BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(8)

Evolutionary analysis of circumsporozoite surface protein and merozoite surface protein-1 (CSP and MSP-1) sequences of malaria parasites

Vijay Tripathi[1]* & Dwijendra Gupta[1, 2]

[1]Center of Bioinformatics, University of Allahabad, Allahabad, India; [2]Department of Biochemistry, University of Allahabad, Allahabad, India; Vijay Tripathi - Email: vijaytripathi84@gmail.com; Phone: +91-9452599741; *Corresponding author

Received June 09, 2011; Accepted June 28, 2011; Published July 06, 2011

Abstract:
Malaria, one of the world's most common diseases, is caused by the intracellular protozoan parasite known as Plasmodium. In this study, we have determined the evolutionary relationship of two single-copy proteins, circumsporozoite protein (CSP) and merozoite surface protein-1 (MSP-1), among *Plasmodium* species using various bioinformatics tools and softwares. These two proteins are major blood stage antigens of *Plasmodium* species. This study demonstrates that the circumsporozoite protein of *Plasmodium falciparum* shows similarity with *Plasmodium cynomolgi* and *Plasmodium knowlesi.* The merozoite surface protein-1 of *Plasmodium coatneyi* forms a monophyletic group with *Plasmodium knowlesi*, demonstrating their close relationship and these two species also reveal similarity between the human malaria *Plasmodium vivax*. This *Plasmodium* phylogenetic arrangement is evidently crucial to identify shared derived characters as well as particular adaptation of *plasmodium* species from inside and between monophyletic groups.

Keywords: Circumsporozoite protein, Merozoite surface protein-1, Phylogenetic relationship, Plasmodium species.

Background:
Malaria is a common and life-threatening disease in many tropical and subtropical areas. Five species of the plasmodium parasite can infect humans: the most serious forms of the disease are caused by *P. falciparum*. Malaria caused by *P. vivax*, *P. ovale* and *P. malariae* causes milder disease in humans that is not generally fatal. A fifth species, *P. knowlesi*, is a zoonosis that causes malaria in macaques but can also infect humans. *P. falciparum* is the most dangerous of these infections as *P. falciparum* (or malignant) malaria has the highest rates of complications and mortality. The closest relative of *P. falciparum* is *P. reichenowi*, a parasite of chimpanzees. *P. falciparum* and *P. reichenowi* are not closely related to the other *Plasmodium* species that parasitize humans **[1, 2]**. It appears that *P. reichenowi* was the ancestor of *P. falciparum. P. cynomolgi*, which infects macaque monkeys, is very closely related to *P. vivax* and shares the characteristics of hypnozoites, relapse parasitemias, reticulocyte specificity and caveolae vesicular complexes in the infected erythrocyte membrane **[1, 3]**. The circumsporozoite (CS) protein is the most abundant protein on the surface of malaria sporozoites in the early 1980s **[4]**. It is a multifunctional molecule that plays a crucial role at various points of the malaria life cycle. This abundant protein forms a dense coat on the circumsporozoite, the form of the parasite that is released into the human bloodstream upon the bite of a mosquito **[5]**. The single-copy CSP gene encoding the highly immunogenic CSP surface antigen expressed at the parasite's sporozoite stage has proven to be a useful marker for defining the phylogenetic relationship of *Plasmodium* species **[6]**. The merozoite surface protein 1 (MSP-1) of *Plasmodium* plays an important role in erythrocyte invasion by the merozoite and is another potential malaria vaccine candidate antigen **[7]**, which has been characterized from *P. vivax*, *P. cynomolgi*, *P. knowlesi*, *P. falciparum* and *P. yoelii* **[8, 9]**. In this study, the evolutionary relationships of two single-copy proteins, circumsporozoite protein and merozoite surface protein-1 of malaria parasite was examined because these proteins are potential malaria vaccine for pre-erythocytic stage. This analysis confirms the CS protein of *P. falsiperum* and *P. reichenowi, P. yoelli* and *P. berghei, P. coatneyi* and *P. Knowlesi* demonstrates a close relationship and this may also show similar function and structure among them. The MSP-1 protein of *P. cynomolgi* and *P. knowlesi* are very much similar and demonstrates close relationship with *P. falsiparum.*

Methodology:
The protein sequences of CSP and MSP-1 of various malarial parasites were taken from GenBank Database of NCBI (National Center for Biotechnology Information). All the sequences were taken in their FASTA format. These sequences have been listed in **Table 1** and **Table 2 (see Supplementary material)**.

Sequence analysis:
Amino acid sequence alignments of these CS and MSP-1 proteins for phylogenetic analysis were generated using *Clustal W* server **[10]** and the data obtained was analyzed to determine the similarity between the protein sequences of different *Plasmodium* species. For multiple sequence alignments, gap open penalty was -7 and gap extension penalty was -1. BLOSUM weight matrix was used for substitution scoring **[11, 12]**. Hydrophilic gap penalties were used to increase the chances of a gap within a run (5 or more residues) of

hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common. BioEdit 7.0.2 **[13]** was also used to perform the protein sequence alignment to determine the number of conserved sites, parsimony info sites, variable sites and singleton sites.

Determination of the Entropy and hydrophobicity:
BioEdit 7.0.2 was used to determine the entropy plot and hydrophobicity values of CS and MSP-1 proteins. From information theory, Entropy can be defined as a measure of the unpredictable nature of a set of possible elements. The higher level of variation within the set, higher the entropy. Entropy is then calculated as given in **supplementary material**. The information content of a position l, then, is defined as a decrease in uncertainty or entropy at that position. As an alignment improves in quality, the entropy at each position (especially conserved regions) should decrease. This gives a measure of uncertainty at each position relative to other positions. Maximum total uncertainty will be defined by the maximum number of different characters found in a column. The mean hydrophobicity value was plotted for the middle residue of the window. Eisenberg method **[14]** and Kyte Doolittle method **[15]** were used to plot hydrophobic moment profile with a window size of 13 residues having six residues on either side of the current residue and rotation angle, θ =100 degrees. **(see Supplementary material)**

Phylogenetic Analysis:
Phylogenetic and molecular evolutionary analysis was conducted using MEGA version 4.0 **[16]**. The multiple alignments of sequences of CS and MSP-1 proteins were used to create phylogenetic trees. The evolutionary history was inferred using the Neighbour-Joining method **[17]**. All the characters were given equal weights. The bootstrap consensus tree inferred from 10000 replicates was taken to represent the evolutionary history of the taxa analyzed **[18]**. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). Phylogenetic analysis was conducted in MEGA4.

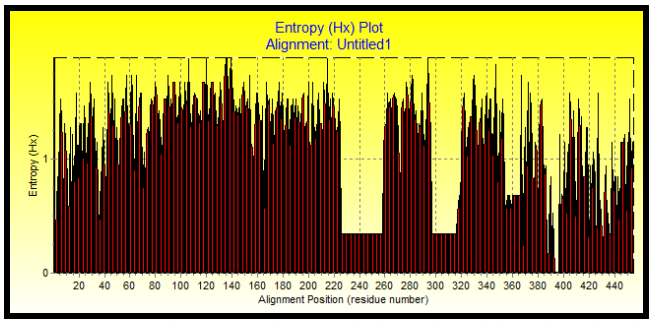

Figure 1: Entropy plot (CS protein). X-axis shows the positions of residues in MSA profile and the Y-axis shows entropy scores for individual positions in MSA profile. In this plot conserved regions in the profile are to be found from 220-260, 287-320 and 387-400 amino acid positions.

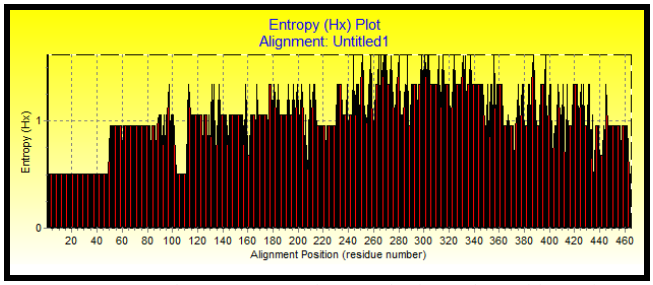

Figure 2: Entropy plot (MSP 1 protein). X-axis shows the positions of MSA and the Y-axis shows entropy scores for individual positions in MSA. In this plot conserved regions in the profile are to be found from 90-100, 130-190, 240-260 and 360-380 amino acid positions.

Results and Discussion:
Multiple Sequence Alignment: In the case of CS protein, the protein resulted into 446 positions out of which 297 are parsimony informative, 378 are variable sites, and conserved sites were 36 and 71 singleton sites. The most frequent amino acid of these sequences are Aparagine, Alanine, Glycine, Proline, Aspartic acid, Lysine, Glutamine and its composition observed 15.46, 13.34, 11.19, 10.07, 6.37, 5.79, 5.28 respectively. MSA of MSP-1 protein resulted into 465 positions out of which 25 are parsimony informative, 383 are variable sites, and conserved sites were 20 and 291 singleton sites. By statistical analysis of multiple aligned sequences it is observed that Lysin, Glutamic acid, Lucine, Serine, Asparagine, Alanine, Isolucine, Glutamine, Aspartic acid are the most frequently present amino acids with frequency percentage of 11.04, 9.76, 8.89, 7.21, 6.73, 6.26, 6.06, 5.86, 5.05 respectively.

Conserved Domain Search:
A conserved domain region search in CS protein from amino acid 220-260, 287-320, 387-400 region and we have found four conserved regions from amino acid positions 90-100, 130-190, 240-260, 360-380 in MSP-1 protein sequences. This conservation has already been upheld by minimal entropy shown by respective positions of the previous half of the MSA results.

Entropy Plot of CS proteins and MSP-1 Proteins:
An entropy plot, measure of the lack of the information content and the amount of variability, was generated for all the aligned positions. The plot shows that entropy rarely touches a scale of 1, showing minimal entropy at several positions subjected to previous half of the protein sequences, which is a sign of higher similarities in the region. The lower entropy value implies that the randomness of these amino acid residues in that particular column of MSA profile is less and hence those are the most conserved region. These results have already been upheld by finding conserved regions. According to entropy plot the conserved region in CS proteins **(Figure 1)** to be found from 220-260, 287-320 and 387-400 amino acid positions and in MSP-1 proteins **(Figure 2)** the conserved regions belongs from these regions: 90-100, 130-190, 240-260, 360-380 because entropy values of these amino acids are less than 1.

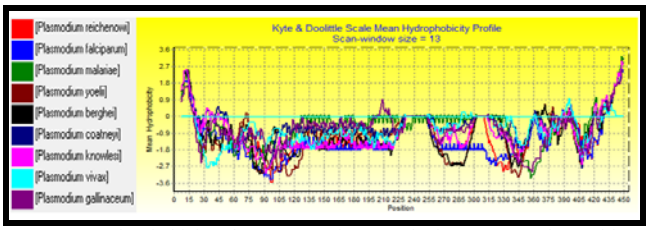

Figure 3: Hydrophobicity plot (CS protein). X-axis demonstrates the positions of MSA profile and the Y-axis shows hydrophobicity scores for individual positions in MSA profile. In this plot conserved regions in the profile are to be found from 135-200, 220-260, 360-400 and 420-435 amino acid positions.

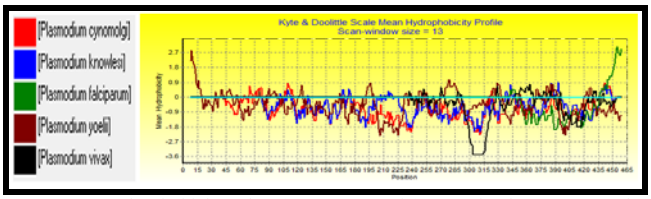

Figure 4: Hydrophobicity plot (MSP 1 protein). X-axis demonstrates the positions of MSA profile and the Y-axis shows hydrophobicity scores for individual positions in MSA profile. In this plot conserved regions in the profile are to be found from 165-185, 240-290, 330-390 and 425-450 amino acid positions.

Hydrophobicity profile of CS proteins and MSP-1 proteins:
A hydrophobicity profile plot shows that mean hydrophobicity of the protein for most of the positions in all the species is below zero; occasionally it turns to be positive. From the profile it is clear that the regions related to conserved positions also have a characteristic of possessing residues in a balanced way and the profile is always around zero value. However, the amino acid regions from 135-200, 220-260, 360-400, 420-435 in CS proteins and 165-185, 240-290, 330-390, 425-450 amino acid regions in MSP-1 proteins have lower hydrophobicity, subjected to more conserved part of the protein **(Figure 3, 4)**.

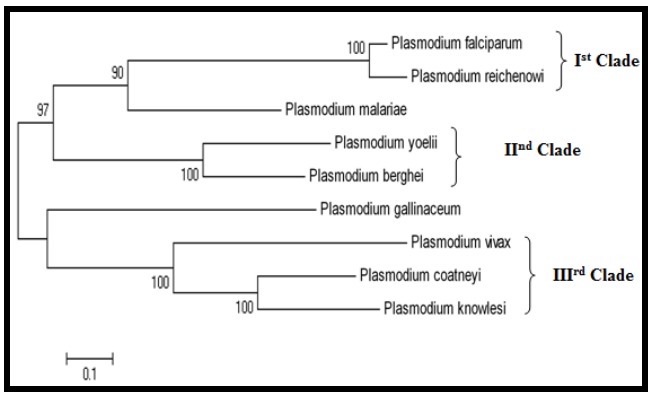

Figure 5: Phylogenetic tree of CS Protein.

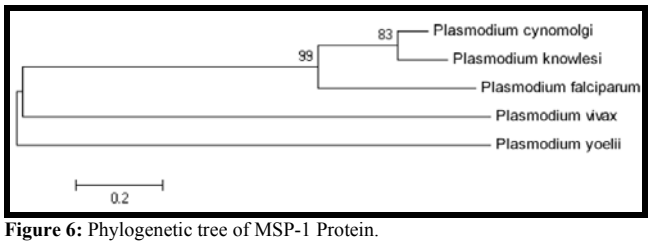

Figure 6: Phylogenetic tree of MSP-1 Protein.

Phylogenetic Analysis of CS proteins and MSP-1 proteins:

The phylogenetic trees were constructed by using Neighbour–joining method. In the case of CS proteins the tree is divided into three different clades **(Figure 5)**. In the first clade, *P. falsiperum* shows very much similarity with *P. reichenowi* and the bootstrap percentage of this similarity is 100. *P. yoelli* shows similarity with *P. berghei* in a second clade of this tree with 100 boot strap percentage. But in the third clade *P. coatneyi* shows very close relationship between *P. knowlesi* with 100 boot strap percentage and these two malaria parasites show similarity with *P. vivax* and the boot strap percentage of this similarity is 100. Esmeralda Vargas-Serrato *et al.* **[19]** study also shows that these two species (*P. coatneyi and P. Knowlesi*) are very much similar. The phylogenetic tree of MSP-1 proteins **(Figure 6)** demonstrates that *P. cynomolgi* is highly similar with *P. knowlesi* and the boot strap percentage of this similarity is 83. These two malarial parasites show similarity with *P. falsiparum* with boot strap percentage of 99. *P. yoelli* appears with a totally diverged branch from the main tree. The CS and *MSP-1* proteins derived phylogenetic trees reveal different topologies, conclusively showing that *P. coatneyi* and *P. knowlesi* are evolutionarily closely related and these two malaria parasites show similarity with *P. vivax* but in case of MSP 1 the *P. cynomolgi* is highly similar with *P. knowlesi* and the boot strap percentage of this similarity is 83. *P. falsiparum* shows very much similarity with these two malarial parasites.

Conclusion:

This study presents a comparative proteomics study and evolutionary analysis of the CS and MSP-1 proteins based on molecular phylogeny across different species of malaria parasite. According to this study we have concluded that CS protein of *P. coatneyi and P. knowlesis* demonstrate a close relationship with *P. vivax* and thus it may also show similarity in their structure and function. The MSP-1 protein of *P. cynomolgi* and *P. knowlesi* is very much similar and demonstrate close relationship with *P. falsiparum.*

References:

[1] Coatneyi RS *et al.* Plasmodium coatneyi. In: The Primate Malarias. US Government Printing Office, Washington, DC. 1971, 289–299

[2] Escalante AA *et al. Proc Natl Acad Sci U S A.* 1998 **95**: 8124 [PMID: 9653151]

[3] Barmwell JW & Galinski MR. *Ann Trop Med Parasitol.* 1995 **89**: 113 [PMID: 7605120]

[4] Nussenzweig V & Nussenzweig RS. *Cell* 1985 **42**: 401 [PMID: 2411417]

[5] Good MF. *Trends Parasitol.* 2005 **21**: 29 [PMID: 15639738]

[6] McCutchan TF *et al. Proc Natl Acad Sci U S A.* 1996 **93**: 11889 [PMID: 8876233]

[7] Ferreira MU *et al. Gene* 2003 **304**: 65 [PMID: 12568716]

[8] Barnwell JW *et al. Exp Parasitol.* 1999 **91**: 238 [PMID: 10072326]

[9] Vargas-Serrato E *et al. Mol Biochem Parasitol.* 2002 **120**: 41 [PMID: 11849704]

[10] Thompson JD *et al. Nucleic Acids Res.* 1994 **22**: 4673 [PMID: 7984417]

[11] Altschul SF & Gish W. *Methods Enzymol.* 1996 **266**: 460 [PMID: 8743700]

[12] Henikoff S & Henikoff JG. *Proc Natl Acad Sci U S A.* 1992 **89**: 10915 [PMID: 1438297]

[13] Hall TA. *Nucleic Acids Symposium Series* 1999 **41**: 95

[14] Eisenberg D *et al. J Mol Biol.* 1984 **179**: 125 [PMID: 6502707]

[15] Kyte J & Doolittle RF. *Journal of Molecular Biology* 1982 **157**: 105 [PMID: 7108955]

[16] Tamura K *et al. Mol Biol Evol.* 2007 **24**: 1596 [PMID: 17488738]

[17] Saitou N & Nei M. *Mol Biol Evol.* 1987 **4**: 406 [PMID: 3447015]

[18] Felsenstein J. *Evolution* 1985 **39**: 783

[19] Vargas-Serrato E *et al. Infection Genetics and Evolution* 2003 **3**: 67

Edited by P Kangueane

Citation: Tripathi *et al.* Bioinformation 6(8): 320-323 (2011)

BIOINFORMATION open access

Supplementary material:

Entropy:
H (I) = -Σf(b,l)ln(f(b,l)),
where H(I) = the uncertainty, also called entropy at position s l, b represents a residue (out of the allowed choices for the sequence in question), and f(b, l) is the frequency at which residue b is found at position l.

μH= {[Hn sin(δn)]^2+[Hn cos(δn)]^2}
Where μH is the hydrophobic moment, Hn is the hydrophobicity score of the residue H at position n, δ=100 degrees, n is position within the segment, and each hydrophobic moment is summed over a segment of the same defined window length.

Table 1: Protein sequences of CS protein used for comparative genomics and evolutionary studies

Name of Species	Accession No.	Length
Plasmodium coatneyi	AAN16518.1	341
Plasmodium falciparum	AAN87612.1	408
Plasmodium vivax	ABJ53009.1	281
Plasmodium knowlesi	ABG29610.2	364
Plasmodium malariae	AAA29557.1	429
Plasmodium reichenowi	AAA29561.1	388
Plasmodium yoelii	AAA29558.1	367
Plasmodium berghei	AAA29541.1	332
Plasmodium gallinaceum	AAC47344.1	388

Table 2: Protein sequences of MSP-1 protein used for comparative genomics and evolutionary studies

Name of Species	Accession No.	Length
Plasmodium cynomolgi	AAW65096.1	380
Plasmodium yoelii	AAA66185.1	462
Plasmodium falciparum	ABF18819.1	116
Plasmodium vivax	ABJ53051.1	200
Plasmodium knowlesi	CAA62966.1	327