Prediction of critical residues in PFEMP1 using Information Theoretic measures.

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Manu Kandpal

(2K12/BIO/014)

Delhi Technological University, Delhi, India

Under the supervision of

Prof. B.D Malhotra

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

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **"**Prediction of critical residues in PFEMP1 using Information Theoretic measures**"**, submitted by **Manu Kandpal (2K12/Bio/014)** in partial fulfillment 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 the candidate's own work carried out by her under my guidance.

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

Date:

Prof. B.D Malhotra
(Project Mentor)
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

Dedicated to my dad

Declaration

This thesis is a presentation of my original research work.

Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

The work was done under the guidance of **Dr. Andrew M. Lynn**, at School of Computational & Integrative Sciences (**SCIS**) Jawaharlal Nehru University (**JNU**), New Delhi.

Acknowledgement

Mere words would be insufficient to express my gratitude to the persons whose help has brought me to where I am now. Firstly, I would like to thank the **All Mighty GOD** for surrounding me with such great people and for making this world a great place for me. There are numerous people whose hope, trust and belief in me, gave me the necessary confidence to follow my dreams, the major contribution of which comes from my loving **family**.

The first person that I am grateful to is my mentor **Dr. Andrew M. Lynn**, Associate Professor at the School of Computational & Integrative Sciences (**SCIS**) Jawaharlal Nehru University (**JNU**), New Delhi. Firstly, for understanding my research interests and accepting me in his lab. Secondly for his constant encouragement, valuable suggestions and fruitful criticism that have played a great role in the successful completion of this project. His timely guidance and amicable relation throughout this project helped me to gain a significant research experience through this project.

I express my deepest regards to **Dr. B.D Malhotra**, Professor(HOD), Deptt. of Biotechnology, Delhi Technological Univesity, New Delhi for his valuable guidance at every step of my work.

I would like to thank Ms. **Swati Sinha** Doctoral Scholar from my lab for her prior insight and useful contributions in my project. I am very much grateful to her for efficiently carrying out the validation of my predictions and for her patience which paved way for the successful completion of my project.

I am also grateful to all my other lab mates ***Mr. Basharat Bhat, Mr. Kashif Nawaz, Ms. Shilpi Singh, Mr. Rishi Srivasavat*** for their support throughout my project. I also sincerely wish to thank my faculties ***Dr. Asmita Das, Dr. Yasha Hasija, Dr. Pravir Kumar*** for providing their valuable guidance throughout the course of my master.

A special thanks to my buddy ***Ayushman Srivastava*** for her valuable advice and constant motivation in each of the cafeteria discussion we had. Her support was a great backup, giving me the necessary emotional stability throughout this project and course, thanks a lot.

I am greatly in debt to one of my classmate ***Ms. Jaya Uniyal*** for his support and encouragement throughout my initial period of course work and for giving me a better insight in Biology and sharing his knowledge which was very much helpful during this project, thanks a lot.

Finally, I would like to thank the entire scientific community of Delhi Technological University and Jawaharlal Nehru University for their previous insights in various aspects of Science which have been a major foundation from which I was able to initiate my project.

Manu Kandpal
2k12/Bio/014

Contents

List of figures

List of tables

List of abbreviations

RE	Relative Entropy
CRE	Cumulative Relative Entropy
PFEMP	Plasmodium falciparum erythrocyte membrane protein
CIDR	Cysteine Rich Inter Domain Region
DBL	Duffy binding-like

Prediction of critical residues in PFEMP1 using Information Theoretic measures.

Manu Kandpal

Delhi Technological University, Delhi, India

Abstract:

PfEMP1(*Plasmodium falciparum* erythrocyte membrane protein) is an important target for protective immunity and is implicated in the pathology of malaria through its ability to adhere to host endothelial receptors. PfEMP1 has specific domains which are important in its cytoadherence function. PfEMP1 binds to CD36, an 88 kDa glycoprotein found in several cell types including platelets, monocytes, dendritic cells, and micro vascular endothelial cells. This cytoadherence of PFEMP1 to CD36 receptor is due to a specific domain called CIDR1α domain. We hypothesize that the cytoadherence function of CIDR1α to CD36 receptor is facilitated by various conserved motifs which may be targeted to disrupt the parasite cytoadherence system. In-depth knowledge of structure and function of various conserved motifs of CIDR1α is necessary for effective drug design and vaccine designing.

Herein, we will be employing computational approaches to predict fold and functionally critical residues of CIDR1α domain.For this, information theoretic scores which are variants of Relative Entropy will be calculated from Multiple Sequence Alignment (MSA) by considering distinct physico-chemical properties. The residues of CIDR1α with high RE and CRE will be predicted to be fold and functionally significant respectively.

1. Introduction

Plasmodium falciparum is the most virulent of all other species of microorganisms and responsible for maximum human deaths (Warrell DA *et.al.*, 1990). The distinct pathological characteristic of *Plasmodium falciparum* infection is that the parasite infected erythrocytes attach to host endothelium and are subsequently sequestered from the blood circulation. This enables the parasite to avoid spleen-dependent killing and survive for further transmittance. However, this may produce lethal complications in case sequestration of infected erythrocyte takes place in the vital organs.

Another adaptation of parasite to avoid immune response is by augmenting variability by regularly replacing the antigens expressed on the surface that are exposed to the host immune system. Malaria parasites contain a large family of genes for variant antigens called *var* genes that play a crucial 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 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 encodes the acidic terminal segment or ATS, that is hypothesised to anchor PfEMP1 at knobs (Su XZ *et.al*., 1995).

PFEMP1 contain two distinct adhesive modules: the Cysteine-rich Inter Domain Region(CIDR1 α)(Baruch DI *et.al.*, 1997; Smith JD *et.al.*, 1998) and Duffy binding-like(DBL) domain which is described as adhesive region in distinct Plasmodium proteins which are involved in erythrocyte invasion(Adams JH *et.al.*,1990; Sim BK *et.al.*, 1994). DBL domains bind to different molecules like intercellular adhesion molecule 1(ICAM-1)(Smith JD *et.al.*,1994), chondroitin sulfate A (CSA)(Rowe JA *et.al* 1997) and undefined heparin sulfate molecule on erythrocyte surface (Chen Q *et.al.*, 1998) .The CIDR1α domain binds to the CD36 receptors. Variation in the PFEMP1 primary sequence is such that the function of the protein remains the same. Most of the parasites isolated have the ability to bind to the CD36 receptor. This gives us a clue that there must be some important residues which remain conserved in each variant. To extract the structural and functional residues that exists in the protein family. These residues can be broadly classified as single site residues which involve 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 various subfamilies within a protein family which are responsible for substrate or functional specificity in the protein subfamily.

From a structural standpoint, fold specific residues are those that are responsible for the general scaffold common across a particular protein family and random mutations of residues on these scaffold results in paralogous proteins with a different functional or substrate specificity. The knowledge of these critical residues can lead to the better understanding of the molecular basis of diseases which arise due to altered protein functions. This knowledge also would play a crucial

role in rational protein engineering (Baker D *et.al*., 2010) and drug designing (Tramonotano A. *et.al*. 2005). Further the direct involvement of these critical residues with substrates/ligands and their involvement in maintaining the stability of protein can be efficiently refined. These structural and functional constraints embedded in a particular protein family are efficiently reflected by their Multiple Sequence Alignments (MSA). A multiple sequence alignment serves as a historical record of amino acid variability that has been accumulated 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 majority of the mutations are selectively neutral at the molecular level and do not affect the function and fold of the proteins, whereas those mutations which are deleterious 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 indicates the importance of the residues for maintaining the structure and function of proteins. Traditionally used conservation scores can identify 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 large sets of sequences 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. Therefore in order to overcome the drawbacks of this traditional scoring techniques we have made use of information theory (Christoph Adami *et.al.,* 2004) and explored measures that can accurately distinguish these critical signals with that of the background noises. We have also implemented Hidden Markov Models to estimate the probabilities (Srivastava P.K *et.al.*, 2007) of amino acids which in turn will be used as predictors in various information theoretic measures.

2. Review of Literature

2.1 Malaria

Malaria is the results in between 0.5 and 2 million deaths annually . In human malaria is caused by one of four *Plasmodium* species, namely, *P. falciparum, P. vivax, P. ovale*, and *P. malariae.* Complex life cycle of malaria parasites is completed by passing bothanopheline 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 with the bite of a female anopheline mosquito with this sporozoite stage of parasite gets transmitted. The sporozoites cleared from the circulation in about 45 min before entering hepatocytes. In hepatocyte stage asexual reproduction forming large intracellular schizont. Thousands of merozoites are contained in hepatic schizonts when they are mature (within 5–15 days of inoculation of sporozoites). Numbers of merozoites are discharged into the bloodstream, due to bursting of hepatic schizont, where they rapidly invade erythrocytes to initiate the erythrocytic cycle. Specific erythrocyte surface receptor mediate merozoite attachment to RBCs . In erythrocyte development of the parasite take place inside a membrane bound parasitophorous vacuole first as trophozoite and then as schizont. As schizont matures RBC ruptures releasing number of merozoite which reinfect fresh RBCs.

2.2 Cytoadherence

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

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

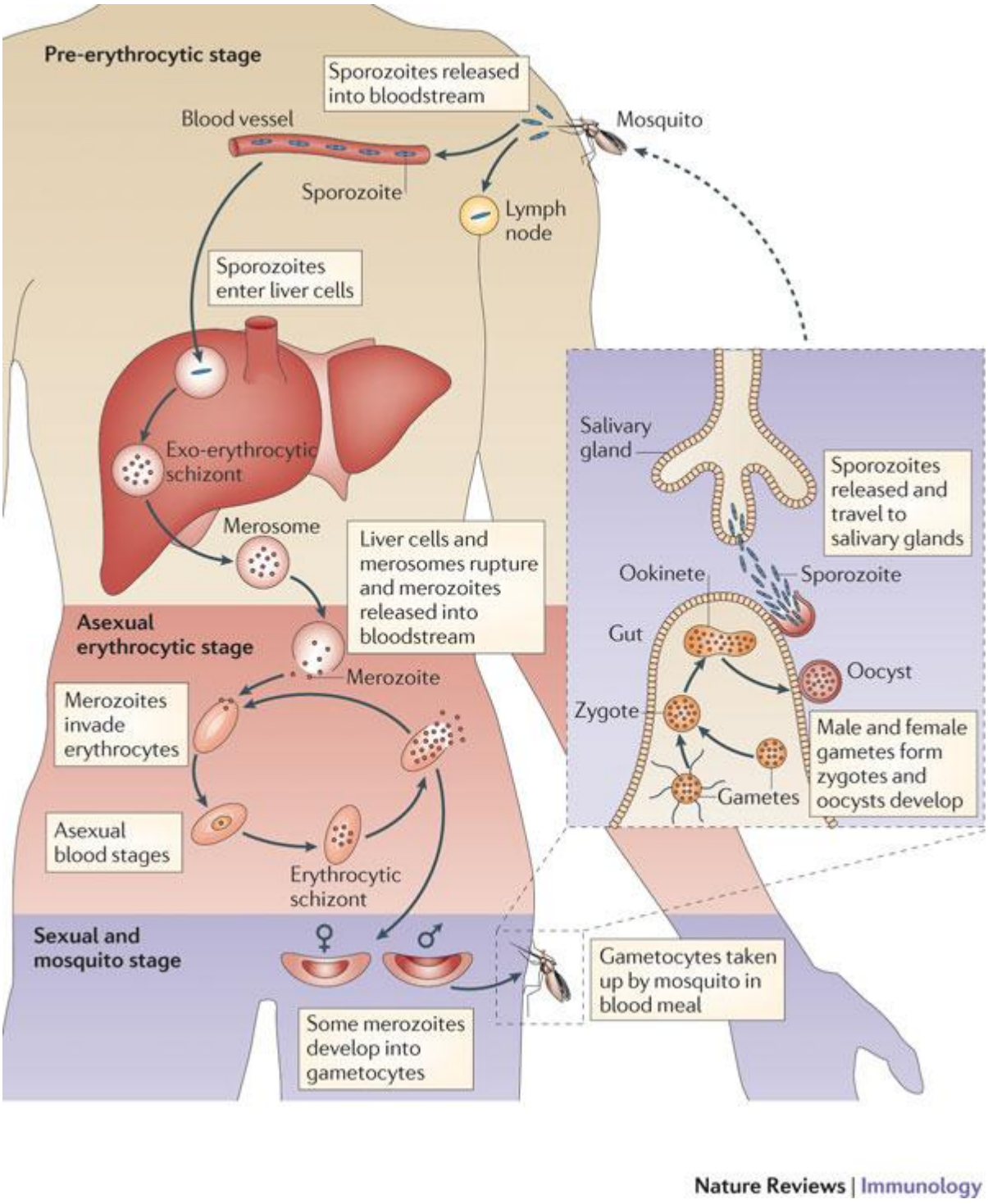


Figure 1 Malaria Life Cycle(Robert W. S et.al., 2011)

2.3 *Plasmodium falciparum* erythrocyte membrane protein 1

During the merozoite stage of *Plasmodium falciparum*, *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) is expressed on the surface of infected RBC and mediates adhesion of infected erythrocytes (IE) to various host cells on the vascular lining. (Baruch DI *et.al.*,1995 , Su XZ *et.al.,1995*).PFEMP1 is encoded by ~60 var genes, majority of which are situated in the sub telomeric regions while the remaining ~40% are found centrally in the chromosomes.(Lavstsen T. *et.al*., 2003 , Kraemer SM *et.al* ., 2003)To a large extent hyper-variable *var*gene repertoire generated by frequent meiotic ectopic recombination in the mosquito abdomen, alignment of *var* genes in the nuclear periphery makes this possible (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 has been reached 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).
CIDR1α domains have been divided into three broad classes: CIDR1α, β, and γ (MacPherson G. G *et.al.,* 1985).

Among these only CIDR1α binds to CD36 receptor (Baruch DI *et.al*.,1995) CIDR1 α 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

Shannon Entropy (H) is one of the simplest and most common information theoretic scores which measures 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 is the amino acid frequency distribution in column i of the alignment.

In case of p(x) = 0 for some amino acids x, the value of the corresponding summation 0 logb0 is taken to be 0. Generally, the log (e) considered is the natural logarithm, log(e). The logarithm base 2 (log2) is also common in use and in this specific case; the unit of entropy is 'bit'. However, it is immaterial which logarithm is used as all logarithms are proportional.

Shannon entropy would be maximum 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) or the Kullback-Leibler divergence (KL divergence) was originally introduced by Solomon Kullback and Richard Leibler in 1951 as the direct divergence between two distribution (KullbacSK *et.al.*,1951). It is often used to compare two probability distributions (Cover T *et.al.*, 2009) and is used to measure the difference of an amino acid distribution P from some background distribution P null. The RE score of a column i is defined as:

$$RE_i = \sum_{x=1}^{20} p_i(x) \log \frac{p_i(x)}{p_{null}(x)}$$

where P null is the background probability of amino acid x which is generally calculated as the probability of finding an amino acid x in all available protein sequences ie, protein sequences in Swiss-Prot database.

Relative Entropy has the property, it is always greater than or equal to zero. The Relative Entropy achieves its maximum value if the amino acid alone is observed which is the least probable according to the background distribution. It is often useful to think RE as the distance between the probabilities of distributions P and P null.

2.4.2 Cumulative Relative entropy (CRE)

Hannenhalli and Russel represented CRE method for identification of Specificity Determining Residues (SDRs) given an alignment and its classification into subfamilies (Hannenhali S *et.al;* 2000). For an alignment position i, the CRE is calculated as:

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

where pi(x,y1) and pi(x,y2)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 preformed a large scale assessment of their method by applying PFAM collections of multiple sequence alignment partitioned into subfamilies by using Swiss-Prot functional assignment. The good performance of the method has been shown by the fact that the predicted SDRs were in close agreement with the experiment.

3. Aim and Objective

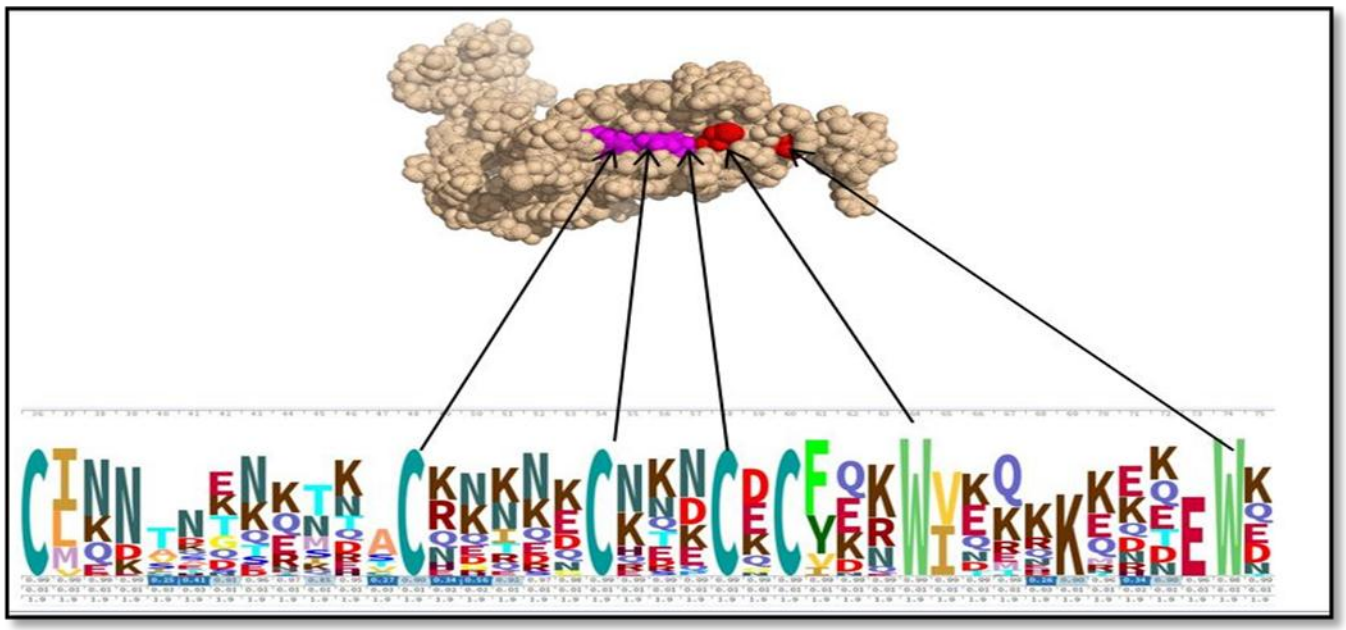
This study is aimed at the identification, modification and implementation of information theoretic measures for the prediction of critical residues from an sequence analysis perspective. To maintain fold and function of a protein family sequences various groups of residues to follow different conservation patterns across various subfamilies which collectively are termed as critical residues.These conservation patterns identified from the multiple sequence alignments.

From a sequence analysis standpoint, fold determining residues are conserved throughout the family while the 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.

Proteins function can be studied hierarchically, e.g.,the broader function of a GPCR family is signal transduction, but at a finer level the binding sites of these signal 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 conversations 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)

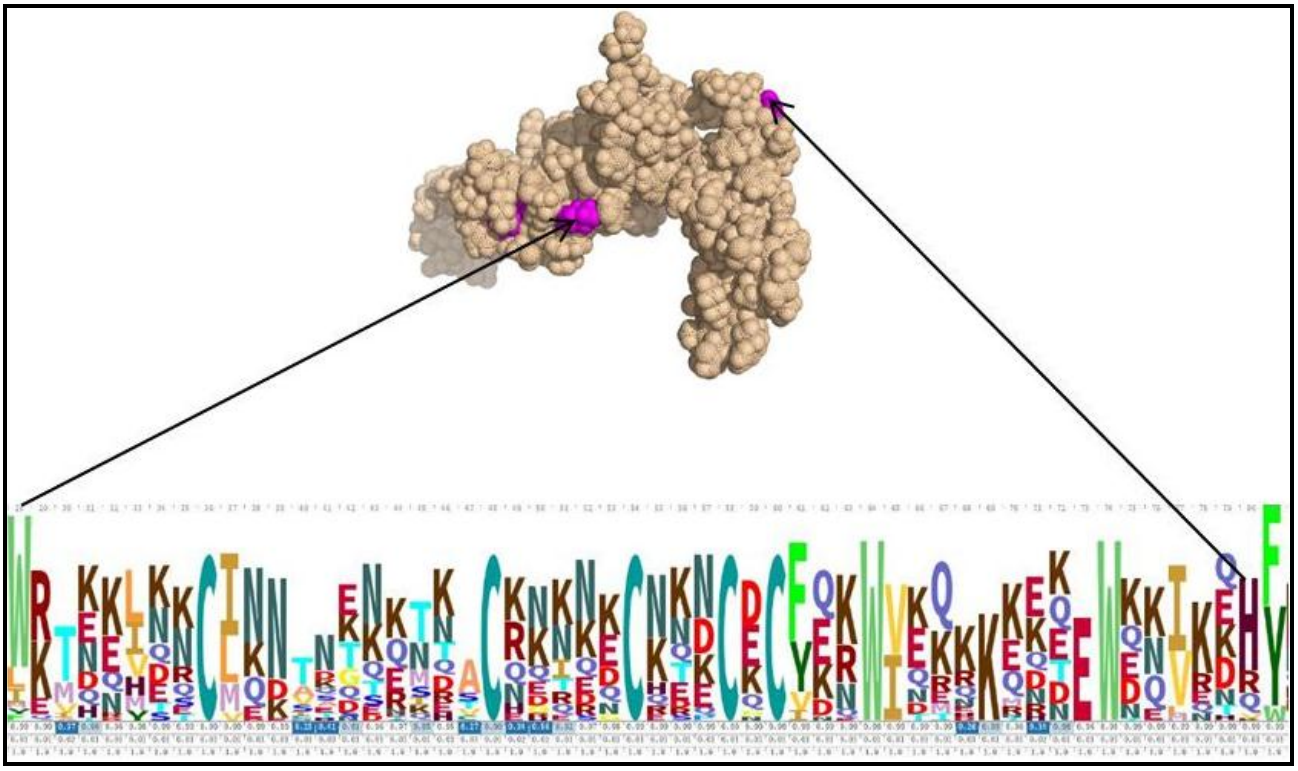


Figure 2. Schematic description of sequence conservation, and its implication on protein function. a) show residues conserved across the alignment. These are responsible for the broadfunction or thermodynamic integration of the protein. b) patterns of differential conservation are seen in the case.

4. Materials and Methods

4.1 CIDR1α domain sequences

The protein sequences of PFEMP were obtained from the CIDR1A(64). These sequences were trimmed to get CIDR1α and CIDR1α domain using local pair alignment using Mafft.
CIDR1α domain was obtained by performing local pair alignment with MC179.

4.2 Building Multiple Sequence Alignment

Mafft is used for the localpair alignment of all CIDR1A1 sequences. We used standalone version of Mafft.

```
$mafft --maxiterate 1000 --localpair  input.fasta > alignment.fasta
```

Where ***input.fasta*** are sequences in fasta format and ***alignment.fasta***is resulting aligned sequences.

4.3 Prediction of Fold and Function specific residues

4.3.1 Calculation of Relative Entropy (RE) Scores

As explained in Section 2.4.1, 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 as shown below:

```
$ perl background_prob.pl alignment.fasta
```

where***alignment.fasta*** is the input alignment file from Section 4.2 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 using *RE.pl*.

```
$ perl RE_family.pl alignment.fasta background.txt
```

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*

$RE_i = RE_i \ X \ S_i$

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

```
$ perl scaling.pl alignment.fasta alignment_RE
```

where***alignment_RE*** is the raw RE score obtained from previous step, *alignment.fa* is the alignment file, the output file***alignment_RE*s** is the scaled RE scores.

```
$ perl mapping_protein.pl alignment_Res alignment.fasta id
```

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 protein 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. This was done by shell editors available in Linux as shown below.

```
$ grep ">" subfamily.fasta |sed -e 's/>//g' > subfamily

$ grep ">" rest.fa | sede 's/>//g' > rest

$ list subfamily rest > list

$ perl sorting_seq.pl alignment.fasta list
```

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

```
$perl RE_subfamily.pl subfamily1.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,

```
$perl RE_subfamily.pl rest.fasta subfamily1.fasta
```

The output file is *subfamily12_RE*s.

```
$ perl scaling.pl rest.fasta subfamily21_RE
```

The output file is *subfamily21_RE*s.

As mentioned earlier in Section 4.3 *scaling.pl* makes correction for gaps in the scores obtained.

```
$ perl RE_family.pl subfamily1.fasta
```

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 *subfamily1_RE* which is later scaled as:

```
$ perl scaling.pl subfamily1.fasta subfamily1_RE
```

whose output is subfamily1_REs.

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 weigh more for these residues than for the others, thereby giving this formula for CRE calculations:

$CRE_i = (RE12_i + RE21_i) \times (RE1_i)$

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

```
$perl CREs.pl subfamily12_REs subfamily1_REs subfamily21_REs
alignment.fasta id
```

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

4.3.4 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 criterias (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.

4.3.5 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 (2.4.1, 2.4.2) 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 might 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.

Rows shuffling were done using the *script rand1.pl*

```
$ perl rand1.pl <alignment.fa><output.fa>
```

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

The above procedures for RE and CRE calculations (Section 4.3, 4.4, 4.5 and 4.7) were later implemented on the randomized datasets, which apartfrom 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.

```
$ R RE_random.R RE_result rand_RE_results Z_results
```

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 are those residues that are significant greater than the threshold value obtained from the null model

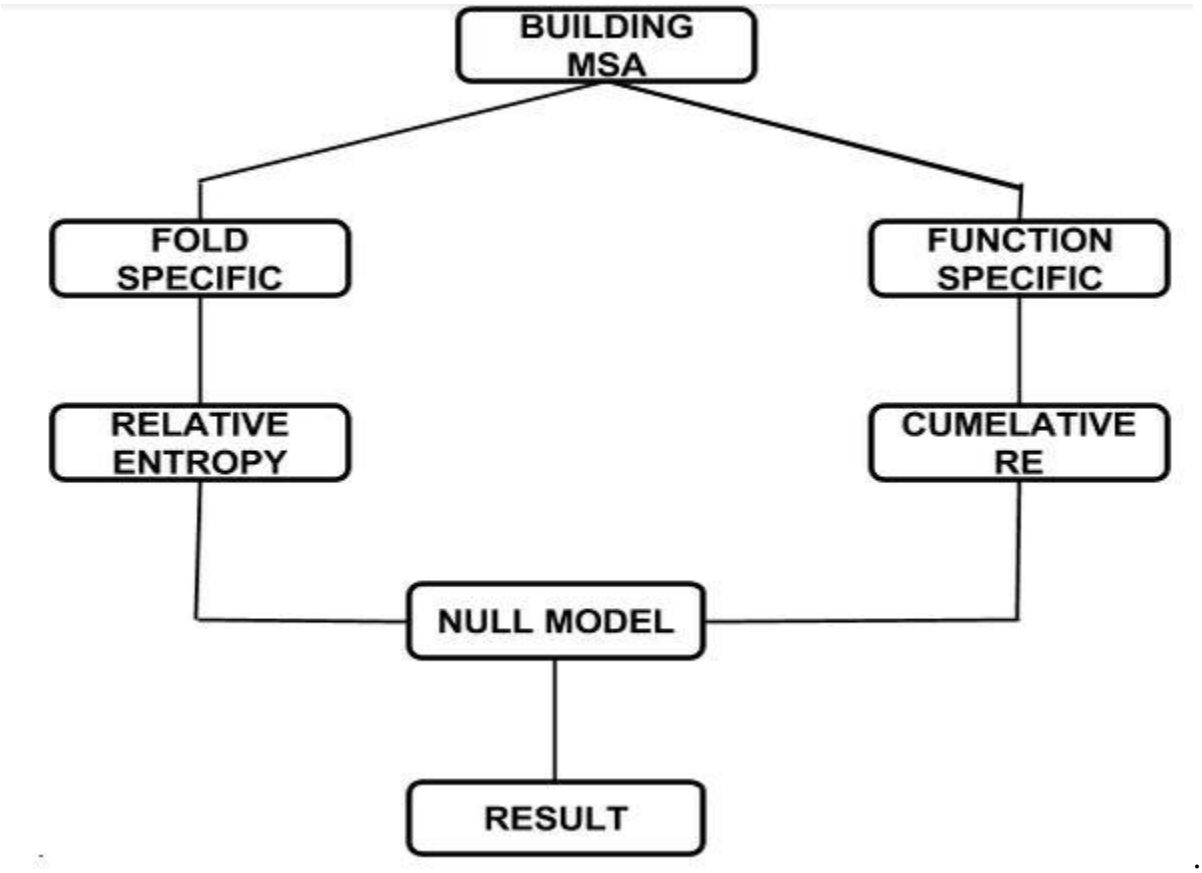


Figure. 3 The work follow of various methodologies that were implement in this thesis.

4.4 Mapping of Residues on Structure

Mark the important residues on the protein with the help of PYMOL.

4.5 Modelling CD36

Homology model of CD36 sequence was generated using Modeller9v7 Package. 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 modl of CD36 sequence are as follows:

4.5.1 Template identification

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

4.5.2 Aligning CD36 with the template

Python script align2d.py is used to align the CD36 with the tamplate structure. It use s 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 similaryty and harbouring more number of gaps in the alignment.

Alignment of CD36 with 4F7B (PDB ID of template)

```
_aln.pos        10       20       30       40       50       60
4F7BA    ------------------------------------IEKKIVLRNGTEAFDSWEKPPLPVYTQFYFF
CD36
   MGCDRNCGLIAGAVIGAVLAVFGGILMPVGDLLIQKTIKKQVVLEEGTIAFKNWVKTGTEVYRQFWIF
_consrvd                            * *  **  ** **  * *   ** **  *

_aln.p   70       80       90      100      110      120      130
4F7BA    NVTNPEEILRGETP-RVEEVGPYTYR-
ELRNKANIQFGDNGTTISAVSNKAYVFERDQSVGDPKIDLI
CD36     DVQNPQEVMMNSSNIQVKQRGPYTYRVRFLAKENVTQDAEDNTVSFLQPNGAIFEPSLSVGTEA-
DNF
_consrvd  * ** *       *   ******    * *      * *      **   ***    *

_aln.pos  140      150      160      170      180      190      200
4F7BA
   RTLNIPVLTVIEWSQVHFLREIIEAMLKAYQQKLFVTHTVDELLWGYKDEILSLIHVFRPDISPYFGL
CD36     TVLNLAVAAASHIYQNQFVQMILNSLINKSKSSMFQVRTLRELLWGYRDPFLSLVPY--P-
VTTTVGL
_consrvd   **  *     *  *  *         *  *  ******* *  ***    *     **

_aln.pos    210      220      230      240      250      260      270
4F7BA
   FYEKNGTNDGDYVFLTGEDSYLNFTKIVEWNGKTSLDWWITDKCNMINGTDGDSFHPLITKDEVLYVF
CD36     FYPYNNTADGVYKVFNGKDNISKVAIIDTYKGKRNLSYW-
ESHCDMINGTDAASFPPFVEKSQVLQFF
_consrvd ** * * *** *   * *      *   **  * *   * ******  *** *  * **  *

_aln.pos     280      290      300      310      320      330      340
4F7BA    PSDFCRSVYITFSDYESVQGLPAFRYKVPAEILAN---TSDNAGFC---IPEGNCLGSGVLNVSICKN
```

```
CD36     SSDICRSIYAVFESDVNLKGIPVYRFVLPSKAFASPVENPDNYCFCTEKIISKNCTSYGVLDISKCKE
_consrvd  ** *** * *     * * *  *   *    ** **  *  **  *** * **

_aln.pos       350      360      370      380      390      400
4F7BA     GAPIIMSFPHFYQADERFVSAIEGMHPNQEDHETFVDINPLTGIILKAAKRFQINIYVKKLDDF-
VET
CD36      GRPVYISLPHFLYASPDVSEPIDGLNPNEEEHRTYLDIEPITGFTLQFAKRLQVNLLVKPSEKIQVLK
_consrvd * *   * *** *     * * ** * * * ** * ** *  *** * *  **     *

_aln.p  410      420      430      440      450      460      470
4F7BA     GDIRTMVFPVMYLNESVHIDKETAS-------------------------------------RLKSM
CD36
    NLKRNYIVPILWLNETGTIGDEKANMFRSQVTGKINLLGLIEMILLSVGVVMFVAFMISYCACRSKTI
_consrvd    *   *  ***  * * *                          * *

_aln.pos
4F7BA     I
CD36      K
_consrvd
```

4.5.3 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 achived by automodel class of Modeller. A total of 100 3D models of CD36 were prepered as result of "model_single.py" and an output file "model-single.log" summarizimg all the models built.

4.5.4 Selection of the best model

The model can be picked with the lowest value of the modeller objective function or the DOPE assesment score and with the highest GA341 assesment score repiorted in log file "model-single.log". GA341 score ranges from 0.0 (worst) to 1.0 (native-like). The DOPE and GA341 scre are not absolute measure, they are only used to rank models calsulated from the sa,e alignment. However, DOPE score is better than GA341 in distinguishing 'good' models from 'bad' models. The GA341 was highest, 1 for all the hundered models, So, 5 lowest DOPE scoring models are picked out of 100 further evaluation.
Python script evaluate_model.py uses complete_pdb to resd in a PDB file and prepare it for energy calculation. The DOPE energy is calculated with the assess_dope command and the energy profile is smoothednedoevr a 15 residue window and normalised by the number of restrains acting on each

residue is written to the output file “4F7B.profile(for template)”, which can be ploted using python script **plot_profiles.py.**

4.5.4 Model evaluation

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

4.6 Energy Minimization of energy model

Energy minimization of 3D model was done using YASARA energy minimization server.

4.6 Protein-Protein Docking

Protein-protein docking was performed using HADDOCK web server (Sjoerd J de Vries*et.al* 2010)

5. Result

5.1 Alignment

Local pair alignment among 105 sequence of CIDR1domain was performed.

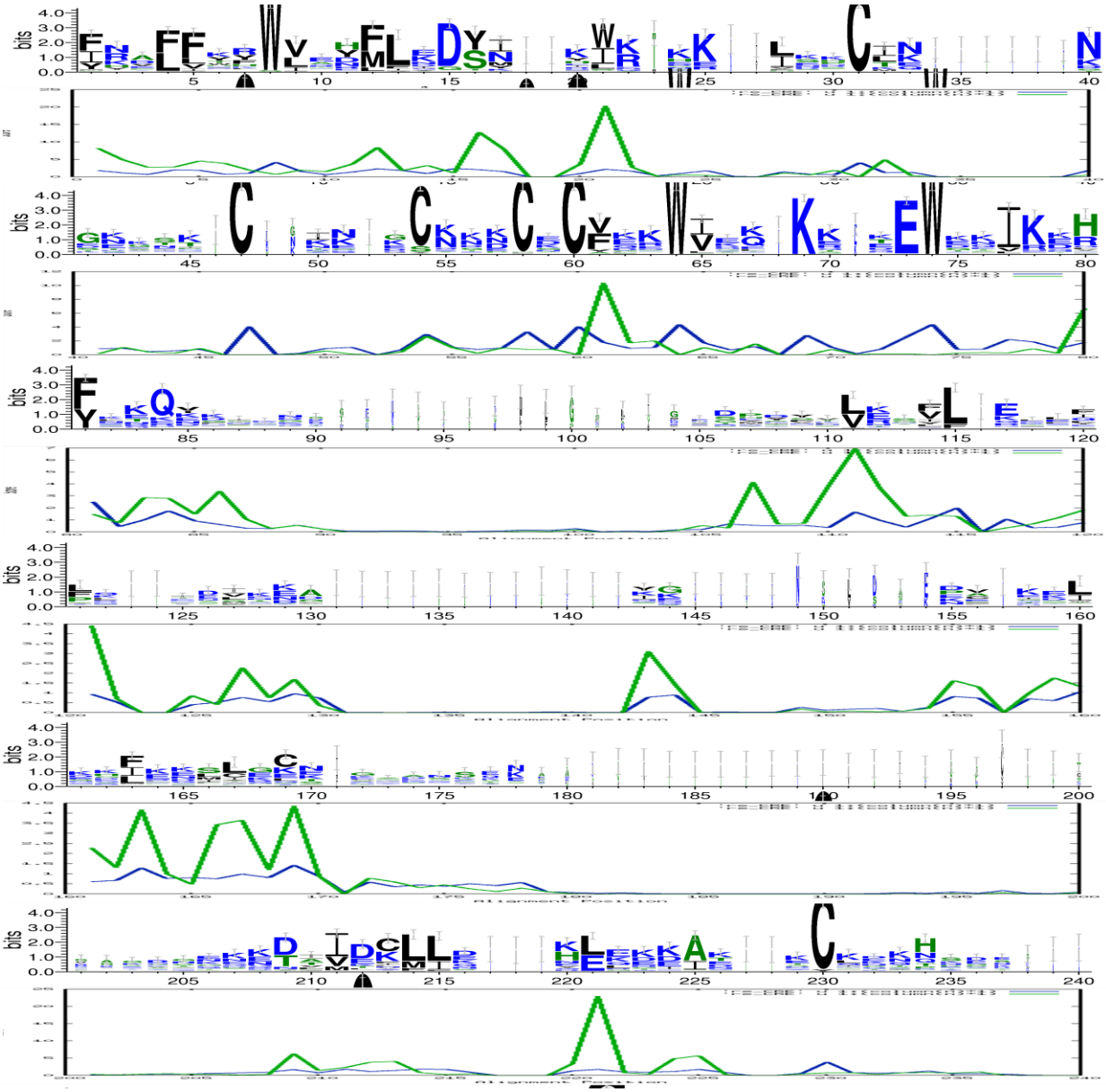


Figure 4 Web logo of MSA of 105 sequences of CIDR1α,graph showing RE(blue) and CRE(green) score

5.2 Prediction of Fold specific Residues – Results 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.

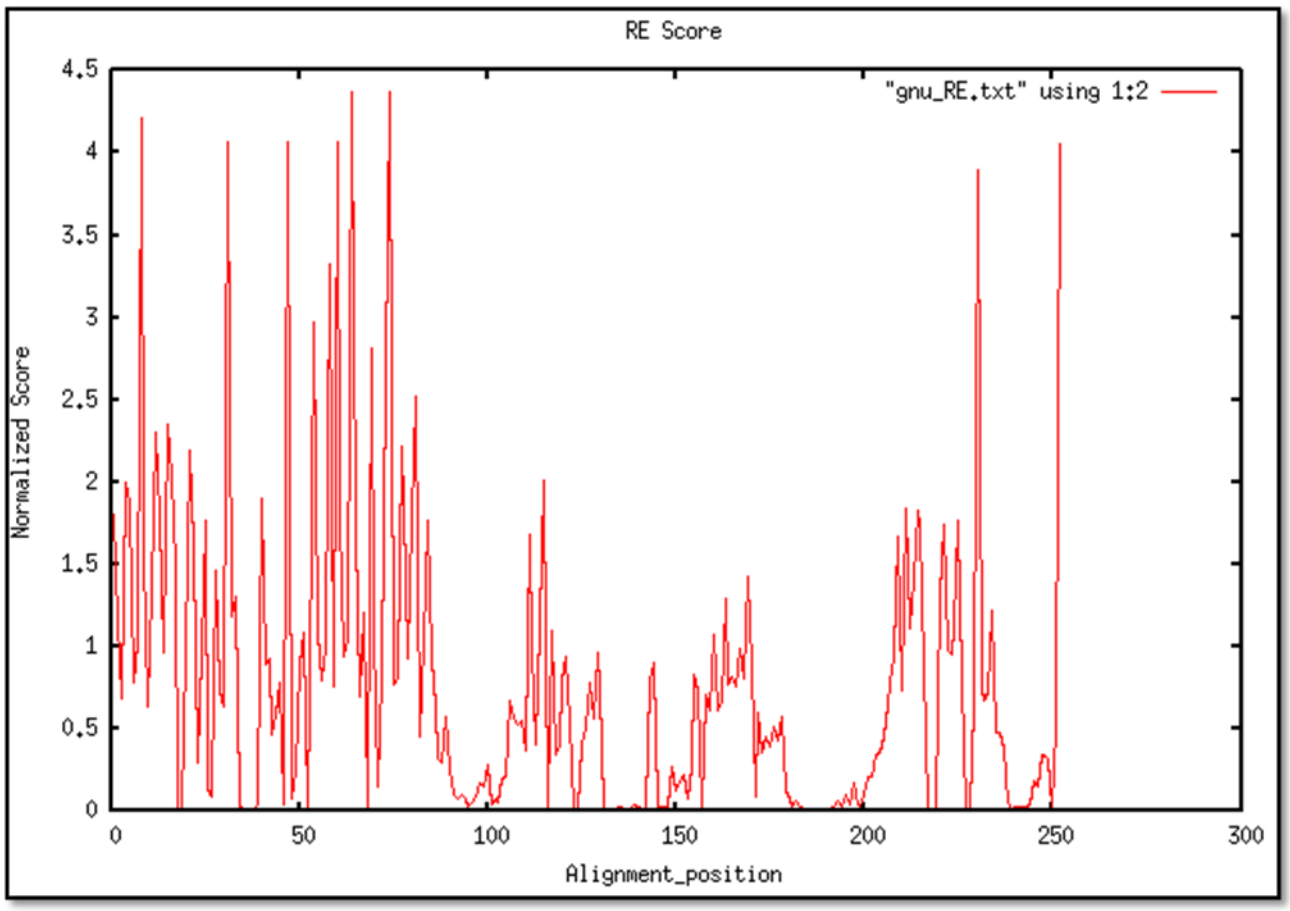


Figure 5 The Relative Entropy results of level of whole CIDR1α alignment

Relative Entropy calculation similar to conservation calculations, can predict residues that are significantly conserved throughout the subfamilies when compared to their background frequencies.Figure 5 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 across 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 columns 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 Residues – 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.

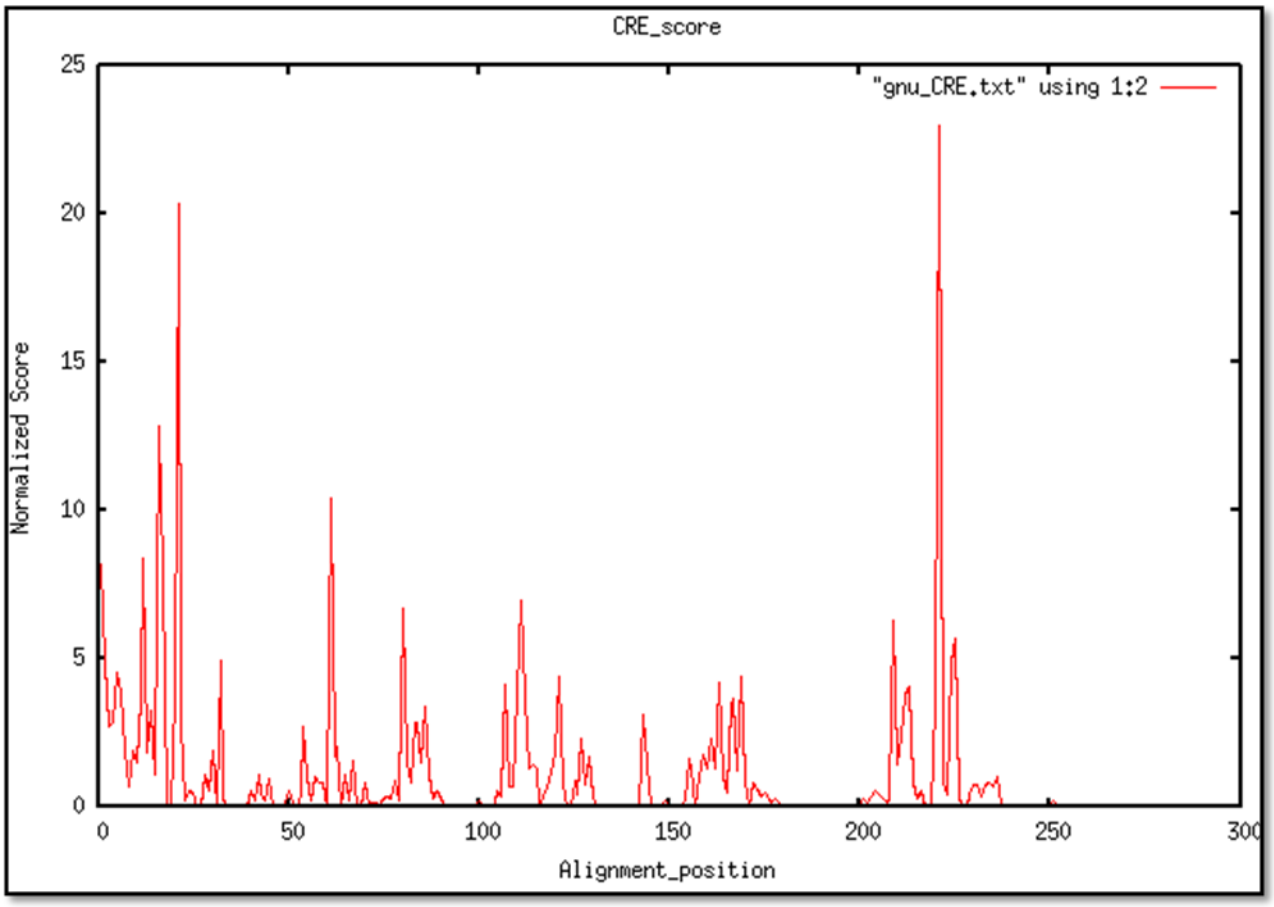


Figure 6 The Cumulative Relative Entropy results of level of whole CIDR1α alignment

A listing of residue ordered by conservation (fold and function) is provided as table 2,3.

5.4 Significance of Prediction – Null Model comparison

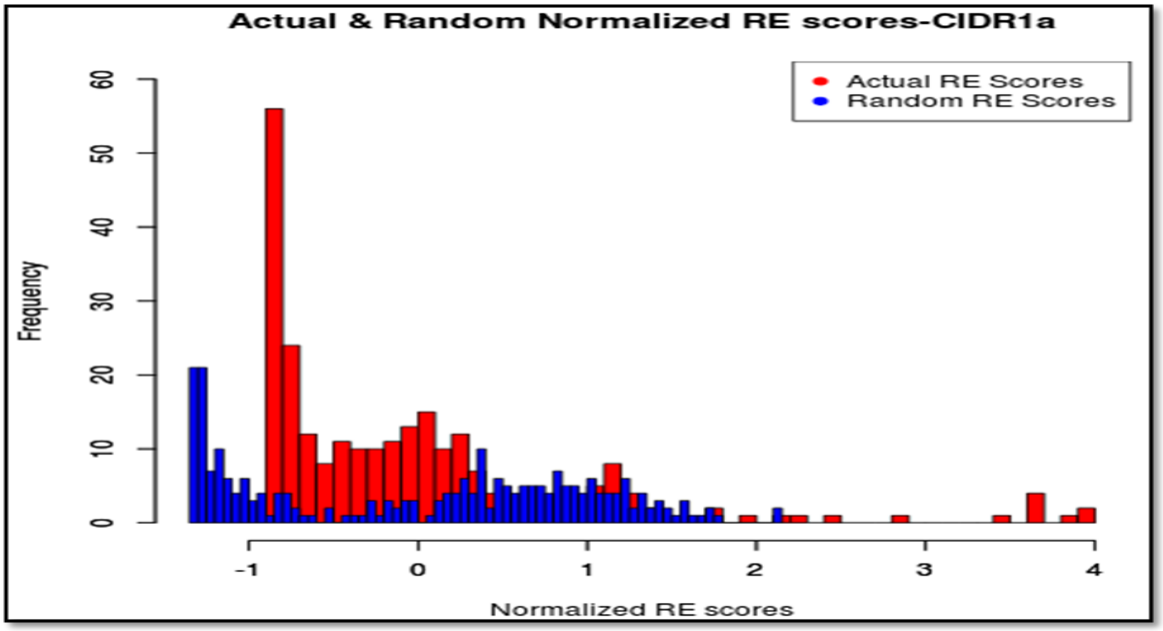
5.4.1 RE - Fold specific residues

In order to gauge the significance of these predictions, and to identify a proper threshold value to be used as a cut off, shortlisting those high scoring residues mentioned above we generated a Null model, as explained inSection 4.3.5. The results in from the native and the null model are later compared as shown in figure 7a and figure 7 b.

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. In the case of RE the cut off was found to be 1.Therefore those residues with normalized RE scores greater than 1 were considered to be significantly contributing to the fold of the protein.

a)



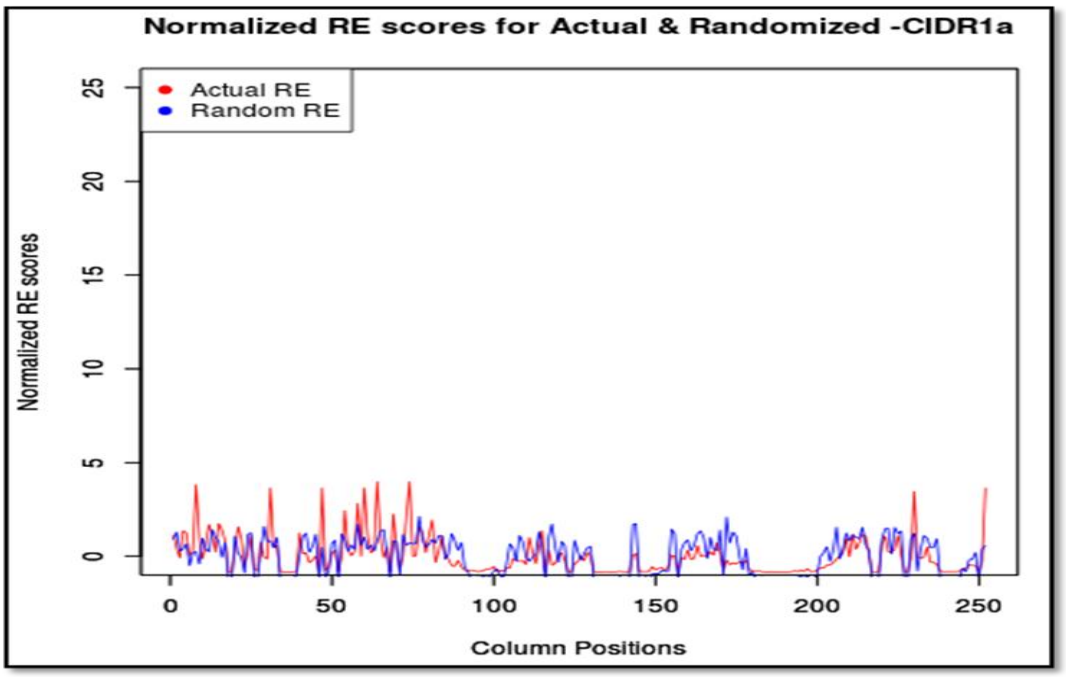


Figure 7 (a) (b) Comparison of native and null model results for RE

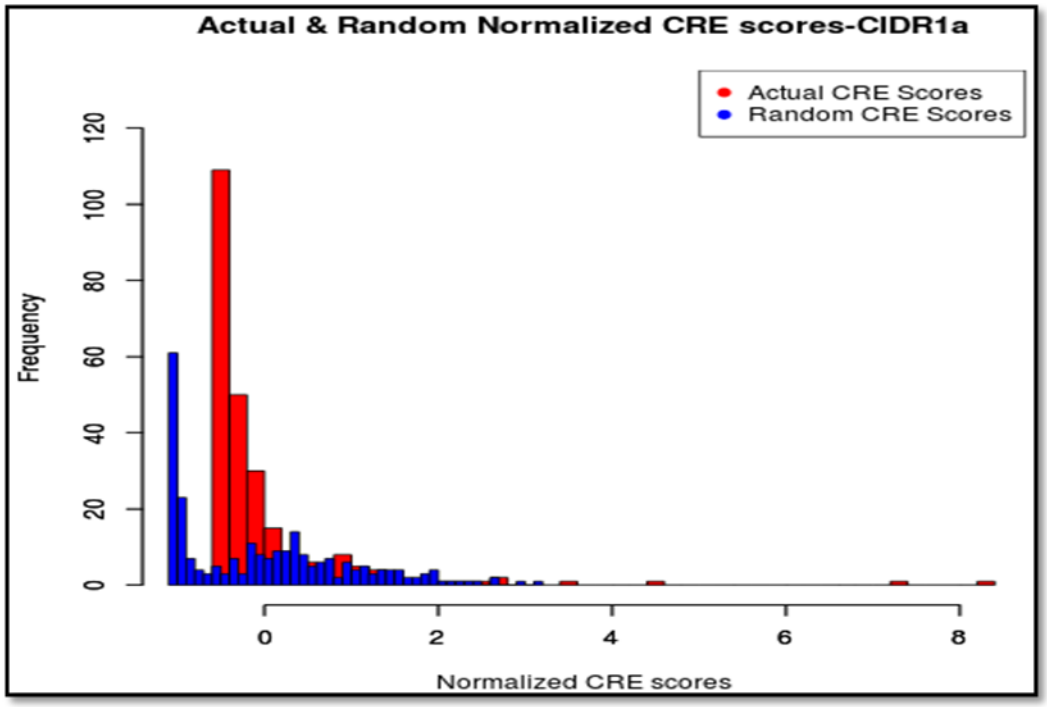
5.4.2 CRE – Function specific residues

Same as in the case of RE to identify a proper threshold value to be used as a cut off, shortlisting those high scoring residues mentioned above we generated a Null model, as explained in Section 4.3.5. The results in from the native and the null model are later compared as shown in figure 8a and figure 8 b.

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. In the case of CRE the cut off was found to be 1.Therefore those residues with normalized CRE scores greater than 3 were considered to be significantly contributing to the fold of the protein.

a)



b)

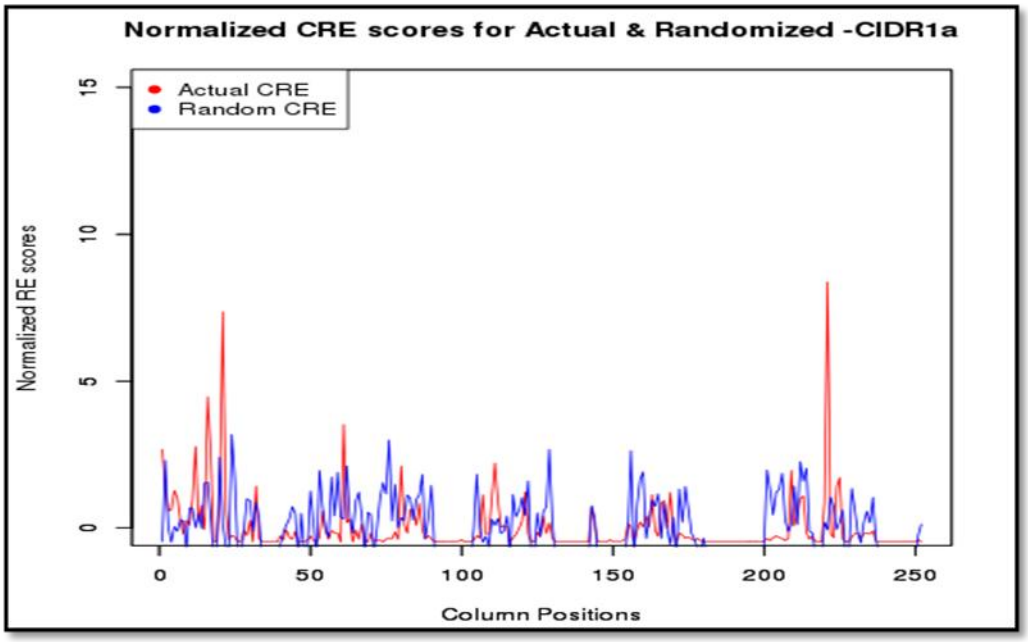


Figure 8 (a) (b) Comparison of native and null model results for CRE

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.

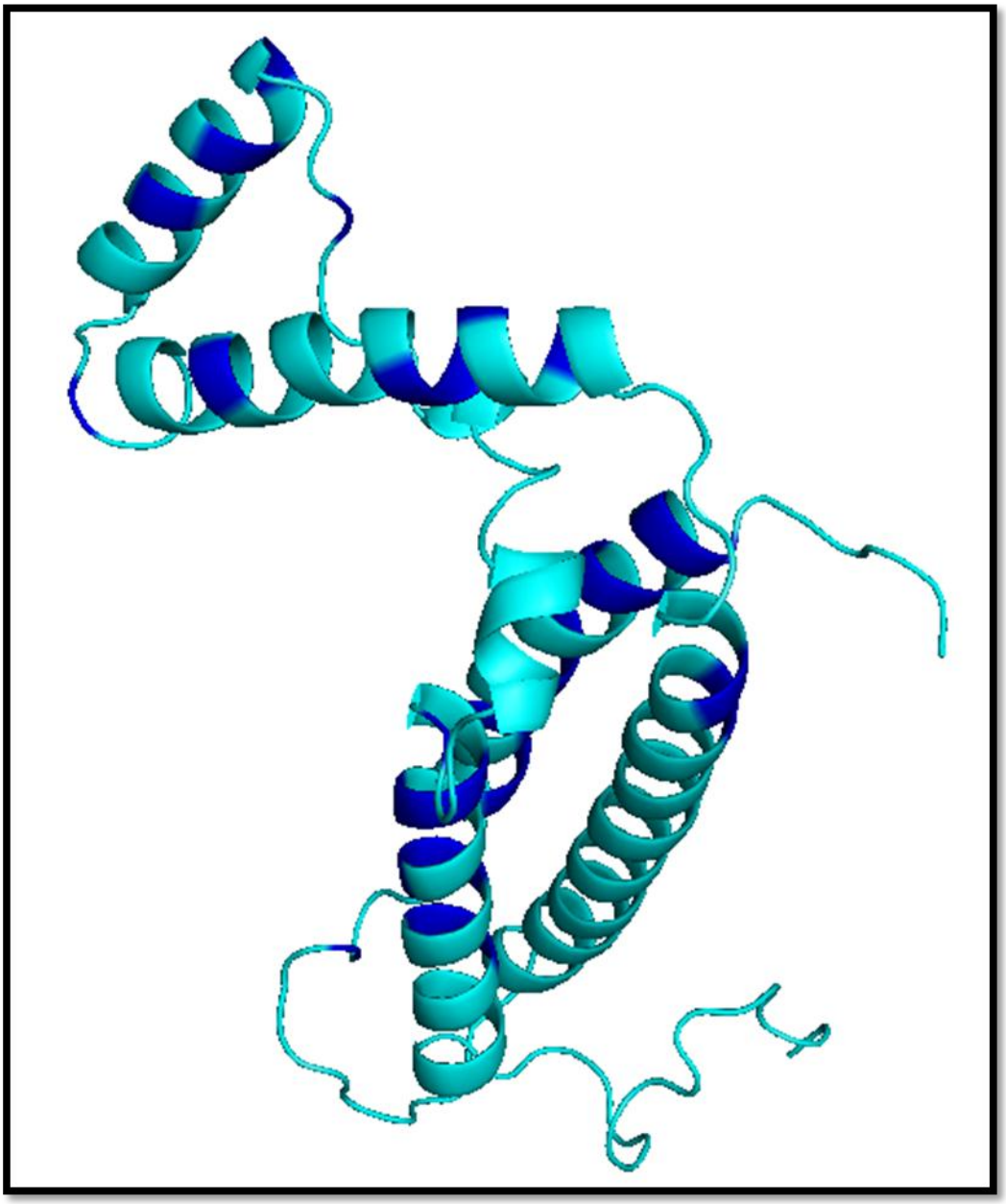


Figure 9 CIDR1α structure is shown in cartoon representation with functionally important residues in color blue.

5.6 Modeled Protein

CD 36 was modeled using modeler 9.v7 Package.

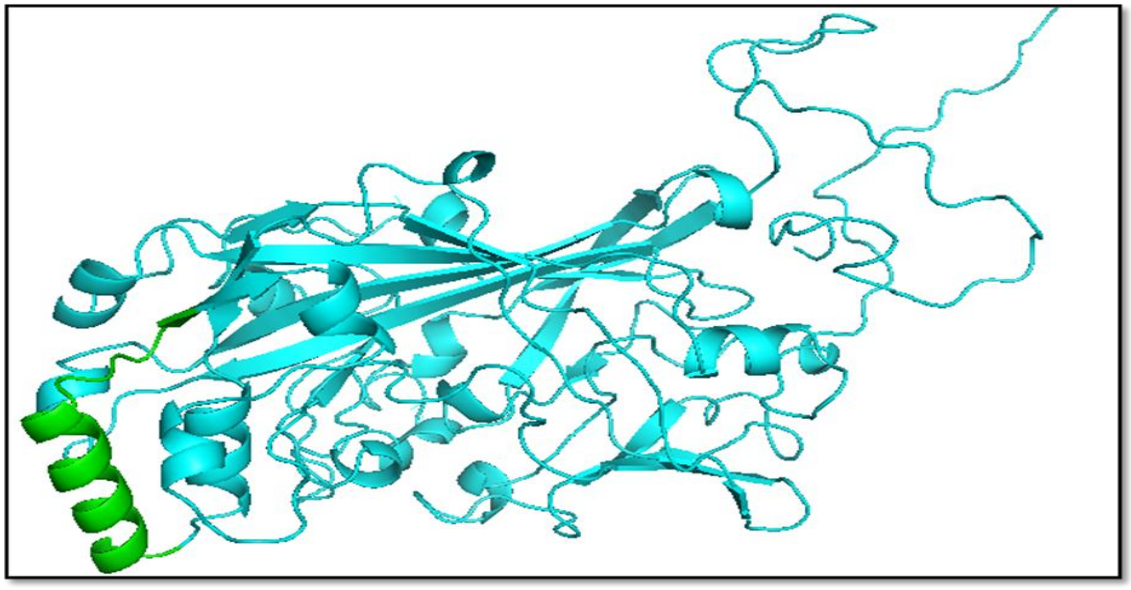


Figure 10 CD 36 structure is shown in cartoon representation with functionally important regions in color green.

5.6 Protein-protein Docking

Docked CD36 and CIDR1a complex, obtain using haddock.

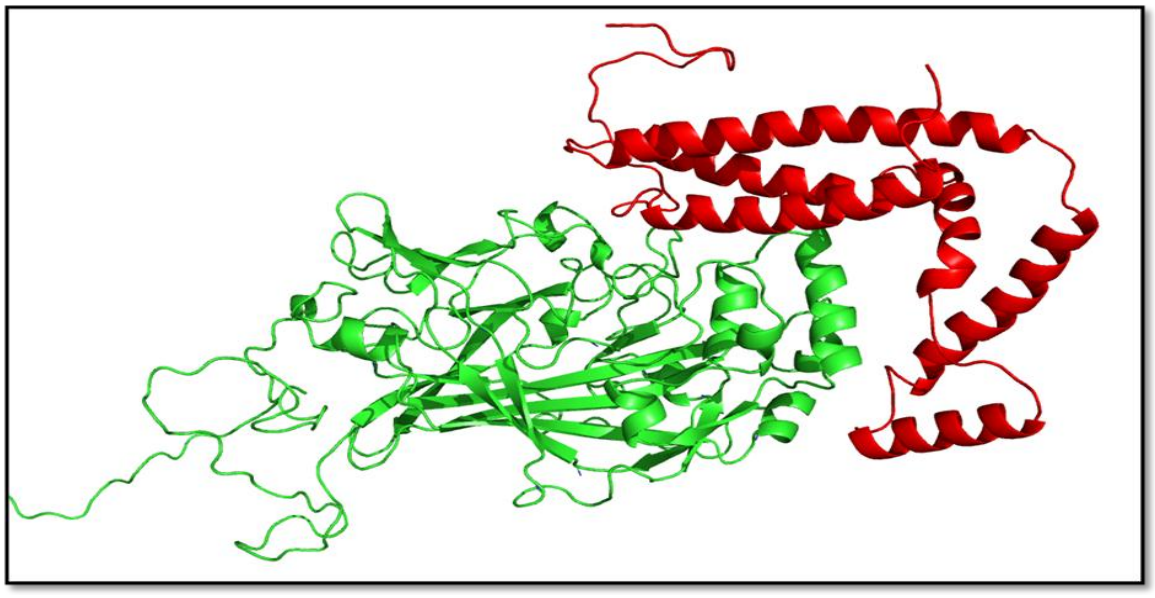


Figure 11 CIDR1α -CD 36 in color red, green respectively docked

6. Conclusion

We replaced traditional conservation scores with the Kulback-Leibler distance to predict the conservation patterns. It was found that this approach facilitates the selection of residues that were critical for the fold and function of the protein. TheseKulback-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 region which proposed as important for CIDR1α interaction to CD36.

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.

8. Appendix

Table 1 Summary of the codes and their usage and reference for their downloads

CODE	APPLICATION	REFRENCES
background_prob.pl	Calculates the background probability distribution from the alignment	http://www.jnu.ac.in/Faculty/andrew/
RE_family.pl	Calculates the Relative Entropy Scores for the alignment	http://www.jnu.ac.in/Faculty/andrew/
scaling.pl	Scales the scores based on the number of gaps	http://www.jnu.ac.in/Faculty/andrew/
mapping_protein.pl	Maps the results onto the sequence of interest	http://www.jnu.ac.in/Faculty/andrew/
sorting_seq.pl	Sorting the subfamily specific sequences from the entire alignment	http://www.jnu.ac.in/Faculty/andrew/
RE_subfamily.pl	Calculates the subfamily specific Relative entropy scores	http://www.jnu.ac.in/Faculty/andrew/
CREs.pl	Calculates the Two class CRE scores of different subfamilies	http://www.jnu.ac.in/Faculty/andrew/
rand1.pl	Generates the Null model for RE calculation	http://www.jnu.ac.in/Faculty/andrew/

RE_random.R	Identifies the Thresholds for RE calculation	http://www.jnu.ac.in/Faculty/andrew/

System Requirements:
All the codes except *rand2.pl* were run in an X86_64 Linux OS with Fedora, with i7 processors and 4GB ram with a normal run time for the codes. The all the plotting were carried out by codes written in R and are not presented here.

Table 2 The Fold Specific residues for CIDR1α

Fields: 1) Alignment Column Position, 2) RE score, 3) Amino Acids of the protein sequence mapped 4) Sequence position in the alignment

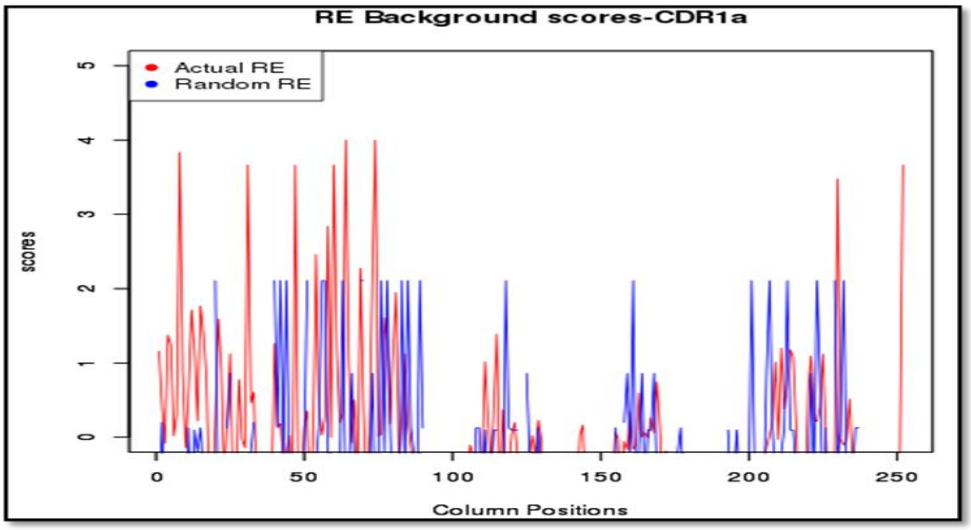
74	4.365618	W	57
8	4.21736	W	8
31	4.064422	C	26
47	4.064422	C	35
60	4.064422	C	45
252	4.064422	C	162
230	3.895122	C	153
58	3.319628	C	43
54	2.976108	C	39
69	2.809653	K	53
73	2.66462	E	56
81	2.517301	F	64
15	2.351927	D	15
12	2.302494	M	12
77	2.210297	I	60
21	2.192974	W	19
16	2.028999	S	16
115	2.009328	L	84

Table 3 The function Specific residues for CIDR1α

Fields: 1) Alignment Column Position, 2) CRE score, 3) Amino Acids of the protein sequence mapped 4) Sequence position in the alignment

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

Figure 12 The results for the comparison of background score with RE and CRE



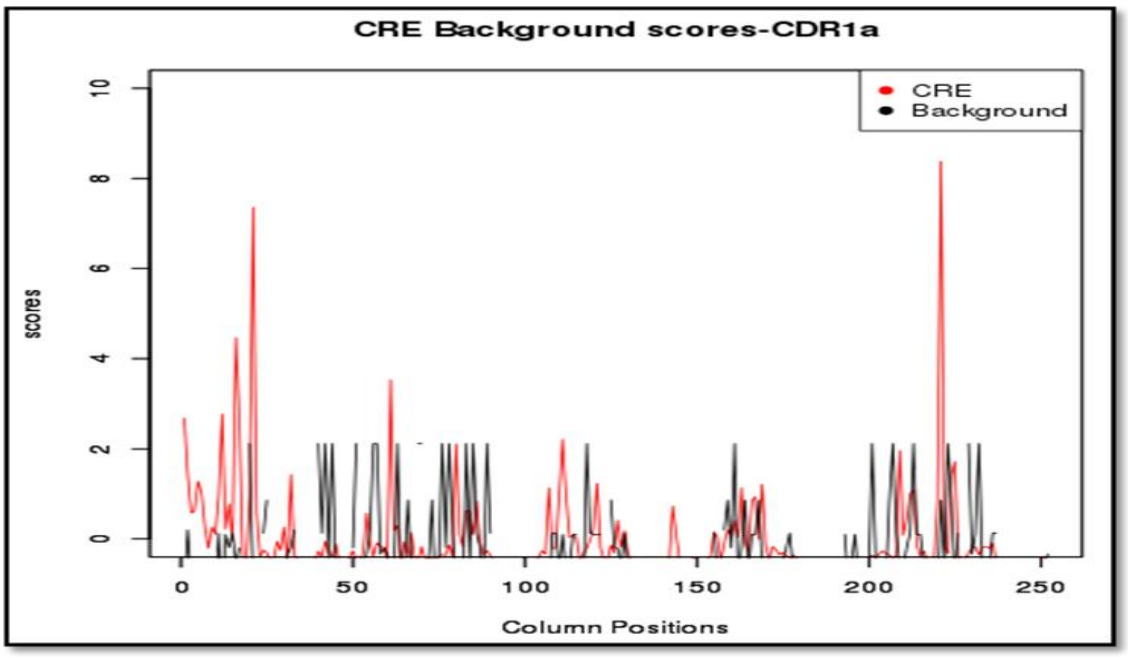


Figure 12 (a) Comparison of Background and RE scores

(b)Comparison of Background and CRE scores

Figure 13 The results for the comparison of dope score of template and query

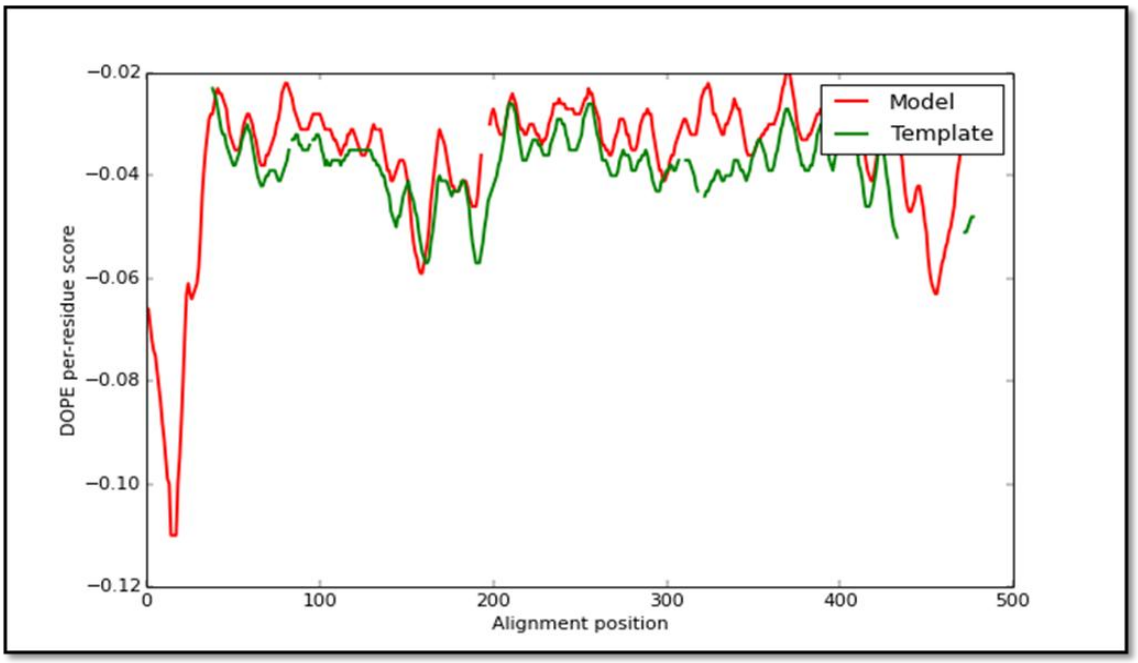


Figure 13 DOPE score profiles for the model and template

9. 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 micronemes 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, PasloskeBL,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, Pasloske BL, Singh HB, Bi X, Ma XC, 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, SundstromA,Schlichtherle M, Sahlen A, Carlson J, Datta S, Wahlgren M. (1998)Identification of Plasmodium falciparum erythrocyte membrane protein 1 (PfEMP1) as the rosetting ligand of the malaria parasite P. falciparum. J Exp Med;**187**:15 – 23.
6. Chen Q, Heddini A, Barragan A, Fernandez 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. ChristophAdami. (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. DanteNeculai, MichaelSchwake, ManiRavichandran, FriederikeZunke, RichardF.Collins, JudithPeters, MirelaNeculai, JonathanPlumb, PeterLoppnau, JuanCarlosPizarro, AlmaSeitova, WilliamS.Trimble, PaulSaftig, SergioGrinstein&SiranoDhe-Paganon(2013) Structure of LIMP-2 provides functional insights with implications for SR-BI and CD36 Nature **504**,172–176
10. Freitas-Junior LH, Bottius E, PirritLA, Deitsch KW, Scheidig C, et al. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature **407**: 1018–1022.
11. Hannenhali S, Russel R B. (2000). Analysis of functional Sub-types from Protein Sequence Alignment. J. Mol. Biology (**303**, 61-76).
12. Ho M., Schollaardt T., Niu X., Looareesuwan S., Patel K. D., Kubes P. (1998) Characterization of*Plasmodium falciparum*-infected erythrocyte and P-selectin interaction under flow conditions.*Blood* **91**:4803–4809.

13. Ho M., White N. J., Looareeswuan 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.
14. Kimura, *et.al.* (1983). The Neutral Theory of Molecular Evolution. Cambridge. University Press Cambridge.
15. 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.
16. 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. MolMicrobiol **50**: 1527–1538.
17. Kullback S, Leibler RA. (1951). On Information and Sufficiency. Annals of Mathematical Statistics (**22**, 79-86).
18. Kullback S, Leibler RA. (1951). On Information and Sufficiency. Annals of Mathematical Statistics (**22**, 79-86).
19. Laskowski RA, Rullmannn JA, MacArthur MW, KapteinR,Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.J Biomol NMR **8**:477–486.
20. 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 regions. Malar J **2**: 27.
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 regions. 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 ultrastructural analysis of parasitized erythrocyte sequestration. *Am. J. Pathol.* **119**:385–401.
24. Prashant K. Srivastava, 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. Rowe JA, Claessens A, Corrigan RA, Arman M (2009) Adhesion of Plasmodium falciparum-infected erythrocytes to human cells: molecular mechanisms and therapeutic implications. Expert Rev Mol Med **11**: e16.
28. Rowe JA, Moulds JM, Newbold CI, Miller LH.(1997) P. falciparum rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. Nature;**388**:292 – 5.
29. Sander S, Schneider R. (1991). Database of homologous-derived structures and the structural meaning of the sequence alignment. Proteins (**9**,56-68).
30. Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH. (1994) Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum. Science; **264**:1941– 4.
31. Sjoerd J de Vries, Marc van Dijk&Alexandre M J JBonvin (2000) The HADDOCK web server for data-driven biomolecular docking Nature Protocols **5**, - 883 - 897
32. Smith JD, Craig AG, Kriek N, Hudson-Taylor D, Kyes S, Fagen T, Pinches R, Baruch DI, Newbold CI, Miller LH. (2000)Identification of a Plasmodium falciparum intercellular adhesion molecule-1 binding domain: A parasite adhesion trait implicated in cerebral malaria. ProcNatlAcadSci USA;**97**:1766 – 71.
33. Smith JD, Kyes S, Craig AG, Fagan T, Hudson-Taylor D, Miller LH, Baruch DI, Newbold CI. (1998)Analysis of adhesive domains from the A4VAR Plasmodium falciparum erythrocyte membrane protein-1 identifies a CD36 binding domain. MolBiochemParasitol;**97**:133 – 48.
34. Smith JD, Subramanian G, Gamain B, Baruch DI, Miller LH (2000) Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. MolBiochemParasitol **110**: 293–310.
35. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell **82**: 89–100.
36. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell;**82**:89–100.
37. Taylor HM, Kyes SA, Newbold CI (2000) Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. MolBiochemParasitol **110**: 391–397.
38. Tramonotano A. (2005) .The ten most wanted solutions in protein bioinformatics. CRC Press.
39. Valdar WS.(2002). Scoring residue conservation.Proteins(**48**, 227-241).
40. Warrell DA, Molyneux ME, Beales PF. Severe and com-plicated malaria. Trans R Soc Trop Med Hyg 1990; **84**(suppl. 2):1 –65.

41. 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