# CHAPTER 1

## Introduction – A review of literature

Malaria is one of the most common infectious diseases caused by protozoan parasites of the genus *Plasmodium*. The disease has infected humans for over 50,000 years, and was possibly a human pathogen for the entire history of the species (Joy et al. 2003). The term malaria originates from medieval Italian: *mala aria* which means "bad air"; and the disease was formerly called ague or marsh fever due to its association with swamps and marshland.

In humans, the infection is caused by four major species namely *Plasmodium falciparum*, *Plasmodium vivax, Plasmodium malariae* and *Plasmodium ovale* (Mendis et al. 2001; Mueller, Zimmerman, and Reeder 2007). Among these, *Plasmodium falciparum* causes the deadliest form of malaria and is responsible for the death of about 2.7 million people annually (Breman 2001; Gardner et al. 2002). The malaria parasites not only have a complex life cycle (**Figure 1.1**) that alternates between vertebrate (human) and invertebrate (female anopheline mosquitoes) host environments, but they are also capable of developing within highly specialized red blood cells, a challenge met by only a few intracellular pathogens (Przyborski and Lanzer 2004).

The extent of human suffering caused by malaria and its devastating costs has long been recognized by international bodies, and many initiatives have been taken over the years to tackle this insidious microbe (Doolittle 2002). Despite the massive efforts made to eradicate malaria, the ambiguity of the genome has restricted our understanding of the disease, to defeat it permanently (Wirth 2002).

There is a continuing upsurge of malaria over the past 35 years due to the emergence and spread of drug resistant parasites (Hastings, Bray, and Ward 2002; Sidhu, Verdier-Pinard, and Fidock 2002; Wootton et al. 2002) and a variety of other mutually reinforcing factors like the evolution of pesticide-resistant mosquitoes (Bradley 1998; Hemingway and Ranson 2000; Hemingway, Field, and Vontas 2002) and global warming that poses new challenges for scientific research and malaria control strategies respectively (Hartl 2004).

These stark facts emphasize the need to understand the biology of the complex protozoan parasite from a new perspective so as to identify new and effective drug and vaccine targets to combat the disease. An international effort was therefore launched in the year 1996 to sequence the *Plasmodium falciparum* genome with the expectation that the genome sequence would open new avenues for malaria research.

## 1.1 The *Plasmodium falciparum* genome

The complete genome of the *Plasmodium falciparum* clone 3D7 was successfully sequenced in the year 2002 (Gardner et al. 2002). The 23-megabase nuclear genome consists of 14 chromosomes, encodes about 5,300 genes and is the most (A+T)-rich genome sequenced to date (Gardner et al. 2002). The overall (A+T) composition of this genome is 80.6% and rises to ~90% in introns and intergenic regions (Gardner et al. 2002). The parasite appears to be relatively distant to other eukaryotes with most of its proteins lacking any notable sequence similarity to other organisms.

Due to the nucleotide complexity the *Plasmodium falciparum* genome presented substantial technical hurdles; severe problems were encountered during the assembly of primary sequence reads (Hyman et al. 2002). The annotation process also proved cumbersome with functional assignments to only 40% of the protein encoding genes. The remaining 60% of the predicted proteins that could not be assigned functions either encode functions that are unique to *Plasmodium* species, are conserved proteins of unknown function, or have diverged to an extent that they fail to match known proteins from other organisms so that they could be recognized by simple BLASTP searches using standard methods (McConkey et al. 2004). About 52% of the predicted gene products of *Plasmodium falciparum* were detected in cell lysates prepared from several stages of the parasite life cycle by high resolution liquid chromatography and tandem mass spectrometry (Florens et al. 2002; Lasonder et al. 2002). These included many predicted proteins with no similarity to proteins from other organisms (Gardner et al. 2002).

The malaria parasite also harbors a plastid like genome, a characteristic feature of the phylum Apicomplexa, homologous to the chloroplasts of plants and algae (McFadden et al. 1996; Wilson et al. 1996; Kohler et al. 1997). Known as the 'apicoplast' this organelle arose through a process of secondary endosymbiosis (Roos et al. 1999; Stoebe and Kowallik 1999; Fast et al. 2001; Wilson 2002) and is important for the survival of the malaria parasite (Fichera and Roos 1997; He et al. 2001). The 35-kb apicoplast genome encodes only 30 proteins (Wilson et al. 1996) and its proteome is supplemented by proteins encoded in the nuclear genome that

are post-translationally targeted into the organelle (Gardner et al. 2002). About 60% of the putative apicoplast targeted proteins are of unknown function (Gardner et al. 2002).

The evolution of Apicomplexa, especially *Plasmodium* is a subject of controversy (Escalante and Ayala 1995). There seems to be no clear proof of the evolutionary origin of the parasite. Various schools of thought have evolved in this regard; while some believe they might have been derived from the gut parasites of vertebrates, others think that they may have evolved from originally monogenetic parasites of dipterans (Escalante and Ayala 1995). Considering its relationship to organisms ranging from plants to animals and those as primitive as red algae, its evolutionary history seems ambiguous. One of the evident fact is that, in terms of the overall genome content, among the completely sequenced eukaryotes (excluding *E. cuniculi*), *Plasmodium falciparum* is more similar to *Arabidopsis thaliana* than to other taxa (Gardner et al. 2002). This is possibly due to the presence of the genes derived from plastids or from the nuclear genome of the secondary endosymbiont (Gardner et al. 2002). Around 237 proteins match strongly to proteins in all completed eukaryotic genomes but they do not match to any of the prokaryotes even at low stringency (Gardner et al. 2002). These proteins have roles in cytoskeleton construction and maintenance, cell cycle regulation, chromatin packaging and modification, intracellular signaling, transcription, translation, replication and many proteins of unknown function (Gardner et al. 2002). About 3.9 % proteins are known to be involved in evasion of host immune

system, and ~14% are identified as metabolic enzymes (Gardner et al. 2002).

The existence of proteins in the *Plasmodium falciparum* genome that fail to show sequence matches to known proteins raise many interesting questions that seek an answer; Is this due to the apparent failure of sequence search algorithms with complex genomes, or, is this a pure outcome of the uniqueness of the organism? Summarizing our current knowledge of the genome, the enigma of the parasite biology presents a definitive scope for further research.

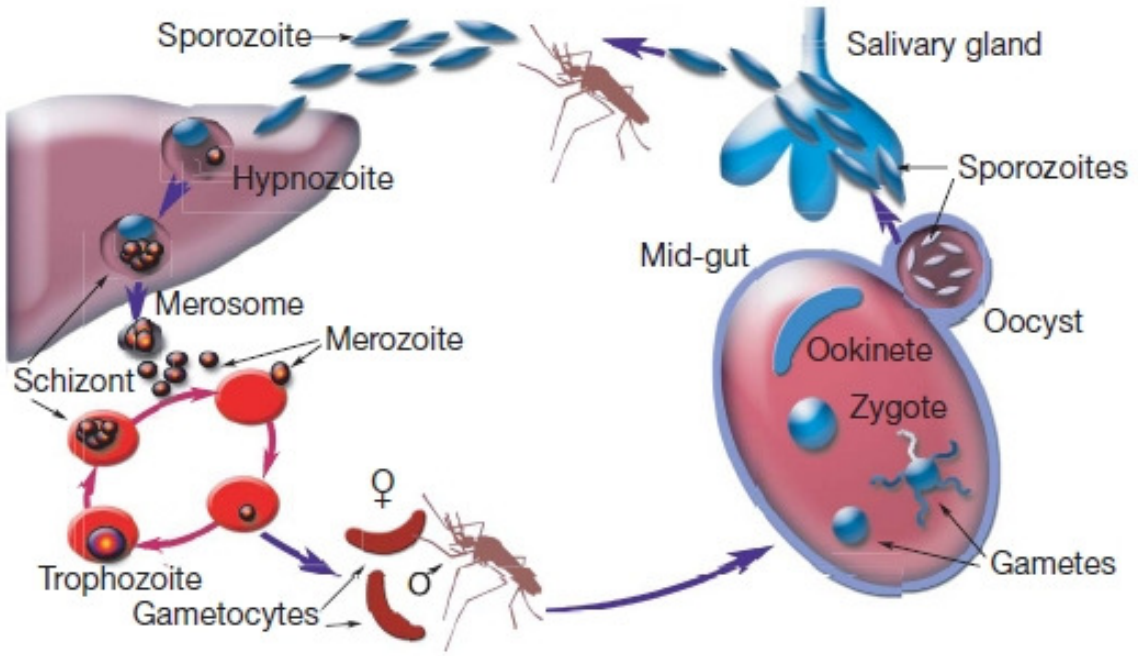


**Figure 1.1: Life cycle of the malaria parasite** (Adapted from Winzeler 2008)

The malaria disease is transmitted by the bite of a female mosquito in which hundreds of sporozoites enter the host's circulatory system. In due course, the parasites penetrate the liver cell forming parasitophorous vacuoles. At this stage, they either remain dormant (hypnozoite form of *P. vivax* or *P. ovale*), or initiate asexual development that results in the production of merozoites. Next, the parasites induce the detachment of the infected hepatocyte, which then migrates to the liver sinusoid. Here, budding of parasite-filled vesicles called merosomes76 occurs. The new merozoites invade erythrocytes and replicate. In response to some signal, some parasites differentiate into male and female gametocytes, which are the forms taken up by the mosquito. Once they enter the mosquito via a blood meal they rapidly undergo a change into activated male and female gametes. The ookinete, which is the diploid form of the parasite, migrates to the mid-gut wall where an oocyst is formed. Within this oocyst, a meiotic reduction in chromosome number occurs, and sporozoites are formed. Eventually the oocyst ruptures and the sporozoites migrate to the salivary gland of the mosquito, where they await transfer to a new host (Winzeler 2008).

## 1.2 Post-genomic era

The process of genome sequencing is followed by a longer, more gradual process of genome annotation. The annotated genomes of organisms define a 'blueprint' of their possible gene products (Lasonder et al. 2002). Post-genomic analyses attempt to confirm and modify the annotation and impose a sense of the spatial, temporal and developmental usage of genetic information by the organism (Lasonder et al. 2002). This era marks the exploitation of the sheer volumes of *Plasmodium falciparum* genomic data which is incomplete in terms of annotation. This is particularly important in a system like *Plasmodium* that has its experimental restrictions in terms of genetics and biochemistry, relative to other model organisms. The most important aspect of the post-genomic era is the careful analysis of the primary sequence data for deriving meaningful biological information. While efforts are being made in making advancements in the areas of transcriptomic and proteomics technologies in malaria research (Ginsburg 2006), the basic bio-informatic methods still need to be improved, considering the complexity of the parasite genome and the restriction it imposes to sequence analysis.

## 1.3 Amino acid substitution in proteins

As a process of evolution, a protein's amino acid sequence may undergo insertions, deletions and amino acid replacements. The amino acid composition of proteins varies substantially between various taxonomical classes, leading to evolution (Jordan et al. 2005). Studies of protein evolution also suggest that structure and

function can be well conserved even as protein sequence diverges extensively (Brooks and Fresco 2002). Conversely evolution of amino acid composition may have its impact on the protein structure in newly arising proteins of the proteome (Brooks and Fresco 2002).

It is a well known fact that individual genes as well as complete genomes vary markedly in their nucleotide compositions (Bernardi 1986; Muto and Osawa 1987). While some genomes are biased towards the Guanine and cytosine (G+C) content of their nucleotides (e.g. Actinobacteria), others are disproportionately rich in adenine and thymine (A+T) nucleotides (e.g. *Plasmodium falciparum*). Variation in nucleotide composition is usually most pronounced at the synonymous codon positions of genes. However, due to the redundancy in the genetic code, these variations in DNA content may have little effect on the amino acid content of the encoded proteins (Loomis and Smith 1990; Lockhart et al. 1992). If, however, compositional bias at the nucleotide level affects the non-synonymous sites in protein-coding genes, proteins are anticipated to change their amino acid composition over evolutionary time, in a direction predicted by the underlying nucleotide bias (Singer and Hickey 2000).

Several previous studies suggest that protein evolution is affected by the nucleotide composition of the encoding genes. The first evidence was provided many years ago by the finding that there was a correlation between the nucleotide content and the amino acid content of bacterial cells (Sueoka 1961). Subsequently, a number of surveys of molecular sequences (Bernardi 1986; D'Onofrio et al. 1991;

Collins and Jukes 1993; Berkhout and van Hemert 1994; Porter 1995) identified correlations between the nucleotide composition of DNA and the amino acid content of the encoded proteins. A positive correlation was also observed between the degree of amino acid bias and the magnitude of protein sequence divergence in genomes, including that of *Plasmodium falciparum* (Singer and Hickey 2000).

Considering the fact that the nucleotide compositional constraints of genomes can dramatically affect the amino acid composition of its encoded proteins (Singer and Hickey 2000), the malaria parasite would possibly allow for substitutions within its protein repertoire suggesting a likely sequence divergence of the encoded proteins such that sequence analysis limits the identification of similar proteins in this organism. Identification of such proteins is thus a computational challenge in a genome like *Plasmodium falciparum*.

## 1.4 Sequence analysis

Sequence analysis is useful for discovering functional, structural, and evolutionary information in biological sequences that are made available by large scale sequencing projects. The preliminary analysis in the process of gene annotation is purely alignment based, weighed both at nucleotide and amino acid levels. Two popular alignment tools are widely used for sequence alignments to infer homology, namely BLAST (Altschul et al. 1990) and FASTA (Pearson and Lipman 1988) that use a dynamic programming approach for achieving optimal alignments. However, alignments involving protein are more preferable as the nucleotide alignment methods often lead to an over-estimation of divergence, which is purely a

consequence of synonymous mutations within the DNA. On the contrary, certain substitutions are known to commonly occur in related proteins from different species as the substituted amino acids are compatible with the protein structure and function. The protein alignment tools thus make use of substitution tables that represent the likelihood of change from one amino acid to another in homologous protein sequences during evolution. These likelihood ratios are computed from standard protein databases having standard amino acid compositions.

A direct assignment of function for a protein may be made if homology-based gene finding produces a good match to a known protein. The malaria parasite however has evolutionarily diverged to a good extent from other organisms and has introduced rare substitutions in its proteome. This is obvious from instances of supposedly missing proteins of the malaria parasite that could not be identified at sequence level but were later detected from crude extracts of the parasite (McConkey et al. 2004). These very facts emphasize the need to improve upon the existing sequence analysis tools (sequence alignments, motif detection, phylogeny), such that the functional predictions are improved. The exploration of these areas may give new insights into the *Plasmodium* biology that remains to be investigated.

## 1.5 Substitution matrices

Scoring matrices are a fundamental component of most of the currently available protein comparison and alignment tools that are used to score alignments. Such are the amino acid substitution matrices or symbol comparison tables that depict substitution rates of each amino acid change in homologous protein sequences over

time. These are of potential importance as the sensitivity of most protein sequence alignment methods depends strongly on the quality of the comparison matrices used (Vogt, Etzold, and Argos 1995).

One of the earliest substitution matrices were developed by Margaret Dayhoff in the late 1960's and 1970's known as the Dayhoff amino acid substitution matrices (Dayhoff 1978) or the percent accepted mutation (PAM), that are based on strong evolutionary principles. An accepted mutation is the one that occurred and was positively selected by the environment, in the sense that within some given species, the mutation has not only arisen but has, over time, spread to essentially the entire species. Using sequences that were available at that time, Margaret Dayhoff constructed multiple alignments of related proteins and compared the frequencies of amino acid substitutions. The similarities between amino acids were represented as a $\log_2$ odd ratio, known as lod score.

In deriving the PAM matrices, each change in the current amino acid at a particular site is assumed to be independent of previous mutational events at that site (Dayhoff 1978). According to this, the amino acid substitutions in a protein sequence follow a Markov model such that the substitutions observed over a relatively short period of evolutionary time can be extrapolated to longer periods of evolutionary time. For simplicity, the direction of evolution is usually ignored in this case such that the scores are symmetrical.

An improvement over the PAM model was the amino acid scoring matrix, BLOSUM (BLOcks Substitution Matrix), developed by Henikoff (Henikoff and

Henikoff 1992). These matrices were constructed by extracting un-gapped amino acid patterns or blocks from a set of multiple alignments of protein families. The amino acid changes observed in each column of the block are tabulated, regardless of the overall degree of similarity between the protein sequences, unlike the PAM matrices that are based on scoring all amino acid positions. These matrices are more empirical and were derived from a larger dataset of diverse proteins compared to the Dayhoff PAM matrices. The extrapolation employed by the PAM model has been a drawback since the forces governing the sequence evolution over short evolutionary times are supposedly different from those shaping sequences over longer intervals. The distant substitution frequencies could thus be estimated appropriately from alignments of distantly related proteins which are an advantage with the BLOSUM series of matrices.

In addition to the well known PAM and BLOSUM substitution matrices, a number of other PAM related matrices were compiled like that of Gonnet and coworkers (Gonnet, Cohen, and Benner 1992), Benner and coworkers (Benner, Cohen, and Gonnet 1994), and, Jones and coworkers (Jones, Taylor, and Thornton 1992) for protein sequence alignments. Scoring matrices were also built based on the amino acid properties like that of Grantham (Grantham 1974), Miyata and coworkers (Miyata, Miyazawa, and Yasunaga 1979) and Rao (Mohana Rao 1987). Apart from this a number of protein structure based matrices like that of Risler and coworkers (Risler et al. 1988) and, Johnson and Overington (Johnson and Overington 1993) were also compiled. While Johnson matrices were derived from

tertiary structural alignments of 65 protein families in a database of 235 structures, the Risler matrices were derived from the alignments of 32 three dimensional structures from 11 protein families by rigid body superposition of the backbone topology (Vogt, Etzold, and Argos 1995).

The sub-optimal performances of the general scoring matrices for specialized searches lead researchers to formulate new methods for computing matrices for specific searches. Examples of such matrices are those developed for detecting trans-membrane proteins e.g. JTT trans-membrane matrix by Jones and coworkers (Jones, Taylor, and Thornton 1994), PHAT matrix (Ng, Henikoff, and Henikoff 2000) and the SLIM matrix series (Muller, Rahmann, and Rehmsmeier 2001). The rational aspect behind the making of these matrices was the non-standard amino acid composition of the trans-membrane proteins that hampered their detection with the general-purpose scoring matrices. Very recently there have also been efforts by some research groups to formulate methods to overcome the compositional bias of biased genomes and correct for the log-odd ratios of substitution matrices (Yu and Altschul 2005; Brick and Pizzi 2008).

It is routinely assumed that extant proteins are in a detailed equilibrium and their evolution is a stationary and reversible process: reciprocal fluxes of amino acid substitutions are equal, amino acid frequencies are constant, and nothing would change if time were to flow backwards (Muller and Vingron 2000; Goldman and Whelan 2002; Veerassamy, Smith, and Tillier 2003). Accordingly, symmetric substitution matrices are used for protein sequence alignments (Henikoff and

Henikoff 2000). However, it has been shown that proteins are not in a detailed equilibrium and their evolution is irreversible (Jordan et al. 2005). The utility of asymmetric matrices was long realized and there were some schools of thought regarding the need to develop asymmetric matrices. For example, Feng and Doolittle had suggested the use of non-symmetric matrices for phylogeny-based alignment of multiple protein sequences (Feng and Doolittle 1996) while Muller and co-workers had recommended the use of asymmetric matrices for database searches, where one attempts to discriminate subject homologs from unrelated subjects relative to the query (Muller, Rahmann, and Rehmsmeier 2001). However, there is an ample need for the existing sequence analysis tools to be modified such that they support the asymmetric score matrices as an input for their usage and application.

## 1.6 Objective and overview of the present work

A review of literature as provided in this chapter indicates that sequences of *Plasmodium falciparum* proteins are difficult to analyze using conventional tools. This is due to the striking amino acid bias observed in the *Plasmodium* proteins, which reflects the nucleotide compositional bias of the malarial genome. As a result, the homology of a *Plasmodium falciparum* sequence with a related sequence from another organism would be 'blurred' by the magnitude of the amino acid bias and many of these may no longer be detected by conventional automatic similarity search procedures (Bastien et al. 2004). However, the use of modified computational tools would provide effective means to pierce this veil and glean

interesting insights.

Many proteins encoded by the genome of *Plasmodium falciparum* are of unknown function. A majority of them bear no acceptable sequence homology with known proteins in other organisms. Under such circumstances, the genomic analysis provides a firm basis for constructing cell biological hypothesis rather than making unfounded conjectures (Aravind et al. 2003). This is particularly important in a system like *Plasmodium falciparum* that has its experimental restrictions. It is thus essential that an attempt be made to annotate the unknown gene products and understand interactions between them. The characterization of these proteins in turn would provide ways to develop combative strategies against malaria based on the parasite biology. The true exploration of these areas will lead to major advances in understanding the biology of *Plasmodium falciparum* and their translation to new tools for malaria control.

The primary objective of my study is to understand the amino acid substitutions in the malaria parasite, *Plasmodium falciparum,* and develop a novel series of organism specific amino acid substitution matrices that would perform better than the standard matrices in the detection of orthologs and improve sequence alignments of the *Plasmodium* proteins.

As a first step towards this goal, a study was carried out as to how the amino acid composition of the *Plasmodium falciparum* proteins is driven by a nucleotide bias of the genome. Second, the amino acids substituted in ortholog proteins of the parasite compared to its distant relatives were explored and how these changes

affected *Plasmodium falciparum* specific substitution matrices was studied. As a third step, methods were designed and an algorithm developed to arrive at a novel series of symmetric and asymmetric substitution matrices that could be used specifically for sequence analysis of the *Plasmodium falciparum* proteins. These matrices were built from a unique dataset of annotated proteins of *Plasmodium falciparum* and its distant orthologs picked up from all three kingdoms of life. As a final step, the performance of these matrices was analyzed and reported and an attempt made to add functional annotations to some of the hypothetical proteins of the malaria parasite based on these novel matrices. A further attempt has been made to understand the pathways to which these proteins could be mapped and fill up gaps/holes in the metabolic pathway of the malaria parasite.

There has been a preconceived notion that conventional alignments using symmetric substitution matrices are pragmatically valid when homologous sequences are evolutionarily close. However, these are not theoretically accurate for comparing *Plasmodium falciparum* proteins with evolutionarily distant proteins from other species (Bastien et al. 2004). The asymmetric substitution matrix developed by us takes into account the directionality of the amino acid substitutions. A web server has been developed in this respect for pairwise alignments of potential orthologs of *Plasmodium falciparum* proteins with this novel asymmetric matrix. The alignments obtained with this matrix have been shown to be biologically more significant in terms of the improved alignment statistics and significant motif overlaps that were achieved for known proteins. The users can test

potential ortholog pairs for the biological relevance of the alignment obtained through our web server. Users also have an option to look for potential orthologs of *Plasmodium falciparum* proteins against other genomes with our Smat matrices based on BBH method.

The future scope of this work is to utilize these matrices in various sequence analysis tools for use. However, it demands the incorporation of these matrices in an appropriate way which can be achieved only by modifying the current tools/algorithms to accept user defined matrices. The annotation results presented in this thesis generate experimentally testable hypotheses and could serve as new drug targets, which would be of immense value for the rational design of anti-malarials.

# CHAPTER 2

---

## Amino acid substitution in *Plasmodium falciparum*

## 2.1 Introduction

The malaria parasite, *Plasmodium falciparum* has an extremely biased genome composition of A & T nucleotides. It is the most AT-rich genome sequenced to date. Previous studies (Musto et al. 1999; Singer and Hickey 2000) have shown that a nucleotide bias is likely to affect the codon usage and amino acid composition of proteins. For a majority of *Plasmodium falciparum* genes, codon usage is driven mainly by compositional constraints, while a small number of genes exhibit translational selection (Musto et al. 1999). Earlier studies carried out on the nuclear coding sequences of the malaria parasite showed the codon usage to be biased with respect to the 3rd codon position of genes with an abundance of A & T nucleotides (Musto, Rodriguez-Maseda, and Bernardi 1995). At the same time, certain amino acids like K, N, E, L & I were specially observed to be over-represented in the encoded proteins (Musto, Rodriguez-Maseda, and Bernardi 1995). There were also reports of compositionally biased genomes (AT-rich and GC-rich genomes) that exhibited amino acid compositional bias at the protein level (Singer and Hickey 2000). Studies carried out on 21 prokaryotic genomes; the *Saccharomyces cerevisiae* genome; and two of the chromosomes from the malaria parasite showed that the genomes with an extremely biased composition of AT nucleotides usually have a higher proportion of AT-rich codons in their genes that encode the FYMINK amino acids and the GC-rich genomes have proteins that are rich in GARP amino acids (Singer and Hickey 2000).

It is likely that the amino acid substitution in the malaria parasite is driven

by a nucleotide bias. Since proteins are in a constant state of evolution across species, certain changes at the nucleotide level (non-synonymous) in ortholog genes may possibly lead to amino acid replacements in corresponding proteins without affecting its function. However, it is also possible that these changes may affect the protein's overall structure and function, leading to an entirely new class of novel proteins. The complexity of *Plasmodium falciparum* genome and the availability of its complete genome sequence make it an excellent model to study amino acid substitution in a biased genome.

Since a genome wide nucleotide bias may affect the amino acid composition of proteins (Singer and Hickey 2000), it was interesting to study this effect in *Plasmodium falciparum*. This chapter deals with studies done to understand how the amino acid composition is driven by a nucleotide bias in *Plasmodium falciparum* (AT-rich) compared to a GC-rich genome. The overall affect of nucleotide composition on codon usage was also studied for ortholog proteins. To understand amino acid substitution, correlation of protein amino acid content and AT-rich nucleotide content of the corresponding genes (with respect to different codon positions) was studied for genomes differing in their overall nucleotide composition. Further amino acids that are substituted in ortholog proteins of the malaria parasite compared to its distant relatives were explored. The work presented in this chapter, has important implications in understanding amino acid changes in *Plasmodium falciparum* and assessing the need to re-compute scoring matrices for sequence analysis of biased genomes.

## 2.2 Methods

### 2.2.1 Amino acid composition and codon usage across diverse genomes

To understand the role of nucleotide bias on the amino acid and codon usage of an organism, the GC-rich and AT-rich genomes of *Mycobacterium tuberculosis* and *Plasmodium falciparum* respectively were selected for comparison. A list of *Plasmodium falciparum* proteins was obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/) (download date – 26th June 2006) by ftp (ftp://ftp.ncbi.nih.gov/genomes/Plasmodium_falciparum/). The incomplete annotations viz. putative, predicted and hypothetical were filtered out to give a final set of 302 proteins. A protein BLAST (version 2.2.10) search was performed with this set of annotated proteins as query against the *M. tuberculosis* (H37Rv) proteins as a database, at an E-value cut off equal to $10^{-5}$. An exclusive set (redundant) of 88 protein hits was identified for which the amino acid composition was calculated per protein. The statistical t-test for correlated samples was performed for each amino acid fraction obtained from this set for these organisms. In order to study the differences in the codon choice, the coding sequences for the same set of 88 proteins was retrieved and the fraction of codons coding for each individual amino acid was calculated. The amino acids Met and Trp were omitted from the analysis as these are coded by a single codon. The stop codons have been included though they do not code for any amino acids. The protein translation table and ORF information files (.ptt and .ffn respectively) of the organism available at NCBI's ftp site were used for this purpose. On this basis, codon frequencies were calculated and the

highly represented codons for each amino acid were tabulated for both the genomes.

### 2.2.2 Correlation studies of AT-rich codons

As protein evolution follows a universal trend of amino acid loss and gain, the composition of proteins is expected to vary substantially between taxa. Nucleotide bias in turn seems to cause an amino acid change (Singer and Hickey 2000). Acknowledging the divergence of *Plasmodium falciparum*, the effect of AT-codon content was studied along with its impact on the F, Y, M, I, N, K amino acid composition, that is known to be over-represented in AT-rich genomes. An analysis was carried out on an ortholog set of proteins and their corresponding coding regions across seven genomes. The organisms that were selected for this study were *Mycobacterium tuberculosis*, *Treponema pallidum*, *Escherichia coli K12*, *Helicobacter pylori*, *Lactobacillus johnsonii*, *Mycoplasma mycoides* and *Plasmodium falciparum.* These organisms had different AT nucleotide compositions of their genome (**Table 2.1**). The protein and coding sequences for each of these genomes were downloaded from NCBI's ftp site (ftp://ftp.ncbi.nih.gov/genomes/). A protein BLAST search was performed for the 302 completely annotated proteins (proteins that were not putative, hypothetical or predicted) of *Plasmodium falciparum* versus the rest of the organisms. A set of 36 proteins common to all organisms were picked, that had similar annotations. This protein dataset was used to estimate the F, Y, M, I, N, K amino acid composition. The AT-rich codon compositions with respect to the 1$^{st}$ and 2$^{nd}$ codon position, i.e. AT12 (non-synonymous), the third

position, i.e. AT3 (synonymous) and all three codon positions, i.e. AT123 was calculated from the corresponding coding sequences. Perl scripts were written for these calculations. A correlation was calculated for the AT-rich codon fractions so obtained and the F, Y, M, I, N, K amino acid composition of proteins.

**Table 2.1 List of organisms and their AT content**

| Organism | AT Content of genome |
|---|---|
| *Mycobacterium tuberculosis* | 34% |
| *Treponema pallidum* | 46% |
| *Escherichia coli K12* | 48% |
| *Helicobacter pylori* | 60% |
| *Lactobacillus johnsonii* | 65% |
| *Mycoplasma mycoides* | 76% |
| *Plasmodium falciparum* | 80% |

### 2.2.3 Dataset of *Plasmodium falciparum* protein orthologs

A sequence dataset was prepared that consisted of proteins with complete annotations from *Plasmodium falciparum* genome and its orthologs from both the microbial and eukaryotic species, for the study of amino acid substitutions. A list of *Plasmodium falciparum* proteins was obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/) through anonymous file transfer protocol (ftp) (ftp://ftp.ncbi.nih.gov/genomes/Plasmodium_falciparum/). The annotated proteins of *Plasmodium falciparum* (described in section 2.2.1) was considered for

ortholog search with the genomic BLASTp option available at NCBI. Distantly related orthologs were picked (E<1 was considered) manually, for this set of 302 *Plasmodium falciparum* proteins from 10–20 taxa representing all three domains of life. The criterion for the manual selection of orthologs was to select only those hits that had annotation similar to the query sequence, irrespective of the order of their E-values. The rationale was to pick distant relatives of *Plasmodium falciparum* proteins such that the amino acid substitutions were better studied.

Clustering of sequences was performed to remove redundancy in the ortholog protein set with the BLASTCLUST program from blast-2.2.10 package. Sequences were clustered at 90% identity over 80% of the sequence length. Proteins that showed few (proteins that gave ortholog hits to less than 10 organisms) or biased representation (proteins that gave hits to a biased group of organisms only, e.g. proteins showing hits to only *Plasmodium* genus) of orthologs to a particular kingdom were eliminated, reducing the working set to only 265.

### 2.2.4 Generation of blocks

Blocks are highly conserved un-gapped regions of a multiple sequence alignment of closely related group of protein sequences (Henikoff and Henikoff 1991). Protein blocks were derived from 265 annotated proteins of *Plasmodium falciparum* and their orthologs (a total of 4696 sequences) for which the protomat (Henikoff and Henikoff 1991) program was used. This program accepts a group of related proteins as input and produces a set of un-gapped alignments called blocks, representing the group. Protomat program was obtained from the BLIMPS package

that was available by anonymous ftp at NCBI (ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blimps/). The protein blocks generated in this study were later used for the compilation of substitution matrices as discussed in Chapter 3. Out of the total blocks obtained, a subset of blocks representing 79 protein sets common across *Plasmodium falciparum* and the three model genomes viz. *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* was identified. Though there were more proteins in common across the four genomes, the working set was restricted to only 79, to ensure that a representative sequence was present from all four genomes at the block level and the absence of even one, eliminated the protein from analysis. Amino acid substitutions occurring within the protein block were calculated with respect to each of these four model organisms. While calculating substitutions for a particular organism, the sequence depictive of that organism was made the first sequence of each block and the substitutions tabulated column wise with respect to the other sequences following it. Substitutions across different classes of amino acids, i.e. polar (S, T, Y, H, C, N, Q), hydrophobic (G, A, V, L, I, F, M, P, W) and charged amino acids (D, E, K, R), as well as substitutions within each class of the amino acids was computed and statistical tests were performed to weigh its relevance.

### 2.2.5 Statistical Analysis

All the statistical tests like paired sample t-test and ANOVA for correlated samples were performed using a website for statistical computation, VassarStats (http://faculty.vassar.edu/lowry/VassarStats.html). While paired sample t-test was

used to assess the statistical difference in the means of two correlated samples (here, individual amino acid composition, and, amino acid substitution pair frequencies, for a set of ortholog proteins across a pair of genomes was tested), ANOVA was used to calculate the variance across more than two correlated samples (here, differences in amino acid substitution pair frequencies across four genomes was analyzed). The analysis of variance used in this study was 'one way ANOVA' so as to study the variance in substitution pair frequencies of *Plasmodium falciparum* with respect to three other genomes.

## 2.3 Results and Discussion

### 2.3.1 Effect of genome complexity in *Plasmodium falciparum*

A comparison of the amino acid composition of ortholog proteins from the AT-rich and GC-rich genomes of *Plasmodium falciparum* and *Mycobacterium tuberculosis* respectively, showed significant differences. Differences were also observed in the choice of codons by both these organisms. While the average amino acid composition obtained for a redundant set of protein orthologs from both these genomes is represented in **Figure 2.1, Figure 2.2** demonstrates the most highly represented codon for each individual amino acid in the two genomes which clearly depicts that *Plasmodium falciparum* has a greater preference for possible AT-rich codons, considering all three codon positions.

The amino acid compositional differences obtained as represented in **Figure 2.1,** were highly significant for 75% of the amino acids (A, C, E, F, G, I, K, N, P, R, S, T, V, W, and Y) for which a P-value of <0.0001 was obtained for a two tailed t-test. The

differences were quiet significant for the amino acids D, L, M, and Q (P-value of 0.001–0.004) with the exception of H. The fraction of F, Y, M, I, N, K amino acids was found to be higher in *Plasmodium falciparum* as expected of AT-rich genomes (Singer and Hickey 2000). Apart from the above mentioned amino acids, the amino acid fractions of E, S and C were found to be significantly high in this study, compared to the corresponding *M. tuberculosis* fractions. 'E' is one of the six most ancient amino acids' that has been observed to be universally lost during evolution, which the malaria parasite seems to retain. However, it loses the other three amino acids i.e. P, A and G that are also known to be consistently lost in all three domains of life (Jordan et al. 2005). This may imply that 'E' codons, not being GC-rich, are retained by the parasite. Moreover, the most preferred codon for 'E' is GAA which is AT-rich (considering all codon positions) (**Figure 2.2**). The amino acids C, M, H, S and F are acquired with time (Jordan et al. 2005) and *Plasmodium falciparum* seems to have gained four of these amino acids in the course of evolution, as is evident from the increased fractions of C, M, S and F as compared to proteins of *Mycobacterium tuberculosis* that branches early in evolution. Much could not be deduced from the 'H' fractions as no significant difference was obtained in this case. The 'Y' amino acid fractions were also high in *Plasmodium falciparum* that are commonly held to be late additions to the genetic code (Brooks and Fresco 2002).

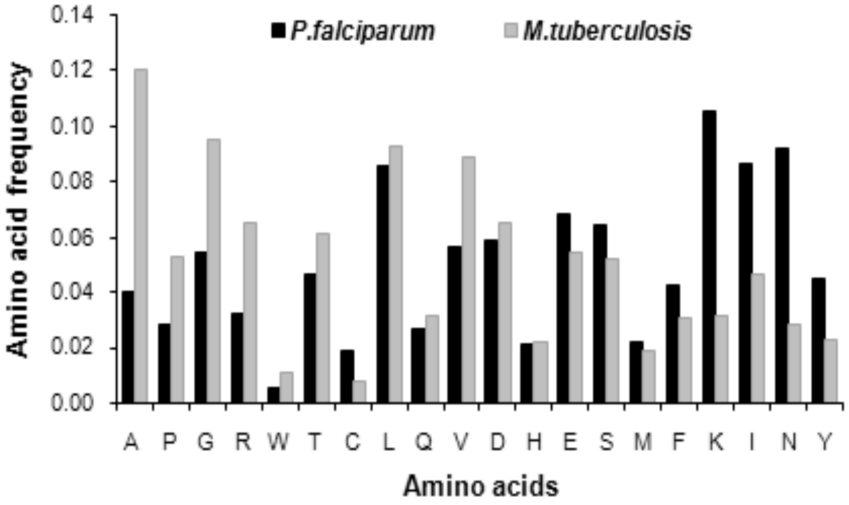


**Figure 2.1 Differences in the amino acid frequencies of an AT-rich and a GC-rich genome**. x-axis: Amino acids arranged in the increasing order of the AT-richness of their respective codons; y-axis: Average amino acid frequencies of protein orthologs from '*Plasmodium falciparum*' and '*Mycobacterium tuberculosis*'. The fractions of A, P, G, R, W, T and V amino acids are less in *P. falciparum* whereas the F, Y, M, I, N, K, S, E and C fractions are high compared to *M. tuberculosis.*

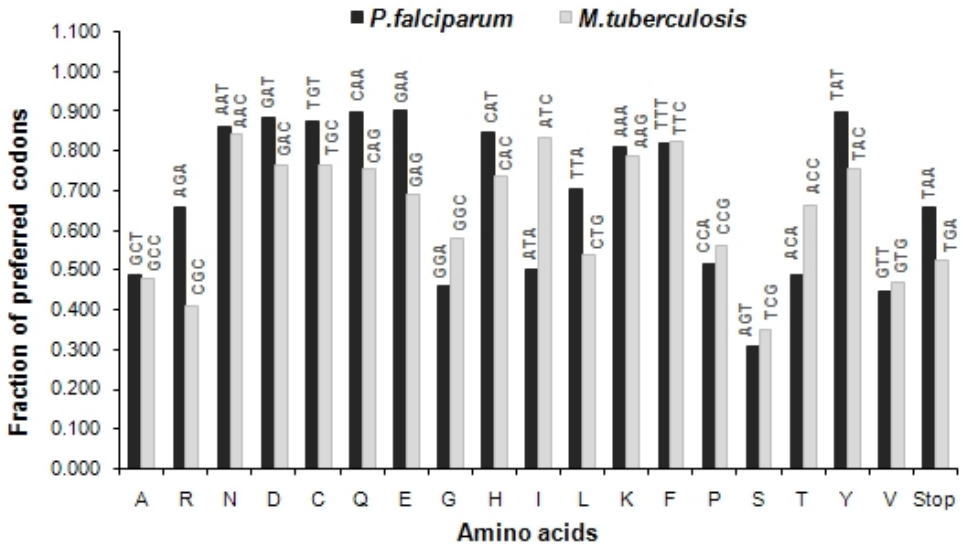


**Figure 2.2 Amino acid codon preferences for an AT-rich (*Plasmodium falciparum*) versus a GC-rich (*Mycobacterium tuberculosis*) genome**. The codon preferences clearly show the bias towards AT and GC codons for *P. falciparum* and *M. tuberculosis*, respectively.

### 2.3.2 Nucleotide bias directs amino acid substitution

Correlation studies between the F, Y, M, I, N, K amino acid composition of ortholog proteins from diverse genomes and the AT-rich codon fraction of the corresponding coding sequences revealed a significant relationship. **Figure 2.3** represents this correlation with respect to different codon fractions. The response of the F, Y, M, I, N, K amino acid composition to AT content at the 1$^{st}$ and 2$^{nd}$ codon positions (AT12) was found to be the highest (slope 0.748). Moreover, the co-efficient was maximum for AT12 (**Figure 2.3a**) indicating that the degree of variation in the F, Y, M, I, N, K amino acid usage could be well understood in terms of AT content at the 1$^{st}$ and 2$^{nd}$ codon positions (R2=0.98), compared to AT3 (R2=0.80) (**Figure 2.3b**). This implies that a bias at the non-synonymous position of a codon has affected the amino acid composition of the protein leading to a protein evolution. Conclusively, nucleotide bias in *Plasmodium falciparum* directs the amino acid substitution in a protein largely. However, when the total fraction of AT3 and AT12 were compared in *Plasmodium falciparum*, the fraction of AT3 was found to be more than 1.5 times that of AT12. This implies that though the organism allows for changes which may lead to a substitution; it still maintains a balance by having a high fraction of AT3 that would lead to only synonymous mutations.

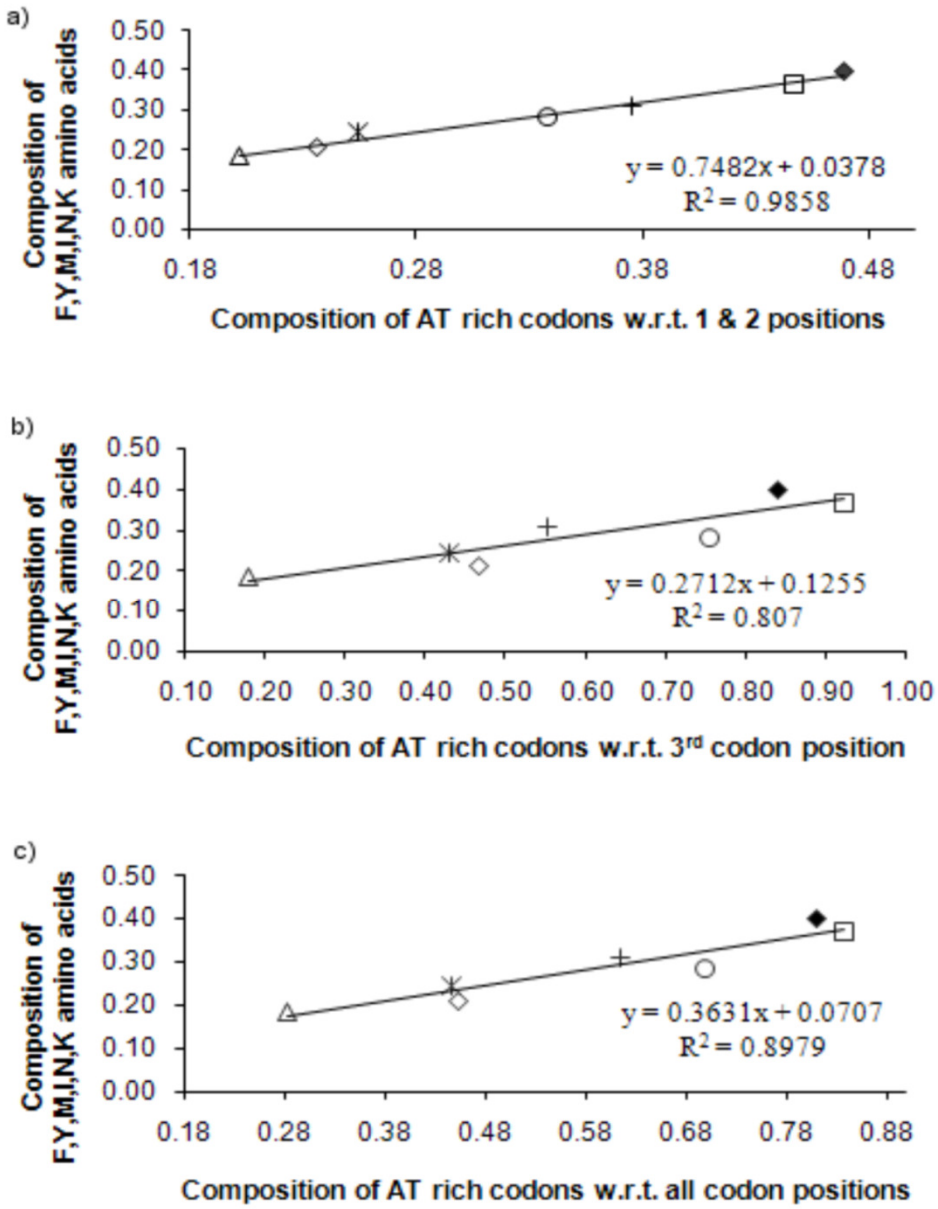


**Figure 2.3 Correlation of AT-rich codon composition and F, Y, M, I, N, K amino acid frequencies**. (**a**) Correlation between AT12 fraction and the composition of F, Y, M, I, N, K amino acids. (**b**) Correlation between AT3 fraction and the composition of F, Y, M, I, N, K amino acids. (**c**) Correlation between AT123 fraction and F, Y, M, I, N, K amino acid composition. Note: the symbols used here are; *Mycobacterium tuberculosis* (triangle), *Escherichia coliK12* (asterisk), *Treponema pallidum* (diamond), *Helicobacter pylori* (cross), *Lactobacillus johnsonii* (circle), *Plasmodium falciparum* (solid diamond), *Mycoplasma mycoides* (square).

### 2.3.3 Amino acid preferences in protein blocks

Having shown that the AT bias of the genome has important influence on amino acid choices in a biased genome, it was interesting to study what amino acid substitutions are observed in the protein blocks of *Plasmodium falciparum* as compared to other model genomes like *Drosophila melanogaster*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. The study of substitutions across and within different classes of amino acids, i.e. polar (S, T, Y, H, C, N, Q), hydrophobic (G, A, V, L, I, F, M, P, W) and charged amino acids (D, E, K, R) gave some interesting observations. It was observed that in *Plasmodium falciparum*, the hydrophobic amino acid fractions substituted for polar and charged amino acids were less as compared to *A. thaliana*, *S. cerevisiae* and *D. Melanogaster* (**Figure 2.4**). A two-tailed paired sample t-test values of P for these fractions were, $P<0.0001$ for *Plasmodium falciparum* versus *A. thaliana*; $P<0.0001$ for *Plasmodium falciparum* versus *S. cerevisiae*; and, $P=0.001$ for *Plasmodium falciparum* versus *D. melanogaster*. Conversely, the polar amino acids substituted for the hydrophobic and charged amino acids were higher compared to the other three organisms (**Figure 2.4**). The P-values for these fractions were as follows: $P<0.0001$ for *Plasmodium falciparum* versus *A. thaliana*; $P=0.0006$ for *Plasmodium falciparum* versus *S. cerevisiae*; $P<0.0001$ for *Plasmodium falciparum* versus *D. melanogaster*.

In order to understand the individual substitutions which were actually significant among these groups, the fractions of each type of substitution were compared for all the four organisms. Among the hydrophobic substitutions, the

fractions of 'I' substituted for R, K, C, T and N was high in *Plasmodium falciparum*. On the other hand, the fractions of 'A' substituted for K; 'G' substituted for Q; and 'P' substituted for N was less (**Figure 2.5**). A one-way analysis of variance, ANOVA, for correlated samples gave a Tukey's HSD post hoc test value of P as follows; $P<0.05$ for RI, CI, NI, QG, NP and $P<0.01$ for KI, TI, KA, for *Plasmodium falciparum* versus other three organisms. In case of the polar substitutions, the fractions of 'N' substituted for F, G, and P; and the fractions of 'Y' substituted for K and R was high in *Plasmodium falciparum*, whereas the fractions of 'S' substituted for K was less (**Figure 2.6**). The Post hoc test values of 'P' for *Plasmodium falciparum* versus others were; <0.05 for FN, PN, KY, RY and; $P<0.01$ for GN and KS. Among the substitutions occurring within the same class of amino acids, only the polar to polar fractions were significant (**Figure 2.7**). A post hoc test showed $P<0.05$ for the polar to polar fractions when an ANOVA test was performed for *Plasmodium falciparum* versus other organisms. Again, among these, only the SN (N in *Plasmodium falciparum* substituted for S in others) and QN (N in *Plasmodium falciparum* substituted for Q in others) fractions were high in *Plasmodium falciparum* (**Figure 2.8**). A post hoc test showed $P<0.01$ when ANOVA was performed. All these differences were thus statistically significant.

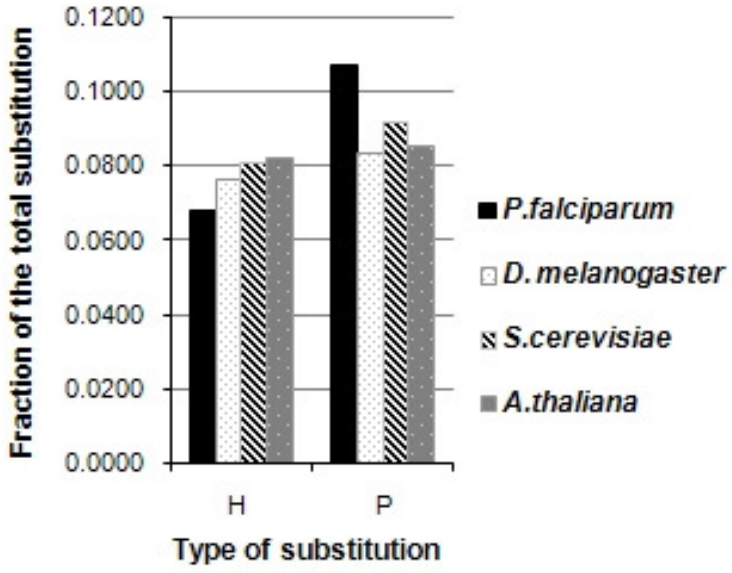


**Figure 2.4 Differences in substitution across four genomes for hydrophobic and polar amino acid fraction of proteins.** 'H' along the x-axis stands for hydrophobic residues substituted to, in the organism, for polar or charged residues. 'P' stands for polar residues substituted to, in the organism, for charged or hydrophobic residues.

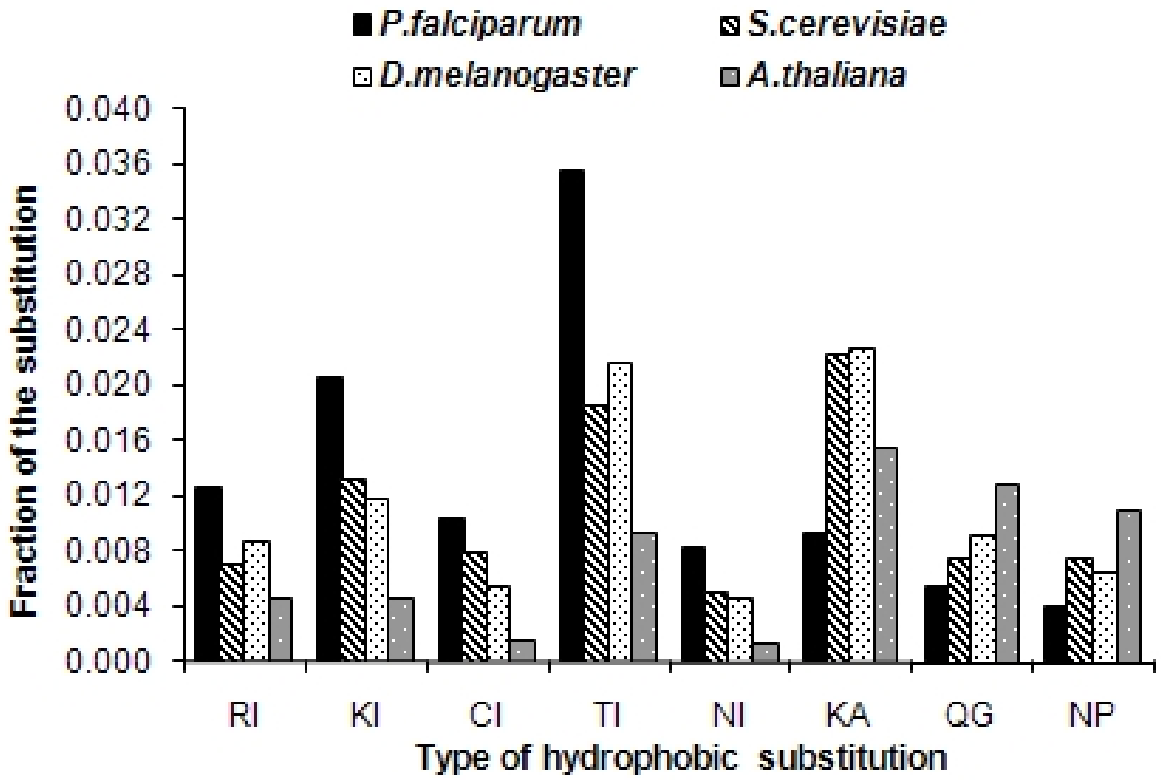


**Figure 2.5 Type of hydrophobic substitutions significant in *Plasmodium falciparum* compared to other organisms**. The type of hydrophobic substitutions are represented along the x-axis where, the first alphabet represents the amino acid substituted for and the second alphabet is the one substituted to in the representative organism.

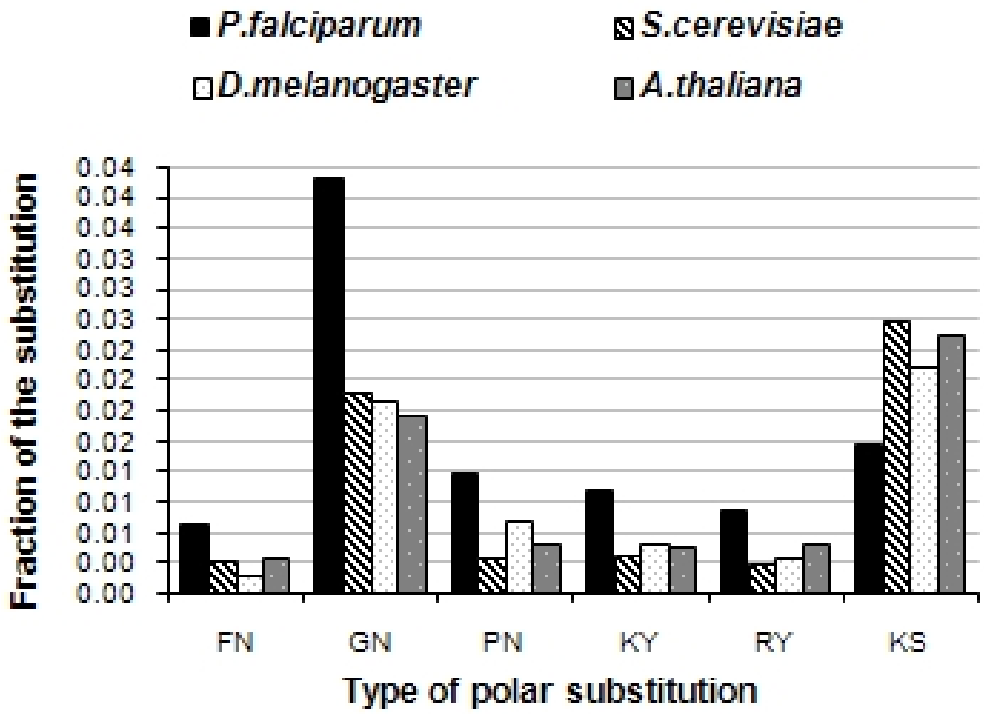


**Figure 2.6 Polar substitutions significant in *Plasmodium falciparum* compared to the other organisms.** The type of polar substitutions are shown along the x-axis where, the first alphabet represents the amino acid substituted for and the second alphabet is the one substituted to in the representative organism.

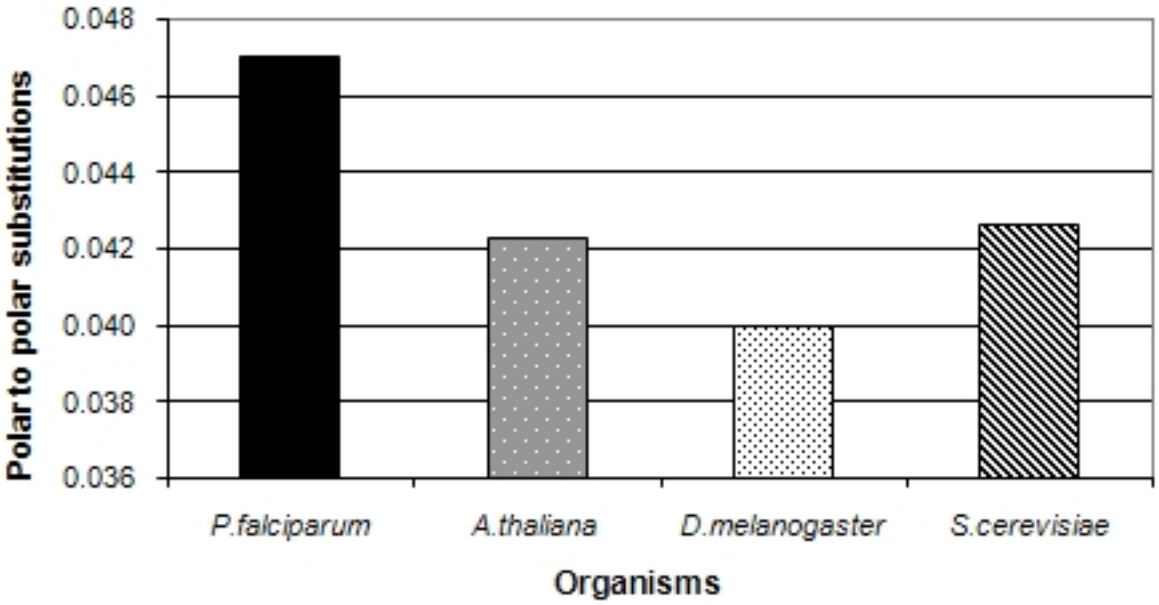


**Figure 2.7 The differences in the total fraction of polar to polar substitutions across four genomes**.

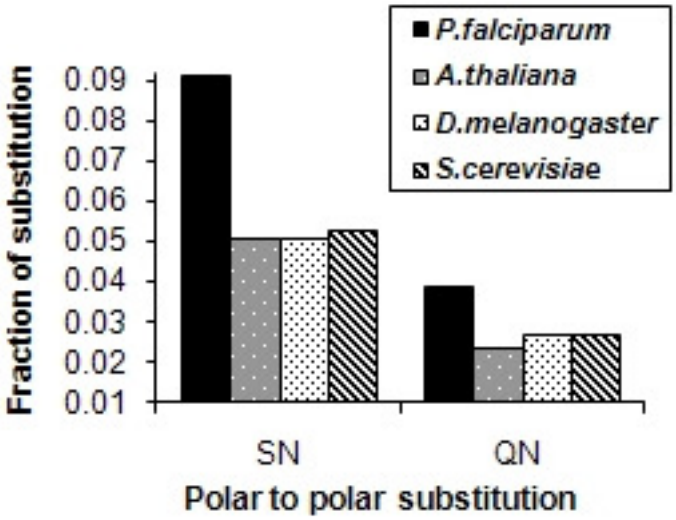


**Figure 2.8 Significant polar to polar substitutions across four genomes**. The type of polar to polar substitutions are represented along the x-axis where, the first alphabet represents the amino acid substituted for and the second alphabet is the one substituted to, in the representative organism.

## 2.4 Conclusion

The present chapter deals with understanding the influence of the extreme nucleotide bias observed in the *Plasmodium falciparum* genome on the amino acid composition of the encoded proteins and the codon usage of the parasite genome. The key idea was to know these changes in proteins of *Plasmodium falciparum* in comparison with ortholog proteins from other genomes that have an overall nucleotide composition different from that of the malaria parasite. The inference out of this work is that certain amino acids are preferentially represented in the proteins of the malaria parasite that are usually encoded by AT-rich codons and are thus a direct outcome of the nucleotide bias (80.6% AT) of the genome. Apart from this, the choice of the codons for amino acids is also influenced by the overall AT content of the genome.

The correlation between the AT-codon content of synonymous and non-synonymous codon positions and the FYMINK amino acid composition of the proteins across different genomes showed that the degree of variation in the F, Y, M, I, N, K amino acid usage could be well understood in terms of AT-content at the non-synonymous codon positions. A nucleotide bias thus directs amino acid substitution in *Plasmodium falciparum.*

An attempt to understand the nature of amino acid substitution in protein blocks of *Plasmodium falciparum* orthologs revealed some interesting results. The parasite showed preferential substitutions across different classes of amino acids in conserved regions of ortholog proteins, when compared to its distant relatives. The

study presented in the current chapter was significant to understand how these amino acid changes would affect the calculation of *Plasmodium falciparum* specific substitution matrices, discussed in the subsequent chapter.

# CHAPTER 3

## Compilation of a novel series of amino acid substitution matrices

## 3.1 Introduction

The protein sequence comparison methods make use of a frequency table of amino acid changes occurring in protein families known as 'amino acid substitution matrices'. These consist of log-likelihood scores which reflect how likely one amino acid is substituted over the other. Some of the popular general purpose matrices used in sequence alignments was initially developed by Margaret Dayhoff, and Henikoff & Henikoff known as the PAM (Dayhoff 1978) and the BLOSUM (Henikoff and Henikoff 1992) series respectively, which are symmetric in nature. While Dayhoff PAM matrices are based on an evolutionary model of protein change, the BLOSUM matrices are designed to identify members of the same family and are obtained from a large diverse set of proteins. But, are these substitution matrices constructed from standard databases appropriate for the comparison of compositionally drifted proteins of *Plasmodium falciparum*? When the scoring system for proteins is considered, usually the identical residues and conservative substitutions are found to have positive values in the matrix. Rare substitutions are given a negative score and are usually avoided by the alignment algorithm (Vingron and Waterman 1994). In Chapter 2, it has been shown that *Plasmodium falciparum* has apparently diverged from other organisms, hence rare substitutions may be expected in this parasite. Could this be one of the reasons why alignment programs fail to show good sequence similarities with the standard matrices, for majority of the parasite proteins?

The use of standard matrices for the comparison of proteins with non-

standard compositions was a matter of concern for a long time, though no appropriate solution was immediately available to deal with this issue (Sutormin, Rakhmaninova, and Gelfand 2003; Yu, Wootton, and Altschul 2003; Yu and Altschul 2005). A new rationale for the compositional adjustment of amino acid substitution matrices was however proposed by Yu and coworkers (Yu, Wootton, and Altschul 2003; Yu and Altschul 2005), where the target frequencies of the standard matrices were transformed to frequencies appropriate in a non-standard context. The method employed by this group was indirect, i.e. while the basic matrix used for searches was the same; the compositional adjustment was made only as a final step to improve the E-values and scores. As a result, this method rarely altered the matching sequences that appeared in the output (Altschul et al. 2005). Yet another approach was where the authors had proposed a method for the construction of asymmetric matrices for proteins with biased amino acid distribution, where basically sequence pairs from two different genomes were compared (Bastien, Roy, and Marechal 2005). In the light of evolution, such asymmetric matrices are usually superior to the symmetric matrices due to their directionality. A new aspect of compositional bias was addressed recently by Brick and Pizzi who proposed a novel procedure to construct a series of symmetric substitution matrices for the alignment of proteins from similarly biased *Plasmodium* proteomes (Brick and Pizzi 2008). These matrices were generated by selecting only those blocks from the BLOCKS database that had compositional bias similar to the *Plasmodium falciparum* and *Plasmodium yoelli* genomes. However, no one has actually addressed ways to

capture the substitutions in *Plasmodium falciparum* proteins with respect to distant relatives and re-compute substitution matrices from an entirely new dataset of proteins that would represent unique substitutions for the specific organism in question.

In the previous chapter it was shown how the amino acid substitution in the AT-biased *Plasmodium falciparum* genome could differ from other model organisms. As a further step it was interesting to explore computation of a novel series of *Plasmodium falciparum* specific substitution matrices and study its performance in sequence alignments of the parasite proteins. In this context, a series of amino acid substitution matrices (symmetric and non-symmetric) named as the PfSSM (*Plasmodium falciparum* specific substitution matrix) series were developed based on the BLOSUM approach of Henikoff & Henikoff (Henikoff and Henikoff 1992) which has been described in the present chapter.

This chapter explains that for biased genomes like *Plasmodium falciparum*, substitution matrices derived from a unique ortholog set of *Plasmodium* biased proteins is more appropriate for organism specific sequence searches, as it is expected to resolve the enigma of inconsistent background and target frequencies, which is a common problem with the generalized matrices. Unlike earlier methods for overcoming compositional bias (Yu, Wootton, and Altschul 2003), the substitutions were recalculated from conserved blocks built out of this new dataset. These blocks represent both the *Plasmodium falciparum* sequence and its distant orthologs for matrix construction. While the symmetric series were built on

BLOSUM model, the asymmetric matrices were modified to represent directional amino acid substitutions with respect to *Plasmodium falciparum*. The asymmetric series were compiled with the assumption that the pair-wise alignments of *Plasmodium falciparum* proteins and its potential orthologs would give better alignments. Finally, the performance of these matrices was validated and reported in terms of the alignment quality and statistics obtained for some of the *Plasmodium falciparum* proteins.

## 3.2 Methods

### 3.2.1 Dataset for matrix compilation

In order to arrive at a unique dataset of proteins for matrix compilation that represented amino acid substitutions across distant relatives, *Plasmodium falciparum* proteins with complete annotations were selected and its distant orthologs were identified as described earlier in chapter 2. As previously mentioned the orthologs were picked manually for the annotated proteins of *Plasmodium falciparum* with the genomic BLASTp option at NCBI (E<1) from 10–20 taxa representing all three domains of life. The ortholog proteins thus represented both microbial and eukaryotic genomes.

The manual selection of the orthologs for the *Plasmodium* proteins was done in a stringent way and is discussed as follows. First, only those proteins were selected that had annotation similar to the query sequence. Second, annotated hits were picked up irrespective of the order of their E-values, to get distant orthologs. Third, over-representation of subject hits to a particular taxonomic group was

avoided (to represent wide range of substitutions), and, lastly, in case of hypothetical hits that were picked up (to represent a particular taxa that lacked an annotated hit), E-value near to zero ($<10^{-5}$) and length similar to the query was considered. However, the third option was rarely used and the total hypothetical proteins constituted only 6–7% of the total sequences used to build the matrices. This protein dataset is identical to the one used for amino acid substitution study in protein blocks as described in Chapter 2.

Clustering was performed with BLASTCLUST program from the blast-2.2.10 package to remove redundancy in the ortholog protein set. Sequences were clustered at 90% identity over 80% of the sequence length. Proteins that showed few (proteins that gave ortholog hits to less than 10 organisms) or biased representation (proteins that gave hits to a biased group of organisms only, e.g. proteins showing hits to only *Plasmodium* genus) of orthologs to a particular kingdom were eliminated, reducing the working set to only 265.

### 3.2.2 Protein blocks

Substitution matrices derived from the highly conserved regions of the protein are known to perform better in alignments and homology searches (Henikoff and Henikoff 1992). Protein blocks were derived from 265 annotated proteins of *Plasmodium falciparum* and their orthologs (a total of 4696 sequences) using the protomat program from the BLIMPS package as described in the previous chapter. In order to reduce the over-representation of amino acid pair frequencies from closely related members of a group of sequences, segment clustering within the

block was performed at different clustering percentages of 50%, 60%, 70%, 80% and 90%, over the entire block width. Approximately 1500 blocks were obtained that were processed for the calculation of a substitution matrix.

### 3.2.3 *Plasmodium falciparum* Specific Substitution Matrices (PfSSM)

The substitution matrices were computed from the protein blocks derived from the unique dataset of annotated *Plasmodium falciparum* orthologs using an in-house developed Perl script that was based on Henikoff's formalism for BLOSUM matrix compilation (Henikoff and Henikoff 1992) and was as follows:

**General formula:**

$$S_{ij} = \frac{1}{\lambda} \ln (q_{ij}/p_i p_j)$$

Where, $S_{ij}$ is the substitution score in bits; $\lambda$ is the scaling factor; $q_{ij}$ is the target frequency; $p_i p_j$ is the background frequency; and, ln is the natural logarithm.

**Target or observed frequencies:** It is the relative frequency, $q_{ij}$ of occurrence of an amino acid pair $A_iB_j$, and is given as;

$$q_{ij} = f_{ij} \Big/ \sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}$$

Where, $f_{ij}$ is the observed substitution frequency of the pair $A_iB_j$.

In the present study, the target frequencies were calculated from the dataset of 265 protein orthologs of *Plasmodium falciparum*.

**Background or expected frequencies:** The frequency of occurrence of an individual amino acid, $A_i$ in a pair $A_iB_j$ is calculated as;

$$p_i = q_{ii} + \sum_{i \neq j} q_{ij}/2$$

Similarly, the frequency of occurrence of the amino acid $B_j$ in the pair $A_iB_j$ is given as,

$$p_j = q_{jj} + \sum_{i \neq j} q_{ij}/2$$

The background or expected frequency $e_{ij}$ for the substitution of the pair $A_iB_j$ is then given as,

$$e_{ij} = (p_i p_j + p_j p_i) \text{ or, } e_{ij} = 2p_i p_j \quad (\text{for } i \neq j)$$

$$\text{and, } e_{ij} = p_i p_j \quad (\text{for } i = j)$$

**Odd ratio (likelihood ratio):** The odd ratio is calculated as a ratio of the observation frequency to the expected frequency for each 20x20 matrix entry leading to data normalization and is given as,

$$r_{ij} = q_{ij}/e_{ij}$$

where, $q_{ij}$ is the observation frequency of a substitution pair and $e_{ij}$ is the expected frequency of that pair.

**Raw score:** The natural log of the odd ratio is the raw score value for the matrix and is given as follows,

$$S_{ij} = \ln (q_{ij}/e_{ij})$$

**Bit score:** The raw score is normalized with a factor lambda such that the matrix could be directly used for alignment studies. In this case, the value of lambda was taken as the natural log of 2 which is equal to 0.693 such that the base for the log

odds matrix changes to that of 2 as shown below,

$$S_{ij} = \frac{1}{\lambda} \ln (q_{ij}/e_{ij})$$

$$S_{ij} = \frac{1}{(\ln 2)} \ln (q_{ij}/e_{ij})$$

$$\text{or, } S_{ij} = \log_2 (q_{ij}/e_{ij})$$

where, $\ln 2 = 0.693$

**Scaling:** Each matrix entry was rounded off to the nearest integer and all the values scaled to half-bit units by multiplying with the factor 2.

$$S_{ij} = 2 \log_2 (q_{ij}/e_{ij}) \quad \text{(Half-bit units)}$$

**The relative entropy:** The relative entropy or the relative information content of the matrix was calculated as

$$\mathrm{H} = \sum_{i,j} q_{ij} S_{ij}$$

where, H is the entropy in bits; $q_{ij}$ is the observation frequency; and, $S_{ij} = \log_2(q_{ij}/p_i p_j)$ i.e. substitution score or lod score in bits.

**Expected Score** One of the important parameters of the scoring matrix of statistical use is the expected score (E) of the matrix where the score for each amino acid pair ($S_{ij}$) is multiplied by the fractional occurrences of each amino acid ($p_i$ & $p_j$) and the weighted score summed over all the amino acid pairs. E was thus calculated as,

$$E = \sum_{i=1}^{20} \sum_{j=1}^{i} p_i p_j S_{ij}$$

where, $E$ is the expected score; $p_i p_j$ is the expected frequency; and, $S_{ij}$ is the raw score.

The Perl code was appropriately modified for calculating the asymmetric *Pf* fixed and the scaled version of the substitution matrices that are discussed later in this chapter. The entire matrix series was named as ***P****lasmodium* ***f****alciparum* **S**pecific **S**ubstitution **M**atrix i.e the PfSSM series.

#### 3.2.3.1 Symmetric matrix series

A symmetric matrix was calculated initially where the value for the substitution pair $A_i B_j$ (where $i \neq j$) was given the same as the pair $A_j B_i$. For e.g. if the observation count for A to L substitution is 'x' and that of L to A substitution is 'y' then it is assumed that AL = LA = (x+y) and thus the observation frequency for the pair AL or LA is (x+y/N), where N is the total number of pair observations. Consequently, all possible pair-wise substitutions are tabulated across protein blocks in this case. These matrices were generated at varying block clustering percentages of 50, 60, 70, 80, & 90; scaled to half-bit values, rounded off, and named as the Smat (**S**ymmetric **mat**rix) series. A diagrammatic illustration of symmetric matrix compilation is provided in **Figure 3.1.**

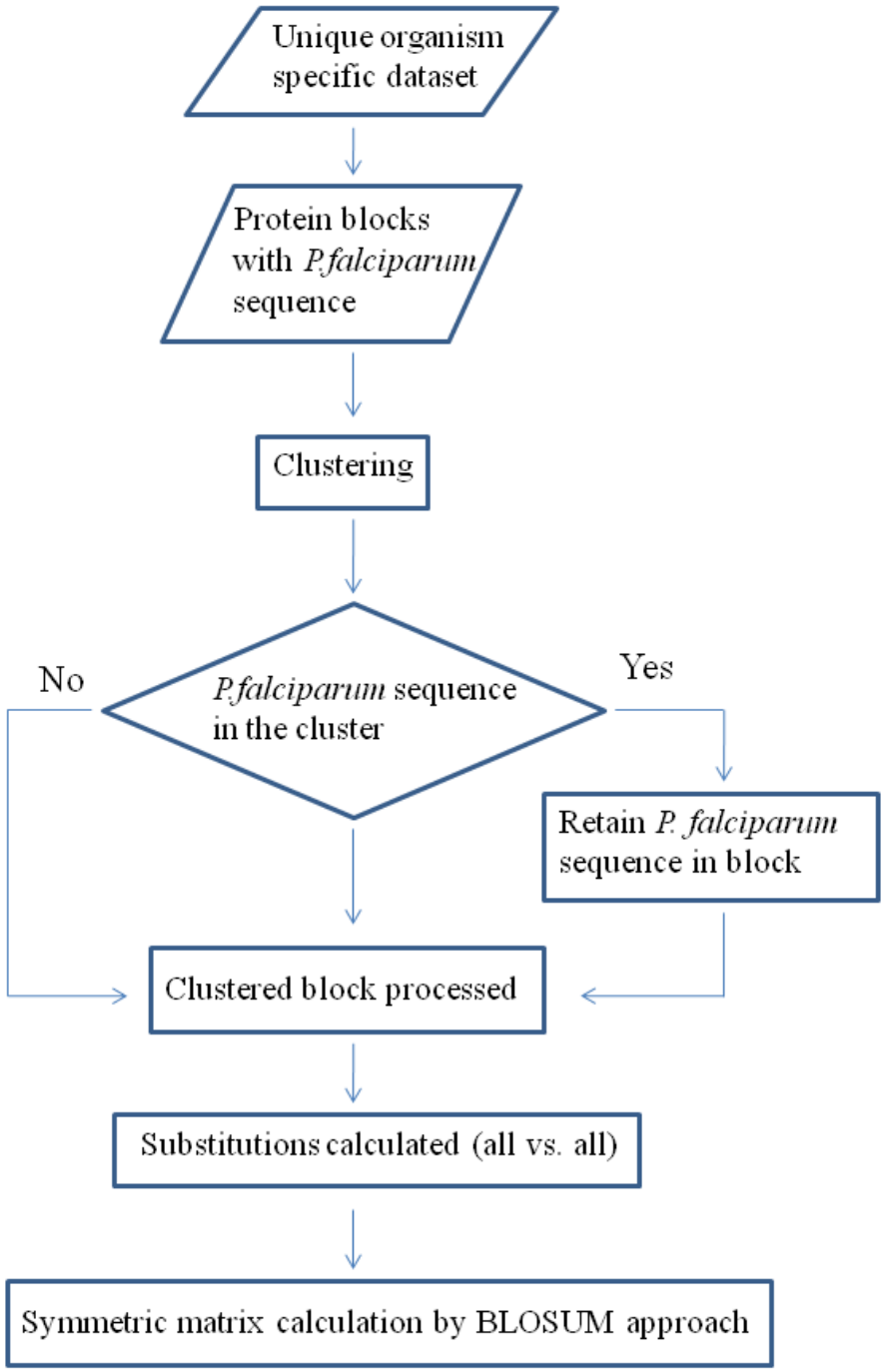


**Figure 3.1 A schematic representation for the calculation of a symmetric substitution matrix for *Plasmodium falciparum***

### 3.2.3.2 Asymmetric matrix series

An asymmetric matrix goes with the assumption that the observation $A_i B_j \neq A_j B_i$ and thus the values across the matrix diagonal are asymmetric. Thus considering the same e.g. if the observation count for A to L substitution is 'x' and that of L to A substitution is 'y' then the observation counts are AL = x and LA = y and their observation frequencies are x/N and y/N respectively, where N is the total number of pair observations. The script was modified to incorporate these changes giving rise to an asymmetric matrix. The program places the *Plasmodium falciparum* protein sequence at the first position of every block that PROTOMAT generates and then the substitutions tabulated one against all. In case of a tie between more than 2 sequences in eliminating sequences at block/sequence level clustering, *Plasmodium falciparum* sequence was retained as the cluster representative. These matrices were generated at varying block clustering percentages of 50, 60, 70, 80, & 90 (similar to the Smat series); scaled to half-bit values and rounded off. Since the *Plasmodium* sequence is fixed as the first sequence of every block and the substitutions calculated with respect to it, they were named as the one-way substitution or the ***P****lasmodium* ***f****alciparum* **F**ixed **mat**rix series (PfFmat). A diagrammatic illustration for asymmetric matrix calculation is given in **Figure 3.2.**

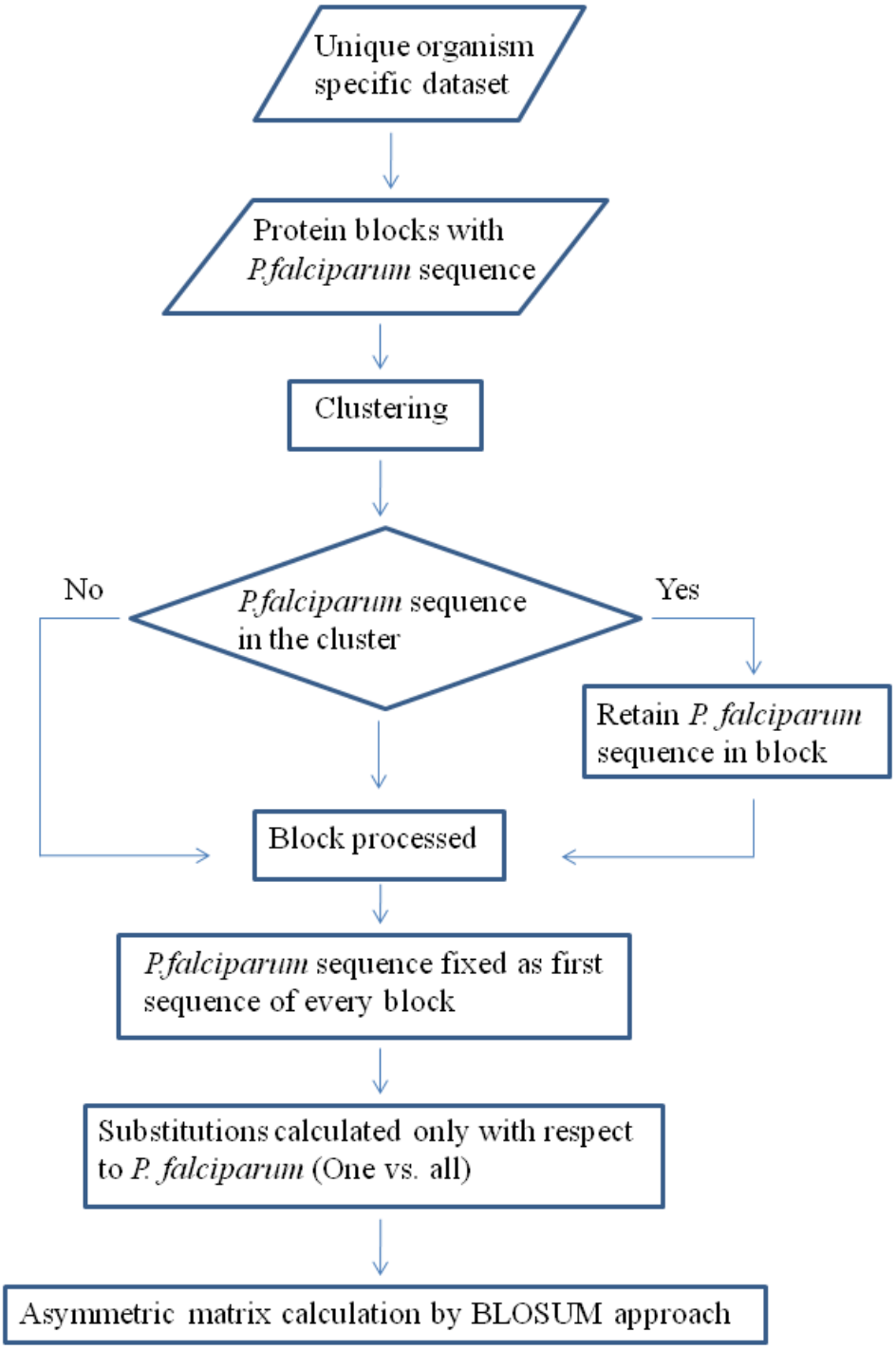


**Figure 3.2 A schematic representation for the calculation of an asymmetric substitution matrix for *Plasmodium falciparum***

#### 3.2.3.3 Scaled matrices

Substitution matrices used for studying local similarities are known to have no essential effect with the addition of a constant to all matrix values (Altschul 1991). A scaled version of the PfSSM matrices was calculated by adding a constant positive number to all the matrix values; +1 for symmetric and, +3 for the asymmetric *Plasmodium falciparum* fixed substitution matrices. These were termed as the SSmat (**S**ymmetric **S**caled **mat**rix) and the PfFSmat (***P****lasmodium* ***f****alciparum* **F**ixed **S**caled **mat**rix) series respectively.

### 3.2.4 Performance evaluation of PfSSM matrices

The performance of the PfSSM series was tested with alignment programs from the FASTA package (version3). SSEARCH (Smith and Waterman 1981) that uses William Pearson's implementation of the method of Smith and Waterman was used to search the Uniprot/Swiss-Prot protein database obtained from the EBI ftp site (http://www.ebi.ac.uk/FTP/) for relevant hits to *Plasmodium falciparum* queries and the FASTA program for pair-wise alignments. All the alignments were performed at a gap opening and extension penalty of -12 and -2, respectively, since FASTA programs are known to work best at these parameters. The bit-scores, E-values and alignment overlap obtained for these alignments with the standard and test matrices were compared.

## 3.3 Results and Discussion

### 3.3.1 Characteristics of the PfSSM series

The PfSSM series compiled as described in the methods section consisted of two types of substitution matrices; the symmetric (Smat) series and the asymmetric (PfFmat) series. While the Smat series was an outcome of all vs. all substitutions calculated from protein blocks the PfFmat series reflects substitutions (*Plasmodium falciparum* vs. all) that have occurred in *Plasmodium falciparum* over time compared to its other orthologs. It is known that substitution matrices used for studying local similarities have no essential effect with the addition of a constant to all matrix values (Altschul 1991). Hence, a scaled version of the above two matrix series (SSmat and PfFSmat) was compiled by adding a constant value to all the matrix elements such that the lowest negative value was in the range of BLOSUM matrices and the overall average of the matrix was slightly negative. The choice of a negative matrix average was to ensure that the alignments fell in the logarithmic region of a phase transition curve such that the alignment statistics were better understood (Vingron and Waterman 1994). Good alignments are usually expected in the logarithmic region while in the linear region longer, biologically irrelevant alignments take over. Hence, the Smat and PfFmat matrices were scaled to bring the matrix average to slightly negative values. The matrix averages for the PfSSM series is provided in **Table 3.1**.

To ensure data sufficiency in building the amino acid substitution matrices, a comparison was made between the PfSSM series and the standard BLOSUM

matrices in terms of the total protein blocks used in building the substitution matrices (as given in **Table 3.2)** and the number of amino acid contributing pairs for each matrix (**Table 3.3)**.

Apart from this various other parameters of the scoring matrices like the relative entropy, expected score and trace (Sutormin, Rakhmaninova, and Gelfand 2003) of matrices was calculated for the PfSSM series.

The relative entropy of the matrix, calculated as shown in the methods section is important for the comparison of matrices for the study of their performance. Usually matrices that are equivalent in either their relative entropy values or the trace of the matrix are comparable (Henikoff and Henikoff 1992; Sutormin, Rakhmaninova, and Gelfand 2003). The trace of a matrix is the sum of the diagonal values of the observation frequency matrix elements and represents the average identity of proteins used to construct the matrix. The relative entropy and trace values (along with the average protein identity) for the PfSSM series and some standard matrices are tabulated in **Table 3.4** and **Table 3.5** respectively.

Similarly, another important parameter of the scoring matrices is the expected score (E) which should be negative for a substitution matrix if the alignment scores for a matrix are to be used for statistical tests (Karlin and Altschul 1990; Altschul 1991). If the expected score is positive for a matrix, the growth of the score of the local alignments is linear even under strong gap penalties (Vingron and Waterman 1994). The expected scores of the PfSSM series calculated in this case were negative as represented in **Table 3.6**.

**Table 3.1 Matrix averages for the PfSSM series**

| | Clustering % | | | | |
|---|---|---|---|---|---|
| **Matrix** | **50** | **60** | **70** | **80** | **90** |
| Smat | -2.130 | -2.110 | -2.115 | -2.215 | -2.400 |
| SSmat | -1.133 | -1.110 | -1.115 | -1.215 | -1.400 |
| PfFmat | -3.203 | -3.260 | -3.300 | -3.440 | -3.653 |
| PfFSmat | -0.203 | -0.215 | -0.300 | -0.440 | -0.653 |

**Table 3.2 Total contributing blocks in calculating the substitution matrix**

| | BLOSUM | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **30** | **35** | **40** | **50** | **80** | **90** | **Smat/SSmat** | **PfFmat/PfFSmat** |
| **Total contributing blocks** | 261 | 439 | 672 | 1161 | 1941 | 2022 | 1482 | 1283 |

**Table 3.3 Total pairs contributing to the substitution matrix**

| **Clustering %** | **30** | **35** | **40** | **50** | **60** | **70** | **80** | **90** |
|---|---|---|---|---|---|---|---|---|
| **BLOSUM** | 70 394 | 136 328 | 227 179 | 557 053 | 1 087 140 | 1 930 219 | 2 907 119 | 5 196 613 |
| **Smat/SSmat** | - | - | - | 1 775 418 | 1 931 436 | 2 426 200 | 3 385 995 | 4 770 031 |
| **PfFmat/PfFSmat** | - | - | - | 251 757 | 274 192 | 332 829 | 427 484 | 538 600 |

**Footnote:** For the Smat/SSmat series the total contributing pairs is equal to half-diagonal values including the main diagonal (380/2+20), while for the asymmetric PfFmat/PfFSmat series it is given as (20x20) values.

**Table 3.4** The relative entropy (bits) of PfSSM series and the comparable standard matrices

| C% or PAM dist. | 2 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|
| Smat | - | 1.13 | 1.10 | 1.10 | 1.16 | 1.27 | - |
| SSmat | - | 2.13 | 2.10 | 2.10 | 2.16 | 2.27 | - |
| PfFmat | - | 1.68 | 1.67 | 1.70 | 1.77 | 1.88 | - |
| PfFSmat | - | 4.68 | 4.67 | 4.70 | 4.77 | 4.88 | - |
| BLOSUM | - | 0.48 | 0.66 | 0.83 | 0.98 | 1.18 | 1.45 |
| PAM | 3.9 | 2.0 | 1.79 | 1.60 | 1.44 | 1.30 | 1.18 |

**Footnote:** C% – clustering percentage; PAM dist. – PAM distance (in case of PAM matrices)

**Table 3.5** Trace values and average identity of proteins used to build PfSSM series and the standard BLOSUM matrices

| Clustering % | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|
| **Smat/SSmat** | tr = 0.50 (50.3%) | tr = 0.50 (49.6%) | tr = 0.49 (49.5%) | tr = 0.51 (50.9) | tr = 0.54 (53.7) | - |
| **PfFmat/PfFSmat** | tr = 0.66 (65.7%) | tr = 0.65 (65.4%) | tr = 0.66 (65.8%) | tr = 0.67 (67.2%) | tr = 0.69 (69.4%) | - |
| **BLOSUM** | tr = 0.27 (27.1%) | tr = 0.32 (32.1%) | tr = 0.37 (36.9%) | tr = 0.41 (40.7%) | tr = 0.46 (45.9%) | tr = 0.53 (52.7%) |

**Footnote:** The values in the bracket are the average identity of proteins used to build the matrices; tr = trace of the matrix

**Table 3.6** The expected scores of the PfSSM matrix series

| Clustering % | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|
| Smat/SSmat | -1.6367 | -1.6055 | -1.6082 | -1.6895 | -1.8276 |
| PfFmat/PfFSmat | -2.4295 | -2.4222 | -2.4682 | -2.5792 | -2.7342 |

### 3.3.2 Comparison of PfSSM and the standard BLOSUM matrices

To understand the substitution preferences of BLOSUM and PfSSM matrices, BLOSUM90 and Smat80 (symmetric matrix calculated at 80% clustering that performed best for database search) matrices having relative entropies of 1.18 and 1.16 bits, respectively were compared. The substituting pairs were sorted in the decreasing order of their lodscore values and their substitution preferences were studied as represented in **Figure 3.3**. Noticeably, the pattern of substitution was different for both the matrices and was more evident for the rows R, N, D, C, H and K. Next, due to similar scaling of Smat80 and BLOSUM90 matrices, the lodscore values were directly compared as a difference (**Figure 3.4**), where the lower half-diagonal represents the BLOSUM90 lodscore values and the upper half-diagonal represents the difference in the BLOSUM90 and Smat80 matrix values. The large number of positive values in the difference matrix (upper half-diagonal of **Figure 3.4**) shows that most of the Smat80 values are less than the corresponding values of BLOSUM90. More than half of the values along row 'C' which are negative, e.g. CS, CT, CA, etc. imply to the Cysteine substitutions that are more frequent in Smat80. Few of the W substitutions like WN, WH and WF are also more frequent in Smat80 matrix. To understand the difference between PfFSmat60 (*Plasmodium falciparum* fixed scaled matrix calculated at 60% clustering that performed best for pairwise alignments) and the BLOSUM50 matrix (popular matrix with FASTA); the odd-ratios calculated from their respective observation frequencies were compared. Since the scaling of PfFSmat60 and BLOSUM50 matrix was not similar, lodscores

could not be quantitatively compared. A similarity index was computed as a ratio of the odd-ratio of PfFSmat60 and the odd-ratio of BLOSUM50 matrix (Sutormin, Rakhmaninova, and Gelfand 2003). The values obtained are tabulated in **Table 3.7**, where the vertically displayed amino acids are those substituted in *Plasmodium falciparum*. The low ratios obtained indicate that most of the BLOSUM50 values are more than that of PfFSmat60, possibly an over-representation of the substitutions with respect to a biased genome like *Plasmodium falciparum*, except for CA and SA (row versus column), the ratios of which are greater than one. Among the least frequent substitutions, W is rarely substituted for E and P; and G for I, W and Y as compared to BLOSUM50. However, the disparity could have been better understood if PfFSmat60 was compared to a non-symmetric matrix, which is not the case with BLOSUM50.

.

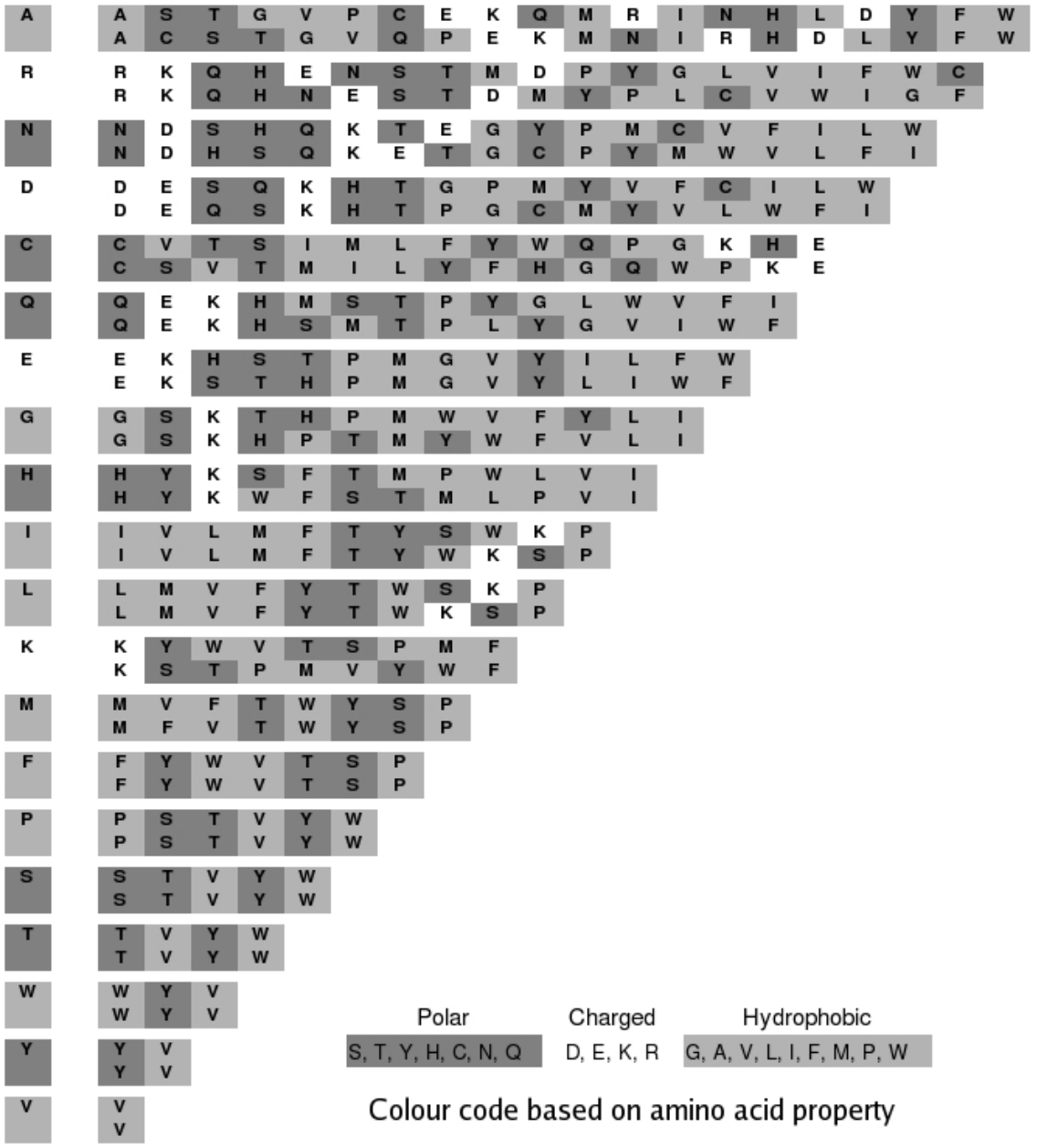


**Figure 3.3 A half-diagonal representation of BLOSUM90 and Smat80 matrices showing differences in their substitution preferences.** The boxes in the left represent each of the 20 amino acids pairing with all other amino acids (adjacent rows) in the BLOSUM90 (odd rows of alphabets) and Smat80 (even rows of amino acid alphabets) matrices compared here. The amino acid pairs are arranged in a decreasing order of their lodscore values i.e most preferred followed by the least preferred. Colour code is used to represent different classes of amino acids i.e. dark grey fill for polar amino acids, no fill for the charged amino acids and light grey fill for the hydrophobic amino acids.

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -2 | -2 | 0 | -2 | -1 | -2 | -1 | -1 | 0 | -2 | -1 | 0 | -1 | -1 | 0 | -2 | 0 | -2 | 0 | C |
| | | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | S |
| C | 9 | | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 2 | 1 | T |
| S | -2 | 5 | | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | P |
| T | -2 | 1 | 6 | | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | A |
| P | -4 | -2 | -2 | 8 | | -1 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | G |
| A | -1 | 1 | 0 | -1 | 5 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | -1 | N |
| G | -4 | -1 | -3 | -3 | 0 | 6 | | 0 | -1 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 2 | 1 | 0 | D |
| N | -4 | 0 | 0 | -3 | -2 | -1 | 7 | | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | E |
| D | -5 | -1 | -2 | -3 | -3 | -2 | 1 | 7 | | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | Q |
| E | -6 | -1 | -1 | -2 | -1 | -3 | -1 | 1 | 6 | | -1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | -2 | H |
| Q | -4 | -1 | -1 | -2 | -1 | -3 | 0 | -1 | 2 | 7 | | -1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | R |
| H | -5 | -2 | -2 | -3 | -2 | -3 | 0 | -2 | -1 | 1 | 8 | | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | K |
| R | -5 | -1 | -2 | -3 | -2 | -3 | -1 | -3 | -1 | 1 | 0 | 6 | | -1 | 0 | 0 | 0 | -1 | 0 | 0 | M |
| K | -4 | -1 | -1 | -2 | -1 | -2 | 0 | -1 | 0 | 1 | -1 | 2 | 6 | | 0 | 0 | 0 | 0 | 1 | 0 | I |
| M | -2 | -2 | -1 | -3 | -2 | -4 | -3 | -4 | -3 | 0 | -3 | -2 | -2 | 7 | | 0 | 0 | 0 | 0 | 0 | L |
| I | -2 | -3 | -1 | -4 | -2 | -5 | -4 | -5 | -4 | -4 | -4 | -4 | -4 | 1 | 5 | | 0 | 0 | 0 | 1 | V |
| L | -2 | -3 | -2 | -4 | -2 | -5 | -4 | -5 | -4 | -3 | -4 | -3 | -3 | 2 | 1 | 5 | | 0 | 0 | -1 | F |
| V | -2 | -2 | -1 | -3 | -1 | -5 | -4 | -5 | -3 | -3 | -4 | -3 | -3 | 0 | 3 | 0 | 5 | | 0 | 0 | Y |
| F | -3 | -3 | -3 | -4 | -3 | -5 | -4 | -5 | -5 | -4 | -2 | -4 | -4 | -1 | -1 | 0 | -2 | 7 | | -1 | W |
| Y | -4 | -3 | -2 | -4 | -3 | -5 | -3 | -4 | -4 | -3 | 1 | -3 | -3 | -2 | -2 | -2 | -3 | 3 | 8 | | |
| W | -4 | -4 | -4 | -5 | -4 | -4 | -5 | -6 | -5 | -3 | -3 | -4 | -5 | -2 | -4 | -3 | -3 | 0 | 2 | 11 | |
| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | |

**Figure 3.4 The difference in the lodscore values of BLOSUM90 and Smat80 matrices**. The lower half-diagonal represents the BLOSUM90 lodscore values and the upper half-diagonal shows the difference in the lodscore values of BLOSUM90 & Smat80 (i.e. BLOSUM90-Smat80). The half-bit values have been considered here. Comparison is made with BLOSUM90 as the entropy and scaling is similar to Smat80.

**Table 3.7 Similarity index of PfFSmat60 and Blosum50 matrices (comparison of odd ratios)**

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | **1.661** | 0.218 | 0.158 | 0.181 | 0.788 | 0.241 | 0.175 | 0.469 | 0.114 | 0.181 | 0.230 | 0.148 | 0.315 | 0.208 | 0.463 | 0.459 | 0.403 | 0.223 | 0.140 | 0.460 |
| R | 0.320 | **1.495** | 0.207 | 0.183 | 0.255 | 0.338 | 0.237 | 0.189 | 0.292 | 0.155 | 0.212 | 0.288 | 0.274 | 0.137 | 0.245 | 0.292 | 0.310 | 0.123 | 0.122 | 0.168 |
| N | 0.607 | 0.666 | **0.659** | 0.609 | 0.318 | 0.723 | 0.557 | 0.683 | 0.473 | 0.160 | 0.404 | 0.460 | 0.356 | 0.366 | 0.518 | 0.529 | 0.484 | 0.435 | 0.286 | 0.419 |
| D | 0.430 | 0.370 | 0.292 | **0.850** | 0.253 | 0.525 | 0.624 | 0.327 | 0.345 | 0.128 | 0.224 | 0.282 | 0.308 | 0.255 | 0.421 | 0.429 | 0.326 | 0.242 | 0.190 | 0.231 |
| C | 1.917 | 0.572 | 0.517 | 0.502 | **0.725** | 0.303 | 0.412 | 0.559 | 0.558 | 0.678 | 0.670 | 0.196 | 0.717 | 0.417 | 0.407 | 0.970 | 0.748 | 0.567 | 0.495 | 0.961 |
| Q | 0.598 | 0.535 | 0.390 | 0.464 | 0.252 | **1.472** | 0.516 | 0.353 | 0.424 | 0.243 | 0.291 | 0.376 | 0.416 | 0.267 | 0.423 | 0.367 | 0.491 | 0.179 | 0.154 | 0.336 |
| E | 0.604 | 0.394 | 0.304 | 0.551 | 0.268 | 0.549 | **0.916** | 0.307 | 0.339 | 0.203 | 0.271 | 0.313 | 0.380 | 0.203 | 0.392 | 0.393 | 0.345 | 0.144 | 0.137 | 0.346 |
| G | 0.507 | 0.220 | 0.169 | 0.197 | 0.302 | 0.149 | 0.214 | **0.962** | 0.168 | 0.089 | 0.114 | 0.133 | 0.188 | 0.118 | 0.263 | 0.265 | 0.165 | 0.094 | 0.077 | 0.156 |
| H | 0.576 | 0.485 | 0.241 | 0.444 | 0.275 | 0.562 | 0.371 | 0.342 | **1.068** | 0.149 | 0.269 | 0.250 | 0.286 | 0.279 | 0.360 | 0.365 | 0.378 | 0.490 | 0.302 | 0.380 |
| I | 0.559 | 0.500 | 0.207 | 0.241 | 0.522 | 0.502 | 0.328 | 0.185 | 0.255 | **0.668** | 0.516 | 0.316 | 0.745 | 0.335 | 0.366 | 0.258 | 0.403 | 0.440 | 0.191 | 0.782 |
| L | 0.501 | 0.413 | 0.230 | 0.236 | 0.484 | 0.471 | 0.308 | 0.151 | 0.226 | 0.345 | **0.915** | 0.308 | 0.738 | 0.387 | 0.409 | 0.261 | 0.297 | 0.265 | 0.185 | 0.484 |
| K | 0.638 | 0.820 | 0.370 | 0.426 | 0.350 | 0.633 | 0.473 | 0.399 | 0.419 | 0.270 | 0.357 | **0.700** | 0.378 | 0.236 | 0.547 | 0.465 | 0.470 | 0.361 | 0.187 | 0.455 |
| M | 0.388 | 0.295 | 0.202 | 0.232 | 0.341 | 0.311 | 0.371 | 0.115 | 0.162 | 0.417 | 0.530 | 0.261 | **1.908** | 0.415 | 0.182 | 0.252 | 0.367 | 0.225 | 0.181 | 0.514 |
| F | 0.478 | 0.280 | 0.189 | 0.218 | 0.425 | 0.318 | 0.207 | 0.247 | 0.384 | 0.286 | 0.496 | 0.173 | 0.669 | **0.907** | 0.386 | 0.268 | 0.277 | 0.715 | 0.444 | 0.399 |
| P | 0.574 | 0.362 | 0.192 | 0.257 | 0.243 | 0.281 | 0.186 | 0.165 | 0.211 | 0.143 | 0.203 | 0.180 | 0.200 | 0.141 | **1.048** | 0.387 | 0.229 | 0.195 | 0.102 | 0.325 |
| S | 1.038 | 0.381 | 0.316 | 0.440 | 0.547 | 0.403 | 0.384 | 0.455 | 0.400 | 0.180 | 0.250 | 0.270 | 0.487 | 0.212 | 0.604 | **1.072** | 0.617 | 0.474 | 0.153 | 0.312 |
| T | 0.655 | 0.450 | 0.264 | 0.340 | 0.426 | 0.525 | 0.313 | 0.290 | 0.301 | 0.206 | 0.243 | 0.252 | 0.394 | 0.224 | 0.450 | 0.584 | **1.297** | 0.148 | 0.168 | 0.384 |
| W | 0.223 | 0.429 | 0.307 | 0.172 | 0.141 | 0.179 | 0.079 | 0.140 | 0.445 | 0.162 | 0.169 | 0.054 | 0.490 | 0.546 | 0.072 | 0.194 | 0.178 | **1.330** | 0.343 | 0.232 |
| Y | 0.504 | 0.369 | 0.260 | 0.288 | 0.289 | 0.279 | 0.296 | 0.207 | 0.416 | 0.210 | 0.341 | 0.250 | 0.478 | 0.580 | 0.520 | 0.341 | 0.317 | 0.696 | **0.754** | 0.320 |
| V | 0.637 | 0.352 | 0.197 | 0.212 | 0.579 | 0.372 | 0.313 | 0.220 | 0.380 | 0.400 | 0.375 | 0.247 | 0.511 | 0.254 | 0.377 | 0.290 | 0.430 | 0.239 | 0.146 | **1.219** |

**Footnote**: Each element of the table is a ratio of the odd ratios of PfFSmat60 and Blosum50 substitution matrices, a measure of similarity of the two matrices. The text highlighted in bold are the diagonal elements.

### 3.3.3 Comparative performance of the PfSSM matrices

Qualitative differences were observed in the performance of the PfSSM series compared to standard matrices when tested for some protein alignments. Alignment examples are provided in the following sections. For comparison, the scores obtained with the best performing BLOSUM series has been reported along with the scores with similar entropy matrices, PAM2 (H=3.97) and BLOSUM90 (H=1.18) that have entropies close to PfFSmat60 (H=4.67) and Smat80 (H=1.16) matrices, respectively.

#### 3.3.3.1 Experimentally characterized *Plasmodium falciparum* cyclin-3

Recently, Cyclin-3 protein of *Plasmodium falciparum*, PFE0920c, was experimentally characterized by a research group (Merckx et al. 2003) that showed low sequence match to known cyclins. SSEARCH was performed for PFE0920c against the Uniprot/Swiss-Prot database with BLOSUM and PfSSM series. The scores for the first hit obtained with PfSSM series are given in **Table 3.8**. While

BLOSUM100 (the best performing BLOSUM matrix) gave a score of 113.3 bits at an E-value of 1.3e-24, Smat80 gave an alignment score equal to 117.8 bits for the same length of alignment overlap (124 residues) at an improved E-value of 6.1e-26. BLOSUM90 (H=1.18) matrix with similar entropy values to Smat80 (H=1.16) gave a score of 112.6 bits at an E-value of 2.1e-24 for a 128 amino acid overlap. The performance of Smat80 was thus better.

A pair-wise local alignment of PFE0920c and the first hit from SSEARCH gave the best alignment score equal to 239.8 bits with PfFSmat60 at an E-value of 4.4e-68 for a 190 amino acid overlap. This spanned the entire cyclin domain of both the sequences. The alignment score with the best performing standard matrix BLOSUM50 was 119.8 bits at an E-value of 5.6e-32 (128 amino acid overlap) that failed to span the entire domain region. The score was very poor with PAM2 (H=3.97) compared to PfFSmat60 (H=4.67) and was equal to 16.2 bits at an E-value of 0.57. **Table 3.9** shows the bit scores obtained for the PfSSM series with the FASTA local alignment program.

**Table 3.8 Alignment scores for PFE0920c protein of *P. falciparum* and yeast cyclin, P40186, with PfSSM series**

| Clustering % | Smat | SSmat | PfFmat | PfFSmat |
|---|---|---|---|---|
| 50 | 116.4 (124) | 74.4 (128) | 106.3 (124) | 37.0‡, 35.6 |
| 60 | 116.0 (124) | 75.2 (128) | 107.7 (124) | 36.7‡, 35.3 |
| 70 | 116.2 (124) | 75.3 (128) | 108.2* (124) | 37.7‡, 36.1 |
| 80 | 117.8* † (124) | 79.6 (128) | 103.2 (122) | 37.9 (190) |
| 90 | 114.2 (124) | 90.1* (128) | 98.9 (124) | 39.4* (190) |

The first column of the table represents the clustering % at which the matrices (represented in the subsequent columns along the first row) were calculated. Columns 2-5 represent the alignment scores obtained for each matrix series at different clustering percentages. The value in the closed bracket is the amino acid overlap for the alignment. The values represented with symbol '‡' are the cyclin family first hits obtained from a different organism and the values following it are the yeast cyclin hits which take up a second position. We did not mention the overlap lengths for these instances. The standard equivalent matrix values for each corresponding * marked cells from column 2-5 are 112.4 (128), 96.1 (124), 101.9 (124) and 54.7(55) respectively.

* The highest obtained scores for the respective columns
† The score for the best performing matrix series
‡ The scores of the first hit from a different organism

**Table 3.9 FASTA alignment scores for *P. falciparum* Cyclin, PFE0920c, and *yeast c*yclin protein, with PfSSM series**

| Clustering % | Smat | SSmat | PfFmat | PfFSmat |
|---|---|---|---|---|
| 50 | 101.1 (124) | 136.8 (128) | 98.7 (124) | 238.0 (190) |
| 60 | 98.7 (124) | 135.0 (128) | 100.5* (124) | 239.8* † (190) |
| 70 | 98.7 (124) | 135.0 (128) | 100.5* (124) | 238.0 (190) |
| 80 | 101.7* (124) | 137.9* (128) | 97.0 (122) | 226.3 (190) |
| 90 | 99.9 (124) | 136.2 (128) | 97.6 (124) | 221.6 (190) |

The first column of the table represents the clustering % at which the matrices (represented in the subsequent columns along the first row) were calculated. Columns 2-5 represent the alignment scores obtained for each matrix series at different clustering percentages. The values in the closed brackets are the amino acid overlap for the alignment. The standard equivalent matrix values for each corresponding * marked cells from column 2-5 are 99.3(128), 97(124), 94.1(124), 16.2 (11) respectively.

* The highest obtained scores for the respective columns
† The score for the best performing matrix series

### 3.3.3.2 A hypothetical protein, a probable DnaJ ortholog

The *Plasmodium falciparum* protein, PFB0090c, annotated as a hypothetical protein in PlasmoDB (http://www.plasmodb.org/plasmo/), has a weak DnaJ motif forming a part of the heat shock protein machinery (Date and Stoeckert 2006). HMMer, an implementation of profile HMM methods for database searches (Eddy 1998) gave DnaJ profile hits for the protein. PFB0090c was used as a query in the search against the Uniprot/Swiss-Prot database to look for best hits, probably a DnaJ protein. The bit scores obtained against the first hit (human DnaJ) with the PfSSM series is tabulated in **Table 3.10**. Smat80 (H=1.16) gave the highest score equal to 238.9 bits at an E-value of 4.3e-62 for a 348 amino acid overlap, while the best performing similar entropy matrix, BLOSUM90 (H=1.18) scored 234.7 bits for a 338 amino acid overlap (E=7.8e-61). Smat80 which was symmetric thus seemed to work best for database searches.

Pair-wise alignment of PFB0090c and human DnaJ (Q9UDY4), gave an alignment score of 810.7 bits (E=0; 343 residue overlap) with PfFSmat60. The alignment spanned the DnaJ domain regions of both the sequences. PAM2 (H=3.97) compared to PfFSmat60 (H=4.67) gave a score of 57.7 bits and E-value of 5.9e-13 for only an 83 amino acid overlap that failed to span the entire domain region. The best performing BLOSUM50 aligned these regions but gave a score and E-value worse than PfFSmat60 (score in bits 428.6 and E-value of 1.3e-124 for a 339-residue overlap). **Table 3.11** shows the FASTA results for the DnaJ protein (bit score values) with the PfSSM series. The non-symmetric, PfFSmat60 matrix thus

seemed to perform best for pair wise local alignments.

**Table 3.10 Alignment scores of the PFB0090C protein of *Plasmodium falciparum* and human DnaJ homolog, Q9UDY4, with PfSSM series of matrices**

| Clustering % | Smat | SSmat | PfFmat | PfFSmat |
|---|---|---|---|---|
| 50 | 235.8 (338) | 142.7 (339) | 199.6 (346) | 53.5 (343) |
| 60 | 235.9 (348) | 148.0 (339) | 203.5* (346) | 53.0 (343) |
| 70 | 236.8 (348) | 148.0 (339) | 201.4 (346) | 54.8 (343) |
| 80 | 238.9* † (348) | 151.9 (339) | 196.2 (337) | 57.7 (343) |
| 90 | 230.2 (348) | 179.4* (339) | 183.0 (346) | 62.5* (340) |

The first column of the table represents the clustering % at which the matrices (represented in the subsequent columns along the first row) were calculated. Columns 2-5 represent the alignment scores obtained for each matrix series at different clustering percentages. The values in the closed brackets are the amino acid overlap for the alignment.

* The highest obtained scores for the respective columns
† The score for the best performing matrix series

**Table 3.11 Alignment scores for the FASTA local pairwise alignment of *Plasmodium falciparum* hypothetical protein, PFB0090C and human DnaJ protein, Q9UDY4, with the PfSSM series of matrices**

| Clustering % | Smat | SSmat | PfFmat | PfFSmat |
|---|---|---|---|---|
| 50 | 428.6* (339) | 479.5 (339) | 239.1 (346) | 806.0 (343) |
| 60 | 288.2 (348) | 477.2 (339) | 243.7* (346) | 810.0* † (343) |
| 70 | 287.6 (348) | 476.6 (339) | 239.6 (346) | 806.0 (343) |
| 80 | 295.2 (348) | 484.2* (339) | 226.8 (346) | 792.6 (343) |
| 90 | 274.2 (348) | 463.2 (339) | 212.7 (346) | 775.6 (340) |

The first column of the table represents the clustering % at which the matrices (represented in the subsequent columns along the first row) were calculated. Columns 2-7 represent the alignment scores obtained for each matrix series at different clustering percentages. The values in the closed brackets are the amino acid overlap for the alignment.

* The highest obtained scores for the respective columns
† The score for the best performing matrix series

### 3.3.3.3 Other interesting pair-wise alignments

With the aim to assess the performance of PfFSmat60 in generating accurate alignments in the twilight zone of protein pairs (Rost 1999), some proteins that fell in this zone were analyzed. FASTA was used to generate the pair-wise alignments. The alignments were tested for, firstly, supposedly missing genes of *Plasmodium falciparum* metabolic pathway that had other lines of their existence and secondly, hypothetical proteins with a suspected clue of functional homology based on various bio-informatics' approaches, e.g. profile based-methods, functional clusters in a network, etc. These proteins otherwise gave poor alignments with likely orthologs when the standard matrices were used. An improvement was achieved in the number of aligned amino acids with PfFSmat60. Further, the authenticity of these alignments was proved in terms of the motifs, domains, and secondary structure elements that were initially identified for these proteins. It was noteworthy to find that PfFSmat60 could align domains and functional motifs with a good conservation. However, for such instances BLOSUM gave shorter alignments that mostly did not fall within these regions. The results from PfFSmat60 matrix have been compared with BLOSUM100 (the only highest entropy BLOSUM matrix comparable to PfFSmat60) BLOSUM50 (known to be the best performing BLOSUM matrix with FASTA programs) and PAM2 (matrix with entropy equivalent to PfFSmat60).

#### 3.3.3.3.1 Bi-functional enzyme of the shikimate pathway

The shikimate kinase pathway plays a vital role in the survival of Apicomplexans. Moreover, the absence of this pathway in mammals makes it an attractive target for the development of anti-parasitic drugs. The first six enzymes of the shikimate pathway were missing in the initial genome annotation of *Plasmodium falciparum*. However, the presence of these enzyme activities were detected in the crude extracts of the parasite (McConkey et al. 2004). Recently, a hypothetical protein, PFB0280w was identified, presumably a bi-functional protein with EPSP (5-enolpyruvylshikimate-3-phosphate) and SK (shikimate kinase) activities (fifth and sixth enzymes of this pathway). The predictions were based on a bioinformatics' approach with a moderate level of confidence (McConkey et al. 2004). This protein was used to test the sensitivity of our matrix. A local alignment was generated for PFB0280w and the yeast AROM complex, P08566; known to have the EPSP and SK activities. The signature motifs for EPSP and SK were obtained for P08566, identical to the PROSITE pattern and for PFB0280w with a mismatch of 2 and 5, respectively. Motifs were detected with fuzzpro, a program from the EMBOSS package (Rice, Longden, and Bleasby 2000). While BLOSUM100 gave an insignificant alignment (23.4 bits score at an E-value of 0.31 for a 95 amino acid overlap), BLOSUM50 aligned only the EPSP motif region of these proteins (60.8 bit score at an E-value of 2e-12 for a 131 amino acid overlap). On the other hand, PfFSmat60 successfully extended the alignment spanning the SK motif of the yeast AROM complex. The region of overlap was the probable SK motif of *Plasmodium falciparum* obtained

with a score equal to 1065.5 bits and E-value of 0, for a 1781 amino acid overlap (**Figure 3.5**). The equivalent PAM2 matrix gave an insignificant alignment with a score of 12.2 bits and E-value of 1.0 for an overlap of 5 amino acids. It is worth mentioning that though an alignment overlap was achieved for what was supposed as the SK motif, gaps were observed in the motif alignment. A multiple sequence alignment of yeast AROM complex and the well-characterized SK's from other organisms having known crystal structures reveal that gaps are not uncommon in the SK motif.

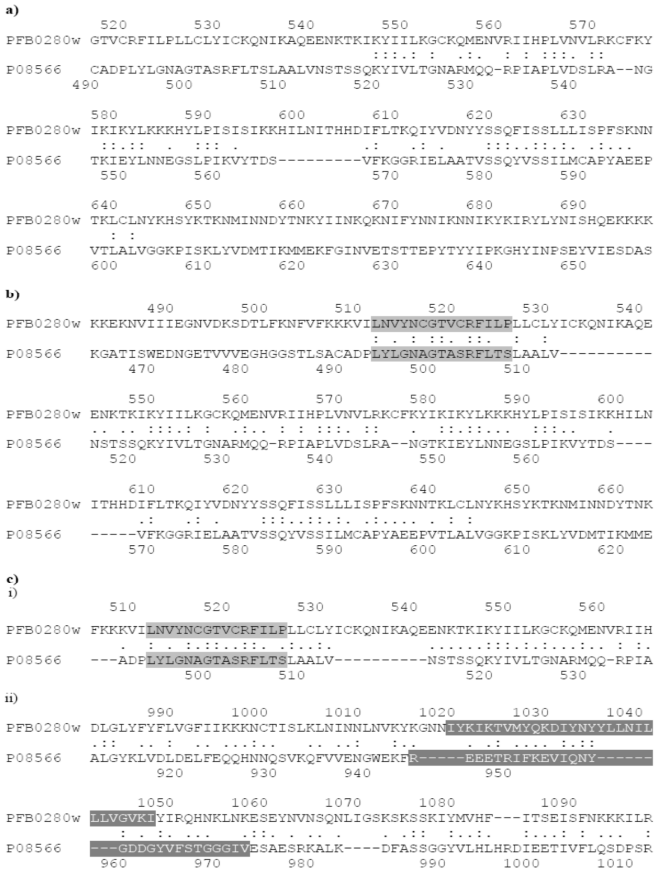


**Figure 3.5 Alignment extensions with PfFSmat60 for *P. falciparum* bi-functional enzyme of the shikimate pathway** The sequences compared here are the *P. falciparum* hypothetical protein, PFB0280w and yeast multifunctional protein, P08566. (a) The alignment yielded with BLOSUM100, showing no alignment overlap for the motif regions, EPSP synthase I and shikimate kinase. (b) The alignment with BLOSUM50 showing the aligned motif regions for only EPSP synthase I motif (gray shading). (c) The alignment extended by PfFSmat60 for both the EPSP synthase I and shikimate kinase motifs represented as (i) and (ii), respectively. The text shaded in gray corresponds to the EPSP synthase I motif. The text in white with dark gray shading represents the shikimate kinase motif. PFB0280w gave only hypothetical hits with BLASTp at the default parameters.

#### 3.3.3.3.2 A missing metabolic enzyme - TPK

In an attempt to reconstruct the malaria metabolic pathway for *Plasmodium falciparum*, Limviphuvadh (Limviphuvadh 2003) attempted to identify some missing enzymes from the parasites genome, using the virtual enzyme system and KEGG ortholog clusters. One of them was a hypothetical protein, PFI1195c, which formed an ortholog cluster with proteins annotated as thiamine pyrophosphokinase (TPK). The pair-wise alignment of PFI1195c and TPK of S. cerevisiae (P35202) was analyzed with PfFSmat60. A possible TPK motif was obtained for PFI1195c, at position 97-116 with a mismatch 3 and for P35202 with a mismatch of 4, spanning the residues 52-71. HMMer analysis identified PFI1195c as a likely TPK, agreeing with the article. Surprisingly, while neither BLOSUM50 (bit score: 75.1; E-value: 3.1e-18 for a 293 amino acid overlap) nor BLOSUM100 (bit score: 40.6; E-value: 7.6e-08 for a 42 amino acid overlap) gave meaningful alignments within the motif region, PfFSmat60 gave a biologically significant and lengthier alignment, spanning the TPK motif. The bit score achieved was 250.6 at an E-value of 4.4e-71 (**Figure 3.6**). On the other hand, PAM2 aligned one fourth of the motif region with a bit score of 32.4 and E-value equal to 2.2e-05 for a 11 amino acid overlap. High entropy matrices are known to be better at detecting short regions of strong similarity (Altschul 1991). This might be a possible explanation for the alignments achieved in the motif regions of protein with PfFSmat60 matrix which has relatively high entropy.


**Figure 3.6 Alignment of PFI1195c protein of *P. falciparum* with the yeast TPK protein, P35202.** (a) Alignment with BLOSUM100 showing an insignificant overlap. (b) A portion of the alignment with BLOSUM50 showing unaligned motif regions (c) Alignment with PfFSmat60 that has successfully aligned the predicted motif regions. The text shaded in gray indicates the motif regions of both the sequences.

#### 3.3.3.3.3 Asparagine synthetase

The compositional adjustment substitution matrix of Yu and coworkers (Yu, Wootton, and Altschul 2003) was shown to significantly increase the bit score and alignment length of proteins. The sequences tested, were the putative asparagine synthetase from *Plasmodium falciparum* and the PurF protein from *M. tuberculosis* that share a common domain, glutamine amidotransferase (GATase). A comparison of PfSSM matrices with the new adjusted matrix of Yu and coworkers was desired but a direct comparison could not be made since a modified matrix was not available for standalone programs, in this case. An attempt was thus made to compare the results indirectly by aligning the same sequence pair with the test matrix and study its performance. It was interesting to find that, in spite of the differences in the AT content of these genomes; good consistency was achieved with PfFSmat60 in terms of the secondary structure elements of the proteins compared. The results were similar to the compositionally adjusted substitution matrix of this group (**Figure 3.7**), though the improvement in score and alignment length achieved in this case was much better over the standard (369.6 bits at an E-value of 1.7e-106 for a 539 amino acid overlap). The comparable PAM2 matrix gave a poor score of 10.9 bits at a high E-value of 1 for only a seven amino acid overlap while the best performing BLOSUM matrix (BLOSUM50) gave a bit score of 42.5 at an E-value of 5.1e-08 (78 amino acid overlap). The secondary structure elements were obtained by comparing known crystal structures of the ortholog proteins, predicted using FUGUE server (Shi, Blundell, and Mizuguchi 2001). This is

consistent with our earlier observation for the GntR family of proteins which show secondary structure conservation despite high degree of sequence variation in the effector binding region of these proteins (Vindal, Ranjan, and Ranjan 2007; Vindal, Suma, and Ranjan 2007).

**a)**

```
            110       120       130       140       150       160
gi|16805184| VIPNLFKIYKEKIPSMLDGIFAGVISDKKNNTFFAFRDPIGICPLYIGYAADGSIWFSSE
                                              ::..: ::: :. :: .:
gi|15607948| HGAADSTLEQAALDLLPTVRGAFCLTFMDENTLYACRDPYGVRPLSLGRLDRGWVVASET
              180       190       200       210       220       230
```

**b)**

```
              100       110       120       130       140       150
gi|16805184| NLNKLKSCSDCAVIPNLFKIYKEKIPSMLDGIFAGVISDKKNNTFFAFRDPIGICPLYIG
                                              : :  .. :.  ::..: ::: :. :: .:
gi|15607948| DSDILGALLAHGAADSTLEQAALDLLPTVRGAFCLTFMDE--NTLYACRDPYGVRPLSLG
              170       180       190       200         210       220

              160       170       180        190       200       210
gi|16805184| YAADGSIWFSSEFKALKDNCIRYVI-FPPGHYYKNNKNKGEFVRYYNPNWWSLNNSIPNN
                 : .  .::  ::      .:  . ::.    . .  . .:. ::
gi|15607948| RLDRGWV-VASETAALDIVGASFVRDIEPGELLAIDADGVRSTRFANPTPKGCVFEYVYL
               230       240       250       260       270       280
```


**Figure 3.7 Alignment extensions with PfFSmat60 for *P. falciparum* putative asparagine synthetase and *M. tuberculosis* PurF protein.** The sequences compared are *P. falciparum* putative asparagine synthetase (NCBI gi 16805184) (top line) and *M. tuberculosis* PurF protein (NCBI gi 15607948). a) Alignment with BLOSUM100 with a bit score of 23.2. b) The alignment with BLOSUM50 with a bit score of 42.5. c) The alignment region with PfFSmat60 where the structural elements highlighted corresponds closely to the three-dimensional structural superposition of the common domain shared between PurF and asparagine synthase families. The bit score was equal to 369.6 at an E-value of 1.7e-106. The known crystal structures of *E. coli* asparagine synthetase B (1CT9, chainA) and *B. subtilis* PurF protein (1GPH, chain 3) were used to assign the secondary structure elements. The beta sheets are shaded grey and the alpha elements are shown in white text and a black background.

#### 3.3.3.3.4 Fructose-bisphosphate aldolase

The glycolytic pathway is the major source of energy (ATP) for the blood stage of *Plasmodium falciparum* due to the absence of a functional TCA cycle in the parasite. Aldolase is a key enzyme of the glycolytic pathway and is an attractive drug target due to its high degree of sequence diversity and structural differences (Kim et al. 1998) as compared to the human counterpart. Aldolase activity was observed in a *Plasmodium falciparum* protein, with protective properties as early as 1988 (Certa et al. 1988) and drug targets were defined by in-vitro mutagenesis studies (Certa, Itin, and Dobeli 1992). Fructose bisphosphate aldolase protein of *Plasmodium falciparum* was expressed in its active form by Knapp and coworkers (Knapp, Hundt, and Kupper 1990) while Dobeli and coworkers had expressed this enzyme in its tetrameric form and suggested that the *Plasmodium falciparum* aldolase can associate with the cytoskeleton of the parasite or of the host. (Dobeli et al. 1990). Later, aldolase was shown to provide an unusual binding site for thrombospondin-related anonymous protein in the invasion machinery of the malaria parasite (Bosch et al. 2007).

An alignment of fructose-bisphosphate aldolase from AT-rich genomes of *Plasmodium falciparum* (P14223) and *Fusobacterium nucleatum* were compared. Both these genomes tend to favour amino acids with AT-rich codon sets and hence a comparison with the standard and PfFSmat60 matrices should have made a difference. **Figure 3.8** shows the alignment of aldolases from both these organisms with the PfFSmat60 and standard matrices. The alignment with PfFSmat60

extended left and right of the alignment obtained with the standard BLOSUM50 matrix, with a good conservation of residues. The alignment length (325aa) improved 2 fold with an increase in the percentage of similarity and identity compared to BLOSUM50 at an improved E-value of 2.7e-78 (E = 5.2e-14 with BLOSUM50). It should be noted here that though BLOSUM50 was not equivalent to the PfFSmat60 matrix in terms of entropy, the comparison has been made to show that PfFSmat60 performed better than the best performing standard matrix i.e. BLOSUM50. With the equivalent PAM2 matrix, an overlap of only 8 amino acids at an E value of 0.43 was achieved while with BLOSUM100 there was an overlap of 40 amino acids at an E value of 0.071.

**a)**

```
>>gi|19703667|ref|NP_603229.1| fructose-bisphosphate ald  (295 aa)
 initn:  76 init1:  50 opt:  50  Z-score: 50.0  bits: 17.6 E(): 0.43
Smith-Waterman score: 50;  87.500% identity (87.500% similar) in 8 aa overlap (81-88:69-76)

               60        70        80        90       100       110
P14223 DNIKLENTIENRASYRDLLFGTKGLGKFISGAILFEETLFQKNEAGVPMVNLLHNENIIP
                                      :::::: :
gi|197 YSNDKEMFDLIHKMRTRIIKSPAFNESKILGAILFEQTMDSKIDGKYTADFLWEEKKVLP
       40        50        60        70        80        90

              120       130       140       150       160       170
P14223 GIKVDKGLVNIPCTDEEKSTQGLDGLAERCKEYYKAGARFAKWRTVLVIDTAKGKPTDLS

gi|197 FLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFGTKMRSVIKKASPAGIARVVEQ
      100       110       120       130       140       150
```

**b)**

```
>>gi|19703667|ref|NP_603229.1| fructose-bisphosphate ald  (295 aa)
 initn: 153 init1:  90 opt: 124  Z-score: 284.0  bits: 60.9 E(): 5.2e-14
Smith-Waterman score: 124;  26.708% identity (59.006% similar) in 161 aa overlap (79-235:67-216)

       50        60        70        80        90       100
P14223 RFDNIKLENTIENRASYRDLLFGTKGLGKFISGAILFEETLFQKNEAGVPMVNLLHNENI
                                     : ::::::.:. .: ..       : .....
gi|197 NEYSNDKEMFDLIHKMRTRIIKSPAFNESKILGAILFEQTMDSKIDGKYTADFLWEEKKV
         40        50        60        70        80        90

      110       120         130       140       150       160
P14223 IPGIKVDKGLVNIPCTDEE--KSTQGLDGLAERCKEYYKAGARFAKWRTVLVIDTAKGKP
       .: .:.:::: ..     .  : .  :  : .: .: .  :.   : :.:.     :..:
gi|197 LPFLKIDKGLNDLDADGVQTMKPNPTLADLLKRANERHIFGT---KMRSVI----KKASP
        100       110       120       130          140

        170       180       190       200       210       220
P14223 TDLS-IHETAWGLARYASICQQNRLVPIVEPEILADGPHSIEVCAVVTQKVLSCVFKALQ
       . .. . :  . .:  :..   . :.::.:::.  ..  ... :  . .  .   ..::
gi|197 AGIARVVEQQFEVA--AQVVAAG-LIPIIEPEVDINNVDKVQ-CEEILRDEIRKHLNALP
     150       160         170       180       190       200

          230       240       250       260       270       280
P14223 E-NGVLLEGALLKPNMVTAGYECTAKTTTQDVGFLTVRTLRRTVPPALPGVVFLSGGQSE
       : ..:.:. .:
gi|197 ETSNVMLKLTLPTVENLYEEFTKHPRVVRVVALSGGYSREKANDILSKNKGVIASFSRAL
         210       220       230       240       250       260
```

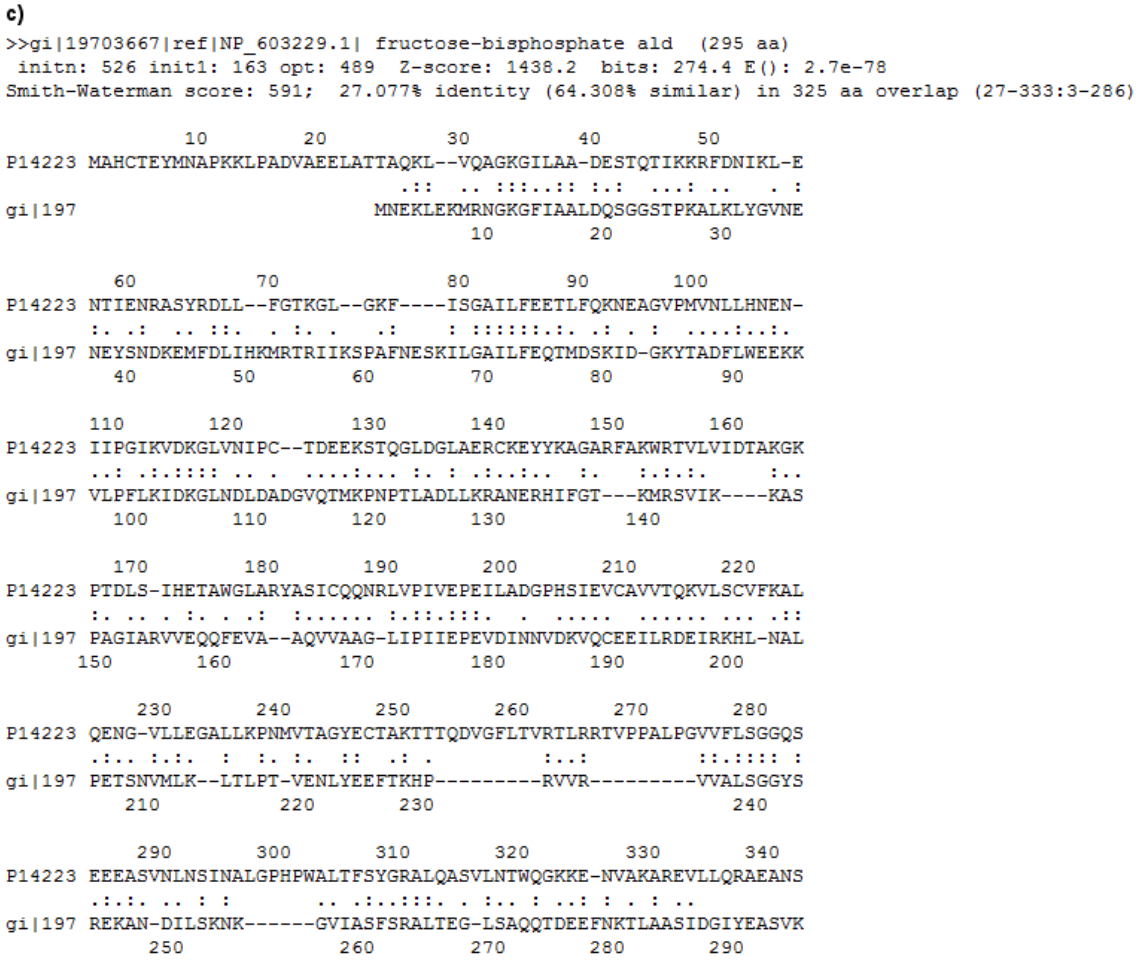


**Figure 3.8 Alignment of fructose-bisphosphate aldolase from AT-rich genomes of *P. falciparum* (P14223) and *Fusobacterium nucleatum*** a) Alignment with PAM2 matrix at a gap opening and extension penalty of -12, -2 respectively b) Alignment with BLOSUM50 at similar parameters c) Alignment with PfFSmat60 (at similar parameters) showing an extension, left and right of the alignment obtained with the standard matrices, with a good conservation of residues. The alignment length improved 2 fold compared to BLOSUM50 with an improvement in similarity and identity and a 40 fold increase in alignment length with respect to PAM2.

#### 3.3.3.3.5 Peroxiredoxin – putative

Cellular redox metabolism is considered to be involved in the pathophysiology of diseases caused by protozoan parasites such as *Toxoplasma, Trypanosoma, Leishmania*, and *Plasmodia* (Deponte 2007). Redox reactions furthermore are thought to play a major role in the action of and the resistance to some clinically used antiparasitic drugs (Deponte 2007). Interestingly, in the absence of catalase and glutathione peroxidase, the parasites rely primarily on peroxiredoxin-linked systems for protection (Boucher et al. 2006b). The role of 1-Cys peroxiredoxin in haem detoxification has been shown earlier in *Plasmodium falciparum* (Kawazu et al. 2005).

Here an alignment of the *Plasmodium falciparum* peroxiredoxin (1-Cys peroxiredoxin) and the yeast homologue was performed with the PfFSmat60 and standard matrices. As is evident from **Figure 3.9**, the alignment with PfFSmat60 extended left and right of the alignment obtained with the standard matrices. The alignment overlap with PfFSmat60 was 1.5 times more than that obtained with BLOSUM50 at an improved E-value of 1.4e-86 (E-value with BLOSUM50 was 2.8e-31). With the equivalent PAM2 matrix only a 6 amino acid overlap was obtained at a poor E-value of 1 while with the BLOSUM100 matrix there was an overlap of 81 amino acids at an E-value of 6.8e-15.

**a)**

```
>>P38013|AHP1_YEAST Peroxiredoxin type-2 - Saccharomyces  (176 aa)
 initn:  37 init1:  37 opt:  37  Z-score:  8.9  bits: 8.6 E():    1
Smith-Waterman score: 37;  83.333% identity (83.333% similar) in 6 aa overlap (112-117:57-62)

              90       100       110       120       130       140
MAL7P1 DTDGSPNDFTSIDTHELFNNKKILLISLPGAFTPTCSTKMIPGYEEEYDYFIKENNFDDI
                                       :: :::
P38013 DSESCKMPQTVEWSKLISENKKVIITGAPAAFSPTCTVSHIPGYINYLDELVKEKEVDQV
         30        40        50        60        70        80

             150       160       170       180       190       200
MAL7P1 YCITNNDIYVLKSWFKSMDIKKIKYISDGNSSFTESMNMLVDKSNFFMGMRPWRFVAIVE

P38013 IVVTVDNPFANQAWAKSLGVKDTTHIKFASDPGCAFTKSIGFELAVGDGVYWSGRWAMVV
         90       100       110       120       130       140
```

**b)**

```
>>P38013|AHP1_YEAST Peroxiredoxin type-2 - Saccharomyces  (176 aa)
 initn: 213 init1: 183 opt: 222  Z-score: 593.9  bits: 116.8 E(): 2.8e-31
Smith-Waterman score: 222;  30.081% identity (67.480% similar) in 123 aa overlap (101-218:46-164)

              80        90       100       110       120       130
MAL7P1 MIDVRNMNNISDTDGSPNDFTSIDTHELFNNKKILLISLPGAFTPTCSTKMIPGYEEEYD
                                     :::... . :.::.:::... :::: .  :
P38013 FQYIAISQSDADSESCKMPQTVEWSKLISENKKVIITGAPAAFSPTCTVSHIPGYINYLD
          20        30        40        50        60        70

             140       150       160          170         180
MAL7P1 YFIKENNFDDIYCITNNDIYVLKSWFKSMDIK---KIKYISDGNSSFTESMN--MLVDKS
        ..::.. :..  .: .. .. ..: ::. .:   .::. :: . .::.:..  . :  .
P38013 ELVKEKEVDQVIVVTVDNPFANQAWAKSLGVKDTTHIKFASDPGCAFTKSIGFELAVGDG
          80        90       100       110       120       130

         190       200       210       220       230       240
MAL7P1 NFFMGMRPWRFVAIVENNILVKMFQEKDKQHNIQTDPYDISTVNNVKEFLKNNQL
        .. :    :.. .:::.:..   .: .   ..
P38013 VYWSG----RWAMVVENGIVTYAAKETNPGTDVTVSSVESVLAHL
          140           150       160       170
```

```
c)
>>P38013|AHP1_YEAST Peroxiredoxin type-2 - Saccharomyces  (176 aa)
 initn: 518 init1: 315 opt: 536  Z-score: 1586.9  bits: 300.6 E(): 1.4e-86
Smith-Waterman score: 536;  28.261% identity (73.913% similar) in 184 aa overlap (62-235:2-176)

             40        50        60        70         80        90
MAL7P1 IVSKRGNSKNRFSQKVYESKNIDLENDIKENDLIPNVKVMIDVR-NMNNISDTDGSPNDF
                                  .::. .   . : . .  .::..:.....
P38013                          MSDLVNKKFPAGDYKFQYIAISQSDADSESC
                                        10        20        30

                 100       110       120       130       140
MAL7P1 TSIDTHE----LFNNKKILLISLPGAFTPTCSTKMIPGYEEEYDYFIKENNFDDIYCITN
       .. .: :    . .:::.....:.::.:::... :::: . .: ..::...:.. ..:
P38013 KMPQTVEWSKLISENKKVIITGAPAAFSPTCTVSHIPGYINYLDELVKEKEVDQVIVVTV
              40        50        60        70        80        90

        150       160          170         180       190       200
MAL7P1 NDIYVLKSWFKSMDIKK---IKYISDGNSSFTESMN--MLVDKSNFFMGMRPWRFVAIVE
       .. .. ..:.::. .:.   ::..:: ...::.:..  ..: .. .. :    :....::
P38013 DNPFANQAWAKSLGVKDTTHIKFASDPGCAFTKSIGFELAVGDGVYWSG----RWAMVVE
             100       110       120       130       140

             210       220       230       240
MAL7P1 NNILVKMFQEKDKQHNIQTDPYDISTVNNVKEFLKNNQL
       :.:..     ..:..:  :: ...:.:..: ..:
P38013 NGIVTY----AAKETNPGTD-VTVSSVESVLAHL
       150           160        170
```

**Figure 3.9 Alignment of a putative peroxiredoxin of *P. falciparum* (MAL7P1.159) with the yeast peroxiredoxin** a) Alignment with PAM2 matrix at a gap opening and extension penalty of -12, -2 respectively b) Alignment with BLOSUM50 at similar parameters c) Alignment with PfFSmat60 (at similar parameters) showing an extension left and right of the alignment obtained with the standard. The alignment overlap improved 1.5 times over BLOSUM50 and there was a 30 fold increase in alignment length compared to PAM2.

#### 3.3.3.3.6 Fe-superoxidase dismutase

The intra-erythrocytic malaria parasite is under constant oxidative stress due to both endogenous and exogenous processes caused by the chemical progeny of exogenous reactive, oxygen (ROS) and nitrogen species (RNS) produced by the immune system of the host and the endogenous ROS generated during the digestion of host cell haemoglobin and concomitant biochemical reactions (Bozdech and Ginsburg 2004). Superoxide dismutases (SODs) are regarded as one of the defending enzymes against oxygen toxicity and they may be expected to have special significance in *Plasmodium falciparum*, since part of the parasite life cycle is spent in red blood cells where the formation of reactive oxygen species is likely to be promoted by the products of haemoglobin breakdown (Boucher et al. 2006a; Boucher et al. 2006b). Haemoglobin degradation produces free haem groups leading to oxidation of the iron from the ferrous ($Fe^{2+}$) to the ferric ($Fe^{3+}$) state. This oxidation liberates electrons, which promote the formation of reactive oxygen intermediates, including superoxide.

It was earlier thought that malaria parasites had no requirement for an endogenous superoxide dismutase and merely exploited the activity of the host's enzyme in the red blood cell (Fairfield, Meshnick, and Eaton 1983). However, in 1996 a *Plasmodium falciparum* iron-dependent SOD (*Pf*FeSOD) was identified in parasites that were isolated from infected blood cells (Becuwe et al. 1996). The SOD gene is known to express at its highest level during the intra-erythrocytic stage of the parasite life cycle.

In this context a pairwise alignment was performed at a gap opening and extension penalty of -12, -2 respectively for a typical *Plasmodium falciparum* FeSOD that was earlier identified and biochemically characterized (Becuwe et al. 1996) with an ortholog from the GC-rich *Xanthomonas campestris* genome. As evident from **Figure 3.10,** the alignment of SOD's shows an overall improvement in E-value (6.3e-155) and alignment length (205 aa) with the PfFSmat60 matrix (**Figure 3.10c**). Compared to BLOSUM50 alignment (**Figure 3.10b**) an improvement in identity (additional identities marked in yellow in **Figure 3.10c**), similarity and an alignment extension of 2 residues is observed with PfFSmat60 matrix. While the PAM2 matrix gave an alignment overlap of 8 aa at an E-value of 0.0029, BLOSUM100 gave an alignment overlap of 194 amino acids at an E-value of 2.8e-56. On the other hand the best performing standard matrix, BLOSUM50 gave an alignment overlap of 203 residues at an E-value of 2.1e-86.

**a)**

```
>>B0RS18|B0RS18_XANCB Superoxidase dismutase - Xanthomon  (203 aa)
 initn: 102 init1:  63 opt:  63  Z-score: 91.1  bits: 23.7 E(): 0.0029
Smith-Waterman score: 63;  87.500% identity (87.500% similar) in 8 aa overlap (158-165:164-171)

       130       140       150       160       170       180
PF08_0 LNNNNKLVILQTHDAGNPIKDNTGIPILTCDIWEHAYYIDYRNDRASYVKAWWNLVNWNF
                                        : ::::::
B0RS18 SVTPDKKVVVESTANQDSPLFEGNTPILGLDVWEHAYYLKYQNRRPDYIGAFYNVVNWDE
           140       150       160       170       180       190

       190
PF08_0 ANENLKKAMQK

B0RS18 VERRYHAAIA
           200
```

**b)**

```
>>B0RS18|B0RS18_XANCB Superoxidase dismutase - Xanthomon  (203 aa)
 initn: 530 init1: 255 opt: 535  Z-score: 1583.7  bits: 299.9 E(): 2.1e-86
Smith-Waterman score: 535;  36.946% identity (73.892% similar) in 203 aa overlap (1-196:1-202)

               10        20        30        40        50        60
PF08_0 MVITLPKLKYALNALSPHISEETLNFHYNKHHAGYVNKLNTLIKDTPFAEKSLLDIVKES
       :. :::.: :: .:: :.:. .:...:..:::  :.:..:. .. : .:.  . ..:..
B0RS18 MAYTLPQLPYAYDALEPNIDAQTMEIHHTKHHQTYINNVNAALEGTEYADLPVEELVSKL
               10        20        30        40        50        60

                     70        80        90       100       110
PF08_0 S-------GAIFNNAAQIWNHTFYWDSMGPDCGGEPHGEIKEKIQEDFGSFNNFKEQFSN
       .       : . ::..   ::...:  :.:. ::::.::. . :..:.:.:..::: :..
B0RS18 KSLPENLQGPVRNNGGGHANHSLFWTVMSPNGGGEPKGEVAKAIDKDIGGFEKFKEAFTK
               70        80        90       100       110       120

           120       130       140       150       160       170
PF08_0 ILCGHFGSGWGWLALNNNNKLVILQTHDAGNPIKDNTGIPILTCDIWEHAYYIDYRNDRA
          ..:::::.::... ..:.:. .: .  .:. ...  :::  :.:::::::. :.: :
B0RS18 AALSRFGSGWAWLSVTPDKKVVVESTANQDSPLFEGN-TPILGLDVWEHAYYLKYQNRRP
              130       140       150       160       170

           180       190
PF08_0 SYVKAWWNLVNWNFANENLKKAMQK
       .:. :..:.:::. ...  . :.
B0RS18 DYIGAFYNVVNWDEVERRYHAAIA
     180       190       200
```

c)

```
>>B0RS18|B0RS18_XANCB Superoxidase dismutase - Xanthomon  (203 aa)
 initn: 956 init1: 409 opt: 924  Z-score: 2813.8  bits: 527.5 E(): 6.3e-155
Smith-Waterman score: 924;  38.537% identity (83.902% similar) in 205 aa overlap (1-197:1-203)

               10        20        30        40        50
PF08_0 MVITLPKLKYALNALSPHISEETLNFHYNKHHAGYVNKLNTLIKDTPFAEKSLLDIV-KE
       :. :::.:.:: .::.:.:...:...:..:::  :.:..:.... : .:. .. ..: :
B0RS18 MAYTLPQLPYAYDALEPNIDAQTMEIHHTKHHQTYINNVNAALEGTEYADLPVEELVSKL
               10        20        30        40        50        60

      60              70        80        90       100       110
PF08_0 SS------GAIFNNAAQIWNHTFYWDSMGPDCGGEPHGEIKEKIQEDFGSFNNFKEQFSN
       .:       :.. ::..   ::...:. :.:. ::::.::....:..:.:..:::.:..
B0RS18 KSLPENLQGPVRNNGGGHANHSLFWTVMSPNGGGEPKGEVAKAIDKDIGGFEKFKEAFTK
               70        80        90       100       110       120

           120       130       140       150       160       170
PF08_0 ILCGHFGSGWGWLALNNNNKLVILQTHDAGNPIKD-NTGIPILTCDIWEHAYYIDYRNDR
       .....:::::.::......:.:. .:..  .:. . ::  ::: .:.:::::::..:.:.:
B0RS18 AALSRFGSGWAWLSVTPDKKVVVESTANQDSPLFEGNT--PILGLDVWEHAYYLKYQNRR
              130       140       150         160       170

            180       190
PF08_0 ASYVKAWWNLVNWNFANENLKKAMQK
       ..:. :..:.:::. .... ..:..
B0RS18 PDYIGAFYNVVNWDEVERRYHAAIA
      180       190       200
```

**Figure 3.10 Alignment of a typical Fe-SOD of *P. falciparum,* with the *Xanthomonas campestris* ortholog** a) Alignment with the PAM2 matrix at a gap opening and extension penalty of -12, -2 respectively b) Alignment with BLOSUM50 at the same parameters c) The alignment with PfFSmat60 matrix (identical parameters) showing an improvement in E-value, similarity, identity & extension (2 additional identical residues marked in yellow colour) compared to BLOSUM50 and an overall improvement in E-value and alignment length compared to PAM2.

## 3.4 Conclusion

In this chapter, the idea of compiling a novel series of amino acid substitution matrices was explored with an aim to improve the sequence alignments of the malaria parasite proteins. The novel approach taken to build a series of *Plasmodium falciparum* specific substitution matrices (PfSSM); symmetric as well as asymmetric (having directionality) matrices have been summarized. The ideology of using a novel dataset of *Plasmodium falciparum* distant orthologs for tabulating substitutions, unlike earlier methods (Yu, Wootton, and Altschul 2003; Bastien, Roy, and Marechal 2005; Yu and Altschul 2005; Brick and Pizzi 2008), seemed to be promising. Various physical parameters i.e. total contributing pairs, total number of blocks used to build the matrices; the relative entropy, trace values, expected score, and matrix averages were calculated for the PfSSM series and compared with the standard matrices.

Comparative study of the standard verses the PfSSM matrices revealed variation in substitution preferences. Further, a performance evaluation of these matrices with sequence alignments revealed qualitative differences in terms of the alignment statistics (E-values and scores). Database searches were improved with the symmetric Smat80 matrix of the PfSSM series. Proteins falling under the twilight region showed improvement in alignment length (spanning important motif regions) with the asymmetric PfFSmat60 matrix.

In summary, the current chapter highlights the approach taken to tackle a

genome bias by compiling a new series of substitution matrices that seemed to supersede the general method of matrix calculation for an organism-specific sequence search. The preliminary results presented in this chapter strengthen the dogma of using a novel organism specific substitution matrix for sequence analysis and annotation of biased genomes like that of *Plasmodium falciparum.* An analysis performed on the complete genome of *Plasmodium falciparum* is discussed in the next chapter with an aim to understand the global performance of PfSSM compared to the conventional matrices.

# CHAPTER 4

---

# Sequence analysis and performance evaluation

## 4.1 Introduction

Sequence analysis has implication in assigning functions to genes and proteins by comparing patterns in sequences, secondary structure features, compositional similarities and alignments (alphabet match) with known sequences. The genome sequences resulting from large scale sequencing projects are subjected to various tools for sequence comparisons and the alignment product is analyzed to infer evolutionary and functional relationships between these sequences.

The basic tools for sequence alignments are BLAST (Altschul et al. 1990) and FASTA (Pearson and Lipman 1988) that are extensively used heuristics which make use of dynamic programming for achieving the best possible alignments. These programs use statistical theory to produce a bit score and expect value (E-value) for each alignment pair. While the bit-score gives an estimate of how good the alignment is (higher the score, better the alignment), the E-value gives an indication of the statistical significance of the alignment and portrays the probability of occurrence of either a chance alignment or a biologically significant one.

There are fundamentally two methods of aligning sequences, namely 'global' and 'local'. A general global alignment technique is the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), which is based on dynamic programming. Global alignments are performed to align sequences that are quite similar to each other or are close homologs such that the entire sequence is aligned using as many characters/alphabets as possible. Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or

similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm (Smith and Waterman 1981) is a general local alignment method that is also based on dynamic programming. SSEARCH is a full implementation of the Smith-Waterman algorithm with well validated statistical estimates.

In case of genomes like *Plasmodium falciparum,* the information achieved using the general methods of sequence analysis is limited due to its extreme compositional divergence. Moreover, the average length of the *Plasmodium* proteins is usually large compared to its orthologs and generally comprises of multiple domains that probably confers a multi-functional facet to the protein. Local alignments thus seem to be more meaningful for aligning *Plasmodium falciparum* proteins, as regions of local similarity, viz. motifs and domains could be identified. In such a situation, however, the statistical significance of the alignments becomes important with a need for a proper estimation of statistical parameters for local alignments.

In the earlier chapter, it was shown as to how the symmetric Smat80 matrix and the asymmetric PfFSmat60 matrix of the PfSSM series performed over the standard matrices in database searches and pair-wise alignments respectively. In the current chapter, a global analysis of the *Plasmodium falciparum* proteome has been presented using these best performing matrices in sequence alignments. Database searches of the *Plasmodium falciparum* hypothetical proteins were performed with Smat80 matrix, using SSEARCH, which uses the Smith Waterman local alignment algorithm. To check the improvement in the alignment statistics of

pair-wise alignments, PfFSmat60 was used to align known/annotated proteins of the parasite with its potential orthologs from other organisms. Next, the precision and specificity of the un-scaled PfSSM matrices was tested with BLAST searches and compared with results obtained for the standard matrices.

The main objective of this study was to evaluate the global performance of these new set of matrices compared to the conventional amino acid substitution matrices in sequence alignments of the *Plasmodium falciparum* proteins. The study gives an insight into the usefulness of the corrections made in the substitution preferences of *Plasmodium falciparum* that are represented in the PfSSM series of amino acid substitution matrices.

## 4.2 Methods

### 4.2.1 Database search

In order to validate the performance of Smat80 on a global scale, a database search of the hypothetical/putative proteins of *Plasmodium falciparum* was performed against the non-redundant database. The *Plasmodium falciparum* genome was downloaded (version 2002, download date – Oct 2007) and the complete list of the protein files was obtained from the ftp site of NCBI (ftp://ftp.ncbi.nih.gov/genomes/Plasmodium_falciparum/). The protein files were filtered to obtain only hypothetical/putative sequences that comprised of 4410 proteins. This was used as the query set for the database searches. The non-redundant (NR) (Oct 2007 version) database was downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/blast/db/). SSEARCH was used from version 3 of the

FASTA package. All the searches with the symmetric Smat80 and the comparable BLOSUM90 matrix were performed at a gap opening and extension penalty of -12 and -2 respectively.

### 4.2.2 Sequence alignments

The overall performance of PfFSmat60 was assessed by performing pair-wise alignments of the annotated proteins (excluding 'hypothetical') of *Plasmodium falciparum* and its potential orthologs. The ortholog set of proteins was generated, using the common best non-self hits obtained for 1340 of the annotated proteins with the BLOSUM90 and Smat80 matrices, against the UniprotKB/Swiss-Prot protein database obtained from the EBI ftp site (http://www.ebi.ac.uk/FTP/). The protein sequence alignments were performed with the fasta34 program of the FASTA package (Pearson and Lipman 1988) at gap opening and extension penalties of -12 and -2 respectively and a ktup equal to 2. The performance of PfFSmat60 in pair-wise alignments was studied in comparison to the alignments obtained with PAM2, BLOSUM100 and BLOSUM50 matrices. While PAM2 and BLOSUM100 were selected considering their closeness to PfFSmat60 in terms of relative entropy, BLOSUM50 was used for its popularity with the FASTA programs.

### 4.2.3 Sensitivity studies with BLAST

#### 4.2.3.1 Calculating Karlin-Altschul's statistics

The statistical parameters, lambda, K, H, alpha and beta were derived for the unscaled symmetric matrices of the PfSSM series at gap penalties of -10 (gap opening) and -1 (gap extension) using the Island algorithm that estimates the

statistical parameters for local alignment score distributions (Altschul et al. 2001). For the unscaled asymmetric matrices, a modified version of the Island C++ code (provided by Stephen Altschul on personal communication) was used to calculate similar statistics. The above mentioned gap penalties were defined best for the PAM70 (comparable to PfFmat series) and BLOSUM90 matrices (comparable to Smat series) in the blast_stat.c file of the NCBI toolkit and were used subsequently for calculating the statistical parameters for the unscaled PfSSM matrices as well.

#### 4.2.3.2 Incorporation of PfSSM series with stand alone BLAST

The standalone version of BLASTP (version 2.2.13) contained in the NCBI toolkit build 06/12/2005, was obtained from the NCBI website (ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/old/20051206). The source code was modified by altering two files, namely blastkar.c and blast_stat.c, to incorporate the statistical details of the new matrices. The modifications required for adding a new matrix were personally communicated by Tom Madden (BLAST software group) and Kevin Brick (Brick and Pizzi 2008) and are also provided in the comment lines of the blastkar.c file of the NCBI toolkit. The BLAST source code was re-compiled on a linux platform (fedora) with these modifications. BLAST searches were performed for the user defined matrices with the –M option at gap opening and extension penalties of -10 and -1 respectively for which the statistics were predefined.

#### 4.2.3.3 Composition based statistics

The performance of Smat80 was compared with the compositionally adjusted substitution matrix of Altschul and coworkers (Altschul et al. 2005). The search was

performed at default E-value and a gap opening and extension penalty of -10 and -1 respectively, for a set of 100 annotated proteins of *Plasmodium falciparum* representing different family/super family of proteins, against the uniprotKB/swissprot protein database. The performance of Smat80 matrix was tested with the modified BLASTp program that could accept the user defined matrix. Version 2.2.15 of BLASTp with the C3 option for the composition based score adjustment (Yu and Altschul 2005) was used for the equivalent entropy BLOSUM90 matrix, for comparison. The performance of the matrix was compared in terms of precision/positive predictive value (PPV= TP/TP+FP) at an E-value cut off equal to 1E-005. The true positive and false positive hits were calculated as the number of true (proteins belonging to the same family or having similar annotation) or false hits found at this E-value cut off, respectively, for both the standard and Smat80 matrices.

#### 4.2.3.4 Precision of PfSSM series

Precision values were used to evaluate the overall sensitivity of the unscaled matrices of the PfSSM matrix series (asymmetric PfFmat, and the symmetric Smat) with the BLAST searches. The study was performed at different E-value cut-offs equal to 10, 1, 1E-001, 1E-003 and 1E-005. Gap penalties (-10, -1) defined best for the above two standard matrices were used for all the BLAST searches performed. The TPR (TP/P) and FPR (FP/P) values obtained for each Super-family of proteins was averaged over all 164 instances and reported at each E-value cut-off for all the matrices. The positive predictive value (TP/TP+FP) or precision of the matrices was

calculated by summing all the TP and FP instances across all the 164 super-families studied. In order to correctly record the true/false number of hits and the missed cases in a database search, a database of known proteins with a pre-assigned Super-family classification was used as described in the section 4.2.3.4.1.

#### 4.2.3.4.1 Super-family dataset

Super family assignments were made for the annotated proteins of *Plasmodium falciparum* using the sequence search option with the Superfamily 1.69 server (Gough et al. 2001) (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/). Proteins for which SF assignments were returned, redundancy was checked and only a single protein was selected as a representative of each assigned superfamily, resulting in 164 proteins of *Plasmodium falciparum* with unique superfamily assignments. These sequences were prefixed with an arbitrary SF number and were used as a query set for database searches.

A list of the protein identification numbers for each of the 164 super-families was retrieved from the 'pirsfinfo.dat' file obtained by ftp download at PIR website (ftp://ftp.pir.georgetown.edu/databases/pirsf/).The SF member sequences were retrieved by batch download at PIR. A simple Perl script was used to prefix each sequence with its arbitrary SF number corresponding to its *Plasmodium* counterpart. This protein set consisted of 104744 proteins and was used as a database for sensitivity/specificity studies with BLAST.

#### 4.2.3.4.2 Stand-alone blast

Database searches were performed with the stand-alone version of BLAST that was modified for accepting the PfSSM series of matrices. Standard matrices (supported by BLAST) i.e. PAM70 (H=1.6) and BLOSUM90 (H=1.18) with relative entropies closest to the entire series of symmetric (Smat) and asymmetric (PfFmat) matrices respectively, were used as standards, for comparison. Blast was run at gap opening and extension penalties of -10 and -1 respectively, that has been defined best for the PAM70 and BLOSUM90 matrices.

## 4.3 Results and Discussion

### 4.3.1 Global performance of Smat80

A database search of the hypothetical/putative proteins of *Plasmodium falciparum* with the Smat80 and BLOSUM90 matrices yielded similar hits (best non-self hits were compared) for 4165 of 4410 (94%) proteins with Smat80 and BLOSUM90 matrices, while 245 (6%) proteins gave non-identical hits (either different proteins or similar proteins from different organisms) (**Figure 4.1**). A further analysis of the identical hits confirmed an improvement in both the E-values and scores for 72% (2988/4165) of these identical cases (BEBS in **Figure 4.2**) with Smat80, while 22% (931/4165) showed either, an improved E-value and similar score or vice versa (IEBS/BEIS in **Figure 4.2**). Proteins that performed poor with respect to both E-values and scores (PEPS in **Figure 4.2**) or at least with respect to one (IEPS /PEIS in **Figure 4.2**), was only 6%. In case of the non-identical hits a difference in subject hit (the best non-self hit was compared) ranking was observed. **Table 4.1, Table 4.2**

and **Table 4.3** show the differences observed in the ranking of subjects by the Smat80 and BLOSUM90 matrices for the non-identical hits. The above three tables are based on the E-value and bit-score performance of the Smat80 matrix. A comparison of the subject annotation for these hits showed an improved annotation for some proteins with Smat80 as represented in **Table 4.4**, **Table 4.5** and **Table 4.6**. These tables are again based on the E-value and bit-score performance of the Smat80 matrix.

The observed improvement in score compared to standard matrix raised the question of the other possible factors which might be responsible for such an increment. The improvement in alignment score with Smat80 matrix compared to BLOSUM90 could have been alternatively explained if optimized alignment parameters were used for Smat80 alignments. However, since all the alignments reported here were performed using the same moderate parameters for both the standard and test matrices, this possibility could be ruled out completely. Moreover, the alignments were tested at various parameters and it was observed that even at the best-optimized parameter for BLOSUM; the alignment score with Smat80 was higher. The only considerable reason that could be figured out was the biased protein dataset used for matrix compilation that could possibly resolve the inconsistency in the target and background frequencies, which is a basic problem with standard matrices. Secondly, the non-identical hits obtained with Smat80 signify the presence of differences in substitution preferences that are represented in the *Plasmodium* specific matrices. The differences observed in the rank order of

subject hits with both these matrices (**Tables 4.1, 4.2 and 4.3**), clearly demonstrates this tendency. Thirdly, the poor performance of Smat80 observed in certain cases (poor E-values and bit-scores) symbolizes the over-representation of certain substitutions in the standard BLOSUM matrix compared to Smat80 and vice versa.

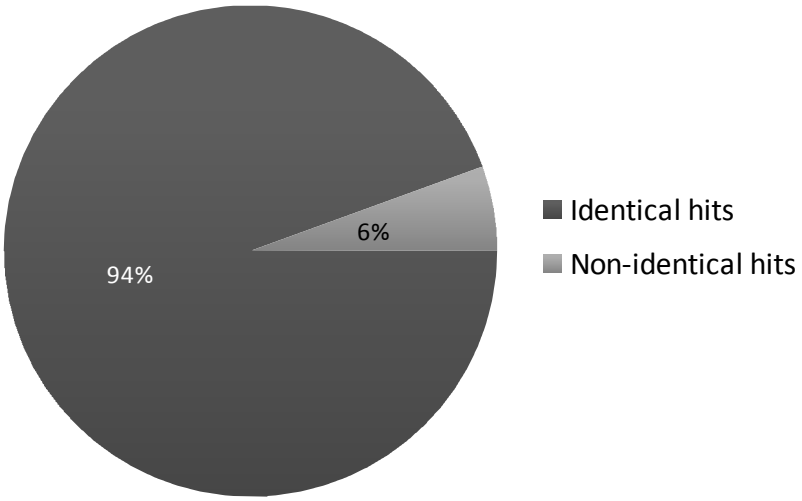


**Figure 4.1 Database search results with Smat80 matrix** The pie chart shows the percentage of best non-self hits obtained with Smat80 that were either identical or non-identical to the BLOSUM90 hits in a database search of the hypothetical/putative proteins of *Plasmodium falciparum* against the non-redundant database.

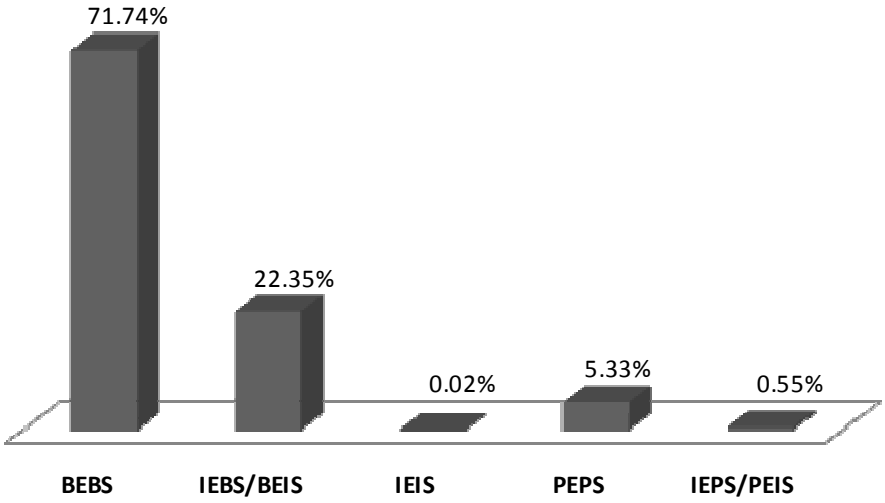


**Figure 4.2 Smat80 alignment statistics for the identical hits** The bar graph shows the percentage of identical hits that gave better/poor/similar statistics with Smat80 compared to the BLOSUM90 matrix. The E-values and bit scores have been compared here. **BEBS** - Better E-values better scores; **IEBS** - Identical E-values better scores; **BEIS** - Better E-values identical scores; **IEIS** - Identical E-values identical scores; **PEPS** – Poor E-values poor scores; **IEPS** -Identical E-values poor scores; **PEIS** – Poor E-values identical scores

# TABLE 4.1

# TABLE 4.1

# TABLE 4.1

# TABLE 4.1

# TABLE 4.1

# TABLE 4.1

**Table 4.2 Comparison of subject hit ranking for the non-identical hits obtained with BLOSUM90 and Smat80 matrices - Smat80 matrix gave better E-values and poor bit-scores and vice versa for the following**

| *Plasmodium falciparum* query (gi) | BLOSUM90 subject hit (gi) | Smat80 subject hit (gi) | BLOSUM90 ranking of Smat80 subject hit | Smat80 ranking of BLOSUM90 subject hit |
|---|---|---|---|---|
| 124802939 | 124808847 | 68492502 | 2nd | 2nd |
| 124804955 | 150398763 | 46358439 | 6th | 5th |
| 124808190 | 86170913 | 126304686 | 3rd | 2nd |
| 124810365 | 124808958 | 70936748 | 3rd | 4th |
| 124512628 | 68075521 | 60651170 | 2nd | 3rd |
| 124505683 | 66805187 | 123455321 | 4th | 5th |
| 124505885 | 124513494 | 124507028 | 2nd | 2nd |
| 124504917 | 124808898 | 70953423 | 2nd | 3rd |
| 124505217 | 83314370 | 68063533 | 2nd | 2nd |
| 124505499 | 150387395 | 109499053 | 9th | 8th |
| 124506359 | 118084385 | 149262837 | 7th | Not found |
| 124507091 | 124513104 | 70946640 | 2nd | 7th |
| 124507247 | 124810376 | 66826105 | 2nd | 3rd |
| 124803528 | 73993547 | 124513342 | 3rd | 5th |
| 124809024 | 26892020 | 126273182 | 3rd | 2nd |
| 124505199 | 15892426 | 53715658 | 3rd | 2nd |

**Footnote**: Ranking numbers in columns 4 & 5 are given considering the reported non-self hit as rank 1. "Not found" means that the particular subject hit was not reported among the first 20 hits that were reported.

# TABLE 4.3

**TABLE 4.3**

**TABLE 4.4**

# TABLE 4.4

**Table 4.5 Significant differences in subject annotation for the non-identical hits obtained with BLOSUM90 & Smat80 matrices - Smat80 matrix gave better E-values and poor bit-scores and vice versa for the following**

| Query | Subject hit with BLOSUM 90 | Subject hit with Smat80 | Subject annotation with BLOSUM90 | Subject annotation with Smat80 |
|---|---|---|---|---|
| 124804955 | 150398763 | 46358439 | Protein of unknown function DUF6 transmembrane | NADH dehydrogenase subunit 6 |
| 124808190 | 86170913 | 126304686 | Hypothetical | PREDICTED: similar to KIAA1962 protein |
| 124809024 | 26892020 | 126273182 | Hypothetical | Similar to ectonucleoside triphosphate diphosphohydrolase 1 |

**Table 4.6 Significant differences in subject annotation for the non-identical hits obtained with BLOSUM90 & Smat80 matrices - Smat80 matrix gave poor/similar E-values and poor/similar bit-scores for the following**

| Query | Subject hit with BLOSUM9 | Subject hit with Smat80 | Subject annotation with BLOSUM90 | Subject annotation with Smat80 |
|---|---|---|---|---|
| 124802361 | 829215 | 156098211 | unnamed protein product | early transcribed membrane protein |
| 124802391 | 157768881 | 150387395 | Hypothetical | Chromosome segregation ATPases-like protein |
| 124803004 | 15004730 | 116335017 | Hypothetical | Ribosomal protein L33 |
| 124804938 | 109119000 | 87119643 | Similar to chromatin modifying protein 6 | Putative protein translocase subunit |
| 124810460 | 156100567 | 121944764 | Serine/threonine protein kinases, putative | Immunoglobulin A heavy chain variable region |
| 124810517 | 157768881 | 156093215 | Hypothetical | tryptophan-rich antigen |
| 124810579 | 83282688 | 66813688 | Hypothetical | Bromodomain containing protein |
| 124511642 | 34495238 | 83314798 | Hypothetical | CCAAT-box DNA binding protein subunit B |

### 4.3.2 Global performance of PfFSmat60

A comparison of the *E*-values, bit-scores and the alignment overlap obtained for the pair-wise alignments of the *Plasmodium falciparum* proteins and their orthologs showed better performance of PfFSmat60, as compared to standard matrices. Compared to the similar entropy PAM2 matrix, the *E*-values with PfFSmat60 were better for 92% of the cases studied, while the remaining 8% showed similar *E*-values. The bit-scores improved for 99.9% of the cases while 0.1% showed scores similar to that obtained with PAM2. The alignment extension with PfFSmat60 improved for 90% of the alignments compared. About 6% performed similarly, while 4% gave poor alignment extensions, compared to PAM2.

The PfFSmat60 matrix showed improved bit-scores compared to BLOSUM100, for 99.9% of cases while 0.1% scored similar. The E-values improved for 89% of the pair-wise alignments while 11% scored similar. The alignment overlap was better for 89% cases, while it was similar for 10% and poor for 1% of the protein alignments.

When compared to the popular BLOSUM50 matrix, the bit-scores with PfFSmat60 improved for 99.8% of the alignments. While 0.1% performed similarly, the remaining 0.1% of the proteins performed poorly. The E-values with PfFSmat60 were better for 85% of the cases, similar to BLOSUM50 E-values for 15% of cases and was poor for the remaining 0.1%. The alignment extension on the other hand improved for 86% of the alignments, while 13.5% performed similar and 0.5% performed worse than BLOSUM50. Conclusively, the performance of PfFSmat60

was globally better than the standard matrices in pair-wise alignments. The number of better/same/poor instances of E-values, bit-scores and alignment extension obtained with the PfFSmat60 in comparison to the PAM2, BLOSUM100 and BLOSUM50 matrices are represented in **Figure 4.3**, **Figure 4.4**, **Figure 4.5** respectively.

Acknowledging the improvement in bit-scores and alignment length with PfFSmat60, it was important to rule out the other parameters that could play a role in the resultant changes. Alignment scores are known to increase linearly for a positive matrix where all the matrix values are greater than zero (Vingron and Waterman 1994). Since PfFSmat60 was a scaled matrix (scaled by a factor of +3), there was a possibility of such a linear relationship. However, PfFSmat60 matrix scaling was stringent where the matrix values were scaled only to the range of BLOSUM values and it was not a positive matrix. Second, one may argue that the alignment extension achieved might be due to the use of low gap penalties. Nevertheless, moderate gap opening and extension penalties of -12 and -2 respectively were used and hence this chance could also be ruled out. Lastly, since PfFSmat60 has high entropy, it is expected to easily distinguish real from chance alignments (Friedberg et al. 2007). Hence, it could be concluded that the choice of *Plasmodium* biased protein blocks and uni-directional substitutions for building asymmetric matrices, probably solved the problem of twilight regions (Rost 1999).

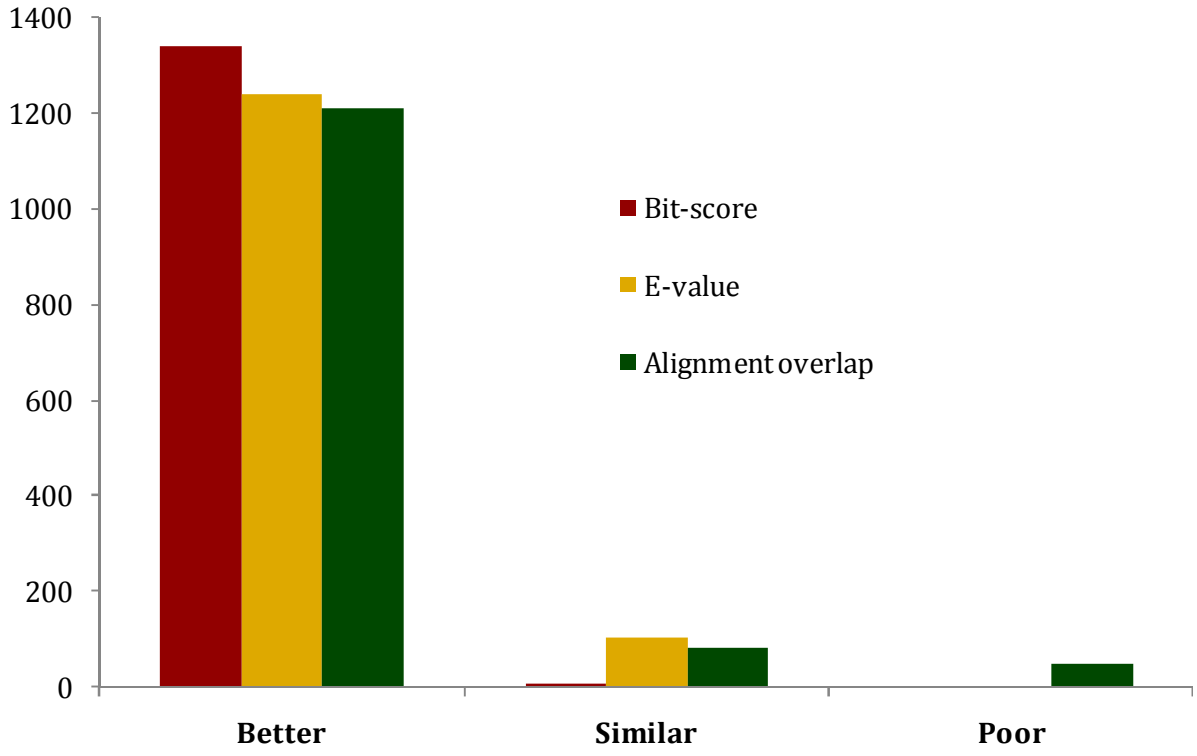


**Figure 4.3 The performance of PfFSmat60 versus PAM2 matrix** The values compared were the Bit-scores, E-values and the alignment overlap for 1340 pair-wise alignments.

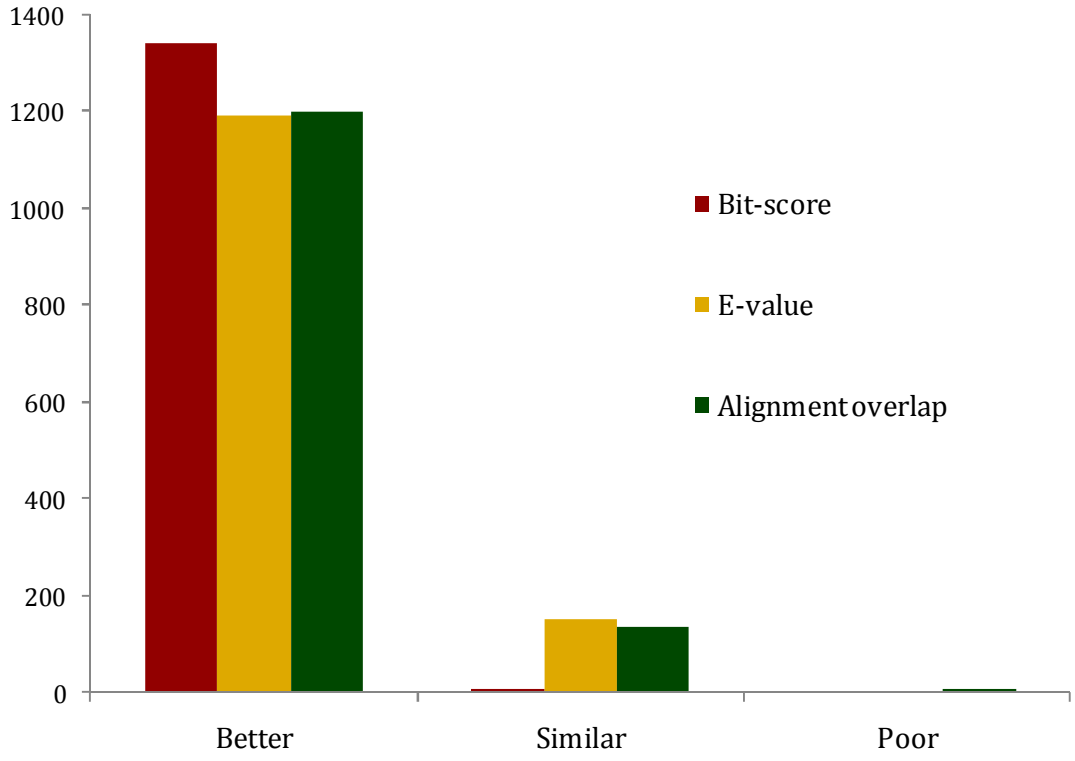


**Figure 4.4 The performance of PfFSmat60 versus BLOSUM100 matrix** The values compared were the Bit-scores, E-values and the alignment overlap for 1340 pair-wise alignments.

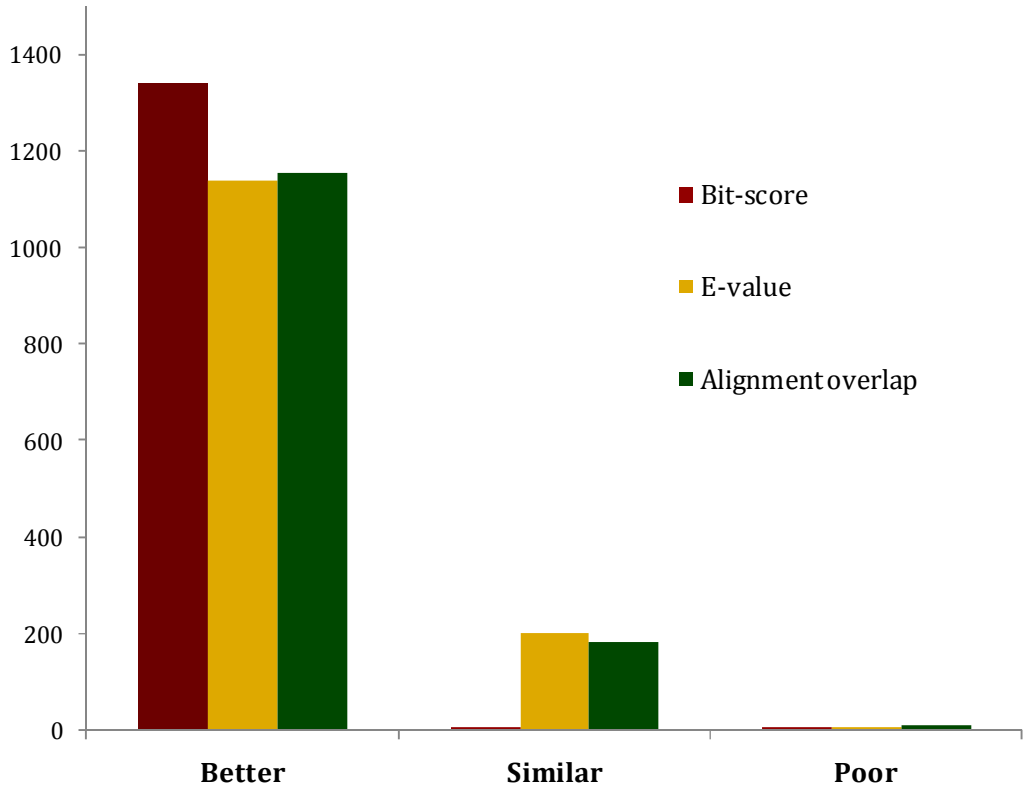


**Figure 4.5 The performance of PfFSmat60 versus BLOSUM50 matrix** The values compared were the Bit-scores, E-values and the alignment overlap for 1340 pair-wise alignments

### 4.3.3 Composition based statistics and sensitivity of Smat80

Comparison of precision values at an E-value of 1E-005 obtained in a database search for the Smat80 and compositional adjusted (Yu, Wootton, and Altschul 2003) BLOSUM90 matrix, yielded similar PPV (positive predictive value) values for 96% of the protein super-families. For 3% of the proteins, Smat80 showed a better performance while only 1% showed a poor performance. The results are summarized in the **Figure 4.6**. The positive predictive value of Smat80 was thus as good as the compositionally adjusted BLOSUM90 matrix. A comparison of the total number of true hits that could be picked up at the default E value of 10 resulted in similar performance for 43% of the proteins, better performance for 35% of the proteins and worse for the rest 22%.

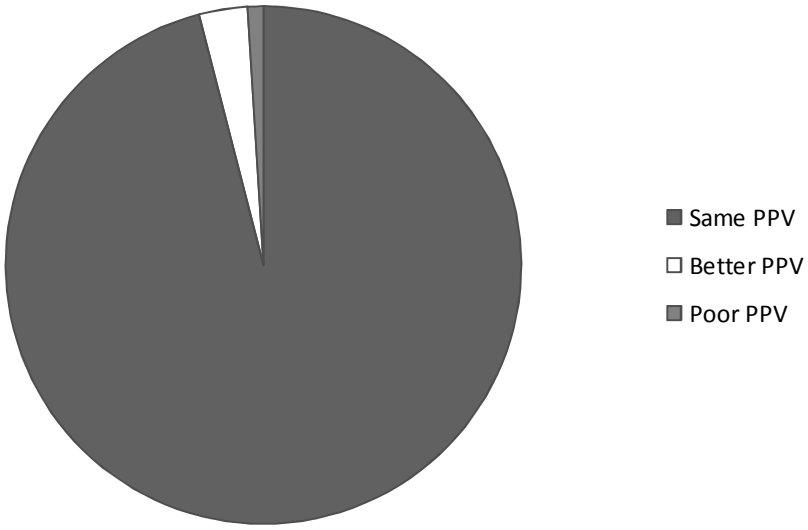


**Figure 4.6 A comparison of precision for Smat80 and composition adjusted BLOSUM90 matrix in database searches**

### 4.3.4 Precision of unscaled PfSSM matrices

Compared to the standard matrices, the overall precision of the unscaled PfSSM matrices was found to be better, preferably at higher E-values. Differences were observed in the precision rate for the Smat series and BLOSUM90 with an observable improvement at a higher expectation value. At an E-value of 1E-05 and 1E-03 the highest precision attained with Smat series was 96.46% (Smat50) while for BLOSUM90 it was 96.48% and 96.47% respectively, which were comparable. Similarly at an E-value of 1E-01, the PPV% was comparable for Smat80 (96.44%) and BLOSUM90 (96.40%). However, with an increasing E-value (1 and 10) the differences in the positive predictive value (%) became more apparent for these matrices with a better precision demonstrated by the Smat series. The precision improved for Smat series (96.26% for Smat80 and Smat90) compared to BLOSUM90 (96.08%) at an E-value of 1 while at an E-value of 10, the precision was markedly improved for Smat series (Smat60 PPV% equal to 94.53) compared to BLOSUM90 (PPV% equal to 94.06). This observation indicates that all the matrices performed similarly when aligning highly similar proteins with a significant difference in performance for less similar sequences. The percentage of positive predictive values for the Smat series and BLOSUM90 are tabulated in **Table 4.7**.

A comparison of the precision rates for the asymmetric matrices, PfFmat, showed almost a similar trend of improvement over the standard. There was an overall improvement in the precision rate at all E-values compared to PAM70, and the difference was well marked at an E-value of 1 and 10. While the precision rates

were slightly improved with the PfFmat series at an E-value of 1E-05 (96.46% for PfFmat70 and 96.41% for PAM70) and 1E-03 (96.47% for PfFmat50 & PfFmat70; 96.42% for PAM70), they were comparable at an E-value of 1E-01(96.44% for PfFmat50 and 96.36% for PAM70). The precision rates were however much better with PfFmat series at E-values of 1 (96.34% for PfFmat90 and 96.20% for PAM70) and 10 (95.84% for PfFmat90 and 94.73% for PAM70). The percentage of positive predictive values for the PfFmat series and PAM70 are represented in **Table 4.8**.

The overall precision of the unscaled PfSSM matrices (both symmetric and asymmetric) thus seemed to improve with a relaxation in the E-value employed for the BLAST searches. The probability of false positives is usually expected to increase drastically with increasing expectation value. However, a characteristic feature that was observed in the case of PfSSM matrices was its greater discriminatory power to predict positives correctly at a higher E-value compared to the standard matrix.

**Table 4.7 Positive predictive values (%) of Smat series & BLOSUM90 in a database search**

| | 1E-05 | 1E-03 | IE-01 | 1E | 1E+001 |
|---|---|---|---|---|---|
| **Smat50** | 96.46 | 96.46 | 96.38 | 96.09 | 94.44 |
| **Smat60** | 96.43 | 96.42 | 96.37 | 96.10 | 94.53 |
| **Smat70** | 96.42 | 96.41 | 96.36 | 96.10 | 94.50 |
| **Smat80** | 96.45 | 96.45 | 96.44 | 96.26 | 95.20 |
| **Smat90** | 96.38 | 96.41 | 96.38 | 96.26 | 95.39 |
| **BLOSUM90** | 96.48 | 96.47 | 96.40 | 96.08 | 94.06 |

**Footnote**: E-value cut-off for BLAST searches represented in row 1; Matrices used for search listed in column 1; Figures highlighted in red shows the highest PPV% obtained by a Smat series at a particular E-value cut-off; Text highlighted in blue are the BLOSUM90 PPV%

**Table 4.8 Positive predictive values (%) of PfFSmat series and PAM70 in a database search**

| | 1E-05 | 1E-03 | IE-01 | 1E | 1E+0011 |
|---|---|---|---|---|---|
| **PfFmat50** | 96.45 | 96.47 | 96.44 | 96.29 | 95.83 |
| **PfFmat60** | 95.85 | 96.02 | 95.96 | 95.87 | 95.32 |
| **PfFmat70** | 96.46 | 96.47 | 96.43 | 96.31 | 95.77 |
| **PfFmat80** | 96.41 | 96.39 | 96.40 | 96.29 | 95.83 |
| **PfFmat90** | 96.40 | 96.42 | 96.43 | 96.34 | 95.84 |
| **PAM70** | 96.41 | 96.42 | 96.36 | 96.20 | 94.73 |

**Footnote**: E-value cut-off for BLAST searches represented in row 1; Matrices used for search listed in column 1; Figures highlighted in red shows the highest PPV% obtained by a PfFmat series at a particular E-value cut-off; Text highlighted in blue are the PAM70 PPV%

A comparison of false positives picked in BLAST searches by the PfFmat series and PAM70 matrices for the 164 families of proteins studied, revealed, that on average for 98% of the proteins studied, the rate of false positives predicted were similar for these matrices at an E value of 1E-05 (**Table 4.9**). However, at a higher E-value of 10, the ratio of the number of better (low FPR) to similar predictions (same FPR) improved significantly for the PfFmat series (**Table 4.10**). Conclusively, at a higher E-value a majority of proteins showed a comparatively low FPR i.e. the number of false positives picked by PfFmat series was less than those picked by PAM70. A direct comparison of the average FPR/TPR ratios for this dataset at an E-value of 10 also showed similar results, where, comparatively higher ratios were obtained with PAM70.This particular feature of a matrix is of significance in profile studies like PSI-BLAST (Altschul et al. 1997), where the presence of false positives in the results set used to build the search matrix can skew the search profile. A repetition of this analysis with the Smat series of matrices gave similar results. On an average, 99% of the proteins showed similar false positive rates with both the Smat and BLOSUM90 matrices at an E value of 1E-05 (**Table 4.11**). But, at an E-value of 10, the ratio of the number of better (low FPR) to similar predictions (same FPR) improved significantly for the series (**Table 4.12**).

**Table 4.9 Comparative performance of PfFmat series in predicting false positives at an E-value of 1E-05**

| PL | PfFmat50 | PfFmat60 | PfFmat70 | PfFmat80 | PfFmat90 | Average | Protein% |
|---|---|---|---|---|---|---|---|
| Better | 1 | 2 | 1 | 1 | 1 | 1.2 | 0.73 |
| Similar | 161 | 160 | 161 | 162 | 162 | 161.2 | 98.29 |
| Poor | 2 | 2 | 2 | 1 | 1 | 1.6 | 0.98 |

**Footnote:** Column 1 shows the performance level (PL) of PfFmat series with respect to PAM70 FPRs. The values in columns 2-6 show the number of instances of occurrence (better/similar/poor) corresponding to the matrices represented in row2. Column 7 represents the average of columns 2-6. Column 8 represents the average percentage (column7/164*100) of the total protein families that show better/similar/poor FPR compared to PAM70 values.

**Table 4.10 Comparative performance of PfFmat series in predicting false positives at an E-value of 1E+001**

| PL | PfFmat50 | PfFmat60 | PfFmat70 | PfFmat80 | PfFmat90 | Average | Protein% |
|---|---|---|---|---|---|---|---|
| Better | 78 | 81 | 76 | 80 | 83 | 79.6 | 48.50 |
| Similar | 72 | 62 | 72 | 73 | 70 | 69.8 | 42.60 |
| Poor | 14 | 21 | 16 | 11 | 11 | 14.6 | 8.90 |

**Footnote**: Column 1 shows the performance level (PL) of PfFmat series with respect to PAM70 FPRs. The values in columns 2-6 show the number of instances of occurrence (better/similar/poor) corresponding to the matrices represented in row2. Column 7 represents the average of columns 2-6. Column 8 represents the average percentage (column7/164*100) of the total protein families that show better/similar/poor FPR compared to PAM70 values.

**Table 4.11 Comparative performance of Smat series in predicting false positives at an E-value of 1E-05**

| PL | Smat50 | Smat60 | Smat70 | Smat80 | Smat90 | Average | Protein% |
|---|---|---|---|---|---|---|---|
| Better | 1 | 0 | 0 | 1 | 2 | 0.8 | 0.49 |
| Similar | 162 | 162 | 161 | 163 | 161 | 161.8 | 98.66 |
| Poor | 1 | 2 | 3 | 0 | 1 | 1.4 | 0.85 |

**Footnote**: Column 1 shows the performance level (PL) of Smat series with respect to BLOSUM90 FPRs. The values in columns 2-6 show the number of instances of occurrence (better/similar/poor) corresponding to the matrices represented in row2. Column 7 represents the average of columns 2-6. Column 8 represents the average percentage (column7/164*100) of the total protein families that show better/similar/poor FPR compared to BLOSUM90 values.

**Table 4.12 Comparative performance of PfFmat series in predicting false positives at an E-value of 1E+001**

| PL | Smat50 | Smat60 | Smat70 | Smat80 | Smat90 | Average | Protein% |
|---|---|---|---|---|---|---|---|
| Better | 61 | 61 | 58 | 76 | 79 | 67 | 40.85 |
| Similar | 78 | 79 | 79 | 78 | 78 | 78.4 | 47.80 |
| Poor | 25 | 24 | 27 | 10 | 7 | 18.6 | 11.34 |

**Footnote**: Column 1 shows the performance level (PL) of Smat series with respect to BLOSUM90 FPRs. The values in columns 2-6 show the number of instances of occurrence (better/similar/poor) corresponding to the matrices represented in row2. Column 7 represents the average of columns 2-6. Column 8 represents the average percentage of the total protein families that show better/similar/poor FPR compared to BLOSUM90 values.

A comparison of the average specificity obtained in database searches of the 164 *Plasmodium* proteins showed better performance with PfFmat series over PAM70, with PfFmat60 showing the highest specificity (**Table 4.13**). However, the overall specificity with Smat series was much higher with a slight increment in specificity compared to BLOSUM90 (**Table 4.14**).

A point to be noted here is that, a vast number of *Plasmodium falciparum* proteins show SF assignments to more than one SF due to the multi-domain nature of the proteins. Since in this case, such assignments have been ignored, there is always a possibility that the BLAST hits that were regarded as false positives were actually true, considering the secondary SF assignments to the protein. Our results would have been well defined if this particular tendency was also checked.

**Table 4.13 The average specificity of PAM70 and PfFmat series with BLAST at an E-value of 1E+001**

| PAM70 | PfFmat50 | PfFmat60 | PfFmat70 | PfFmat80 | PfFmat90 |
|---|---|---|---|---|---|
| 48.86 | 49.35 | 57.38 | 49.17 | 49.70 | 50.00 |

**Footnote**: The values are given as percentages

**Table 4.14 The average specificity of BLOSUM90 and Smat series with BLAST at an E-value of 1E+001**

| BLOSUM90 | Smat50 | Smat60 | Smat70 | Smat80 | Smat90 |
|---|---|---|---|---|---|
| 99.983 | 99.984 | 99.984 | 99.984 | 99.986 | 99.987 |

**Footnote**: The values are given as percentages rounded to third decimal for precision.

## 4.4 Conclusion

The present chapter highlights the global performance of Smat80 and PfFSmat60 matrices that were discussed in the previous chapter. The overall performance of these matrices in database searches and pair-wise alignments respectively was observed to be better over the standard matrices. The alignment statistics improved significantly with Smat80 for 94% of the protein alignments. The ranking of subject hits varied strikingly for the non-identical hits, being indicative of variation in substitution. A remarkable difference in subject annotation was also observed in some instances. In pair-wise alignments with PfFSmat60 a significant improvement in E-values (92%), bit-scores (99.9%) and alignment overlap (90%) was achieved compared to the standard matrix having similar entropy. The un-scaled symmetric matrix series, Smat and the asymmetric PfFmat matrices showed an overall better precision and specificity at higher E-values in database searches. Furthermore, these matrices demonstrated comparatively lower FPR/TPR ratios signifying their relevance in profile searches.

In summary, when compared on a global scale, the overall performance of PfSSM matrices was found to be better than the standard matrices like BLOSUM and PAM. A difference in annotation of the subject hit observed for the non-identical hits, lead to a further effort in this direction, to identify potential orthologs for the hypothetical proteins of the malaria parasite as discussed in the next chapter.

# CHAPTER 5

---

# Annotation of hypothetical proteins of *Plasmodium falciparum*

## 5.1 Introduction

The process of protein annotation involves attaching relevant biological information to predicted protein sequences. The functional assignments include assigning probable biological function, involved interactions, expression and possible role in different cellular/biological pathways of the organism under study. Database searches are used to transfer functional features from annotated proteins to the query sequences. The transfer of functional annotation from annotated proteins of model organisms is one of the main applications of comparative genomics.

In the early days of comparative biology, relationships between different species were studied using morphological characters (Kuzniar et al. 2008). With the emergence of sequencing techniques and, in particular, the high-throughput techniques of the past decade, the amount of molecular characters in the form of fully sequenced genomes has increased enormously (Kuzniar et al. 2008). A wide range of bioinformatics tools have been developed to interpret the sequence data from evolutionary and functional perspectives. The knowledge of molecular phlogeny in general and orthology in particular has become an integral component of many genome-scale studies like gene order, expression, regulatory networks, metabolic pathways and functional genome annotation (Kuzniar et al. 2008).

Various methods are used to analyze cross-species orthologous relationships according to an operational definition of orthology. Orthologs are considered to be homologous sequences derived by a speciation event from a single

ancestral sequence in the last common ancestor of the species being compared (Kuzniar et al. 2008). They have been demonstrated to typically perform equivalent functions in closely related species. Hence ortholog prediction is widely used to infer putative functions for unknown proteins, though the quality of the prediction is an important factor in the transfer of functional annotations. Like any other method, the use of orthologs in function annotation has got its own drawbacks in case of molecular events like domain shuffling, presence or absence of a domain, lineage specific gene duplication and gene loss (Sjolander 2004).

Some of the methods that are commonly used as first-pass approximations to find putative orthologs using the essence of the 'best' genome-wide matches between two species are; best hit (BeT), reciprocal best hit (RBH), bi-directional best hit (BBH), symmetrical best hit (SymBeT) and reciprocal smallest distance (RSD) (Kuzniar et al. 2008). The 'best bidirectional hit' (BBH) method (Hulsen et al. 2006) is the most frequently applied method to determine orthologous pairs. It assumes that a cross-species protein pair in which each protein gives back the other protein as being the best hit in the whole other proteome is an orthologous pair.

One of the major impacts of sequence divergence in *Plasmodium falciparum* is the acknowledged absence of some enzymes in metabolic pathways. While many enzymes in metabolic processes have been identified, several examples exist of incomplete pathways. One such example is that of the shikimate pathway for which the enzyme activities were later experimentally detected from crude extracts of the parasite (McConkey et al. 2004). A plausible explanation for this would be that

some enzymes could have evolved independently of their counterparts in other organisms so that, although they catalyse the same reaction, their sequences are unrelated. Second, they are too distantly related to known sequences to be readily identified by general methods of orthology. Nonetheless, it might be possible to identify these missing enzymes using novel methods in comparative genomics.

In the previous chapter it was shown as to how a significant difference in subject ranking and annotation was obtained with Smat80 in database searches with BLAST. These observations lead to the search for potential orthologs for the hypothetical proteins of *Plasmodium falciparum* with the entire Smat series based on best bi-directional hits, the results of which are summarized in the current chapter. The resultant functional assignments were cross validated with the *Plasmodium falciparum* interactome data where ever possible. The primary aim of this exercise was to improve the current annotation of the hypothetical proteins of *Plasmodium falciparum* with a further attempt to identify gap fillers of the parasite's metabolic pathways that are either incomplete or are considered absent in the parasite.

## 5.2 Methods

### 5.2.1 Ortholog detection

A variant of the Best bi-directional hit (BBH) method was employed to search for potential orthologs for the hypothetical/putative proteins of *Plasmodium falciparum* against the manually annotated UniprotKB/Swiss-Prot database. Though bi-directional hit method is usually employed in the context of two

genomes, in this study the UniprotKB/Swiss-Prot database was used because, firstly, the hits obtained would have high quality manual annotations and secondly, a wide range of proteins from different organisms would be available for search which is an important criterion for *Plasmodium falciparum* that has distant relatives across a range of organisms. A point to be noted here is that only cross species hits to proteins were considered and within species hits (if any) were discarded from the study. The search was performed with the modified version of BLAST at default E-value of 10 and a gap opening and extension penalty of -10 and -1 respectively with the entire Smat series (Smat50, Smat60, Smat70, Smat80, and Smat90) and the equivalent BLOSUM90 matrix. The BBH results obtained with each Smat series was compared to the results obtained with the standard matrix. The uncommon BBH's obtained with the Smat series were sorted from which a list of exclusive BBH's (*Plasmodium falciparum* proteins that gave best bi-directional hits with Smat and not with BLOSUM90) was tabulated. A list of *Plasmodium falciparum* protein identifiers was made which consisted of proteins that gave BBH's exclusively with each Smat matrix series.

Similarly, in order to identify orthologs of the hypothetical proteins of the malaria parasite in model organisms, BLASTp search was done to identify BBHs against the *Arabidopsis thaliana* and the *Saccharomyces cerevisiae* genomes for which the complete protein (.faa) files were obtained from the NCBI ftp site.

### 5.2.2 Pathway mapping

The BBH's obtained for the *Plasmodium falciparum* hypothetical/putative proteins as described in the previous section were subjected to Gene ontology (GO) annotations provided by the GOA group at EBI. The 'QuickGO' browser available at http://www.ebi.ac.uk/ego, was used for this purpose that provides high quality manual as well as electronic annotations for the proteins in the UniProt knowledgebase. The GO term 'process' was listed for these proteins and these annotations were transferred to the hypothetical proteins of *Plasmodium falciparum*. The protein protein interaction data of the malaria parasite available at PlasmoMAP (http://cbil.upenn.edu/plasmoMAP/index-v1.html) and the interactions of unknown proteins with known proteins (Date and Stoeckert 2006) were used to confirm some of the GO annotations assigned to the *Plasmodium falciparum* proteins. Apart from this, the interacting partners involved in a process similar to the query protein were identified. For this, *Plasmodium* proteins involved in a particular process were picked using 'term search' at PlasmoDB, from which the interacting members were identified for the respective query protein.

## 5.3 Results and Discussion

### 5.3.1 Detection of potential orthologs with Smat series using BBH

A search for potential orthologs against the annotated database with the Smat80 matrix gave best bi-directional hits for 256 hypothetical/putative proteins of *Plasmodium falciparum* for which no BBH's were picked with the BLOSUM90 matrix. Similarly, excluding the overlapping hits, Smat50 gave BBH's for 49

proteins, Smat60 gave for 34, Smat70 gave for 14 and Smat90 gave BBH's for 148 proteins exclusively. These results are summarized in **Table 5.1**. The annotations (based on the subject) so obtained for the hypothetical proteins of *Plasmodium falciparum* were compared with the protein annotations provided in PlasmoDB (http://plasmodb.org/plasmo/) database (version 6.3) which are summarized in **Table 5.2**. It was observed that for some of the hypothetical proteins, the Smat annotations matched the updated annotations available on this website. The BBH's obtained were further categorized based on the Gene Ontology terms (process) of the subject proteins, such that the role of these hypothetical proteins in various pathways of the malaria parasite could be better understood. This data is summarized in **Table 5.3** from which it is clear that a majority of the hypothetical proteins were mapped to metabolic/biosynthetic processes while a good number of them were related to transcription, cell division/repair processes, signalling and processes involving development/conjugation/reproduction.

**Table 5.1 The number of *Plasmodium falciparum* proteins for which Smat series exclusively gave BBH's**

| Matrix | No. of proteins that gave BBH exclusively |
|---|---|
| Smat50 | 49 |
| Smat60 | 34 |
| Smat70 | 14 |
| Smat80 | 256 |

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.2**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

**TABLE 5.3**

#### 5.3.1.1 Comparison with interactome data

Proteins that are part of an interaction network are often known to be involved in similar functions (McDermott and Samudrala 2004). Proteins that are uncharacterized can be thus associated with known pathways based on the interaction of unknown proteins with functionally characterized proteins. To substantiate the results obtained as in **Table 5.3**, some of the *Plasmodium falciparum* proteins from each category of the listed GO term – process were further analyzed based on their interacting partners as a further step for confirmation. **Table 5.4** represents the proteins for a particular GO process and their interacting partners that were found to be involved in a similar process. A point to be noted here is that most of the proteins were not a part of the interaction network, and even if they were, their interacting partners were mostly hypothetical. For some of the proteins from **Table 5.3**, the data from the interaction map of the unknown proteins to know proteins (data available at PlasmoMAP) was compared, and the possible role of the query protein in the assigned GO term was studied. These results are discussed in the following section.

**Pathogenesis (PF11_0154):** Some of the interacting partners of this hypothetical protein were rifins and an erythrocyte membrane protein that are related to pathogenesis. Other interacting partners were Cg7 (homolog of this protein in *Plasmodium vivax* is involved in drug resistance), glycophorin_binding_protein 130 precursor (one of the proteins identified in *Plasmodium falciparum* infected RBC's) (Wu 2006), Fe-superoxide_dismutase (may be indirectly related to pathogenesis as

it is involved in antioxidant defence for a protective mechanism), phosphatidyl-inositol glycan (involved in GPI anchor synthesis), and heat shock proteins (chaperone involvement in cytoadherence is known) (Pavithra, Kumar, and Tatu 2007). This protein is annotated as a conserved protein of unknown function in PlasmoDB database.

**Proteolysis (PFB0600c):** A good number of interacting partners for PFB0600c were proteins that are subunits of the proteasome complex, involved in the process of degradation of misfolded proteins. The other interacting partners were ubiquitin protein ligase, ubiquitin carboxyl-terminal hydrolase and ubiquitin like proteins that are involved in ubiquitin mediated proteolysis. Proteins like vacuolar sorting protein 35 (these are known to be involved in ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway in some organisms), methionine aminopeptidase (removes the amino-terminal methionine from nascent proteins), glutathione peroxidase (may have an indirect role since they are known to be involved in the proteasome inhibitory activity in humans), Skp1 family protein (involved in ubiquitination and subsequent proteasomal degradation of target proteins) and SEL-1 (involved in ubiquitin-dependent degradation of misfolded endoplasmic reticulum proteins) were also part of the interaction network. This protein in *Plasmodium* is annotated as a conserved protein of unknown function in the PlasmoDB database.

**Autophagic cell death (PFB0490c):** The interacting partners of this protein might have a role in autophagic cell death, namely subunit of proteasome activator

complex, and ubiquitin conjugating enzyme, that degrade short lived proteins. Yet, there are a number of interacting partners that are heat shock proteins/chaperones that may have a role in micro autophagy which is chaperone mediated. A good number of the interacting partners are ribosomal proteins, which are known to have extra ribosomal functions and their involvement in programmed cell death like apoptosis is well marked (Warner and McIntosh 2009). However, their involvement in autophagic cell death is not known. This protein is also annotated as a conserved protein of unknown function in the PlasmoDB database.

**Repair (PFC1035w):** Most of the interacting partners of this protein show involvement in DNA repair. Interacting partners are DNA polymerase alpha and delta subunits, helicases, RAD51, Msh2p, PCNA, RFC, RPA, 3′ -5′ exonuclease, 5′-3′ exonuclease, uracil DNA glycosylase, exoribonuclease III, smc proteins (some of the smc proteins viz. smc5 and smc6 are implicated in DNA repair and checkpoint responses), Apn1 (apurinic-apyrimidinic endonuclease) and yet others annotated as DNA repair proteins. This protein is also annotated as a conserved protein of unknown function in the PlasmoDB database.

**Phospholipid metabolic process (PFB0250w):** The interacting partners are myo-inositol 1-phosphate synthase, glycerol-3-phosphate acyltransferase, and choline kinase that are related to inositol and phospholipid biosynthetic processes. The other related interacting partners are acyl carrier protein, malonyl coa-acyl carrier, phospholipase, glycogen synthase kinase, acyl-CoA synthetases, diacylglycerol O-acyltransferase (triglyceride biosynthetic process), N-acetylglucosaminyl-

phosphatidylinositol de-n-acetylase (GPI anchor biosynthesis), & phosphatidyl-inositol glycan (GPI anchor biosynthesis). This protein is annotated as a conserved protein of unknown function in PlasmoDB database.

**RNA metabolic/catabolic process (PFC1011c):** RNA metabolism is broadly defined as the compendium of all processes that involve RNA, including transcription, processing and modification of transcripts, translation and RNA degradation and its regulation (Anantharaman, Koonin, and Aravind 2002). The PFC1011c protein's possible role in RNA metabolism is quiet indicative from its interacting partners that are mostly ribosomal proteins. The other interacting proteins are, dimethyladenosine transferase (rRNA processing), Ser/Arg-rich splicing factor (mRNA processing), phosphoribosylpyrophosphate synthetase, RNA helicases, pre-mRNA splicing factors, ribosomal phosphoprotein, tRNA ligases, u6 snRNA-associated sm-like protein, cyclophilin like protein, RNAse L inhibitor protein (mRNA catabolic process), nucleoside transporter 1, U2 snRNP auxiliary factor, queuine tRNA ribosyltransferase, a homolog of Mago nashi protein (non-sense mediated mRNA decay) and a 3'-5' exoribonuclease Csl4 homolog (involved in rRNA processing), all of which are involved in RNA metabolism. PFC1011c is defined as a conserved *Plasmodium* membrane protein of unknown function in the PlasmoDB database.

**TABLE 5.4**

#### 5.3.1.2 Transaldolase - A missing link

One of the most interesting best bi-directional hits to one of the *Plasmodium* proteins was an important enzyme, Transaldolase, of the pentose shunt pathway which is assumed to be a missing link (http://sites.huji.ac.il/malaria/maps/ppcpath.html), to date (**Figure 5.1**). This protein is not a part of the interaction network and hence its interacting partners could not be studied. This protein was picked up by the Smat50 matrix against the Enterobacteriaceae member, *Wigglesworthia glossinidia brevipalpis.* Transaldolase is an enzyme of the non-oxidative phase of the pentose phosphate shunt pathway that generates R5P and links this pathway to glycolysis. Gene that codes for transaldolase could not be found in the genome of *Plasmodium falciparum* though, biochemical evidence indicates that this phase is active in the parasite (Bozdech and Ginsburg 2005).

The erythrocytic stage of the malaria parasite is engaged in intensive synthesis of nucleotides and is subjected to endogenously produced oxygen radicals that must be detoxified. Nucleic acid synthesis requires ribose-5-phosphate which is provided by the non-oxidative arm of the pentose phosphate shunt, involving the enzyme, transaldolase. Subsequently, any cell that proliferates rapidly needs large quantities of NADPH that is required for the conversion of ribonucleotides to deoxyribonucleotides. NADPH is also required for reductive antioxidant protection (Atamna, Pascarmona, and Ginsburg 1994) and is a by-product of the oxidative phase of the pentose phosphate shunt. Hence, this pathway is of utmost importance

for the parasite. The identification of the probable Transaldolase enzyme (GID 124807078) in *Plasmodium falciparum* creates an intriguing possibility for further studies. Clearly this effort will be enhanced by the biochemical characterization of the gene in question. A point to be noted here is that this protein is annotated as a rhoptry neck protein by PlasmoDB.

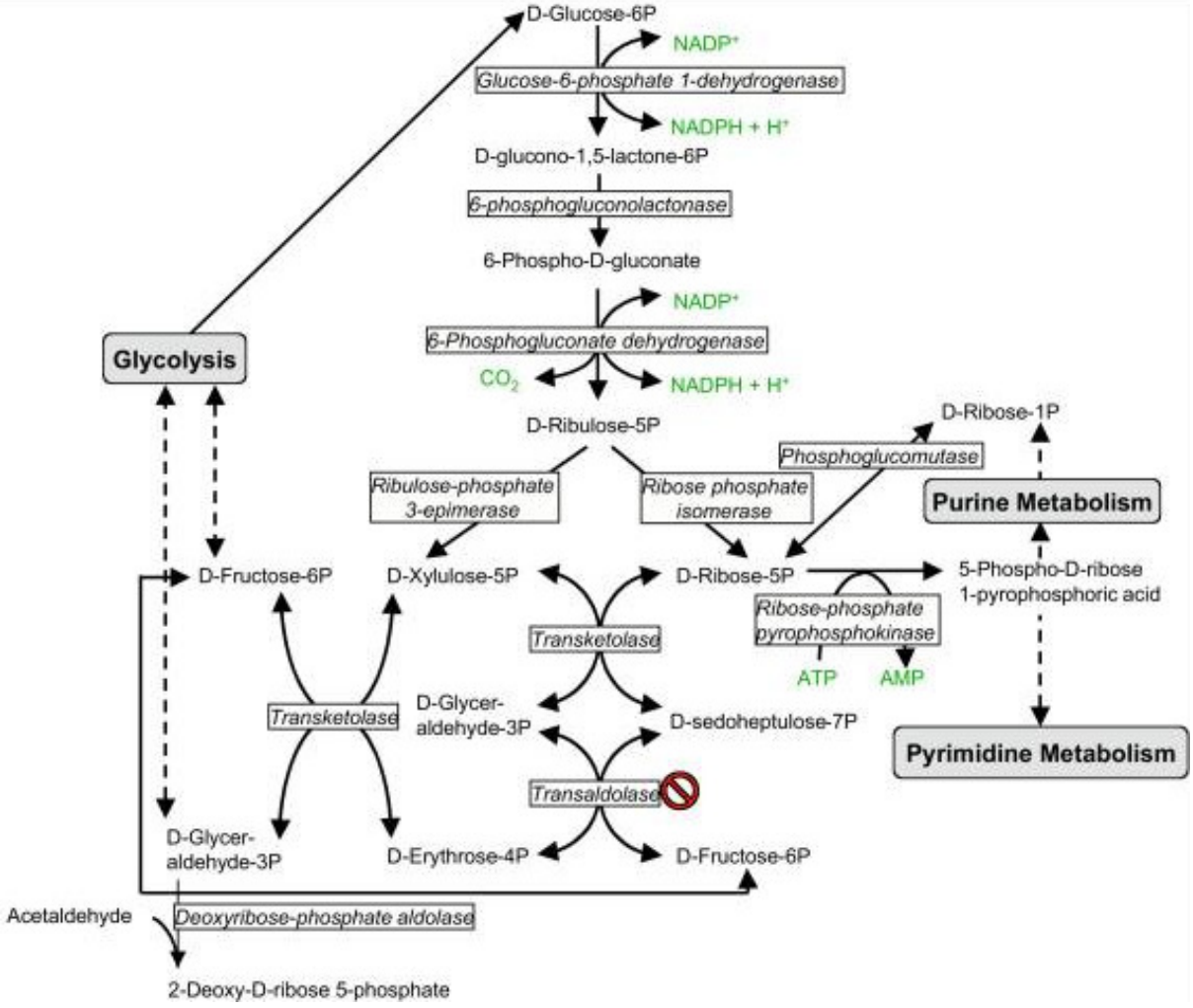


**Figure 5.1 Pentose phosphate shunt pathway of *Plasmodium falciparum* showing the missing enzyme, Transaldolase** (Adapted from Bozdech and Ginsburg 2005)

### 5.3.2 Identification of *Plasmodium falciparum* orthologs across a plant model - *Arabidopsis thaliana*

The malaria parasite is known to be evolutionarily close to the model plant, *Arabidopsis thaliana* perhaps due to the secondary endo-symbiontic origin of apicoplast (Gardner et al. 2002). A search against the *Arabidopsis* proteome was thus interesting to identify potential orthologs in this organism. A total of 465 exclusive *Plasmodium falciparum* proteins gave BBH's to plant proteins out of which 340 gave hits to annotated proteins of *Arabidopsis thaliana*. A majority of the plant specific orthologs obtained for the hypothetical proteins of *Plasmodium falciparum* were metabolic enzymes, transcription factors, and F-box proteins. F-box proteins consist of a protein motif that functions as a site for protein-protein interaction. These were first characterized as components of SCF ubiquitin-ligase complexes, in which they bind substrates for ubiquitin-mediated proteolysis. However, these have been recently discovered to function in a variety of other cellular functions (Kipreos and Pagano 2000). This reflects a possible role of these proteins in protein-protein interactions of the malaria parasite. A list of *Plasmodium falciparum* hypothetical/putative proteins that gave hits (BBH's) to annotated proteins of *Arabidopsis thaliana* has been summarized in **Table 5.5**.

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

**TABLE 5.5**

#### 5.3.2.1 Missing links in metabolic pathways

A comparison with *Arabidopsis* proteome gave clues of the possible presence of certain missing enzymes in the malaria parasite. In view of the missing links in the metabolic pathways of *Plasmodium falciparum*, the orthologs of interest were the arginine decarboxylase and glutamate-tRNA ligase of *Arabidopsis* that were best bi-directional hits to *Plasmodium falciparum* proteins, PFC0176c (gi|124504743|) and MAL8P1.1 (gi|124512174|) respectively.

**Arginine decarboxylase:** Arginine decarboxylase enzyme has not been identified yet in *Plasmodium falciparum*. However, this enzyme was identified in *Cryptosporidium* (Keithly et al. 1997), which is a close relative of the malaria parasite. Arginine is a precursor of polyamine biosynthesis which is decarboxylated to ornithine by Arginine decarboxylase in mammals. Ornithine in turn is decarboxylated by ornithine decarboxylase to putrescine. In case of trypanosomatids, plants and some bacteria, arginine is converted to putrescine via agmatine by the same enzyme, Arginine decarboxylase, in the absence of a functional ornithine decarboxylase (Ramya, Surolia, and Surolia 2006). However, the malaria parasite has a bi-functional enzyme with ornithine decarboxylase and S-adenosylmethionine decarboxylase and thus it might not take up the agmatine path for putrescine formation. The presence of polyamine biosynthesis in this parasite is also well acknowledged (**Figure 5.2**) and polyamine metabolism of *Plasmodium* is seen as a potential target for malaria chemotherapy (Muller et al. 2008). Hence, presence of Arginine decarboxylase in *Plasmodium* is not surprising.

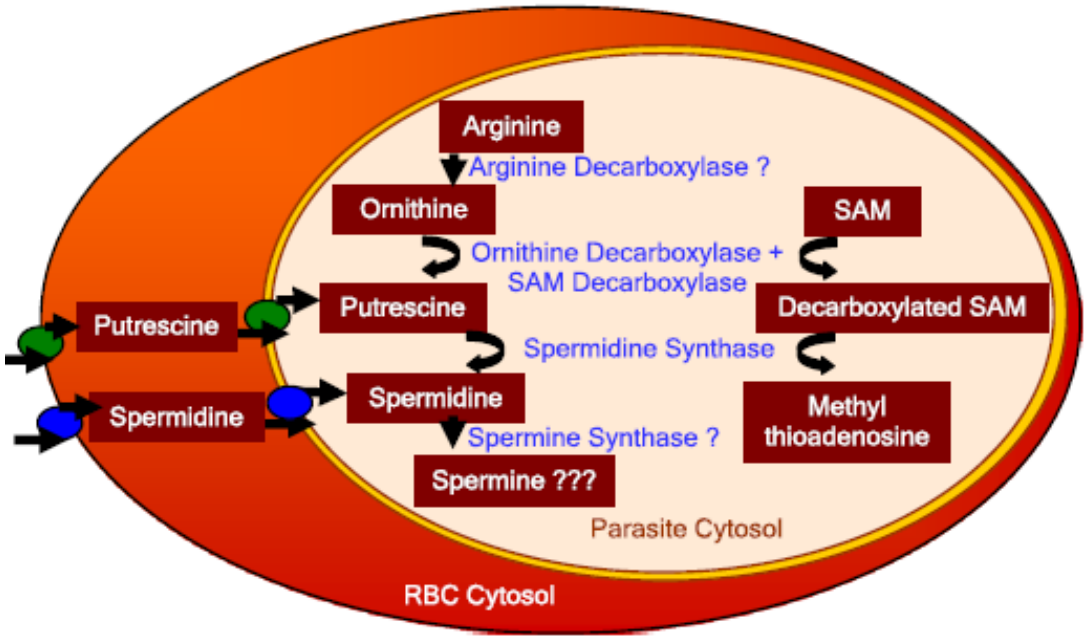


**Figure 5.2 Schematic representation of a possible path followed in polyamine biosynthesis by *Plasmodium falciparum*.** Arginine or ornithine is the precursor for the synthesis of putrescine, spermidine (and spermine) (adapted from Ramya, Surolia and Surolia, 2006).

**Glutamate-tRNA ligase:** A second missing enzyme for which an ortholog was obtained in *Plasmodium* using Smat matrices was a glutamate-tRNA ligase or the non-discriminating glutamyl-tRNA synthetase that is involved in glutamate metabolism and catalyzes the aminoacylation of glutamyl-tRNA (**Figure 5.3**). When this enzyme acts on tRNA(Glu), it catalyzes the same reaction as EC 6.1.1.17. It has, however, diminished discrimination, so that it can also form glutamate-tRNA(Gln). This is the only missing enzyme in glutamate metabolism of *Plasmodium falciparum* as listed in the malaria metabolic pathway (http://sites.huji.ac.il/malaria/ ). MAL8P1.1 seems to be a potential ortholog of *Arabidopsis* glutamate-tRNA ligase as per best bi-directional hits given by the Smat series of matrices.

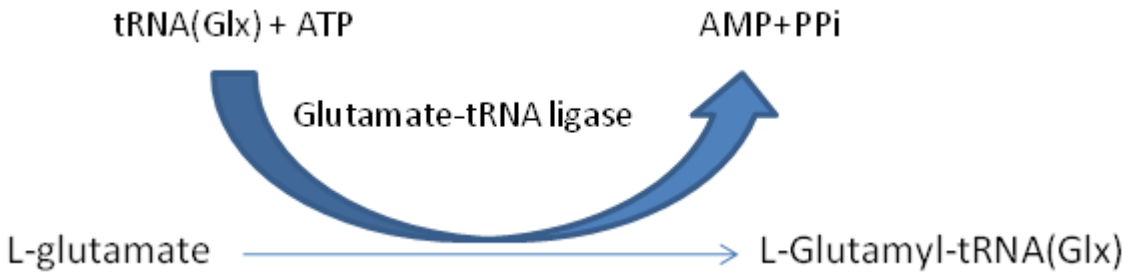


**Figure 5.3 Reaction path of glutamate-tRNA ligase, a missing enzyme of *Plasmodium falciparum* glutamate metabolism**

### 5.3.3 Identification of *Plasmodium falciparum* orthologs in *Saccharomyces cerevisiae*

A comparison with the model organism *Saccharomyces cerevisiae* gave BBH's for 256 *Plasmodium falciparum* proteins (that were either hypothetical or putative) with the Smat series. These were exclusive proteins for which no BBH's were picked with the standard matrix. A list of *Plasmodium falciparum* proteins for which annotated *Saccharomyces cerevisiae* orthologs could be picked, using BBH has been summarized in **Table 5.6**. Few proteins of interest were a cysteine protease and a scaffold protein (responsible for pre-autophagosomal structure organization) both of which are involved in autophagy in yeast. The NCBI gene identifiers of their potential orthologs in *Plasmodium falciparum* are, 124808553 and 124506779, both of which are termed conserved hypothetical proteins of unknown function in the latest version of PlasmoDB database. It is known that many host-parasite interactions are regulated in part by the programmed cell death of host cells or the parasite (Hurd and Carter 2004). Malaria parasites undergo several dramatic reorganisations during their life cycle and it is conceivable that they employ autophagy as a means to remove dispensable cellular material. To an effect a process similar to autophagy has also been recently described in the malaria parasite (Totino et al. 2008). Other hypothetical proteins of interest were; a probable centrin binding protein required for spindle pole body duplication that is required for progression through G(2)-M transition (124505607); a protein involved in DNA damage and replication checkpoint pathway (124506177); a meiosis specific protein required for double strand break repair and pairing

between homologous chromosomes (124505519). The former is listed as a homolog of reticulocyte binding protein 4, Rh4, according to which it is a rhoptry neck associated protein homolog. However, GO terms have not been assigned and information is not available for the sub-cellular localization of this protein in PlasmoDB database. The later two are conserved proteins of unknown function as per PlasmoDB latest annotations. Some of the other interesting BBH results that were obtained are discussed in the following sections.

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

# TABLE 5.6

#### 5.3.3.1 Missing links

**Polyamine oxidase:** Yet another enzyme linked to the polyamine biosynthetic pathway for which a *Plasmodium* ortholog was obtained was a polyamine oxidase. This enzyme is involved in the conversion of the polyamine, spermine to spermidine which is an essential step for the hypusination of eIF-5A. Hypusine (product of hypusination) is an unusual amino acid that is a result of a unique post-translational modification event that occurs in only one cellular protein i.e. eIF-5A. Functionally eIF-5A has possible roles in cell proliferation and apoptosis and is a characterized protein in *Plasmodium falciparum* (Molitor et al. 2004). Since the 4-aminobutyl moiety of hypusine is derived from spermidine, hypusine synthesis defines an absolute requirement for the polyamine spermidine in eukaryotes (Park et al. 2009). Though the existence of hypusine pathway in *Plasmodium falciparum* is known (Frommholz et al. 2009), yet there is no clear evidence of the presence of a polyamine oxidase (enzyme not listed in PlasmoDB database) in this parasite. The *Plasmodium* ortholog for this enzyme in yeast is the protein with the NCBI gene identifier, 124512726 which is defined as a conserved *plasmodium* protein of unknown function in PlasmoDB database (version 6.3).

**Ceramide synthase and TORC2:** Recently, sphingolipids and their precursors have attracted great attention because these compounds have been shown to play roles in cell signalling, heat stress response, calcium homeostasis, and have been implicated in the formation of specialized membrane microdomains (Vallee and Riezman 2005). Sphingolipids also function in membrane trafficking, influencing

the intracellular targeting of glycosylphosphatidylinositol-anchored proteins and regulating the internalization step of endocytosis (Vallee and Riezman 2005). Ceramide is also known to play an important role in the cell signalling pathway involved in apoptosis (Basnakian et al. 2005). These sphingolipids may either be synthesized de novo (requires ceramide synthase) or by the hydrolysis of sphingomyelins.

An interesting protein of importance in this regard was the yeast ortholog to a *Plasmodium* protein (124512152) that was a subunit of ceramide synthase (Lip1p), related to sphingolipid metabolism. This protein in yeast (Lip1p) is a single-span ER membrane protein associated with Lag1p and Lac1p and is required for ceramide synthase activity (Vallee and Riezman 2005). The null mutant in yeast is known to grow extremely slowly and is defective in ceramide synthesis. De novo synthesis of ceramide in *Plasmodium falciparum* is well acknowledged (Gerold and Schwarz 2001) and the Lag1 component is known (http://sites.huji.ac.il/malaria/maps/sphingometpath.html). Another interesting BBH (124506213) to a yeast protein was the protein subunit of TORC2. In yeast, TORC2 is known to regulate the de novo ceramide and sphingolipid synthesis, wherein TORC2 senses growth signals and activates the protein kinase, Ypk2, which then activates ceramide synthase (Dickson 2008).

**Glycosylphosphatidylinositol:** GPI or glycosylphosphatidylinositol: of parasite origin functions as the dominant malarial toxin in the context of infection (Delorenzi et al. 2002). Two related proteins were identified in *Plasmodium falciparum* (124514088

and 124506769) that were orthologs to yeast proteins, Gaa1p (6323117) and Gpi2p (6325181) respectively. The former is a subunit of the GPI:protein transamidase complex that removes the GPI-anchoring signal and attaches GPI to proteins in the ER. The later is a protein involved in the synthesis of N-acetylglucosaminyl phosphatidylinositol (GlcNAc-PI), the first intermediate in the synthesis of GPI anchors. This protein is a homolog of human PIG-C protein. While the former protein in *Plasmodium falciparum* is termed as a 'conserved membrane protein of unknown function' the second one is annotated as a putative phosphatidylinositol N-acetylglucosaminyltransferase in the latest version of PlasmoDB database.

**Phosphoribosylaminoimidazole carboxylase:** This enzyme catalyzes the conversion of 5-Aminoimidazole ribonucleotide to 5-Amino-4-carboxy-imidazole ribonucleotide, an intermediate step in the 'de novo' purine nucleotide biosynthetic pathway. An ortholog of this enzyme in yeast detected in *Plasmodium falciparum* by the BBH method was surprising because de novo purine synthesis in the malaria parasite is considered absent (Wang and Simashkevich 1981; Kicska et al. 2002). However, it is quite possible that the parasite takes up some alternative pathways and synthesizes purines de novo conditionally as observed in *Lactobacillus leichmannii* (Craven and Downing 1963).

A list of the probable *Plasmodium falciparum* enzymes identified by Smat matrices (as discussed in this chapter), that were either unidentified or missing is summarized in **Table 5.7**. However, these are preliminary results of first pass approximations and require further investigation and experimental validation for

drawing any reliable conclusions.

**Table 5.7 Missing or un-identified enzymes of *Plasmodium falciparum* identified by Smat matrices**

| Gene name | Enzyme | EC number |
|---|---|---|
| PFL2505c | Transaldolase | 2.2.1.2 |
| PFC0176c | Arginine decarboxylase | 4.1.1.19 |
| MAL8P1.1 | Glutamate tRNA ligase | 6.1.1.24 |
| MAL8P1.154 | Polyamine oxidase | 1.5.3.11 |
| MAL13P1.88 | Phosphoribosylaminoimidazole carboxylase | 4.1.1.21 |

## 5.4 Conclusion

In the present chapter, an ardent effort has been made to provide first pass approximations of the role played by the huge list of hypothetical proteins in *Plasmodium falciparum* that remain unannotated to date. It has been demonstrated in this chapter as to how the Smat series of matrices were able to pick BBHs for the hypothetical proteins of the malaria parasite for which the standard matrices failed to give hits. The substitution matrices constructed in the context of a biased genome like *Plasmodium falciparum* proved to be productive since apparent functions and probable pathways could be assigned for many unknown proteins. Some of the results correlated with the interaction data of the malaria parasite that proves the predictions to be accurate. Apart from this, some of the missing links in metabolic pathways of the parasite were identified based on the BBH method using our novel Smat matrices.

In summary, probable functions could be assigned to *Plasmodium falciparum*

hypothetical proteins for which annotations were not available or for which the standard matrices failed to give potential orthologs. Though the results from this chapter require a proper validation in the form of experiments, the clues provided in this study would certainly help an experimental biologist in the field of malaria research.

# CHAPTER 6

---

# PlasmoAlign - A sequence alignment web server for *Plasmodium falciparum*

## 6.1 Introduction

Re-annotation of the *Plasmodium falciparum* genome is ongoing even after 7 years of the genome being sequenced; the major challenge being the unique genome composition of the parasite. The organism hosts a large number of genes in its genome, whose protein products are poorly understood. Sequence analysis, in general, provides preliminary clues for functional annotation. However, the limitations remain with the unusual genome of *Plasmodium falciparum* that shows an atypical amino acid composition of proteins and is distantly related to known organisms in terms of evolution. As a result, many proteins show no clue of any sequence match to the existing proteins to infer functional annotations.

Several efforts have been made in this context, primarily being an effort towards the compositional adjustment of substitution matrices (Yu, Wootton, and Altschul 2003; Bastien, Roy, and Marechal 2005; Brick and Pizzi 2008) that would improve sequence analysis of *Plasmodium falciparum* genome. However, none of these methods claimed improvement in annotation of hypothetical proteins based on improved ortholog detection across other species. Further, the asymmetric matrix developed in this study (as described in chapter 3), for pair-wise alignment of *Plasmodium falciparum* proteins was novel unlike the above mentioned methods. The novelty of our method in constructing the substitution matrices and the improvement achieved with these in ortholog detection and pair-wise alignment of proteins, lead us to develop an alignment server for *Plasmodium falciparum* that makes use of the parasite specific matrices. The current chapter presents the

implementation of the web server, its usage and results that would be of immense interest for those in the field of malaria research.

## 6.2 Methods

PlasmoAlign is a CGI based web application written in Perl language. The server provides a web based form for input submission. The three components of PlasmoAlign are: i) a front end web interface for submitting the protein sequences or selecting genomes of choice; ii) A search engine for searching orthologs or aligning sequences; iii) a reporting system that lists out potential orthologs or gives local alignments of sequence pairs.

The webserver has two CGI applications, one being ortholog detection for *Plasmodium falciparum* proteins and the other being the option for pair-wise alignment (local alignment) of the parasite proteins with its potential orthologs. These have been described as follows.

### 6.2.1 Ortholog detection

The method of BBH (Bi-directional best hit) (Hulsen et al. 2006) has been employed for the search of potential orthologs of *Plasmodium falciparum* across a range of organisms. The parasite specific symmetric matrices i.e. the Smat series consisting of Smat50, Smat60, Smat70, Smat80 and Smat90 are provided for ortholog detection. Smat matrices have been earlier demonstrated to work best for database searches (as shown in chapter 4). For comparison, the standard BLOSUM62 matrix and the standard matrices having relative entropy similar to the Smat series of matrices is provided. Basically, the server uses the standalone version of BLASTp

program that was modified to accept the user defined matrices.

### 6.2.2 Pair-wise alignments

The pair-wise alignment option uses the fasta program (FASTA package, version 3) for performing local alignments of the *Plasmodium falciparum* query protein and its potential ortholog. The asymmetric parasite specific matrix, PfFSmat60 (PF60 in the matrix pull down menu) is provided for performing these alignments. This matrix has been demonstrated to perform best for pair-wise alignments (as is evident from chapter 4), where the alignments span motif like regions of the protein. The standard matrices PAM2, BLOSUM50 and BLOSUM100 have been provided for comparison.

## 6.3 Results and discussion

### 6.3.1 The web interface for input/output

The input to bi-directional ortholog detection is a protein sequence file of the query genome (here, *Plasmodium falciparum*) and that of the subject in the fasta format. The subject proteomes may be either selected from the list of the organisms provided in the web page or in case of a user specific sequence file, it may be uploaded through the file upload option provided for the same. The output for this feature is a list of the BBHs obtained for the queried proteins and consists of the gene identifiers for the query and that of the potential ortholog, along with the subject annotation. The results of this search are available as an html link on the browser.

In case of pair-wise alignments, the input is a single protein sequence in the

fasta format, each for the query and the subject. The query protein is expected to be a protein sequence from *Plasmodium falciparum* only. The output is a pair-wise alignment of the query and the subject, generated by an html and displayed on to the web browser.

### 6.3.2 Limitation

One of the limitations of the pair-wise alignment is that the query sequence is restricted only to *Plasmodium falciparum*, due to the uni-directionality of the substitution matrix, PfFSmat60 (here PF60). Hence, the query and subject proteins should not be reversed in their order. Large sequence files may take a longer run time in case of ortholog detection. There is a limit to the size of the sequence file, in the file upload option provided for ortholog detection, which should not exceed 5 MB.

### 6.3.3 Performance and availability

The performance of the matrices provided through this web server has been tested for the complete genome of *Plasmodium falciparum* and validated in terms of overall performance compared to the similar entropy conventional matrices.

The web server is freely available at http://www.cdfd.org.in/plasmoalign/ or http://210.212.215.203/plasmoalign/. A screen shot of the web server page for the best bi-directional ortholog detection and pair-wise alignments are provided in the figures, **Figure 6.1 and Figure 6.2** respectively.

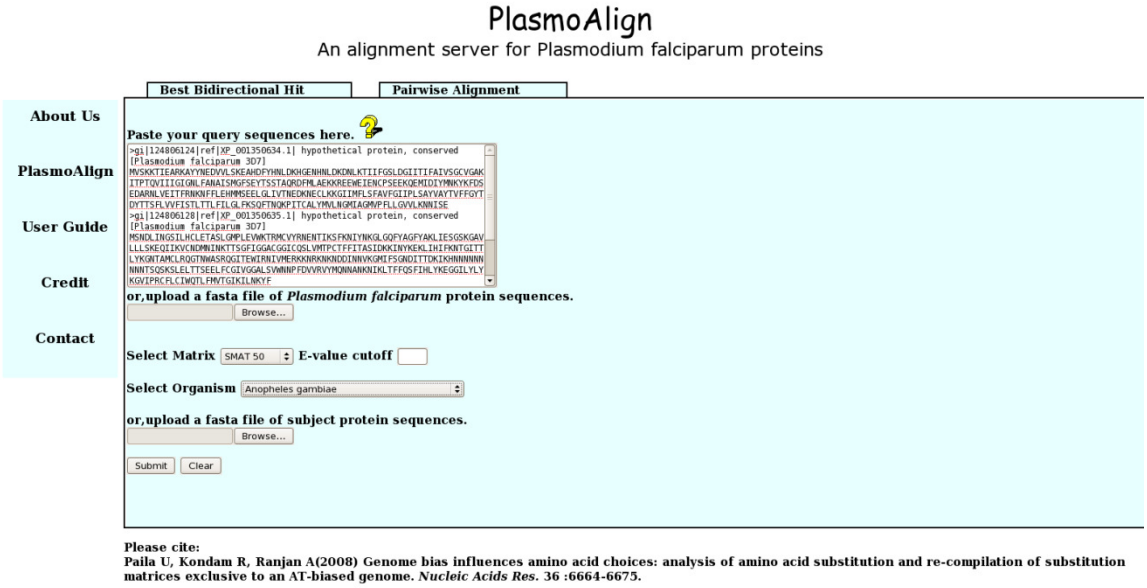


**Figure 6.1 PlasmoAlign search page for performing ortholog detection with BBH method.** The user may paste protein sequences in the fasta format for input query or may upload a sequence file in fasta format. The subject genomes and the matrices may be selected from the pull down menu. Subject protein sequences may also be uploaded in case the genome of choice is not listed.

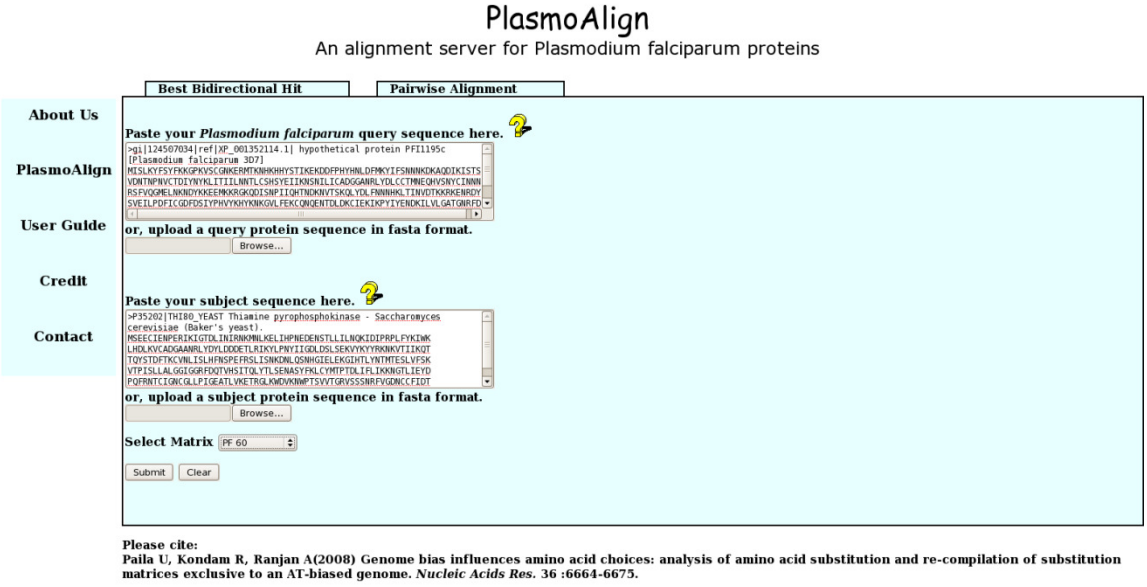


**Figure 6.2 PlasmoAlign search page for performing pair-wise alignments.** The user may either paste protein sequences in the fasta format for query and subject or may upload a sequence file in fasta format. The matrices may be selected from the pull down menu.

## 6.4 Conclusion

PlasmoAlign is an alignment web server designed specifically for *Plasmodium falciparum* proteins. The server provides potential orthologs of the parasite proteins in other organisms based on BBHs and performs pair-wise local alignments of these protein pairs, to reflect the sequence conservation. It uses the parasite specific substitution matrices developed by us. This tool is of immense value for finding orthologs of hypothetical proteins of the malaria parasite and for identifying motif like regions in the proteins of interest through pair-wise alignments.

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol **219**:555-565.

Altschul, S. F., R. Bundschuh, R. Olsen, and T. Hwa. 2001. The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res **29**:351-361.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**:3389-3402.

Altschul, S. F., J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schaffer, and Y. K. Yu. 2005. Protein database searches using compositionally adjusted substitution matrices. FEBS J **272**:5101-5109.

Anantharaman, V., E. V. Koonin, and L. Aravind. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res **30**:1427-1464.

Aravind, L., L. M. Iyer, T. E. Wellems, and L. H. Miller. 2003. Plasmodium biology: genomic gleanings. Cell **115**:771-785.

Atamna, H., G. Pascarmona, and H. Ginsburg. 1994. Hexose-monophosphate shunt activity in intact Plasmodium falciparum-infected erythrocytes and in free parasites. Mol Biochem Parasitol **67**:79-89.

Basnakian, A. G., N. Ueda, X. Hong, V. E. Galitovsky, X. Yin, and S. V. Shah. 2005. Ceramide synthase is essential for endonuclease-mediated death of renal tubular epithelial cells induced by hypoxia-reoxygenation. Am J Physiol Renal Physiol **288**:F308-314.

Bastien, O., S. Lespinats, S. Roy, K. Metayer, B. Fertil, J. J. Codani, and E. Marechal. 2004. Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. Gene **336**:163-173.

Bastien, O., S. Roy, and E. Marechal. 2005. Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. C R Biol **328**:445-453.

Becuwe, P., S. Gratepanche, M. N. Fourmaux, J. Van Beeumen, B. Samyn, O. Mercereau-Puijalon, J. P. Touzel, C. Slomianny, D. Camus, and D. Dive. 1996. Characterization of iron-dependent endogenous superoxide dismutase of Plasmodium falciparum. Mol Biochem Parasitol **76**:125-134.

Benner, S. A., M. A. Cohen, and G. H. Gonnet. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng **7**:1323-1332.

Berkhout, B., and F. J. van Hemert. 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res **22**:1705-1711.

Bernardi, G. 1986. Compositional constraints and genome evolution. J Mol Evol **24**:1-11.

Bosch, J., C. A. Buscaglia, B. Krumm, B. P. Ingason, R. Lucas, C. Roach, T. Cardozo, V. Nussenzweig, and W. G. Hol. 2007. Aldolase provides an unusual binding site

for thrombospondin-related anonymous protein in the invasion machinery of the malaria parasite. Proc Natl Acad Sci U S A **104**:7015-7020.

Boucher, I. W., A. M. Brzozowski, J. A. Brannigan, C. Schnick, D. J. Smith, S. A. Kyes, and A. J. Wilkinson. 2006a. The crystal structure of superoxide dismutase from Plasmodium falciparum. BMC Struct Biol **6**:20.

Boucher, I. W., P. J. McMillan, M. Gabrielsen, S. E. Akerman, J. A. Brannigan, C. Schnick, A. M. Brzozowski, A. J. Wilkinson, and S. Muller. 2006b. Structural and biochemical characterization of a mitochondrial peroxiredoxin from Plasmodium falciparum. Mol Microbiol **61**:948-959.

Bozdech, Z., and H. Ginsburg. 2004. Antioxidant defense in Plasmodium falciparum--data mining of the transcriptome. Malar J **3**:23.

Bozdech, Z., and H. Ginsburg. 2005. Data mining of the transcriptome of Plasmodium falciparum: the pentose phosphate pathway and ancillary processes. Malar J **4**:17.

Bradley, D. J. 1998. The particular and the general. Issues of specificity and verticality in the history of malaria control. Parassitologia **40**:5-10.

Breman, J. G. 2001. The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. Am J Trop Med Hyg **64**:1-11.

Brick, K., and E. Pizzi. 2008. A novel series of compositionally biased substitution matrices for comparing Plasmodium proteins. BMC Bioinformatics **9**:236.

Brooks, D. J., and J. R. Fresco. 2002. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol Cell Proteomics **1**:125-131.

Certa, U., P. Ghersa, H. Dobeli, H. Matile, H. P. Kocher, I. K. Shrivastava, A. R. Shaw, and L. H. Perrin. 1988. Aldolase activity of a Plasmodium falciparum protein with protective properties. Science **240**:1036-1038.

Certa, U., C. Itin, and H. Dobeli. 1992. In vitro mutagenesis defines drug targets in aldolase of Plasmodium falciparum. Mem Inst Oswaldo Cruz **87 Suppl 3**:263-264.

Collins, D. W., and T. H. Jukes. 1993. Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J Mol Evol **36**:201-213.

Craven, G. R., and M. Downing. 1963. Vitamin B-12 and purine metabolism in Lactobacillus leichmannii. Glycine-2-C-14 incorporation into ribonucleic and deoxyribonucleic acid. J Biol Chem **238**:1464-1466.

D'Onofrio, G., D. Mouchiroud, B. Aissani, C. Gautier, and G. Bernardi. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol **32**:504-510.

Date, S. V., and C. J. Stoeckert, Jr. 2006. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. Genome Res **16**:542-549.

Dayhoff, M. O. 1978. Atlas of protein sequence and structure
National Biomedical Research Foundation, Washington, DC.

Delorenzi, M., A. Sexton, H. Shams-Eldin, R. T. Schwarz, T. Speed, and L. Schofield. 2002. Genes for glycosylphosphatidylinositol toxin biosynthesis in Plasmodium falciparum. Infect Immun **70**:4510-4522.

Deponte, M. 2007. Peroxiredoxin Systems. Springer.

Dickson, R. C. 2008. More chores for TOR: de novo ceramide synthesis. Cell Metab **7**:99-100.

Dobeli, H., A. Trzeciak, D. Gillessen, H. Matile, I. K. Srivastava, L. H. Perrin, P. E. Jakob, and U. Certa. 1990. Expression, purification, biochemical characterization and inhibition of recombinant Plasmodium falciparum aldolase. Mol Biochem Parasitol **41**:259-268.

Doolittle, R. F. 2002. The grand assault. Nature **419**:493-494.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics **14**:755-763.

Escalante, A. A., and F. J. Ayala. 1995. Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci U S A **92**:5793-5797.

Fairfield, A. S., S. R. Meshnick, and J. W. Eaton. 1983. Malaria parasites adopt host cell superoxide dismutase. Science **221**:764-766.

Fast, N. M., J. C. Kissinger, D. S. Roos, and P. J. Keeling. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol Biol Evol **18**:418-426.

Feng, D. F., and R. F. Doolittle. 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Methods Enzymol **266**:368-382.

Fichera, M. E., and D. S. Roos. 1997. A plastid organelle as a drug target in apicomplexan parasites. Nature **390**:407-409.

Florens, L., M. P. Washburn, J. D. Raine, R. M. Anthony, M. Grainger, J. D. Haynes, J. K. Moch, N. Muster, J. B. Sacci, D. L. Tabb, A. A. Witney, D. Wolters, Y. Wu, M. J. Gardner, A. A. Holder, R. E. Sinden, J. R. Yates, and D. J. Carucci. 2002. A proteomic view of the Plasmodium falciparum life cycle. Nature **419**:520-526.

Friedberg, I., T. Harder, R. Kolodny, E. Sitbon, Z. Li, and A. Godzik. 2007. Using an alignment of fragment strings for comparing protein structures. Bioinformatics **23**:e219-224.

Frommholz, D., P. Kusch, R. Blavid, H. Scheer, J. M. Tu, K. Marcus, K. H. Zhao, V. Atemnkeng, J. Marciniak, and A. E. Kaiser. 2009. Completing the hypusine pathway in Plasmodium. FEBS J **276**:5881-5891.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature **419**:498-511.

Gerold, P., and R. T. Schwarz. 2001. Biosynthesis of glycosphingolipids de-novo by the human malaria parasite Plasmodium falciparum. Mol Biochem Parasitol **112**:29-37.

Ginsburg, H. 2006. Progress in in silico functional genomics: the malaria Metabolic Pathways database. Trends Parasitol **22**:238-240.

Goldman, N., and S. Whelan. 2002. A novel use of equilibrium frequencies in models of

sequence evolution. Mol Biol Evol **19**:1821-1831.
Gonnet, G. H., M. A. Cohen, and S. A. Benner. 1992. Exhaustive matching of the entire protein sequence database. Science **256**:1443-1445.
Gough, J., K. Karplus, R. Hughey, and C. Chothia. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol **313**:903-919.
Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862-864.
Hartl, D. L. 2004. The origin of malaria: mixed messages from genetic diversity. Nat Rev Microbiol **2**:15-22.
Hastings, I. M., P. G. Bray, and S. A. Ward. 2002. Parasitology. A requiem for chloroquine. Science **298**:74-75.
He, C. Y., B. Striepen, C. H. Pletcher, J. M. Murray, and D. S. Roos. 2001. Targeting and processing of nuclear-encoded apicoplast proteins in plastid segregation mutants of Toxoplasma gondii. J Biol Chem **276**:28436-28442.
Hemingway, J., L. Field, and J. Vontas. 2002. An overview of insecticide resistance. Science **298**:96-97.
Hemingway, J., and H. Ranson. 2000. Insecticide resistance in insect vectors of human disease. Annu Rev Entomol **45**:371-391.
Henikoff, S., and J. G. Henikoff. 2000. Amino acid substitution matrices. Adv Protein Chem **54**:73-97.
Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A **89**:10915-10919.
Henikoff, S., and J. G. Henikoff. 1991. Automated assembly of protein blocks for database searching. Nucleic Acids Res **19**:6565-6572.
Hulsen, T., M. A. Huynen, J. de Vlieg, and P. M. Groenen. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol **7**:R31.
Hurd, H., and V. Carter. 2004. The role of programmed cell death in Plasmodium-mosquito interactions. Int J Parasitol **34**:1459-1472.
Hyman, R. W., E. Fung, A. Conway, O. Kurdi, J. Mao, M. Miranda, B. Nakao, D. Rowley, T. Tamaki, F. Wang, and R. W. Davis. 2002. Sequence of Plasmodium falciparum chromosome 12. Nature **419**:534-537.
Johnson, M. S., and J. P. Overington. 1993. A structural basis for sequence comparisons. An evaluation of scoring methodologies. J Mol Biol **233**:716-738.
Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.
Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994. A mutation data matrix for transmembrane proteins. FEBS Lett **339**:269-275.
Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. 2005. A universal trend of amino acid gain and loss in protein evolution. Nature **433**:633-638.
Joy, D. A., X. Feng, J. Mu, T. Furuya, K. Chotivanich, A. U. Krettli, M. Ho, A. Wang, N. J. White, E. Suh, P. Beerli, and X. Z. Su. 2003. Early origin and recent expansion of Plasmodium falciparum. Science **300**:318-321.

Karlin, S., and S. F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87**:2264-2268.

Kawazu, S., N. Ikenoue, H. Takemae, K. Komaki-Yasuda, and S. Kano. 2005. Roles of 1-Cys peroxiredoxin in haem detoxification in the human malaria parasite Plasmodium falciparum. FEBS J **272**:1784-1791.

Keithly, J. S., G. Zhu, S. J. Upton, K. M. Woods, M. P. Martinez, and N. Yarlett. 1997. Polyamine biosynthesis in Cryptosporidium parvum and its implications for chemotherapy. Mol Biochem Parasitol **88**:35-42.

Kicska, G. A., P. C. Tyler, G. B. Evans, R. H. Furneaux, V. L. Schramm, and K. Kim. 2002. Purine-less death in Plasmodium falciparum induced by immucillin-H, a transition state analogue of purine nucleoside phosphorylase. J Biol Chem **277**:3226-3231.

Kim, H., U. Certa, H. Dobeli, P. Jakob, and W. G. Hol. 1998. Crystal structure of fructose-1,6-bisphosphate aldolase from the human malaria parasite Plasmodium falciparum. Biochemistry **37**:4388-4396.

Kipreos, E. T., and M. Pagano. 2000. The F-box protein family. Genome Biol **1**:REVIEWS3002.

Knapp, B., E. Hundt, and H. A. Kupper. 1990. Plasmodium falciparum aldolase: gene structure and localization. Mol Biochem Parasitol **40**:1-12.

Kohler, S., C. F. Delwiche, P. W. Denny, L. G. Tilney, P. Webster, R. J. Wilson, J. D. Palmer, and D. S. Roos. 1997. A plastid of probable green algal origin in Apicomplexan parasites. Science **275**:1485-1489.

Kuzniar, A., R. C. van Ham, S. Pongor, and J. A. Leunissen. 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet **24**:539-551.

Lasonder, E., Y. Ishihama, J. S. Andersen, A. M. Vermunt, A. Pain, R. W. Sauerwein, W. M. Eling, N. Hall, A. P. Waters, H. G. Stunnenberg, and M. Mann. 2002. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature **419**:537-542.

Limviphuvadh, V., Okuno, Y., Katayama, T., Goto, S., Yoshizawa, A.C., and Kanehisa, M. 2003. Metabolic pathway reconstruction for malaria parasite Plasmodium falciparum. Genome Informatics **14**:368-369.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol **34**:153-162.

Loomis, W. F., and D. W. Smith. 1990. Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. Proc Natl Acad Sci U S A **87**:9093-9097.

McConkey, G. A., J. W. Pinney, D. R. Westhead, K. Plueckhahn, T. B. Fitzpatrick, P. Macheroux, and B. Kappes. 2004. Annotating the Plasmodium genome and the enigma of the shikimate pathway. Trends Parasitol **20**:60-65.

McDermott, J., and R. Samudrala. 2004. Enhanced functional information from predicted protein networks. Trends Biotechnol **22**:60-62; discussion 62-63.

McFadden, G. I., M. E. Reith, J. Munholland, and N. Lang-Unnasch. 1996. Plastid in

human parasites. Nature **381**:482.
Mendis, K., B. J. Sina, P. Marchesini, and R. Carter. 2001. The neglected burden of Plasmodium vivax malaria. Am J Trop Med Hyg **64**:97-106.
Merckx, A., K. Le Roch, M. P. Nivez, D. Dorin, P. Alano, G. J. Gutierrez, A. R. Nebreda, D. Goldring, C. Whittle, S. Patterson, D. Chakrabarti, and C. Doerig. 2003. Identification and initial characterization of three novel cyclin-related proteins of the human malaria parasite Plasmodium falciparum. J Biol Chem **278**:39839-39850.
Miyata, T., S. Miyazawa, and T. Yasunaga. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol **12**:219-236.
Mohana Rao, J. K. 1987. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Pept Protein Res **29**:276-281.
Molitor, I. M., S. Knobel, C. Dang, T. Spielmann, A. Allera, and G. M. Konig. 2004. Translation initiation factor eIF-5A from Plasmodium falciparum. Mol Biochem Parasitol **137**:65-74.
Mueller, I., P. A. Zimmerman, and J. C. Reeder. 2007. Plasmodium malariae and Plasmodium ovale--the "bashful" malaria parasites. Trends Parasitol **23**:278-283.
Muller, I. B., R. Das Gupta, K. Luersen, C. Wrenger, and R. D. Walter. 2008. Assessing the polyamine metabolism of Plasmodium falciparum as chemotherapeutic target. Mol Biochem Parasitol **160**:1-7.
Muller, T., S. Rahmann, and M. Rehmsmeier. 2001. Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics **17 Suppl 1**:S182-189.
Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J Comput Biol **7**:761-776.
Musto, H., H. Rodriguez-Maseda, and G. Bernardi. 1995. Compositional properties of nuclear genes from Plasmodium falciparum. Gene **152**:127-132.
Musto, H., H. Romero, A. Zavala, K. Jabbari, and G. Bernardi. 1999. Synonymous codon choices in the extremely GC-poor genome of Plasmodium falciparum: compositional constraints and translational selection. J Mol Evol **49**:27-35.
Muto, A., and S. Osawa. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A **84**:166-169.
Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48**:443-453.
Ng, P. C., J. G. Henikoff, and S. Henikoff. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics **16**:760-766.
Park, M. H., K. Nishimura, C. F. Zanelli, and S. R. Valentini. 2009. Functional significance of eIF5A and its hypusine modification in eukaryotes. Amino Acids.
Pavithra, S. R., R. Kumar, and U. Tatu. 2007. Systems analysis of chaperone networks in the malarial parasite Plasmodium falciparum. PLoS Comput Biol **3**:1701-1715.
Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85**:2444-2448.
Porter, T. D. 1995. Correlation between codon usage, regional genomic nucleotide

composition, and amino acid composition in the cytochrome P-450 gene superfamily. Biochim Biophys Acta **1261**:394-400.

Przyborski, J., and M. Lanzer. 2004. Parasitology. The malarial secretome. Science **306**:1897-1898.

Ramya, T. N., N. Surolia, and A. Surolia. 2006. Polyamine synthesis and salvage pathways in the malaria parasite Plasmodium falciparum. Biochem Biophys Res Commun **348**:579-584.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**:276-277.

Risler, J. L., M. O. Delorme, H. Delacroix, and A. Henaut. 1988. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J Mol Biol **204**:1019-1029.

Roos, D. S., M. J. Crawford, R. G. Donald, J. C. Kissinger, L. J. Klimczak, and B. Striepen. 1999. Origin, targeting, and function of the apicomplexan plastid. Curr Opin Microbiol **2**:426-432.

Rost, B. 1999. Twilight zone of protein sequence alignments. Protein Eng **12**:85-94.

Shi, J., T. L. Blundell, and K. Mizuguchi. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol **310**:243-257.

Sidhu, A. B., D. Verdier-Pinard, and D. A. Fidock. 2002. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. Science **298**:210-213.

Singer, G. A., and D. A. Hickey. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol **17**:1581-1588.

Sjolander, K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics **20**:170-179.

Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. J Mol Biol **147**:195-197.

Stoebe, B., and K. V. Kowallik. 1999. Gene-cluster analysis in chloroplast genomics. Trends Genet **15**:344-347.

Sueoka, N. 1961. Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harb Symp Quant Biol **26**:35-43.

Sutormin, R. A., A. B. Rakhmaninova, and M. S. Gelfand. 2003. BATMAS30: amino acid substitution matrix for alignment of bacterial transporters. Proteins **51**:85-95.

Totino, P. R., C. T. Daniel-Ribeiro, S. Corte-Real, and M. de Fatima Ferreira-da-Cruz. 2008. Plasmodium falciparum: erythrocytic stages die by autophagic-like cell death under drug pressure. Exp Parasitol **118**:478-486.

Vallee, B., and H. Riezman. 2005. Lip1p: a novel subunit of acyl-CoA ceramide synthase. EMBO J **24**:730-741.

Veerassamy, S., A. Smith, and E. R. Tillier. 2003. A transition probability model for amino acid substitutions from blocks. J Comput Biol **10**:997-1010.

Vindal, V., S. Ranjan, and A. Ranjan. 2007. In silico analysis and characterization of GntR family of regulators from Mycobacterium tuberculosis. Tuberculosis (Edinb) **87**:242-247.

Vindal, V., K. Suma, and A. Ranjan. 2007. GntR family of regulators in Mycobacterium

smegmatis: a sequence and structure based characterization. BMC Genomics **8**:289.

Vingron, M., and M. S. Waterman. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol **235**:1-12.

Vogt, G., T. Etzold, and P. Argos. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol **249**:816-831.

Wang, C. C., and P. M. Simashkevich. 1981. Purine metabolism in the protozoan parasite Eimeria tenella. Proc Natl Acad Sci U S A **78**:6618-6622.

Warner, J. R., and K. B. McIntosh. 2009. How common are extraribosomal functions of ribosomal proteins? Mol Cell **34**:3-11.

Wilson, R. J. 2002. Progress with parasite plastids. J Mol Biol **319**:257-274.

Wilson, R. J., P. W. Denny, P. R. Preiser, K. Rangachari, K. Roberts, A. Roy, A. Whyte, M. Strath, D. J. Moore, P. W. Moore, and D. H. Williamson. 1996. Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum. J Mol Biol **261**:155-172.

Winzeler, E. A. 2008. Malaria research in the post-genomic era. Nature **455**:751-756.

Wirth, D. F. 2002. Biological revelations. Nature **419**:495-496.

Wootton, J. C., X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu, D. I. Baruch, A. J. Magill, and X. Z. Su. 2002. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature **418**:320-323.

Yu, Y. K., and S. F. Altschul. 2005. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics **21**:902-911.

Yu, Y. K., J. C. Wootton, and S. F. Altschul. 2003. The compositional adjustment of amino acid substitution matrices. Proc Natl Acad Sci U S A **100**:15688-15693.