In Silico analysis of malaria parasite databanks for specific genes and motifs associated with immune evasion

by

Jaco de Ridder

2001

Submitted in partial fulfillment of the degree M.Sc. (Biochemistry)

in the Department of Biochemistry
Faculty of Biological and Agricultural Sciences
University of Pretoria

ACKNOWLEDGEMENTS

My thanks to

My Father above
My wife Cecilia for her patience and love
My parents for their support and love
My promotor Prof Louw for all his ideas and support
The head of the Department Prof Neitz for his encouragement
My fellow students for their friendship and support

CONTENTS

ABBREVIATIONS

A ..... Adenine

AMA..... Apical Merozoite Antigen

BLAST ..... Basic Local Alignment Search Tool

C ..... Cytosine

cDNA ..... Complementary Deoxyribonucleic Acid

CDS..... Coding Sequence

CSP ..... Circumsporoziote Protein

DDT ..... 1,1,1-Trichloro-2,2-bis-(p-chlorophenyl)ethane

DHFR-TS..... Dihydrofolate Reductase - Thymidylate Synthase

DNA ..... Deoxyribonucleic Acid

EMBL..... European Molecular Biology Laboratory

EST..... Expressed Sequence Tag

ExPaSy..... Expert Protein Analysis System

FASTA..... Fast-align

G..... Guanine

GB ..... Genbank

GST ..... Genomic Sequence Tag

HRP ..... Histidine-rich Protein

html..... HyperText Markup Language

http..... Hypertext Transfer Protocol

LAMA..... Local Alignments of Multiple Alignments

MDR ..... Multi-drug Resistant

mRNA..... Messenger Ribonucleic Acid

MSA..... Merozoite Surface Antigen

O/E ..... Observed over Expected Ratio

ORF ..... Open Reading Frame

P. falciparum..... *Plasmodium falciparum* - causative agent of cerebral malaria

PCR ..... Polymerase Chain Reaction

PfEMP ..... *Plasmodium falciparum* Erythrocyte Membrane Protein

PSSM ..... Position-specific Scoring Matrix

RAPD..... Random Amplified Polymorphic DNA

RESA..... Ring-infected Surface Antigen

RNA ..... Ribonucleic Acid

RT-PCR ..... Reverse Transcriptase Polymerase Chain Reaction

SP..... SwissProt

SPf66..... Vaccine candidate derived from the 35, 55 and 83kDa malaria antigens

T ..... Thymine

TIGR ..... Tulane Institute for Genomic Research

WHO..... World Health Organization

WWW ..... World Wide Web

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

GENERAL INTRODUCTION

1.1 History of malaria

Malaria parasites have been with us since the dawn of time. Hippocrates was the first, in the fifth century BC, to describe in detail the clinical picture of malaria and some complications of the disease and relate them to the time of year and where the patients lived. The association with stagnant waters (breeding grounds for its *Anopheline* vector) led the Romans to begin drainage programs, the first intervention against malaria. The first recorded treatment dates back to 1600, where the native Peruvian Indians used the bitter bark of the cinchona tree in Peru (Desowitz, 1991). By 1649, the bark was available in England, as 'Jesuits powder' so that those suffering from 'agues' (malaria) might benefit from the quinine it contained. Malaria in the United Kingdom could have been clustered around stagnant marshes, and the invading Roman soldiers could certainly have brought the disease with them. There have been no recent cases of infective mosquito bites within the UK so this is certainly not an endemic region any more (Bruce-Chwatt, 1985).

It was not until the 1880s, that the protozoal cause of malaria was elucidated. The French army surgeon, Charles Louis Alphonse Laveran was the first to observe the malaria parasite in the blood of an Algerian patient. Only in 1897 was the *Anopheles* mosquito demonstrated to be the vector for the disease. At this point, the major features of the epidemiology of malaria seemed clear, and control measures started to be implemented (Brown *et al*, 1986).

1.2 Life cycle and vector

Malaria is a protozoal infection transmitted to human beings by female mosquitoes. In total there are nearly 120 species of *Plasmodia*, including at least 22 found in primate hosts, and 19 in rodents, bats and other mammals. About 70 other *Plasmodia* species have been described in birds and reptiles. Human malaria is caused by four species of *Plasmodium*, namely *P. falciparum, P. vivax, P. ovalae* and *P. malariae.* These four species of human malaria parasites differ morphologically, immunologically, in geographical distribution, relapse pattern and drug response. Of the four human malaria parasites, *Plasmodium falciparum* is the most virulent and causes the most fatalities.

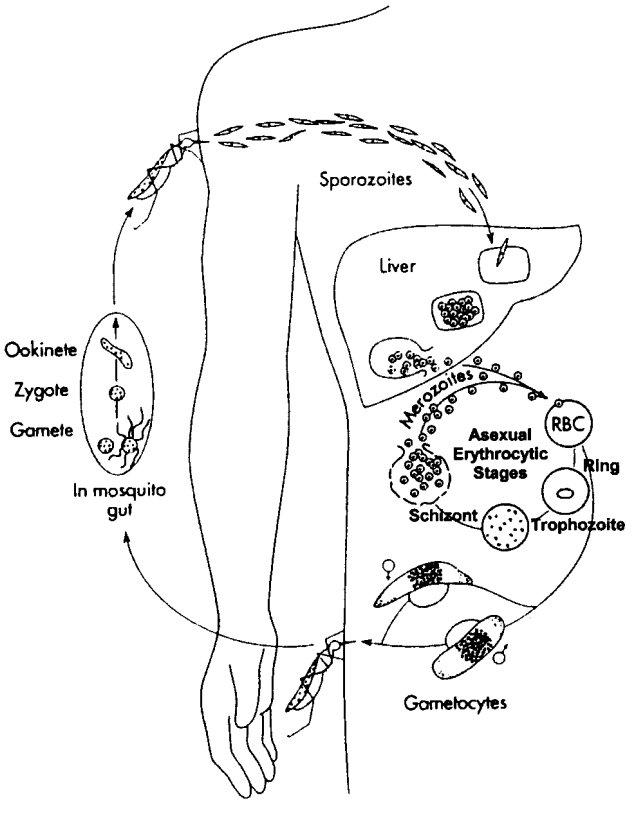

Figure 1.1: The life cycle of the human malaria parasite, *Plasmodium spp.* (Adapted from Good *et al*, 1998)

The malarial life cycle (Figure 1.1) is split between a vertebrate host and an insect vector. These four *Plasmodium* species, with the exception of *P. malariae* (which may also affect the higher primates) are parasites exclusive to man (Good *et al*, 1998).

The mosquito is always the vector, and is always an *Anopheline* mosquito, although, out of the 380 species of *Anopheline* mosquito, only 60 can transmit malaria, some 30 of which are of major importance. Only female mosquitoes are involved, as the males do not feed on blood (Good *et al*, 1998).

1.3 The epidemiology of the disease

Approximately 300 million people world-wide are affected by malaria and between 1 and 1.5 million people die from it every year. Previously extremely widespread, the malaria is now mainly confined to Africa, Asia and Latin America. Inadequate health structures and poor socio-economic conditions aggravate the problems of controlling malaria in these countries. The situation has become even more complex over the last few years with the increase in drug-resistant parasites (Greenwood, 1999).

Malaria is currently endemic in 91 countries with small pockets of transmission occurring in a further eight countries. *Plasmodium falciparum* is the predominant parasite. The following map indicates the worldwide distribution of indigenous malaria according to the World Health Organisation (Greenwood, 1999).

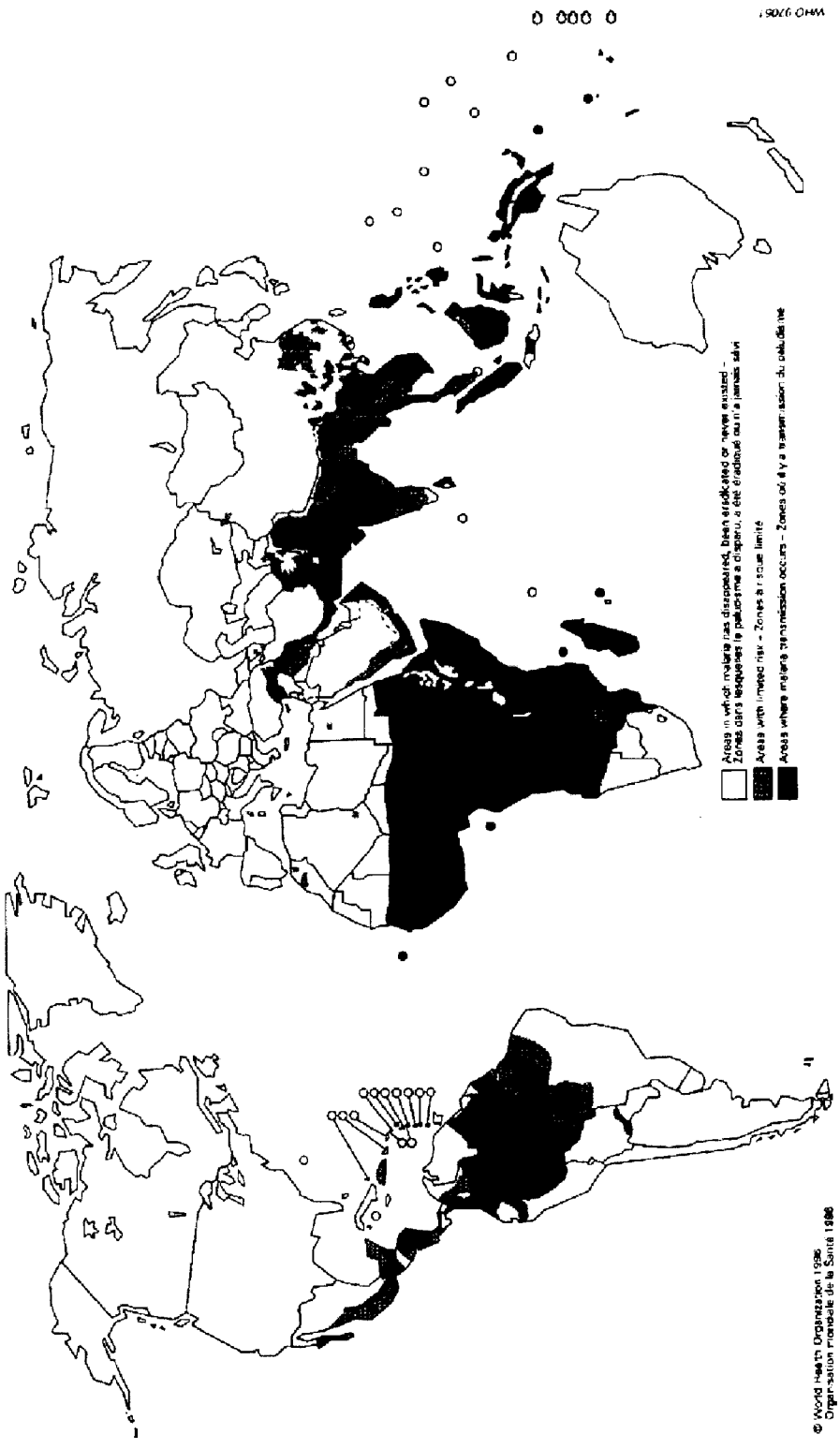

Figure 1.2: World map indicating the epidemiological assessment of the status of malaria in 1994 (WHO, 1997).

Even in South Africa the areas of high risk is on the increase (Figure 1.3), following the spread of malaria strains resistant to the present prophylactic drugs.

Figure 1.3: The spread of Malaria in South Africa (MRC, 1996).

The significance of malaria as a health problem is increasing in many parts of the world. Epidemics are even occurring around traditionally endemic zones in areas where transmission have been eliminated. These outbreaks are generally associated with deteriorating social and economic conditions, and main victims are underprivileged rural populations. Demographic, economic and political pressures compel entire populations (seasonal workers, nomadic tribes and farmers migrating to newly developed urban areas or new agricultural and economic developments) to leave malaria free areas and move into endemic zones. Non-immune people are at high risk of severe disease. Unfortunately, these population movements and the

intensive urbanisation are not always accompanied by adequate development of sanitation and health care. In many areas conflict, economic crises and administrative disorganisation can result in the disruption of health services. The absence of adequate health services frequently results in recourse to self-administration of drugs often with incomplete treatment. This is a major factor in the increase in resistance of malaria parasites to previously effective drugs (WHO, 1997).

Malaria is a complex disease but it is curable and preventable. Lives can be saved by early detection and adequate treatment of the disease. The actions that are necessary to prevent the disease and to avoid or contain epidemics and other critical situations are known. The technology to prevent, monitor, diagnose and treat malaria exists. It needs to be adapted to local conditions and to be applied through local and national malaria control programmes.

1.5 Strategies of combating malaria

There are two main strategies in combating malaria. The first is by disrupting the transmission of the disease i.e. vector control. In controlling the mosquito vector, the most important method still is the spraying of insecticides in the mosquito breeding grounds as well as houses of people living in high-risk endemic areas.

The discovery of the insecticide DDT in 1942 and its first use in Italy in 1944 raised the hope that the global eradication of malaria is possible. Subsequently, widespread systematic control measures such as spraying with DDT, coating marshes with paraffin (to block *Anopheles* mosquito larvae spiracles), draining stagnant water, and the widespread use of nets and cheap, effective drugs such as chloroquine were implemented - with impressive results (Brown *et al*, 1986). The hope of global eradication of malaria was however finally abandoned in 1969 when it was recognised that this was unlikely ever to be achieved due to the increase in resistance of

the vector to the insecticides used (Good *et al*, 1998). Ongoing control programs thus remain essential in endemic areas (WHO, 1997). Although insecticides like pyrethroids have been developed, that are safer, more effective and cheaper than DDT, the eradication of the disease appears to be unlikely (Collins *et al*, 1995).

The second strategy involves combating the parasite itself. The preferred method of treatment would be vaccination. However, an effective, practical, and affordable vaccine has yet to be developed. There are three groups of vaccine candidates being investigated at present. The first of these is pre-erythrocytic vaccines, directed against the sporozoite and hepatic stages of the parasite. These vaccines directed against the circumsporozoite protein (CSP-1 and CSP-2; Dame *et al*, 1984) and irradiated sporozoites were designed to prevent the invasion of red blood cells so that the individual does not contract the disease. Any mosquito feeding on such an individual also does not become infected and the spread of the disease is thus being reduced. Such a vaccine would however not protect against invasion by merozoites, for example by blood transfusion (Brown, 1992).

The second group is directed against the asexual blood stage antigens on merozoites or infected erythrocytes. Antigens like the merozoite surface antigen (MSA-1, MSA-2, MSA-3), apical merozoite antigen (AMA-1), erythrocyte membrane protein (PFEMP-3) and the histidine rich protein (HRP-2) are targeted (Mitchell, 1989). An anti-merozoite vaccine should reduce illness and infection by blocking the reinfection of erythrocytes by merozoites, interrupting the asexual life cycle. The patient would however remain vulnerable to sporozites and gametocytes.

The third and last type of vaccination strategy is directed at the sexual stage antigens. This type of transmission-blocking vaccine should prevent gametocytes from sexual reproduction, preventing the transmission of the disease via mosquitoes, but would not prevent the disease

itself. It is unlikely that such vaccines will be used on their own but rather as a component of a multi-stage vaccine cocktail (Good *et al*, 1998).

One of the more promising vaccine candidates, Manuel Patarroyo's synthetic peptide vaccine, SPf66, derived from the 35, 55 and 83kDa antigens and linked by the well-known asparagine-alanine-asparagine-proline (NANP) motif, has given highly varying results, giving poor antibody responses to the 35 and 83 kDa peptides as well as Spf66 (Amador *et al*, 1996; Ash, 1991).

The only viable alternative to a vaccine is still chemotherapeutic drugs used as prophylaxis and curative treatment. On the chemotherapeutic side however, the picture is also getting bleaker along with the increase in resistance to the present spectrum of antimalarial drugs (Basco *et al*, 1995a; Borst *et al*, 1995b). The need to design novel drugs against which resistance will not as easily be obtained, is greater than ever. Another novel method of combating malaria that is currently being investigated is the use of antisense therapy. Here highly specific oligonucleotides are employed in an attempt to block the synthesis of various essential parasite proteins (Stein, 1996).

1.6 Drugs - action and effectiveness

For a drug to be effective, it must be present in appropriate concentrations at its site of action. There are various factors influencing the effective concentration of a drug. Firstly, it is a factor of the dose administered and secondly, it also depends on the extent and rate of its absorption, distribution, binding, biotransformation and excretion (Benet *et al*, 1985).

Antimalarial drugs can be divided into three basic groups according to their mechanism of action. The first is the quinine-type of derivatives. Drugs in this group include the older agents

like quinine, primaquine and chloroquine, in addition to relatively newer agents like mefloquine. The pharmacological action of these drugs is directed against the food vacuole of the malaria parasite (Krogstad *et al*, 1985).

Antimalarial drugs such as chloroquine and mefloquine that are given to cure malaria infections do not eliminate mature *Plasmodium falciparum* gametocytes from the bloodstream. A person who has been successfully treated with schizontocidal antimalarial drugs may thus be healthy but infective for on average two months until the gametocytes die off naturally, or until another drug such as primaquine is given that does eliminate the gametocytes.

Resistance to the quinine-type of drugs is mostly a result of the ability of the parasite to prevent the intracellular accumulation of these drugs to toxic levels. At least two MDR (multidrug-resistant) efflux pumps, implicated in resistance to quinine-compounds have been identified and characterised (*pfmdr1* and *pfmdr2*; Krogstad *et al*, 1987; Cowman, 1991). These transporters play an important but not exclusive role in the acquisition of chloroquine resistance by the parasite. The locus of chloroquine resistance has since been mapped to a 40kb region on chromosome 7 (Wellems *et al*, 1990; Wellems *et al*, 1991; Wellems, 1992). Recently Wellems and Plowe showed that chloroquine resistance is associated in vitro with point mutations in two genes, pfcrt and pfmdr 1, which encode the *P. falciparum* digestive-vacuole transmembrane proteins PfCRT and Pgh1, respectively (Djimde *et al*, 2001).

The second group of drugs is directed against the folate-metabolism of the parasite. Pyrimethamine, its derivatives, chloroguanides as well as sulphonamides and sulphones are examples of drugs in this group. The mechanism of action of this group of drugs is clearly defined. Agents in this group interfere either with the incorporation of p-aminobenzoate into folate (inhibitors of the folate biosynthesis) or bind to and inhibit dihydrofolate reductase-thimidylate synthetase (DHFR-TS; Peterson *et al,* 1990; Basco *et al*, 1995b).

The last group comprises the newer compounds like artheter and artemisinin. The mechanism of their action is as yet largely unknown. It has however been postulated that the action of artemisinin is due to its interaction with hemozoin. Hemozoin is the polymerised form of haem released from red cell haemoglobin, found in the parasite food vacuole when the protein part of the haemoglobin is digested to serve as amino acid source for the malaria parasite (Hong, 1994).

1.7 Drug-resistance

Drug-resistant *P. falciparum* was first reported in Thailand in 1961. Drug-resistant malaria has become one of the most important problems in malaria control in recent years. Resistance *in vivo* has been reported to all antimalarial drugs except artemisinin and its derivatives. Drug-resistance necessitates the use of drugs that are more expensive and may have dangerous side effects. In some parts of the world, artemisinin drugs are the first line of treatment, and are used indiscriminately for self-treatment of suspected uncomplicated malaria - so we can expect to see malaria forms resistant to artemisinin within a decade (Greenwood, 1999).

The areas most affected by drug-resistance are the Indo-Chinese peninsula and the Amazon region of South America. The problem of drug-resistance can be attributed primarily to increased selection pressures on *P. falciparum* in particular, due to indiscriminate and incomplete drug use for self-treatment. In areas such as Thailand and Vietnam, mosquitoes of the *Anopheles dirus* and *A. minimus* species spread the drug-resistant parasites. These mosquitoes adapt their feeding activity to human behaviour patterns, and maintain intense transmission capacity (Greenwood, 1999).

In man, the problem of resistance to the common antimalarial drugs such as chloroquine and pyrimethamine, and the decreasing effectiveness of quinine is mainly limited to *P. falciparum* infection; chloroquine remains the treatment of choice for *P. vivax*. Several mechanisms can account for changes in drug sensitivity in the malaria parasites, for example, physiological adaptations due to non-genetic changes, selection of previously existing drug-resistant cells from a mixed population under drug pressure, spontaneous mutations, mutations of extranuclear genes, or the existence of plasmid-like factors (Krogstad *et al*, 1987; Cowman, 1991; Wellems *et al*, 1990).

Selection of mutants by the drugs themselves appears an important mechanism. In an environment where subtherapeutic levels of the antimalarial drugs are present, those parasites that have resistance through their natural variation or through mutations clearly have an important biological advantage. This means that even though the resistant forms were initially in the minority, the continued drug mediated elimination of intraspecies competition from the non-resistant forms has allowed the resistant forms to attain numerical superiority - to the point that drugs such as chloroquine are officially considered to be ineffective. The majority of studies indicate that drug pressure selection is to blame for the emergence of resistant malaria. The subcurative plasma levels of drugs found in many areas where there is uncontrolled and irresponsible prophylaxis and treatment will kill the most drug-sensitive forms of the parasite, but select the less sensitive ones. Spontaneous mutations in these forms tend to further reduce the sensitivity of the parasites to the drug. Fortunately, the problem of irresponsible prophylaxis has been recognised, and precautions taken. Not only must the release of the prophylactic drug be better regulated, but certain drugs can be kept in reserve for the treatment of infections. As drug-resistance seems to be genetically determined, gametocytes produced by resistant populations will produce more resistant parasites, promoting the spread of the resistant forms (WHO, 1997).

The *Plasmodium* parasites have extremely complex genomes, and the ease with which they can switch between the microenvironments in different hosts and adapt to the metabolic changes required, illustrates the difficulty in studying the exact modes of action of the antimalarial drugs on parasite metabolism. Resistance develops more quickly where a large population of parasites are exposed to drug pressure. The increasingly rapid spread of resistant malaria may be due to an increasingly efficient mosquito vector. This phenomenon may be explained by the increased oocyst formation efficiency that has been observed with resistant species. At any rate, the resistant forms undoubtedly have a biological advantage with transmission (Brown *et al*, 1986).

In order to appreciate the physiological nature of resistance, it is necessary to look in more detail at the metabolism of the parasites, and the modes of action of the antimalarial drugs. Intraerythrocytic stages of malaria ingest haemoglobin into the parasitic food vacuole. Here exopeptidases and endopeptidases break down haemoglobin into hemozoin (malaria pigment), of which the cytotoxic ferriprotoporphyrin IX is a major component. A parasite synthesised binding protein, 'haembinder', seems to sequester the membranolytic ferriprotoporphyrin IX into the inert hemozoin complex to protect the *Plasmodium* membranes from damage (Oliaro *et al*, 1995). It is now appropriate to discuss a number of antimalarials and apparent adaptations seen in resistant strains.

1.8 Treatment and immunity

It is obvious that resistance is an ongoing problem. By 1973, sulfadoxine-pyrimethamine cocktails had replaced chloroquine, but by 1985, this too was ineffective. Though quinine remains effective, there is a 50% failure rate unless it is supplemented with tetracycline. Between 1985 and 1990, the recommended treatment for malaria in Thailand was mefloquine, combined with sulfadoxine-pyrimethamine, but by 1990 the cure rate had fallen to 71% in adults

and 50% in children. This treatment can no longer be used in that area due to resistance (WHO, 1997). The future of chloroquine is not clear as a recent report suggests that due to the current absence of chloroquine drug pressure, chloroquine sensitivity may well be returning. In patients from an area where chloroquine has not been used for some time, only one in five of the *P. falciparum* infections were truly resistant to chloroquine- presumably due to the lack of selection pressures (WHO, 1997).

Malaria prophylaxis may be prescribed to protect against clinical symptoms. The type of prophylaxis however depends on the area, local species of malaria, local pattern of antimalarial drug-resistance as treatment is a function of the infecting species, the degree of drug-resistance, the severity of infection, and personal allergies and contraindications. No antimalarial prophylaxis regimen gives complete protection and malaria may be contracted despite taking antimalarial prophylaxis. In general, pregnant women and parents travelling with young children should weigh the necessity of the trip when travelling to areas where malaria parasites have become highly resistant to chloroquine (Greenwood, 1999).

Immunity is achieved only partially and for a duration that is a function of the intensity and frequency of prior infections. In areas with seasonal or epidemic malaria where disease is infrequent, adequate protective immunity may never develop. In endemic areas with high levels of transmission, new-born children are protected in their first months of life by the antibodies from their immune mothers. After that, they gradually develop their own immunity over the years with repeated non-lethal infections of the disease. The immunity is reversible, and fully "immune" adults who leave malarious areas are known to return to a state of non-immunity over a period of 1 to 2 years. In persons with sickle cell anaemia or the sickle cell trait, the abnormal haemoglobin S offers some protection against *Plasmodium falciparum* infection (Good *et al*, 1998).

P. vivax and *P. ovalae* can remain quiescent in the liver for many months. Relapses caused by the persistent liver forms may appear months, and occasionally up to 4 years, after exposure. Untreated or partially treated blood infection with *P. malariae* may be present for many years before giving rise to a symptomatic episode, and can be carried for a lifetime (Mons *et al*, 1997).

In areas with emerging drug-resistant *P. falciparum*, recrudescence of the infection may occur up to a month or more after what initially seemed to be a successful clinical cure of the infection. The boundaries of malaria transmission are determined by the presence and abundance of Anophelines, their susceptibility to malaria infection, the type of hosts they select for blood meals, and whether they live long enough to serve as effective transmitters of infection, which in turn is largely determined by ambient temperature and humidity. Although currently largely confined to what would be considered tropical conditions, in the past, malaria (*P. malariae* and *P. vivax*) was endemic as far north as Finland (Sherman, 1979).

The goal of malaria control is to prevent mortality and reduce morbidity and social and economic losses, through the progressive improvement and strengthening of local and national capabilities. The clinical diagnosis of malaria is difficult under the best of circumstances. Definite diagnosis is based on light microscopic observation of parasites in the red blood cells of the patient. Less widely used diagnostic tools include fluorescent staining, genetic probes, PCR, and antigen detection in the form of a dipstick (Greenwood, 1999).

The rate of reproduction of malaria is the estimated number of secondary malaria infections potentially transmitted within a susceptible population from a single non-immune individual. This number represents the theoretical estimate of the intensity of transmission. In practice transmission will depend amongst others on the parasite species involved, fluctuations of the source of infection (gametocyte carriers), the density and infectivity of the *Anopheles* species involved (Collins *et al*, 1995).

Factors making up the mathematical formula of the basic reproduction rate are: number of feeding sessions per person per night by vector population, multiplied by the expectation of infective life of the vector population, multiplied by the expectation of life of female vectors (making up the vectorial capacity of the vector population), multiplied by the mosquito's receptivity to infection, and the days of infectivity per case (Collins *et al*, 1995).

1.9 What the future may hold

During the last three years, the overall number of malaria cases has remained fairly steady at between 2.6 and 2.7 million cases reported annually. The real number of cases is likely to be nearer 19 million cases per annum (WHO, 1997).

As global eradication of malaria is not a possibility today, alternative ways of combating this disease must be studied. As effective anti-malarial vaccines will not be available in the foreseeable future it is of prime importance to get ahead of the parasite in the drug-resistance race. To achieve this goal it is essential to identify, isolate and study novel drug targets, the mechanisms of interaction between drugs and parasite proteins as well as to better understand the mechanisms utilised by the parasite to modify itself to become resistant to present and future drugs (Meutener *et al*, 1999).

Drug-resistance, however, is a menacing threat to malaria control. The solution to the malaria problem includes the identification of novel drug targets and the design of drugs against which resistance will not be as easily obtained. It will certainly remain a problem in the future, and it is unlikely that the drug manufacturers will ever produce a drug that can be described as a flawless antimalarial with no possibility of resistance. *Plasmodium falciparum* has demonstrated that it is extremely good at adapting to any drugs we may use against it - and there is no reason

to suspect that this would be different for new compounds. Responsible use of new drugs will be important for the future if artemisinin and whatever follows artemisinin are not to become as clinically compromised as chloroquine (Meutener *et al*, 1999; WHO, 1997).

This study focuses on the use of data from the *Plasmodium* genome-sequencing project in order to identify and isolate sequences of interest as potential vaccine candidates and drug targets. A better understanding of the metabolism of the parasite as well as the mechanisms by which it achieves drug-resistance is needed in the search for new drugs and drug-targets. Similarly, a better understanding of the mechanisms employed by the malaria parasite to achieve antigenic variation is essential for the rational design of vaccine candidates. The use of bioinformatic and molecular biological techniques to identify, isolate and study a novel drug target is described in the second chapter. It was decided to focus on a nutrient transport protein of the parasite. The glucose transporter protein was selected, due to its metabolic importance as well as it's likely localisation on the parasite membrane surface, which will make it more readily accessible as a drug target. Obtaining sequences of interest by means of EST-mining is described. In the third chapter we analysed the genomic signature of the malaria parasite. The genomic signature of the parasite is shown to not only be a usefull tool in studying its evolutionary relationship to other organisms, but also in the identification of genes obtained from other sources.

In addition to curative and prophylactic drugs, an effective and affordable vaccine must be developed. In order to achieve this, suitable vaccine targets must be identified. Various attempts have been made to design an effective malaria vaccine, but unfortunately no real successes have yet been obtained. One reason is the parasite's ability to achieve antigenic polymorphism and variation. The role that interspersed sequence motifs play in achieving antigenic polymorphism, as well as the mechanisms by which it is achieved, is addressed in the fourth chapter. The concluding discussion follows in chapter 5.

CHAPTER 2

AN IN SILICO APPROACH TO THE IDENTIFICATION AND ISOLATION OF PUTATIVE DRUG TARGETS

2.1 Introduction

Computer-based methods are used to effectively access and utilise the exponentially increasing number of deposited protein and nucleotide sequences in the databanks. A wide range of public databases is available and it is essential to understand the compatabilities and limitations of each of these. Sequence manipulation and search tools for these databases also abound. Computer-aided analysis has become an essential tool for the analysis and comparison of sequences and structures of proteins. The principal stages in for example, a drug-discovery programme, for which the use of these databases is indispensable, are shown in Figure 2.1.

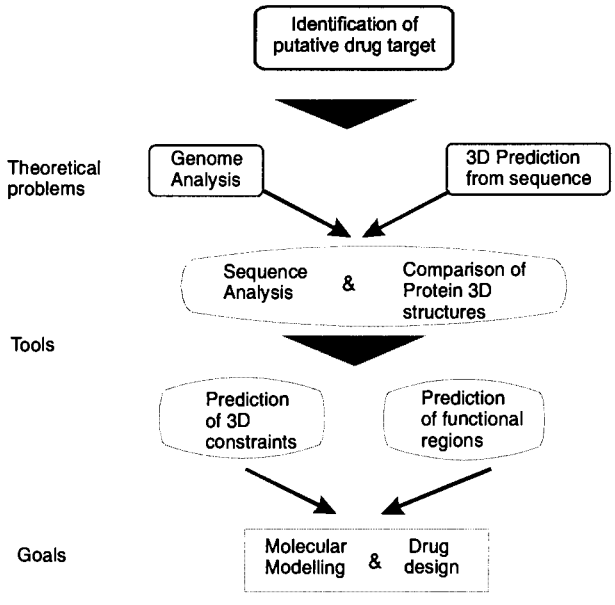

Figure 2.1: The use of Biocomputing in research towards drug-discovery.

An essential prerequisite when searching for potential drug targets is an intimate knowledge of the metabolism of the relevant pathogen and its host. In most instances such knowledge is still at a rudimentary stage and the proteins involved in specific pathways still need to be identified and characterised.

Genomic sequencing projects and easier access to sequence data have opened new doors for bioinformatic applications. A strategy for protein studies is to retrieve sequence information from the relevant databases with web-based search tools. Sequences can be retrieved either as nucleotide (e.g. Genbank; Burks *et al*, 1991) or protein data (e.g. Swiss-Prot; Bairoch *et al*, 1991) and used for comparison with BLAST or FASTA searches to identify similar sequences.

The next step is to explore protein family relationships. Sequence databanks are constantly increasing in size and redundancy, but the number of protein families identified has been levelling off. When searching for proteins with known structure and function, the use of motif- and family-specific databases increases the searching sensitivity and specificity. Protein family features in a sequence of interest can be efficiently identified by searching against any of several family-specific databases. Regions of sequence similarity characteristic to a protein family can then be used to detect more distant homologues in the sequence databanks. These regions are identified from multiple sequence alignments.

To get the most from an alignment, informative displays are essential. Logos display positions of multiple alignments as stacks of residue letters whose heights are indicative of the degree of conservation. Evolutionary information is traditionally displayed as trees derived from multiple alignments, which are valuable for discerning subfamily relationships (Schneider *et al*, 1990).

2.2 Identification of genes/ gene families of interest

When searching for vaccine candidates or putative drug targets, genes that are suitable candidates for further study must firstly be identified. In this chapter the focus was on nutrient transporter genes because of their suitability as drug targets. These proteins were deemed suitable chemotherapeutic targets due to their metabolic importance as well as their extracellular localisation, increasing their availability to any inhibitory drugs and circumventing the problem of intracellular multiple drug resistance (MDR) mechanisms. The glucose transporter was selected for further study because of its metabolic importance.

Infected erythrocytes are capable of metabolising up to 75 times more glucose than uninfected cells. The parasite must have a means by which it can obtain this supply of glucose (Sherman, *et al*, 1974). Kirk *et al* (1999) showed that the glucose uptake is via an equilibritative (passive) process. Sequences of passive glucose transporters from other organisms can thus be used to design primers or probes for the malarial glucose transporter.

2.3 Finding data of interest

Protein families / Superfamilies

Despite the tremendous theoretical and practical efforts devoted to understand how proteins fold and their functions, the main questions are still unsolved. Indeed, the major current limitation in Protein Engineering is our limited understanding about the structure and function of proteins. It is therefore of great importance for the future of Biotechnology to understand the factors important for protein folding to identify functional regions. Many open reading frames (ORF's) of sequences still have unknown functions. The prediction of three-dimensional structures and functional regions in protein families from the sequence data are major challenges. With the avalanche of sequences coming from the different genome projects and

the increased number of solved 3D structures, more proteins and families of proteins for which neither the function nor the structure are known, will be identified. Computational biology i.e. databases, sequence analysis, structural comparisons and protein-protein interactions are thus expected to play a major role in the characterisation of these proteins (Orengo *et al*, 1994; Tatusov *et al*, 1997).

2.4 Databases

Databases available

There are numerous nucleotide, protein, and other specialised databases freely available on the Internet. In most instances they are cross-referenced to each other as indicated in Figure 2.2.

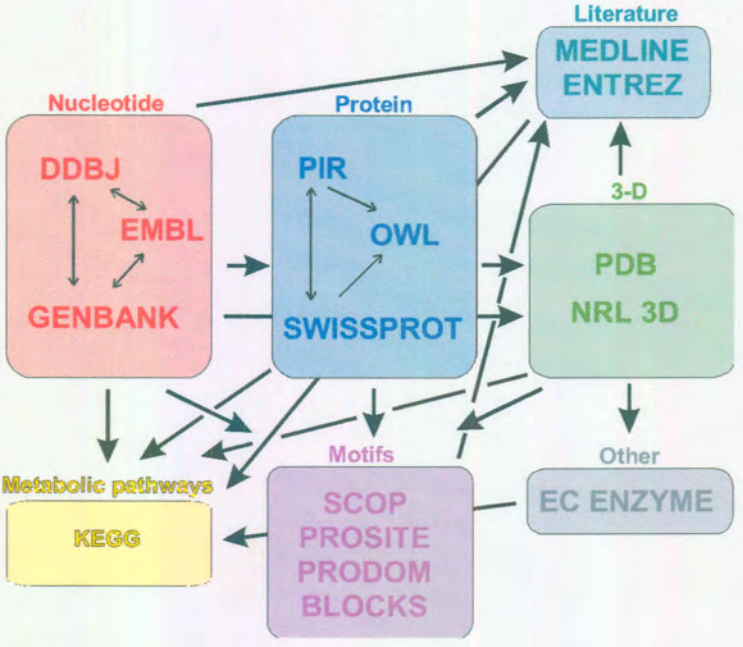

Figure 2.2: A simplified schematic representation of the most prominent databases and repositories available through the Internet as well as their interrelationships.

Nucleotide (DNA &RNA) databases

There are three main nucleotide sequence databases available on the net, namely Genbank, EMBL (European Molecular Biology Laboratory) Datalibrary and DDBJ (DNA Databank of Japan; Burks *et al*, 1991; Stoehr *et al*, 1991). These three organisations form the *International Nucleotide Sequence Database Collaboration*. They exchange data on a daily basis and can in theory be seen as just different names for the same database. In reality timelags in the propagation cause minor differences in these databases. A complete set of nucleotide sequences is however formed in the nr (non-redundant) nucleotide database of the NCBI (National Center for Biotechnology Information) by the merging of Genbank with the updates of the others. This database is maintained by the NCBI as target for their BLAST search services (Burks *et al*, 1991).

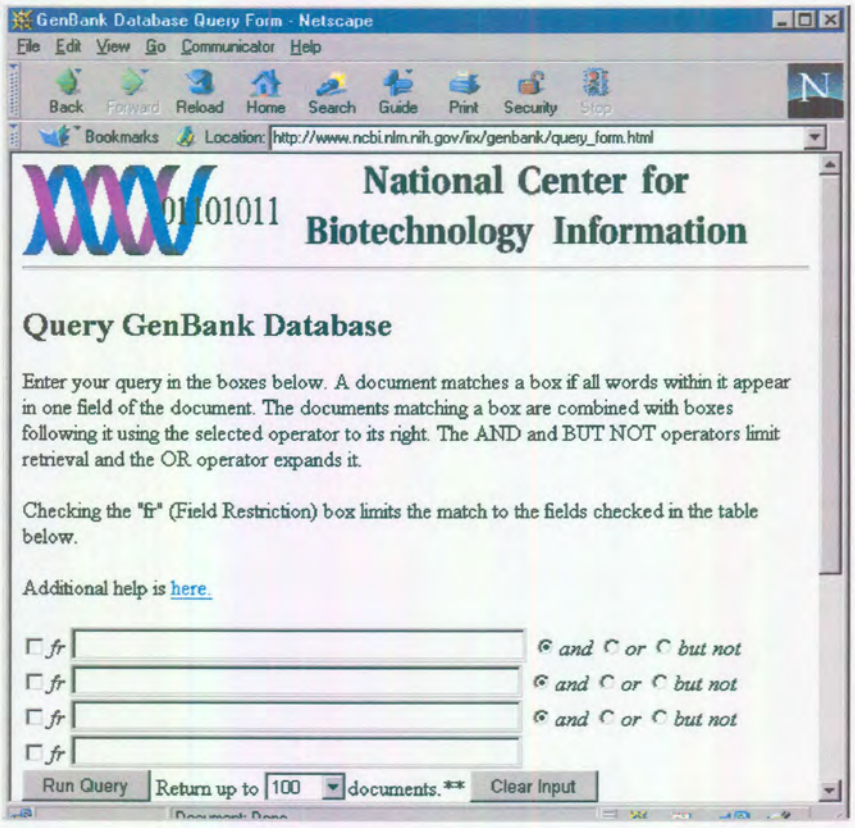

Figure 2.3: Access to Genbank using the WWW interface. The Genbank query page at the NCBI is shown here, using Netscape 4.0 as WWW-browser.

Genbank queries:.....http://ncbi.nlm.nih.gov/genbank/query_form.html

EMBL queries:.....http://www.ebi.ac.uk/queries/queries.html

DDBJ queries: http://ftp2.ddbj.nig.ac.jp:8000/cgi-bin/getent/

Protein databases

SwissProt

SwissProt is the best known of the protein databases. It is maintained at the University of Geneva as a highly curated, cross-referenced and non-redundant database. Unfortunately, due to cost-implications, not every sequence is in SwissProt (Bairoch *et al*, 1991).

SwissProt homepage:.http://expasy.hcuge.ch/sprot/sprot-top.html

Protein Identification Resource (PIR):

PIR1 is a totally non-redundant database with only one entry per protein product. The complete PIR database (PIR1 + PIR2 + PIR 3), however, has many redundancies (Barker *et al*, 1991).

PIR homepage: http://www.bis.med.jhmi.edu/Dan/proteins/pir.html

PIR query: .....http://www.bis.med.jhmi.edu/Dan/fields/pir.form.html

OWL:

OWL is a non-redundant composite of four publicly available primary sources: SWISS-PROT, PIR (1-3), Genbank (translation) and NRL-3D. SWISS-PROT is the highest priority source, all others being compared against it to eliminate identical and trivially different sequences. The strict redundancy criteria render OWL relatively "small" and hence efficient in similarity searches (Bleasby *et al*, 1994).

OWL homepage: .....http://www.bis.med.jhmi.edu/Dan/proteins/owl.html or
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/OWL.html

OWL queries:.....http://www.bis.med.jhmi.edu/bio/search/FILT/owlref.html

GenPept:

GenPept is produced by parsing the corresponding Genbank release for translated coding regions of Genbank sequences. The GenPept data can be accessed via the Entrez search engine at the NIH web site (www.ncbi.nlm.nih.gov/Entrez) as discussed later in this chapter (Burks *et al*, 1991).

Three-dimensional (3-D) protein structures

Protein Data Bank (PDB)

The PDB is maintained by Brookhaven National Laboratory and contains all publicly available solved protein structures. Searches against this database can be done to determine whether any known 3D structures are similar to the query protein. This database is non-redundant – only the "best" determination of a protein structure is left in the database (Bernstein *et al*, 1977).

PDB homepage: .....http://pdb.pdb.bnl.gov/

PDB query: .....http://pdb.pdb.bnl.gov/pdb-bin/pdbmain

NRL_3D

NRL_3D is a sequence-structure database derived from the 3 dimensional structure of proteins deposited with the Brookhaven National Laboratory's Protein Data Bank. The Web version derived from NRL_3D has hot links among its own entries and to the following Databases: PDB - The Protein Databank (3D structures), EC-Enzyme - The EC Enzyme Classification Database and Refbase - A Protein Sequence Citation Database (Barker *et al*, 1991).

NRL_3D homepage:.....http://www-nbrf.georgetown.edu/pirwww/dbinfo/nrl3d.html

Protein Motifs

Structural Classification of Proteins (SCOP)

Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification (Barton, 1994).

SCOP homepage:.....`http://scop.mrc-lmb.cam.ac.uk/scop/`

PROSITE database

The Prosite database of protein motifs is maintained at the University of Geneva. Each motif is defined by a regular expression or a profile and is accompanied by a description of the motif, a listing of the true positive, false negative and false positive SwissProt entries, as well as information about it's biology (Barioch *et al*, 1995).

PROSITE homepage:...`http://expasy.hcuge.ch/sprot/prosite.html`

BLOCKS database

The BLOCKS database is a database of gap-free multiple alignments of sequences based on the Prosite database developed by Steve Henikoff. Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the BLOCKS database are made automatically by looking for the most highly conserved regions in groups of proteins represented in the PROSITE database. These blocks are then calibrated against the SWISS-PROT database to obtain a measure of the chance distribution of matches.

It is these calibrated blocks that make up the BLOCKS database (Henikoff *et al*, 1994).

BLOCKS homepage: ...`http://www.blocks.fhcrc.org/blocks/`

ProDom database

ProDom - The ProDom (Protein Domain) database is a comprehensive collection of protein families. It was constructed by clustering all complete protein sequences in SwissProt. The novelty of ProDom is that the modular arrangement of proteins has been taken into account and whenever domain boundaries were detected the sequences were cut to produce consistent families of domains (Corpet *et al*, 1998).

ProDom homepage:.....`http://protein.toulouse.inra.fr/prodom.html`

Other databases

EC Enzyme

The Web version of the EC Enzyme database has hot links among its own entries and to the following Databases: OMIM - Online Mendelian Inheritance in Man, SwissProt - The Swiss Protein Database (Bairoch, 1993).

EC Enzyme homepage: `http://www.bis.med.jhmi.edu/Dan/proteins/ec-enzyme.html`

EC Enzyme query: .....`http://www.bis.med.jhmi.edu/bio/search/FILT/enzyme.html`

KEGG

The Kyoto Encyclopaedia of Genes and Genomes provides links from the gene catalogues generated by the genome sequencing projects to the biochemical pathways with links to pathway and genome maps as well as the relevant genes or molecules in other databases (Kanehisha *et al*, 2000).

KEGG homepage: `http://www.genome.ad.jp/kegg/`

Literature databases

MEDLINE

MEDLINE (MEDlars onLINE) is the National Library of Medicine's (NLM) premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. The MEDLINE file contains bibliographic citations and author abstracts from over 3,800 current biomedical journals published in the United States and 70 foreign countries. The file contains over 8.6 million records dating back to 1966.

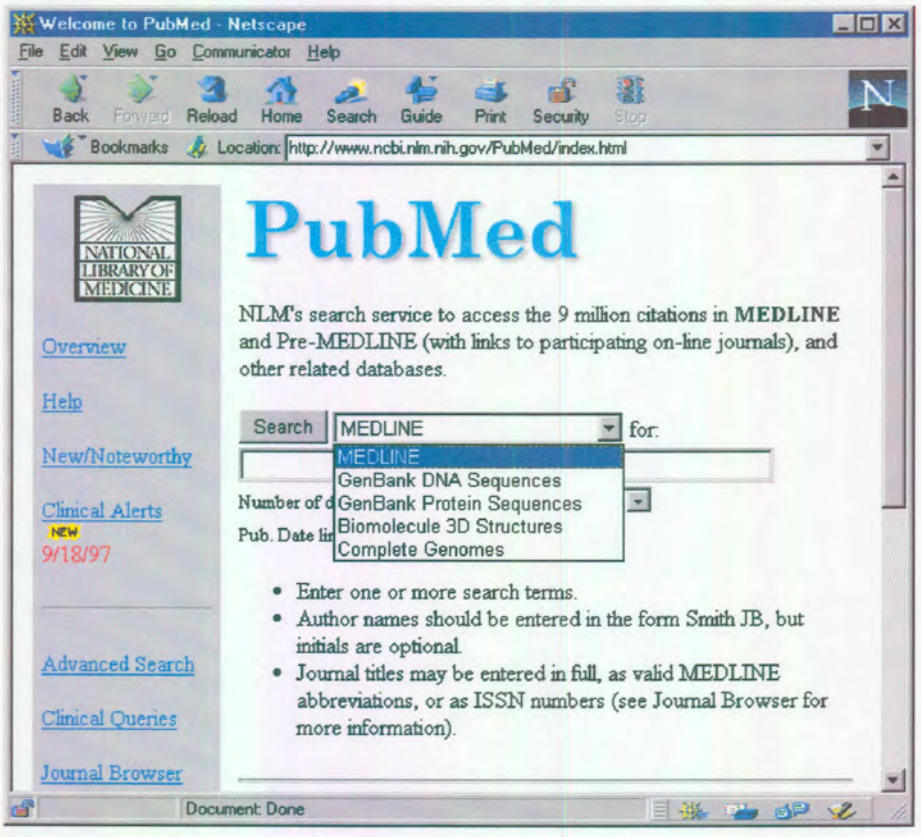

Figure 2.4: The WWW-Entrez interface for Public Medline searches (PubMed).

Coverage is world-wide, but most records are from English-language sources or have English abstracts. Each MEDLINE record is identified with a unique identifying number called a MEDLINE UID (MUID in PubMed). Citations for MEDLINE are created by the National Library of

Medicine, International MEDLARS partners, and co-operating professional organisations. MEDLINE records are incorporated into PubMed weekly, and are also assigned a PubMed unique identifier (PMID) (Greenberg, 1999; McEntyre, 1999).

MEDLINE homepage/query: `http://www.ncbi.nlm.nih.gov/PubMed/`

ENTREZ

ENTREZ is a search engine designed by the National Center for Biotechnology Information (NCBI) to search the Medline, Protein, Nucleotide, 3D structures and the Genomes databases at NCBI. The Entrez databases can be accessed on the Internet, using a dedicated ENTREZ client or a WWW client (e.g. Netscape). Links to related proteins, genes or publications can be accessed from here (McEntyre, 1999).

ENTREZ WWW server: `http://www.ncbi.nlm.nih.gov/Entrez/`

E-MAIL ACCESSING OF DATABASES:

Searches of Genbank / SwissProt databases can also be done by e-mail. The Retrieve E-mail server allows users to retrieve records via e-mail from Genbank and other sequence databases.

Retrieval may be based on accession number, locus name, keywords, author name or a variety of other search strategies. The format of queries is very simple, consisting of as few as three lines. A "DATABASE" operator and terms gb (Genbank), sp (SwissProt), etc., is used to specify the database to be searched. Retrieval hits can then be limited using multiple search terms combined with the "AND" and "BUT NOT" operators and widened using the "OR" operator or wildcarded search terms in the search.

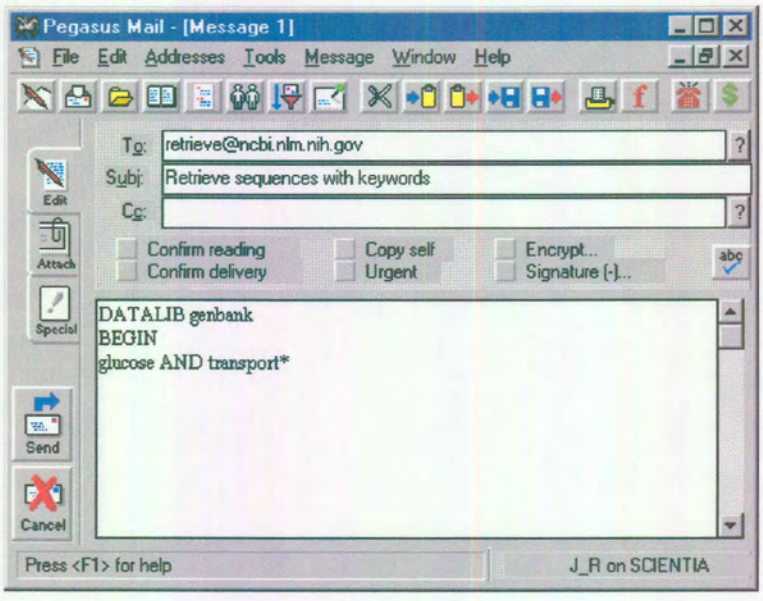

Figure 2.6: An example of the E-mail retrieval of Genbank sequences. The database searched is specified by the DATALIB delimiter and the search terms entered after the BEGIN command. Boolean delimiters and wildcards can be used in the search terms.

2.5 Plasmodium sequencing projects

Plasmodium genome project.

Detailed information about the parasite genome might uncover targets for drug and vaccine development in the form of vital and unique gene products. The genome approach is particularly useful because there are several stages in the parasite's life cycle from mosquito vector to human host, during which different sets of genes are expressed as proteins.

The *Plasmodium falciparum* genome is relatively small (2.5 to 3.0x 10^7 bp) and distributed over 14 chromosomes. The parasite also contains two stretches of extrachromosomal DNA (a 35kb plastid-like circular DNA as well as a 5966bp mitochondrial DNA). These extrachromosomal DNA sequences have been completely determined except for a small fragment (~10bp in length) in an inverted repeat of the 35kb plastid-like DNA (Dame *et al*, 1996).

The Sanger Centre, Tulane Institute for Genomic Research (TIGR) and Stanford University collectively undertook the task of sequencing the *Plasmodium* genome. A summary of distribution of the workload as well as the progress of the project is given in table 2.1.

Recognising the recent advances that have been made in high throughput sequencing of DNA and in bioinformatics and especially their application to the sequencing of the genomes of bacterial pathogens, a number of investigators, in late 1995/early 1996, undertook the task to completely sequence the genome of *Plasmodium falciparum* (Dame *et al*, 1996). All sequencing groups agreed to work on the 3D7 strain of *P. falciparum*.

A shotgun search engine for the sequences generated by the project can be found on the web at: `http://www.sanger.ac.uk/Projects/P_falciparum/shotgun_blastx_search.shtml`

Table 2.1: Summary of the progress of the *Plasmodium falciparum* genome sequencing project. (25 January 2001)

Chromosome	Size(Mb)	Institute	Status
1	0.7	Sanger Centre	Closure
2	0.95	TIGR	Finished
3	1.06	Sanger Centre	Finished
4	1.2	Sanger Centre	Closure
5	1.4	Sanger Centre	Closure
6	1.6	Sanger Centre	Shotgun Complete
7	1.7	Sanger Centre	Shotgun Complete
8	1.7	Sanger Centre	Shotgun Complete
9	1.8	Sanger Centre	Closure
10	2.1	TIGR	Closure
11	2.4	TIGR	Closure
12	2.4	Stanford Univ.	Closure
13	3.2	Sanger Centre	Closure
14	3.4	TIGR	Closure

Plasmodium sequence tag project

Knowledge of the sequence, identity and genomic location of the structural genes and virulence factors of the human malaria parasite will expedite efforts to identify new approaches for preventing, curing, and eradicating malaria. The sequencing of Expressed Sequence Tags (EST's) of cDNA's from blood stage parasites and Genome Sequence Tags (GST's) of a mung-bean nuclease library has been undertaken by John Dame and collaborators at the University of Florida. The goal of this project is to identify approximately 80% of *P. falciparum* genes. At present (December 2000), the project has sequenced approximately 3,000 tags. Genes are tentatively identified by homology to characterised genes in GenBank's non-redundant (nr) database using BLAST. An attempt is also made to map the physical location of each tagged gene in the genome.

Table 2.2: The progress of the Plasmodium sequence tag project. (December 2000)

	Genomic (Mung-Bean)	cDNA (asexual blood stage)
Clones Analysed	1356	1098
Sequence Tags	1733	1117
Significant Match (Score >100)	16% (23%[a])	23%
Current Redundancy Rate	14%	20%[b]
Estimate of Genes Tagged	1002	878
Average Tag Length	402 bp[c]	297 bp
Total DNA Sequence	696 kb	331 kb

[a] Last 169 clones. [b]Reduced from 32% by pre-screening for abundant messages. [c]For last 169 tags average length 582 bp.

These sequence tags are derived from cloned genomic and cDNA libraries. The *Plasmodium falciparum* (Mung Bean Nuclease-Digested) Genomic DNA library in pBlueScript SK(+) vector transformed into *E. coli* XL1 Blue host cells was used. Plasmid clones estimated to contain inserts of >500 bp were selected randomly for analysis. The *Plasmodium falciparum* cDNA library was prepared by reverse transcribing PolyA+ RNA from asynchronous bloodstage parasites of the Dd2 cloned isolate cultured *in vitro* using an oligo dT-Xho I primer. The second strand was prepared using RNase H and DNA polymerase I. The fragments were ligated into lambda ZAP II vector and transformed into *E. coli* XL-1 blue cells. The sequence tag data can be downloaded from the sites:

`http://goodman.jax.org/malaria/welcome.html` or

`http://parasite.vetmed.ufl.edu/falc.htm`

Either the genomic sequence tags, expressed sequence tags or a combined file is available for browsing or downloading. The lists are sorted by either the BLAST P-value (P(N)), Score of Best High-Scoring Segment Pair (HSP) Found by BLAST, Name of Sequence Tag or the Number of HSPs used by BLAST in Calculating P-value (the N in P(N)).

Efforts have been underway for several years to provide maps of the P. falciparum chromosomes.

Plasmodium chromosome maps: `http://www.wehi.edu.au/biology/malaria/`

Plasmodium EST project: `http://parasite.arf.ufl.edu/malaria.html`

Malaria genome mapping data:

`http://www.wehi.edu.au/biology/malaria/genomeInfo/MapData/MapData.html`

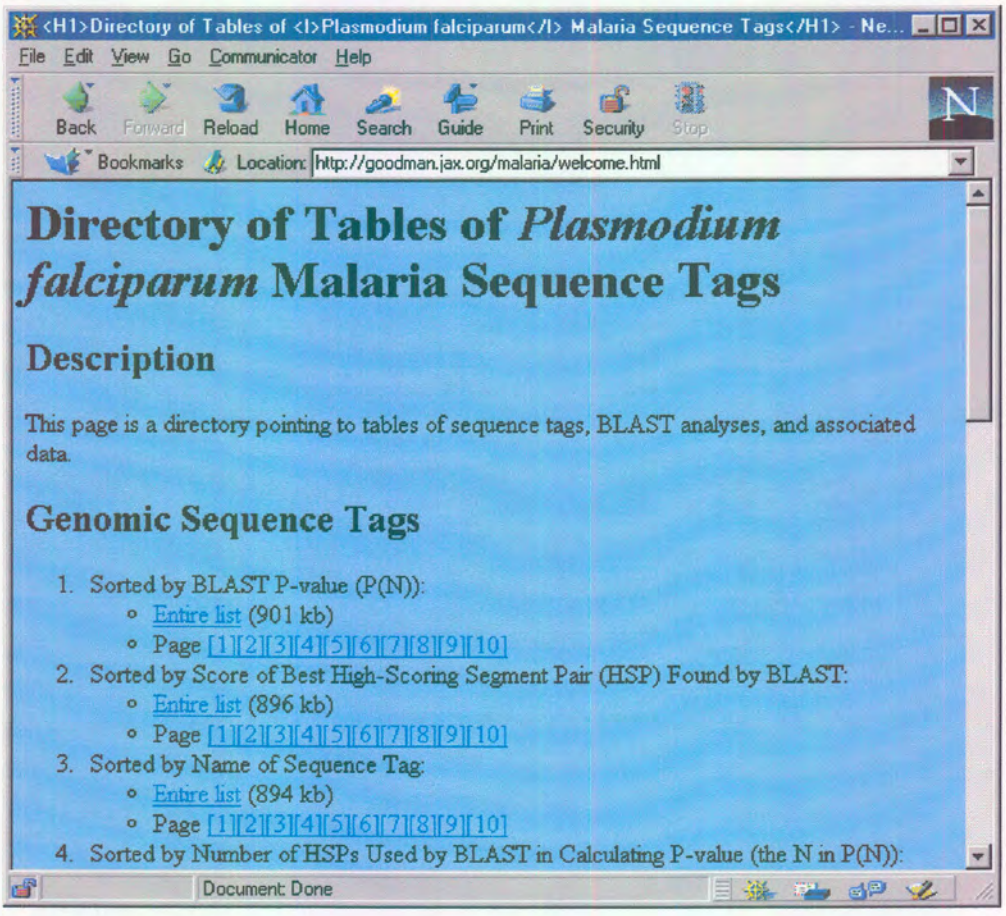

Figure 2.7: Plasmodium falciparum sequence tag accession/browsing site.
(`http://goodman.jax.org/malaria/welcome.html`)

Methods for isolating genes of interest (other than external antigen genes) include probing a cDNA library with gene-specific probes or RT-PCR with gene-specific primers. In both these instances similar problems with regard to the selection of sequences for probing or priming are

encountered. Problems that have to be overcome in the design of primers/probes include the identification of suitable consensus areas and the codon bias of the parasite DNA (>80 A+T rich; Goman, *et al,* 1982).

2.6 Methods

The following steps were followed in order to identify the putative gene of the permease type glucose transporter:

2.6.1 Collection of sequences of interest

The SwissProt database was queried using text queries e.g. "glucose AND transporter" to obtain amino acid sequences of interest. The sequences obtained from this query as well as sequences classified by the server as "related protein sequences" were then downloaded. These sequences were saved in the FASTA format for input in a multiple alignment program like CLUSTALW (Thompson *et al*, 1994). The query was then further refined manually to contain only the sequences of interest.

The shotgun-sequenced data of the *Plasmodium* genome project (at the Sanger web site: www.sanger.ac.uk) was also searched for possible transporter protein sequence tags using various keywords. The keywords "TRANSPORTER", "PERMEASE" and "GLUCOSE" were used. Only the amino acid sequences were used at this stage to eliminate variation due to differences in codon preferences between different organisms.

2.6.2 Sequence manipulations

Multiple alignment of these selected sequences was obtained with CLUSTALX (version 1.64). These CLUSTALX multiple alignments were then formatted for input to the LAMA (Local Alignments of Multiple Alignments) searcher using the block formatter tool on the WWW at:

http://fhcrc.org/block_formatter.html

Alternatively the sequences could be sent directly to the BLOCKMAKER server in FASTA format (http://fhcrc.org/blocks/block_formatter.html).

The formatted blocks were then submitted to the LAMA server where they were matched against the BLOCKS database (`http://blocks.fhcrc.org/blocks-bin/LAMA-search`). Alternatively, query text terms can be sent to the BLOCK searcher server to obtain previously identified consensus blocks for known protein families (`http:// www.blocks.fhcrc.org/blocks-bin/getblock.www`).

2.6.3 Identification of Genomic sequence tags (GST's) and Contigs containing identified motifs

Databases containing malaria sequences were searched for the presence of degenerate reverse translations of the identified amino-acid motifs. This was done on the nucleotide sequences deposited at NCBI, using the BLAST search engine (Altschul *et al*, 1990) as well as on the downloaded pre-release Shotgun sequence data from the Sanger Centre using text-based search utilities.

2.6.4 Identification and further studies on genes of putative glucose transporters

Sequences found containing consensus motifs were screened for open reading frames and translated to the conceptual amino acid sequences. These conceptually translated products were screened against the non-redundant nucleotide as well as protein databases at the NIH using the BLAST search engine to identify closely matching proteins. Hydrophobicity plots, as well as membrane-spanning predictions were also done using various available Web sites as well as the Antheprot software package.

2.7 Results

2.7.1 Identification of genes/ gene families of interest

Nutrient transporter proteins of the malaria parasite were identified as the main target of interest. The focus was on transporters for hexose sugars (glucose), an essential exogenous metabolite needed by the parasite for its asexual growth cycle.

2.7.2 Collection of sequences of interest

The sequences for a diverse range of organisms was downloaded after selection from the results obtained from the text word search of the SwissProt database. A number of relevant keywords were selected to search the database for matching XBLAST proteins. XBLAST matches a six-base reading frame translation of a given nucleotide sequence against the non-redundant protein databank at NIH. The results of the searches are given in Tables 2.3, 2.4 and 2.5, respectively.

Table 2.3: The following relevant genomic sequence tag matches were obtained when using the SHOTGUN BLASTX search engine at the Sanger genomic sequencing site with the keyword search term "TRANSPORTER":

Nearest BLAST search match	Sequence tag	SwissProt match	Reading frame	Consensus motif
Probable Formate Transporter (formate channel)	M3B4c10.r1t	P21501	+1	**A-(EQD)-(GHK)-K-(MIV)-HHT-(FW)-(VIFT)-E**
Probable amino acid ABC transporter	M3Cb3.r1t	P54537	-1	**(KQ)-Q-R-- --L-(LF)-D-E-P-T**
MDR transporter/ SER protease	mal3l2a3.s1t	Q23868	+1	**V-G-(KPH)-S-G-(SC)-G-K**
Hypothetical ABC transporter ATP-binding protein	mal3M2g2.r1t	P42423	+0	**S-G-S-G-K**
Na^+ bile acid cotransporter	mal3X2e8.r1t	P26435	-2	**P-L-(TAS)-LAG)-F-(VL)-L**
Ca^{2+} transporting ATP-ase (a)	mal3R1g2.r1t		-2	**G-(ASG)-N-D-- --(VI)-G-(VI)-G-(VI)**
Ca^{2+} transporting ATP-ase (b)	M3Rf5.s1t		-1	**G-(AG)-N-D-V-(ASG)-M-I**

Table 2.4: The following relevant genomic sequence tag matches were obtained when using the SHOTGUN BLASTX search engine at the Sanger genomic sequencing site with the keyword search term "PERMEASE":

Nearest BLAST search match	Sequence tag	SwissProt match	Reading frame	Consensus motif
Dipeptide transport system permease protein (DPPB)	M3N4h9.r1t	P37316	+1	**D-F-G-x-S**
Nitrate transport permease protein (NRTB)	mal3K1b6.r1t	P38044	-2	**I-V-A-A-E-M-T-L-- --G-I-G-- --V-G-L-(LS)-L**
Lactose transport system permease protein (LACG)	M3P4h1.s1t	P29824	?	?
Spermidine/ Putrescine transport system permease protein (POTC)	M3P4h1.s1t	P23859	?	?

Table 2.5: The following relevant genomic sequence tag matches were obtained when using the SHOTGUN BLASTX search engine at the Sanger genomic sequencing site with the keyword search term "GLUCOSE":

Nearest BLAST search match	Sequence tag	SwissProt match	Reading frame	Consensus motif
Glucose transporter type 2, liver. [0]	mal3Z1f2.r1t	P12336	?	?
Glucose transporter type 3, brain. [0]	mal3Z1f2.r1t	P47843	?	?
Glucose transporter (sugar carrier). [0]	mal3Z1f2.r1t	P23586	?	?
Glucose transporter type 5, small intestine (fructose transporter) [0]	M3Z3a12.s1t	P22732	?	?

In Table 2.5 the first three matches found were only different BLAST matches of the same genomic sequence tag as can be seen when comparing their sequence tag ID numbers.

2.7.3 Sequence manipulations

The multiple alignments were done using CLUSTALX(1.64b). Regions of high consensus were identified as family-specific transporter motifs. The multiple alignment for the sugar transporter protein is given in figure 2.8, showing the QQXXGIN consensus motif.

The multiple alignments were formatted for comparison with the BLOCKS database. (www.blocks.fhcrc.org) The match obtained with the input of the sugar-transporter alignments is shown in figure 2.9. The matching block with Block BL00216: Sugar_transport_1 is shown in

figure 2.10, revealing the QQ(LF)(ST)GIN consensus motif (Figure 2.11).

```
CLUSTAL X (1.64b) multiple sequence alignment

HXT2_YEAST      PSIVAEMDTIMANVETERLAGNASWGELFSNKG-AILPRVIMGIMIQSLQQLTGNNYFFY
HXT1_YEAST      PYIQYELETIEASVEEMRAAGTASWGELFTGKP-AMFQRTMMGIMIQSLQQLTGDNYFFY
Athalian        EEFQDLIDASEESKQVK-----HPWKNIMLPR---YRPQLIMTCFIPFFQQLTGINVITF
Ntabacum        EEFNDLVVASEASRKIE-----NPWRNLLQRK---YRPHLTMAIMIPFFQQLTGINVIMF
Ggallus         KEIAEMEKEKQEAASEKR----VSIGQLFSSS--KYRQAVIVALMVQISQQFSGINAIFY
Hsapiens        HDLQEMKEESRQMMREKK----VTILELFRSP--AYRQPILIAVVLQLSQQLSGINAVFY
ARAE_ECOLI      EKAREELNEIRESLKLKQ----GGWALFKINR--NVRRAVFLGMLLQAMQQFTGMNIIMY
ARAE_KLEOX      EKARDELNEIRESLKLKQ----GGWALFKVNR--NVRRAVFLGMLLQAMQQFTGMNIIMY
GALP_ECOLI      AEAKRELDEIRESLQVKQ----SGWALFKENS--NFRRAVFLGVLLQVMQQFTGMNVIMY
Tvivax          G--EGVLDPNEYSVRQM-------------------LGPLAVGAVTAGTLQLTGINAVMN
TH12_TRYBB      A--DGGMDPNEYGWGQM-------------------LWPLFMGAVTAGTLQLTGINAVMN
TRYPCRU         D--GTALDPNEYSYLQM-------------------LGPLAMGLVTSGTLQLTGINAVMN
Tsolium         EEEIGELLAEQENESENHT--KFPLKDLFRVK--ALRLALFVAVVAHLAQQFSGINAALF
Smansoni        DTFIGELREEIEVAKNQP---VFKFTQLFTQR--DLRMPVLIACLIQVLQQLSGINAVIT
GTR1_LEIDO      D--LCEFQEGDELPSVR-----IDYRPLMARD---MRFRVVLSSGLQIIQQFSGINTIMY
B_VULGAR        SLEVNEIKRSVASSSKRT---TIRFAELRQRR---YWLPLMIGNGLLILQQLSGINGVLF
Pfalc           EPLNAIKEAVEQNESAKK----NSLSLLSALKIPSYRYVIILGCLLSGLQQFTGINVLVS
                                                          .       *..* *

HXT2_YEAST      YGTTIFNAVGM---KDSFQTSIVLGIVNFASTFVALYTVDKFGRRKCLLGG-SASMAICF
HXT1_YEAST      YGTIVFQAVGL---SDSFETSIVFGVVNFFSTCCSLYTVDRFGRRNCLMWG-AVGMVCCY
Athalian        YAPVLFQTLGFG-SKASLLSAMVTGIIELLCTFVSVFTVDRFGRRILFLQG-GIQMLVSQ
Ntabacum        YAPVLFKTIGFG-ADASLMSAVITGGVNVLATVVSIYYVDKLGRRFLFLEG-GIQMLICQ
Ggallus         YSTNIFQRAGVG---QPVYATIGVGVVNTVFTVISVFLVEKAGRRSLFLAG-LMGMLISA
Hsapiens        YSTSIFEKAGVQ---QPVYATIGSGIVNTAFTVVSLFVVERAGRRTLHLIG-LAGMAGCA
ARAE_ECOLI      YAPRIFKMAGFTTTEQQMIATLVVGLTFMFATFIAVFTVDKAGRKPALKIG-FSVMALGT
ARAE_KLEOX      YAPRIFKMAGFTTTEQQMVATLVVGLTFMFATFIAVFTVDKAGRKPALKIG-FSVMAIGT
GALP_ECOLI      YAPKIFELAGYTNTTEQMWGTVIVGLTNVLATFIAIGLVDRWGRKPTLTLG-FLVMAAGM
Tvivax          YAPEIMRNIGM----DPMEGNSAVMSWNFVTALVAIPLVSRFTMRQLFLACSFMASCACL
TH12_TRYBB      YAPKITENLGM----DPSLGNFLVMAWNFVTSLVAIPLASRFTMRQMFITCSFVASCMCL
TRYPCRU         YAPKIMGNLGM----VPLVGNFVVMAWNFVTTLVSIPLARVLTMRQLFLGASLVASVSCL
Tsolium         YSTSLFESIGLT--SQAVYATLGVGSMIVVITVASIFLIERVGRRILLIGG-LSVMLFSA
Smansoni        YSSLMLELAGIP-DVYLQYCVFAIGVLNVIVTVVSLPLIERAGRRTLLLWP-TVSLALSL
GTR1_LEIDO      YSSVILYDAGFRDAIMPVVLSIPLAFMNALFTAVAIFTVDRFGRRRMLLISVFGCLVLLV
B_VULGAR        YSSTIFKEAGVT---SSNAATFGLGAVQVIATVVTTWLVDKSGRRLLLIVS-SSGMTLSL
Pfalc           NSNELYKEFLDS--HLITILSVVMTAVNFLMTFPAIYIVEKLGRKTLLLWG-CVGVLVAY
                  .  .                          .  :         .
```

Figure 2.8: Selection from the CLUSTALX (1.64b) multiple alignment of glucose transporter sequences, showing the consensus family-specific sugar-transport motif as shaded. A consensus alignment is indicated with "*" and similar residues with ":".

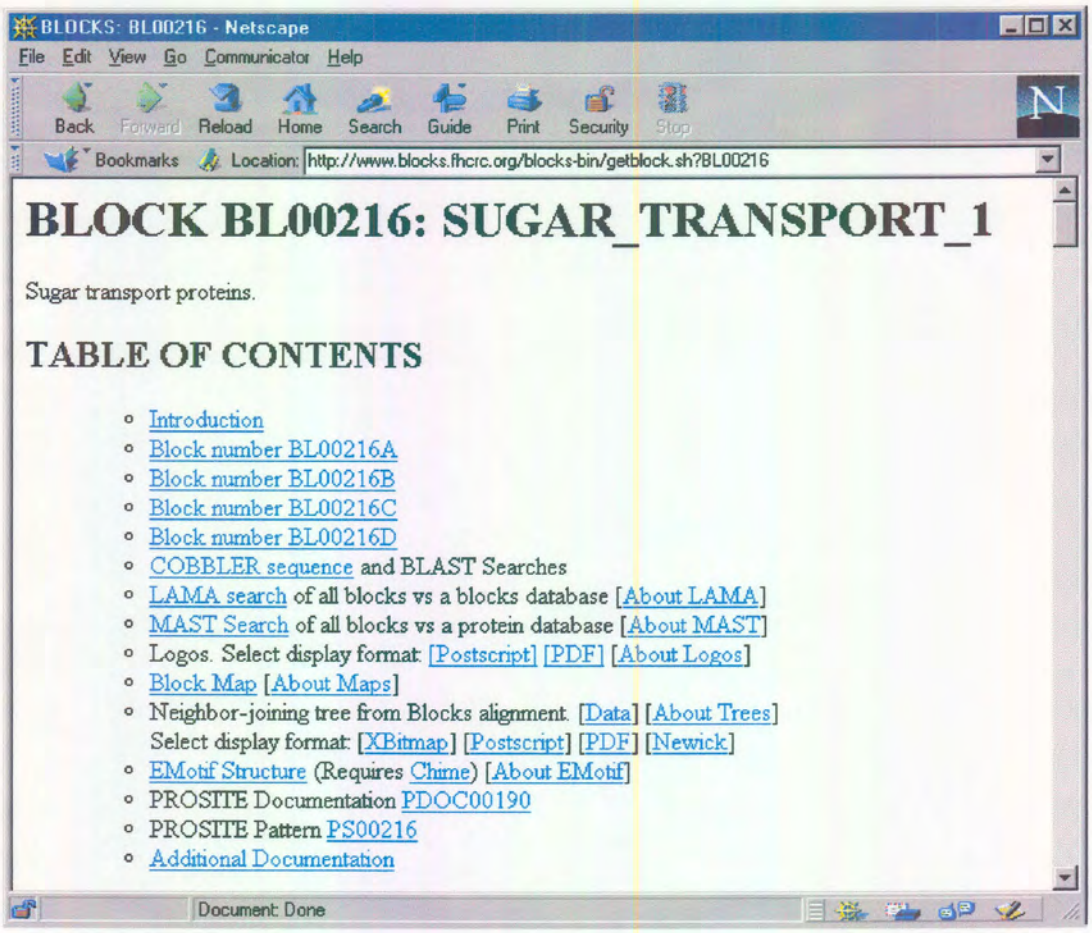

Figure 2.9: Example of the results obtained from the blocks server for the sugar transporter.

The protein consensus block (BL00216) is shown in figure 2.10.

Block BL00216C

```
ID   SUGAR_TRANSPORT_1; BLOCK
AC   BL00216C; distance from previous block=(12,94)
DE   Sugar transport proteins.
BL   QYG motif; width=21; seqs=57; 99.5%=1137; strength=1320
GLCP_SYNY3 ( 277) QQFVGINVIFYYSSVLWRSVG  20

GLF_ZYMMO  ( 279) QQLVGINAVLYYAPQMFQNLG  24

GTR1_LEIDO ( 250) QQFSGINTIMYYSSVILYDAG  20

GTR7_RAT   ( 312) QQTSGVNGIFYYHQHIYKQAG  33

HUP1_CHLKE ( 297) QQFTGINAIIFYVPVLFSSLG  22

PH84_YEAST ( 382) QTIGYAGSKNVYKKLYDTAVG  92

SNF3_YEAST ( 366) QQFSGINFIFYYGVNFFNKTG  21

STL1_YEAST ( 302) QQFTGCNAAIYYSTVLFNKTI  37

STP1_ARATH ( 294) QQLTGINVIMFYAPVLFNTIG  13

XYLE_ECOLI ( 288) QQFVGINVVLYYAPEVFKTLG  17

Y418_HAEIN ( 240) LEVVTTHFKPFFLGMLVCIAG 100

YB91_YEAST ( 307) QQFCGINSIIFYGVKVIGKIL  47

YCY8_YEAST ( 254) RNIPYFLALKFYWKRLLGTCG  79

YFE0_YEAST ( 286) VQFSGINIILGYITYICEIVG  36

YXBC_BACSU ( 253) QQAVGINTVIYYAPTIFTKAG  18

ARAE_ECOLI ( 269) QQFTGMNIIMYYAPRIFKMAG  12
ARAE_KLEOX ( 269) QQFTGMNIIMYYAPRIFKMAG  12
GALP_ECOLI ( 262) QQFTGMNVIMYYAPKIFELAG  13

GAL2_YEAST ( 341) QQLTGNNYFFYYGTVIFKSVG   8
HXT0_YEAST ( 319) QQLTGCNYFFYYGTTIFNAVG   8
HXT1_YEAST ( 335) QQLTGDNYFFYYGTIVFQAVG  10
HXT2_YEAST ( 326) QQLTGNNYFFYYGTTIFNAVG   8
HXT3_YEAST ( 332) QQLTGDNYFFYYGTTVFNAVG   7
HXT4_YEAST ( 341) QQLTGDNYFFYYGTTVFTAVG   8
HXT5_YEAST ( 356) QQLTGDNYFFYYGTTIFQAVG   6
HXT6_YEAST ( 335) QQLTGDNYFFYYGTTIFKAVG   6
HXT7_YEAST ( 335) QQLTGDNYFFYYGTTIFKAVG   6
HXT8_YEAST ( 337) QQLTGDNYFFYYGTTIFKSVG   6
HXTA_YEAST ( 331) QQLTGDNYFFYYGTTIFKSVG   6
HXTC_YEAST ( 328) LQLTGENYFFYYGTTIFKSVG  16
HXTD_YEAST ( 327) QQLSGINYFFYYGTSVFKGVG  10
RAG1_KLULA ( 333) QQLTGDNYFFYYGTTIFQSVG   6
YIR0_YEAST ( 221) QQLTGDNYFFYYGTTIFKSVG   6

GTR1_BOVIN ( 282) QQLSGINAVFYYSTSIFEKAG   5
GTR1_CHICK ( 281) QQLSGINAVFYYSTSIFEKSG  12
GTR1_HUMAN ( 282) QQLSGINAVFYYSTSIFEKAG   5
GTR1_MOUSE ( 282) QQLSGINAVFYYSTSIFEKAG   5
GTR1_RABIT ( 282) QQLSGINAVFYYSTSIFEKAG   5
GTR1_RAT   ( 282) QQLSGINAVFYYSTSIFEKAG   5
GTR2_HUMAN ( 314) QQFSGINGIFYYSTSIFQTAG   7
GTR2_MOUSE ( 313) QQFSGINGIFYYSTSIFQTAG   7
GTR2_RAT   ( 312) QQFSGINGIFYYSTSIFQTAG   7
GTR3_CHICK ( 281) QQLSGINAVFYYSTGIFERAG  10
GTR3_HUMAN ( 280) QQLSGINAVFYYSTGIFKDAG   7
GTR3_MOUSE ( 280) QQLSGINAVFYYSTGIFKDAG   7
GTR3_RAT   ( 280) QQFSGINAVFYYSTGIFQDAG   7
GTR4_HUMAN ( 298) QQLSGINAVFYYSTSIFETAG   5
GTR4_MOUSE ( 300) QQLSGINAVFYYSTSIFESAG   5
GTR4_RAT   ( 298) QQLSGINAVFYYSTSIFELAG   7

GTR5_HUMAN ( 288) QQLSGVNAIYYYADQIYLSAG  17
GTR5_RAT   ( 287) QQLSGVNAIYYYADQIYLSAG  17

ITR1_YEAST ( 348) QQFTGWNSLMYFSGTIFETVG  20
ITR2_YEAST ( 374) QQFTGWNSLMYFSGTIFETVG  20

MA3T_YEAST ( 372) GQCSCGASLIGYSTYFYEKAG  36
MA6T_YEAST ( 372) GQCSCGASLIGYSTYFYEKAG  36

QAY_NEUCR  ( 296) QNGSGINAINYYSPTVFRSIG  18
QUTD_EMENI ( 293) QNGSGINAINYYSPRVFKSIG  17
//

[Return to top]
```

Figure 2.10: An example of a sugar transporter consensus block of sequences as produced by the BLOCKS server. The consensus motif (Figure 2.11) is indicated by the shaded sequences.

To better visualise these consensus motifs they were viewed in the logos format. A sequence logo is a graphical representation of aligned sequences where at each position the size of each

residue is proportional to its frequency in that position and the total height of all the residues in the position is proportional to the conservation (information content) of the position (Schneider, *et al*, 1990).

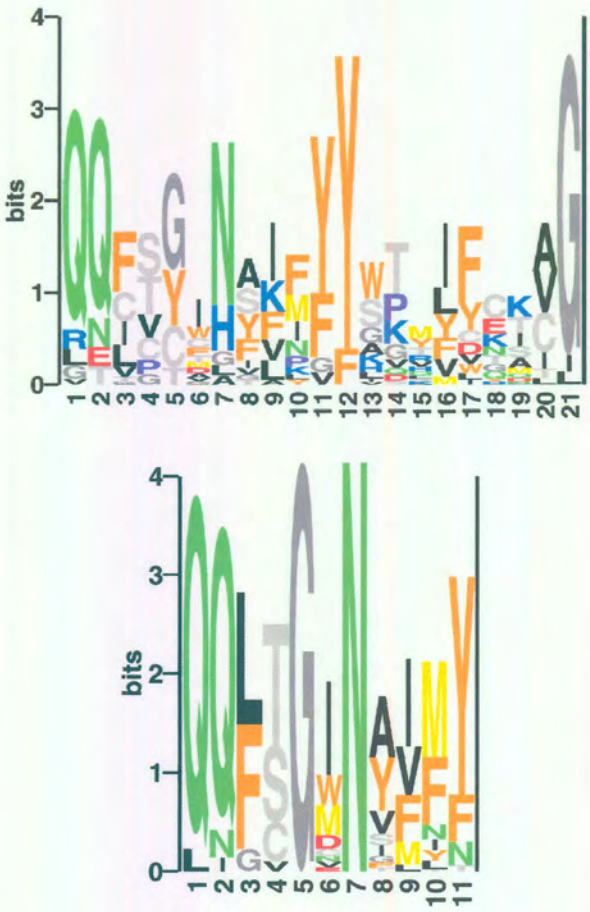

Figure 2.11: The logos graphical representation of the position-specific sorting matrices (PSSMs) of blocks in the sugar transporter family. All these blocks are showing a "QQ(LF)(ST)GIN" consensus motif. The colour scheme for the amino acids in the LOGOS representations is according to their physiochemical representations. Red for acidic amino acids (Glu and Asp); Blue for basic amino acids (Lys, Arg and His); Light grey for polar OH/SH amino acids (Ser, Thr and Cys); Green for amide amino acids (Asn and Gln); Yellow for Methionine; Black for hydrophobic amino acids (Ala, Val, Leu and Ile); Orange for aromatic amino acids (Tyr, Phe and Trp); Purple for Proline; Grey for Glycine.

The consensus blocks obtained with ClustalW alignments and Blocks searches were then used for the design of the gene-specific primers.

2.7.4 Identification of GSTs / Contigs containing identified motifs

Genomic sequence tag mal3Z1f2.r1t was identified in Table 2.5 as a glucose-transporter related sequence. All six reading frames of this sequence tag was analised and when the translation product of this sequence tag (from the *Plasmodium* genome sequencing project at Sanger Centre, `http://www.sanger.co.uk` - Figure 2.12) was compared with known sequences on the specialised malaria blast server (`http://www.ncbi.nlm.nih.gov/malaria/`), it was found to match the combined contig 7920 (116629bp long) on chromosome 2. The match was found to be on the complemantary strand of the mal3Z1f2.r1t sequence tag sequence.

Shotgun Read : mal3Z1f2.r1t

Predicted Gene :

Brief_ID :

Keyword :

[A+T] : 74

```
DNA : "mal3Z1f2.r1t"
          tgactatctaaaaattctttatataattcatttgaattggacactaaaac
          atttatacctgtaaattgttgtaaaccagataacaaacatcctaatatta
          taacatatctatatgatgggatttttaatgctgataataaagataaagaa
          ttttttctttgctgattcattttgttcaacagcttcttttatagcattcaa
          tggttcatctacattatctgtttcataaatttntttcaaaatgtttttgg
          attcttcaattcttcctttctcanaaagaaaatatggggtgtcttcttta
          aaaaaaaacaactaaagctaatatacctatttatgatatgacngagggga
          gatanaaactt
```

Figure 2.12: An example of the output of a genomic sequence tag identified with BLAST match to the glucose transporter type 3 Brain (mal3Z1f2.r1t).

The Blast match for this GST was confirmed by analysis on the BLOCKS (Figure 2.14) and BLAST (Figure 2.15) servers of the conceptually translated product from the sequence tag (Figures 2.13).

The preliminary genomic sequence data of the genome sequencing project was screened using the GST sequence (Figure 2.12). A contig containing a 1515bp open reading frame (Figure 2.16) was identified coding for a putative glucose transporter protein. The translated product of this open reading frame was compared to the SwissProt database, doing a BLAST-search and was found to highly match with proteins in the sugar transporter family.

```
        Nucleic Acid Sequence of and Protein coded by mal3Z1f2.r1t

                                                                    aa      2

gtttntatctcccctcngtcatatcataaataggtatattagctttagttgtttttttttaaagaagaca     72
V  ?  I  S  P  ?  S  Y  H  K  <  V  Y  <  L  <  L  F  F  F  K  E  D  T
24

ccccatattttctttntgagaaaggaagaattgaagaatccaaaaacattttgaaanaaatttatgaaac    142
  P  Y  F  L  ?  E  K  G  R  I  E  E  S  K  N  I  L  K  ?  I  Y  E  T
47

agataatgtagatgaaccattgaatgctataaaagaagctgttgaacaaaatgaatcagcaaagaaaaat    212
 D  N  V  D  E  P  L  N  A  I  K  E  A  V  E  Q  N  E  S  A  K  K  N     70

tctttatctttattatcagcattaaaaatcccatcatatagatatgttataatattaggatgtttgttat    282
S  L  S  L  L  S  A  L  K  I  P  S  Y  R  Y  V  I  I  L  G  C  L  L  S
94

ctggtttacaacaatttacaggtataaatgttttagtgtccaattcaaatgaattatataaagaatttttt    352
  G  L  Q  Q  F  T  G  I  N  V  L  V  S  N  S  N  E  L  Y  K  E  F  L
117

agatagtca     361
 D  S  ?      119
```

Figure 2.13: The conceptual amino acid translation of the reverse complement sequence of the mal3Z1f2.r1t sequence tag. The sugar transport consensus motif (QQFTGIN) is indicated in bold. The symbol "?" indicates a codon coding for more than one possible amino acid and "<" indicates a pause site.

The conceptually translated products were converted to the BLOCKS format and submitted to the BLOCKS WWW-site for comparison with other sequence-blocks in the database. As shown in Figures 2.14 matches to various sugar-transporter blocks were obtained.

```
Query=mal3Z1f2.r1t ,
 Size=112 Amino Acids
Blocks Searched=4460

1.--------------------------------------------------------------------------
Block     Rank Frame Score Strength   Location (aa) Description
PR00171B   346   0    894  1217          81-    100 SUGAR TRANSPORTER
SIGNAT
PR00171C     1   0   1175  1240          90-    100 SUGAR TRANSPORTER
SIGNAT

1175=99.42th percentile of anchor block scores for shuffled queries
P not calculated for single block PR00171C
                        |---  202 amino acids---|
   PR00171 A::::::::::::::::::BB:::::::::::::::::.....C::::::::::..DDDEE
mal3Z1f2.r                                  ::::::::::C
mal3Z1f2.r                                           BB

PR00171C    <->C   (231,463):89
ITR1_YEAST 348     QQFTGWNSLMY
                   |||||  | |
mal3Z1f2.r 90      QQFTGINVLvs

2.--------------------------------------------------------------------------
Block     Rank Frame Score Strength   Location (aa) Description
PR00172A     2   0   1095  1550          80-    101 GLUCOSE TRANSPORTER
SIGN
PR00172F   162   0    916  1373          77-     97 GLUCOSE TRANSPORTER
SIGN

1095=90.85th percentile of anchor block scores for shuffled queries
P not calculated for single block PR00172A
                        |---   76 amino acids---|
   PR00172 AAAAAAA:::::.BBBBBBB:::CCCCCCC::::.DDDDDDDD::.EEEEEE:::FFFFFFF
mal3Z1f2.r AAAAAAA
mal3Z1f2.r FFFFFFF

PR00172A    <->A   (231,316):79
GTR3_HUMAN 270     IIISIVLQLSQQLSGINAVFYY
                   ||    |   ||  |||
mal3Z1f2.r 80      IIlgclLsGlQQFtGINvlvsn
```

Figure 2.14: Block searcher Query of the amino acid translation of the reverse complement strand of the mal3Z1f2.r1t sequence tag to the Prints Database in blocks format. The sugar transport consensus motif (QQFTGIN) is indicated in bold.

The conceptually translated sequence tag (mal3Z1f2.r1t) was also compared to the SwissProt protein sequence database and significant matches to sugar transporters could also be observed in the results from a BLAST search of the translated sequence tag on the database (Figure 2.15).

```
sp|P45598|ARAE_KLEOX ARABINOSE-PROTON SYMPORT (ARABINOSE TRANSPORTER)
            pir||S47089 arabinose-proton symporter - Klebsiella oxytoca
            gi|498920 (X79598) arabinose-proton symporter [Klebsiella oxytoca]
            Length = 472
  Minus Strand HSP's:
 Score = 70 (32.2 bits), Expect = 0.00016, Sum P(2) = 0.00016
 Identities = 12/31 (38%), Positives = 21/31 (67%), Frame = -3
Query:    110 RYVIILGCLLSGLQQFTGINVLVSNSNELYK 18
              R  + LG LL  +QQFTG+N+++  +  ++K
Sbjct:    256 RRAVFLGMLLQAMQQFTGMNIIMYYAPRIFK 286

 Score = 56 (25.8 bits), Expect = 0.00016, Sum P(2) = 0.00016
 Identities = 11/31 (35%), Positives = 20/31 (64%), Frame = -3
Query:    311 LFFFKEDTPYFLXEKGRIEESKNILKXIYET 219
              L  F  ++P +L EKGR  E++ +L+ + +T
Sbjct:    194 LVIFLPNSPRWLAEKGRHVEAEEVLRMLRDT 224

sp|P09830|ARAE_ECOLI ARABINOSE-PROTON SYMPORT (ARABINOSE TRANSPORTER)
            pir||B26430 arabinose transport protein - Escherichia coli
            gi|145321 (J03732) arabinose-proton symporter [Escherichia coli]
            gi|1789207 (AE000368) L-arabinose isomerase [Escherichia coli]
            prf||1303337A arabinose transport protein [Escherichia coli]
            Length = 472

  Minus Strand HSP's:
 Score = 70 (32.2 bits), Expect = 0.00034, Sum P(2) = 0.00034
 Identities = 12/31 (38%), Positives = 21/31 (67%), Frame = -3
Query:    110 RYVIILGCLLSGLQQFTGINVLVSNSNELYK 18
              R  + LG LL  +QQFTG+N+++  +  ++K
Sbjct:    256 RRAVFLGMLLQAMQQFTGMNIIMYYAPRIFK 286

 Score = 54 (24.8 bits), Expect = 0.00034, Sum P(2) = 0.00034
 Identities = 11/31 (35%), Positives = 20/31 (64%), Frame = -3
Query:    311 LFFFKEDTPYFLXEKGRIEESKNILKXIYET 219
              L  F  ++P +L EKGR  E++ +L+ + +T
Sbjct:    194 LVVFLPNSPRWLAEKGRHIEAEEVLRMLRDT 224
```

Figure 2.15: The BLAST search output of the genomic sequence tag in figure 2.12 (Glucose transporter type 5 of small intestine) against non-redundant protein database, indicating the homology matches of the queried malaria tag sequence (mal3Z1f2.r1t) and the matches found in the databases. The sugar transport consensus motif (QQFTGIN) is indicated in bold.

The sequence tags were screened against the unpublished shotgun contig sequences and a full-length conceptual sequence, with start and stop-codons as well as the sugar-transporter motif could be obtained (Figure 2.16).

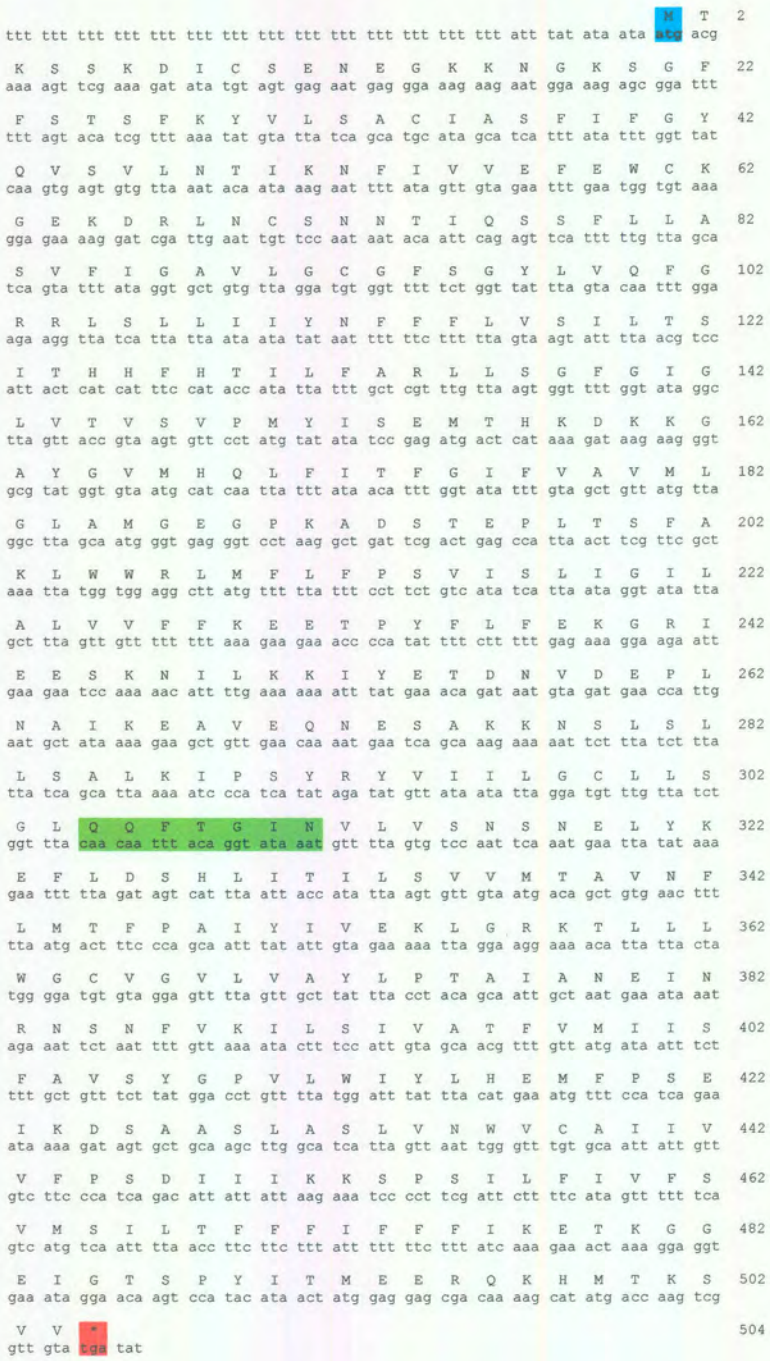

Figure 2.16: The fragment of contig 7920 containing the putative glucose transporter gene, with the conceptual translation thereof. The start and stop codons are shown in blue and red respectively and the glucose transporter consensus motif in green.

2.7.8 Identification and further studies on putative glucose transporter gene

Phylogenetic analysis was done using the UPGMA Neighbour-joining method in PROTPARS with the PHYLIP package (Felstein, 1989). Evolutionary distances of the different branches of the unrooted tree are shown. The putative *Plasmodium* sugar transporter could be classed with other sugar transport proteins but as shown in Figure 2.17, it did not match closely with any member of the sugar transporter group. As shown by Saier *et al* (1994), there is a clustering of groups (e.g bacteria clustering together) but the prokaryotes occur on their own branches.

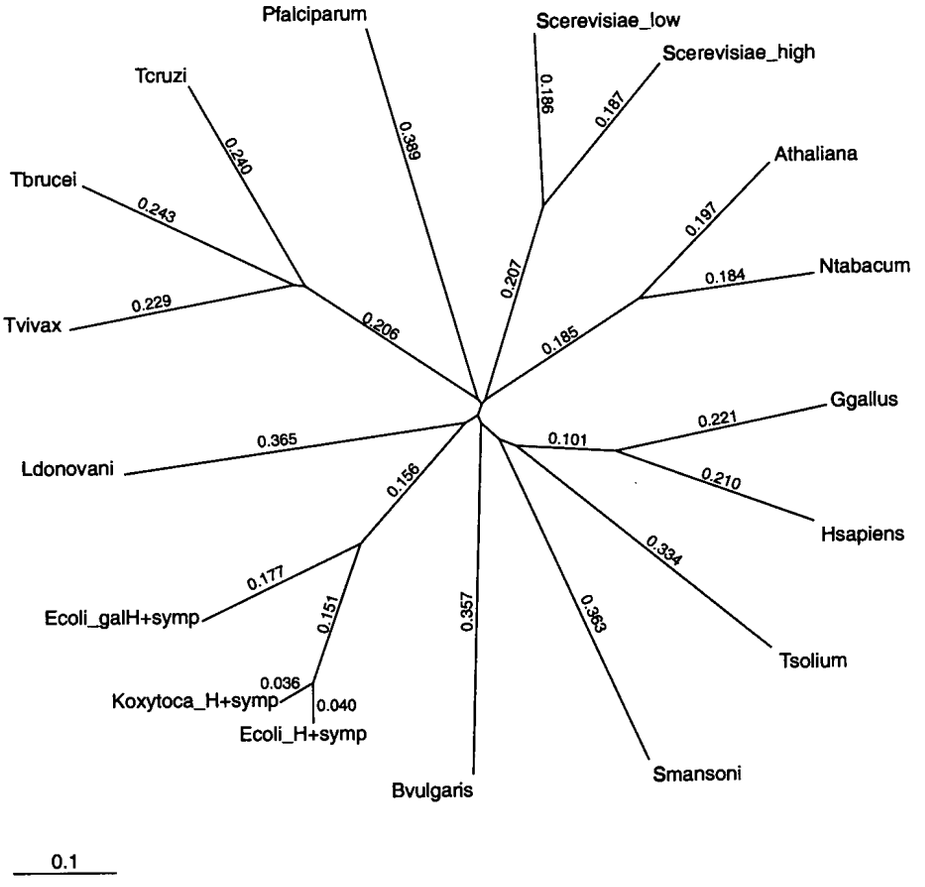

Figure 2.17: Phylogenetic analysis was done on the malaria glucose transporter as well as other genes in the sugar transporter family, showing evolutionary distances between the different proteins. Evolutionary distances are indicated on the branches.

Helix prediction plots were done using the Antheprot software package (Geourjon *et al*, 1995). The predicted transmembrane helix-profile (Figure 2.18) corresponded with those of other known sugar transport proteins (Figure 2.19).

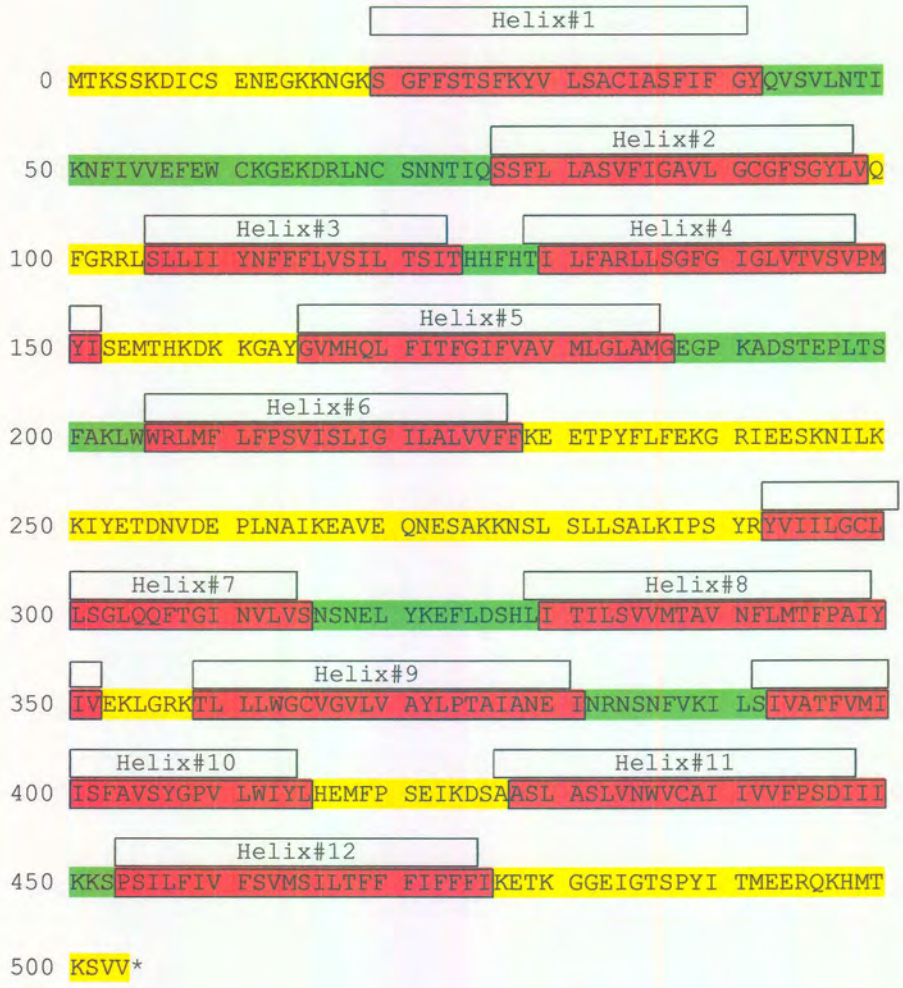

Figure 2.18: Diagram indicating the predicted transmembrane helices in red, cytoplasmic regions in yellow and extracellular regions in green.

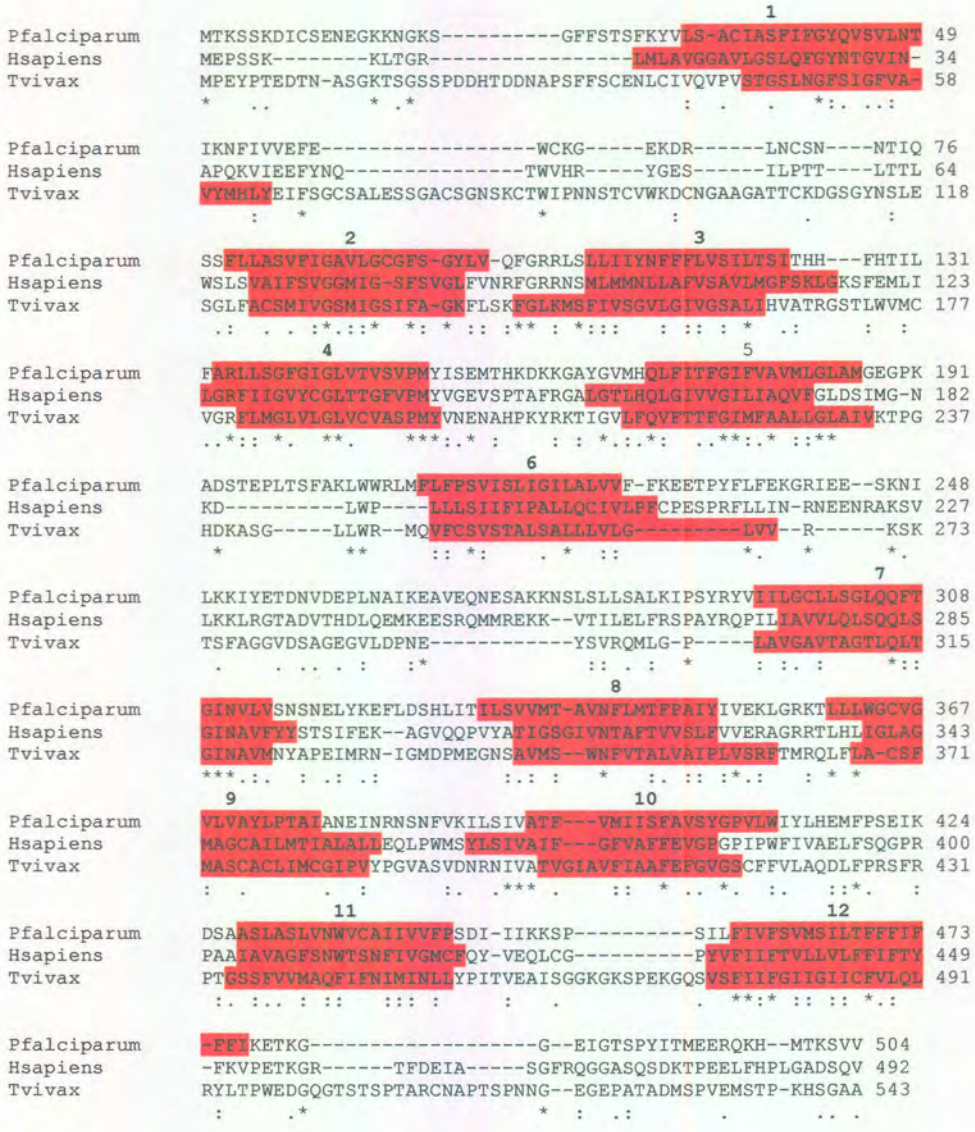

Figure 2.19: Comparison of transmembrane helices between the putative malarial sugar transporter (*P. falciparum*; Genbank Accession Number AF101827) and known sugar transporters of *Trypanosoma vivax* (*T. vivax*; Genbank Accession Number L47540) and *Homo sapiens* (*H. sapiens*; Genbank Accession Number AJ245937). The shaded areas indicate predicted transmembrane helices. The multiple alignment of the sequences was performed with ClustalW.

The theoretical pI and molecular weight were calculated as: 8.80 and 56416.66, respectively by the Compute pI/Mw server at the ExPASy Web site (http://expasy.hcuge.ch/ch2d/pi_tool.html; Bjellqvist *et al*, 1993). These values compare closely with the pI and molecular weight values of other known glucose transporter proteins as shown in Table 2.6.

Table2.6: Comparison of pI values and molecular weights of putative transporter with that of known glucose transporters.

SwissProt Accession	Description	pI	Molecular weight
Gene product	Malarial putative glucose transporter	8.80	56416.66
GTR1_BOVIN (P27674) (Bovine)	GLUCOSE TRANSPORTER TYPE 1, ERYTHROCYTE/BRAIN.	8.94,	54131.82
GTR1_CHICK (P46896) (Chicken)	GLUCOSE TRANSPORTER TYPE 1 (GT1).	8.82,	54086.62
GTR1_HUMAN (P11166) (Human)	GLUCOSE TRANSPORTER TYPE 1, ERYTHROCYTE/BRAIN.	8.93,	54117.79
GTR1_RABIT (P13355) (Rabbit)	GLUCOSE TRANSPORTER TYPE 1, ERYTHROCYTE/BRAIN.	8.94,	54097.72
GTR1_RAT (P11167) (Rat)	GLUCOSE TRANSPORTER TYPE 1, ERYTHROCYTE/BRAIN.	8.94,	53962.51
GTR2_CHICK (Q90592) (Chicken)	GLUCOSE TRANSPORTER TYPE 2, LIVER.	8.90,	57699.17
GTR4_MOUSE (P14142) (Mouse)	GLUCOSE TRANSPORTER TYPE 4, INSULIN-RESPONSIVE (GT2).	8.81,	54951.68

Primers were designed from the obtained contig sequence to amplify the *Plasmodium* sugar transporter mRNA sequence by RT-PCR. The amplified sequence was subsequently cloned and sequenced. The obtained experimental sequence was shown to be identical to the contig-sequence (Nel, 1998).

2.8 Discussion

The intraerythrocytic stages of the malaria parasite pose a great demand on the host cell for the supply of nutrients and building blocks. The malaria parasite requires external sources of certain amino acids, purine nucleotides and glucose in order to grow and proliferate. The parasite increases its volume 25-fold and reproduces itself manifold over a 48-hour period. These enormous demands for nutrients and building blocks can not be met by intraerythrocytic supply alone and the parasite profoundly alters the host cell's permeability to these compounds (Homewood *et al*, 1974; Gero *et al*, 1992; Desai *et al*, 1993). Transporters of these essential nutrients would be ideal drug targets as they are both essential to parasite proliferation and readily accessible to chemotherapeutic agents due to their extracellular membrane-localisation.

Sequence motifs are identified by an iterative method. Starting with a few sequences of interest homology searches are done against available databases (Genbank and SwissProt) to obtain a large number of similar sequences. A multiple alignment of these sequences are then produced using the ClustalW program. Possible consensus motifs are identified both by eye and by sending the alignments to the BLOCKS server. Identified motifs are then used in database searches (LAMA server). Sequences are then added or removed on the basis of their weighted similarity to the other sequences in the initial group and the process is repeated. A range of motifs for transport proteins was identified using the text search function of the SHOTGUN BLAST server at the Sanger genomic sequencing centre.

The differences between the conventional and the alternative strategies of design and gene identification employed in the identification of the glucose transporter are depicted in Figure 2.20. The tools used to facilitate the faster identification of consensus motifs in gene families are highlighted in green.

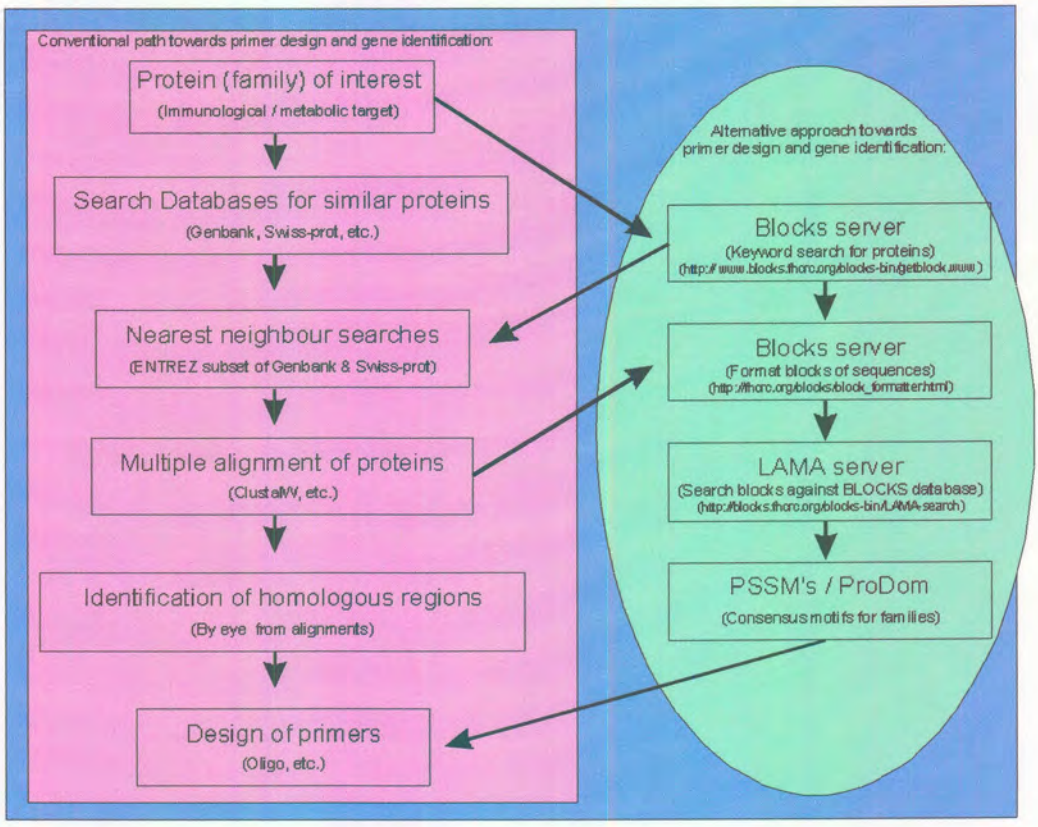

Figure 2.20: Strategies that can be followed when searching for consensus sequence motifs and in the design of primers.

Using this strategy one main sugar transporter motif "QQ(FL)(TS)GIN" was identified from the CLUSTALW multiple alignment of the relevant family of sequences. In general no less than four sequences should be used to produce a reliable alignment for the identification of consensus motifs to avoid the false-positive identification of sequence-family motifs. ClustalW firstly does a pairwise alignment of the sequences, calculating a distance matrix. From this it creates a rooted neighbour-joining tree (guide tree for the further alignments) followed by a progressive alignment along the guide tree.

As a result of the accelerating expansion of sequence databanks, it becomes increasingly probable that a search for similarity will succeed in detecting a relationship between any newly determined sequence and one or more known sequences. Often such relationships are

important clues to gene or protein function. However, sometimes the similarity is too weak for a potentially interesting relationship to be detected above the background of chance alignments. Hence, the high incidence of unmatched sequence tags from the sequencing projects. The background increases with the growth of sequence databanks, making distant relationships even more difficult to detect with confidence. Detection of distant relationships is aided by the presence of multiple members of a single protein family in a database. Alternatively, a database in which relationships are explicitly represented can be searched. An example of this latter approach is a database of protein "blocks" where each block is a local multiple alignment of ungapped segments from a group of related proteins. A query sequence is searched against this database of blocks by calculating a position-specific scoring matrix representing each block and scoring every possible position in the query for all blocks in the database. Searching a database of blocks provides information on local relationships, useful for identifying sequence motifs. These searches are more specific than are searches of sequence databases because blocks represent only the most highly conserved regions of proteins, a much smaller set than the set of sequences. Most protein families are characterised by multiple local motifs indicative of more global relationships. The QQ(FL)(TS)GIN motif was found as part of a sugar transporter block in the BLOCKS database. Block00216 (Sugar_transport_1) is a highly conserved block amongst sugar transporter proteins (Figure 2.11).

Global information is present in the Blocks Database as multiple blocks and distances between them observed for the sequences in the protein family. If a query sequence belongs to a family with multiple blocks, then at least a subset of these blocks should score highly in a search and be arranged in a compatible way along the query.

The multiple alignments (generated either with a program like ClustalW or on the Blocks server) are first transformed into position specific scoring matrices (PSSMs). Each column in the PSSM corresponds to a position in the alignment and has the amino acid distribution of that position.

The transformation into the PSSM is done with position-based sequence weights (Henikoff *et al*, 1994) and odd ratios between the amino acid frequencies observed in the multiple alignments and the frequencies expected from protein databases (Henikoff *et al*, 1997). The transformation corrects possible overrepresentation of some sequences by sequence weighting and considers the background frequencies of the amino acids. The method was tested and calibrated with ungapped local multiple alignments (blocks) from the BLOCKS Database.

The data that was obtained from the subsequent LAMA search included the sorting and matching of the sequences with existing protein families, confirming the classification of the sequences as glucose transporter proteins. Matching blocks in the blocks database, with links to the ProDom protein domain database could also be identified. Calculation of PSSMs and corresponding LOGOs of the relevant high homology blocks gave a highly visual representation of the homology within this conserved area.

When searching genomic sequences for coding regions there are three main factors to be taken into consideration, namely: The location of the coding regions themselves, the strand or orientation of the sequence and the reading frame of the sequence and its translated product. There are two main methods used for finding coding regions in genomic sequences (Staden, 1990). The first method is to identify signal sequences such as splice junctions and promoters that surround coding regions. The second method is to examine long stretches of DNA to see if they show a higher similarity to coding or non-coding sequences. The simplest method would be to search for stretches of DNA containing no stop-codons in a specific reading frame. There must then be distinguished between open reading frames that exist by chance and those coding for proteins. In the case of *Plasmodium* criteria like the A+T content - which is markedly higher in noncoding regions (>80%) than coding regions (~70%) - can be used to further discriminate between coding and non-coding regions.

In this case a variation of the first method was used by searching for regions containing a family-specific motif whereafter the length and reading frame of the genomic fragment were analysed. The identified sugar transporter motif was found in the translated product of the malaria sequence tag mal3Z1f2.r1t. Comparison of the malaria genomic sequence tag mal3Z1f2.r1t with known sequences on the specialised malaria blast server matched it to the combined contig 7920 on chromosome 2. Upon analysis of open reading frames in the contig, a 1515bp open reading frame was identified. The translated product of this open reading frame was compared to the SwissProt database, doing a BLAST-search and was found to match highly with proteins in the sugar transporter family.

Helix prediction plots were done on the conceptually translated product and it was found to contain twelve predicted transmembrane regions with both the C- and N-terminal regions of the protein predicted to be on the cytoplasmic side of the membrane. The predicted membrane structure of the putative glucose transporter was found to be similar to those of other known glucose transporter proteins, containing 12 transmembrane-spanning helices.

After confirmation of the proposed sequence by automated sequencing of a number of cDNA clones, this sequence was deposited in Genbank on 27 October 1998 under the accession number AF101827. The sequence was subsequently submitted to Genbank on 2 November 1998 by Gardner *et al* (1998) as part of the complete chromosome 2 sequence under the accession number AE001381. The proposed substrate specificity for glucose as well as the structural division of the protein in 12 membrane-spanning helices was confirmed the next year by Woodrow *et al* (2000; Genbank accession number: PFA131457).

The Genomic-sequencing project will continue to provide a wealth of data that only need to be utilised. As the project releases more data, genes for amino acid and nucleoside transporters will probably also be found in a similar way. As the genome sequencing project is projected to

release the complete genome sequence of *Plasmodium falciparum* later in 2001, the focus will shift to functional genomics.

2.9 Acknowledgements

Sequence data for *P. falciparum* chromosomes were obtained from The Sanger Centre website (www.sanger.ac.uk/Projects/P_falciparum/), The Institute for Genomic Research website (www.tigr.org) and The Stanford DNA Sequencing and Technology Center website (www-sequence.stanford.edu/group/malaria). Sequencing of *P. falciparum* chromosomes were accomplished as part of the International Malaria Genome Sequencing Project with support by The Wellcome Trust, the Burroughs Wellcome Fund and the U.S. Department of Defence.

CHAPTER 3

IN SILICO ANALYSIS OF OLIGONUCLEOTIDE SIGNATURES

3.1 Introduction

Our understanding of the human malaria parasite, *Plasmodium falciparum* at the molecular genetic level is rather rudimentary. The *P. falciparum* genome project is therefore a powerful resource for comprehensive understanding of the parasite at the molecular level. Eukaryotic genomic DNA contains numerous repetitive nucleotide sequences in coding and non-coding regions (e.g. transposons and sub-telomeric repeats; Levin, 1997). These characteristics of an organism have definite functions, whether it is to regulate expression of proteins with either multiple gene copies or polymorphic repetitive, immunodominant regions. Coding regions of genes dominate in prokaryotic genomes whereas the eukaryotic genome is dominated by noncoding sequences. Families of nucleotide repeats derived from transposable elements constitute a major part of eukaryotic genomes (Levin, 1997).

Base composition, di-, tri- and even tetranucleotide frequencies are important characteristics of an organism's genome. The base composition and nearest-neighbour frequencies of sequences have been shown to exhibit organism-specific genomic signatures (Santibanez-Koref *et al*, 1986; Karlin *et al*, 1995). For example, elevated GG•CC levels found in RNA-viri, thermophylic prokaryotes and organellar DNA but not in other prokaryotes or in eukaryotes are postulated to be due to positive selection of C + G content. Lower levels of C+G would be more advantageous for organisms with smaller genomes and high replication rates due to the lower energy required for stacking and unstacking in each replication cycle. The genomic signature of

an organism is an array of all its dinucleotide relative abundance values. While maintaining a relative uniformity throughout an organism's genome, it is an organism-specific property that can be used to discriminate between sequences of different organisms (Campbell *et al*, 1999).

Various models have been proposed to explain the relative over- and under-representation of certain dinucleotides in DNA sequences. The CG-suppression model proposes that CG-levels in vertebrates are decreased by a methylation-demethylation process. Methylation of the cytosine at position 5 in the CG dinucleotide and subsequent deamination converts it to thymidine resulting in elevated TG levels (Bird, 1980; Burge *et al*, 1992). Due to CG being a methylation "hot-spot" it is also argued that a reduction of CG dinucleotides would be evolutionary selected for in order to minimise potential harmful consequences of mutation effects (Karlin *et al*, 1995).

It has been shown that normalised CG and TA dinucleotides (calculated as an Observed / Expected - O/E ratio) are generally underrepresented in the coding regions of *P. falciparum*. Significantly lower than expected frequencies were also found for AC, TA and CG dinucleotides with weight-averaged O/E ratios of 0.85, 0.8 and 0.5, respectively (Hyde *et al*, 1987; Weber, 1987; Musto *et al*, 1997). The TG, CC and CA dinucleotides exhibited a higher than expected frequency with O/E ratios of 1.45, 1.4 and 1.15, respectively. In non-coding regions, only three dinucleotides were found to differ significantly from the expected values. The GG dinucleotide frequency was significantly higher (O/E of 1.5) and the CT and AG dinucleotides frequencies lower (O/E of 0.8 and 0.7, respectively) than expected. The observed dinucleotide frequencies of *P. falciparum* (higher than expected TG, CC and CA levels and lower than expected levels of the CG and TA dinucleotides) correlate to dinucleotide patterns of higher vertebrates rather than those of lower vertebrates and invertebrates (Karlin *et al*, 1992; Karlin *et al*, 1995).

Differing frequencies of nucleotides can also be accounted for by restrictions placed on neighbouring bases by the codon preference of an organism (Santibanez-Koref *et al*, 1986). Viewing the triplet codon as comprised of two dinucleotide sequences (e.g. ATG comprising of the two dinucleotides AT and TG) its effect on the dinucleotide composition is quite apparent. The relative over- and under-representation of tetranucleotides will be likewise associated or influenced by the composition of their dinucleotide subunits. However, probable constraints imposed on the codon preference of an organism by its dinucleotide composition cannot be excluded.

Neither codon preference (Santibanez-Koref *et al*, 1986) nor the methylation /demethylation hypothesis (Karlin *et al*, 1992) can adequately explain the almost universal depression of TA and CG dinucleotides, since CG levels in mitochondria are also depressed although they have no mechanism for methylation (Karlin *et al*, 1995). The overrepresentation of homodinucleotides (AA·TT and GG·CC) in organellar and prokaryotic DNA is postulated to be due to polymerase slippage events (Karlin *et al*, 1992). A number of dinucleotide sequences (AT·AT; CA·TG) have also been implicated in minor instabilities in microsatellite and dinucleotide repeat sequences (Mitas, 1997). The relative abundance of dinucleotides (normalised for G+C content) have been shown to reflect the evolutionary selection pressure for a certain organism and can as such provide insight into its evolutionary roots (Karlin *et al*, 1995). The avoidance of or preference towards specific dinucleotides in an organism could thus be due to a combination of regulatory, structural as well as evolutionary factors.

Since the previous studies (Hyde *et al*, 1987; Weber, 1987) the quantity of *Plasmodium falciparum* sequence data has increased exponentially (http://www.sanger.ac.uk; http://www.tigr.org). The complete sequence of *Plasmodium falciparum* chromosome 2, being both larger and more representative of the whole genome, was used in this study for analysis of oligonucleotide frequencies. The relative dinucleotide abundance values that reflect the

genomic signature of an organism could be due to a number of constraining factors. To study the origin of these constraints, possible nearest-neighbour influences of tri- and tetranucleotides on the dinucleotide genomic signature of *Plasmodium falciparum* chromosome 2 were also analysed. As neighbouring bases and codon-preferences have been previously proposed to impose constraints on the dinucleotide composition of an organism (Nussinov, 1984; Santibanez-Koref *et al*, 1986), possible constraints placed on the genomic signature by amino acids abundancies were also investigated. Further insights are expected to be gained into the evolutionary origins of the malaria parasite by comparison of the *Plasmodium falciparum* genomic signature with that of other organisms (Karlin *et al*, 1995 and Karlin *et al*, 1992). The base composition as well as the di-, tri and tetranucleotide genomic signatures would furthermore be useful tools in the identification of genes with foreign origins, incorporated in the malaria genome by means of lateral transfer (Wren *et al*, 2000).

3.2 Methods

3.2.1 Nucleotide sequences used

The complete sequence of chromosome 2 of *Plasmodium falciparum* was used for analysis of di-, tri- and tetranucleotide abundancies. Sequence data for *P. falciparum* chromosomes were obtained from The Institute for Genomic Research (TIGR) website (http://www.tigr.org). Sequencing of *P. falciparum* chromosomes was accomplished as part of the International Malaria Genome Sequencing Project with support by The Wellcome Trust, the Burroughs Wellcome Fund and the U.S. Department of Defence.

3.2.2 Analysis of di-, tri- and tetranucleotide abundancies

The counts of di-, tri- and tetranucleotide sequences in the complete nucleotide sequence as well as the coding regions of chromosome 2 were determined using the SeqSearch program, developed for these studies (unpublished; Appendix E). The program analyses sequences in plain text format using text string searches. The abundance of a specific oligonucleotide is defined as the number of instances of that specific oligonucleotide is counted in the dataset. The counting window (2 for dinucleotides, 3 for trinucleotides, etc) is shifted with increments of one base to count all instances of an oligonucleotide.

For the whole chromosome 2 and its coding sequence (CDS) subset, the average counts of short nucleotide sequences and their reverse complements were determined (e.g. TGGA combined and averaged with its reverse complement TCCA to yield the average count of TGGA•TCCA). This was necessary as both strands of the whole chromosome contain coding sequences. The counts for the non-coding regions were calculated by subtracting the counts in the CDS subset from the counts in whole chromosome 2. The observed counts were compared to the expected counts (Equation 7) as the Observed/Expected (O/E) or odds ratio (Equation

11). The expected counts were calculated for a subset of randomly distributed nucleotides of similar size and G+C content. The frequency of these sequences was defined as the number of occurrences of the specific sequence per total number of nucleotides in the dataset.

3.2.3 Equations used in the analysis of the sequence data

The frequency (f) of a specific oligonucleotide sequence was defined as follows:

$$f = \frac{Number\ of\ occurences\ of\ the\ specific\ sequence}{Number\ of\ total\ nucleotides} \quad \text{.....(1)}$$

The occurrence of the different bases as fractions of the total nucleotide composition was calculated using the following equations to correct for the high (A+T):(G+C) ratio of the malaria genome:

$$\text{Adenine fraction of total nucleotide bases} = \frac{\text{Observed number of Adenine bases}}{\text{Total number of bases}} \quad \text{.....(2)}$$

$$\text{Thymine fraction of total nucleotide bases} = \frac{\text{Observed number of Thymine bases}}{\text{Total number of bases}} \quad \text{.....(3)}$$

$$\text{Guanine fraction of total nucleotide bases} = \frac{\text{Observed number of Guanine bases}}{\text{Total number of bases}} \quad \text{.....(4)}$$

$$\text{Cytosine fraction of total nucleotide bases} = \frac{\text{Observed number of Cytosine bases}}{\text{Total number of bases}} \quad \text{.....(5)}$$

Where a+t+g+c=1

The encounter rate (K_{Exp}) of a specific oligonucleotide sequence was defined as the number of times a specific oligonucleotide sequence is expected to occur in a DNA fragment with length L (number of bases) and calculated using equation 6:

$$K_{Exp} = a^u c^v g^w t^x \quad \text{.....(6)}$$

with: K_{Exp} = Encounter rate of oligonucleotide with length L

where L=u+v+w+x

u = number of positions in sequence occupied by Adenine
v = number of positions in sequence occupied by Thymine
w = number of positions in sequence occupied by Guanine
x = number of positions in sequence occupied by Cytosine
a, *t*, *g*, *c* = fractions of respective nucleotide bases
(with $a + t + g + c = 1$) ..... (from Equations 2 to 5)

The theoretically expected counts of a specific oligonucleotide sequence were calculated as follows:

$$C_{Exp} = K_{Exp} \cdot \text{Subset size} \quad \text{.....(7)}$$

with: *Subset size* = The size of the Sequence searched in nucleotides.
C_{Exp} = The number of instances (counts) the specific oligonucleotide is expected to occur in the Sequence searched.
K_{Exp} = Encounter rate of oligonucleotide (Equation 6)

The observed frequency of a specific dinucleotide sequence was calculated using equation 8.

$$f_{XY} = \frac{C_{Obs}}{\text{Subset size}} \quad \text{.....(8)}$$

with: *Subset size* = The size of the Sequence searched in nucleotides.
C_{Obs} = The number of instances (counts) the specific dinucleotides counted experimentally in the Sequence searched.
f_{XY} = Experimentally determined (observed) frequency of dinucleotide sequence XY.

Chi-square values were calculated using the standard equation:

$$\chi^2 = \sum \frac{(C_{Obs} - C_{Exp})^2}{C_{Exp}} \quad \text{.....(9)}$$

C_{Obs} = Observed Counts (Experimentally determined)
C_{Exp} = Expected Counts (Theoretically calculated)

Odds ratios are used to compare the actual dinucleotide frequency to the expected frequency calculated from the frequencies of its base components. The odds ratio of the dinucleotide XY was calculated by (Burge *et al*, 1992). An odds ratio of 1 indicates that the dinucleotide occurs at the expected level as calculated by its base composition. Values higher or lower will indicate relative over- or underrepresentation of the dinucleotide.

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y} \quad \text{..... (10)}$$

f_X = Frequency of nucleotide X (A, C, G or T) in the sequence XY
f_Y = Frequency of nucleotide Y (A, C, G or T) in the sequence XY
f_{XY} = Frequency of dinucleotide XY
ρ_{XY} = Odds ratio of dinucleotide XY

Strand asymmetry (SA; purine/pirimidine distribution) was calculated with equation 11 (Hyde *et al*, 1987). A strand asymmetry of zero would indicate an equal distribution of purines and pirimidines in both strands. Strand asymmetry values higher or lower than zero will indicate preference in the coding strand for purines and pirimidines respectively.

$$SA = \frac{A+G-T-C}{A+G+T+C} \times 100\% \quad \text{..... (11)}$$

3.2.4 Amino acid composition

The amino acid composition of chromosome 2 was determined from the total amino acid composition of the individual genes in the CDS-subset using the SeqSearch program (Appendix E).

3.3 Results

3.3.1 Base-Analysis of sequences

Chromosome 2 of *Plasmodium falciparum* contains 229 open reading frames (ORF's) coding for gene-products with a combined length of 440031 base pairs. The high A+T content (80.6%) of chromosome 2 shown in Table 3.1 corresponds to previously published data (~82%; Weber, 1987).

Table 3.1: Base composition of Chromosome 2 sequence.

	Coding regions (CDS subset)		Non-coding regions (Introns, 3' & 5' UTR's, etc)		Whole chromosome	
	Total (C_{Obs})	% of Total	Total (C_{Obs})	% of Total	Total (C_{Obs})	% of Total
A	197291	44.8	174225	36.0	371516	40.2
T	135689	30.8	237154	49.0	372843	40.4
C	43539	9.9	46490	9.6	90029	9.7
G	63512	14.4	25748	5.3	89260	9.7
A+T	332980	75.7	411379	85.1	744359	80.6
G+C	107051	24.3	72238	14.9	179289	19.4
Total	440031	100.0	483617	100.0	923648	100.0
Strand Asymmetry (SA):		18.54		*		-0.227

* The Strand Asymmetry (SA) could not be calculated for the non-coding regions as their counts were obtained by subtraction of the counts for the open reading frames from the whole chromosome counts. The CDS open reading frames occur on both strands of chromosome 2. The coding region (CDS subset) was searched on the sense strand only. The SA of the non-coding regions (obtained by subtraction of base-counts in the CDS-subset from that in the whole chromosome 2) could thus not be calculated.

The A+T ratio in the non-coding regions (85.1%) of chromosome 2 is higher than the ratio in the coding regions (75.7%). Strand asymmetry (SA) of the coding regions (18.54%) was also much higher than the asymmetry evident in the whole chromosome (-0.23%).

3.3.2 Sequence anomalies

The average counts of oligonucleotide sequences and their respective reverse complementary sequences were used for analysis of the observed di-, tri- and tetranucleotide frequencies of the whole chromosome 2 as opposed to counts in open reading frames of the CDS subset.

3.3.2.1 Observed/Expected Dinucleotide ratios

The Observed/Expected (O/E) ratio gives an indication of the abundance of an oligonucleotide relative to the expected abundance (corrected for the high (A+T):(G+C) ratio) of the sequence subset analysed. Figure 3.1 shows that the GG•CC, TG•CA and AT dinucleotides occur at higher than expected levels (O/E values higher than 1), while CT•AG and CG have lower than the expected values in both coding and non-coding regions of chromosome 2. The dinucleotides GA•TC and TT•AA were relatively more abundant than expected in the coding subset but were less abundant than expected in the non-coding subset. The opposite held true for TA and GC, having O/E values higher than 1 in the non-coding subset but lower than 1 in the coding regions. The GT•AC dinucleotide pair occurred at the expected levels in the non-coding subset, but was less abundant than expected in coding regions.

The complete data for the actual and expected counts of dinucleotides in the whole chromosome 2 as well as the coding and non-coding regions are given in Appendix A.

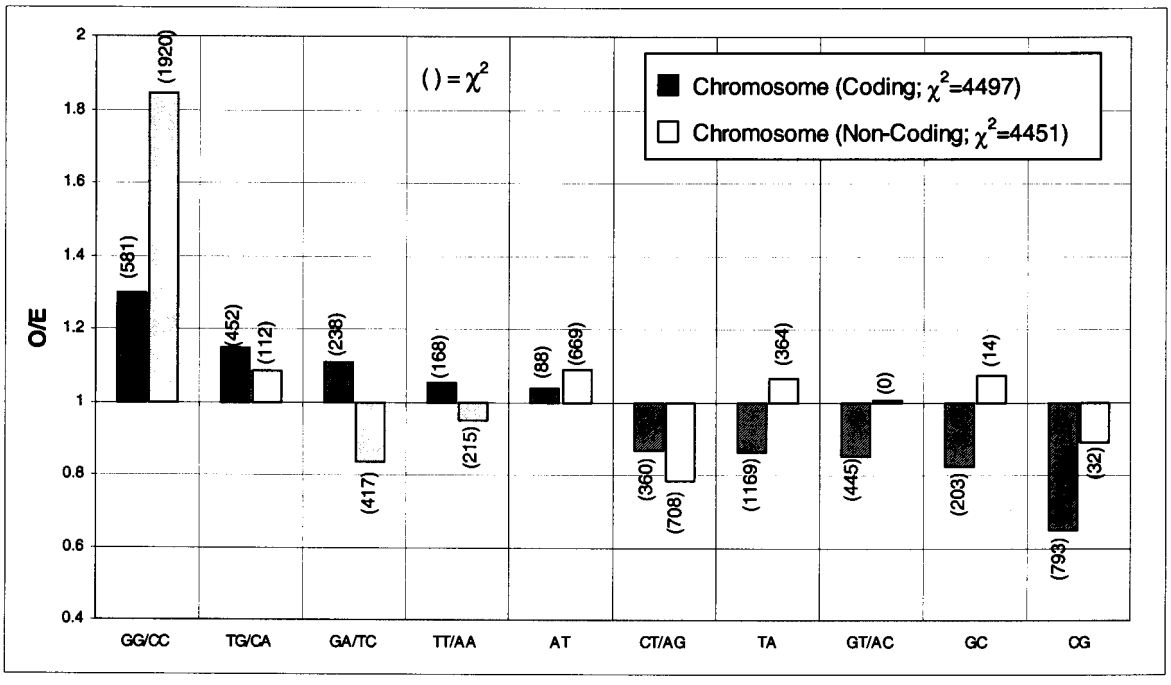

Figure 3.1: Observed dinucleotide frequencies over the expected frequencies in Chromosome 2 of coding and non-coding sequences. The O/E values for the coding and non-coding sequences are given separately, with the chi-square values in brackets. For the individual O/E values, there is one degree of freedom. Individual O/E values with χ^2 values higher than 11 will thus be significant, while χ^2 values less than 2.7 is not significantly different from the expected values. For the complete dinucleotide signature, there are eight degrees of freedom. A χ^2 value higher than 26.1 for each subset (coding and non-coding) will thus be significant at $p<0.001$.

3.3.2.2 Observed/Expected Tetranucleotide ratios

The relative tetranucleotide abundance showed similar trends as the relative dinucleotide abundance. The tetranucleotides in figure 3.2 are arranged according to the decreasing O/E ratios observed for dinucleotides in the coding region (Figure 3.1). For the individual tetranucleotides, there are three degrees of freedom. Observed over expected ratios with χ^2 values higher than 11 and 16 will thus be significant at $p<0.01$ and $p<0.001$, respectively for individual tetranucleotides. For the tetranucleotide signature there are 132 degrees of freedom and subsets (coding or non-coding) with χ^2 values higher than 188 will be significant at $p<0.001$. A relative over-representation of G+C rich tetranucleotides was found in all cases. The

oligonucleotide pairs with the highest O/E ratio always contained only G or C (GG•CC for dinucleotides and CCCC•GGGG for tetranucleotides).

Tetranucleotides containing the dinucleotide pair GG•CC (with the highest dinucleotide O/E ratio) are indicated with a black bar in figure 3.2. All GG•CC containing tetranucleotides exhibited O/E ratios higher than 1 (χ^2 from 0.3 to 1373.7) for the coding region, with the exception of ACCG•CGGT. The lower O/E ratio (0.95) of the ACCG•CGGT tetranucleotide pair appears to be associated with its dinucleotide components - CG and GT•AC - which had the lowest and third lowest O/E ratios, respectively.

The significantly lower than expected O/E ratios of the ACGG•CCGT ($p<0.001$), ACCG•CGGT ($p<0.001$), CCGC•GCGG ($p<0.01$), CGGC•GCCG ($p<0.01$), TCCG•CGGA ($p<0.001$) and TCGG•CCGA ($p<0.001$) tetranucleotides in the non-coding regions (indicated with ← in figure 3.2) appears to be associated with their CG dinucleotide component (O/E = 0.4) rather than their GG•CC (O/E = 1.9) or GC (O/E = 1.05) dinucleotide components. All CG containing tetranucleotides (indicated with * in figure 3.2) have O/E ratios of less than one for the non-coding regions and are always lower than the O/E ratios for the coding regions. The CG suppression effect is more pronounced in the non-coding than in the coding regions.

Eight tetranucleotides not containing a GG•CC dinucleotide component exhibited O/E ratios higher than 2 (indicated with † in figure 3.2). The elevated O/E ratio (2.2) of the AAAA•TTTT tetranucleotide pair ($p<0.001$) in the non-coding region does not correlate to the O/E ratios of its dinucleotide components TT•AA (O/E = 0.9). The high O/E ratios of ATAT (O/E = 2.9; $p<0.001$) and TATA (O/E = 2.7; $p<0.001$) in the non-coding regions, might be due to the presence of a number of long AT repeat sequences in the non-coding regions of chromosome 2 (Gardner *et al*, 1998). Four of the remaining five non-GG•CC tetranucleotides with O/E values higher than 2 contain the CA•TG dinucleotide, which has the second-highest dinucleotide O/E ratio (1.15).

These are CATC•GATG (O/E = 2.9; p<0.001), CAGC•GCTG (O/E = 2.5; p<0.001), CTGC•GCAG (O/E = 2.25; p<0.001) and GCAC•GTGC (O/E = 2.1; p<0.001). The fifth tetranucleotide, CTTC•GAAG (O/E = 2.95; p<0.001), contains the dinucleotide pairs - GA•TC and TT•AA with the third and fourth-highest dinucleotide O/E ratios.

The complete data of the occurrences of tetranucleotides in the whole chromosome 2 as well as the coding and non-coding regions are given in Appendix C.

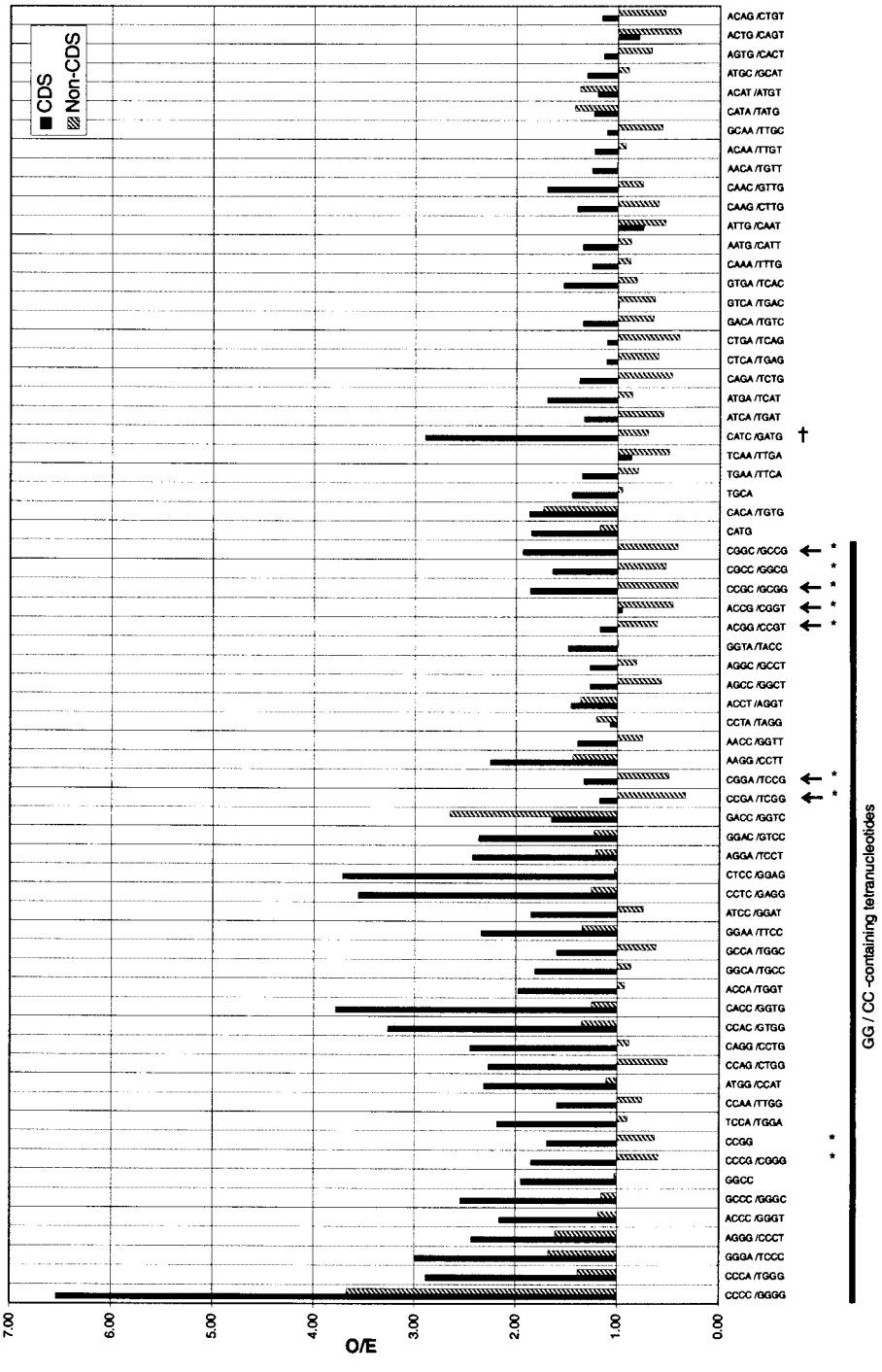

Figure 3.2: Observed frequencies divided by the expected frequencies of tetranucleotide sequences of Chromosome 2. The O/E values for the coding sequences (grey) and non-coding sequences (diagonal) are given separately. GG•CC containing tetranucleotides are indicated by the black bar; CG containing tetranucleotides are indicated with * and non-GG•CC containing tetranucleotides with O/E ratios higher than 2 are indicated with †. (Continued on next page)

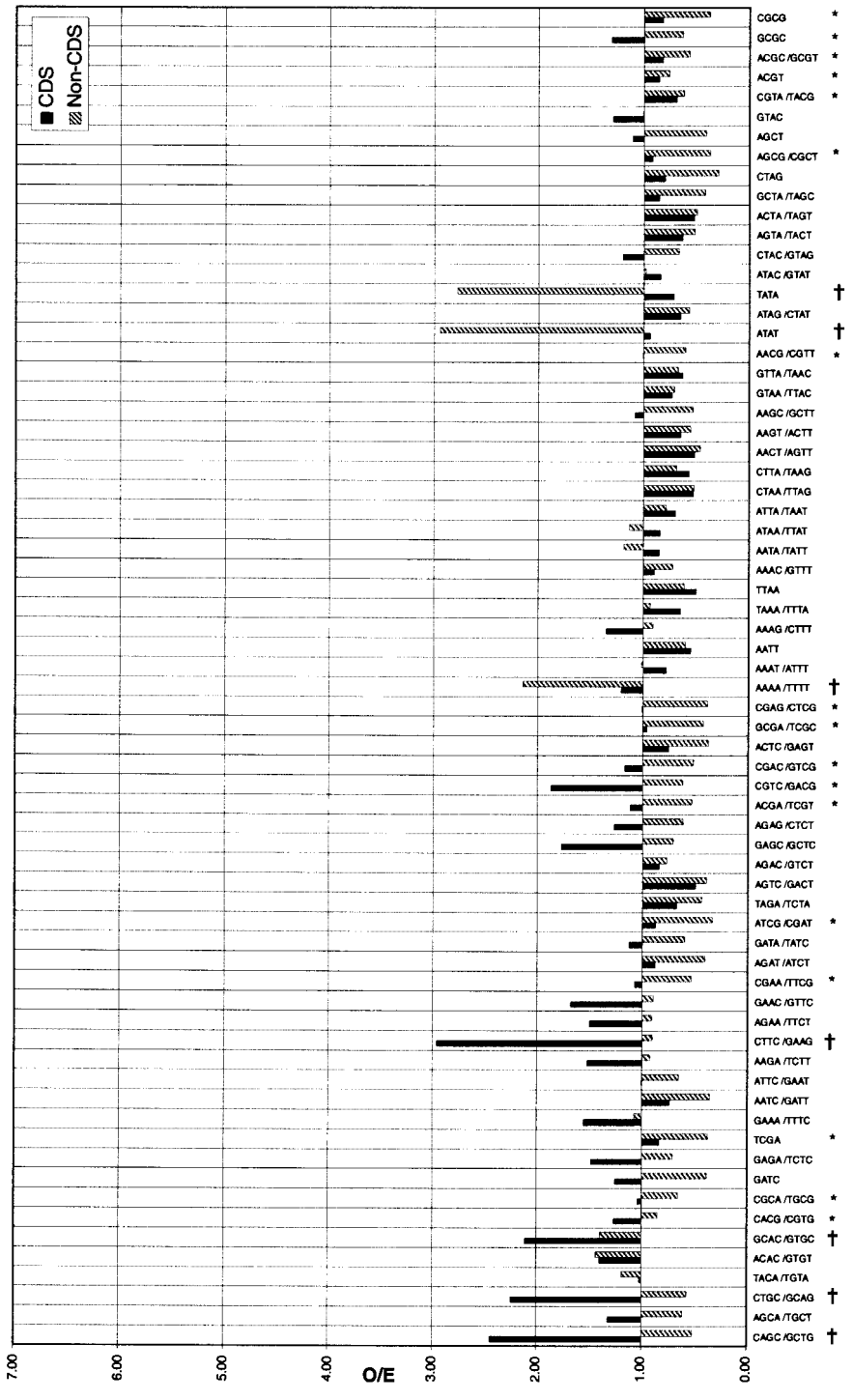

Figure 3.2(continued): Observed frequencies divided by the expected frequencies of tetranucleotide sequences of Chromosome 2. The O/E values for the coding sequences (grey) and non-coding sequences (diagonal) are given separately. CG containing tetranucleotides are indicated with * and non-GG•CC containing tetranucleotides with O/E ratios higher than 2 are indicated with †.

3.3.3 Amino-acid composition

As shown in Table 3.2 the most abundant amino acids coded for by chromosome 2 are asparagine (13.37%), lysine (11.87%) and isoleucine (9.05%). As expected for an A+T rich genome, all three these abundant amino acids are encoded by AT-rich codons in the malaria parasite (AAT for Asn; AAA for Lys and ATW for Ile). Tyrosine and phenylalanine appear to be exceptions since their counts are close to the expected mean of 5% even though their preferred codons (TAT and TTT, respectively) are also A+T rich.

Table 3.2: Amino-acid composition of putative proteins encoded by genes located on chromosome 2.

	Amino acid	Counts	% of Total	Codon for Amino Acid
1	Asparagine	19597	13.37	A A TC
2	Lysine	17394	11.87	A A AG
3	Isoleucine	13256	9.05	A T TCA
4	Leucine	11135	7.60	C T ATCG
5	Glutamic acid	10812	7.38	G A AG
6	Serine	9439	6.44	A G TC; T C ATGC
7	Aspartic acid	9159	6.25	G A TC
8	Tyrosine	8072	5.51	T A TC
9	Threonine	6425	4.38	A C ATCG
10	Phenylalanine	6323	4.31	T T TC
11	Valine	6028	4.11	G T ATCG
12	Glycine	4735	3.23	G G ATCG
13	Glutamine	4037	2.75	C A AG
14	Arginine	3810	2.60	A G AG; C G ATCG
15	Histidine	3540	2.41	C A TC
16	Methionine	3308	2.25	A T G
17	Alanine	3126	2.13	G C ATCG
18	Proline	2959	2.02	C C ATCG
19	Cysteine	2648	1.80	T G TC
20	Tryptophan	664	0.45	T G G
Total AA's counted:		146467	100	

It is expected that trinucleotides containing predominantly A or T will have a higher than mean abundance for the A + T rich genome of the malaria parasite. Likewise a lower than mean abundance is expected for C or G rich trinucleotides. The distribution of trinucleotide abundancies in chromosome 2 (Table 3.3) displays the expected higher abundance for A/T containing trinucleotides and lower abundance for C/G containing trinucleotides in the coding regions.

In Figure 3.3, the fractions of a trinucleotides in coding frame are compared to the codon preference. A randomly drawn trinucleotide is expected to have a fraction of 0.33 (indicated with an arrow in Figure 3.3). If a trinucleotide is drawn at random from the sequence, then the chance that the first base falls on the first, second or third codon position should be equal to 0.33. Trinucleotide fractions that are either higher or lower than this value can be an indication of constraints imposed by the trinucleotide base composition or the codon preference on trinucleotide abundance.

A factor causing the up-regulation of the relative abundance of an oligonucleotide is referred to as a positive constraint. A factor leading to the down-regulation of the relative abundance of an oligonucleotide is conversely referred to as a negative constraint. The codon preference fraction for equal usage of codons for a particular amino acid depends on the number of codons for that amino acid. Codons with usage fractions higher than this random value will have the greatest positive constraining influence on the nucleotide sequence. This conclusion will be confirmed if the fraction of trinucleotide in reading frame is also higher than the random value of 0.33. If the fraction of trinucleotide in reading frame is however lower than this random value it would indicate that the codon preference has little if any influence on the nucleotide sequence. Similarly, for codons with usage fractions lower than the expected value, the greatest negative constraints are expected. These negative constraints will be confirmed if the trinucleotide fraction in the reading frame is also lower than its random value.

Table 3.3: The abundance (counts) of trinucleotides in coding strands of the CDS-subset of chromosome 2.

Higher than mean abundance		Intermediate abundance		Lower than mean abundance	
AAA	42891	AAC	7704	CTT	4254
ATA	28441	CAT	7104	AGG	4220
AAT	26980	TGT	7104	TGG	4104
TAA	22022	GTA	6358	TCT	3967
TAT	21031	TCA	6239	CTA	3668
ATT	17037	ATC	5942	GTG	3420
TTA	15945	TAC	5759	GAG	3318
TTT	15762	TTG	5734	ACT	3304
GAA	15554	TAG	5600	GGT	2988
ATG	13928	GTT	5379	CCA	2956
AAG	12338	AGT	5265	ACC	2649
AGA	12139	GGA	5260	CAG	2539
TGA	11612	TTC	4991	GCA	2418
GAT	10027			TCC	2355
ACA	9893			CAC	2267
CAA	9596			TGC	2230
				CCT	2214
				GAC	2166
				AGC	2150
				CGA	2089
				ACG	2050
				CTG	1968
				GCT	1822
				GGG	1539
				GTC	1445
				CTC	1416
				CGT	1245
				TCG	1233
				CCC	841
				GCC	642
				GGC	638
				CGG	561
				GCG	480
				CCG	476
				CGC	344

Mean abundance	6869
$p < 0.1$ (confidence of 90%)	± 1935
Mean + (1 confidence interval)	8804
Mean - (1 confidence interval)	4934

The trinucleotide fractions in the coding frame of the three stop-codons have low values, as expected, since each open reading frame (ORF) contains only a single stop-codon. All the CG-containing trinucleotides (lowest dinucleotide O/E values; indicated with ✠ in figure 3.3) have fractions less than 0.36 in the reading frame. The codon preference fractions of CG-containing codons are correspondingly, also lower than the random fraction of these codons. This indicates negative constraints by the amino acid abundance and coding preference on the nucleotide sequence (resulting in the low O/E values for CG).

Four of the fourteen GG•CC (O/E = 1.3) containing codons have both a higher than random codon preference fraction and a higher than random trinucleotide fraction in the coding frame (marked with ✪ in Figure 3.2). This indicates a constraining effect by these codons on the nucleotide sequence. Since Met (ATG) and Trp (TGG) both have single codons their usage fractions are one. The trinucleotide fraction in reading frame in both cases is however less than 0.33. No conclusions regarding the constraining effects of their codons on either the nucleotide sequence or amino acid composition can thus be drawn.

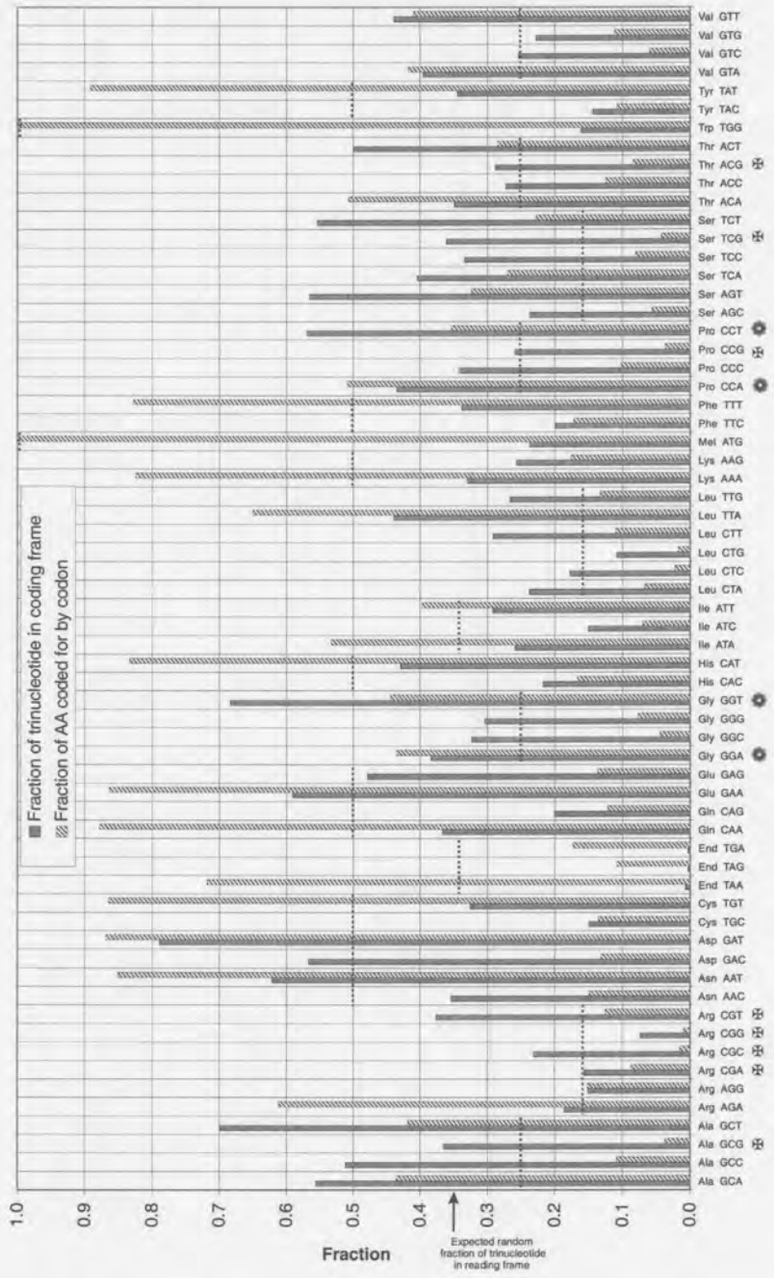

Figure 3.3: Correlation between codon preference and in-frame trinucleotides. The fraction of trinucleotides that occur in the coding frame is shown with grey bars and was calculated by dividing the codon abundance by the sum of the corresponding trinucleotide abundance in frames 1, 2 and 3. The fraction of the codon for the indicated amino acid is shown with diagonally striped bars. Codons that constrain CG and GG•CC dinucleotides are marked with ✠ and ✪ respectively. The fraction for random distribution of trinucleotides is indicated by an arrow. The fractions for randomly distributed codons are indicated by dotted lines.

3.3.4 The A+T content of genes encoding selected antigens

Genes encoding the antigens MSA2, HRP2 and CSP were compared to the coding regions of chromosomes 2 and 3 (Bowman *et al*, 1999; Gardner *et al*, 1998; Table 3.4).

Table 3.4: A+T content of chromosomes and antigen-encoding genes

	A+T content
Chromosome 2 (Coding region)	75.7%
Chromosome 3 (Coding region)	76.4%
Circumsporozoite protein (CSP)	65.2%
Histidine rich protein 2 (HRP2)	45.6%
Merozoite surface antigen 2 (MSA2)	58.8%

The A+T content of all three genes were lower than that of the coding regions of chromosomes 2 and 3. The CSP and MSA2 genes (65.2 and 58.8% A+T, respectively; Table 3.4) exhibited a higher A+T content than the HRP2 gene.

Comparison of the genomic signatures of the same genes with that of coding regions of chromosomes (Figure 3.4), yielded even greater differences between the three sequences. The genomic signatures for both chromosomes as well as the individual genes exhibited a lower than expected CG content (O/E < 1). The discriminating dinucleotide pair of the genomic signature was GG·CC. The GG·CC O/E ratio of CSP correlated with that of the whole chromosomes (all with significantly higher O/E values; $p<0.001$). The GG·CC O/E values of both MSA2 (1.05) and HRP2 (0.65) however differed from that of the whole chromosomes (GG·CC O/E = 1.3; Figure 3.4).

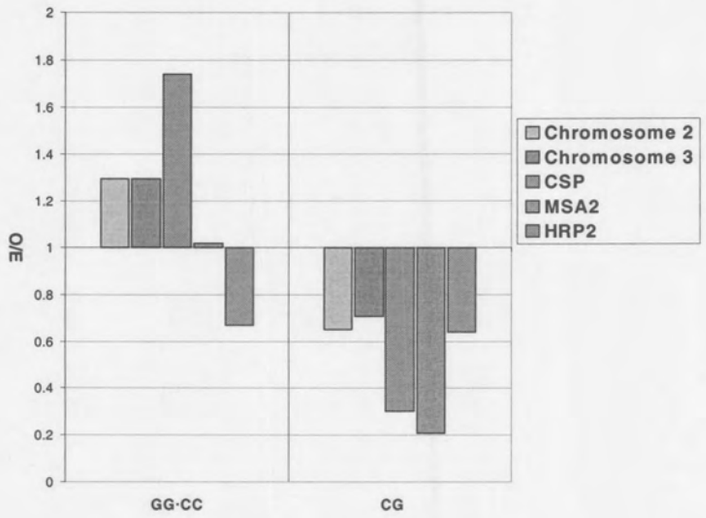

Figure 3.4: Genomic signature extremes (highest and lowest O/E values) of the CDS subset of chromosomes 2 and 3 compared to genes of CSP, HRP2 and MSA2.

3.4 Discussion

The genomic signature of an organism is an array of all its dinucleotide relative abundance values (O/E-values; Karlin *et al*, 1995) and is a valuable tool for studying evolutionary relationships between species and other organisms. The identification of constraints that underpin organism-specific genome signatures are important to understand as it will give us better insights in the factors influencing the regulation of transcription as it reflects the species-specific properties of DNA-modification, replication and repair mechanisms (Karlin *et al*, 1995).

The 80.6% A + T content in chromosome 2 reflected the composition of both coding as well as non-coding regions. The previously reported values of 69% A + T for coding and 86% A + T for non-coding regions from a smaller dataset (Weber, 1987) correlate with the experimentally obtained values of 75.7% A + T for the coding regions and 85.1% A + T for the non-coding regions of chromosome 2.

The malaria genome is ~80% A + T rich and it is predictable that certain A + T -containing sequences will occur more frequently than others will. The dinucleotide O/E ratios (corrected for the A + T rich composition) obtained for chromosome 2 are in agreement with values obtained in previous studies (Hyde *et al*, 1987; Weber, 1987). The significant GG•CC elevation ($p<0.001$) observed in both coding and non-coding regions, was similar to the data reported by Hyde *et al* (1987). However, Weber (1987) found no significant difference between the observed and expected values.

The elevated O/E values of the GG•CC dinucleotide pair, observed in coding and non-coding regions, are similar to those of eukaryotes, thermophylic prokaryotes and mitochondrial genomes (Karlin *et al*, 1992; Karlin, *et al*, 1995). These findings link *Plasmodium* as

phylogenetically closer to the higher eukaryotes than the lower eukaryotes and prokaryotes. The trinucleotide relative abundance data (Appendix B) further support these findings since elevated O/E values for GG•CC containing trinucleotides were also observed for chromosome 2.

The tetranucleotides with the highest O/E ratios in the coding regions were CCCC•GGGG (O/E = 6.5) and CACC•GGTG (O/E = 3.8). A good correlation was evident between these tetranucleotides and their corresponding dinucleotide components. For example, the CCCC•GGGG sequence with the highest O/E value (6.5) in the coding region, correlated with its dinucleotide component GG•CC, having the highest O/E ratio (1.3) of all the dinucleotides.

The lower than expected O/E ratio in the non-coding region for the three tetranucleotide pairs, ACCA•TGGT (0.9), GGCA•TGCC (0.8) and GCCA•TGGC (0.6) can not be explained by constraints imposed by their dinucleotide components – all of which have neutral or elevated O/E ratios (ranging from 0 to 1.85) in the non-coding regions. Other functional or structural constraints, as yet still unidentified, seem to be responsible for the lower O/E ratios of these tetranucleotides.

The trinucleotides with the highest abundance contain only A/T-bases or two A/T and a single G/C base, as expected (Table 3.3). Correspondingly, trinucleotides containing two or more G/C bases have the lowest abundance. The strand-asymmetry is also reflected by A- and G-containing trinucleotides having higher abundancies than the corresponding T- and C-containing trinucleotides (e.g. AAA having an almost three-fold higher abundance than TTT).

The dominant constraining factor on amino acid composition appears to be the high A + T content (75.7% for the CDS) selecting preferentially for A or T rich codons. This A + T skewness is not only evident in the higher abundance of amino acids coded for by A+T rich

codons (Table 3.2), but also in the codon preference of the amino acids. This is most notable on the third codon-position with, for example GCA and GCT being the most preferred codons encoding Alanine. Similar conclusions can be drawn for all other amino acids coded for by more than one codon.

The relative overrepresentation of GG•CC appears to be determined by the codons used by the malaria parasite for glycine (GGN) and proline (CCN; Figure 3.3). This can be inferred as both the codon preference fraction and the trinucleotide fraction in the coding frame are higher than their respective random values. The high A + T content of the malaria genome however tends to mask the constraints imposed by these codons due to the low abundance of the corresponding amino acids (Table 3.2). The actual abundance of GG•CC is thus low - in spite of its high O/E value (1.3). These results suggest that the relative overrepresentation of GG•CC can be attributed to constraints imposed by the amino acid content of the organism.

The TA-suppression observed in the coding regions corresponds to the earlier data, but the TA-elevation in the noncoding region is in contrast with the reported insignificant or lower than expected values (Hyde *et al*, 1987; Weber, 1987). These differences are probably due to the use of a larger dataset that is more representative of both coding and non-coding regions of the parasite genome. The TA-suppression found in the coding regions also correlate with the almost universal TA-suppression found in all previously studied organisms (Karlin *et al,* 1992; Karlin *et al*, 1995).

This universal TA-suppression phenomenon has been postulated to be due to conformational preferences and low stacking energy of this dinucleotide (Karlin *et al*, 1995). This explanation is however inadequate to account for the marked difference in TA -frequency between the coding (O/E = 0.76) and non-coding regions (O/E = 1.19) of chromosome 2. The proposal by Karlin *et al* (1992) that the prominent regulatory role of the "TATA"-box imposes a functional constraint to

minimise inappropriate binding of transcription and termination factors, seems even more unlikely (O/E = 2.7 in non-coding regions). A more likely explanation for the high TATA levels in chromosome 2 of *Plasmodium falciparum*, is the presence of long stretches of TA - repeats in the sub-telomeric regions (Gardner *et al*, 1998).

Two of the eight TA-containing codons (TAA and TAG) are stop-codons. The fractions of the corresponding trinucleotides in the reading frame are minimal as expected. No discernible pattern was evident from correlation of the fractions of trinucleotide in reading frame with the remaining six TA-containing codons. The lower relative abundance of TA is thus constrained by other as yet still unidentified factors.

Vertebrate genomes tend to show a dramatic deficiency in CG dinucleotides, which is a likely consequence of high levels of DNA methylation. This involves the methylation of C in the CG dinucleotide, which greatly elevates the mutation rate of C to T through spontaneous deamination of the resultant 5-methylcytosine (Sved *et al,* 1990). Thus, the relative frequency of CG dinucleotide in genes or genomes can be taken as an index of the methylation effect. DNA methylation is a ubiquitous biochemical process that has also been observed for the malaria parasite. Methylation in *Plasmodium* spp. was found to occur mainly in association with CG sequences (Pollack *et al*, 1991). The CG dinucleotide frequency was significantly lower than expected ($p<0.001$) in both the whole chromosome 2 as well as in its coding and non-coding subsets (O/E=0.65 - coding; O/E=0.90 - non-coding). A corresponding elevated TG·CA level is expected if the lower than expected CG frequency is as a result of a methylation effect. Elevated levels of the TG·CA dinucleotide pair (O/E=1.15 - coding; O/E=1.10 - non-coding) thus appear to support the methylation hypothesis. The CG - suppression corresponds to that encountered in higher eukaryotes and some mitochondrial genomes whereas it is rarely observed in prokaryotes (Karlin *et al*, 1992; Karlin *et al*,1995). The fact that CG-suppression

was much more pronounced in the coding regions might be due to the fact that methylation as regulatory signal is of more importance in coding regions.

The constraints imposed by the low O/E ratio of the 26 CG containing tetranucleotides are however more pronounced in the non-coding regions than in coding regions. For example, the CGCG tetranucleotide has O/E values of 0.4 and 0.8 in the non-coding and coding regions, respectively. All eight CG-containing codons are the least preferred codons for their respective amino acids, having codon preference fractions lower than expected for random codon usage. Their corresponding trinucleotide fractions in the reading frame are also either at random or lower values. This indicates definite negative constraints on the nucleotide sequence imposed by the codon preferences of the malaria parasite. Three of the five amino acids coded for by CG-containing codons are furthermore less abundant than expected (Arg - 2.6%; Ala - 2.13% and Pro - 2.02% of total amino acid composition). The CG suppression effect in coding regions thus appears to be related to constraints imposed by both the amino acid abundance and codon preference (Figure 3.3). This indicates that the "methylation"-hypothesis (Karlin *et al*, 1995) is unlikely to be the sole reason for the avoidance of CG dinucleotides in the malaria genome. The constraints imposed on the CG suppression signature in the non-coding region thus are still not resolved and needs to be addressed in other studies.

The amino acid abundancies in Table 3.2 correlate well with previous results obtained by Hyde *et al.* (1987). Two exceptions are however evident. Isoleucine and alanine comprised 9.05% and 2.13 % of the total amino acid abundance, respectively, compared to 5.41% and 6.54% found by Hyde *et al* (1987). A better correlation is demonstrated in this study between the A + T content of their respective codons (ATH - Ile and GCN - Ala) and actual abundance. These differences can probably be attributed to the larger and more representative set of proteins used in our study since large variations in amino acid composition are apparent between individual malaria proteins (Appendix D).

Tyrosine and phenylalanine appear to be an exception to the constraints imposed by A + T composition, since their counts are close to the expected mean of 5% even though their preferred codons (TAT and TTT, respectively) are also A+T rich. The strand asymmetry (18.54) in the coding regions reflects the preference for codons containing purine bases (A and G) and probably constrain the abundance of these amino acids.

Codons for hydrophilic amino acids are more A+T rich than codons for hydrophobic amino acids (Verra *et al*, 1999). A preference for hydrophilic amino acids over hydrophobic amino acids is thus expected for the malaria parasite due to its A +T content. The total amino acid content of proteins coded for by chromosome 2 (Table 3.2), showed the expected higher abundance of hydrophilic amino acids and lower abundance of hydrophobic amino acids. A recent study of antigenic repeat regions of the malaria parasite showed a significant reduction of hydrophobic amino acids but not the expected higher abundance of hydrophilic amino acids (Verra *et al*, 1999). The reason for these differences is not readily apparent but may be due to the larger size and more representative composition of the set of proteins used in this study.

The relative abundancies of the short oligonucleotide sequences (O/E values) of chromosome 2 provide a unique genomic signature of the parasite genome (Figure 3.1). The high O/E value of GG•CC (1.3) and low O/E value of CG (0.65) are the most prominent features of this signature. The low relative abundance of CG corresponded best with that of higher eukaryotes (Human, *Xenopus*, etc) and almost all mitochondrial genomes. The high relative abundance of GG•CC however is found almost universally in chloroplasts and mitochondria and in some viri (e.g. Cytomegalo virus and Epstein-Barr virus; Karlin *et al*, 1992; Karlin *et al*, 1995). This signature, although showing close links to its human host (low CG O/E value), places *P. falciparum* closer to intracellular organelles and viri. This is not unexpected as genomic signatures have been postulated to be influenced by the organism's environment (Karlin *et al*, 1995).

The constraining factors influencing the genomic signature of the malaria parasite have not been fully elucidated. The high relative abundance of GG•CC was shown to be positively constrained by the parasite's codon preference and amino acid abundancies. The other prominent characteristic of this signature - the low relative abundance of CG – was also shown to be negatively constrained by the parasite's codon preference and amino acid abundancies. This signature can not only be employed in tracing evolutionary roots of the malaria parasite as a whole, but also by studying the origins of individual genes based on their signatures compared to that of other organisms. The prominent differences between the genomic signatures of the coding and non-coding regions may prove to be an useful tool in the identification of ORF's.

Lateral gene transfer has been shown to be an important factor in the diversification of virulence in a number of bacterial pathogens (Ochman *et al*, 2000). Laterally transferred genes can be identified by analysis of the A+T content of DNA. The DNA containing transferred genes can then be identified as regions with A+T content different from that of the average of an organism (Wren *et al*, 2000).

The A+T content of genes encoding the antigens CSP, HRP2 and MSA2 were found to differ significantly from the average A+T content of the *P. falciparum* genome (Table 3.4). The A+T content of the histidine rich protein (45.6%) was clearly different from that of chromosomes 2 and 3 (~75%). For the circumsporozoite protein (65.2% A+T) and merozoite surface antigen 2 (58.8% A+T), the case is however less clear-cut. Although their A+T content were somewhat lower than that of chromosomes 2 and 3, it is not evident from base content alone whether these genes could have been transferred from other sources. These genes contain repetitive polymorphic regions and it is possible that the A+T content can be skewed by the codon preference of repetitive amino acids (e.g. high Histidine content of repeats in HRP2). Genomic

signatures (Figure 3.4) are not skewed due to normalising for base composition and thus a better measure for comparative purposes.

It is also clear from Figure 3.4 that only the circumsporozoite protein (CSP; found on chromosome 3) conforms to these genomic signature extremes. The signature for the histidine rich protein 2 (HRP2; found on chromosome 8) exhibited a low GG·CC O/E value (0.65), considerably different from those of chromosomes 2 and 3 (GG·CC, O/E = 1.3). It is thus possible that the parasite might have acquired the HRP2 protein by means of lateral transfer from another unknown source. This conclusion is supported by recent evidence for the spontaneous uptake and expression of DNA in erythrocye cultures of the malaria parasite (Deitsch *et al*, 2001). The GG·CC signature extreme for the merozoite surface antigen 2 (MSA2; found on chromosome 2) also exhibited a similar deviation from the genomic signature extremes of chromosomes 2 and 3, although less than HRP2. The CG genomic signature extreme was found to be lower in all cases (both in the antigens and whole chromosomes). The GG·CC and CG genomic signature extremes of chromosome 2 showed a good correlation with that of chromosome 3 (Figure 3.4). The possible representative nature of the signature throughout the whole genome is thus corroborated.

The genomic signature can not only be employed to trace the evolutionary roots of the parasite - linking it as closer to the higher eukaryotes, but it is also a powerful tool for studying the origins of individual genes in an organism. In the next chapter the occurrence of interspersed repeat oligonucleotide sequences in genes coding for prominent malaria antigens was investigated.

W 12	W 13	W 14	W 15	W 16	W 17	W 18	W 19	W 20
100	100	100	100	100	70	70	70	70
100	100	100	100	100	70	70	70	70
100	100	100	100	100	70	70	70	70
100	100	100	100	100	70	70	70	70
70	60	40	30	0	0	0	0	0
100	100	100	100	70	70	70	70	70
80	70	70	70	0	0	0	0	0
100	100	100	100	60	60	60	60	50

W 21	W 22	W 23	W 24	W 25	W 26	W 27	W 28	W 29	W 30
70	70	70	70	70	70	70	70	70	70
70	70	70	70	70	70	70	70	70	70
70	70	70	70	70	70	70	70	70	70
70	70	70	70	70	70	70	70	70	70
0	0	0	0	0	0	0	0	0	0
70	70	40	30	30	20	20	10	10	10
0	0	0	0	0	0	0	0	0	0
50	50	40	30	30	10	10	10	10	10

W 31	W 32	W 33
70	70	70
70	70	70
70	70	70
70	70	70
0	0	0
0	0	0
0	0	0
10	10	0

CHAPTER 4

ANALYSIS OF REPETITIVE SEQUENCES IN MALARIA ANTIGEN MSA-2

4.1 Introduction

The malaria parasite evades the immune system by antigenic variation and polymorphism (Newbold, 1999). The immune response of the human host to malaria differs from its immune response to viral or bacterial infections. Protective immunity to malaria is only developed after exposure to multiple infections over a period of several years. Non-immune individuals have very high levels of antibodies to repetitive immunodominant antigens and exhibit an associated hypergammaglobulinemia, slow development of unstable immunity and autoantibody production. These observations suggest that hyperstimulation of irrelevant B-cell responses leads to a less effective response against critical, essential epitopes. As the repetitive regions on many antigens are highly immunogenic, they compete with the host's protective immune responses. In a situation of immune pressure, gene duplication and variation by random mutations in repetitive sequences would increase the probability of immune evasion. This "smokescreen" effect is hypothesised to be due to the fact that the immunodominant repeat sequences in antigens divert the immune response away from the more critical epitopes (Kemp *et al,* 1987).

Antigens such as *Plasmodium falciparum* erythrocyte membrane antigen 1 (PFEMP1) coded for by the *Var*-gene family are presented by the parasite on the host cell membrane to further the sequestration of infected erythrocytes and to protect it from detection and destruction by the reticuloendothelial system of the host. Although this shields the parasite from destruction in the spleen, it is rendered vulnerable to the immune system of the host (Saul, 1999; Borst *et al,*

1995a; Mitchell, 1989). *Var*-genes are a diverse family of 50 to 150 genes whose expression switches between different members of the family over the course of an infection. Intramolecular recombination within *var*-gene clusters results in both switching of *var*-gene expression and changes in the *var*-gene repertoire. This brings about changes in the antigenic and cytoadherent properties of the infected cells, enabling the parasite to limit exposure of antigens to the immune system of the host (Deitsch *et al*, 1999). It was proposed that the expression of these antigens is a deliberate mechanism of the parasite to modulate parasite growth in order to establish a chronic infection and increase its chances for further transmission via subsequent infections of its mosquito vector (Saul, 1999).

Three basic mechanisms of antigenic switching were observed in large families of surface antigens like the *var*-genes described above. The first method is pre-transcriptional (replacement of an antigen) by means of gene conversion or reciprocal recombination without altering the promoter. The second method is transcriptional, promoter-switching between the promoter of the previously active antigen gene and a previously silent gene. The last method of antigen switching occurs on the post-transcriptional level, in changing the reading frame of the antigen mRNA. Another speculated method comprises the regulation at RNA expression level, with the differential degradation of the mRNA species coding for the antigens (Newbold, 1999; Brannan *et al*, 1994).

Whereas the *Var*-gene antigens are a multi-copy gene family, polymorphic antigens are encoded by single copy genes. Several polymorphic antigens are being investigated as vaccine candidates. Immunisation with antigen cocktails selected from Merozoite surface antigen 1 (MSA1), Merozoite surface antigen 2 (MSA2), Ring-infected surface antigen (RESA), Histidine-rich protein (HRP), Plasmodium falciparum Erythrocyte membrane protein 1 (PfEMP1), Circumsporozoite protein (CSP) and Apical membrane antigen 1 (AMA1) is currently under investigation in a number of trials in progress (Cooke, 2000). Although there are a number of

candidate vaccines in various stages of testing, a truly effective vaccine appears to be still far away. A feature lacking in all these vaccine candidates is a long-lasting protective immune response. The success of protection offered by conserved, non-immunodominant regions of individual candidate vaccines in earlier vaccine trials varied greatly (Brown, 1992).

Merozoite surface antigen 1 (MSA1) can be broadly divided into three allelic types (MAD20, K1 and RO33 types; Da Silveira, 1999; Miller *et al*, 1993). Tanabe (1987) proposed a division of the antigen into 17 distinct blocks (Figure 4.1) according to the level of conservation in each block. Blocks 1, 3, 5, 12 and 17 are conserved blocks between all allelic types. The rest of the sequence was then divided into semi-conserved blocks (blocks 7, 9, 11, 13, and 15) and variable blocks (blocks 2, 4, 6, 8, 10, 14 and 16). Both the semi-conserved as well as the variable blocks can be divided into two distinct dimorphic groups according to the different allelic types.

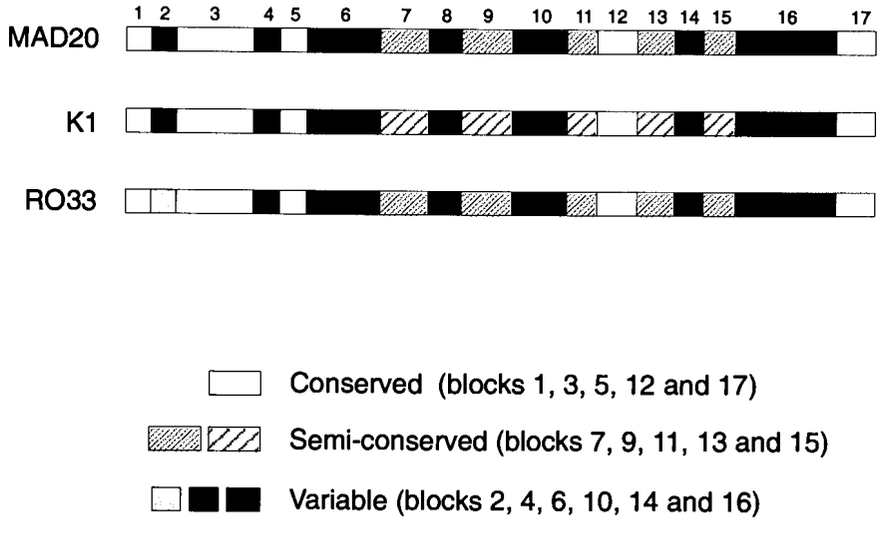

Figure 4.1 The three allelic types of Merozoite surface antigen 1 (MAD20, K1 and RO33). Conserved regions are in white, semi-conserved regions are hatched and variable regions are shaded (adapted from Tanabe, 1987).

Merozoite surface antigen 2 (MSA 2) have only two central variable repeat regions in blocks 2 and 3 and can be divided into two main allelic types (3D7 and FC27; Ranford-Cartwright *et al*, 1993). The repeat-blocks consist of family-specific, non-repetitive regions as well as allele-specific repetitive regions. The central repetitive region is flanked by conserved regions (Figure 4.2).

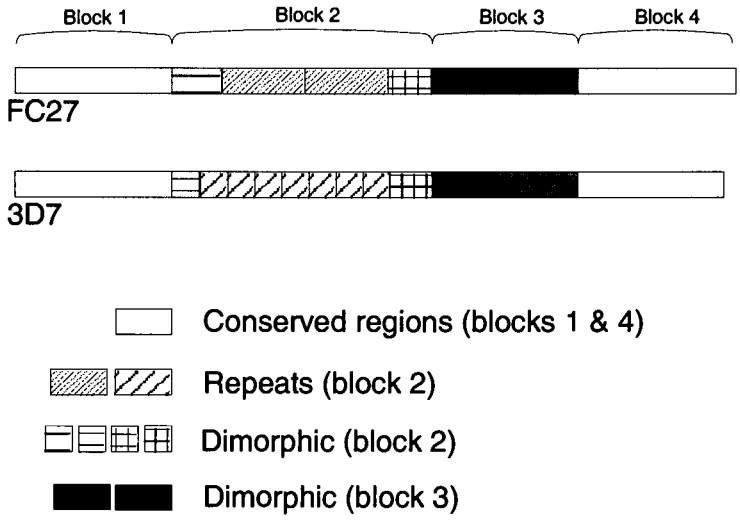

Figure 4.2 The two allelic types of Merozoite surface antigen 2 (FC27 and 3D7). Conserved regions are in white, family-specific, non-repetitive regions are hatched and allele-specific repetitive regions are shaded (adapted from Ranford-Cartwright *et al*, 1993).

The circumsporozoite protein (CSP) has a single central repetitive region flanked by conserved regions as shown in Figure 4.3 (Contamin *et al*, 1995). The histidine-rich protein 2 (HRP-2; Figure 4.4) has a similar structure to that of the CSP protein, with a central region containing variable polymorphic repeats bordered by constant regions (Ranford-Cartwright *et al*, 1993).

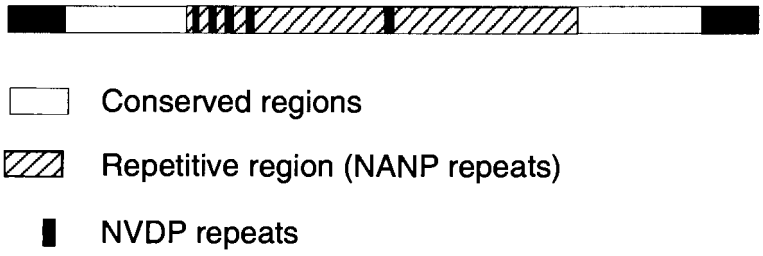

Figure 4.3 Organisation of the circumsporozoite protein, indicating the conserved regions flanking a repetitive central region. The repetitive blocks (NANP repeats) are represented as hatches with the variant repeat NVDP (one-letter amino acid code) as black bands (adapted from Contamin, 1995).

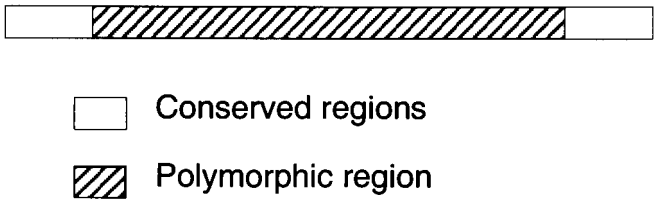

Figure 4.4 The Organisation of the histidine-rich protein 2 (HRP-2), indicating the conserved regions bordering a central histidine-rich repetitive region (adapted from Ranford-Cartwright *et al*, 1993).

A common feature shared between these polymorphic antigenic proteins is amino-acid repeats in immunodominant regions. Understanding of the molecular mechanisms by which the parasite achieves both the variation in type and length of repeat will be integral to the design of a successful vaccine.

Several authors (Ranford-Cartwright *et al,* 1999; Rich *et al*, 2000a and Conway *et al*, 1999) suggested that polymorphisms are produced by several recombination mechanisms. No recombination motifs or hot-spots were however identified. There are two main types of intragenic recombination events. Intrahelical recombination refers to events like slipped-strand mispairing, that is thought to be the principal source of variation in repetitive sequences of the

malaria parasite (Rich *et al*, 2000b). Interhelical recombination refers to interhelical crossover events taking place during the sexual replication cycle in the mosquito vector (Conway *et al*, 1999). The exact molecular mechanism that enables the malaria parasite to rapidly modify its antigenic profile is still unknown.

Translational anomalies like readthrough and frameshifting, allowing organisms to adapt more rapidly to changing environments, were shown to be associated with interspersed palindromic AGCT tetranucleotide pairs in *E. coli* and *B. subtilus* (Henaut *et al,* 1998). The distribution of tetranucleotide motifs in these organisms was postulated to present binding sites for a DNA-binding protein that guide the transition between aerobic and anaerobic growth conditions. The palindromic nature of these tetranucleotides was directly implicated in recombination mechanisms.

In this chapter, we analysed the occurrence of interspersed tetranucleotide pairs in a collection of genes and ESTs of the malaria parasite deposited in various databanks. Subsequently, genes exhibiting higher than random occurrence of interspersed repeats were identified. The localisation of one of these interspersed repeats in relation to the allelic differences between MSA-2 antigen sequences was further analysed to identify potential hotspots for variation and recombination events in this *P. falciparum* surface antigen protein.

4.2 Methods

4.2.1 Nucleotide sequences used

A list of malaria sequences and expressed sequence tags maintained at the department of Microbiology, Monash University and the Walter and Eliza Hall Institute of Medical Research, Australia was used. This dataset of sequences is representative of genome sequences of all chromosomes. The coding strands for all the sequences in the dataset are however not known. (ftp://ftp.wehi.edu.au/pub/biology/who_dbase/current_version/pcnuc.dat; downloaded Dec 1997)

4.2.2 Interspersed tetranucleotides:

The sequence subset was screened for possible clustering of tetranucleotide sequences by analysis of the frequency of interspaced tetranucleotide pairs in the malaria parasite genome (e.g. $TGCA(N)_xTGCA$ where x=2,5,8,..). The frequency of these interspersed repeats was determined using the interspersing search function of the SeqSearch program (Appendix E). The program performs regular expression text string searching functions on the sequence subset in a plain text format.

The obtained results were then compared with theoretically expected frequencies for a subset of randomly distributed nucleotides of similar size and A+T content. Sequences containing selected interspersed repeats were identified and grouped according to function (Appendix G).

Intervening regions (i.e. the $(N)_x$ part of the $TGCA(N)_xTGCA$ motif) of selected sequences were then inspected by eye for similarities of interspersed sequences.

4.2.3 Translation, ORF-searching and Alignment of Sequences

Merozoite surface antigen 2 (MSA-2) type FC27 sequences were translated in all six possible reading frames, using the PROPHET software package (designed for the National Institute of Health by BNN Software Corporation) and analysed for open reading frames. The translated products were further inspected by eye to determine the reading frame in which the interspaced repeats occurred as well as to locate any possible frameshift points. The genes were also analysed for the presence of both direct as well as inverse repeat sequences by searching the sequence subset, using the repeat-searcher function of the SeqSearch program (Appendix E).

Sequences representing the different allelic forms of MSA-2 type FC27, containing multiple occurrences of interspaced tetranucleotide repeats were compared by multiple alignment with the ClustalW program.

4.3 Results

4.3.1 Interspersed tetranucleotides

Counts of tetranucleotide sequences with (3n-1) interspersing were found to be significantly higher than the background in the case of the interspaced TGCA and CATG nucleotide sequences (Figure 4.5). The marked increase above the background is prominent with regular interspacing lengths of 2, 5, 8, 11, 14, 17 or 20 base pairs.

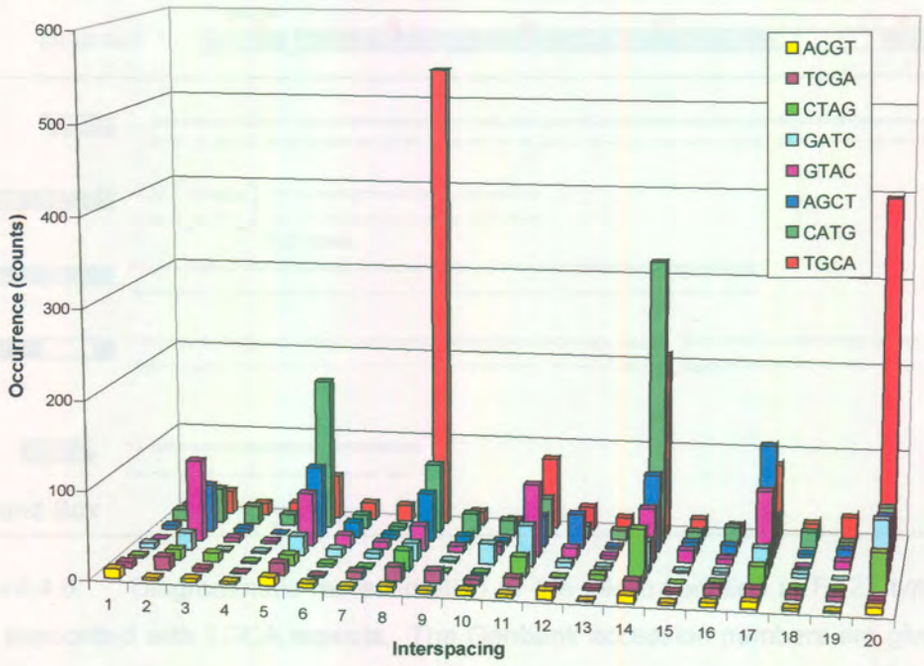

Figure 4.5: Interspacing of palindromic tetranucleotide pairs. The frequency of interspaced tetranucleotide pairs peaked with an interspacing of (3n-1).

A total of 96 sequences exhibited higher than expected interspaced TGCA frequencies (Appendix F). The interspaced TGCA frequencies in three genes encoding prominent antigenic proteins are given in Table 4.1 and compared with the expected frequency normalised for A+T

content (calculated by equations given in Chapter 3). The distribution of TGCA interspaced in the merozoite surface antigen 2 (MSA-2; AF033859 gb) was selected for further analysis in this study. This interspersed repeat was also localised to the polymorphic repeat regions of the sequences coding for the CSP and HRP-II proteins (Figures 4.3 and 4.4, respectively) but is not addressed in this study.

Table 4.1: Three prominent antigenic proteins selected from the list of Proteins, Sequence tags and Contigs containing interspersed TGCA repeats (Appendix F).

	Protein	Total bp in gene	Instances of TGCA-repeats	Frequency (/10^6 bp)	Observed/ Expected fraction
1	>gi|488899|emb|A13159.1|A13159 P. falciparum HRPII antigen	631	29	45959	1.93
2	>gi|294141|gb|M83150.1|PFACSPM Plasmodium falciparum circumsporozoite protein (CSP) gene, complete cds	1315	38	28897	1.90
3	>gi|2645872|gb|AF033859.1|AF033859 Plasmodium falciparum merozoite surface protein 2 gene, partial cds	444	5	11261	1.67

4.3.2 Translation and ORF-searching

Interspersed TGCA repeats for different allelic forms of genes of MSA2 type FC27 proteins are compared in Figure 4.6.

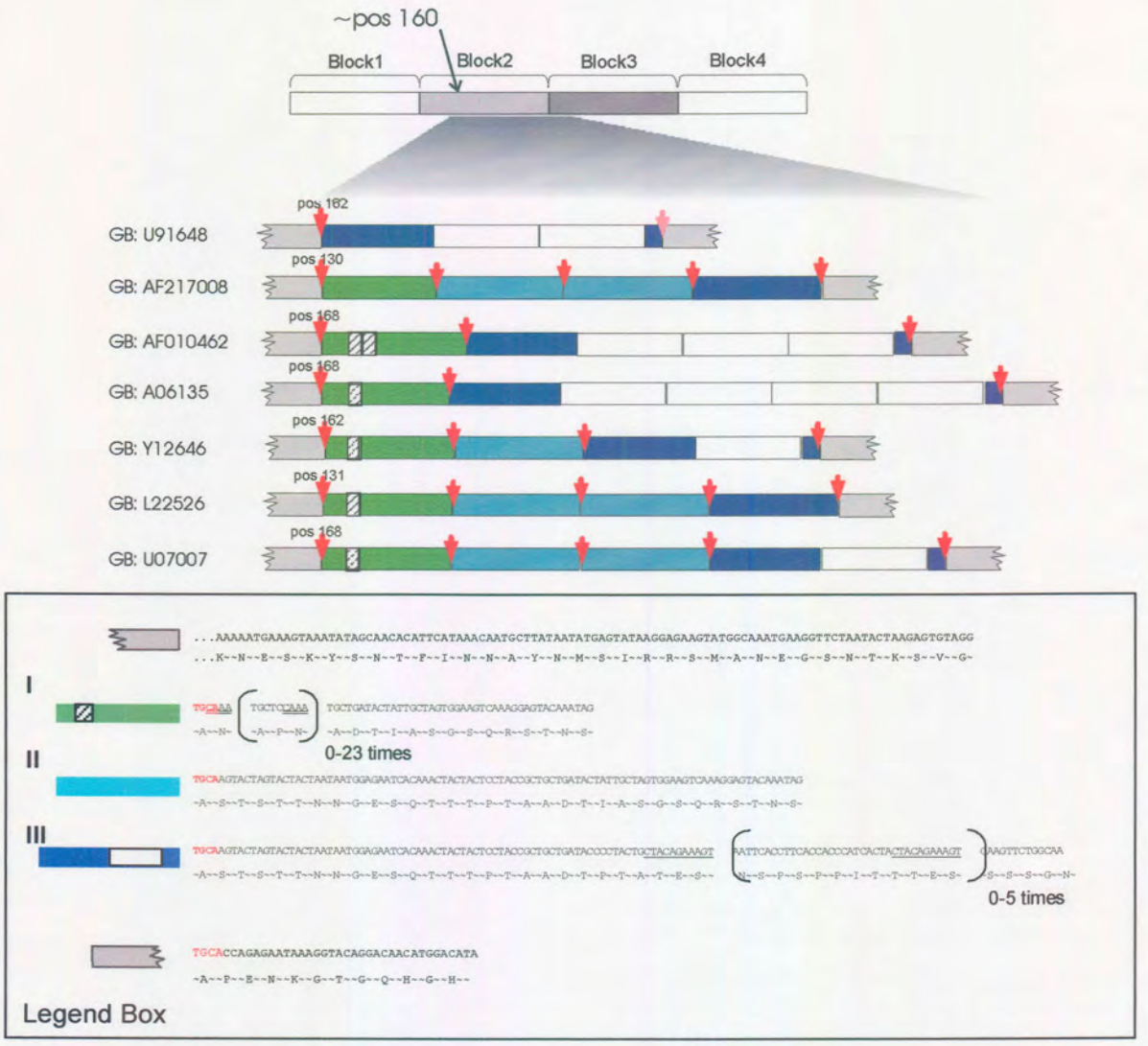

Figure 4.6: Diagrammatic representation of the allelic variation in FC27 type MSA-2 DNA, associated with TGCA repeats. The Genbank accession numbers are given to the left of the respective sequences in the diagram. Identical intervening sequences bordered by the TGCA motif are indicated in the same colour. The respective sequences are given in the legend box. Red arrows indicate the positions of the TGCA motifs. The position of the first TGCA repeat in block 2 of MSA2 is given above each sequence. A synonymous mutation where TGCA is replaced by TGCT is indicated with a pink arrow. Repeat units within the green and dark blue blocks are indicated by striped and white blocks, respectively and their sequences are given in brackets in the legend box as well as their copy numbers. The two direct repeat sequences within and bordering the tandem repeat units are underlined. The interspersed sequences (identified with roman numerals in the legend box) correspond to those in Table 4.2.

Table 4.2: Mutations in the intervening sequences bordered by TGCA repeats in the MSA2 protein. A total of 79 MSA2 sequences were analysed. The roman numerals of the different intervening sequences correspond to those in Figure 4.6. For the three intervening sequences only mutations compared to the most abundant sequence are indicated. Repetitive blocks in these regions are shaded in grey. The crossed, striped and chequered blocks above the sequences correspond to the respective blocks in Figure 4.10. Intervening sequence III is divided into subtypes a and b due to a (+/-) frameshift on an A and C, respectively.

	Nucleotide sequence	Translated amino acid sequence	Number of times found	% of all repeats
I	TGCAAATGCTGATACTATTGCTAGTGGAAGTCAAAGGAGTACAAATAG	AMADTIASGSQRSTNS	25	11.0
	·····A·····T·····	··K·····S····	11	4.8
	·····A·····G·····G·T·····T·····	··K··V·RV··S····	11	4.8
	·····T·····	·····S····	11	4.8
	·····A·····G·····T·····	··K··V·····S····	4	1.8
	·····A·····A·····T·····	··K····N···S····	2	0.9
	·····A·····G·····T·····T·····	··K··V··V··S····	2	0.9
	·····A·····	··K·····	1	0.4
	·····A·····G·····G·T····A····T·····	··K··V·RV·KS····	1	0.4
	·····A···T·····T·····	··KY·····S····	1	0.4
	·····C·····T····	·····T··Y·	1	0.4
	·····T·····G··	·····W··S·	1	0.4
	·····C·····	·····H·····	1	0.4
	·····T·····T·····	·····V··S····	1	0.4
	····G·A··A·····T·····	··EN·····S····	1	0.4
Repeat in I			124	-
			6	-
			1	-
II	TGCAAGTACTAGTACTACTAATAATGGAGAATCACAAACTACTACTCCTACCGCTGCTGATACTATTGCTAGTGGAAGTCAAAGGAGTACAAATAG	ASTSTTNNGESQTTTPTAADTIASGSQRSTNS	71	31.1
	····C·····	·T·····	1	0.4
	·····A·····	·····R·····	1	0.4
	·····A·····	·····R·····	1	0.4
	·····T·····	·····V·····	1	0.4
III	TGCAAGTACTAGTACTACTAATAATGGAGAATCACAAACTACTACTCCTACCGCTGCTGATACTATTGCTACAGAAAGTTCAAGTTCTGGCAA	ASTSTTNNGESQTTTPTAADTPTATESSSSGN	45	19.7
	·····A·····	·····K·····	10	4.4
	·····A·····	·····N·····	3	1.3
	·····A·····	·····K·····	1	0.4
	·····C·····	·····	1	0.4
	·····A·····	·····R·····	1	0.4
	·····A·····	·····N·····	1	0.4
	·····G···T·····	·····F·	1	0.4
	·····G·····	·····G·····	1	0.4
	·····T·····A·····	·····I·····	1	0.4
	TGCAAGTACTAGTACTACTAATAATGGAGAATCACAAACTACTACTCCTACCGCTGCTGATACCCCTACTGCTACAGAAAGTTCAAGTTCTGGCAA	ASTSTTNNGESQTTTPTAADTPTATESSKFWQ	9	3.9
	·····A·····	·····T·····	2	0.9
	·····A··G·····	·····	1	0.4
	·····G·····	·····R	1	0.4
	·····G·····	·····G·····	1	0.4
Repeat in III	AATTCACCTTCACCACCCATCACTACTACAGAAAGT	NSPSPPITTTES	58	-
	·····A····	·····K·	20	-
	·····G·····	··R·····	16	-
	·····T·····	I·····	12	-
	··C·····A····	·····K·	1	-
	··C·····	·····	1	-
	···A·····	·T·····	1	-
	·····C·····A····	·····K·	1	-
	·····A·····	·····	1	-
	·····G·····A····	·····A··	1	-
	·····G·····	·····	1	-
	·····T·····	·····S·····	1	-
	·····CA·····	·····T·····	1	-
	·····C·····	I·····R	1	-

The sequences in Figure 4.6 are representative of 79 MSA 2 type FC27 nucleotide sequences, deposited in Genbank. Three different sequences (I, II and III) bordered by TGCA repeats are evident. The order of these intervening sequences in relation to each other appears to be a consistent feature. The copy number of the second of these sequences (indicated in light blue in Figure 4.6) differs between different allelic forms of the protein. In some alleles, intervening sequences represented by light blue boxes were completely absent. Noteworthy is that the reading frame of the interspersed repeats were always T GCA (encoding alanine) except for a few instances noted below.

Intervening sequence III consists of two subtypes (a and b). These sequences have identical sequences up to a (+1) frameshift due to the insertion of a single adenine base by a slippage event at a position three bases after the repeat units (indicated by an asterisk in Table 4.2; TC**A**AAG instead of TCAAG). This frameshift however only affected six codons before a subsequent base-deletion slippage event, adjacent to the interspersed TGCA repeat (TGCACA instead of TGCA**C**CA) resulted in a -1 frameshift, corrected it. This +1 frameshift gave rise to the amino acid sequence TESS**KFWQCT**NKT instead of TESS**SSGNAP**NKT. This frameshift was noted in 14 of the 79 available sequence (18%). It is also the only instance in which the reading frame of TGCA was changed from an alanine to cysteine.

Several point-mutations were observed in the intervening sequences bordered by TGCA. These mutations as well as the number of times that they occur are summarised in Table 4.2. In total 45 different loci where point-mutation changes occurred, were observed. Eleven mutations at these loci were synonymous substitutions whereas the remainder led to amino acid sequence changes.

The order of the interspersed sequences in MSA2 type FC27 alleles appear to be fixed. The copy number of the intervening sequence I (indicated with green in Figure 4.6) encoding the

amino acid sequence ANADTIASGSQRSTNS varied from zero to one times. Intervening sequence II (indicated with light blue in Figure 4.6) encodes the amino acid sequence ASTSTTNNGESQTTTPTAADTIASGSQRSTNS and has a copy number of between zero and three per MSA2 gene. Intervening sequence III subtypes a and b (indicated with dark blue in Figure 4.6) encodes the amino acid sequences ASTTTNNGESQTTTPTAADTPTATESSSSGN and ASTTTNNGESQTTTPTAADTPTATESSKFWQ, respectively and occurs only once.

The sequence TGCTGATAC (marked with crossed blocks above the sequence in Table 4.2) is a second motif shared between the three sequences bordered by TGCA. This motif divides intervening sequence II into two distinct sequences. The upstream sequence is identical to the sequence upstream of this motif in intervening sequence III. The downstream sequence of intervening sequence II on the other hand is identical to the sequence downstream of this motif in intervening sequence I. It is apparent that this motif is a intra- / intergenic recombination hot-spot and that intervening sequence II is the result of a recombination event between intervening sequences I and III. The hot-spot sequence is perfectly conserved in intervening sequences II and IIIa. Two single point-mutations in the hot-spot are evident in two of fourteen intervening sequences of subtype IIIb. A point-mutation in this motif (TGCTGATAC to **A**GCTGATAC) occurred in 45% of intervening sequence I. Only single instances of the two other mutations (TGCT**G**ATAC to TGCT**T**ATAC or TGCT**A**ATAC) occurred in intervening sequence I.

In addition to the interspersed TGCA and TGCTGATAC motifs, two other interspersed repeat sequences were also identified. The first interspersed repeat sequence CAAA (double-underlined in Figure 4.6) appears to be a recombination hot-spot for the sequence TGCTCCAAA (striped block in Figure 4.6) that is repeated in tandem between zero and 23 times in intervening sequence I. This CAAA motif is essentially conserved since only one sequence with a CAGA mutation was observed. The tandemly repeated unit encoding the amino acid sequence APN (Table 4.2) occurred immediately upstream of the sequence

TGCTGATAC identified as a hot-spot for recombination between sequences bordered by TGCA. The mutation TGCT to AGCT noted above for the intra- / intergenic hot-spot are associated with low copy-numbers (1 or 2) of the tandem repeat unit TGCTCCAAA.

The other interspersed repeat sequence, CTACAGAAAGT (single-underlined in the legend box of Figure 4.6), borders another tandemly repeated sequence encoding the amino acid sequence NSPSPPITTTES (Table 4.2). The latter sequence occurred internally in interspersed sequence III and is repeated in tandem between zero to five times (indicated in white in Figure 4.6). Three point-mutations were observed in the interspersed repeat. The most abundant was the substitution of a G with an A (CTACAGAAAGT to CTACAAAAAGT), observed in 19 of the 79 sequences analysed. Two other mutations CTACAGAAAGT to CTACAGAAAGG and CTACAGAAAGT to CTACAGAAGGT were also observed, but occurred only once each. It is apparent that this interspersed repeat sequence is a recombination hot-spot for the insertion or deletion of the tandem repeat unit in interspersed sequence III.

No correlation could be demonstrated between the interspersed TGCA repeat and polymorphic repeats in block 2 of the MSA2 3D7 allelic type. Further analysis also did not reveal any other four nucleotide interspersed repeats associated in a discernible manner with polymorphism. The (A/T)GCTGATAC hot-spot described above for the FC27 allelic type is apparently also a hot-spot position for recombination between MSA2 allelic types FC27 and 3D7 (Figure 4.7). The recombination between the FC27 and 3D7 allelic types appears to be an infrequent occurrence, since only 11 such hybrid-sequences are available in Genbank. The point-mutations found in the FC27 allele sequences downstream from the recombination site (Table 4.3) correspond to those found in intervening sequence I. Multiple interspaced occurrences of the T(T/G)CTGGTGC hot-spot for intergenic recombination between 3D7 and FC27 were found in 80% of the available MSA2 type 3D7 genes. In the remaining 20% of the MSA2 type 3D7 allele genes the putative hot-spots had no more than 2 point-mutations (results not shown).

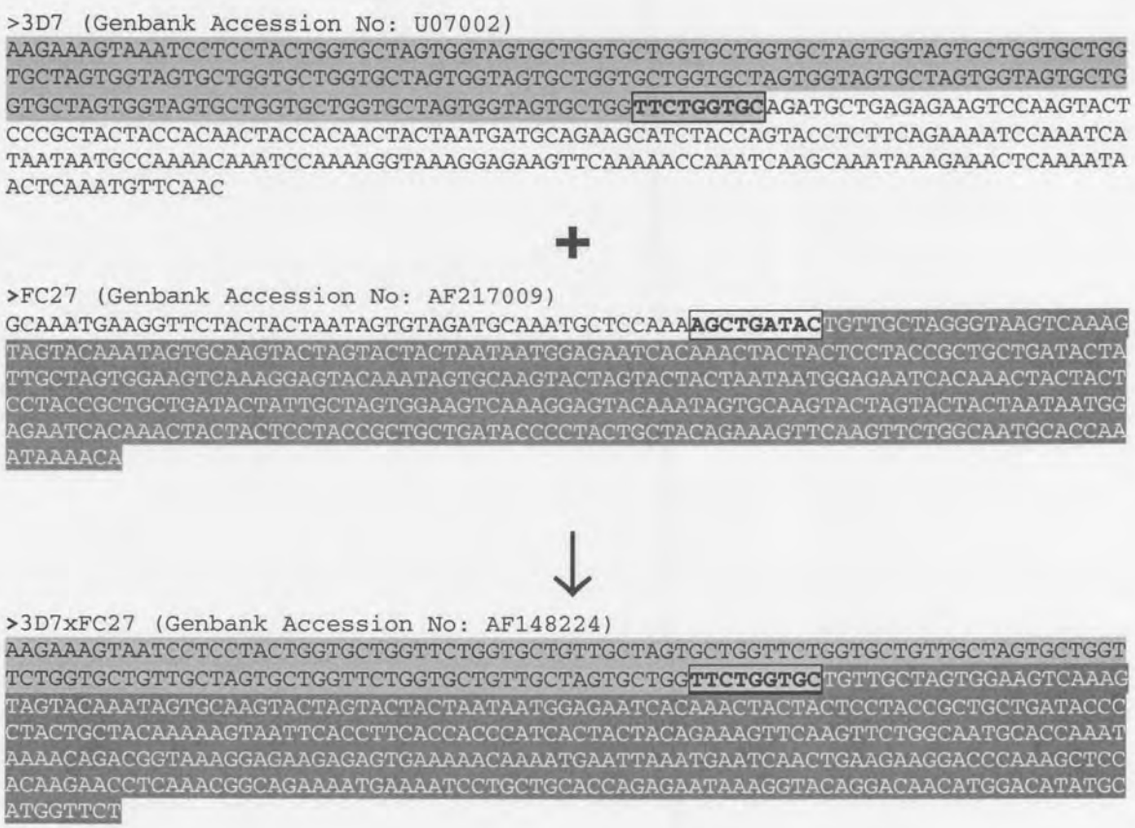

Figure 4.7: Hot-spot for recombination (boxed) observed in mixed alleles of crosses between FC27 and 3D7 allelic types of the MSA2 protein. The MSA2 3D7 allelic type is shaded in light grey and the FC27 type in dark grey. No actual intergenic recombination took place between the 3D7 (U07002) and FC27 (AF217009) sequences but they are used to illustrate the result of a recombination. The 3D7/FC27 hybrid sequence (AF148224) is an actual sequence of such an intergenic recombination.

Table 4.3: Sequences downstream of the recombination site between 3D7 and FC27. The recombination site is shaded. Point-mutations corresponding to those previously observed in intervening sequence I of FC27 is indicated in bold. Novel point-mutations are underlined.

Genbank Accession Numbers	FC27 sequences downstream of the recombination site between 3D7 and FC27
>AF217013.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>U91675.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>AF148224.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>AF148223.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>U91648.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>AF010457.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>M58414.1	TTCTGGTGCT**G**TTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>AF010460.1	TTCTGGTGCT**G**GTGCTAGTGGAAGTCAAAGGAGTACAAATAG
>AF010456.1	TGCTGGTGCT**G**GTGGTAGTGGAAGTCAAAGGAGTACAAATAG
>AF148222.1	TGCTGGTGCT**G**GTGCTAGTGGAAGTCAAAG**T**AGTACAAATAG
>U91674.1	TGCTGGTGCTAGTGGTAGTGGAAGTCAAAGGAGTACAAATAG

The FC27-3D7 recombination hot-spot (T[G/T]CTGGTGC) exhibited a degeneracy in the second position with 8 of the 11 hybrid sequences (73%) having the sequence TTCTGGTGC. The nucleotide sequence of this intergenic recombination hot-spot and that of the intragenic FC27 recombination hot-spot differed at only three loci. The FC27 hot-spot sequence (TGCTGATAC) can be converted to that of the less abundant 3D7 hot-spot (TGCTG**G**T**G**C) by only two point-mutation events, with a single other G to T point-mutation on the second position resulting in the more abundant hot-spot sequence.

4.4 Discussion

The malaria parasite has the ability to evade its host's immune system by varying the presented antigens of multi-gene copy proteins or by modifying the length, composition and sequence of repetitive regions in polymorphic proteins. Antigenic variation in single-gene copy proteins has been attributed to duplication and/or deletion of repeated segments within the genes. The length variation generated in repetitive regions was proposed to be achieved by interhelical or intrahelical recombination events or strand-slippage mechanisms (Rich *et al*, 2000a). Henaut *et al* (1998) showed an association between interspersed tetranucleotides and frameshifting events for prokaryotes. Their results revealed high occurrences of a palindromic tetranucleotide motif (AGCT), repeated at regular intervals (3n-1 base pairs). As similar mechanisms might account for the antigenic variation observed in malaria antigens, we analysed a large set of malaria sequence data (pcnuc.dat) for the presence of interspersed repeat sequences.

The most abundant palindromic repeat sequence in the malaria genome was TGCA whereas the second most abundant repeat sequence was CATG. The TGCA interspersed repeats were found to occur with higher than random frequency in 96 malaria proteins (Appendix F). The observed reading frame of the TGCA interspaced repeats in the coding regions of the MSA2 type FC27 antigen investigated here was always T GCA encoding for Alanine with the exception of a +1 and -1 frameshift event and one silent point-mutation (TGCA to TGCT; Figure 4.6 and Table 4.2). These repeats always exhibited an interspacing of 3n-1 and appear to be associated with the insertion/deletion of the intervening sequences. This indicates a level of intrinsic conservation that can not be linked to sequence alone.

In MSA2 sequences of the 3D7 allelic type, interspersed TGCA repeats did not occur at levels higher than the expected random. Furthermore, no association between the TGCA or any other tetranucleotide sequence and polymorphic repeat sequences in block 2 of the 3D7 allelic type

could be demonstrated.

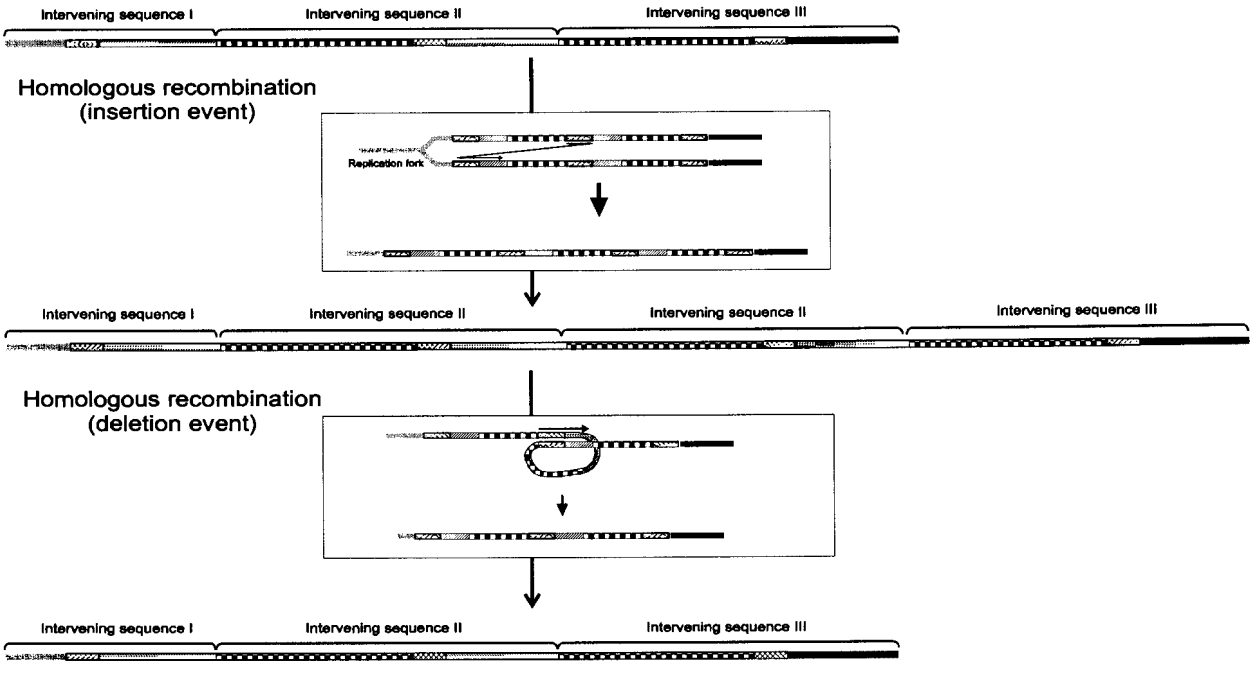

Figure 4.8: Proposed mechanism of Insertion and deletion of intervening sequences in MSA2 type FC27 by recombination events at the TGCTGATAC motif (indicated with crossed block). The crossed, striped and chequered blocks correspond to the respectively marked sequences in Table 4.2.

The second repeat motif (TGCTGATAC; indicated in bold in Table 4.2) was identified as a hot-spot for homologous recombination events between intervening sequences I and III by which intervening sequence II was created. Deletion and insertion events at this hot-spot (Figure 4.8) can also account for the variation in copy number of the first two intervening sequences (Figure 4.6; Green and light blue blocks). The copy number of the first intervening sequence (green sequence box) varied between zero and 1 while that of the second (light blue sequence box) between zero and 2. This hot-spot was conserved in 54% of intervening sequence 1. The other 46% exhibited a single T to A point mutation at the first position (TGCTGATAC to AGCTGATAC). Two of these sequences had an additional (GAT to TAT or GAT to AAT) point-

mutation. The hot-spot was however 100% conserved for all intervening sequences II and III, with the exception of one A to G mutation in intervening sequence III subtype b. A general mechanism for deletions / insertions on a hot-spot is shown in Figure 4.9. The reading frame was always preserved during recombination events at the inter-/intragenic recombination hot-spot as well as during insertion and deletion of both tandem repeat units in intervening sequences I and III.

The 32 codon repeats blocks proposed by Felger *et al*, 1997 differ in position from the 32 codon intervening sequence II (Figure 4.6). Their 32 codon repeat block stretches from codon position 6 in intervening sequence I to codon position 18 in intervening sequence II. A proposed recombination hot-spot (G C/G TAGTGGA encoding the amino acid sequence ASG) for the MSA2, type FC27 (Marshall *et al*, 1991; Irion *et al*, 1997) is only present in the first two intervening sequences bordered by TGCA, at codon positions 10 and 23, respectively of intervening sequence I and intervening sequence II. It is however absent from both subtypes of the third intervening sequence III. This hot-spot furthermore occurred in only 60 of the 79 MSA2 type FC27 sequences analysed.

Two further recombination hot-spots occurred internally in the first and third of the intervening sequences delimited by the TGCA motif. Both these hot-spots led to the insertion or deletion of tandem repeats units by means of homologous recombination. The first hot-spot CAAA (double-underlined in Figure 4.6) resulted in variable copy numbers of between zero and 23 of the repeat unit TGCTCCAAA, encoding the tripeptide APN in intervening sequence I. The propagation or deletion of this short direct repeat unit once created, by a slipped-strand mechanism, cannot be excluded. The second of these recombination hot-spots (CTACAGAAAGT; single-underlined in Figure 4.6) is responsible for the insertion / deletion of tandem repeat units that has a copy number of between zero and 5 in intervening sequence III. The 12 codon repeat sequence AATTCACCTTCACCACCCATCACTACTACAGAAAGT,

encodes the peptide sequence NSPSPPITTTES. The frameshift mutation downstream from this repeat unit appears to be due to slippage events on A (insertion) and C (deletion). It was previously described by Smythe *et al* (1991) but was thought to be due to a sequencing error. Since a total of 14 out of 79 different sequences exhibited the same +1/-1 frameshift, the possibility of sequencing errors can probably be excluded. Ninety-five percent of the tandem repeat units in intervening sequence I exhibited no point-mutations, with only 2 point-mutation loci in the remaining 5%. Only a single sequence contained a non-synonymous point-mutation. Fifty percent of the tandem repeat unit in intervening sequence III exhibited point-mutations (13 loci of which 7 are non-synonymous).

The product of a recombination event between intervening sequences I and III has not yet been observed. A total of 48 point-mutation loci were observed for the polymorphic block 2 of MSA2 type FC27 allele. Thirty of these loci (62.5%) exhibited non-synonymous mutations. This high level of non-synonymous substitution events is probably due to immune pressure, as the MSA2 protein is a highly exposed antigen on the merozoite surface. There is however still a significant number of loci (18) exhibiting silent (synonymous) substitutions.

Eleven sequences apparently generated by sexual recombination between the FC27 and 3D7 allelic types were deposited in Genbank. The site of recombination is at the FC27 recombination hot-spot TGCTGATAC. This recombination hot-spot is however replaced in the hybrids by the 3D7 sequence, T[G/T]CTGGTGC that differed in sequence from the FC27 hot-spot at three loci (TGCTGATAC to T[G/T]CTGGTGC). As the 3D7 hot-spot exhibits a degeneracy on the second position only two point-mutation events are needed to convert the FC27 hot-spot (TGCTGATAC) to that of 3D7 (TGCTGGTGC). The 3D7 sequence is always upstream in the combined 3D7/FC27 MSA2 gene. The 3D7 T(T/G)CTGGTGC recombination hot-spot sequence occurs with regular interspacing throughout the polymorphic block in 80% of the 106 available MSA2 type 3D7 sequences. In the remaining 20% of the sequences a similar

hot-spot was observed with no more than two point-mutations (Table 4.2). The recombination between the FC27 and 3D7 allelic types appears to be an infrequent occurrence due to the limited hybrid-sequences available. One explanation for the low level of recombination between the two allelic subtypes could be that the protein(s) mediating this recombination has a low binding affinity to the FC27 hot-spot. The other possible explanation could be that mutated copies of the FC27 recombination hot-spot are rare and therefore also inter-allelic recombination. The recombination apparently takes place by recombination of the 3D7 hot-spot with the FC27 recombination hot-spot in intervening sequence I. The repeat hot-spot of intervening sequence I is more degenerate than that of intervening sequence II. In intervening sequence I, 46% of the sequences exhibited one or more point-mutations compared to only 5% of the sequences for intervening sequence II.

Rich *et al* (1998) postulated that the low level of synonymous substitution events in the circumsporozoite antigen is due to an evolutionary bottleneck and argued that the population was thus largely clonal. Conway *et al* (1999) however, indicated a decline in linkage disequilibrium with physical distance in a number of polymorphic sites. This decline is expected in the case of frequently occurring meiotic recombination events in high transmission areas. Conway *et al* (1999) thus linked the level of recombination to the level of transmission. The substantial non-synonymous substitution events observed for MSA2 type FC27 allele in this study supports the latter conclusion.

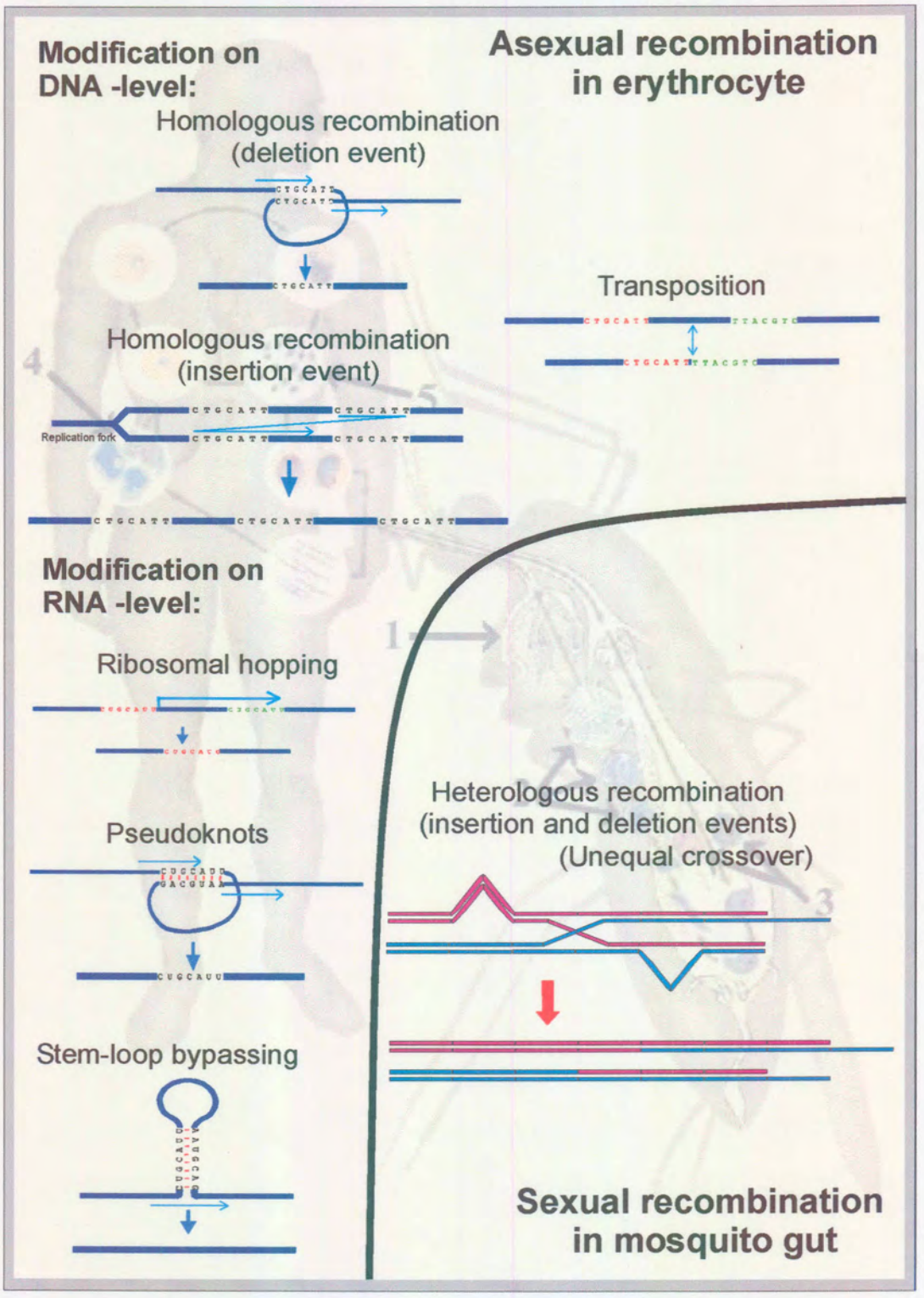

Figure 4.9: Possible mechanisms of recombination on either the asexual level in the erythrocyte, or on a sexual level in the mosquito gut (Recombination mechanisms adapted from Levin, 1997).

No definite conclusions regarding the stage at which these recombination events occur can be drawn from the results presented in this study. Recombination events at the asexual stage would be the most advantageous for the parasite due to the high level of exposure to the host's immune system. Recombination events during the sexual stage in the mosquito vector however can not be excluded. Several mechanisms of possible recombination, readthrough or bypassing events could explain the variation obtained in these antigens (Figure 4.9).

The TGCA motif is essentially completely conserved and delimited distinctive repeated intervening regions in MSA2 type FC27 alleles. The TGCA hot-spot repeats are both direct as well as reverse complementary repeats due to its palindromic nature. As such, more than one possible mechanism of recombination is possible. Variation in the copy number of the interspersing sequences can best be explained by a homologous recombination mechanism (Figure 4.9). This mechanism gives rise to both insertion and deletion of repeat blocks and has a high recombination rate (Rich *et al*, 2000). This high level of conservation and delimitation of sequences suggests an important function for this motif.

In addition to slipped-strand mispairing and unequal crossing-over events (Figure 4.9) ribosomal hopping or readthrough may also occur on interspersed palindromic tetranucleotide motifs (Henaut *et al*, 1998). The sequence of a mRNA-transcript can thus not only differ from that of the genomic DNA, but the translated protein sequence may also differ due to ribosomal hopping on the mRNA. The RNA-editing methods (Figure 4.9) will thus further allow the parasite to rapidly modify the copy number of the intervening sequences in polymorphic regions. Although the interspersed T GCA repeats (encoding alanine) appear to be a hot-spot motif it is unlikely that it is related to alanine-slippage observed on the GCG alanine codon (Henaut *et al*, 1998). Modifications at the mRNA or protein level will however not be carried over to the next generation of parasites after asexual DNA multiplication. Recombination by the parasites on a DNA-level, both in the asexual and sexual stages, will however be passed on to their progeny.

The method by which the parasite achieves variation is of profound importance in the design of an effective vaccine in order to rationally select the most important immunogenic regions as antigens for vaccinations. A number of further studies, investigating recombination events in both sexual as well as asexual stages, are currently under way. These studies will aim to pinpoint these and other recombination events to the site of action (DNA or RNA) as well as life cycle stage (sexual or asexual) at which it occurs.

CHAPTER 5

CONCLUDING DISCUSSION

The battle against malaria must be fought on at least two fronts. The first being the availability and development of prophylactic and chemotherapeutic drugs as well as effective insecticides to prevent malaria. On the second front the metabolism and mechanisms by which resistance is achieved and the immune system is evaded must be understood to effectively combat the disease in the long run. Strategies to sustain both these levels include bioinformatic methods to discover mechanisms utilised by the parasite to develop drug resistance and to evade the immune system of its host.

It was convincingly shown in the preceding chapters that bioinformatic manipulation of sequences and sequence data is an extremely powerful tool to identify the sequences of potential drug targets and to understand regulatory mechanisms of an organism at molecular level. The flood of sequence data generated by genome sequencing projects seem overwhelming, but application of the correct tools is invaluable in identifying such targets.

The exponential growth of data in sequence databanks increases the probability that a BLAST search will succeed in detecting a relationship between any new sequence and one or more known sequences in the databases. However in many cases matches are too weak for a potentially interesting relationship to be detected above the background of chance alignments. The background increases with the growth of sequence databanks, making distant relationships even more difficult to detect with confidence. A range of strategies can be followed to increase the possibility of a match between an unknown sequence and a known protein in the available databases. Detection of distant relationships is aided by the presence of multiple members of a single protein family in a database. Alternatively, a database in which relationships are explicitly

represented can be searched. An example of this latter approach is a database of protein "blocks" (e.g. the BLOCKS database at www.blocks.fhcrc.org/blocks/).

The latter approach was used to identify a malarial glucose transporter gene (Chapter 2). The sequence motif QQ(FL)(TS)GIN (conserved motif for hexose transporters) was obtained by comparison of multiple alignments of the amino acid sequences from a range of organisms. Searches of the available genomic sequence data identified a sugar transporter motif in the translated product of the malaria sequence tag mal3Z1f2.r1t. Comparison of the translation product of the sequence tag with known sequences on the specialised malaria BLAST server, identified matches with contig sequences of chromosome 2. Assembly of these contigs and further analysis thereof identified an 1515bp open reading frame, coding for a protein product matching known hexose transport proteins in the databanks. Thus an open reading frame for a putative malarial sugar transporter could be identified starting from an identified consensus motif against a high background of low-matching proteins. This method of searching the databases is more effective than searching the sequence data using a straightforward BLAST search. Using family-specific consensus sequences to search genome sequencing databases with a sequence tag increases the specificity of a match above a random background.

We are now entering an era when sequence and gene maps of organisms have become a resource rather than a goal. The increase in available sequences from the *Plasmodium falciparum* genomic sequencing project provides a wealth of information that can be utilised to identify genes of interest as well as studying their regulation and metabolism. As knowledge of genome sequences increase, the application of tools and techniques described in this study become essential in order to identify genes and the structure-function relationships of their encoded proteins. There are many different reasons to study patterns and motifs in DNA sequences - not only can the oligonucleotide signature of the genes of an organism provide clues of its evolutionary origin but also of its functional role as shown in Chapter 3.

Comparisons between signatures of genes may also assist in clarifying their functional importance. One of the major challenges of molecular biology today is the clarification of gene regulation, requiring knowledge of both the elements involved as well as the site(s) of action thereof.

Base-analysis of the sequence of chromosome 2 of *Plasmodium falciparum* showed a relative underrepresentation of A+T-rich nucleotide sequences and a relative overrepresentation of G+C-rich nucleotide sequences, compared to theoretically calculated randomly expected values (Pollack *et al*, 1991). The A+T content was 79.87%, an average between the previously reported values of 69% A+T for coding and 86% A+T for non-coding regions (Hyde *et al*, 1987; Weber, 1987).

Genomes from different organisms often differ much in their nucleotide and dinucleotide frequencies, and some of these differences is related to molecular adaptation. An example concerns the evolution of G+C content in prokaryotes. Thermophylic bacteria tend to have a high G+C content, which can be interpreted as an adaptation to protect its DNA against denaturation under high temperature because of the difference in hydrogen bonds between G/C and A/T pairs. Another interpretation is that G/C-rich codons tend to code for thermally stable amino acids and GC-poor codons tend to code for thermally unstable amino acids (Argos, 1979).

Another evolutionary hypothesis that is related to dinucleotide frequencies in the genome concerns T-T dimers produced by UV-radiation. Organisms that are exposed to sunlight are proposed to have a lower frequency of genomic TT dinucleotides to avoid the deleterious effect of T-T dimers produce by exposure to sunlight (e.g., intestinal bacteria such as *E. coli*; Singer *et al*, 1970). This might provide an explanation as to why the malaria parasite has a significantly higher AT/GC ratio than its host organism. The parasite, being intracellular, is not

exposed to the same levels of UV radiation as its host and can therefore maintain a higher level of A+T base content in its genomic makeup.

The genomic signature for *Plasmodium falciparum* was determined and GG·CC and CG was identified as the highest and lowest extremes, respectively. The high O/E value of GG•CC was shown to be positively constrained by the parasite's codon preference and amino acid abundancies. Similarly the O/E value of CG was shown to be due to negative constraints by the codon preference and amino acid abundancies. Due to the variation in genomic signature between different organisms it was also possible to analyse the origins of individual genes based on differences in their genomic signature compared to the average signature of the organism. The genomic signatures of the histidine-rich protein II (HRP-II) and merozoite surface antigen 2 (MSA-2) differed significantly from that of the average genomic signature of the malaria parasite. This led us to propose that these proteins were acquired by means of lateral transfer from an unknown source. This possibility is supported by the demonstration that the malaria parasite spontaneously acquires and expresses foreign DNA (Deitsch *et al* , 2001).

A fortuitous spin-off of this study was the detailed information on the prevalence of tri- and tetranucleotide sequence abundancies. The design of primers for use with the A+T-rich DNA and RNA of the malaria parasite intuitively suggests that their A/T base content and sequences at the 3'-end should be taken into account for polymerase-chain-reaction (PCR) amplification (Frohman, 1993). Primers with higher G+C content at the 3'-end would be more template-specific but not likely to occur in the specific region useful for primer design. The A+T/G+C ratio of the sequence at the 3'-end of primers will therefore be a trade-off between the intended application and specificity required. For differential display methods (Liang *et al*, 1997) primers are needed with a higher probability of priming to several sites in order to provide better screening resolution. In the latter case selecting primers with 3'-ends containing more frequent tri- and/or tetranucleotides would be more advantageous. A list of sequences to be used and to

avoid at the 3'-end of primers for specific amplifications, can be deduced from Table 3.3 in Chapter 3 (Trinucleotide abundancies) and Appendix H (Tetranucleotide abundance).

An association between interspersed tetranucleotides and frameshifting, ribosomal hopping and other modifying phenomena has been indicated for prokaryotes (Henaut *et. al.,* 1998). Counts of interspaced palindromic tetranucleotide pairs were determined in a large subset of malaria genes and EST's and compared with theoretically expected values. The occurrence of interspaced TGCA nucleotide pairs with (3n-1) interspacing was shown to be significantly higher than the expected values. The interspaced tetranucleotides were found to be more frequent at a specific interspacing of TGCA(N)xTGCA (x = 2, 5, 8, 11...). Eighteen functional proteins (structural/enzymes) and 31 antigens / surface proteins (e.g. MSA, CSP, PfEMP and VAR-genes) were shown to contain interspaced repeats sequences (Appendix F). These repeats were localised to the polymorphic regions of three malarial antigens.

The merozoite surface antigen 2 (MSA-2) was selected for further study. The reading frame of the TGCA interspaced repeats in the polymorphic region of the FC27 allele of MSA-2, was always in the T GCA reading frame, encoding alanine. This indicates a level of intrinsic conservation that is not linked to sequence alone. Outside the coding region no preference for any specific reading frame could be demonstrated. Analyses of 79 different allelic forms of MSA-2 type FC27 showed that the interspersed TGCA repeats demarcate three different intervening sequences with high levels of conservation. Only a single +1/-1 frameshift event was observed in intervening sequence III subtype b of the MSA-2 FC27 allelic type. The TGCA interspersed repeat was completely conserved in all of the sequences analysed except for one synonymous mutation to TGCT (sequencing error?). The +1/-1 frameshift caused a reading frame shift of the alanine encoded by T GCA to a cysteine encoded by TGC A.

Even though inter- and intragenic recombination events have been widely proposed to be the mechanism by which the parasite achieves antigenic variation (Rich *et al*, 2000; Ranford-Cartwright *et al*, 1999; Conway *et al*, 1999), no recombination hot-spots were identified. The recombination mechanisms by which the malaria parasite modulates the antigenic repertoire of the FC27 allele of MSA-2 and the hot-spots at which they likely occur have been discovered in this study.

Three recombination hot-spots have been also identified in the MSA-2 type FC27 allele. The first two are located internally in intervening sequences I and III. These hot-spots were shown to be associated with variable copy numbers of tandem repeat units by means of homologous recombination (Figure 4.6). The third recombination hot-spot consists of a 9 nucleotide sequence (A/T)GCTGATAC (Table 4.2) that mediated homologous recombination events between intervening sequences I and III and was shown to be associated with insertions or deletions of intervening sequences I and II. This hot-spot was also shown to be located at he intergenic recombination site between the two MSA-2 allelic types, giving rise to 3D7-FC27 hybrids (Figure 4.7). The sequence of the recombination site (T[G/T]CTGGTGC; Table 4.3) is different from that of the FC27 allele identified in this study (TGCTGATAC; Table 4.2). However, when a degenerate sequence of the recombination hot-spot is deduced from point-mutations observed in that of intervening sequence I, only three differences were apparent (T[G/**T**]CTG**G**T**G**C). The 3D7-FC27 hybrids appear to be rare events since only 11 hybrid sequences out of a combined total of 196 MSA-2 allelic sequences were deposited in Genbank). One likely explanation for this low percentage of hybrids is that the 3D7 and FC27 hot-spots are not exact matches, thus reducing the likelihood of recombination events. The other possibility is that further point-mutations in the recombination hot-spot of MSA2 type FC27 to produce an exact copy of the recombination hot-spot in the 3D7 allele, are rare occurrences. The mutual hot-spot between FC27 and 3D7 identified by Irion *et al* (1997; G[C/G]TAGTGGA

encoding the amino acid sequence ASG) does not correspond to the recombination hot-spot identified in this study.

No intervening sequences were found consisting of recombinants between the sequences upstream and downstream of the repeat hot-spot in intervening sequence I and III, respectively. Such a recombination event would result in a "suicide recombination" since the downstream sequence of intervening sequence I, the whole of intervening sequence II as well as the upstream sequence of intervening sequence III would be deleted. Regeneration of these sequences would not be possible any more thus reducing the ability of the parasite to sustain the polymorphic traits of block 2.

A remarkable feature of the insertion/deletion events in the intervening sequences and even the +1/-1 frameshift event in intervening sequence IIIb is that the reading frame of the protein downstream from the last interspersed repeat is never altered. This observation suggests a very important but yet undefined function for MSA-2 that is protected by the polymorphic nature of block 2.

Although the 3D7-FC27 hybrid sequences provide evidence for intergenic recombination on a sexual level no definite conclusions can be made regarding the stage(s) - sexual and/or asexual - at which the recombinations of the FC27 allele take place. It is highly likely that recombination takes place at least on the asexual level as this is where the parasite is most exposed to the immune system of its host. The rate of recombination at these hot-spots are orders of magnitude higher than point-mutations enabling the parasite to rapidly change its polymorphic profile (Rich *et al*, 2000).

Repeats have been shown to have at least three important functional roles (Levin, 1997). Firstly, some may evolve to become regulatory regions of genes expressed in a tissue-specific

manner. Secondly, repeats play an important role in refashioning the genomic architecture by facilitating homologous recombination, translocations and gene conversions. Lastly, repeats have also been implicated in epigenetic phenomena such as parental imprinting and position-effect variegation, serving as valuable time-markings for unravelling the complexities of molecular archaeology. The exact role of the TGCA motifs is not yet clear, but its importance is evident from its high level of conservation. The possibility that RNA readthrough or ribosomal slipping takes place on these motifs, leading to different protein sequences from the same mRNA fragment, can not be excluded and also need to be addressed in further studies and by comparison of the DNA, mRNA and protein sequences. The proteins associated with these recombination events also need to be identified and characterised.

The antigenic repertoire of the MSA-2 type FC27 polymorphic region has been reduced to three intervening sequences (I, II and III). As these sequences are the building blocks for the reported variable sequences of the MSA-2 type FC27 gene, they are ideal vaccine candidates. Other possible applications of the results of this study is its use in microsatellite typing (Anderson *et al*, 2000) of field isolates to analyse the effects of vaccines on the polymorphism of the MSA-2 protein and also provide insights into population genetics and rates of recombination in high- compared to low transmission areas.

REFERENCES

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

Amador, R. and Patarroyo, M.E. (1996). Malaria vaccines. *J. Clin. Immunol.* **16**, 183-189.

Anderson, T.J.C., Haubold, B., Williams, J.T., Estrada-Franco, J.G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I.D., Brockman, A.H., Nosten, F., Ferreira, M.U., Day, K.P. (2000). Microsattelite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum* *Mol. Biol. Evol.* **17**(10), 1467-1482.

Argos, P., Rossman, M.G., Grau, U.M., Zuber, H., Frank, G., Tratschin, J.D. (1979). Thermal stability and protein structure. *Biochemistry* **18**(25), 5698-5703.

Ash, C. (1991). First Impressions of the Malaria Vaccine. *Parasitol. Today* **7**(4), 63-64.

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19** Supplement, 2241-2245.

Bairoch, A. (1993) The ENZYME data bank. *Nucleic Acids Res.* **21**(13), 3155-3156.

Bairoch, A. Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19** Supplement, 2247-2249.

Bairoch, A., Bucher, P. and Hofman, K. (1995). The PROSITE database, its status in 1995. *Nucleic Acids Res.* **24**, 189-196.

Barker, W.C., George, D.G., Hunt, L.T., Garavelli, J.S. (1991). The PIR protein sequence database. *Nucleic Acids Res.* **19** Supplement, 2231-2236.

Barton, G.J. (1994). SCOP: Structural classification of proteins. *Trends in Biochemical Sciences* **19**, 554-555.

Basco, L.K., Le Bras, J., Rhoades, Z., Wilson, C.M. (1995a). Analysis of *pfmdr1* and drug susceptibility in fresh isolates of *Plasmodium falciparum* from sub-Saharan Africa. *Mol. Biochem. Parasitol.* **74**, 157-166.

Basco, L.K., de Pecoulas, P.E., Wilson, C.M., Le Bras, J., Mazabraud, A. (1995b). Point mutations in the dihydrofolate reductase-thymidylate synthase gene and pyrimethamine and cycloguanil resistance in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **69**, 135-138.

Benet, L.Z., Sheiner, L.B. (1985). Chapter 1: Pharmacokinetics: The dynamics of drug absorption, distribution and elimination. in Goodman and Gilman's: The Pharmacological Basis of Therapeutics 7thedition. Ed's. Gilman A.G., Goodman L.S., Rall T.W., Murad F., Macmillan Publishing Company, NY.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**(2), 319-24.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**(7),1499-1504.

Bjellqvist, B.,Hughes, G.J., Pasquali, C.H., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S. Hochstrasser, D.F. (1993). The focusing positions of polypeptides in immobilised pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023-1031.

Bleasby, A.J., Akrigg, D. Attwood, T.K. (1994). OWL - A non-redundant, composite protein sequence database. *Nucleic Acids Res.* **22** (17), 3574-3577.

Borst, P., Bitter, W., McCulloch, R., Van Leeuwen, F., Rudenko, G. (1995a). Antigenic variation in malaria. *Cell* **82**(1), 1-4.

Borst, P. and Ouellette, M. (1995b). New mechanisms of drug resistance in parasitic protozoa. *Annu. Rev. Micro.* **49**, 427-460.

Bowman, S. L., Basham, D., Brown, D., Chillingworth, T., Churcher, C.M., Craig, A., Davies, R.M., Devlin, K., Feltwell, T., Gentles, S., Gwilliam, R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P., Jagels, K., Jassal, B., Kyes, S., McLean, J., Moule, S., Mungall, K., Murphy, L., Olivier, K., Quail, M.A., Rajandream, M-A., Rutter, S., Skelton, J., Squares, S., Sulston, J.E., Whitehead, S., Woodward, J.R., Newbold, C., Barrell, B.G. (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532-538.

Brannan, L.R., Turuner, C.M.R., Phillips, R.S. (1994). Malaria parasites undergo antigenic variation at high rates *in vivo*. *Proc. R. Soc. Lond. B.Biol. Sci.* **256**, 71-75.

Brown, G.V., Nossal, G.J.V. (1986). Malaria – Yesterday, Today and Tomorrow. *Persp. Biol. Med.* **30**(1), 65-75.

Brown, P. (1992). Who cares about malaria? *New Scientist* (31 October 1992), 37-41.

Bruce-Chwatt, L.J. (1985). *Essential malariology*. 2nd edition. Heinemann Medical ELBS, London.

Burge, C., Campbell, A.M., Karlin, S. (1992). Over- and underrepresentation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358-1362.

Burks, C., Cassidy, M., Cinkosky, M.J., Cumella, K.E., Gilna, P., Hayden, J.E.D., Keen, G.M., Kelley, T.A., Kelly, M., Kristofferson, D., Ryals, J. (1991). GenBank. *Nucleic Acids Res.* **19** Supplement, 2221-2225.

Campbell, A, Mrazek, J, Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* **96**(16), 9184-9189.

Collins, F.H. and Paskewitz, S.M. (1995). Malaria: Current and Future Prospects of Control. *Annu. Rev. Entomol.* **40**, 195-219.

Contamin, H., Fandeur, T., Bonnefoy, S., Skouri, F., Ntoumi, F., Mercereau-Puijalon, O. (1995). PCR typing of field isolates of *Plasmodium falciparum*. *J. Clin. Microbiol.* **33**(4), 944-951.

Conway, D.J., Roper, C., Oduola, A.M.J., Arnot, D.E., Kremsner, P.G., Grobush, M.P., Curtis, C.F., Greenwood, B.M. (1999). High recombination rate in natural populations of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4506-4511.

Cooke, B.M. (2000). Molecular approaches to malaria: seeking the whole picture. *Parasitol. Today* **16**(10), 407-408.

Corpet, F., Gouzy, J., Kahn, D. The ProDom database of protein domain families. (1998). *Nucleic Acids Res.* **26**(1), 323-326.

Cowman, A.F. (1991). The P-glycoprotein Homologues of *Plasmodium falciparum*: Are They Involved in Chloroquine Resistance? *Parasitol. Today* **7**(4), 70-75.

Da Silveira, L.A., Dorta, M.L., Kimura, E.A., Katzin, A.M., Kawamoto, F., Tanabe, K., Ferreira, M.U. (1999). Allelic diversity and antibody recognition of *Plasmodium falciparum* merozoite surface protein 1 during hypoendemic malaria transmission in the Brazilian amazon region. *Inf. Immunity.* **67**(11), 5906-5916.

Dame, J.B., Arnot, D.E., Bourke, P.F., Chakrabarti, D., Christodoulou, Z., Coppel, R.L., Cowman, A.F., Craig, A.G., Fisher, K., Foster, J., Goodman, N., Hinterberg, K., Holder, A.A., Holt, D.C., Kemp, D.J., Lanzer, M., Lim, A., Newbold, C.I., Ravetch, J.V., Reddy, G.R., Rubio, J., Schuster, S.M., Su, X., Thompson, J.K., Vital, F., Wellems, T.E., Werner, E.B. (1996). Mini-review: Current status of the *Plasmodium falciparum* genome project. *Mol. Biochem. Parasitol.* **79**, 1-12.

Dame, J.B., Williams, J.L., McCutchan, T.F., Weber, J.L., Wirtz, R.A., Hockmeyer, W.T., Maloy, W.L., Haynes, J.D., Schneider, I., Roberts, D. (1984). Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. *Science* **225**, 593-599.

Deitsch, K.W., Driskill, C.L., Wellems, T.E. (2001). Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. *Nucleic Acids Res.* **29**(3), 850-853.

Deitsch, KW, del Pinal, A, Wellems, TE (1999). Intra-cluster recombination and var transcription switches in the antigenic variation of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **101**(1-2), 107-116.

Desai, S.A., Krogstad, D.J., McCleskey, E.W. (1993). A nutrient-permeable channel on the intraerythrocytic malaria parasite. *Nature* **362**, 643-646.

Desowitz, R.S. (1991). In: *The Malaria Capers* (More tales of people, research and reality). New York, W.W. Norton & Company.

Djimde, A., Doumbo, O.K., Cortese, J.F., Kayentao, K., Doumbo, S., Diourte, Y., Dicko A., Su, X., Nomura ,T., Fidock, D.A., Wellems, T.E., Plowe, C.V., Coulibaly, D. (2001). A Molecular Marker for Chloroquine-Resistant *Falciparum* Malaria. *New Engl. J. Med.* **344**(4), 257-263.

Felger, I., Marshal, V.M., Reeder ,J.C., Hunt, J.A., Mgone, C.S., Beck, H.P. (1997). Sequence diversity and molecular evolution of the merozoite surface antigen 2 of *Plasmodium falciparum*. *J. Mol. Evol.* **45**(2), 154-160.

Felsenstein, J. (1989). PHYLIP- Phylogeny inference package, Version 3.2. *Cladistics* **5**, 164-166.

Frohman, M.A. (1993). Rapid amplification of complementary DNA ends for generation of full-length complementary DNAs: Thermal RACE. *Methods in Enzymol.* **218**, 340-356.

Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M., Aravind, L., Koonin, E.V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai Z., Schwartz, D.C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G.G., Clayton, R., White, O., Smith, H.O., Fraser, C.M., Adams, M.D., Venter, J.C., Hoffmann, S.L. (1998). Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum*. *Science* **282**, 1126-1132.

Geourjon, C., Deleage, G. and Roux, B. (1991). ANTHEPROT: an interactive graphics software for analysing protein structures from sequences. *J. Mol. Graph.* **9**, 188-190.

Gero, A.M., Upston, J.M. (1992). Altered Membrane Permeability: a New Approach to Malaria Chemotherapy. *Parasitol. Today* **8**, 283-286.

Goman, M., Langsley, G., Hyde, J.E., Yankovsky N.K., Zolg J.W., Scaife J.G. (1982). The establishment of genomic DNA libraries for the human malaria parasite *Plasmodium falciparum* and identification of individual clones by hybridisation. *Mol. Biochem. Parasitol.* **5**(6), 391-400.

Good, M.F., Kaslow, D.C., Miller, L.H. (1998). Pathways and strategies for developing a malaria blood-stage vaccine. *Annu. Rev. Immunol.* **16**, 57-87.

Greenberg, D.S. (1999). National Institutes of Health moves ahead with "PubMed Central" *Lancet* **354**, 1009.

Greenwood, B. (1999). Malaria martality and morbidity in Africa. *Bull. W.H.O.* **77**(8), 617-618.

Hénaut, A., Lisacek, F., Nitschké, P., Moszer, I., Danchin, A. (1998). Global analysis of genomic texts: The distribution of AGCT tetranucleotides in the *Escherichia coli* and *Bacillus subtilis* genomes predicts translational frameshifting and ribosomal hopping in several genes. *Electrophoresis* **19**(4), 515-527.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., Hood, L. (1997). Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**, 609-614.

Henikoff, S. and Henikoff, J.G. (1994). Protein family classification based on searching a database of blocks. *Genomics* **19**, 97-107.

Homewood, C.A., Neame, K.D. (1974). Malaria and the permeability of the host erythrocyte. *Nature* **252**, 718-719.

Hong, Y-L., Yang, Y-Z., Meshnick, S.R. (1994). The interaction of artemisinin with malarial hemozoin. *Mol. Biochem. Parasitol.* **63**, 121-128.

Hyde, J.E., Sims, P.F. (1987). Anomalous dinucleotide frequencies in both coding and non-coding regions from the genome of the human malaria parasite *Plasmodium falciparum*. *Gene* **61**(2), 177-87.

Irion, A., Beck, H., Felger, I. (1997). New repeat unit and hot spot of recombination in FC27-type alleles of the gene coding for *Plasmodium falciparum* merozoite surface protein 2. *Mol. Biochem. Parasitol.* **90**, 367-370.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**(1), 29-34.

Karlin, S., Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends in Gen.* **11**(7), 283-90.

Karlin S., Burge C., Campbell A.M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **20**(6), 1363-1370.

Kemp, D.J., Coppel, R.L., Anders, R.F. (1987). Repetitive proteins and genes of Malaria. *Annu. Rev. Microbiol.* **41**, 181-208.

Kirk, K., Tilley, L., Ginsburg, H. (1999). Transport and Trafficking in the Malaria-infected Erythrocyte. *Parasitol. Today* **15**(9), 355-357.

Krogstad, D.J., Gluzman, I.Y., Kyle, D.E., Oduola, A.M.J., Martin, S.K., Milhous, W.K., Schlesinger, P.H. (1987). Efflux of chloroquine from *Plasmodium falciparum*: Mechanism of chloroquine resistance. *Science* **238**, 1283-1285.

Krogstad, D.J., Schlesinger, P.H., Gluzman, I.Y. (1985). Antimalarials Increase Vesicle pH in *Plasmodium falciparum*. *J. Cell Biol.* **101**, 2302-2309.

Lewin, B. (1997). *Genes* 6th edition. Oxford University Press, Oxford.

Liang, P., Pardee, A.B. (1997). Differential display. A general protocol. *Meth. Mol. Biol.* **85**, 3-11.

Marshall, V.M., Coppel, R.L., Martin, R.K., Oduola, A.M.J., Anders, R.F., Kemp, D.J. (1991). A *Plasmodium falciparum* MSA-2 gene apparently generated by intragenic recombination between two allelic families. *Mol. Biochem. Parasitol.* **45**, 349-352.

McEntyre, J. (1999). Linking up with Entrez. *Trends in Gen.* **14**(1), 39-40.

Medical Research Council (1996). Map of Malaria in South-Africa. Internet on: http://www.travelclinic.co.za/html/malaria_map.html (1 January 2001).

Miller, L.H., Roberts T., Shahabuddin M., McCutchan T.F. (1993). Analysis of sequence diversity in the *Plasmodium falciparum* surface protein-1 (MSP-1). *Mol. Biochem. Parasitol.* **59**, 1-14.

Mitas, M. (1997). Trinucleotide repeats associated with human disease. *Nucleic Acids Res.* **25**(12), 2245-2253.

Mitchell, G.H. (1989). An update on candidate malaria vaccines. *Parasitology* **98** Supplement, 29-47.

Mons, B., Trape, J-F., Hagan, P. (1997). Malaria, Malaria and More Malaria. *Parasitol. Today* **13**(80), 279-284.

Muentener, P., Schlagenhauf, P., Steffen, R. (1999). Imported malaria (1985-95): trends and perspectives. *Bull. W.H.O.* **77**(7), 560-563.

Musto, H., Caccio, S., Rodriguez-Maseda, H., Bernardi, G. (1997). Compositional constraints in the extremely GC-poor genome of Plasmodium falciparum. *Mem. Inst. Oswaldo Cruz* **92**(6), 835-841.

Nel K. (1998). PhD progress report - 1998. Department of Biochemistry, University of Pretoria.

Newbold, C.I. (1999). Antigenic variation in *Plasmodium falciparum*: mechanisms and consequences. *Curr. Opinion Microbiol.* **2**(4), 420-425.

Nussinov, R. (1984). Strong doublet preferences in nucleotide sequences and DNA geometry. *J. Mol. Evol.* **20**(2), 111-119.

Ochman, H., Lawrence, J.G., Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299-304.

Olliaro, P.L., Goldberg, D.E. (1995). The *Plasmodium* Digestive Vacuole: Metabolic Headquarters and Choice Drug Target. *Parasitol. Today* **11**(8), 294-297.

Orengo, C.A., Jones, D.T., Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.

Peterson, D.S., Milhous, W.K., Wellems, T.E. (1990). Molecular basis of differential resistance to cycloguanil and pyrimethamine in *Plasmodium falciparum* malaria. *Proc. Natl. Acad. Sci. U.S.A.* **87**(8), 3018-22.

Pollack, Y., Kogan, N., Golenser, J. (1991). *Plasmodium falciparum*: evidence for a DNA methylation pattern. *Exp. Parasitol.* **72**(4), 339-44.

Ranford-Cartwright, L.C., Walliker, D. (1999). Intragenic recombinants of *Plasmodium falciparum* identified by in situ polymerase chain reaction. *Mol. Biochem. Parasitol.* **102**, 13–20.

Ranford-Cartwright, L.C., Balfe, P., Carter, R., Walliker, D. (1993). Frequency of cross-fertilization in the human malaria parasite *Plasmodium falciparum*. *Parasitology* **107**(1), 11-18.

Rich, S.M., Ayala, F.J. (1998). The recent origin of allelic variation in antigenic determinants of *Plasmodium falciparum*. *Genetics* **150**(1), 515-517.

Rich, S.M., Ayala, F.J. (2000). Population structure and recent evolution of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.* **97**(13), 6994-7001.

Rich, S.M., Ferreira, M.U., Ayala, F.J. (2000). The origin of antigenic diversity in *Plasmodium falciparum*. *Parasitol. Today* **16**(9), 390-396.

Saier, M.H. Jr. (1994). Computer-Aided Analyses of Transport Protein Sequences: Gleaning Evidence concerning Function, Structure, Biogenesis and Evolution. *Microbiol. Rev.* **58**(1), 71-93.

Santibanez-Koref, M., Reich, J.G. (1986). Dinucleotide frequencies in different reading frame positions of coding mammalian DNA sequences. *Biomed. et Biochim. Acta* **45**(6), 737-48.

Saul, A. (1999) The Role of Variant Surface Antigens on Malaria-infected Red Blood Cells. *Parasitol. Today* **15**(11), 455-457.

Schneider, T.D., Stephens, R.M., (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* **18**, 6097-6100.

Sherman, I.W. (1979). Biochemistry of *Plasmodium* (Malarial Parasites). *Microbiol.Rev.* **43**(4), 453-495.

Sherman, I.W., Tanigoshi, L. (1974). Glucose transport in the malarial (*Plasmodium lophurae*) infected erythrocyte. *J. Protozool.* **21**, 603-607.

Singer, C.E., Ames, B.N. (1970). Sunlight ultraviolet and bacterial DNA base ratios. *Science* **170**, 822-825.

Smythe, J.A., Coppel, R.L., Day, K.P., Martin, R.K., Oduola, A.M., Kemp, D.J., Anders, R.F. (1991). Structural diversity in the Plasmodium falciparum merozoite surface antigen 2. *Proc. Natl. Acad. Sci. U.S.A.* **88**(5), 1751-1755.

Staden, R. (1990). Finding protein coding Regions in Genomic Sequences. *Methods in Enzymol.* **183**,163-180.

Stein, C.A. (1996). Exploiting the potential of antisense: beyond phosphorothioate oligodeoxynucleotides. *Chem. and Biol.* **3**, 319-323.

Stoehr, P.J., Cameron, G.N. (1991). The EMBL data library. *Nucleic Acids Res.* **19** Supplement, 2227-2230.

Sved, J., Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U.S.A.* **87**(12), 4692-4696.

Tanabe, K., Mackay, M., Scaife, J.G. (1987). Allelic dimorphism in a surface antigen of the malaria parasite *Plasmodium falciparum. J. Mol. Biol.* **195**, 273-287.

Tatusov, R.L., Koonin, E.V., Lipman, D.J. (1997). A genomic perspective on protein families. *Science* **278**, 631-637.

Thompson, J.D. Higgins, D.G. Gibson, T.J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Verra, F., Hughes, A.L. (1999). Biased amino acid composition in repeat regions of Plasmodium antigens. *Mol. Biol. Evol.* **16**(5), 627-633.

Weber, J.L. (1987). Analysis of sequences from the extremely A + T-rich genome of Plasmodium falciparum. *Gene* **52**(1), 103-109.

Wellems, T.E. (1992). How chloroquine works. *Nature* **355**, 108-109.

Wellems, T.E., Walker-Jonah, A., Panton, L.J. (1991). Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. *Proc. Natl. Acad. Sci. U.S.A.* **88**(8), 3382-3386.

Wellems, T.E., Panton, I.J., Gluzman, I.Y., do Rosario, R.V., Gwadz, R.W., Walker, J.A., Krogstad, D.J. (1990). Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. *Nature* **345**, 253-255.

Woodrow, C.J., Burchmore, R.J., Krishna, S. (2000). Hexose permeation pathways in Plasmodium falciparum-infected erythrocytes. *Proc. Natl. Acad. Sci. U.S.A.* **97**(18), 9931-9936.

World Health Organisation (1997). World malaria situation in 1994. *WER* **36**, 269-274.

Wren, B.W. (2000). Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nature Rev. Gen.* **1**, 30-39.

SUMMARY

In Silico analysis of available biological data is a powerful tool for not only the identification of new genes, but also to study evolutionary relationships and regulatory mechanisms. In this study, a number of bioinformatic tools and techniques were applied on the available sequence data of the malaria parasite, *Plasmodium falciparum*. *In Silico* techniques were used for the identification of a genomic sequence tag (GST) matching the facilitated glucose transporter family as assessed by BLAST. The open reading frame encoding the full-length glucose transporter gene was subsequently assembled from contig sequences of chromosome 2 of the malaria parasite.

The frequency of occurrence of di-, tri- and tetranucleotide sequences in both the coding and non-coding regions of chromosome 2 of *P. falciparum* was also exhaustively analysed. The relative abundance (observed, compared to expected values) of these oligonucleotide sequences, normalised for the nucleotide base composition, was calculated as an odds ratio and compared to those of other organisms. These relative abundancies are referred to as the organism's genomic signature. The CC·GG and CG-dinucleotides exhibited the highest and the lowest odds ratios, respectively. These genome signatures were shown to be constrained by the codon preference and amino acid abundancies. A number of genes with genomic signatures differing significantly from the average signature were also identified and were deduced to be acquired by lateral transfer from unidentified sources.

A definite association between interspaced TGCA tetranucleotides and polymorphic traits of the FC27 allele of merozoite surface antigen 2 (MSA-2) was shown. The observed switching and deletion of a limited number of identical nucleotide sequences of several alleles interspersed between direct repeats, provided clues to potential mechanisms employed by the parasite to affect antigenic polymorphism. The identification of a number of motifs for intragenic (homologous) recombination led us to propose a mechanism by which the parasite achieves

antigenic variation in single copy genes. These results have profound implications for the design of candidate anti-malarial vaccines, microsatellite typing and characterisation of proteins mediating these recombination events.

OPSOMMING

In Silico analise van die beskikbare biologiese data is 'n kragtige middel - nie net vir die identifikasie van nuwe gene nie, maar ook om evolusionêre verwantskappe en regulatoriese meganismes te bestudeer. In hierdie studie is 'n aantal bioinformatiese hulpmiddels en tegnieke toegepas op die beskikbare volgorde data van die malaria parasiet, *Plasmodium falciparum*. *In Silico* tegnieke is gebruik vir die identifikasie van 'n genoom-volgorde teiken (genomic sequence tag; GST) wat lid is van die gefasiliteerde glukose transporter familie soos aangedui deur BLAST. Die oop leesraam wat kodeer vir die vollengte glukose transporter geen is vervolgens saamgestel vanuit "contig"-volgordes van chromosoom 2 van die malaria parasiet.

Die frekwensie van voorkoms van di-, tri- en tetranukleotied-volgordes in beide die koderende en nie-koderende areas van chromosoom 2 van *P. falciparum* is ook bepaal. Die relatiewe voorkoms (waargenome in vergelyking met verwagte waardes) van hierdie oligo-nukleotied volgordes, genormaliseer vir die nukleotied-basis samestelling, is bereken as 'n oneweredigheids-verhouding ("odds ratio") en vergelyk met díe van ander organismes. Daar word na hierdie relatiewe voorkomste verwys as die genomiese handtekening ("genomic signature"). Die CC·GG en CG-dinukleotiede het onderskeidelik die hoogste en laagste oneweredigheids-verhouding vertoon. Beperkings deur kodon-voorkeur en aminosuur voorkoms op die genomiese handtekeninge kon aangetoon word. 'n Aantal gene waarvan die genomiese handtekening van die gemiddelde handtekening verskil is geïdentifiseer. Daar kon afgelei word dat hierdie gene waarskynlik deur laterale oordrag vanaf ongeïdentifiseerde bronne verkry was.

'n Duidelike verband kon tussen gespasieërde TGCA tetranukleotiede en polimorfiese eienskappe van die FC27 alleel van merosoïet oppervlak-antigeen 2 (MSA-2) getref word. Die waargenome antigeniese skommelinge en delesies van 'n beperkte aantal identiese nukleotied-volgordes van verskeie allele, gespasieërd tussen direkte herhalings, het leidrade verskaf ten

opsigte van die potensiële meganismes wat die parasiet gebruik om antigeniese variasie te bewerkstellig. Die identifikasie van 'n aantal motiewe vir intra-geniese (homoloë) rekombinasie het gelei tot die postulering van 'n meganisme waardeur antigeniese variasie in spesifieke gene deur die parasiet bewerkstellig kan word. Hierdie resultate het belangrike implikasies vir die ontwerp van anti-malaria entstof-kandidate, mikrosatteliet-tiperings en karakterisering van die proteïene betrokke by hierdie geen-rekombinasies.