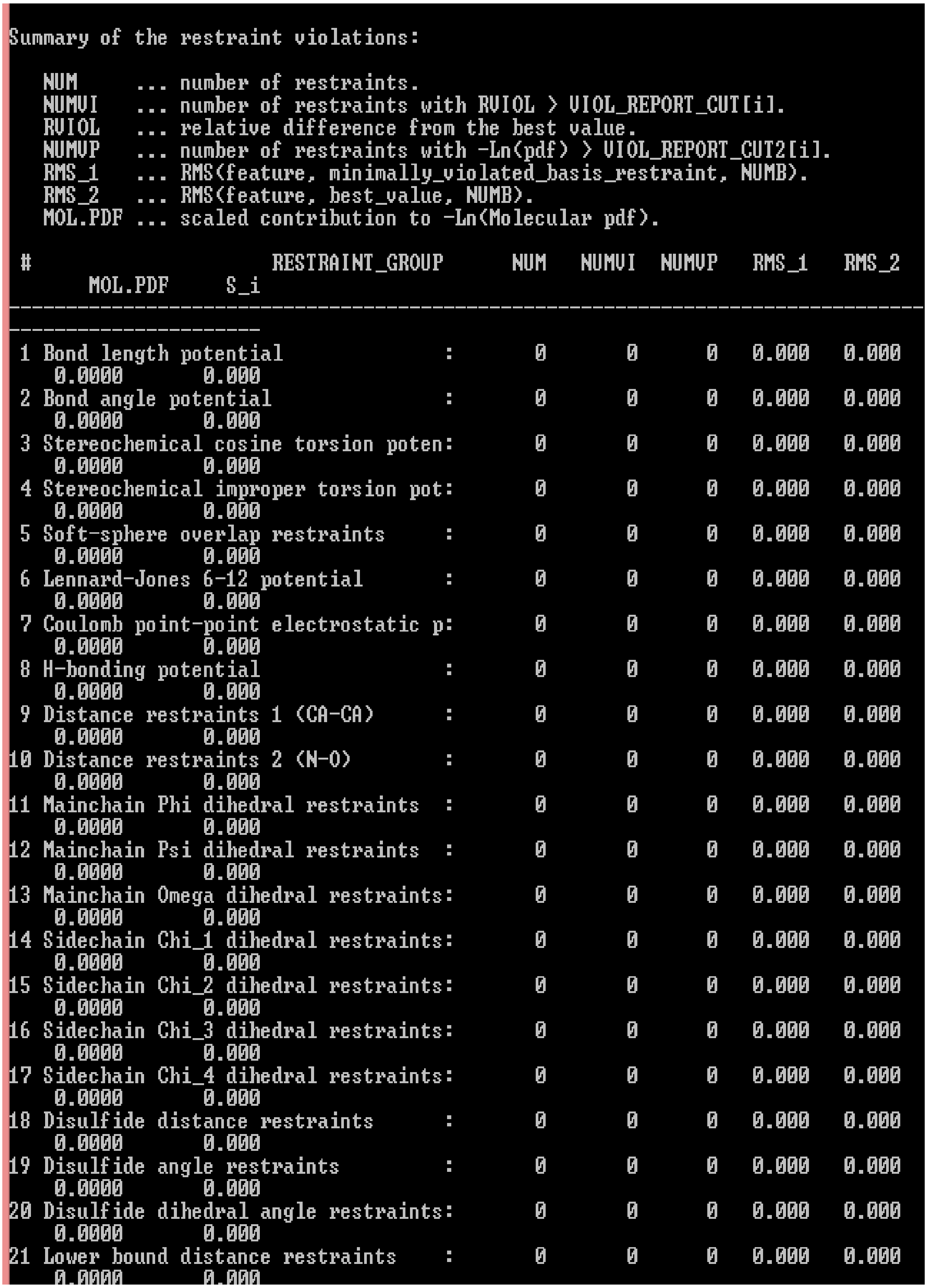
The best model out of these 5 shortlisted model was selected by PROCHECK (Laskowski RA et.al; 1996) server, PROSA-we server.



Ramachandran Plot. Ramachandran Plot is obtained as output by using PROCHECK software.



```
Feature 25                              : Phi/Psi pair of dihedral restraints
List of the RVIOL violations larger than    :       6.5000

    #   ICSR  RESNO1/2 ATM1/2   INDATM1/2    FEAT   restr    viol   rviol   REST
R    VIOL   RVIOL
    1    536   3C   4D C   N      17   19  -72.39  -70.90    3.45    0.12  -63.3
0  166.84   21.16
    1          4D   4D N   CA     19   20  153.41  150.30                  -40.0
0
    2    537   4D   5R C   N      25   27  -76.28  -72.10    6.68    0.44  -63.0
0  172.29   24.19
    2          5R   5R N   CA     27   28  147.12  141.90                  -41.1
0
    3    538   5R   6N C   N      36   38 -111.18 -119.90    8.78    0.39  -63.2
0 -174.56   20.90
    3          6N   6N N   CA     38   39  138.02  137.00                  -41.1
0
    4    540   7C   8G C   N      50   52  149.54 -167.20   43.75    0.70   82.2
0 -174.67   12.15
    4          8G   8G N   CA     52   53 -178.84  174.60                    8.5
0
    5    541   8G   9L C   N      54   56 -106.44 -108.50   19.16    1.06  -63.5
0  172.68   26.62
    5          9L   9L N   CA     56   57  151.55  132.50                  -41.2
0
    6    543  10I  11A C   N      70   72  -81.16  -68.20   22.47    1.43  -62.5
0  156.56   26.62
    6         11A  11A N   CA     72   73  163.66  145.30                  -40.9
0
    7    544  11A  12G C   N      75   77 -119.27  -80.20   52.88    3.78   82.2
0 -155.01   15.94
    7         12G  12G N   CA     77   78  138.46  174.10                    8.5
0
    8    545  12G  13A C   N      79   81 -156.48 -134.00   27.51    0.69  -62.5
0 -177.66   34.90
    8         13A  13A N   CA     81   82  162.85  147.00                  -40.9
0
    9    548  15I  16G C   N      99  101 -163.27 -167.20   29.19    1.09   82.2
0  178.70   13.43
    9         16G  16G N   CA    101  102  145.67  174.60                    8.5
0
   10    549  16G  17A C   N     103  105 -144.04 -134.00   10.51    0.25  -62.5
0 -172.38   35.13
   10         17A  17A N   CA    105  106  150.13  147.00                  -40.9
0
   11    551  18V  19L C   N     115  117 -105.95 -108.50    3.69    0.17  -63.5
0  176.21   22.42
   11         19L  19L N   CA    117  118  129.82  132.50                  -41.2
0
   12    552  19L  20A C   N     123  125 -135.06 -134.00    5.85    0.30  -62.5
0 -178.52   33.64
   12         20A  20A N   CA    125  126  152.75  147.00                  -40.9
0
   13    554  21V  22F C   N     135  137 -120.35 -124.20    6.25    0.35  -63.2
0  176.96   28.37
   13         22F  22F N   CA    137  138  148.23  143.30                  -44.3
```



```
Summary of the restraint violations:

   NUM     ... number of restraints.
   NUMVI   ... number of restraints with RVIOL > VIOL_REPORT_CUT[i].
   RVIOL   ... relative difference from the best value.
   NUMVP   ... number of restraints with -Ln(pdf) > VIOL_REPORT_CUT2[i].
   RMS_1   ... RMS(feature, minimally_violated_basis_restraint, NUMB).
   RMS_2   ... RMS(feature, best_value, NUMB).
   MOL.PDF ... scaled contribution to -Ln(Molecular pdf).

 #                     RESTRAINT_GROUP      NUM   NUMVI  NUMVP   RMS_1   RMS_2
       MOL.PDF     S_i
--------------------------------------------------------------------------------
----------------------
 1 Bond length potential              :       0       0      0   0.000   0.000
    0.0000      0.000
 2 Bond angle potential               :       0       0      0   0.000   0.000
    0.0000      0.000
 3 Stereochemical cosine torsion poten:       0       0      0   0.000   0.000
    0.0000      0.000
 4 Stereochemical improper torsion pot:       0       0      0   0.000   0.000
    0.0000      0.000
 5 Soft-sphere overlap restraints     :       0       0      0   0.000   0.000
    0.0000      0.000
 6 Lennard-Jones 6-12 potential       :       0       0      0   0.000   0.000
    0.0000      0.000
 7 Coulomb point-point electrostatic p:       0       0      0   0.000   0.000
    0.0000      0.000
 8 H-bonding potential                :       0       0      0   0.000   0.000
    0.0000      0.000
 9 Distance restraints 1 (CA-CA)      :       0       0      0   0.000   0.000
    0.0000      0.000
10 Distance restraints 2 (N-O)        :       0       0      0   0.000   0.000
    0.0000      0.000
11 Mainchain Phi dihedral restraints  :       0       0      0   0.000   0.000
    0.0000      0.000
12 Mainchain Psi dihedral restraints  :       0       0      0   0.000   0.000
    0.0000      0.000
13 Mainchain Omega dihedral restraints:       0       0      0   0.000   0.000
    0.0000      0.000
14 Sidechain Chi_1 dihedral restraints:       0       0      0   0.000   0.000
    0.0000      0.000
15 Sidechain Chi_2 dihedral restraints:       0       0      0   0.000   0.000
    0.0000      0.000
16 Sidechain Chi_3 dihedral restraints:       0       0      0   0.000   0.000
    0.0000      0.000
17 Sidechain Chi_4 dihedral restraints:       0       0      0   0.000   0.000
    0.0000      0.000
18 Disulfide distance restraints      :       0       0      0   0.000   0.000
    0.0000      0.000
19 Disulfide angle restraints         :       0       0      0   0.000   0.000
    0.0000      0.000
20 Disulfide dihedral angle restraints:       0       0      0   0.000   0.000
    0.0000      0.000
21 Lower bound distance restraints    :       0       0      0   0.000   0.000
    0.0000      0.000
```

4.4.6 Energy Minimization of Energy Model

Energy minimization of 3D model was done using a multilevel optimization method YASARA server, which runs molecular dynamics simulations of models in explicit solvent, using a new partly knowledge-based all atom force field derived from Amber, whose parameters have been optimized to minimize the damage done to protein crystal structures (Emla K et.al;2009).

Step1: The pdb file of the modelled protein was uploaded on the server.

Step2: Minimization process take place on the YASARA server and the minimized model is received on the mail as YASARA screen, which can be viewed in YASARA viewer.

Step3: This model can be saved in pdb format using save as pdb option of YASARA.

4.4.7 Mapping of Residues on Structure

PyMOL is an opensource tool to visualize molecules available from (www.pymol.org). Mark the important residues on the protein with the help of PYMOL. This was done manually by selecting the residues and then colour of selected residues was changed using [C]olor option of PYMOL.

4.4.8 Protein –Protein Docking

Protein – protein docking was performed using HADDOCK web server (Sjoerd j et.al; 2010).

Step1: Pdb file of receptor (CD36) is uploaded.

Step2: Active residues (directly involved in the interaction).

Step3: Same is done with second molecule i.e. CIDR1α domain.

Step4: Run docking.

CHAPTER 5

RESULT

5.1 Alignment

Local pair multiple sequence alignment among 105 sequence of CIDR domain was performed using MAFFT, two type of conservation pattern are obtain (a) conserved in the CIDR domain of the entire PFEMP family.(Figure 4) (b) conserved in CIDR1α while are not conserved in other e.g. β and γ (Figure 5).

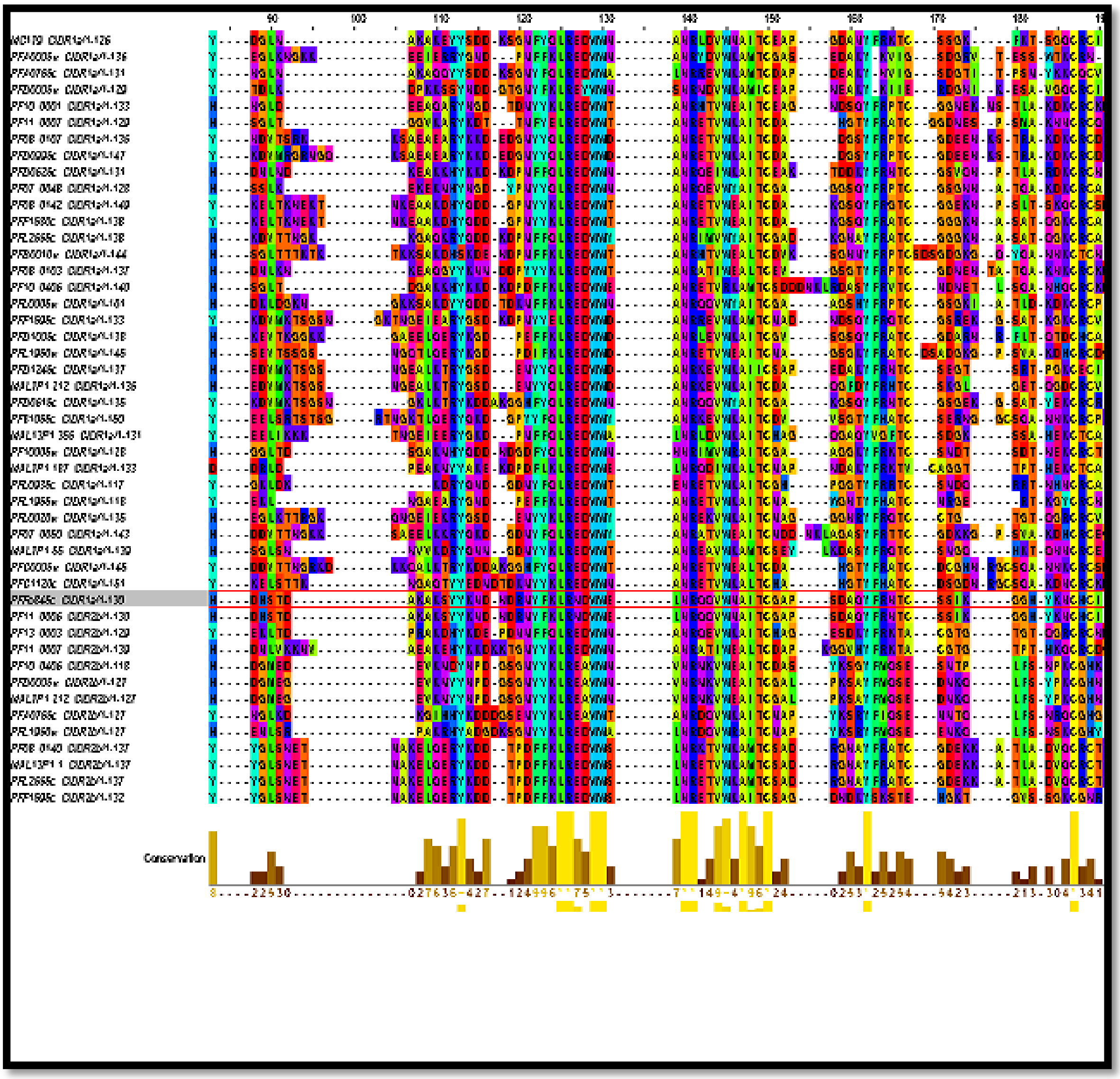


Figure 4: Shows C at 150,187 alignment position is conserved in CIDR domain of entire PFEMP1 family, these residues are structurally critical and thus have high RE scores.

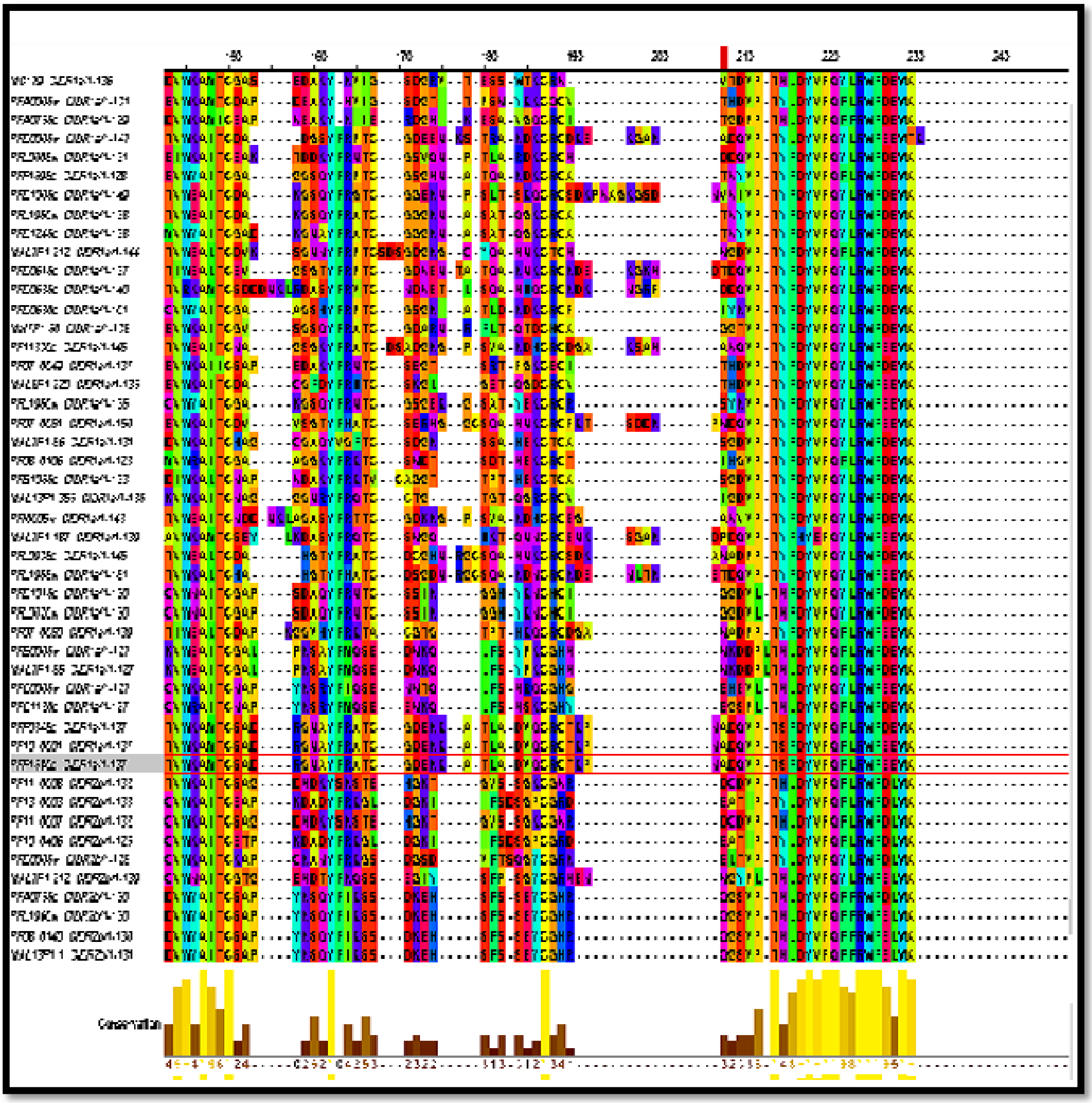


Figure 5: Shows E at 227 alignment position is conserved in CIDR1α domain not conserved in other e.g. β and γ of PFEMP1 family, these residues are functionally critical and thus have high CRE scores.

5.2 Prediction of Fold specific Residue – Result of RE Calculation

Fold specific residues, as defined in this report, are residues which are responsible for maintaining the overall fold of the protein. These residues would be conserved across the CIDR1α alignment, irrespective of the specificity of various subfamilies.

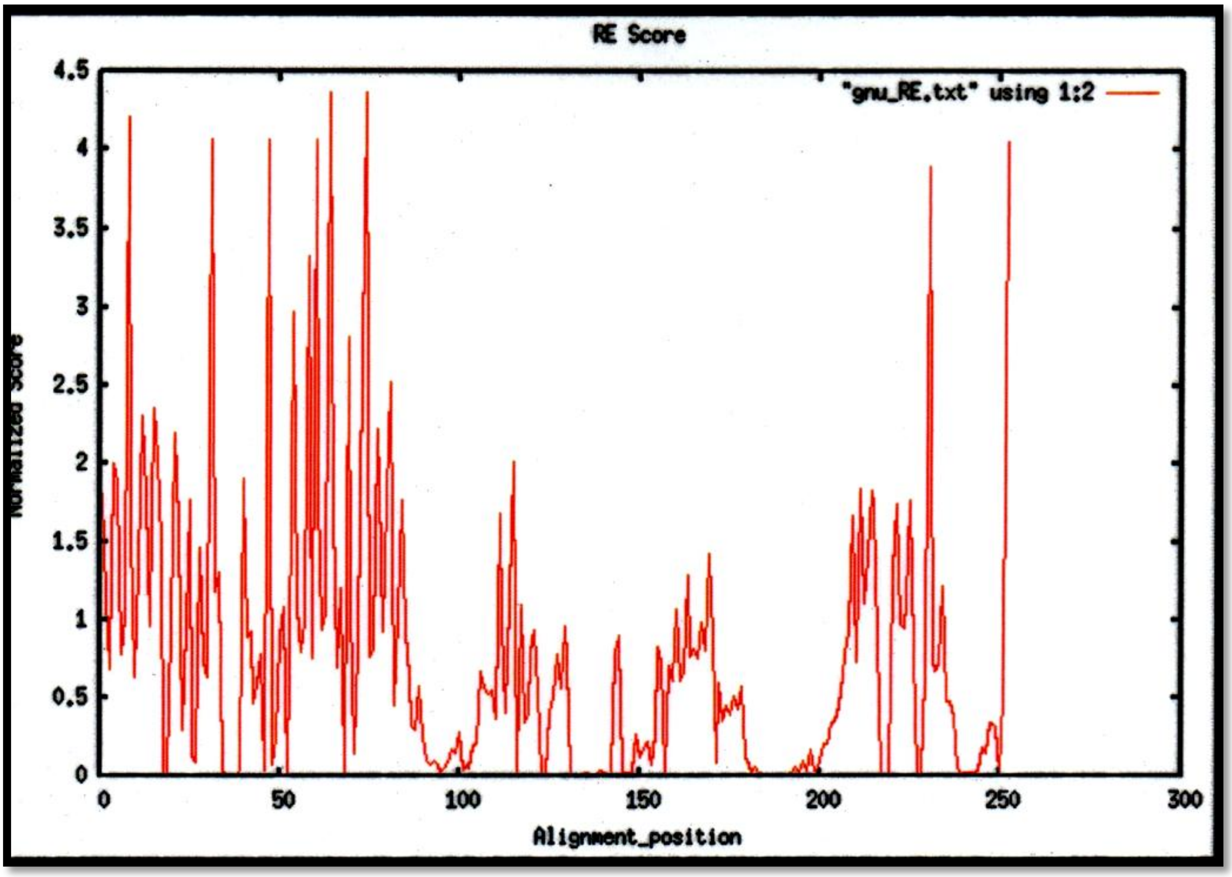


Figure 6: The Relative Entropy result of level of whole CIDR1α alignment

Relative Entropy calculation similar to conservation calculations, can predict residues that are significantly conserved throughout the subfamilies when compared to their background frequencies. Figure 6 shows the Relative Entropy (RE) results for the complete CIDR1α domain.

The x-axis is the alignment column positions of the protein sequence that is mapped in the script mapping_protein.pl. The y-axis is the z normalized RE scores obtained through the RE calculations. The x-axis, in general, spans scores all columns of the alignment from the first to the length of the protein sequence.

The tradition conservation scores consider the frequency distribution of all the amino acids across each column in the alignment. RE calculations identifies those residues whose probability distribution are significantly different from their background probability distribution.

5.3 Prediction of Function specific Residue – Results of CRE Calculation

Functional Specific residues are the residues that are differentially conserved within a subfamily with a specific function. Cumulative Relative Entropy (CRE) as defined previously can be used to identify these functionally critical residues.

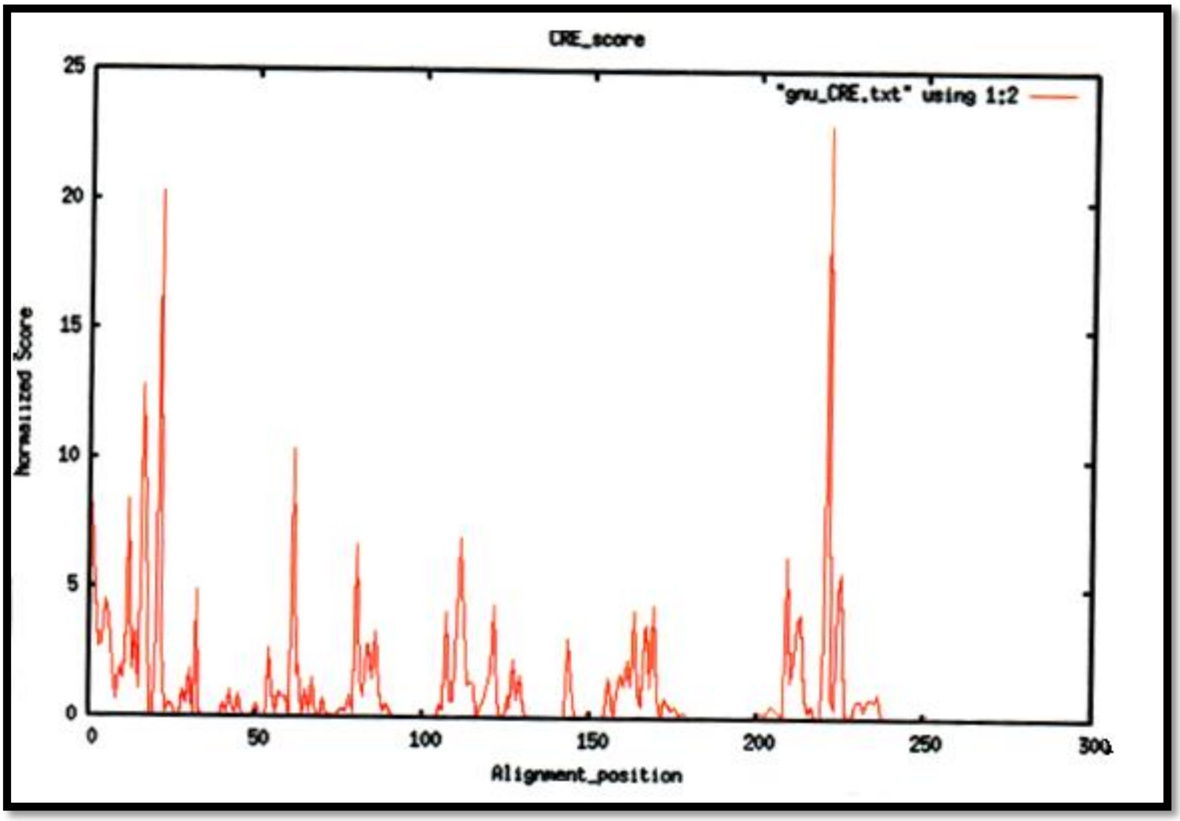


Figure 7: The Cumulative Relative Entropy results of level of whole CIDR1α alignment

A listing of residue ordered by conservation (fold and function) is provided as Table 1 and Table 2 in appendix.

5.4 Significance of Prediction – Null Model comparison

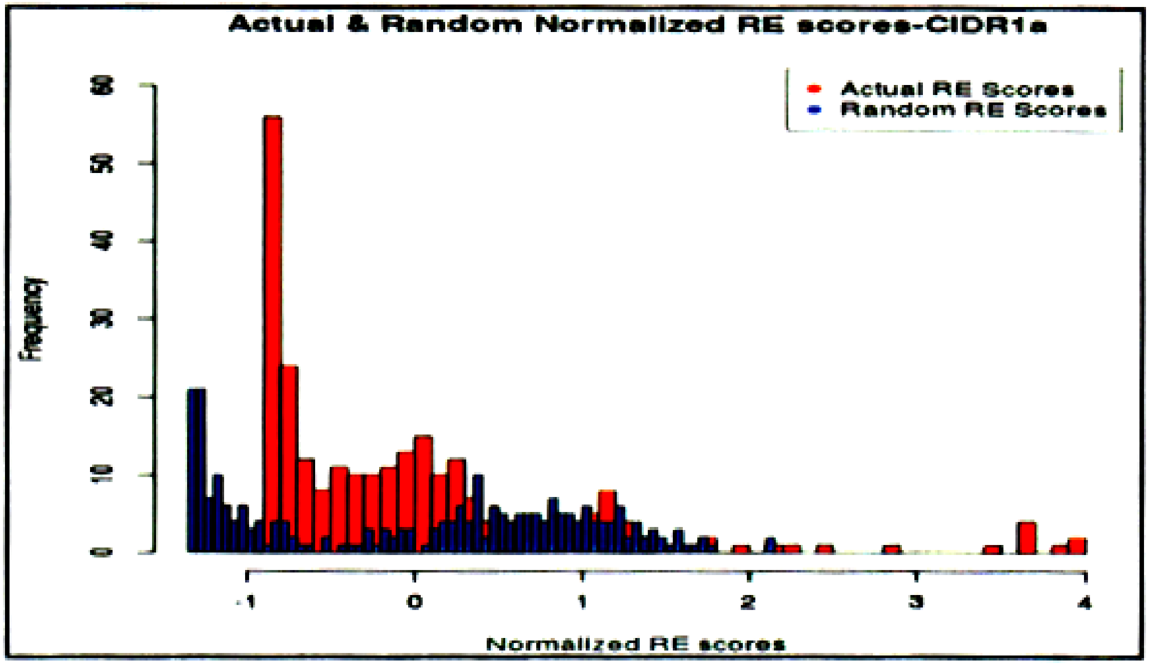
5.4.1 RE – Fold specific residues

In order to gauge the significance of these prediction, and to identify a proper threshold value to be used as a cut off, short listing those high scoring residues mentioned above we generated a null model, as explained in Section 3.4. The results from the native and null model are later compared as shown in figure 8a and figure 8b.

The x-axis is the normalized RE scores and y-axis is the frequency distribution of these RE scores. The null model has a bimodal distribution containing one sharp bar of values close to zero, and shifted to the left extreme after Z-normalization, and another smaller normal distribution, corresponding to the CIDR1α residues.

From the plots a and b the frequency distributions for the null model in contrast to the actual data tends to have a lower distribution value. These threshold points at which the distribution of the null model differs significantly from the actual data is considered as our cut off values. Cut off was interpreted from the frequency distribution graph minimum score which don't overlap with the scores calculated after randomizing the alignment. Therefore those residues with normalized RE scores greater than 1 were considered to be significantly contributing to the fold of the protein.

a)



b)

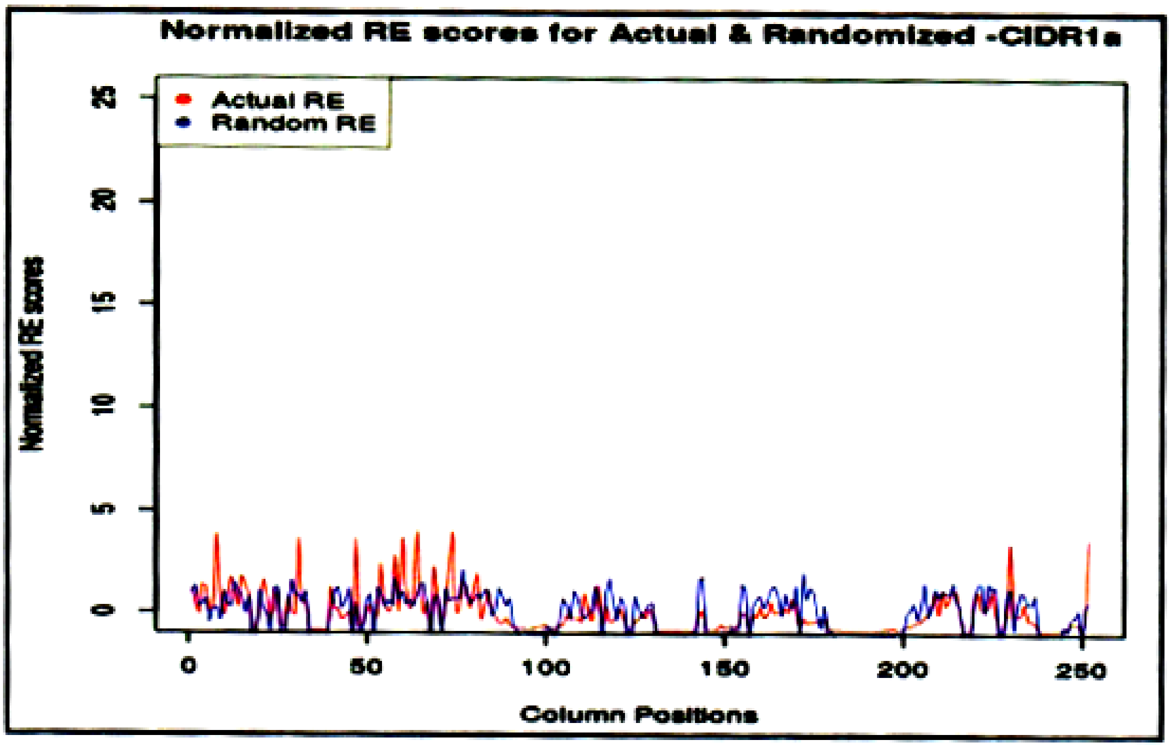


Figure 8 (a) and (b) Comparison of native and null model results for RE

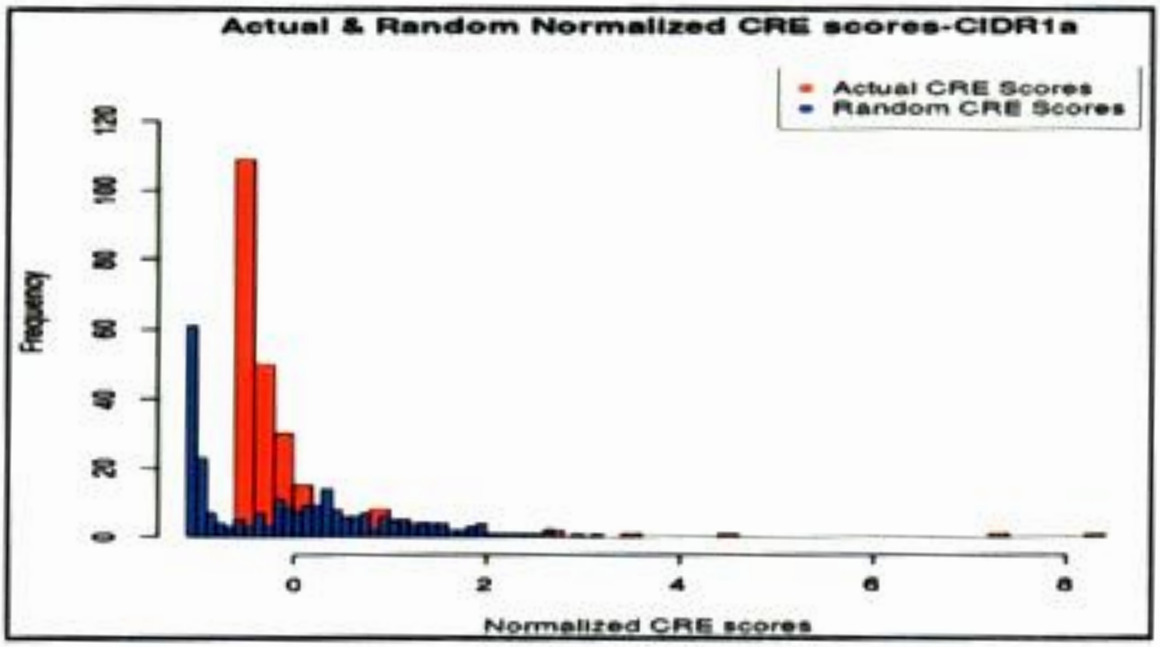
5.4.2 CRE – Function Specific Residues

Same as in the case of RE to identify a proper threshold value to be used as cut off, short listing those high scoring residues mentioned above we generated a Null model, as explained in Section 3.4. The results from the native and the null model are later compared as shown in figure 9a and figure 9b.

The x-axis is the normalized CRE scores and y-axis is the frequency distribution of these CRE scores. The null model has a bimodal distribution containing one sharp bar of values close to zero, and shifted to the left extreme after Z-normalization, and another smaller normal distribution, corresponding to the CIDR1α residues.

From the plots (a) and (b) the frequency distributions for the null model in contrast to the actual data tends to have a lower distribution value. These threshold points at which the distribution of the null model differs significantly from the actual data is considered as our cut off values. Cut off was interpreted from the frequency distribution graph minimum score which don't overlap with the scores calculated after randomizing the alignment. Therefore, those residues with normalized CRE scores are greater than 3 were considered to be significantly contributing to the fold of the protein.

a)



b)

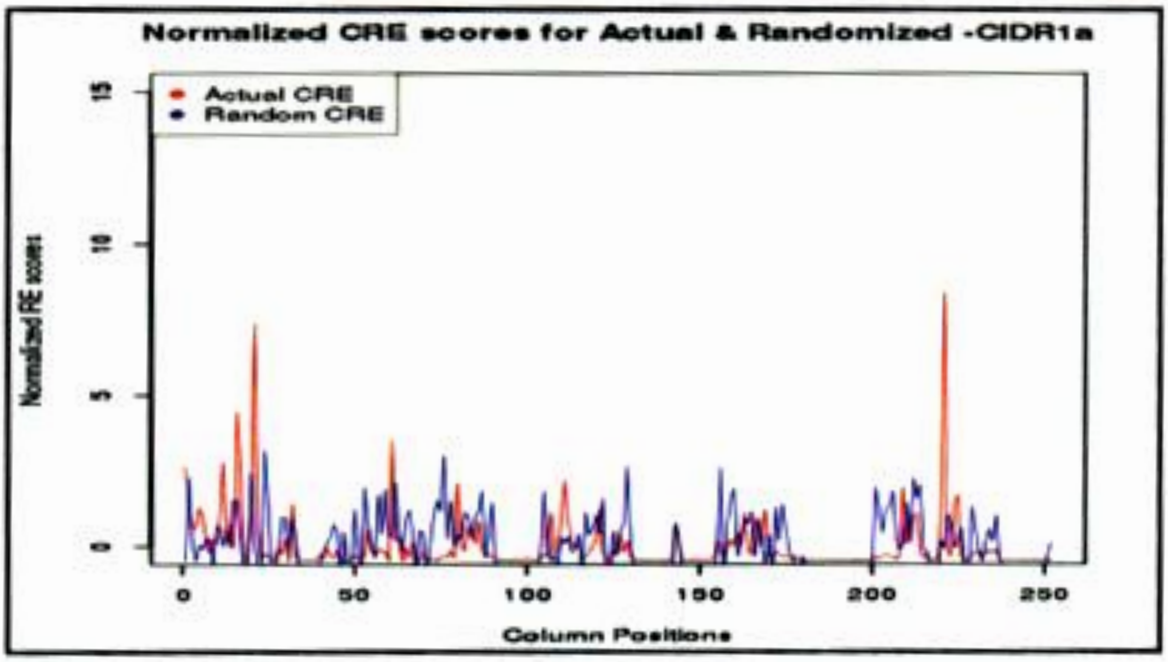


Figure 9(a) and 9(b) Comparison of native and null model results for CRE

5.5 Mapping Residue on Structure

Residues which were having higher score then cut off those residues were mapped on the structure with the help of PYMOL.

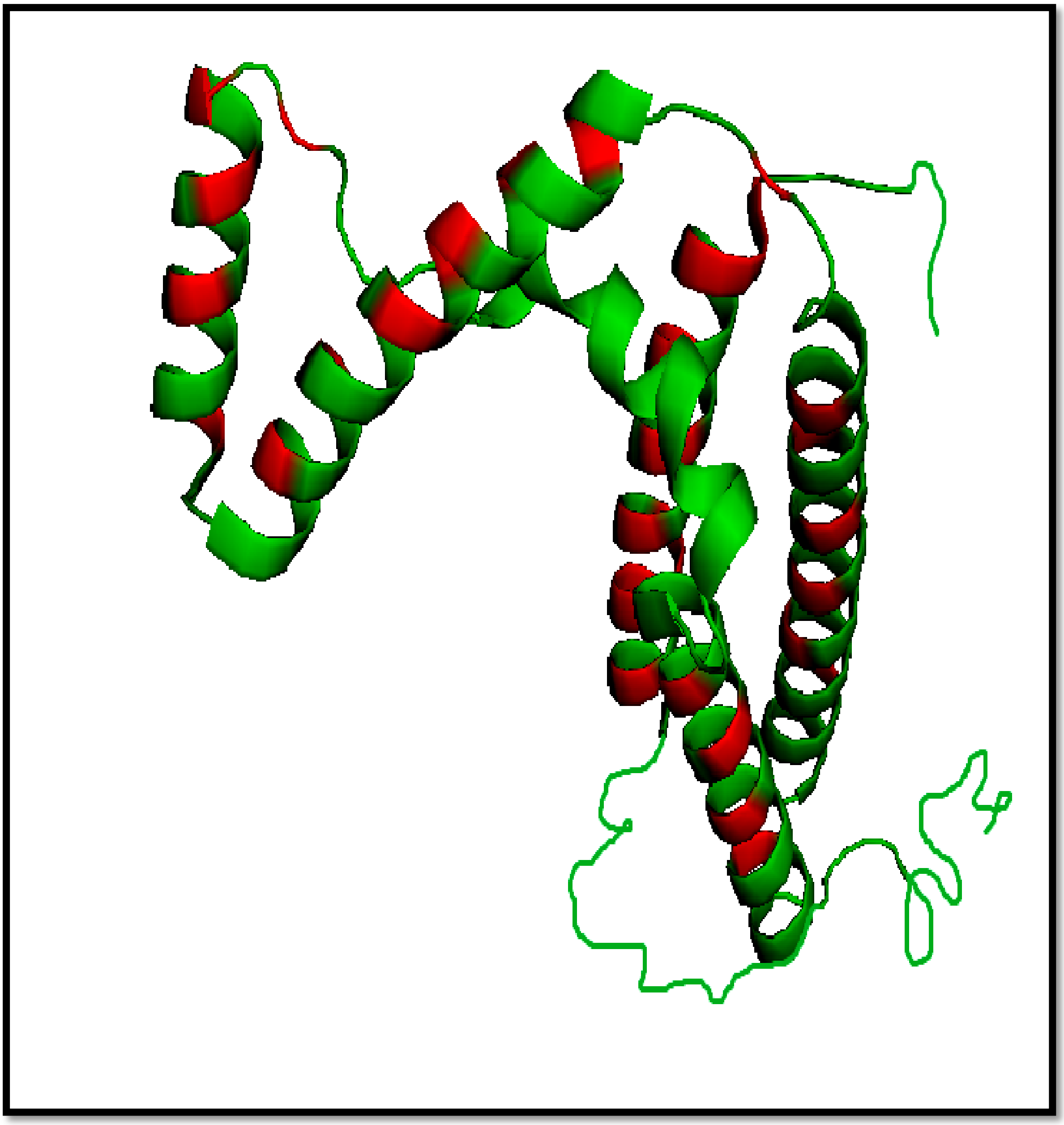


Figure 10: CIRD1α structure is shown in cartoon representation with functionally important residues in colour red.

5.6 Modelled Protein

CD36 was modelled using Modeller 9.v7 Package. The ramachandran calculated from PROCHECK for the model was 85% while the Z-score was -5.6. Which was further improved after the energy minimization ramachandran improved from 85% to 95% and Z-score become -6.3 from -5.6.

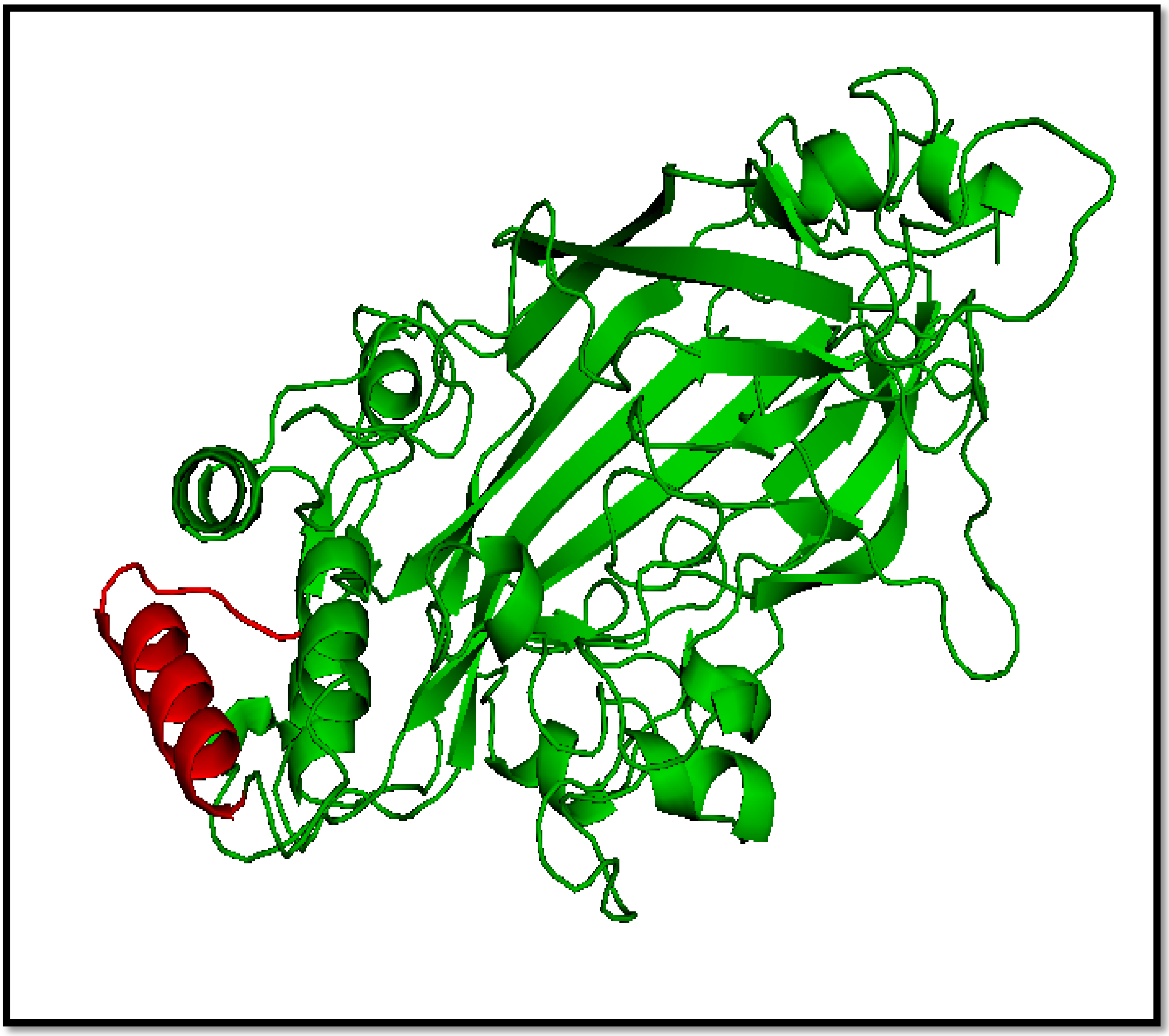


Figure 11: CD36 structure is shown in cartoon representation with functionally important region in colour red.

5.7 Protein – Protein Docking

Docked CD36 and CIDR1α complex, obtain using HADDOCK. Out of five cluster generated by HADDOCK cluster with minimum energy is shown in figure 12. Protein in red is CIDR1α domain while protein in green is CD36 receptor.

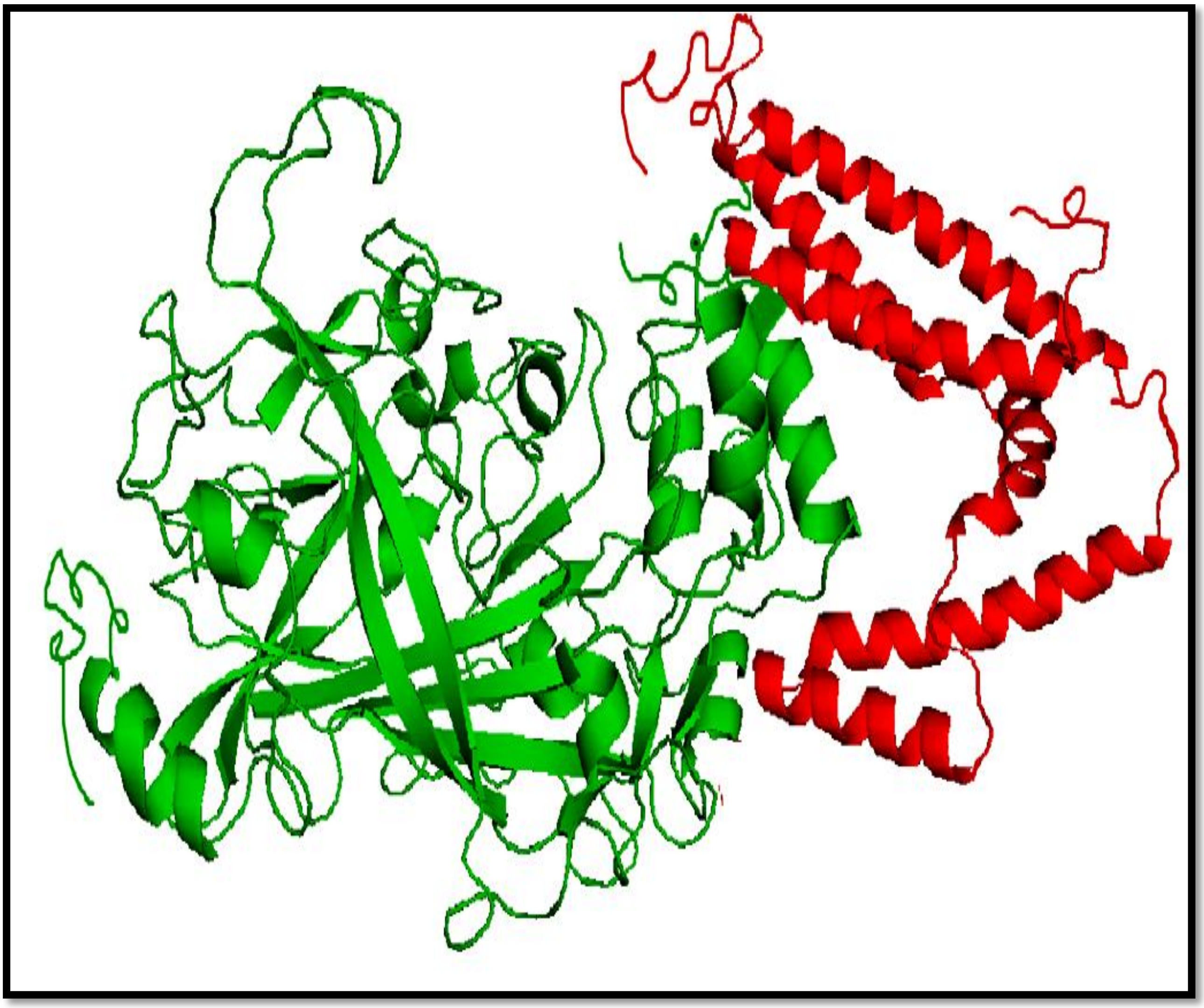


Figure 12: CIRD1α – CD36 in colour red, green respectively docked.

5.8 Comparative result of structurally and functionally critical residues

Figure 13 illustrates the web logo and graph showing the RE and CRE score of various structurally and functionally important residues. Blue line shows the RE values and Green represent CRE scores. Maximum RE score was observed at alignment positions 64 and 74 where Tryptophan ‘W’ was conserved representing structurally important residue.

However, maximum CRE score was observed at alignment position 227 where Glutamic acid ‘E’ was conserved representing functionally important residue. The cut-off value for each score was decided according to the null model. The residues size in the logo indicate the conservation pattern of the residues larger the size of residue more conserved is that residue in the alignment.

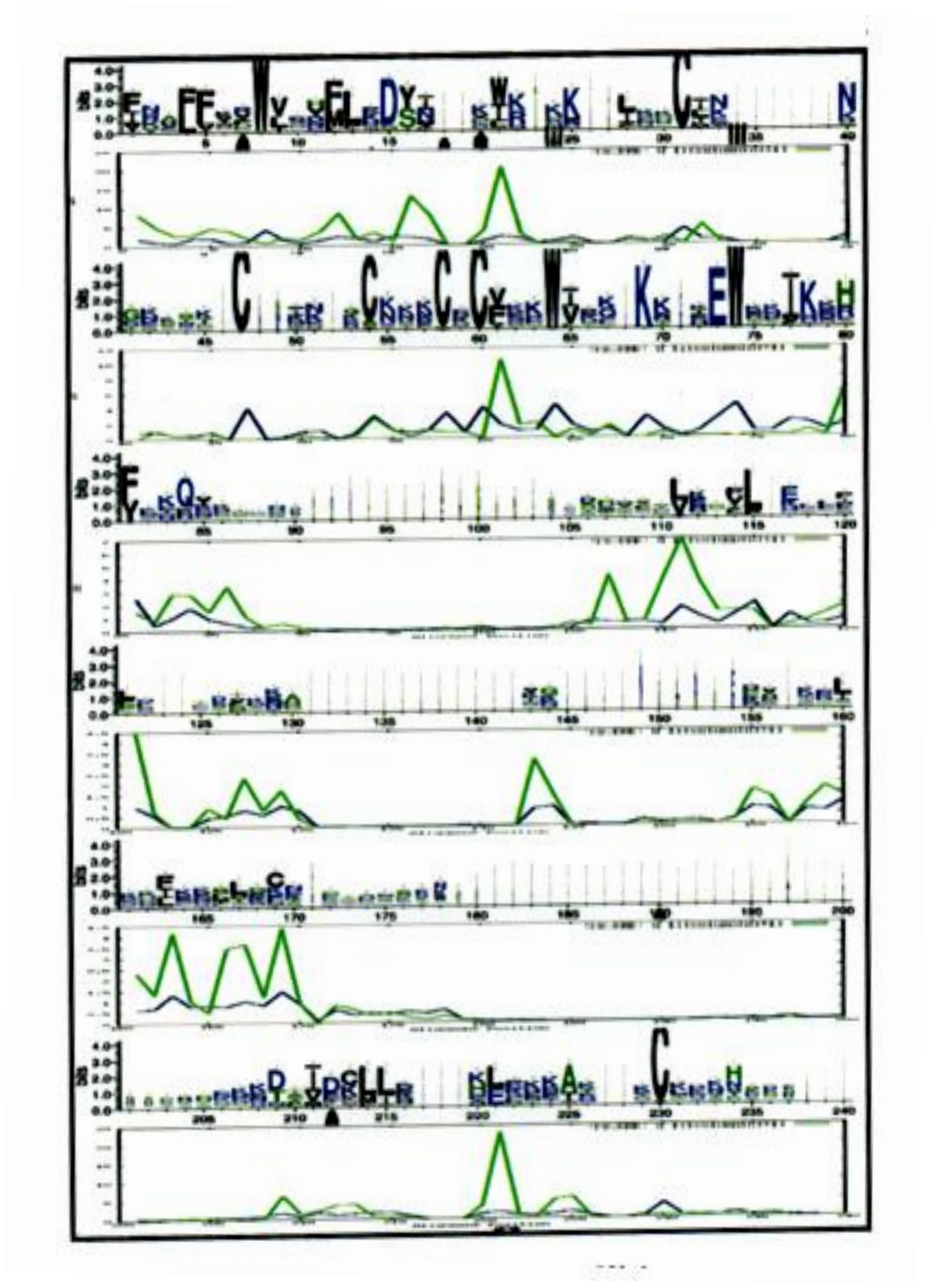


Figure 13: Web logo of MSA of 105 sequences of CIDR1α, graph showing RE (blue) and CRE (green) score.

CHAPTER 6

CONCLUSION

We replaced traditional conservation scores with the Kulback – Leibler distance to predict the conservation patterns. It was found that this approach facilitate the selection of residues that were critical for the fold and function of the protein. These Kulback – Leibler divergence is an improvised information theoretic measure that can identify residues that are conserved, differentially conserved, and residue pairs that are co-evolved, indicating pairwise interactions. There is no efficient method so far that can identify/ differentiate the substrate specific residues which largely constitutes the residues in the active site of a protein and those residues that are responsible for the native fold of the protein. These approaches when compared to the traditional techniques of conservation scores can possibly identify novel binding sites of the protein without the structural information which is necessary in most of the present cases. The use of large sequence datasets allows for the efficient separation of functionally critical residues from phylogenetic conservation, which is a common error from conservation patterns derived from smaller collections of sequences from closely related organisms.

We found out about 8 residues having high CRE scores and are lying in the 106 – 166 amino acid residue regions which proposed as important for CIDR1α interaction to CD36.

CHAPTER 7

FUTURE PERSPECTIVE

1. The critical residues predicted in case of CIDR1α can be validated through site / double site Mutational studies.

2. The knowledge of these functional residues can be useful in vaccine and drug designing.

APPENDIX

System Requirements:

All the software was run in Windows OS with normal run time and was served as window file. All the software's are open- sourced and available online. At back end python was used.

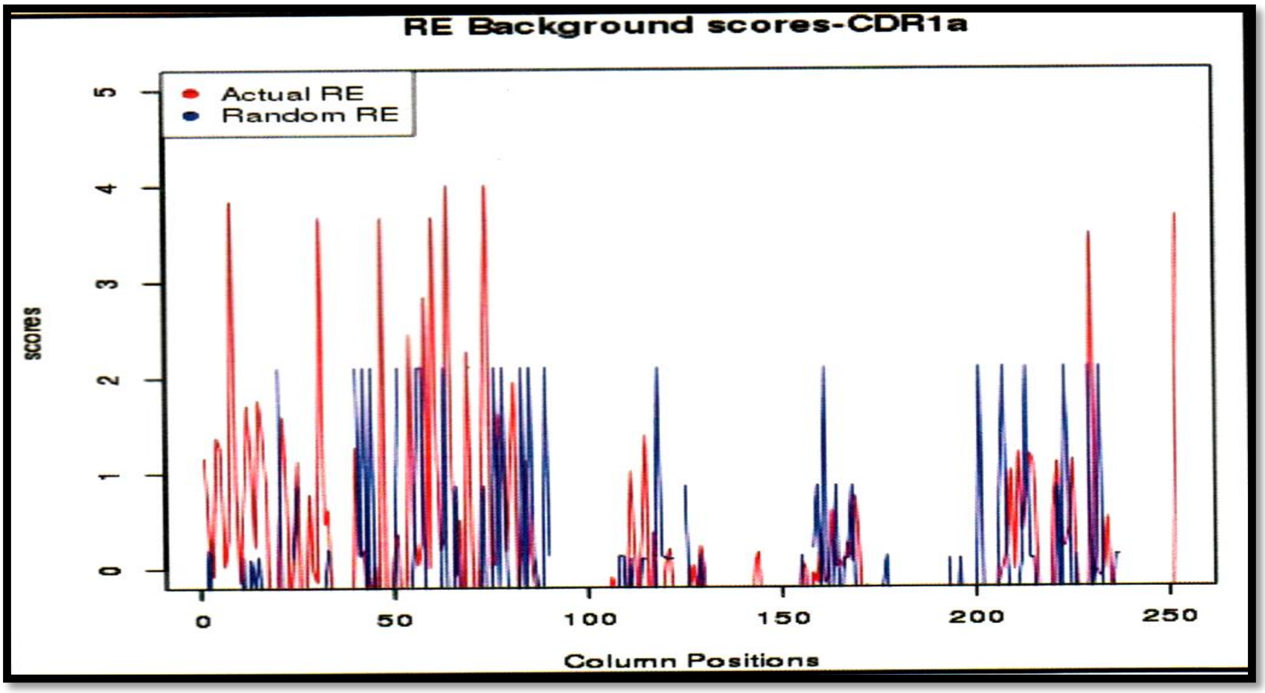
Table 1. The Fold Specific residues for CIDR1α

Alignment Column Position	RE Score	Amino Acids of the protein sequence mapped	Sequence position in the alignment
74	4.365618	W	57
8	4.21736	W	8
31	4.064422	C	26
47	4.064422	C	35
60	4.064422	C	45
25	4.064422	C	16
2			2
23	3.895122	C	15
0			3
58	3.319628	C	43
54	2.976108	C	39
69	2.809653	K	53
73	2.66462	E	56
81	2.517301	F	64
15	2.351927	D	15
12	2.302494	M	12
77	2.210297	I	60
21	2.192974	W	19
16	2.028999	S	16
11	2.009328	L	84
5			

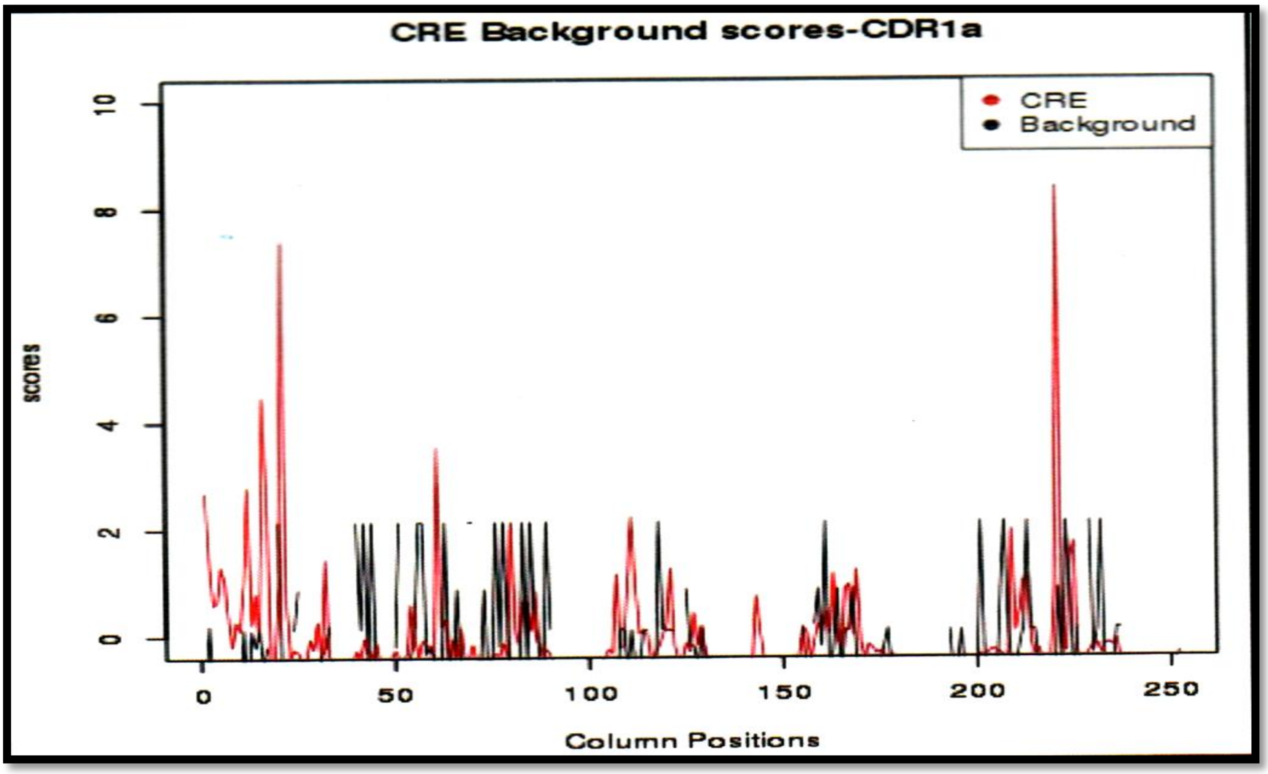
Table 2. The Function Specific Residues for CIDR1α

Alignment Column Position	CRE Score	Amino Acids of the protein sequence mapped	Sequence position in the alignment
221	22.98158	E	146
21	20.34247	W	19
16	12.82155	S	16
61	10.38857	F	46
12	8.414336	M	12
1	8.177313	Y	1
17	7.931737	I	17
111	6.962038	L	80
80	6.688169	H	63
209	6.305531	I	137
225	5.672403	A	150
32	4.926165	I	27
2	4.881423	N	2
224	4.851286	E	149
5	4.544606	F	5
121	4.423214	L	89
169	4.362882	Q	112
163	4.176237	I	106
107	4.154658	H	76
213	4.059553	K	141
6	3.874063	W	6
110	3.872491	F	79
212	3.804157	D	140
20	3.706909	K	18
11	3.670981	D	11
112	3.670781	Q	81
167	3.652108	L	110
220	3.457908	H	145
166	3.424578	L	109
86	3.41335	D	69
14	3.233952	I	14
143	3.116274	Y	97

Figure 14: The results for the comparison of background scores with RE and CRE



(a) Comparison of Background and RE scores



(b) Comparison of Background and CRE scores

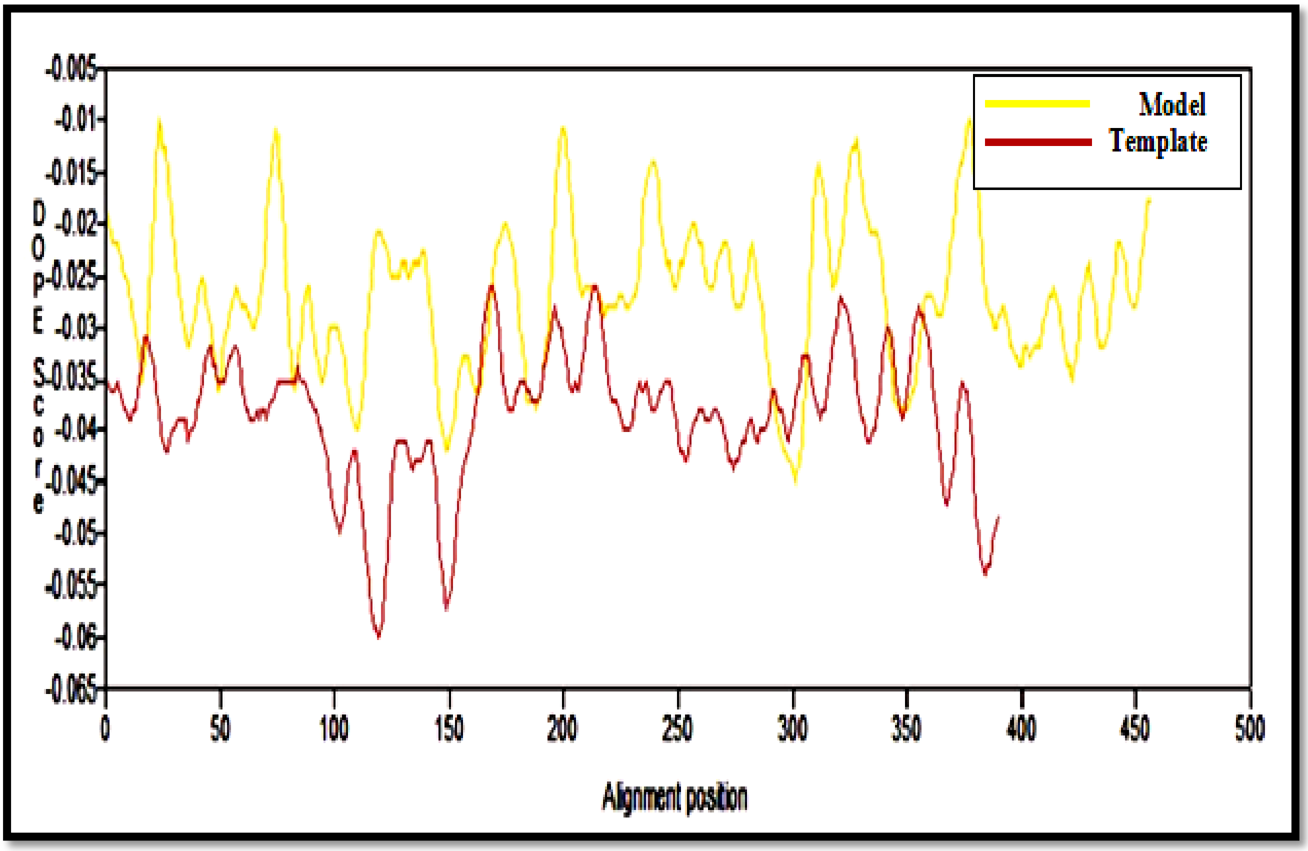


Figure 15: DOPE score profiles for the model and template

REFERENCES

1) Adams JH, Hudson DE, Torii M, Ward GE, Wellems TE, Aikawa M, Miller LH. (1990) The Duffy receptor family of Plasmodium knowlesi is located within the micromeres of invasive malaria merozoites. Cell **63**:141-53.
2) Baker D. (2010). An exciting but challenging road ahead for computational enzyme design. Protein Sciences (**19**(10), 1817-1819).
3) Baruch DI, Ma XC, Singh HB, Bi X, Pasloske BL, Howard RJ. (1997) Identification of a region of PfEMP1 that mediates adherence of Plasmodium falciparum infected erythrocytes to CD36: conserved function with variant sequence. Blood **90**:3766 – 75.
4) Baruch DI, Ma XC, Singh HB, Bi X, Pasloske BL,et al.(1995) Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell **82**: 77 – 87.
5) Chen Q, Barragan A, Fernandez V, Sundstrom A, Schlichtherle M, Sahlen A, J, Datta S, Wahlgren M.(1998) Idendification of plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) as the resetting ligand of the malaria parasite P. faciparum. J Exp Med; **187**: 15 – 23.
6) Chen Q, Heddini A, Barragan A, Frenandez V, Pearce SF, et al.(2000) Thesemiconserved head structure of Plasmodium falciparum erythrocyte membrane protein 1 mediates binding to multiple independent host receptors. J Exp Med **192**: 1 – 10.
7) Christoph Adami. (2004). Information theory in Biology. Physics of Life (Reviews (3 – 22).
8) Cover T, Thomas J. (2009). Elements of Information Theory. John Wiley and Sons Inc. Hoboken, New Jersey.
9) Dante Neculai, Michael Schwake, Mani Ravichandran, Friederike Zunke, Richard F. Collins, Judith Peters, Mirela Neculai, Jonathan Plumb, Peter Loppnau, Juan Carlos Pizarro, Alma Seitova, William S. Trimble, Paul Saftig, Sergio Grinstein & Sirano Dhe-Paganon (2013) Structure of LIMP – 2 provides functional insights with implications for SR – BI and CD36. Nature **504**, 172 – 176.
10) Elmar Krieger, Keehyoung Joo, Jinwoo Lee, Jooyoung Lee, Srivatsan Raman, James Thompson, Mike Tyka, David Baker, and Kevin Karplus (2009) Improving physical realism, stereochemistry and side – chain accuracy in homology modelling: four approaches that performed well in CASP8. Proteins **77**(Suppl 9): 114 – 122.
11) Freitas – Junior LH, Bottius E, Pirrit LA, Deitch KW, Scheidig C, et al. (2000) Frequent etopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature **407**: 1018 – 1022.
12) Hannenhali S, Russell R B. (2000). Analysis of functional Sub – types from Protein Sequence Alignment. J. Mol. Biology (**303**, 61 – 76).
13) Ho M., Schollaardt T., Niu X., Looareesuwan S., Patel K.D., Kubes P. (1998) Characterization of Plasmodium falciparum erythrocyte and P – selection interaction under flow conditions. Blood **91**: 4803 – 4809.

14) Ho M., white N. J., Looareesuwan S., Wattanagoon Y., Lee S. H., Walport M.J., Bunnag D., Harinasuta T. (1990) Splenic Fc receptor function in host defense and anemia in acute Plasmodium falciparum malaria. J. Infect. Dis.**161**: 555 – 561.

15) Kimur, et al. (1990). The Neutral Theory of Molecular Evolution. Cambridge. University Press Cambridge.

16) Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, et al. (2007) Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics **8**: 45.

17) Kraemer SM, Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the Plasmodium falciparum var gene family. Mol Microbial **50**: 1527 – 1538.

18) Kullback S, Leibler RA. (1951). On Information and Sufficiency. Annals of Mathematical Statistics (**22**, 79 – 86).

19) Kullback S, Leibler RA. (1951). On Information and Sufficiency. Annals of Mathematical Statistics (**22**, 79 – 86).

20) Laskowski RA, Rullmanm JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR **8**: 477 – 486.

21) Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG (2003) Sub – grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non coding region. Malar J **2**: 27.

22) Luse S. A., Miller L.H. (1971) Plasmodium falciparum malaria. Ultrastructure of parasitized erythrocytes in cardiac vessels. Am. J. Trop. Med. Hyg. **20**:655 – 660.

23) MacPherson G. G., Warrell M. J., White N. J., Looareesuwan S., Warrell D. A.(1985) Human cerebral malaria: a quantitative ulttratructural analysis of parasitized rytheocyte sequestration. Am. J. Patho. **199**: 385 – 401.

24) Prashant K. Srivatava, Andrew M. Lynn, et al. (2007). HMM-ModE – Improved classification using profile hidden Markov models by optimizing the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinformatics (8).

25) Robert W. Sauerwein, Meta Roestenber & Vasee S. Moorthy (2011) Experimental human challenge infections can accelerate clinical malaria vaccine development. Nature Reviews Immunology **11**: 57 – 64.

26) Rost B, Sander C. (1993). Prediction of protein secondary structure at better than 70% accuracy. J. MolBiol. (**232**, 584 – 599).

27) Sander S, Schneider R. (1991). Database of homologous – derived structures and the structural meaning of the sequence alignment. Protein (**9**, 56 – 68).

28) Sjoerd J de Vries, Marc van Dijk & Alexander M J Bonvin (2000). The HADDOCK web server for data – driven bio molecular docking. Nature Protocols **5**, 883 – 897.

29) Taylor HM, Kyes SA, Newbold CI (2000) Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. Mol Biochem Parasitol **110**: 391 – 397.

30) Tramonotano A. (2005). The ten most wanted solutions in protein bioinformatics.CRC Press.

31) Valdar WS. (2002). Scoring residue conservation. Protein (**48**, 227 – 241).

32) Warrell DA, Molyneux ME, Beales PF (1990). Severe and complicated malaria. Trans R Soc Trop Med Hyg ; **84**(suppl.2): 1 – 65.

33) Yvonne Kalmbach, Matthias Rottmann, Maryvonne Kombila, Peter G. Kremsner, Hans Peter Beck, and Ju¨ rgen F. J. Kun (2010) Differential var Gene Expression in Children with Malaria and Antidromic Effects on Host Gene Expression. JID **202** : 313 – 314.