Plasmodium falciparum dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS): Gene synthesis and recombinant expression

by

Linda Coetzee

Submitted in partial fulfilment of the requirements for the degree *Magister Scientiae*
in the Faculty of Natural and Agricultural Sciences
Department of Biochemistry
University of Pretoria
Pretoria
December 2003

Table of Contents

Chapter 3

Expression of synthetic *P. falciparum* dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS)

Chapter 4

In silico analysis of dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS)

Chapter 5

Concluding discussion

Acknowledgements

To Professor Louw, my promoter and mentor, who made this project as much his own as it was mine, shared in all the ups and downs, from whom I learnt all about science and much about life in general, who believed in my abilities and gave me opportunities far beyond that and to Prof. Neitz, my co-promoter, for his sharp mind and valuable comments. To the NRF for rewarding me a scholarship that enabled these studies.

To Professor Hyde, Dr. Tanya Aspinall and the rest of the UMIST research team, for receiving and hosting me with such warmth, for all their good advice and assistance and their willingness to cooperate and work with me towards our common goal.

To all our staff, lecturers and students at UP, especially Jaco de Ridder for the design of the oligonucleotides, who gave me a new understanding and appreciation of computers, Gordon Wells who generously donated time from his own thesis to help me in the Bioinformatics department, as well as Sandra, Ben and Christine for helping me with the technical and experimental detail and Lyn-Marie for her much valued advice and insights.

To my parents, who realised the importance of a good education and who set some of their dreams aside to provide me with this, for their unconditional love and support. To my grandparents, family and friends for their continual support, interest and understanding. To Heinrich for his undaunted support and belief in my abilities, who shared with me all my failures and successes and loved me nevertheless.

Lastly, to our Father, Who loved us enough to become one of us, Who created this wonderful universe we have yet to begin to understand, Who knows all the answers to our myriad of questions, Who grants us brief glimpses of His perfection through the subjects we study and explore…

Abbreviations

ACP	acyl carrier protein
ARMS PCR	amplification refractory mutation system PCR
ATP	adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	basepair
BSA	bovine serum albumin
CPG	controlled pore glass
CS	circumsporozoite
DHFR	dihydrofolate reductase
DHFS	dihydrofolate synthase
DHPS	dihydropteroate synthase
DMT	4,4'dimethoxytrityl
DNA	deoxyribonucleic acid
DOXP	deoxyxylulose phosphate
dNTP	deoxynucleoside triphosphate
dTMP	deoxythymidine monophosphate
DTT	dithiotreitol
dUMP	deoxyuridine monophosphate
E.C.	Enzyme Commission
EDTA	(ethylenedinitrilo)tetraacetic acid
FPGS	folylpolyglutamate synthase
GTP	guanosine triphosphate
GTP-CH	GTP cyclohydrolase
HLA	human leukocyte antigen
HNN	Hierarchial Neural Networks
HPLC	high performance liquid chromatography
HRP	histidine rich protein
IMAC	immobilised metal affinity chromatography
IPTG	isopropyl-β-D-thiogalactopyranoside
LB	Luria-Bertani medium
LDH	lactate dehydrogenase
MMV	Medicines for Malaria Venture
MVI	Malaria Vaccine Initiative
NMR	nuclear magnetic resonance
nt	nucleotide
OD	optical density
PABA	p-amino benzoic acid
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PPPK	hydroxymethyl pyrophosphokinase
PYR	pyrimethamine
RBS	ribosome binding site
SDS	sodium dodecylsulfate
SDX	sulfadoxine
SERCA	sarcoplasmic-endoplasmic reticulum Ca^{2+} ATPase
SHMT	serine hydroxymethyltransferase

SNP	single nucleotide polymorphism
TEMED	N,N,N'N'-tetramethylenediamine
Tris	tris (hydroxymethyl) aminomethane
tRNA	transcription ribonucleic acid
TS	thymidylate synthetase
U	enzyme units
UTR	untranslated region
X-gal	5-bromo-4-chloro-3-indolyl--β-D-galactopyranoside

List of tables

Chapter 1

Chapter 2

Chapter 3

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 1

Literature Review

"Our vision is a world in which affordable drugs will help eliminate the devastating effects of malaria and help protect the children, pregnant women and vulnerable workers of developing countries from this terrible disease."
-Medicines for Malaria Venture, 2003

1.1 The burden of malaria

In Africa, a child under the age of five dies of malaria every 30 seconds of every day (2003 Africa Malaria Report by the World Health Organisation (http://mosquito.who.int/amd2003). The total malaria mortality in sub-Saharan Africa constitutes about 90% of nearly 3 million deaths caused by the disease per year. Adults living in endemic areas have acquired some immunity to malaria and therefore its impact is mostly observed in children (Marsh *et al.*, 1995). Low birth weight, improper nutrition, low school attendance rates (Brooker *et al.*, 2000), learning disabilities, behavioural disorders and loss of motor functions are disease-related factors that impede proper education and thus the development of the population in general (Holding and Snow, 2001). Apart from the social burden, other repercussions of the disease are immense poverty and lack of development in endemic countries, brought about by escalating public health costs and loss of labour (Gallup and Sachs, 2001). The economic burden of malaria is evident through the direct geographic correlation between disease severity and poverty and resultantly the percentage total income of an endemic country is reduced by up to a half of its potential (Sachs and Malaney, 2002).

1.2 Malaria pathogenesis

Plasmodium falciparum is the parasite species responsible for the most severe form of the disease in humans (Miller *et al.*, 2002). Humans are the hosts for the asexual phase of the parasite life cycle and contract the illness from the bite of an infected female *Anopheles* mosquito, the host for the sexual development phase (Figure 1.1). The exo-erythrocytic cycle starts with the infection of liver cells by sporozoites where they develop into schizonts. Schizonts rupture to release merozoites into the blood stream to re-invade red blood cells. During this process the red blood cells become deformed as the merozoites develop into trophozoites (Miller *et al.*, 2002). Most damage occurs during sequestration when infected red blood cells adhere to capillaries and restricts blood flow to the organs, resulting in organ failure (Miller *et al.*, 2002). Trophozoites develop into gametocytes, which are released through red blood cell lysis, ending the erythrocytic cycle. As the mosquito takes up blood from the human host, the sexual (sporogenic) cycle starts by fusion of the gametes in the mosquito gut. The main clinical symptom responsible for mortality is respiratory distress due to decreased oxygen supply to tissues. Lactic acidosis, severe anaemia and coma are other symptoms resulting ultimately in damage to the cerebral- and central nervous systems, and also the kidneys and lungs (Miller *et al.*, 2002).

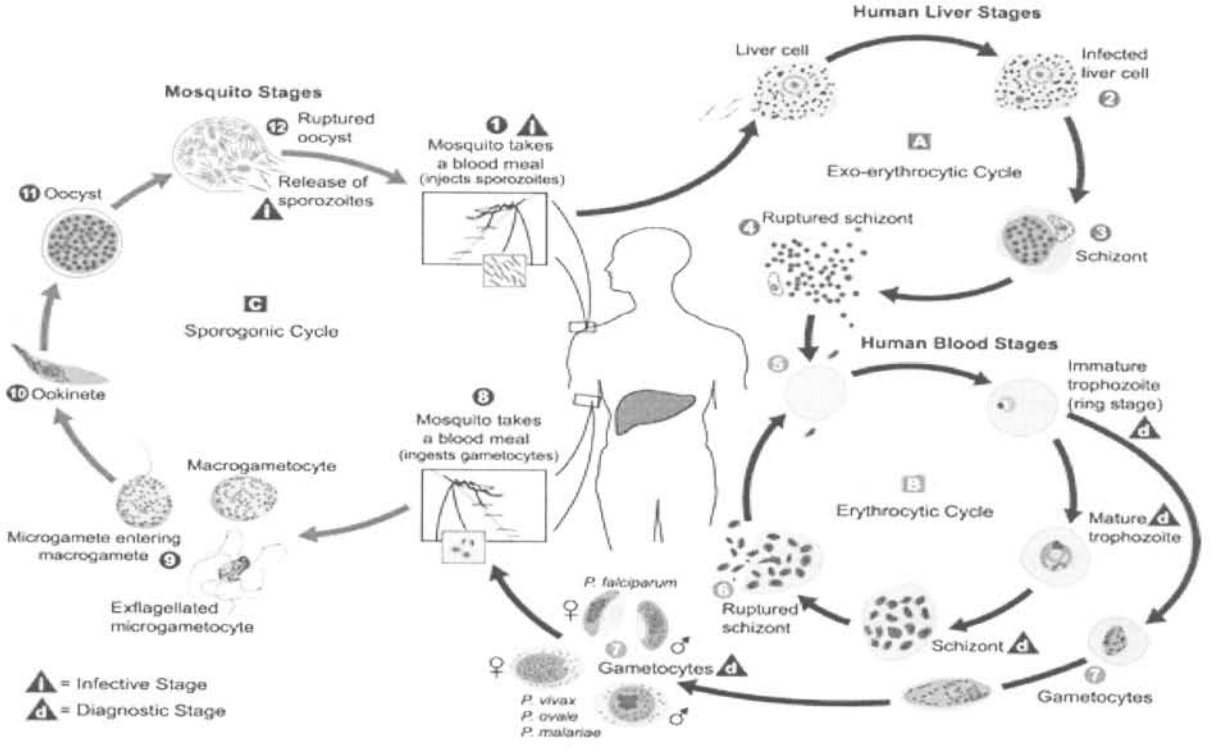


Figure 1.1: The life cycle of the human malaria parasite *P. falciparum*. (www.dpd.cdc. gov/dpdx/HTML/Malaria)

1.3 The multifaceted problem of malaria

As described in section 1.2, the chain of malaria transmission can thus be grouped into three interconnected links. Current malaria control strategies aims at the disruption of either of these links or disruption of contact points between the links (Figure 1.2). Apart from strategies targeting each link separately, transmission control focuses at the contact points between the mosquito and parasite and mosquito and human, whereas disease control focus on the contact points between the parasite and humans. The main problems of malaria control strategies are the resistance of the parasite to chemotherapeutic drugs, insecticide resistance of the mosquito and social and environmental changes brought about by human migration (Muentener *et al.*, 1999). Each of these control strategies will be discussed in more detail in the following sections.

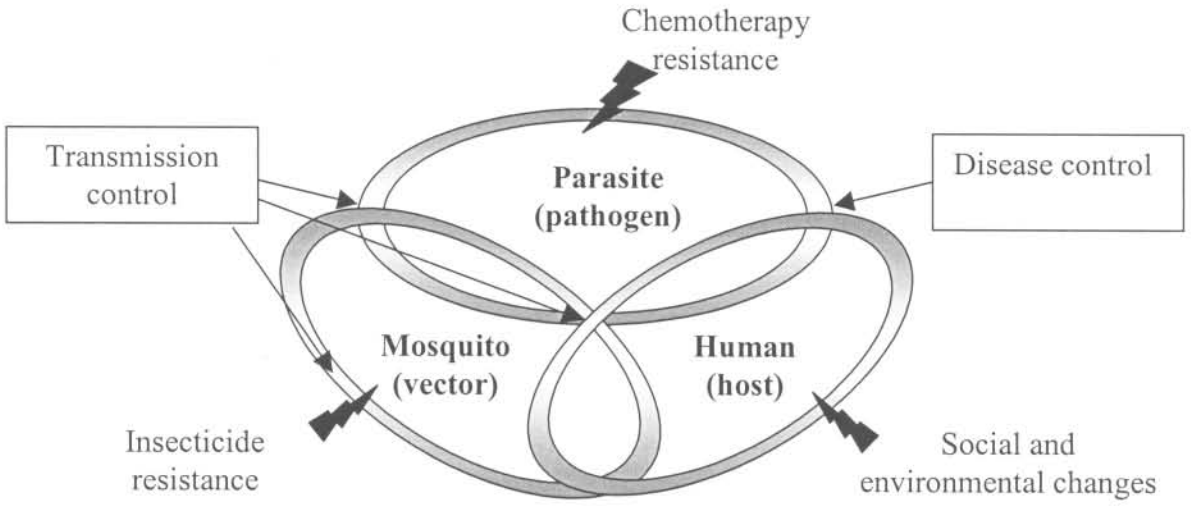


Figure 1.2: The three main links in the cycle of malaria transmission.

1.3.1 Vector control

Transmission control strategies aimed at the main mosquito vector, *Anopheles gambiae,* involves primarily the use of insecticide-treated bed nets to minimise human-vector contact (Armstrong-Schellenberg *et al.*, 2001). Pyrethroid resistance of the mosquito vector largely hampers the use of such insecticides (Kristan *et al.*, 2003). Genetic strategies aimed at genotyping different mosquito species for determination of transmission patterns are curbed by the diverse genotypes of the *Anopheles* species (Gentile *et al.*, 2001). The

availability of the *Anopheles gambiae* genome however, is a recent advance that would provide a platform for better vector characterisation, understanding of the mechanisms of insecticide resistance and pave the way for the development of new effective insecticides (Hoffman *et al.*, 2002). At the contact point between the mosquito and parasite, attempts to develop transgenic mosquitoes incapable of sustaining the parasite, might result in the reduction of parasite transmission (James, 2003). At the contact point between the mosquito and human, satellite imagery is employed to predict the variation in *A. gambiae* distribution and density that is caused by climatic and environmental changes and thus high transmission risk areas (Rogers *et al.*, 2002).

1.3.2 Host genetic factors

Individual genetic traits such as the human leukocyte antigen (HLA)-, haemoglobin- and red blood cell type are responsible for different host susceptibilities (Miller, 1999). Single nucleotide polymorphisms (SNPs) within various immune response genes of infants were analysed by means of the amplification refractory mutation system PCR (ARMS PCR), which showed that a CD36 T1264G mutation was responsible for the hosts' parasite clearing phenotype (Djimde *et al.*, 2002). In another independent study it was also observed that resistance to malaria was linked to glucose-6-phosphate dehydrogenase (G6PD) and CD40 ligand variants containing protective mutations (Sabeti *et al.*, 2002). Since human susceptibility is mostly an individual characteristic, this information though useful, is limited by the patients' genotype.

1.3.3 Pathogen control

1.3.3.1 Cutting-edge technologies

In 2002 the full genome sequence of *P. falciparum* (strain 3D7) was completed (Gardner *et al.*, 2002). This 30 million basepair genome is divided into 14 chromosomes, a mitochondrial genome and a circular plastid genome, which all have an extraordinary high A+T content of approximately 80% (Gardner *et al.*, 2002). Roughly 5400 protein-encoding genes were identified, 60% of which have unknown functions (Gardner *et al.*, 2002). The availability of the sequence data has thus enabled the following functional genomics studies of the parasite:

- Knockouts for the assignment of unknown gene functions, such the functional assignment of knob-formation to the HRP1 gene (Crabb *et al.*, 1997)
- Comparative genomics between different *Plasmodium* species, which led to the construction of the *Plasmodium* genome database (http://PlasmoDB.org).
- Comparative genomics between *Plasmodium* genes and their homologues (Wirth, 2002).
- Assignment of variant antigen families (Carlton *et al.*, 2002)
- Determination of drug resistance loci (Wirth, 2002).
- Microsattelite typing to identify areas under selection pressure (Li *et al.*, 2002)

For the first time it is possible through these cutting-edge technologies to obtain a global view of the functioning of the parasite in terms of the regulation of mRNA production (transcriptome) as well as protein expression (proteome). A better understanding and prediction of the response of the parasite when exposed to various external conditions would enhance the current pathogen control strategies.

1.3.3.2 Vaccines

Malaria is prevalent in countries that lack the financial and social infrastructure to effectively prevent such an illness by the distribution of bed nets and expensive prophylaxis. Compliance by the community is also a serious problem. An effective and safe vaccine would address these issues by means of a single-dosage treatment for sustained pathogen control. Most effective vaccines in the past (e.g. against rabies) have been directed towards infectious agents that elicit immune responses from their host and are then eliminated by the body's own defence system. The malaria parasite, however, has elaborate mechanisms to evade the human immune system and so establish a chronic infection, which is beneficial for the continuance of the parasite life cycle by means of transmission to the mosquito (Saul, 1999). In cases where evasion by the parasite fails, the human immune response manages to control the parasite infection but the exact mechanism by which this happens is not known (Richie and Saul, 2002). A second problem is the immense variability of parasite antigens, which depend on the specific stage of the life cycle of the parasite as well as the polymorphism of such genes (Saul, 1999). Lastly, humans also vary greatly in terms of their immune response towards the parasite,

which depends on their haemoglobin type and HLA (human leukocyte antigen) locus (Richie and Saul, 2002).

On the other hand, the fact that individuals repeatedly infected with malaria develop a natural acquired immunity that prevents the fatal symptoms involved in the disease (as mentioned in 1.3.2), combined with the success of vaccines in animal models, raises the hope that a vaccine is a realistic expectation (Stowers *et al.*, 2001). Current vaccine strategies are aimed at the pre-erythrocytic stage (Kester *et al.*, 2001), the red blood cell stages (merozoite surface protein 1 and 3) (Stowers *et al.*, 2001) of the parasite life cycle or a combination of stages through the use of multiple epitopes (Doolan and Hoffman, 2001). The MVI (Malaria Vaccine Initiative, Maryland, USA) launched clinical stage one trials in children under five years of age for a sporozoite vaccine candidate; RTS,S (CS circumsporozoite protein containing a hepatitis B virus core and adjuvant for enhanced response) in July 2003 and the results will be available at the end of 2005 (Kester *et al.*, 2001). Despite much time, cost and effort, an effective vaccine still remains elusive and will therefore not be available in the near future.

1.3.3.3 Drugs

In contrast to vaccines, drugs have been available for over fifty years and are currently the primary aspect in the parasite control strategy (www.rbm.int). The main problem with the use of drugs is the rapid evolution of drug-resistant parasite strains (Sibley *et al.*, 2001). This is mainly the result of non-compliance by the population where the parasites are not totally eradicated, but remain in undetectable levels in the blood and so develop resistance. This is also influenced by the biological half-life of the drugs. The longer the half-life, the higher the chances of resistance development since the drugs then take longer to be cleared from the body (Ridley, 2002).

Table 1.1: Problems associated with antimalarial drugs (adapted from Ridley 2002).

Antimalarial drug	Main limitations
Prophylaxis	
• Chloroquine	• Resistance
• Amodiaquine	• Safety, resistance
• Mefloquine	• Safety, resistance, cost
• Halofantrine	• Safety, resistance, cost
• Doxycycline	• Low effectivity
Treatments	
• Sulfadoxine-pyrimethamine (treatment)	• Resistance
• Quinine	• Compliance, safety, resistance
• Atovaquone-proguanil (malarone)	• Potential resistance, cost
• Dapsone-proguanil (lapdap)	• Potential resistance, cost
• Lumefantrine-artemether	• Compliance, potential resistance, cost
• Artemisinins (artemether, erteether, artesunate)	• Compliance, safety, cost

For drugs to be a viable control strategy in Africa, they have to meet the criteria of low costs (to enable large scale distribution), short half-lives or single dosage options (to lessen the effect of weak compliance), effectivity and safety (Ridley, 2002).

The genome sequences of the parasite, plastid and mitochondria provide a virtually unlimited source for the discovery of novel drug targets and screening of possible inhibitors. The next section will describe in more detail the focus areas of current drug development, the functions of the mentioned antimalarial drugs and future prospects.

1.4 Drug targets

The effectivity and safety (as determined by parasite specificity) of drugs depend mainly on differences between the parasite and its hosts' metabolism. The molecular targets are not known for all drugs and therefore drugs can be broadly classified according to their subcellular locations (Figure 1.3). The figure below indicates the subcellular compartments

where current antifolates are active as well as explores the options for future drug targets. Current antimalarial drugs are either targeted to the lysosomal food vacuole (quinolines), mitochondrion (atovaquone), apicoplast (doxycycline) or the cytosol (antifolates). For each of these compartments, many other potential drug targets exist such as proteases (lysosomal food vacuole), dihydroorotate dehydrogenase (mitochondrion), fatty acid type II biosynthesis (apicoplast), enzymes involved in folate and nucleotide biosynthesis (cytosol) as well as nutrient transport mediated by the plasma membrane proteins.

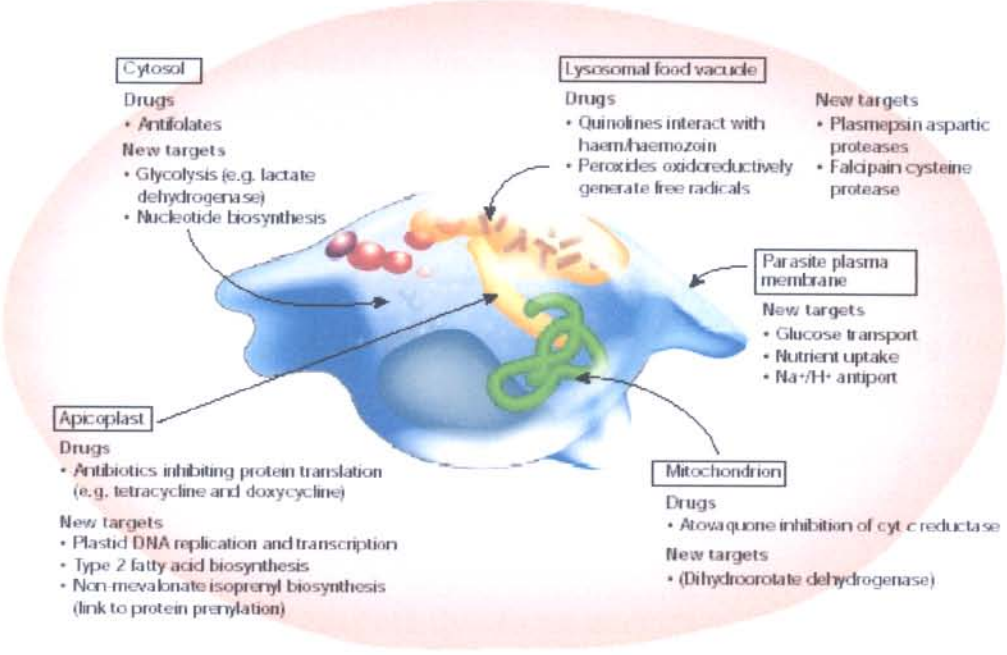


Figure 1.3: The subcellular locations of past, present and future drug targets. A schematic representation of a trophozoite in a red blood cell (Ridley, 2002).

1.4.1 Haemoglobin degradation in the lysosomal food vacuole

Quinoline and artemisinin antimalarial drugs are concentrated in the food vacuole (Figure 1.3). These drugs are expected to interfere with the breakdown of toxic haem, Fe (II) to the oxidised form hematin, Fe (III) and the subsequent formation of hemozoin (Francis *et al.*, 1997). The molecular target of artemisinin drugs was determined as PfATP6, which is a sarcoplasmic-endoplasmic reticulum Ca^{2+} ATPase (SERCA) ortholog but the molecular targets of the quinolines are unknown (Krishna *et al.*, 2003). Novel targets involved in haem metabolism are studied as putative drug targets e.g. the aspartic proteases or plasmepsins, the cysteine protease falcipain, and the metallopeptidases (Coombs *et al.*, 2001) (Shenai *et al.*, 2000) and (Eggleson *et al.*, 1999).

1.4.2 Apicoplast metabolism.

The apicoplast is an organelle unique to the malaria parasite, which is similar to a plant's chloroplast. It also contains prokaryotic transcription and translation elements on its circular chromosome that could possibly explain the antimalarial action of bacteriostatic agents such as doxycycline (Fichera and Roos, 1997). The apicoplast proteins comprise 10% of the total parasite proteins and were identified by means of a signal signature motif shared between these proteins (Foth *et al.*, 2003). This expands the drug target possibilities immensely and opens new possibilities for herbicides as drugs. The herbicide fosmidomycin targeted to 1-deoxy-D-xylulose-5-phosphate synthase, was for instance shown to have anti-malarial activity (Lell *et al.*, 2003). The farnesyltransferase proteins which link up with the deoxyxylulose phosphate (DOXP) pathway through the farnesylation of isopentenyl diphosphate might also be potent drug targets (Ohkanda *et al.*, 2001). The apicoplast is furthermore the centre for type II fatty acid synthesis (Figure 1.3). In humans the fatty acid synthase II system (FASII) is absent. One of the drug discovery projects launched by the Medicines for Malaria Venture (MMV) is the development of inhibitors against enoyl ACP reductase, a key enzyme in this metabolic pathway (MMV Annual report 2002).

1.4.3 Limited electron transport due to an acrystate mitochondrion

The malaria parasite lives in an oxygen poor environment and relies on glycolysis as the main source of ATP. There is thus no real oxidative phosphorylation and enzymes normally associated with electron transport perform other functions. An example of this is cytochrome *c* reductase, which is linked to dihydroorotate dehydrogenase, an enzyme involved in nucleotide biosynthesis (Fry and Beesly, 1991).

1.4.4 Cytosolic and membrane bound drug targets

The dependence of the parasite on glycolysis for energy production (as described in 1.4.3 above) also makes the cytosolic enzyme lactate dehydrogenase (LDH), which is involved in anaerobic glycolysis, an important target (MMV Annual Report 2002). Although humans have LDH, they are less dependent on this enzyme for energy production since the main source of ATP production is the citric acid cycle and electron transport systems (Koukourakis *et al.*, 2003). The primary cytosolic target is the folate pathway. Folate

metabolism is a validated drug target since it differs significantly between the human host and parasite (Ferone, 1977). Past successes with the use of the antifolates furthermore underscores the importance of folate metabolism as a drug target. The current resistance of the malaria parasite to the available antifolates however calls for new drug targets to be identified within this pathway. Section 1.5 will deal with folate metabolism of the malaria parasite, its current status, targets and future possibilities.

Apart from the above-mentioned targets, a variety of membrane proteins involved in e.g. the transport of glucose, other nutrients and essential ions are also potential important points of drug inhibition (Manning *et al.*, 2002) and (Desai *et al.*, 2000).

1.5 The folate pathway

1.5.1 Folate metabolic organisation

Folates are cofactors used in essential reactions of prokaryotic and eukaryotic cells. Tetrahydrofolate (vitamin B9) is used in one-carbon transfer reactions that take place during DNA synthesis (conversion of dUMP to dTMP and purine synthesis), formylmethionine tRNA formation and amino acid biosynthesis (conversion of serine to glycine) (Ferone, 1977). Antifolates have a wide range of applications, not only as antimalarials but also as anti-cancer agents such as methotrexate (Edelman and Gandara, 1996). Through targeting specific enzymes of the folate pathway, DNA base and amino acid synthesis involving a few 1-carbon transfer reactions, are disrupted resulting in cell death (Edelman and Gandara, 1996). Despite the absolute need for folates, different organisms have different ways of obtaining folates. Humans and mammals obtain preformed, reduced folates from their diet whereas prokaryotes, plants, fungi, yeast and the malaria parasite are able to synthesise folates. The *de novo* folate synthesis from guanosine 5'-triphosphate (GTP), p-aminobenzoic acid (PABA) and L-glutamate as well as polyglutamylation of p-aminobenzoylglutamate and pterin-aldehyde (host serum folate degradation products) indicated a unique biosynthetic pathway in the human malaria parasite, as shown in figure 1.4 (Krungkrai *et al.* 1989). Folate metabolism and the enzymes involved thus greatly differ between the parasite and its human host and thus make this metabolic pathway a validated antimalarial drug target (Hitchings, 1971). As shown in figure 1.4, folate biosynthesis can be divided into two main routes, namely *de*

novo synthesis and folate salvage. *De novo* synthesis uses GTP as starting metabolite and ends with the production of dihydrofolate (H_2folate) where it links up with the thymidylate cycle, which is the recycling of dihydrofolate through various steps to provide the essential cofactors needed in metabolism (Figure 1.4).

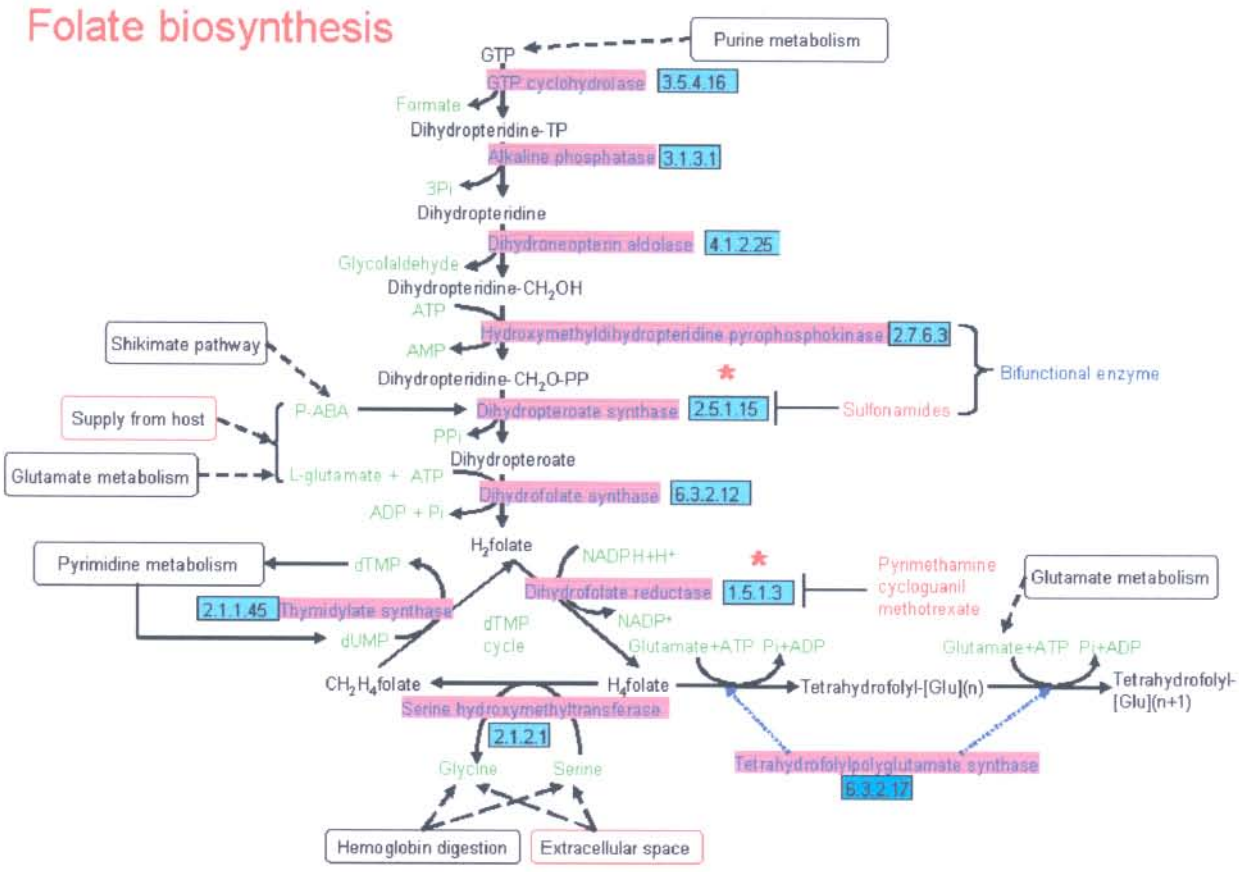


Figure 1.4: Folate metabolism in the human malaria parasite *P. falciparum*. (http://sites.huji.ac.il/ malaria/maps/ folatebiopath.html). Enzyme Commission (E.C.) numbers are indicated next to the enzyme names. Stars indicate enzymes targeted by current antifolates and the links with other metabolic pathways are shown in white boxes.

1.5.2 Current status of the antifolates

Antifolate targets are the dihydropteroate synthase activity of bifunctional hydroxymethylpterin pyrophosphokinase-dihydropteroate synthase (PPPK-DHPS) and the dihydrofolate reductase activity of dihydrofolate reductase-thymidylate synthetase (DHFR-TS) (Brooks *et al.*, 1994; Hyde, 1989). PPPK-DHPS is responsible for the formation of dihydropteroate in the *de novo* pathway. DHFR-TS is responsible for the cycling of dihydrofolate to tetrahydrofolate and *vice versa* in the thymidylate cycle (Ferone, 1977). The antifolates sulfadoxine (SDX) and pyrimethamine (PYR) are inhibitors of DHPS and DHFR, respectively (Brooks *et al.*, 1994; Hyde, 1989). Specific advantages of

pyrimethamine are a much higher affinity for the parasite DHFR than the host homologue and that its' inhibition of the enzyme is late in the erythrocytic cycle when schizonts form and DNA synthesis reaches a maximum (Yuthavong, 2002). The antifolates are also structurally very similar to the respective DHPS and DHFR substrates. This mode of competitive inhibition unfortunately placed the enzymes' active sites under selection pressure, forcing the development of mutations necessary for survival. This resulted in resistant strains with point mutations in the coding sequences of target genes that limited the useful therapeutic life of the drugs (Table 1.2) (Sibley *et al.*, 2001). When comparing the characterized *dhfr-ts* gene (Hyde, 1989) as well as *pppk-dhps* (Brooks *et al.*, 1994) with various clinical isolates, it was evident that certain point mutations in *dhfr* (PYR resistant strains) and *dhps* (SDX resistant strains) were responsible for the reduced affinity of DHFR and DHPS for the drugs, i.e. resistance. It was also discovered in *dhps* that the number of mutations was directly and quantitatively related to sulfadoxine resistance and that a key residue, Glycine 437, was common to all mutants (Triglia *et al.*, 1998). The new resistant strains were subjected to allele-specific tests for the detection of point mutations. Comparison of the point mutations between different strains has led to the construction of maps of mutation patterns and worldwide distribution frequencies (Wang *et al.*, 1995).

Table 1.2: Primary mutations responsible for antifolate resistance (Sibley *et al.*, 2001). Point mutations responsible for resistance are underlined. One-letter amino acid abbreviations are given in brackets.

DHFR gene	**codon 16**	**codon 51**	**codon 59**	**codon 108**	**codon 164**
Wild type	GCA (A)	AAT (N) AAC (N)	TGT (C)	AGC (S)	ATA (I)
Variants	G<u>T</u>A (V)	A<u>T</u>T (I)	<u>C</u>GT (R)	A<u>A</u>C (N) A<u>C</u>C (T)	<u>T</u>TA (L)
DHPS gene	**codon 436**	**codon 437**	**codon 540**	**codon 581**	**codon 613**
Wild type	TCT (S)	GCT (A)	AAA (K)	GCG (A)	GCC (A)
Variants	<u>G</u>CT (A) T<u>T</u>T (F)	G<u>G</u>T (G)	<u>G</u>AA (E)	G<u>G</u>G (G)	<u>T</u>CC (S) <u>A</u>CC (T)

Resistance to the antifolates is addressed through the development of novel inhibitors such as WR99210 that inhibits drug-resistant DHFR, through rational drug design (McKie *et al*, 1998). The recent determination of the *P. falciparum* DHFR-TS crystal structure has revealed the mode of action of this rationally designed inhibitor and will also be useful for the rational design of other antifolates targeted to DHFR (Yuvaniyama *et al.*, 2003).

1.5.3 Drug synergy

In the past, the treatment of resistant strains was modified by the synergistic use of drugs that target different points in the pathway (Sibley *et al.*, 2001). Such drugs were Fansidar, a pyrimethamine/sulfadoxine combination or Maloprim, a pyrimethamine/dapsone combination. Inhibition occurred when the parasite was sensitive to either sulfadoxine or pyrimethamine and therefore the obstacle of resistance to the other drug was eliminated. The success of these drugs could be attributed to the fact that they blocked both possible routes of folate metabolism, namely, *de novo* synthesis and salvage, but the exact mechanism of the synergistic action is unknown and further knowledge of the folate metabolic pathway is thus required. The success of the synergistic application of pyrimethamine and sulfadoxine was however very short-lived (less than five years) as resistance developed to the drug combination (Sibley *et al.*, 2001). The exact mechanisms of resistance to the sulfadoxine-pyrimethamine combination is not yet known, but the minimum resistance requirements are a triply mutant DHFR (S108N, C59R and N51I) as well as a doubly mutant DHPS (Sibley *et al.*, 2001). Such strains are highly pathogenic and the use of combination therapy and synergistic drug treatments should thus be handled with great care since selection for more potent parasites could result from drug misuse (Sibley *et al.*, 2001).

1.5.4 Exogenous folate utilisation

The level of sulfadoxine resistance of parasites under monotherapy treatment fluctuates between various parasite strains depending on the serum concentration of folates of the host (Wang *et al.*, 1997). Additional mechanisms to alter the DHFR and DHPS parts of the bifunctional target enzymes were responsible for this, such as the ability of parasites to use exogenous folic acid (Wang *et al.*, 1999). Genetic crosses between sulfadoxine sensitive and –resistant parasites showed that the folate utilizer phenotype was gained by resistant parasites and was linked to the *dhfr* gene but not dependant on DHFR activity (Wang *et al.*, 1999). Furthermore it was observed that sulfadoxine resistant parasites reverted to sensitivity when a pyrimethamine-sulfadoxine combination was used. It was postulated that this so-called 'folate effect' was responsible for the bypassing of sulfadoxine inhibition of DHPS and the observed synergistic inhibition when pyrimethamine was added.

1.5.5 Putative drug targets within the folate pathway

To date, only two activities within the folate pathway have been targeted by the antifolates, DHFR and DHPS. When considering that almost half of the enzymes in the folate pathway have no human counterparts, the uncharacterised folate metabolising enzymes are promising drug targets. Three novel *P. falciparum* genes within the folate pathway have been identified but their respective enzymes have not yet been characterised (Lee *et al.*, 2001). The first of these is the gene encoding GTP cyclohydrolase (GTP-CH, E.C: 3.5.4.16), which catalyses the rate limiting step in *de novo* folate synthesis: the release of formic acid and formation of a triphosphate pteridine ring structure. In humans this enzyme has a different function, i.e. the formation of reduced biopterin since *de novo* folate synthesis is absent (Stokstad and Koch, 1967). Alignment between the deduced *P. falciparum* GTP-CH and its human homologue revealed a parasite specific N-terminal extension of 130 amino acids as well as 17% identity and 32% similarity in the aligned areas (Lee *et al.*, 2001). The second gene identified encoded serine hydroxymethyltransferase (SHMT, E.C: 2.1.2.1) that converts serine into glycine with the use of tetrahydrofolate as a 1-carbon acceptor (Lee *et al.*, 2001). Alignment with the human SHMT homologue indicates a higher amount of conservation, 43% identity and 63% similarity, which makes it a less favourable drug target when considering selective inhibition (Lee *et al.*, 2001). The last of the three genes identified encoded both dihydrofolate synthase and folylpolyglutamate synthase (DHFS-FPGS EC: 6.3.2.12 and EC: 6.3.2.17 respectively) activities as its bifunctionality was verified in yeast knockouts (Salcedo *et al.*, 2001). The next section will focus on this enzyme as the main topic of the thesis.

1.5.6 Bifunctional dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS) as a possible drug target.

The bifunctional enzyme DHFS-FPGS is used in both *de novo* synthesis to form dihydrofolate by addition of L-glutamate to dihydropteroic acid, as well as in folate salvage to extend the chain of glutamate residues, which is important for the regulation of intracellular folate pools (Stokstad and Koch, 1967). The fact that all organisms have FPGS activity to synthesize folylpolyglutamate, regardless of whether they are capable of *de novo* folate synthesis or not, underlines the important function of FPGS.

The polyglutamate tail addition is thus part of folate salvage and has the following important functions:

- It prevents efflux of cofactors from the cell, thus cellular folates are retained (McGuire and Bertino, 1981)
- It increases the affinity of folate synthesizing enzymes for folates. Most enzymes only use polyglutamated substrates (McGuire and Bertino, 1981)
- It increases the accumulation of folates in the mitochondria where it is used for glycine synthesis (Bognar *et al.*, 1985; Shane, 1980).

Much less is known of DHFS activity except that it is used for *de novo* folate biosynthesis in organisms that synthesise their own folates. DHFS and FPGS catalyse similar reactions in folate metabolism but function in two different routes of the pathway. If these two *Plasmodium* enzymes are co-linear in the bifunctional protein, as is the case in *E. coli* (Keshavjee *et al.*, 1991), then a single inhibitor would have a very powerful effect in blocking both routes of folate metabolism. This would achieve the same result as the synergistic application of drugs such as sulfadoxine and pyrimethamine that target separate points; each within another route of the folate metabolic pathway. On the other hand, it might be preferable to inhibit only the DHFS activity without affecting FPGS activity as it was shown in human cancer cells that FPGS inherently enhances the effect of antifolates (substrate analogues) (Gangjee *et al.*, 2002). This is accomplished by increasing the retention of the antifolates through polyglutamylation as well as the affinity of the folate utilising enzymes for the inhibitors (Gangjee *et al.*, 2002). Considering that both of these activities reside in the same protein it is thus important to determine the structural features necessary for FPGS activity in order to design drugs for the selective inhibition of DHFS.

What makes this enzyme furthermore an appealing drug target is that humans do not have a DHFS homologue. The extent of conservation (30% similarity and 17% identity) between human FPGS and *P. falciparum* DHFS-FPGS is also lower than between GTP-CH and SHMT and their respective homologues (Lee *et al.*, 2001). The whole DHFS-FPGS sequence is aligned against human FPGS, since there are no distinct DHFS and FPGS domains at the primary amino acid or DNA level in *P. falciparum*. The *dhfs-fpgs* gene identified in *P. falciparum*, is also unique since it is the first bifunctional gene

identified in eukaryotes (Lee *et al.*, 2001). The effect of current antifolates on DHFS-FPGS was determined in a gene complementation system. This was based on the fact that DHFS-FPGS and DHFR both contain ATP- and folate binding sites. The *P. falciparum dhfs-fpgs* gene showed no mutations after growth at inhibitory concentrations of antifolates (Salcedo *et al.*, 2001). It seems promising that this enzyme would be less prone to the acquisition of resistance, an advantage for drug development.

To date, *P. falciparum* DHFS-FPGS has been characterized in terms of DNA sequence and bifunctionality verified by means of complementation of *E. coli* and yeast mutants (Lee *et al.*, 2001; Salcedo *et al.*, 2001). No other characteristics of this enzyme are known. To determine the potential value of this enzyme as a novel antifolate target, studies concerning the parasite-specific traits, kinetic parameters, activity and three-dimensional structure are needed. This requires sufficient amounts of correctly folded, soluble protein. The native parasite enzyme is only expressed at undetectable levels from gene-complemented *E. coli* or yeast systems and direct isolation from the parasite results in low yields (personal communication, J. Hyde). Expression of malaria genes in a heterologous system is greatly hampered by their A+T richness and rare codon usage (Baca and Hol, 2000). This study therefore concerns the synthesis of the *dhfs-fpgs* gene adapted to the codon preferences of *E. coli,* and the optimisation of its expression in *E. coli*.

1.6 Research aims

The primary aim of the study is to obtain sufficient quantities of DHFS-FPGS for future kinetic characterisation and crystallisation or nuclear magnetic resonance (NMR) studies.

- Chapter 2 focuses on the PCR mediated synthesis of a synthetic *P. falciparum dhfs-fpgs* gene, which is modified for expression in *E. coli* by the use of preferred *E. coli* codons.
- Chapter 3 focuses on the heterologous expression of synthetic *P. falciparum dhfs-fpgs* in a variety of *E. coli* expression hosts and vector systems to obtain soluble expressed protein and verification of enzymatic activity through functional complementation.
- Chapter 4 describes the *in silico* analysis of the predicted primary amino acid sequence to determine its similarity with the other homologues and predict structural features.
- In Chapter 5 the relationship between the above chapters, their relevance to drug discovery and development and future prospects are discussed.

Chapter 2

PCR-mediated synthesis of the *P. falciparum dhfs-fpgs* gene

2.1 Introduction

The genetic organisation of *dhfs* and *fpgs* varies greatly between organisms. In yeast three genes, each on a separate chromosome, encodes the DHFS and FPGS functions (Cherest *et al.*, 2000). *Arabidopsis thaliana* (Ravanel *et al.*, 2001) also have separate genes encoding for DHFS and the various isoforms of FPGS. In *E. coli* (Bognar *et al.*, 1985) and *P. falciparum* (Lee *et al.*, 2001) the DHFS and FPGS activities are encoded by a single gene. The *P. falciparum* gene consists of four exons (70% of the mRNA transcript) and three introns (Figure 2.1). Three ATG codons, from which differential translation could take place, are found at amino acid positions 1, 17 and 41 (Figure 2.1). It was shown that the whole length of the sequence (from the first ATG codon) was necessary to complement yeast DHFS-FPGS mutants (Salcedo *et al.*, 2001). It is not known whether transcription of the truncated gene (from codons 17 or 41) takes place *in vivo*.

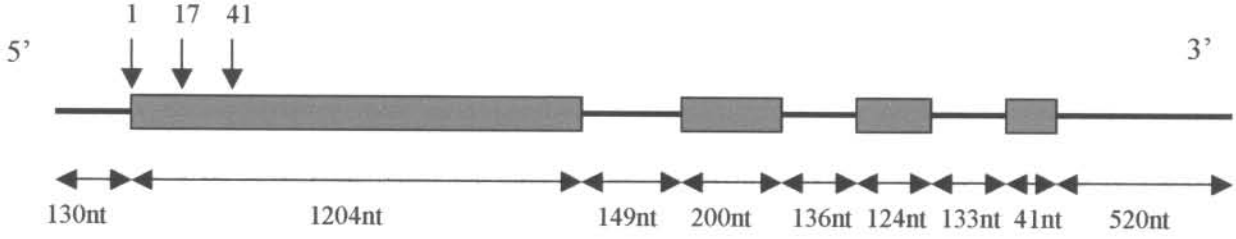


Figure 2.1: Schematic representation of the genetic organisation of *P. falciparum dhfs-fpgs*. Coloured boxes indicate the exons or coding regions. Black lines between the boxes indicate the intron positions and horizontal lines at the 5' and 3' ends indicate the 5' and 3' untranslated regions (UTR's), respectively. Position 1, 17 and 41 indicate ATG codons.

In most organisms DHFS and FPGS are natively expressed only at very low levels, making characterisation of the enzymes very difficult. Isolation of native *E. coli* DHFS-FPGS

encoded by the *folC* gene has such low yields that the gene had to be expressed in a high expression plasmid to obtain sufficient protein for catalytic studies (Bognar *et al.*, 1985). The native *P. falciparum* DHFS-FPGS is only expressed at very low levels in heterologous systems such as *E. coli* or yeast (personal communication, J. Hyde). Heterologous expression of malaria proteins is hampered by codon bias (lysine and arginine preferences) and the fact that the parasite genome is extraordinary A+T rich. The overall percentage of A+T content is 82% (including the plastid genome) and 76% for nuclear genes alone (Gardner *et al.*, 2002). The abundance of poly-A sites can furthermore act as termination signals during transcription, resulting in truncated mRNA and incomplete proteins after translation (Romanos *et al.*, 1991). In order to circumvent codon bias and high A + T content a separate plasmid encoding "rare" tRNAs for the amino acids arginine (AGA, AGG), isoleucine (AUA) and glutamine (GGA) can be introduced together with the plasmid containing the gene of interest (RIG plasmid). This plasmid enhances the expression ability of the host cells (Baca and Hol, 2000).

An alternative to the above method is the synthesis of a modified version of the gene from oligonucleotides, with a lower A+T content and preferred use of *E. coli* codons (Sugiyama *et al.*, 1996). The use of such a strategy has the further advantage that cassette mutagenesis of gene segments is made possible through the use of specific restriction enzyme sites in the gene. A chemically synthesised *P. falciparum* DHFR-TS gene was shown to express 10 times more efficiently from *E. coli* than the native DNA sequence (Prapunwattana *et al*, 1996). The gene was synthesised by cloning duplexes of complementary oligonucleotides ranging from 69 to 116 bp in length into 10 different vectors and restriction-ligation of the cloned fragments with the use of 31 different unique restriction enzyme sites. The oligonucleotides were designed to accommodate the differences in codon preferences in order to decrease the overall A + T content (Prapunwattana *et al.*, 1996).

Other methods describe the application of the polymerase chain reaction (PCR) for the synthesis of genes from overlapping oligonucleotides (Carpenter *et al.*, 1999). Completely overlapping oligonucleotides were assembled in a single PCR reaction, termed assembly PCR, through 'priming' of the complementary oligonucleotide overlaps (Figure 2.2). An additional PCR reaction with primers corresponding to the 5' and 3' outer ends was then used to obtain a sufficient amount of the full-length synthetic gene. This is, however, an

expensive method since the oligonucleotides overlap completely and requires a large amount of oligonucleotides (~125 pmoles of each) for template generation.

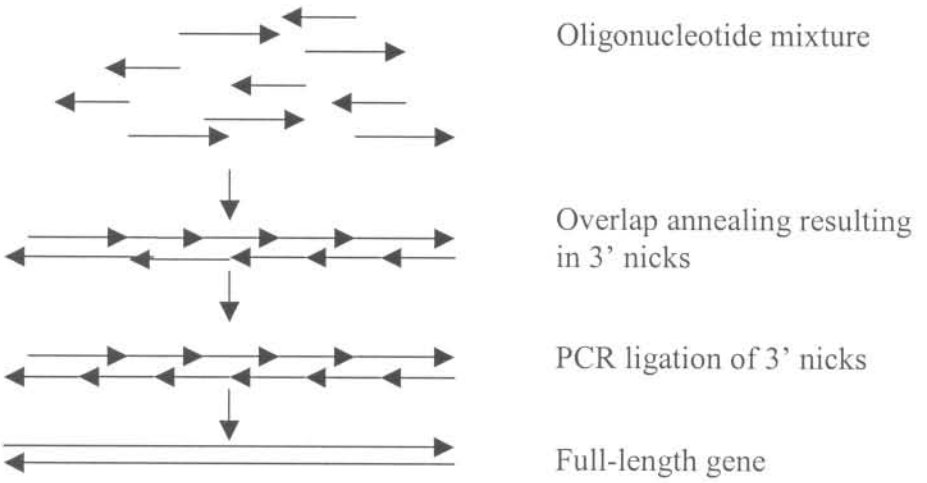


Figure 2.2: Overlap PCR used during gene synthesis (Carpenter *et al.*, 1999).

The strategy used here was to divide the full-length gene into halves, each constructed by means of PCR assembly from partially overlapping (about 20nt) oligonucleotides. The oligonucleotides prime initially only in the overlapping regions (Figure 2.4), leaving single stranded pieces of template which are extended during reiterative PCR cycles to generate the template (as will be described in the following sections). After a limited number of assembly steps and a 10 to 20-fold dilution of the mix, the PCR generated fragment is amplified with gene specific outer primers. Modifications introduced a 3' end reverse oligonucleotide for silencing of the stop codon to enable C-terminal His_6 expression from the pET22b vector. The full-length gene was constructed by overlap extension PCR of the halves and amplification with the designated outer primers.

2.2 Materials and methods

2.2.1 Oligonucleotides and primers

A total of 32 partially overlapping oligonucleotides, sixteen forward (f1-f16) and sixteen reverse (r1-r16) were designed for gene synthesis (Table 2.1). The first eight pairs of forward and reverse oligonucleotides were used for synthesis of the first half-gene and the last eight pairs for the second half-gene. Primers outF (5') and part1R (3') were used to amplify the first half fragment and primers part2F (5') and outR (3') were used to amplify the second half fragment. Oligonucleotide and primer sequences are given in appendix D.

Table 2.1: Forward (f) and reverse (r) oligonucleotides and forward (F) and reverse (R) primers used for the assembly of different gene fragments. The oligonucleotides involved in a specific overlap are indicated in brackets.

First half (nt 1-812)			Second half (nt 789-1586)		
Oligonucleotide name	**Length (nt)**	**Overlap T_m (°C) ***	**Oligonucleotide name**	**Length (nt)**	**Overlap T_m (°C)***
Forward 1 (f1)	72	53 (f1/r1)	Forward 9 (f9)	76	52 (f9/r9)
Reverse 1 (r1)	76	58 (r1/f2)	Reverse 9 (r9)	62	57 (r9/f10)
Forward 2 (f2)	76	53 (f2/r2)	Forward 10 (f10)	69	57 (f10/r10)
Reverse 2 (r2)	76	59 (r2/f3)	Reverse 10 (r10)	72	56 (r10/f11)
Forward 3 (f3)	71	60 (f3/r3)	Forward 11 (f11)	74	61 (f11/r11)
Reverse 3 (r3)	72	57 (r3/f4)	Reverse 11 (r11)	68	59 (r11/f12)
Forward 4 (f4)	69	56 (f4/r4)	Forward 12 (f12)	69	58 (f12/r12)
Reverse 4 (r4)	71	58 (r4/f5)	Reverse 12 (r12)	71	58 (r12/f13)
Forward 5 (f5)	73	58 (f5/r5)	Forward 13 (f13)	74	59 (f13/r13)
Reverse 5 (r5)	71	58 (r5/f6)	Reverse 13 (r13)	73	57 (r13/f14)
Forward 6 (f6)	76	58 (f6/r6)	Forward 14 (f14)	70	58 (f14/r14)
Reverse 6 (r6)	62	58 (r6/f7)	Reverse 14 (r14)	72	58 (r14/f15)
Forward 7 (f7)	76	49 (f7/r7)	Forward 15 (f15)	72	57 (f15/r15)
Reverse 7 (r7)	73	58 (r7/f8)	Reverse 15 (r15)	71	55 (r15/f16)
Forward 8 (f8)	68	57 (f8/r8)	Forward 16 (f16)	72	58 (f16/r16a)
Reverse 8 (r8)	75		Reverse 16a (r16a)	66	
			Reverse 16b (r16b)	69	
Primer name	**Sequence (5' to 3')**			**Length (nt)**	**T_m (°C) ***
outF	CGCGGACATATGGAAAAAAAC			21	56
part1R	CAGTTCAGTTCTTTCGCTTTATC			23	57
part2F	TGATAAAGCGAAAGAACTGA			20	51
outR	CGGATCC**TTA**CACCAGGCTC			20	61
alt4R	CTCGGATCCTG**TTT**CACCAGGCTCGGTTCGTTCAT			35	73

* T_ms were calculated with the formula: T_m=69.3 +0.41(%GC) - 650/ primer or overlap length in nt. Primer alt4R and oligonucleotide r16b were used to silence the stop codon (indicated in bold letters) at the 3' end of the gene.

2.2.1.1 Oligonucleotide design

Comparison of codon frequencies between *P. falciparum* and *E. coli* identified points of codon bias, which were eliminated in the design of the synthetic gene. Codon preference tables (www.kazusa.or.jp/codon), provided in Appendix A, were used to convert the native *P. falciparum* codons to the preferred *E. coli* codons (Figure 2.3). Factors taken into account in the design of the oligonucleotides were similar melting temperatures (<5°C deviation) of overlapping regions and primers to eliminate non-specific priming as well as minimal palindromes that could form secondary structures. The Oligo 4 Primer Analysis Software program (Molecular Biology Insights, MBI) was used to check internal stability and mispriming of the designed oligonucleotides and primers.

2.2.1.2 Oligonucleotide- and primer stock solutions and nucleic acid concentration determination

Oligonucleotides were obtained from MWG (UK) and were dissolved in 20% acetonitrile at 37°C at a stock solution concentration of 100 pmoles/µl. Oligonucleotide solutions were made up containing 1 pmole/µl of each of the oligonucleotides in TE buffer (0,1 mM Tris, 0,1 mM EDTA pH 7,6) and 20 µl aliquots were frozen away to minimise nucleotide damage through repeated freeze-thaw cycles. Primers were dissolved in 10 mM Tris (pH 8,6) at 37°C to give a final stock solution concentration of approximately 100 pmoles/µl each. Silanized tubes were used for primer and oligonucleotide storage at -20°C. The DNA concentration of each oligonucleotide and primer was determined spectrophotometrically at 260 nm. The manufacturer values for oligonucleotide mass per OD was used to determine the concentration according to the following formula:

Concentration = A_{260nm} × weight per OD (ng/µl) ×dilution factor / molecular mass (ng/nmole).

The DNA concentration of plasmids and gel-purified PCR products was determined either by ultraviolet absorbance (260nm) with 1 A_{260nm} unit equivalent to 50 ng/µl double stranded and 33 ng/µl single stranded DNA or by comparison with a range of calf thymus DNA standards spotted on glad wrap and stained with 10 µg/ml ethidium bromide (Sambrook *et al.*, 1989). Alternatively, concentrations of PCR products or isolated plasmids were also estimated by comparison of the fluorescence intensity of the sample

bands to markers of known concentrations (*EcoR*I and *Hind*III digested λ DNA, Promega, Wisconsin, USA). Sample band intensities were compared to marker band intensities of the same size to estimate the concentration.

2.2.2 PCR gene synthesis

The synthetic 1586 bp *dhfs-fpgs* gene was subdivided into ~750 bp halves. Each half fragment was constructed from 16 overlapping oligonucleotides and then amplified with the corresponding primers (Table 2.1). The outer primers were designed to incorporate an N-terminal *Nde*I site and a C-terminal *BamH*I site for the in-frame cloning of the gene into the pET15b and pET22b expression vectors (see Chapter 3). Two alternative C-terminal reverse primers were also designed: outR containing the *P. falciparum* stop codon for N-terminal His tagged expression from pET15b and alt4R for silencing of the stop codon to enable C-terminal His tagged expression from pET22b.

Gene synthesis consisted of a two-step procedure in which each fragment was assembled from 8 pairs of oligonucleotides by overlap extension cycles (refer to figure 2.4) followed by PCR amplification of a 50-fold diluted assembly mixture with the appropriate primers and *Pfu* DNA polymerase (Promega, Wisconsin, USA). The 3' to 5' proofreading ability of the *Pfu* DNA polymerase increased the fidelity of the PCR process.

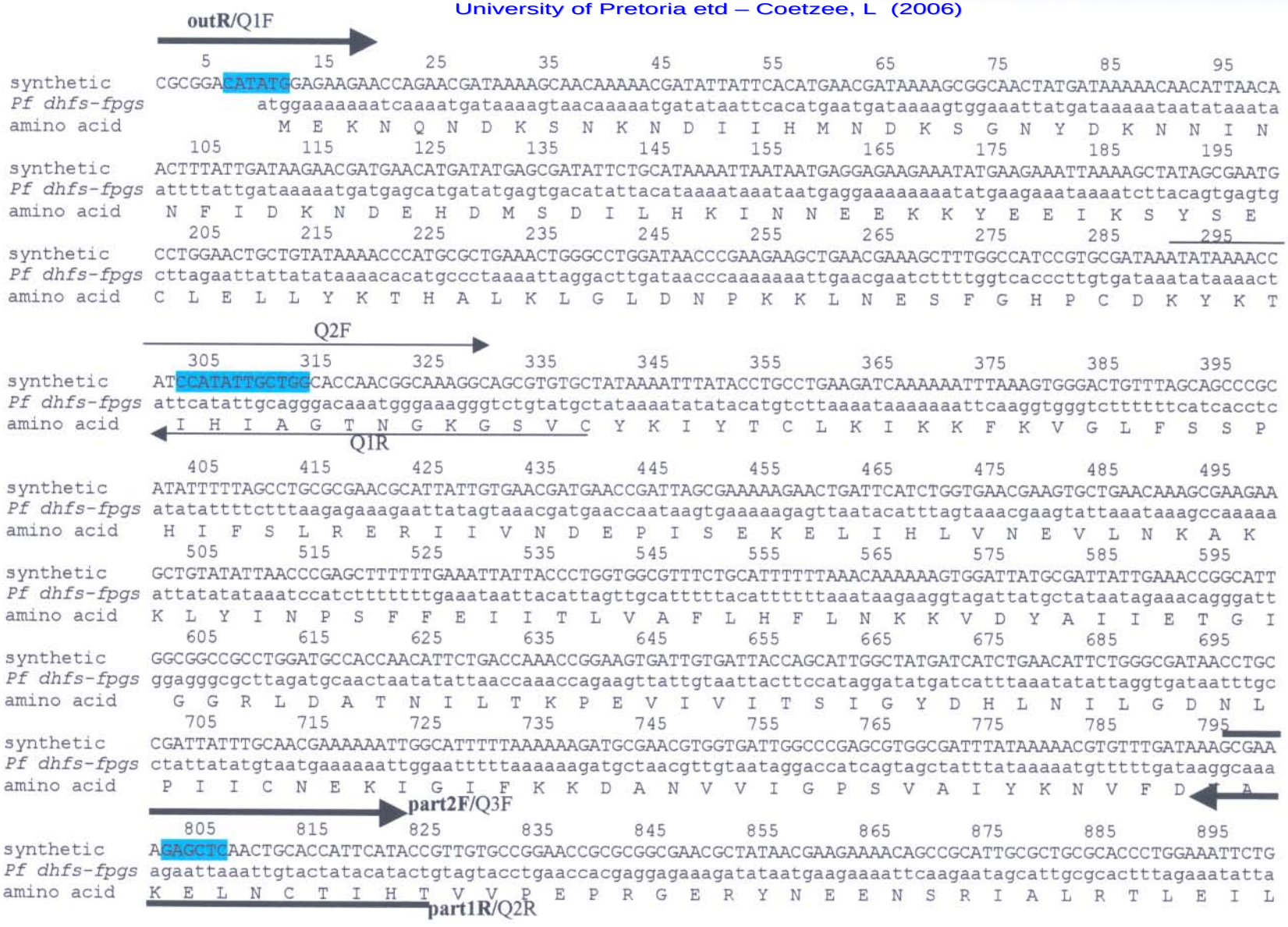

outR/Q1F
synthetic CGCGGACATATGGAGAAGAACCAGAACGATAAAAGCAACAAAAACGATATTATTCACATGAACGATAAAAGCGGCAACTATGATAAAAACAACATTAACA
Pf dhfs-fpgs atggaaaaaaatcaaaatgataaaagtaacaaaaatgatataattcacatgaatgataaaagtggaaattatgataaaaataatataaata
amino acid M E K N Q N D K S N K N D I I H M N D K S G N Y D K N N I N
synthetic ACTTTATTGATAAGAACGATGAACATGATATGAGCGATATTCTGCATAAAATTAATAATGAGGAGAAGAAATATGAAGAAATTAAAAGCTATAGCGAATG
Pf dhfs-fpgs attttattgataaaaatgatgagcatgatatgagtgacatattacataaaataaataatgaggaaaaaaaaatatgaagaaataaaatcttacagtgagtg
amino acid N F I D K N D E H D M S D I L H K I N N E E K K Y E E I K S Y S E
synthetic CCTGGAACTGCTGTATAAAACCCATGCGCTGAAACTGGGCCTGGATAACCCGAAGAAGCTGAACGAAAGCTTTGGCCATCCGTGCGATAAATATAAAACC
Pf dhfs-fpgs cttagaattattatataaaacacatgccctaaaattaggacttgataacccaaaaaaattgaacgaatcttttggtcacccttgtgataaatataaaact
amino acid C L E L L Y K T H A L K L G L D N P K K L N E S F G H P C D K Y K T
Q2F
synthetic ATCCATATTGCTGGCACCAACGGCAAAGGCAGCGTGTGCTATAAAATTTATACCTGCCTGAAGATCAAAAAATTTAAAGTGGGACTGTTTAGCAGCCCGC
Pf dhfs-fpgs attcatattgcagggacaaatgggaaagggtctgtatgctataaaatatatacatgtcttaaaataaaaaaattcaaggtgggtctttttttcatcacctc
amino acid I H I A G T N G K G S V C Y K I Y T C L K I K K F K V G L F S S P
Q1R
synthetic ATATTTTTAGCCTGCGCGAACGCATTATTGTGAACGATGAACCGATTAGCGAAAAAGAACTGATTCATCTGGTGAACGAAGTGCTGAACAAAGCGAAGAA
Pf dhfs-fpgs atatattttctttaagagaaagaattatagtaaacgatgaaccaataagtgaaaaagagttaatacatttagtaaacgaagtattaaataaagccaaaaa
amino acid H I F S L R E R I I V N D E P I S E K E L I H L V N E V L N K A K
synthetic GCTGTATATTAACCCGAGCTTTTTTGAAATTATTACCCTGGTGGCGTTTCTGCATTTTTTAAACAAAAAAGTGGATTATGCGATTATTGAAACCGGCATT
Pf dhfs-fpgs attatatataaatccatctttttttgaaataattacattagttgcatttttacatttttttaaataagaaggtagattatgctataatagaaacagggatt
amino acid K L Y I N P S F F E I I T L V A F L H F L N K K V D Y A I I E T G I
synthetic GGCGGCCGCCTGGATGCCACCAACATTCTGACCAAACCGGAAGTGATTGTGATTACCAGCATTGGCTATGATCATCTGAACATTCTGGGCGATAACCTGC
Pf dhfs-fpgs ggagggcgcttagatgcaactaatatattaaccaaaccagaagttattgtaattacttccataggatatgatcatttaaatatattaggtgataatttgc
amino acid G G R L D A T N I L T K P E V I V I T S I G Y D H L N I L G D N L
synthetic CGATTATTTGCAACGAAAAAATTGGCATTTTTAAAAAAGATGCGAACGTGGTGATTGGCCCGAGCGTGGCGATTTATAAAAACGTGTTTGATAAAGCGAA
Pf dhfs-fpgs ctattatatgtaatgaaaaaattggaatttttaaaaaaagatgctaacgttgtaataggaccatcagtagctatttataaaaatgtttttgataaggcaaa
amino acid P I I C N E K I G I F K K D A N V V I G P S V A I Y K N V F D A
part2F/Q3F
synthetic AGAGCTCAACTGCACCATTCATACCGTTGTGCCGGAACCGCGCGGCGAACGCTATAACGAAGAAAACAGCCGCATTGCGCTGCGCACCCTGGAAATTCTG
Pf dhfs-fpgs agaattaaattgtactatacatactgtagtacctgaaccacgaggagaaagatataatgaagaaaattcaagaatagcattgcgcactttagaaatatta
amino acid K E L N C T I H T V V P E P R G E R Y N E E N S R I A L R T L E I L
part1R/Q2R

Figure 2.3: Alignment of the synthetic and native (*Pf dhfs-fpgs*) *dhfs-fpgs* genes. The amino acid sequence is indicated below these sequences. Bold arrows and bold primer labels indicate the primers used for the generation of the half fragments (Figure 2.4). Thin arrows and primer labels indicate primers used for the generation of the four-quarter segments (Figure 2.5). Restriction enzyme sites incorporated by the primers are highlighted in blue. The stop codon (TAA) is underlined. A portion of the alternative reverse sequence to remove the stop codon is indicated by the sequence annotation '-stop'. Here AAA replaces the stop codon TAA, for readthrough and expression of a C-terminal tag.

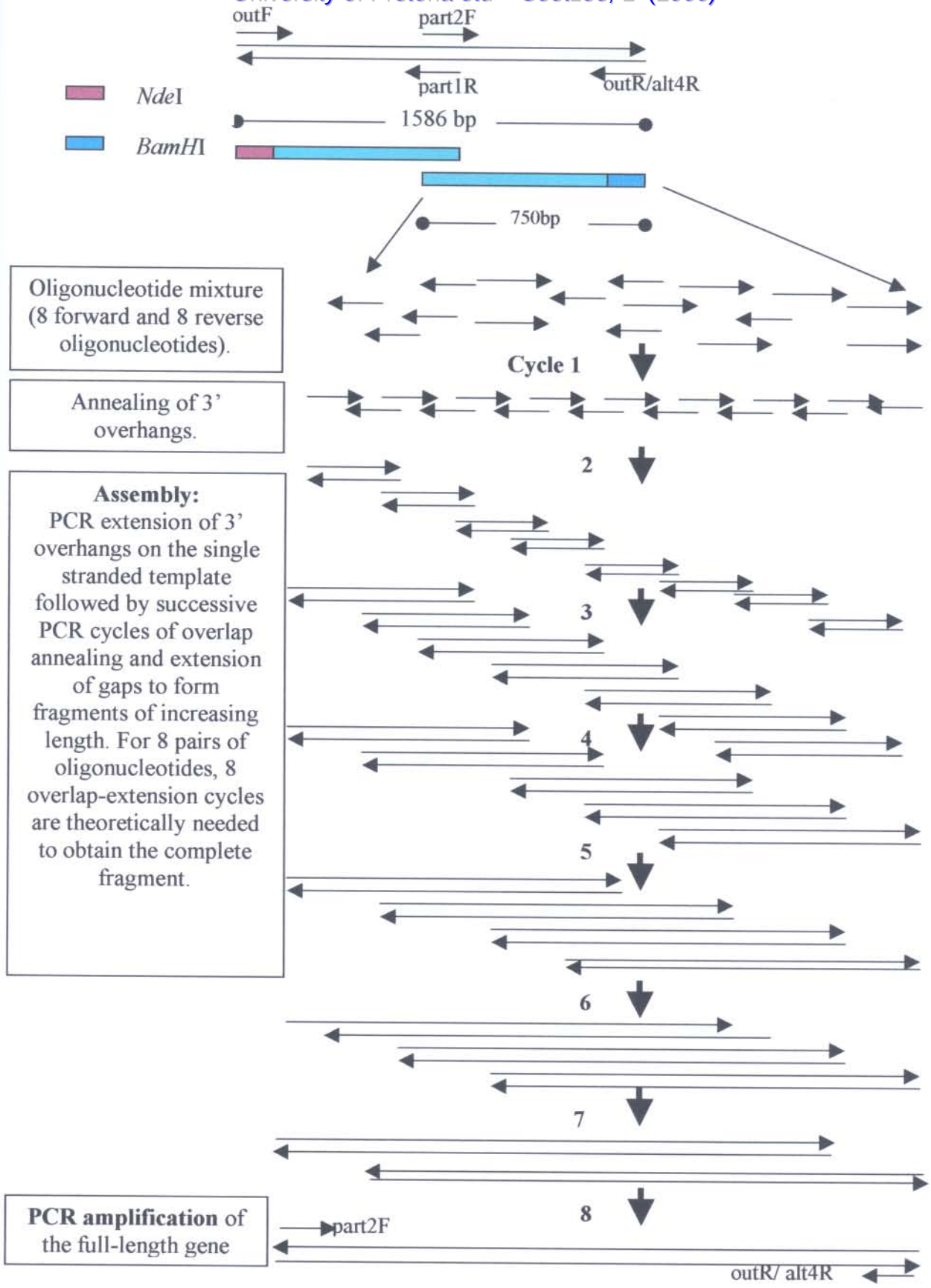


Figure 2.4: Overlap-extension PCR. The gene is divided into halves, each made of 8 forward and 8 reverse partially overlapping oligonucleotides. PCR assembly is used for the full-length construction by extension of single stranded gaps between overlaps, followed by amplification with the designated 5' and 3' end primers (part2F and outR).

In later studies the gene was divided into quarter segments by amplification of half gene segments with internal primers containing mutations to create unique restriction sites at these locations and to generate quarter fragments (Figure 2.5).

The PCR gene synthesis was optimised as follows: The Taguchi method was used to determine broad parameters for the Mg^{2+}, dNTP and *Pfu* DNA polymerase concentrations in the assembly step (Cobb and Clarkson, 1994). These parameters were then further refined with additional optimisation experiments. The final assembly PCR reactions consisted of 1pmole of each oligonucleotide, 4 mM $MgSO_4$ (Promega, Wisconsin, USA), 250 µM of each dNTP (Promega, Wisconsin, USA), 5 µl of an enzyme master mix and 1x *Pfu* DNA Polymerase buffer (Promega, Wisconsin, USA) in a final volume of 50 µl. The final amplification reactions consisted of 10 pmoles of each primer, a 50x dilution of the assembly PCR reaction, 2 mM $MgSO_4$, 250 µM of each dNTP, 1x *Pfu* DNA Polymerase buffer (Promega, Wisconsin, USA) and 5 µl of an enzyme master mix (2.5 U *Pfu* DNA polymerase for assembly or 1.25 U for amplification and 1x *Pfu* DNA Polymerase buffer) in a final volume of 50 µl. Master mixes consisting of all the reaction components except the enzyme were used throughout and added to each reaction in a hot start protocol. Assembly cycles were varied according to the number of oligonucleotides (see results section 2.3.3) and the number of amplification cycles did not exceed 30. The optimised cycling profile was as follows: initial denaturation at 94 °C for 30 sec, hot start addition of *Pfu* DNA polymerase at 80 °C for 20 sec, cycling at 94 °C for 30 sec, lowest overlap T_m (refer to table 2.1) for 30 sec and 72 °C for 2 min and a final extension step at 72 °C for 5 min. PCR products were stored at 4°C. All PCR's were performed in a Gene-Amp 9700 thermocycler (Perkin Elmer, Wellesley, USA) in 0.2 ml thin walled PCR tubes.

2.2.3 Agarose gel electrophoresis and purification

Digested plasmids and PCR products were separated on 1,5% (w/v) or 2% (w/v) agarose (Promega, Wisconsin, USA) gels, respectively. TAE buffer (0.04 M Tris-acetate, 1 mM EDTA) was used for electrophoresis. A 10 µg/ml ethidium bromide (EtBr) stock solution was included in the gel to a final concentration of 1.25 ng/ml. Samples were electrophoresed at 8 V/cm in a Minnie Submarine HE33 Agarose Gel unit (Hoeffer Scientific systems, San Francisco, USA). The bands were visualized on a Spectroline TC-312A UV transilluminator (Spectronics Corporation, Connecticut, USA) at 312 nm. Visualisation of PCR products used for cloning was done with crystal violet to prevent thymine dimers in the A+T rich areas of

the gene by ultraviolet illumination (Clark and Webb, 1955). A final concentration of 10 µg/ml crystal violet was used in TAE for the gels and running buffer. Bands were excised from the agarose gel and purified with the High Pure PCR product purification kit (Roche, Basel, Switzerland) according to the manufacturer's protocol. Gel slices were dissolved at 65°C in binding buffer (4 M guanidinium-HCl, 0.5 M potassium acetate, pH 4.2) and loaded onto a High Pure filter tube. DNA binds to the filter in the presence of chaotropic salts. The filter was washed two times with 500 µl and 200 µl wash buffer II (20 mM NaCl, 2 mM Tris HCl, pH 7.5), respectively, by means of centrifugation 15 000xg for 1 min. Each time the filtrate was discarded. The filter tube was transferred to a clean 1,5 ml tube and the plasmid DNA eluted in 50 µl elution buffer (1 mM Tris-HCl, pH 8.5) and collected by centrifugation at 16000xg for 1min at 22 °C. All buffers mentioned here, were supplied in the kit.

2.2.4. Cloning

2.2.4.1 The pGEM T Easy vector system

An A-tailing reaction was performed to add A overhangs to the blunt-ended PCR products generated by *Pfu* DNA polymerase to enable A/T cloning into the pGEM-T Easy vector according to the manufacturers protocol (Promega, Wisconsin, USA). For the vector map refer to Appendix B. Approximately 190 ng purified DNA was incubated for 30 minutes at 70°C in the presence of 0.2 mM dATPs and 5 U rTaq DNA polymerase (Takara, Shuzo, Japan) with 1x Taq buffer and 2 mM $MgCl_2$. The A-tailed product and pGEM T Easy vector were added together in a 3:1 molar ratio. 3 Weiss units T4 DNA ligase and 1x T4 DNA ligase buffer (provided in the pGEM T Easy vector kit) were added to the vector (50 ng) and the A-tailed product to a final volume of 10 µl. This reaction was incubated on ice overnight. The ligation mixture was cloned into electrocompetent or calcium chloride competent DH5α [genotype F^-Φ80 d*lacz* ΔM15 Δ(*lac*ZYA –argF) U169 *deo*R *rec*A1 *end*A1 *hsd*R17 (r_K^+, m_K^+) *pho*A *sup*E44 λ *thi*-1 *gyr*A96 *rel*A1] *E. coli* cells (see below).

Electrocompetent cells were prepared by the following method (Sambrook *et al.*, 1989): A single DH5α *E. coli* colony was inoculated into 15ml Luria-Bertani liquid medium (1% tryptone, 0,5% yeast extract, 1% NaCl, pH 7) and grown overnight with shaking at 300 rpm at 37°C. The overnight culture was diluted 1/100 into 2 flasks with 500ml LB liquid medium and grown at 30°C with shaking (300rpm) to an OD_{600nm} of 0.5. The cultures were poured into 2 pre-chilled 250 ml centrifuge bottles and incubated on ice for 20 min. All the subsequent steps

were done at 4°C. Cells were pelleted at 5000xg in a Beckman Avanti J-25 centrifuge with a fixed angle rotor at 4 °C for 10min. After removal of the supernatant, the cells were dissolved in 10 ml ice-cold water and washed with 250 ml ice-cold water. The suspension was centrifuged at 5000xg for 10 min at 4°C. This washing step was repeated twice. After the final centrifugation step the supernatant was immediately removed from the loose pellets. The pellets were resuspended in 10 ml ice-cold 10% glycerol each and incubated on ice for 30 minutes. Cells were subsequently pelleted (5000xg for 10 min at 4°C), the supernatant removed with vacuum suction and the pellets resuspended in 800 µl 10% ice cold glycerol. This was divided in 90 µl aliquots and frozen at -70°C.

Plasmids were co-precipitated with 10% tRNA (10mg/ml) for electroporation by the addition of 10% (v/v) sodium acetate of a 3M stock solution, pH 5 and 3 times the sample volume of absolute ethanol. This was incubated at -70°C for 1 hour and then centrifuged at 16 000xg for 20 min at 22°C. The plasmid pellet was washed with 3 volumes 70% ethanol, dried *in vacuo* and dissolved in 10 µl water, to which electrocompetent cells (90 µl) were added. This was then transferred to pre-chilled electroporation cuvettes and a pulse of 2000 V applied for 5 ms in the Multiporator system (Eppendorf, Germany). LB liquid medium (1 ml) was added directly after electroporation and the cells were grown for 1 hour at 30°C with shaking and plated out on LB solid media (1% noble agar, 1% tryptone, 0,5% yeast extract, 1% NaCl, pH 7) containing 0.2 mg/ml ampicillin.

Calcium chloride competent cells (J. Hyde, UMIST personal communication) were prepared as follows: 5ml of an overnight DH5α cell culture was diluted 10 times and grown with shaking to an OD_{600nm} of 0.4 at 30°C. The cells were pelleted by centrifugation for 5 min at 5000xg at 4°C and pellets were resuspended in 30 ml ice cold 0.1 M $CaCl_2$ at 4°C. This was then centrifuged at 5000xg for 5 min at 4°C. The supernatant was removed and the pellet resuspended in 25 ml ice cold 0.1 M $CaCl_2$. The centrifugation was repeated as before and the pellet then resuspended in 2.5 ml ice cold 0.1 M $CaCl_2$ .The competent cells were then incubated on ice for an hour followed by the addition of 15% ice cold glycerol and snap-freezing of 90 µl aliquots in liquid nitrogen.

$CaCl_2$ competent cells were thawed on ice, added to 50 ng of plasmid and incubated on ice for 30 min. A 42°C heat shock was applied for 1 min in a water bath, followed by incubation on

ice for 2 min. 500 µl of LB broth was added to the cells. The cells were then grown with shaking for 1 hour at 30°C and plated out on LB agar plates containing 0.2 mg/ml ampicillin.

2.2.4.2 The pMOS*Blue* vector system

Blunt ended cloning was also conducted with the pMOS*Blue* vector from Amersham Biosciences, Buckinghamshire, UK (refer to Appendix C). An optimal molar vector: insert ratio of 1:2.5 was used as recommended by the manufacturer. Before ligation, a phosphorylation reaction was performed on the purified inserts. The specified amount of insert was added to a reaction tube containing 5 mM dithiotreitol (DTT), a 1x phosphorylase kinase (pk) buffer and 1 µl phosphorylase kinase enzyme mix (provided in the pMOS*Blue* kit) in a final volume of 10µl. The reaction was incubated at 22°C for 5 min, followed by heat inactivation of the phosphorylase at 75°C for 10 min. The reaction was cooled on ice to prepare the mixture for ligation. Ligation mixtures consisted of 10 µl of the phosphorylation reaction, 50 ng of pMOS*Blue* vector and 4 Weiss units of T4 DNA ligase in a final volume of 12µl (one Weiss unit of ligase is equivalent to 200 cohesive end ligation units; Sambrook *et al.* 1989). The reaction was incubated at 22°C for 5 hours.

Competent MOS*Blue* cells (genotype: *end*A1 *hsd*R17 (r_{k12}-m_{k12}^{+}) *sup*E44 thi-1 *rec*A1 *gyr*A96 *rel*A1 lac [F' *pro*$A^{+}B^{+}lac^{q}Z$ ΔM15:Tn10(T_c^{R})]) supplied by the pMOS*Blue* blunt ended cloning kit were used for transformation of inserts ligated into the pMOS*Blue* vector (Amersham Biosciences, Buckinghamshire, UK). The transformation of MOS*Blue* competent cells was done by the addition of 5 ng of ligated plasmid (~2µl of the ligation reaction) to 20 µl competent cells. This was followed by incubation on ice for 30 min, heat shock at 42°C for 40 sec and an ice incubation for 2 min. 80 µl room temperature SOC (2% tryptone, 0.5% yeast extract, 10m M NaCl, 2.5 mM KCl, 1% glucose, pH 7) was added to this and the cells grown with shaking at 37°C for an hour, after which the cells were plated out on LB solid media containing a final concentration of 0.2 mg/ml ampicillin (Roche, Basel, Switzerland), 12.5 µg/ml tetracycline (Roche, Basel, Switzerland), 40 µg/ml 5-Bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) (Promega, Wisconsin, USA) and 0.2 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Ampicillin and tetracycline was used for primary selection and X-gal and IPTG for blue-white screening to identify the recombinant clones.

2.2.5 Plasmid isolation and identification of recombinant clones

2.2.5.1 STET–prep plasmid isolation

An adapted STET (Sucrose, TritonX-100, EDTA, Tris)-prep protocol was used as a quick and convenient method for plasmid isolation and screening of inserts (Quigley and Holmes, 1981). Clones were picked from solid media and grown overnight in liquid culture with shaking at 30°C. 1.5 ml of the liquid cell culture was centrifuged for 10 min, 3500xg at 22°C in an Eppendorf 5415R centrifuge with a 45° fixed angle rotor. The supernatant was partially removed and the pellet was resuspended in the remainder of supernatant. To this 300 µl of STET buffer (8% (w/v) sucrose, 5% (v/v) Triton X100, 50 mM EDTA, 50 mM Tris, pH 8) was added. Lysozyme, 25µl of a freshly prepared 4 mg/ml stock solution, (Roche, Basel, Switzerland) was added to digest the cell walls. The samples were placed in boiling water for 1 min and then centrifuged for 15 min at 16000xg (in an Eppendorf 5415R centrifuge) at 22 °C to pellet cell walls and debris. The genomic DNA is denatured by the high pH of the buffer, while the plasmid DNA reanneals fully. Pellets were discarded and 300 µl isopropanol added to the supernatant for precipitation of the DNA at 22°C. The tubes were centrifuged as above and the supernatant removed. The pellets containing plasmid DNA were air-dried and resuspended in 20 µl TE buffer (10 mM Tris, 1 mM EDTA, pH 8) containing a final concentration of 0.5 mg/ml RNAse (Roche, Basel, Switzerland).

Screening for recombinant clones was done by means of restriction enzyme digestion of STET-prep isolated plasmids. The *EcoR*I restriction enzyme site in the multiple cloning site of the pGEM-T Easy vector was used to screen for recombinant colonies, since it digests the plasmid at both sides of the insert (Appendix B). Double digestion was performed on recombinant MOS*Blue* cells, with *EcoR*I and *Hind*III on opposite sides of the multiple cloning site. Digestion reactions were set up each in a total volume of 20µl, consisting of >50 ng plasmid DNA, 10 U of the specific restriction enzyme and 1x of the appropriate restriction enzyme buffer (Promega, Wisconsin, USA). The tubes were incubated at 37°C for a minimum of 3 hours and the reaction stopped by loading the sample on an agarose gel.

PCR screening was used as an alternative method to identify positive clones (Amersham pMOS*Blue* instruction leaflet). Of the overnight DH5α or MOS*Blue* liquid cultures 100µl cells were centrifuged for 1 min at 16000xg at 4 °C and the pellet resuspended in 100 µl water. This was boiled for 5 min and centrifuged again (1 min at 16000xg 4 °C). The PCR reaction consisted of 10 µl of the supernatant as template, 5 pmoles of insert-specific primers (Tables 2.1 and 2.2), 1.25 U rTaq DNA polymerase (Promega, Wisconsin, USA), 1x Taq

PCR buffer, 2 mM $MgCl_2$ and 0,2 mM of each dNTP in a final volume of 20µl. The PCR profile consisted of 30 cycles at 94°C for 30 sec, 50°C for 30 sec and 72°C for 5 min. PCR products were analysed with agarose gel electrophoresis as described in section 2.2.3.

2.2.5.2. High Pure Plasmid kit plasmid isolation

The High Pure Plasmid isolation kit (Roche, Basel, Switzerland) was used to obtain pure plasmid samples. For plasmid isolation, 7 ml liquid culture was centrifuged for 30 sec, 5000xg at 22°C to pellet the cells. Cell pellets were resuspended in 250 µl suspension buffer (50 mM Tris-HCl, 10 mM EDTA, 10% w/v RNAse, pH 8) and then 250 µl lysis buffer (0.2 mM NaOH, 1% SDS) was added followed by a 5 min incubation at 25°C. After lysis, 350 µl chilled binding buffer (4 M guanidinium-HCl, 0.5M potassium acetate, pH 4.2) was added. The sample was mixed by inversion and incubated on ice for 5 min. Cell walls and large contaminants were pelleted by centrifugation, 16000xg for 10 min at 22°C. The supernatant containing plasmid DNA was passed through a High Pure Filter tube for binding of the plasmid DNA to the filter and purified further as described in section 2.2.3.

2.2.6 Sequencing

Sequencing PCR reactions were done on recombinant plasmids isolated with the High Pure Plasmid kit with T7 promoter and U19 primers (Promega, Wisconsin, USA) specific to the multiple cloning site of the pMOS*Blue* vector or T7 promoter and SP6 primers (Promega, Wisconsin, USA) complementary to the T7 and SP6 promoters flanking the multiple cloning site of the pGEM-T Easy vector (refer to appendix D). Gene-specific primers were also used to obtain internal sequences. Single-primer reactions were set up containing: 200-500 ng pure plasmid DNA as template, 3µl 5x sequencing buffer, 2µl Terminator Big-Dye Ready reaction mix version 3 (Perkin Elmer, Wellesley, USA) and double distilled deionised H_2O to a final volume of 20µl. The PCR was performed in a GeneAmp 9700 thermocycler (Perkin Elmer, Wellesley, USA) and the profile consisted of 25 cycles of: denaturation at 96 °C for 10 sec, annealing at 50 °C for 5 sec and extension at 60 °C for 4 min.

The DNA generated by PCR was isolated from all contaminants such as free dNTPs, protein and truncated DNA fragments by size exclusion filtration of the PCR mixture. Gel filtration media was prepared by addition of 600 µl G50 Superfine Sephadex (66 mg/ml) hydrated overnight in sterile water to the filter tube, followed by centrifugation at 750xg for 2 min. The PCR mixture was loaded onto the column, centrifuged again, collected in a clean tube, dried *in vacuo* and stored at -20 °C in the dark to protect the light sensitive dyes. Sequencing was performed on an ABI PRISM 377 Automatic Sequencer (Perkin Elmer, Wellesley, USA) or the ABI PRISM 3100 capillary sequencer. Another method of PCR product purification was the precipitation of DNA by the addition 3 volumes of ice-cold absolute ethanol followed by centrifugation at 16 000xg for 25-30 min at 4°C. After removal of the supernatant, the pellet was washed with 5 volumes of 70% ethanol followed by centrifugation at 16 000xg for 10 min at 4°C. The wash step was repeated with 4 volumes of 70% ethanol. After centrifugation at 16000xg for 10 min at 4°C, the DNA pellet was dried in the dark *in vacuo*.

The sequences were obtained as ABI files, which were either imported into the BioEdit version 5.0.6 software program (Hall 2001, North Carolina State University, Department of Microbiology, USA) for analysis and alignment with the required sequence, or into the Staden Package (Staden *et al.*, 2000; Medical Research Council Laboratory of Molecular Biology, Cambridge, UK) which was used for alignment of the different clone sequences with the required sequence. The electropherograms were compared with the sequence designated by the software.

2.2.7 Construction of the full-length gene

The full-length gene was assembled from the purified half-gene fragments at different concentrations (1 pmole, 0.1 pmole and 0.01 pmole). A single overlap-extension PCR procedure was used for assembly and amplification. PCR mixtures consisted of 250 mM of each dNTP, 1x of *Pfu* DNA polymerase buffer containing 2 mM $MgSO_4$ and 1.25 U *Pfu* DNA polymerase (Promega, Wisconsin, USA) in a final volume of 50µl. This PCR mixture was assembled for 5 cycles consisting of 94°C for 1 min, 50°C for 30 sec, and 68°C for 3 min. After these cycles, the temperature was lowered to 70°C for the addition of 5 pmoles of each outer primer, followed by 20 amplification cycles consisting of 94°C for 30 sec, 50°C for 30 sec, and 68°C for 2 min. A final extension step of 72°C for 5 min was used.

2.2.8 Gene repair strategies

2.2.8.1 Quarter gene segments

For ease of cloning and sequencing of the PCR generated fragments, the gene was subdivided in later experiments into two 300bp and two 500bp quarters, constructed from six and ten oligonucleotides, respectively. New primers were designed for the generation of the quarter segments (Table 2.2). These primers also incorporated unique restriction enzyme sites for restriction-ligation of the quarter fragments to assemble the full-length gene (Figure 2.5).

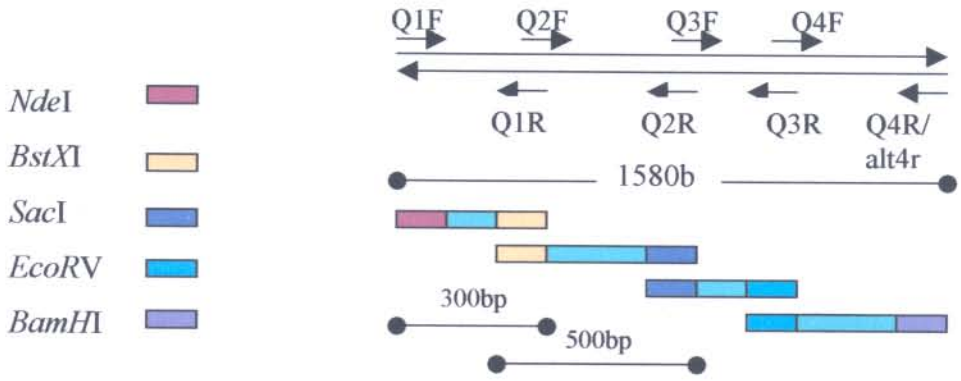


Figure 2.5: Subdivision of *dhfs-fpgs* into quarter segments with unique restriction enzyme sites in the overlaps. Primer positions are indicated by the arrows and primer names are directly above or below the arrows. Specific restriction enzyme sites used for digestion and ligation of the segments are indicated by the corresponding blocks.

Table 2.2: Forward (F) and reverse (R) primers used for the generation of the quarter fragments. T_ms were calculated with the formula: T_m=69.3 +0.41(%GC)-650/(primer length in nt). Restriction enzyme sites are indicated above the underlined sequences and cut sites are indicated with arrows.

Quarter fragment	Primer sequence (5' to 3')	length (nt)	T_m (°C)
Quarter 1 (300bp)	*NdeI* Q1F: CGCGGACATATGGAGAAGAACCAGA Q1R: *BstXI* GTTGGTGCCAGCAATATGGATGGTTTTATATTTATC	25 36	65 65
Quarter 2 (500bp)	*BstXI* Q2F: AAACCATCCATATTGCTGGCACCAAC *SacI* Q2R: CAGTTGAGCTCTTTCGCTTTATCAAACACG	26 30	62 61
Quarter 3 (300bp)	*SacI* Q3F: TGATAAAGCGAAAGAGCTCAACTGCA *EcoRV* Q3R: ATAGTTGATATCGGTGCACAGGCGATCAATC	26 31	62 62
Quarter 4 (500bp)	*EcoRV* Q4F:TGATCGCCTGTGCACCGATATCAACTATTTTCATA *BamHI* Q4R: CCGGATCCTTACACCAGGCTCGGTTCGT	35 28	67 71

2.2.8.2 Site directed mutagenesis of point mutated PCR products

Minor errors (single or triple nucleotide errors) in quarter 4 were corrected by means of site directed mutagenesis (protocol adapted from the Stratagene Quick change mutagenesis protocol). For each mutation a pair of oligonucleotides (>35bases) or megaprimers were designed to incorporate the correct sequence on both DNA strands (Table 2.3). The mutation was designed to be in the middle of each oligonucleotide (with at least 10 bases on each side). Less than 50 ng of plasmid was used as template, while the primer concentration was kept above 125 ng. The rest of the PCR mixture contained 2.5 U *Pfu* DNA polymerase (Promega, Wisconsin, USA), added in a hot start protocol, a 1x *Pfu* DNA polymerase buffer, 2 mM $MgSO_4$ and 250 M dNTPs (Promega, Wisconsin, USA) in a final volume of 50 µl. The PCR profile consisted of an initial denaturation of 95°C for 30 sec, followed by 12 to 18 cycles of 95°C for 30 sec, the primer T_m for 30 sec and 68°C for 6 min (2 minutes of extension time for each kilobase of plasmid). The number of cycles was determined by the size of the mutation, i.e. 12 cycles for a single base change and 18 for a triple base change. After PCR, the tubes were cooled on ice for 2 min and then 10 U *Dpn*I restriction enzyme was added to the PCR mixture to digest the wild type plasmid for 3 hours at 37°C. *Dpn*I only recognizes its site after bacterial methylation, leaving the mutated PCR product intact. The digested mixture was purified with the High Pure PCR purification kit (Roche, Basel, Switzerland) and the PCR product cloned into the pMOS*Blue* vector as described in sections 2.2.3 and 2.2.4.2 respectively. Colonies were screened from agar plates and sequenced to determine whether mutagenesis occurred as described in sections 2.2.6.

Table 2.3: Site-directed mutagenesis primers designed for error corrections in quarter 4.

Primer name	T_m (°C)	Sequence (Positions where errors occur are highlighted)
errors SDM4.1F SDM4.1R	 68 68	nt1279-1308:CTGAAC**C**AACGAAGAA~~~ATTAAAAA... 5′ GTGGAAATGCTGAAC~AACGAAGAA**GAA**ATTAAAAACGAAATTAAAG3′ 3′ CACCTTTACGACTTG~TTGCTTCTT**CTT**TAATTTTTGCTTTAATTTC5′
error SDM 4.2F SDM 4.2R	 65 65	nt1386-1410:GAAGAAGATGCGCTG**G**AAACTGTATA 5′ CGAAGAAGATGCGCTG~AAACTGTATAAACG 3′ 3′ GCTTCTTCTACGCGAC~TTTGACATATTTGC 5′

2.2.8.3 Restriction-ligation of correct fragments

The restriction enzyme *Hha*I (Promega, Wisconsin, USA) was used for the ligation of two correct fragments from different quarter 1 clones as part of the gene repair strategy. The fragments were cloned into the *EcoR*V site of the pMOS*Blue* vector. The restriction enzyme *Hha*I was chosen since it was the only unique restriction site between the correct parts of the two quarter 1 clones. The correct fragments were isolated from the pMOS*Blue* vector by means of *EcoR*I/*Hha*I double digestion for the 5' end correct fragment and *Sal*I/*Hha*I double digestion for the 3' end correct fragment (refer to appendix C). The vectors were first cut with *Sal*I or *EcoR*I restriction enzymes (Promega, Wisconsin, USA), respectively, and the pMOS*Blue* vector was also double digested with *Sal*I and *EcoR*I in a single step. Digestion mixtures contained ~500 ng plasmid, 3 U of each of the restriction enzymes and 1x buffer B (Promega, Wisconsin, USA) in a total volume of 20 µl. The digested bands were purified from agarose gels with the High Pure PCR product purification kit (Roche, Basel, Switzerland) and were then ligated in a 1:1:1 molar ratio as described in section 2.2.4.1.

2.2.8.4. Cassette mutagenesis PCR

Stretches of internal sequence errors were also corrected by means of cassette mutagenesis with short overlapping resynthesised gene fragments (~150 bp). The resynthesised oligonucleotides were also designed to enable future cassette mutagenesis through the use of unique restriction enzyme sites on opposite sides of important features of the DHFS-FPGS protein, such as the ATP-binding P-loop (GTNGKGS; nt 313-333) and an interdomain-stabilising Ω loop (VGLFSSPHIFSLRERI; nt 379-426). Clones containing the correct 5' and 3' areas flanking the errors were identified and areas containing the errors were resynthesised from shorter oligonucleotides (~50 bases each) using the optimised gene assembly method as described in section 2.2.2. Primers were designed for the isolation of the 5' and 3' flanking sequences of each quarter as well as the amplification of the internal assembled parts (Table 2.4 and Figure 2.6). The internal sequences were assembled from resynthesised oligonucleotides using the optimised PCR parameters (Figure 2.6). All the subsequent PCR products were purified and added in equimolar amounts to assemble the correct full-length quarter in a PCR reaction consisting of 250 µM of each dNTP, 4 mM $MgSO_4$, 1x *Pfu* DNA polymerase buffer and 2.5 U *Pfu* DNA polymerase in a final volume of 50µl. PCR cycling consisted of 5 cycles of 94°C for 30 sec, 65°C for 30 sec, 68°C for 2 min, addition of 5

pmoles of each primer at 70°C and 20 amplification cycles of 94°C for 30 sec, 65°C for 30 sec, 68°C for 2 min (Figure 2.6).

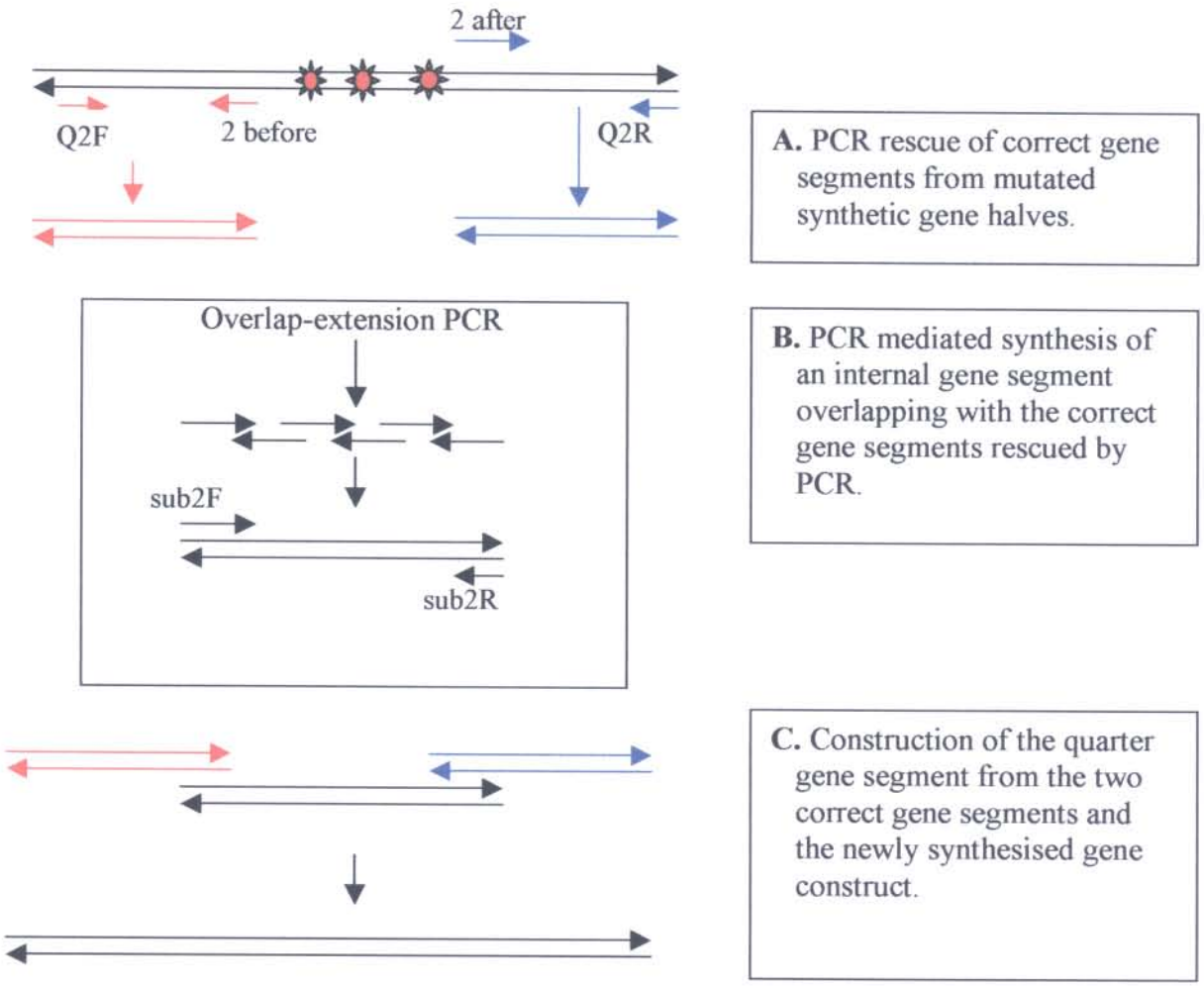


Figure 2.6: Cassette replacement of a mutated gene segment with a newly synthesised internal area of the quarter fragment. The 5' and 3' ends and the primers used for their isolation are indicated in red and blue respectively. Red stars indicate the errors.

Table 2.4: Forward (2f/3f) and reverse (2r/3r) oligonucleotides used for the resynthesis of the quarter 2 and 3 internal fragments. Before (5') and after (3') primers were used for isolation of the correct fragments with existing quarter primers (Table 2.2). Internal forward (sub) and reverse (sub) primers were used for amplification of the resynthesised PCR fragment. The oligonucleotide sequences are given in Appendix D.

Internal fragment 2			Internal fragment 3		
Oligonucleotide name	**Length (nt)**	**Overlap T_m (°C)**	**Oligonucleotide name**	**Length (nt)**	**Overlap T_m (°C)**
2f1	49	58 (2f1/2r1)	3f1	58	50 (3f1/3r1)
2r1	54	55 (2r1/2f2)	3r1	58	64 (3r1/3f2)
2f2	54	57 (2f2/2r2)	3f2	58	53
2r2	54	55 (2r2/2f3)	3r2	59	
2f3	54	61(2f3/2r3)			
2r3	54	55 (2r3/2f4)			
2f4	56	62 (2f4/2r4)			
2r4	54	55 (2r4/2f5)			
2f5	54	55 (2f5/2r5)			
2r5	44				
Primer name	**Sequence (5' to 3')**			**Length (nt)**	**T_m (°C)**
2before	AAACAGTCCCACTTTAAATTTTTTGA			26	55
sub2F	AATTTAAAGTGGGACTGTTT			20	49
sub2R	GTTGCAAATAATCGGCAGGT			20	55
2after	ACCTGCCGATTATTTGCAACGA			22	58
3before	AGAATTTCCAGGGTGCGCAGCGCAAT			26	66
sub3F	TGCGCTGCGCACCCTGGAAATTCT			24	66
sub3R	CACCGCCAGCGGATACTGCACGTT			24	68
3after	ATAACGTGCAGTATCCGCTGGCGGTG			26	68

2.2.9.5 Construction of the full-length gene from correct quarter fragments

The full-length gene was constructed by means of restriction digestion and ligation of the resultant overhangs of the first three correct quarter clones (refer to figure 2.5) followed by PCR assembly with the two alternative fourth quarters and PCR amplification. The 1100 bp fragment and the two alternate versions of the 500 bp fourth segments (Figure 2.5) with and without the stop codon which were cloned into the pMOS*Blue* vector were isolated and purified from their respective vectors by means of *Nde*I/*Eco*RV and *Eco*RV/*Bam*HI restriction digestion, respectively. The full-length gene was constructed by a single overlap-extension assembly of equimolar amounts of the purified 1100 bp fragment and each of the 500 bp fragments. The PCR reaction contained 0.2 pmole of each fragment, 1x *Pfu* DNA polymerase buffer (Promega, Wisconsin, USA) containing 2 mM $MgSO_4$, 1.25 U *Pfu* DNA polymerase and 250 µM of each dNTP in a final volume of 50µl. The PCR profile consisted of 10 cycles

of 94 °C for 30 sec, 50 °C for 30 sec and 72 °C for 2 min. After these cycles, the temperature was taken to 70°C for the addition of 5 pmoles each of the outer primers Q1F and Q4R/alt4R, followed by 10 cycles of 94 °C for 30 sec, 50 °C for 30 sec and 68°C for 2 min. The resulting isolated PCR product was then restriction digested, cloned into the pET15b and pET22b vectors, respectively, using the *Nde*I and *BamH*I sites, and sequenced as described in section 2.2.6.

2.3. Results

2.3.1 Oligonucleotide design

Comparison of the gene sequences for *P. falciparum dhfs-fpgs* and synthetic *dhfs-fpgs* indicated that the synthetic gene has a more equal distribution of the four nucleotides (42.5% G+C and 57.5% A+T) whereas the native gene has a codon bias to A+T (25.49% G+C and 74.51% A+T). This however does not change the amino acid composition, and *P. falciparum* DHFS-FPGS remains skewed towards an abundance of isoleucine (12.3%), lysine (10.2%) and leucine (9.6%). Refer to Appendix D for a list of the oligonucleotides and their complete sequences.

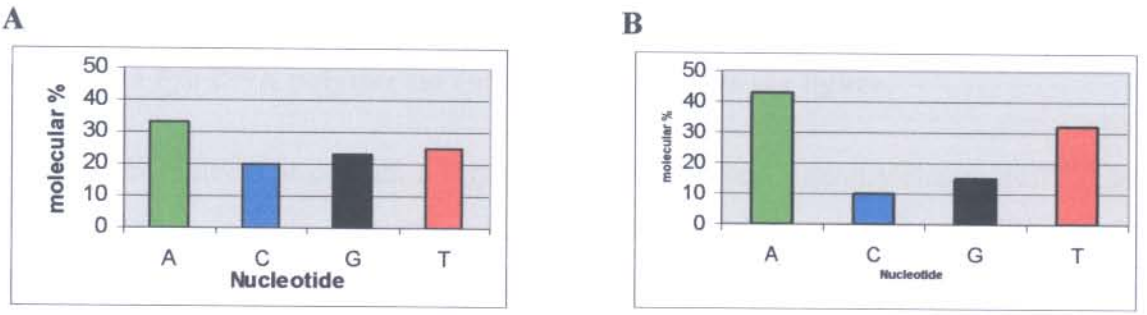


Figure 2.7: Graphs indicating the nucleotide composition of the synthetic *P. falciparum dhfs-fpgs* (A) gene compared to native *P. falciparum dhfs-fpgs* (B).

2.3.2 Optimisation of the overlap extension assembly step

The Taguchi method was used to determine the influence of the Mg^{2+}, dNTP and *Pfu* DNA polymerase concentrations on the efficiency of the PCR assembly step (Cobb and Clarkson, 1994). The overlap-extension PCR protocol was optimised for six oligonucleotides in terms of the Mg^{2+} and dNTP concentrations. The concentrations tested ranged from 150 to 250 µM dNTPs, 2 to 4 mM Mg^{2+} and 1.25 to 2 U *Pfu* DNA polymerase. For this optimisation, the assembly PCR was kept constant at 10 cycles. The amplification was conducted with the standard protocol (section 2.2.2). The results after 20 cycles of amplification are shown in Figure 2.8.

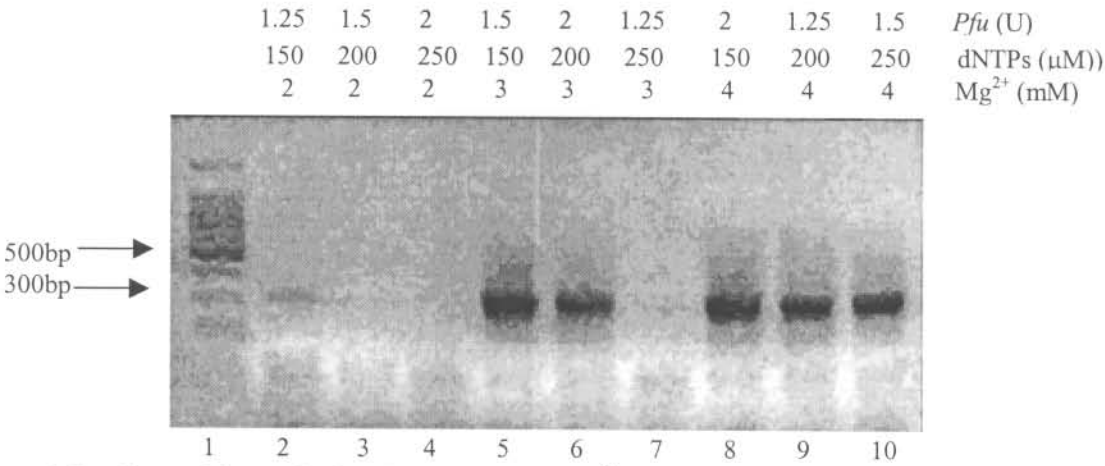


Figure 2.8: Taguchi optimisation of the Mg^{2+}, dNTP and *Pfu* DNA polymerase concentrations for the assembly of six oligonucleotides. Lane 1: 100 bp molecular marker. The arrows indicate the 500bp position and the expected band size (300bp). Mg^{2+} concentrations of each reaction are indicated in the lanes below the figure and dNTP and *Pfu* DNA polymerase concentrations above the figure.

It should be noted that overall, a higher magnesium concentration yields more PCR product. Where the magnesium concentration stays the same, an increased dNTP concentration decreases the PCR efficiency. An increase in the enzyme concentration is not directly linked to an increase in PCR product generation. At a higher magnesium concentration however, a higher enzyme concentration yields more PCR product (Figure 2.8, lane 8)

Using the information gained from the Taguchi experiment, refining of each of the above PCR parameters (the amount of dNTPs, Mg^{2+} and *Pfu* DNA polymerase) was optimised separately in the following three experiments: Firstly, the dNTP concentration (250 μM) was kept constant while using different Mg^{2+} concentrations (4 mM, 4.5 mM and 5 mM). This range was chosen since a high magnesium concentration was observed to give better yields of PCR products (Figure 2.8). 2.5 U *Pfu* DNA polymerase was used since the Taguchi experiment showed that a higher amount of enzyme at 4 mM Mg^{2+} yielded more intense PCR products (Figure 2.8, compare lanes 8, 9 and 10). Secondly, the dNTP concentration was varied (150 μM, 200 μM and 250 μM) at a constant amount of 4 mM Mg^{2+} and 2.5 U *Pfu* DNA polymerase. In the third experiment 1.5 U, 2 U and 2.5 U *Pfu* DNA polymerase was tested in PCR reactions containing 4 mM Mg^{2+} and 250 μM dNTPs. The assembly PCR cycles were kept constant at 10 cycles and amplification at 20 cycles. Refer to section 2.2.2 for the cycling profile.

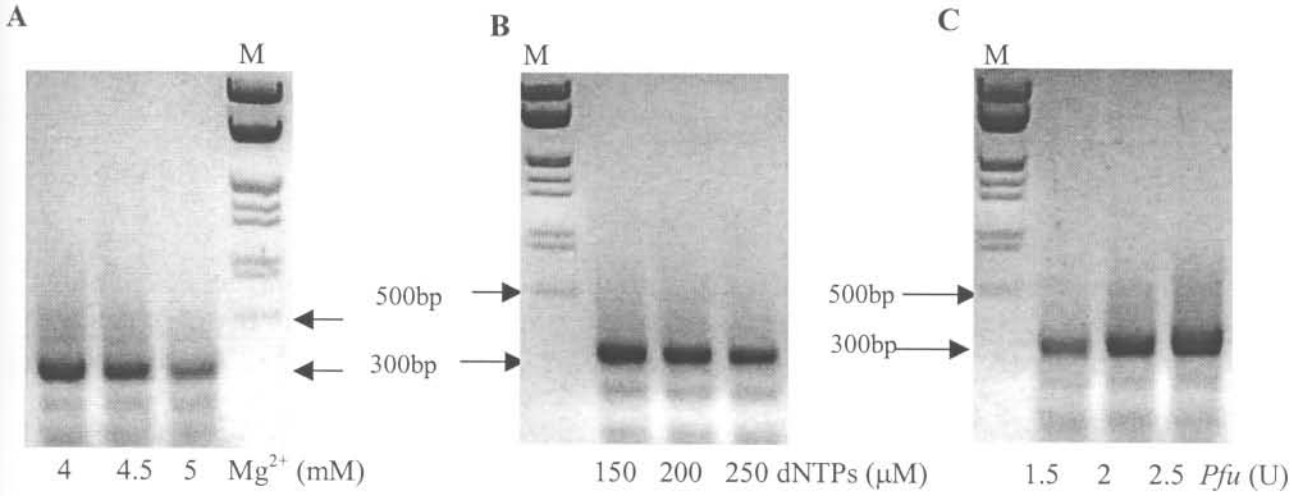


Figure 2.9: Refinement of the PCR parameters for the assembly PCR step. Optimization of the Mg^{2+} concentration (A), dNTP concentration (B) and the amount of *Pfu* DNA polymerase (C). M indicates the molecular marker, *EcoR*I and *Hind*III digested λ DNA. The arrows indicate the 500bp position and the expected band size (300bp).

From the DNA gels above it is evident that the efficiency decreases at 5mM Mg^{2+} (Figure 2.9 A). As the dNTP concentration increases, the PCR efficiency also decreases significantly (Figure 2.9 B). Increasing the amount of *Pfu* DNA polymerase however increases the PCR efficiency (Figure 2.9 C). From the above results the optimal conditions for PCR assembly was thus chosen as 4 mM Mg^{2+}, 250 mM of each dNTP, 1x *Pfu* DNA polymerase buffer and 2.5 U *Pfu* DNA polymerase.

2.3.3 Optimisation of the number of assembly cycles

The parameters in 2.3.2 were optimised for 1 pmole each of six 3' ends at a constant number of 10 assembly cycles. For the synthesis of fragments made up from different numbers of oligonucleotides, the optimal Mg^{2+} and dNTP concentrations as determined in section 2.3.2 were used and the molar amounts of the fragments adjusted to achieve the molar equivalent of six 3' ends. Thus 0.6 pmole of ten 3' ends and 1.6 pmoles of four 3' ends were taken to be equivalent to 1 pmole of six 3'ends (Table 2.5). This ensured a constant ratio of the concentration of 3' ends to the Mg^{2+} concentration under the optimised conditions. The number of assembly cycles required for fragment generation was also extrapolated from the optimised conditions. Fragments assembled from more oligonucleotides than used in the optimised conditions required proportionally more assembly cycles, which was adjusted as shown in Table 2.5.

Assembly of these fragments was thus performed with the optimised conditions of 4 mM Mg^{2+}, 250 mM dNTPs, 1x *Pfu* DNA polymerase buffer and 2.5 U *Pfu* DNA polymerase, as well as the specific amount of oligonucleotides as described above for 7 to 17 assembly cycles. A 50x, 10x and 5x dilution of the assembly mixtures was amplified for 20 cycles (section 2.2.2).

Table 2.5: The theoretical and actual number of cycles needed for complete template generation based on the number of oligonucleotides involved.

Amount of 3' ends	Theoretical amount of cycles	Number of cycles used	Amount of each oligonucleotide required (pmole)
Six 3'ends	3	10 (optimised in 2.3.2)	1
Ten 3'ends	5	17	0.6
Four 3'ends	2	7	1.6

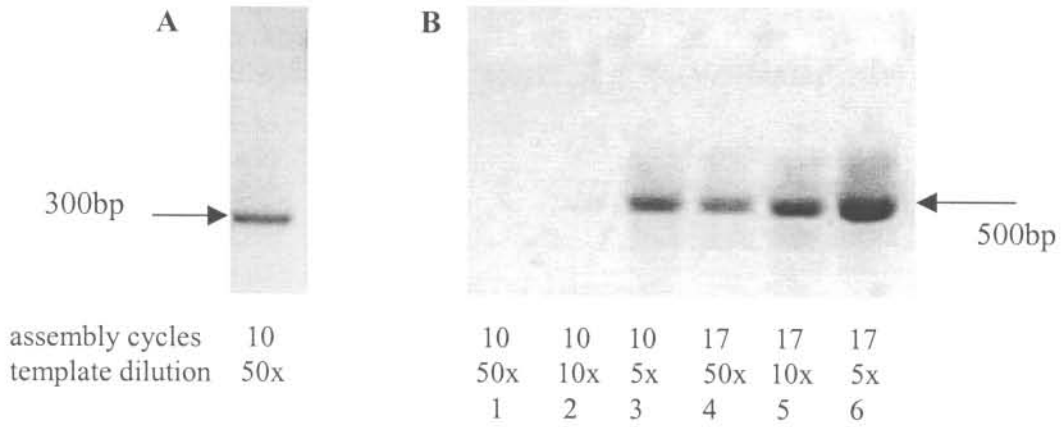


Figure 2.10: PCR products obtained for fragments assembled from 6 oligonucleotides (A) or 10 oligonucleotides (B). B lanes 1 to 6: PCR products obtained from 10 and 17 assembly cycles of 10 oligonucleotides at different dilutions of the assembly mixture (indicated below each lane). Arrows indicate the expected fragment sizes for 6 oligonucleotides (300bp) and ten oligonucleotides (500bp) after 20 amplification cycles.

It was observed that with more oligonucleotides, more assembly cycles were required to obtain PCR products at the same template dilution (compare lanes 1 and 4, lanes 2 and 5 and lanes 3 and 6, Figure 2.10 B). An assembly of 17 cycles for 10 oligonucleotides followed by a 50x dilution of the assembly reaction for amplification was taken to be preferable to an assembly of 10 cycles followed by a 5x dilution of the assembly reaction for amplification (compare lanes 3 and 4, Figure 2.10 B).

2.3.4 Construction of the full-length gene from half fragments

The full-length gene obtained from a single overlap-extension PCR of the gene halves was cloned and sequenced. In this instance 0.01 pmole of the halves gave the most specific PCR product after a total of 25 PCR cycles, when comparing the signal obtained to background 'noise' (Figure 2.11).

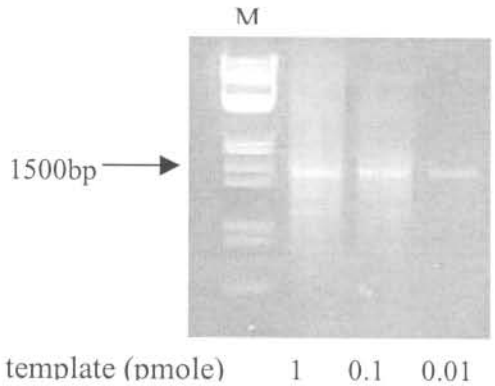


Figure 2.11: The full-length synthetic *dhfs-fpgs* gene obtained from overlap extension PCR of the two 750 bp half fragments. The dilution of each of the half fragments (template dilution) is indicated below the figure; M is the molecular marker (*EcoR*I and *Hind*III digested λ DNA).

2.3.5 Gene repair

Sequencing of gene halves and the full-length gene initially revealed errors in the form of single base insertions and deletions. The collective sequences of 65 clones were aligned in the 5' to 3' direction and the amount of errors per nucleotide position determined as a percentage of the number of oligonucleotides. This showed that the errors were mostly concentrated at the 5'-end overlaps between oligonucleotides. Scoring of the average errors of all the oligonucleotides as a percentage of the total number of errors at specific positions within the oligonucleotide, indicated an increase of the number of errors from the 3'- to the 5'-end (Figure 2.12). Insertions were localised to the 5' end whereas deletions were distributed in a more random fashion. The 3' end contained more deletions than insertions and the 5' end contained more insertions than deletions. Gene synthesis from newly synthesized shorter oligonucleotides (~50nt) revealed no errors.

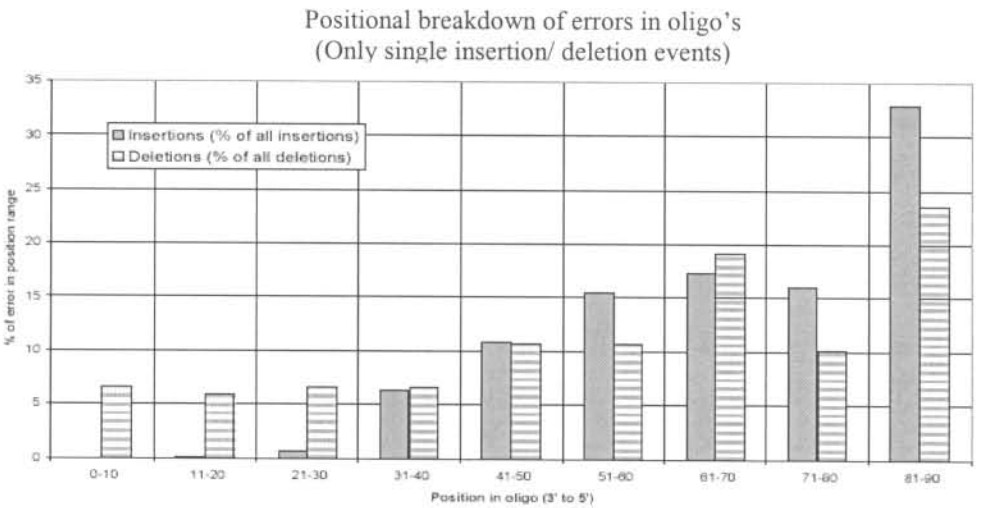


Figure 2.12: The direct relationship between the error rate and the proximity to the 5' end of each oligonucleotide. This graph indicates the relationship between the number of errors (expressed as a percentage of the total number of errors) to its 3' to 5' position in the oligonucleotide. The position ranges are divided in window sizes of 10 nucleotides. Solid bars indicate insertions and striped bars indicate deletions.

Quarters 2 and 3 were corrected by means of cassette mutagenesis of areas containing errors with short overlapping resynthesised gene fragments. The newly synthesised internal fragments used for cassette mutagenesis (Figure 2.13, lane 3) and the correct 5'-end and 3'-end sequences (Figure 2.13, lanes 2 and 4) were combined in an overlap extension PCR assembly of the quarter 2 and 3 fragments.

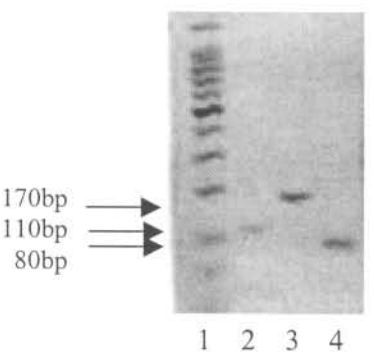


Figure 2.13: Construction of the correct quarter 2 by means of cassette mutagenesis. Lane 1: 100bp molecular marker, lanes 2 and 4: correct 5' and 3' fragments isolated by means of PCR, lane 3: fragment resynthesised from shorter oligonucleotides

After cloning and sequencing at least one correct quarter 2 and 3 clone of three clones sequenced was obtained (Figure 2.14, lanes 3 and 4 respectively).

Quarter 1 was corrected by means of restriction digestion and ligation of two different quarter 1 clones, each with the correct sequence on opposite sides of a unique restriction enzyme site, *Hha*I. This ligation product was cloned and sequenced and one of two clones sequenced contained the correct quarter 1 sequence (Figure 2.14, lane 2)

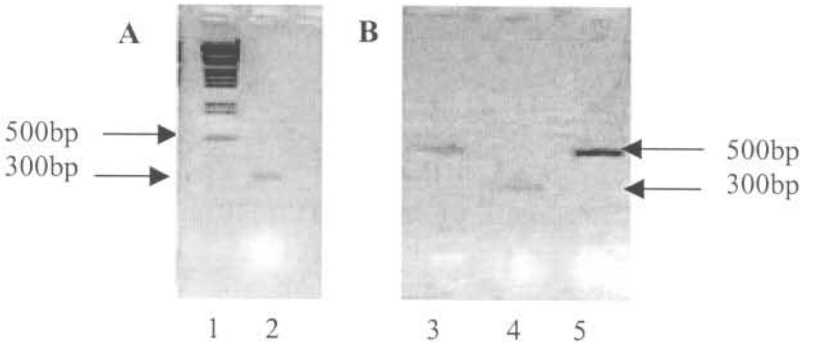


Figure 2.14: Corrected quarter fragments. Lane 1: molecular marker (*EcoR*I and *Hind*III digested λ DNA), lane 2A: quarter 1(300bp) isolated after *Nde*I/*BstX*I digestion, lane3B: quarter 2 (500bp) isolated after *BstX*I/*Sac*I digestion, lane 4: quarter 3 (300bp) isolated after *Sac*I/*EcoR*V digestion and lane 5: quarter 4 (500bp) isolated after *EcoR*V/*BamH*I digestion of the respective plasmids. The arrows indicate the expected fragment sizes.

Quarter 4 was repaired by means of two site-directed mutagenesis steps. Since two of the errors, a single base insertion and triple base deletion were within 10 bases of each other, a single primer pair was sufficient to correct both errors (Table 2.3, SMD4.1F and –R). After cloning and verification of the correct sequence, the second site-directed mutagenesis step was performed on a single base insertion, ~100 bases upstream of the other errors. The corrected quarter 4 was cloned and sequenced (Figure 2.14, lane 5). At least one of every two clones sequenced revealed the correct sequence after site directed mutagenesis.

2.3.6 Construction of the full-length gene from correct quarter fragments.

Once the clones containing the correct quarter sequences were identified, the quarter fragments were isolated from the respective plasmids and the three-quarter gene assembled by means of restriction-ligation, using the engineered restriction sites. An 1100bp fragment was obtained after restriction-ligation and cloning of the first three quarters (Figure 2.15, lane 2).

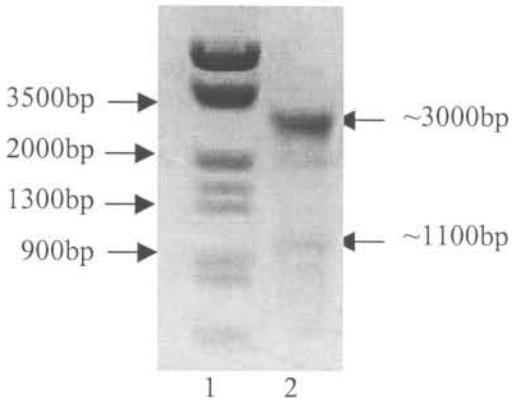


Figure 2.15: Isolation of the 1100bp fragment consisting of the 1st three fragments by means of restriction digestion. Lane 1: molecular marker (*EcoR*I and *Hind*III digested λ DNA), lane 2: *Nde*I/*EcoR*V digested pMOS*Blue* plasmid containing the 1100bp fragment.

The 1100bp fragment was used in a single overlap-extension PCR with the two alternate fourth quarters (with and without the stop codon) for the generation of two versions of the gene to enable N- and C-terminal His_6 tagged expression, respectively (Figure 2.16 lanes 1 and 2).

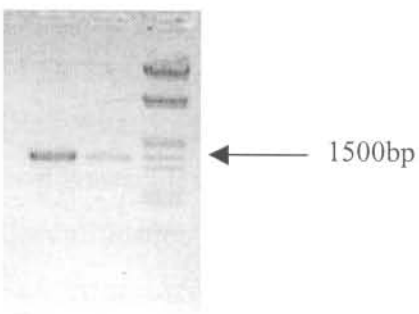


Figure 2.16: Full-length *dhfs-fpgs* constructs obtained after overlap-extension of the 1100bp fragment and alternative 500bp fragments. Lane 1: full-length gene with the stop codon, lane 2: full-length gene without the stop codon, lane 3: molecular marker. The arrow indicates the size of the full-length gene.

2.4 Discussion

2.4.1 Overlap-extension PCR as a gene synthesis method

The value of synthetic genes for the successful expression of proteins, especially those of complicated systems such as the A+T rich malaria genome, is clear through past successes (Prapunwattana *et al.*, 1996; Zhang *et al.*, 2002). Gene synthesis can be achieved through different methods such as the restriction-ligation of cloned double-stranded oligonucleotides or the PCR-based assembly of overlapping oligonucleotides. The first approach involves a large number of complicated cloning and sub-cloning steps and is limited by the length of the gene (Prapunwattana *et al.*, 1996). The longer the gene, the more unique restriction sites, cloning vectors and cloning steps are needed, since extremely long oligonucleotides (>100nt) can not be chemically synthesised without compromising the fidelity of the sequence. The second approach is a simpler method and not restricted by the length of the gene. The assembly of a gene from 125 pmoles of completely overlapping oligonucleotides using a total of 50 PCR cycles was described previously (Carpenter *et al.*, 1999). To cut the costs of this approach, we decided to use only partially overlapping oligonucleotides and to apply the PCR process more efficiently for extension of the gaps between overlapping oligonucleotides. For a 1500bp gene, this reduced the number of nucleotides required from 3000nt (for complete overlapping oligonucleotides) to ~2500nt (for partially overlapping nucleotides), reducing the cost with 17%. A further reduction in the cost was achieved by using less oligonucleotides, 1 pmole of each for only 10 assembly and 20 amplification cycles. The reduction in the number of PCR cycles reduced the possibility of error incorporation during PCR. *Pfu* DNA polymerase with 3' to 5' proofreading activity was also used to increase the fidelity of the PCR protocol.

The 1586 bp *dhfs-fpgs* gene was subdivided into halves, which were constructed by means of a two-step PCR protocol. In the first step each half was assembled from 16 oligonucleotides by means of overlap-extension PCR. In the second step the assembly mixtures were diluted and the halves amplified with specific primers. The full-length gene was then obtained by a single overlap-extension PCR step of the two halves. The 5' and 3' outer primers (Table 2.1) introduced unique restriction enzyme sites into the gene for directional cloning into expression vectors and to enable cassette mutagenesis. The 3' end primers also incorporated alternative 3' gene ends for expression from N- or C-terminal His_6 tagged vectors.

2.4.2 Optimisation of the assembly PCR parameters

The first step of the PCR protocol (assembly of oligonucleotides) was optimised in terms of the Mg^{2+}, dNTP and *Pfu* DNA polymerase concentrations. The Taguchi method indicated that an increased Mg^{2+} concentration promoted PCR efficiency (Figure 2.8). In multiplex PCR reactions, where a large number of different primers anneal to a single template, an increased Mg^{2+} concentration, sometimes up to 8mM, is required for efficient primer annealing and extension (Zangenberg *et al.*, 1999). The PCR assembly is similar to a multiplex reaction due to the large number of oligonucleotide overlaps that have to anneal to each other. Variations in the Mg^{2+} concentration, however, have an effect on other PCR components, such as the dNTP concentration and enzyme activity. Mg^{2+} chelates dNTPs, thus the amount of dNTPs have to be increased as the Mg^{2+} concentration increases. Too high Mg^{2+} concentrations could lead to mispriming due to increased non-specific annealing and also lower the available dNTP concentration. Low dNTP concentrations could result in the misincorporation of bases due to a decreased availability of dNTPs whereas high dNTP concentrations could activate the 3' to 5' exonuclease activity of the *Pfu* DNA polymerase resulting in the degradation of the free 3' ends of the oligonucleotides (*Pfu* DNA polymerase Technical data, Promega). To increase overall PCR efficiency a higher amount of enzyme (2.5 U instead of the recommended amount of 1.25U as indicated by the manufacturer) was used. The refined optimisation indicated the optimal concentrations as 4 mM Mg^{2+}, 250 µM dNTPs and 2.5U *Pfu* DNA polymerase. For all optimisation experiments, the same amplification protocol (section 2.2.2) was used. The PCR products obtained were only visible after the amplification step, since the concentration of assembly products was too low.

2.4.3 Optimisation of the number of assembly cycles

The above parameters (section 2.4.2) were optimised for the assembly of 1 pmole each of six oligonucleotides for ten assembly cycles. These conditions were applied to other fragments consisting of different numbers of oligonucleotides by varying the number of assembly cycles. Based on the number of 3' ends involved in the annealing process the number oligonucleotides used for assembly was thus adjusted in such a manner that the absolute concentration of 3' ends to the Mg^{2+} and dNTP concentration remained the same. For a larger number of oligonucleotides, a proportionally smaller pmole amount was used and vice versa, i.e. if six oligonucleotides were used at a concentration of 1 pmole each, then the mixture consisted of 6 pmoles 3' ends. Similarly, for four oligonucleotides a concentration of 1.6

pmoles of each would give a total of 6 pmoles of 3' ends and for ten oligonucleotides a concentration of 0.6 pmole each would give 6 pmoles of 3'ends. The number of assembly cycles was also varied in proportion to the optimised number of 10 cycles. For ten oligonucleotides the number of assembly cycles was increased from 10 to 17 (Table 2.5). PCR products were obtained after only 20 amplification cycles of a 50x dilution of the assembly mixture. A 50x dilution of the assembly mixture was preferable to a 10x or 5x dilution since the dilution was used to reduce the interference of incompletely assembled products in the assembly mixture during the amplification cycles. The dilution of the assembly mixture together with the addition of 5' and 3'end- specific primers thus drive the generation of the complete assembled product, underlining the importance of the use of a two-step PCR protocol.

2.4.4 Construction of the full-length gene

The full-length gene (~1580 bp) was successfully constructed from the single overlap extension PCR of purified gene halves by means of a one-step PCR protocol. Different concentrations of the halves (1 pmole, 0.1 pmole and 0.01 pmole) were used to determine the optimal template to primer ratio (theoretically the primers have to be in a 10^4 molar excess; Sambrook *et al.*, 1989). 0.01 pmole of each half fragment was sufficient to obtain PCR products within 20 amplification cycles. The lower template concentration was thus optimal to obtain the most specific PCR products. Since there was only a single overlap between the halves, there was thus no interference from incompletely assembled PCR products and no dilution of the assembly mixture was needed before amplification to increase the specificity of the reaction, in contrast to the two-step procedure used for the assembly of >3 oligonucleotides. Furthermore, since this protocol involved the annealing of one overlap, higher Mg^{2+} and enzyme concentrations (used for the two-step protocol) were unnecessary and the manufacturer's recommended amount of 2mM Mg^{2+} and 1.25 U *Pfu* DNA polymerase was used. In theory the single overlap extension occurs during one PCR cycle, but an initial 5 assembly cycles was used for the extension of the single overlap to generate a sufficient amount of the full-length gene required for specific amplification. This was followed by the addition of the outer primers and amplification of the full-length template.

2.4.5 Gene repair

Initial sequencing revealed mostly single base insertions or deletions over the whole length of the gene sequence. PCR generated errors are normally due to misincorporations caused by the decreased fidelity of the polymerase or multiple-base deletions due to secondary structures of the template. Single base insertions and -deletions observed by the sequence alignment of 65 clones revealed that these errors were concentrated at the 5' overlap areas between oligonucleotides. There was also an observed shift from mainly deletions at the 3' end of each oligonucleotide to mainly insertions at the 5' end (Figure 2.12). Taken together all these results indicated that the PCR process was not responsible for the errors. Since the chemical synthesis of oligonucleotides occur in the 3' to 5' direction and on average, the highest number of errors occurred at the 5' end of each oligonucleotide, the errors were attributed to poor quality control by the manufacturer during the chemical synthesis of the oligonucleotides. During phosphoramidite oligonucleotide synthesis the growing oligonucleotide chain is linked through the 3' OH group of the deoxyribose sugar of the first deoxyribonucleotide to a solid matrix, controlled pore glass beads (CPG). The elongation of the oligonucleotide chain by a single oligonucleotide, involves intricate chemical steps. The 4,4'- dimethoxytrityl (DMT) 5' OH protecting group of the deoxyribose sugar of the previous dexoyribonucleotide has to be removed just prior to coupling of the next deoxyribonucleotide. The deprotected 5' end is then coupled to the 3' phosphoramidite derivative of the next deoxynucleoside with the coupling agent, tetrazole. Under prolonged coupling conditions (eg the synthesis of long oligonucleotides) the acidic Tetrazol, used for deprotection during synthesis could remove the protective trityl group from the deoxyribonucleotide, resulting in double-coupling events, observed here as insertions (Fu *et al.*, 2002). Oligonucleotides are normally only guaranteed up to 60 bases in length due to this reason. After coupling, unreacted 5' ends are blocked by acetylation and the phosphite triester resulting from the coupling step is oxidised to a phosphotriester. When synthesis is complete, blocking groups on the bases are removed and the oligonucleotide chain cleaved from the solid support. This can then be purified by HPLC to remove truncated products, but the HPLC is not sensitive enough to detect single base insertions or deletions (Voet and Voet, 1995).

Error-free quarter 2 and 3 sequences could be obtained from the cassette mutagenesis of internal fragments with newly synthesised, shorter (~50nt long) oligonucleotides from as little as three clones. PCR generated errors were few (less than one in 500bp) and were directly proportional to the number of oligonucleotides involved in the assembly. The resynthesised

fragments furthermore include important structural features of the enzyme, such as the P-loop (GTNGKGS; nt 313-333), which is involved in the binding of ATP and the Ω loop (VGLFSSPHIFSLRERI; nt.379-426) that is responsible for stabilisation interactions between domains (Sun *et al.*, 1998). Future structure-function studies are thus possible by cassette mutagenesis of these internal fragments.

Other minor errors were corrected by restriction-ligation or site-directed mutagenesis. The correct first quarter was obtained from the restriction-ligation of two clones containing correct sequences on opposite sides of the unique *Hha*I restriction enzyme site. The correct fourth quarter was obtained after two successive site-directed mutagenesis experiments.

2.4.6 Construction of the full-length gene from quarter fragments

For ease of cloning and sequencing, quarter fragments were used during the repair of the gene. Correct quarter clones were identified by sequencing after the corrections were made as stated above in section 2.4.5. These first three quarter fragments were ligated, using the engineered unique restriction enzyme sites into an 1100bp fragment. The use of unique restriction enzyme sites prevented digestion of the vector or gene at any area besides the overlaps. The two different fourth quarters (with and without the stop codon) were then assembled with the 1100bp fragment into the full-length gene, which was cloned and sequenced. The correct sequences for both *dhfs-fpgs* constructs were obtained.

This chapter described the simple, quick, efficient, low cost synthesis of a codon-adapted gene by means of a two-step PCR method. Costs were cut through the use of partially overlapping oligonucleotides instead of complete overlapping oligonucleotides as well as >100 times less starting material, when compared to previous methods. Provided high quality oligonucleotides are used, correct synthetic genes of virtually any length can be obtained by this method. The oligonucleotides can also be used in future cassette mutagenesis experiments for structure-function studies. The purpose of the synthetic gene described here, *P. falciparum dhfs-fpgs* is to increase recombinant expression in *E. coli*. The synthetic gene was therefore successfully adapted to *E. coli* codon preferences and has a 17% reduction in A+T content. Synthesis of two alternate versions of the gene also enables N- or C-terminal His_6-tagged expression of the protein. The next chapter will focus on the recombinant expression of various tagged versions of this synthetic gene in *E. coli*.

Chapter 3

Expression of synthetic *P. falciparum* dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS)

3.1 Introduction

The two activities DHFS (E.C. 6.3.2.12) and FPGS (E.C. 6.3.2.17) play vital roles (Figure 1.4) in the synthesis and recycling of tetrahydrofolate (vitamin B9), which is used in the synthesis of purines, thymidylate, methionine, formylmethionine tRNA, pantothenate and the conversion of serine to glycine (Stokstad and Koch, 1967). Mammals do not have a DHFS enzyme, but other eukaryotes such as yeast, fungi and plants have separate DHFS and FPGS proteins. The *dhfs-fpgs* identified in *P. falciparum* is the first eukaryotic gene to encode a single bifunctional protein containing both enzyme activities. Other protozoan parasites such as *Leishmania* only have an FPGS enzyme (Fadili *et al.*, 2002). In prokaryotes such as *E. coli* and *Corynebacterium* species the two activities are contained within a single bifunctional protein (Bognar *et al.*, 1985; Shane, 1980). In *E. coli* the DHFS and FPGS functions are furthermore co-linear (Keshavjee *et al.*, 1991). This means that the two functions are not distinguishable at DNA or amino acid level, i.e. there is no clear allocation of activity to certain protein domains in contrast to other bifunctional proteins such as *P. falciparum* DHFR-TS that is divided in an N-terminal domain (nt 1-684) with DHFR activity and a C-terminal domain (nt 969-1824) with TS activity (Bzik *et al.*, 1987). As a result of the co-linearity, a single point mutation in *E. coli* is sufficient to abolish both DHFS and FPGS activities (Keshavjee *et al.*, 1991). It is not yet clear whether the DHFS and FPGS activities are co-linear in *P. falciparum*. If the *P. falciparum* DHFS and FPGS activities were also co-linear it would increase the effectivity of drugs by eliminating both enzyme activities through a single point of inhibition. Considering that the DHFS and FPGS enzymes each function in a different route of folate metabolism, simultaneous inhibition of both activities would be synergistic such as that observed by the combination of sulfadoxine-pyrimethamine, which inhibits the dihydropteroate synthase component of DHPS-PPPK and the dihydrofolate reductase component of DHFR-TS.

The following reactions are catalysed by DHFS and FPGS (refer to figure 1.4):

DHFS

1. dihydropteroate + L-glutamate + MgATP → dihydrofolate + MgADP + P_i

FPGS

2. tetrahydrofolate + L-glutamate + MgATP → tetrahydropteroyldiglutamate + MgADP + P_i

3. 5,10 methylene tetrahydropteroyl $(glutamate)_n$ → 5,10 methylene tetrahydropteroyl $(glutamate)_{n+1}$

Reaction 1 is the addition of the first glutamate residue through an amide linkage to the carboxyl end of the p-amino benzoic acid (PABA) group of dihydropteroate, catalysed by DHFS. Reactions 2 and 3 are catalysed by FPGS. In all the reactions addition takes place at the γ- carboxyl group of glutamate in a K^+- and ATP-dependent manner. The phosphate group of ATP activates the carboxyl group of the dihydropteroate substrate (DHFS) or the glutamate group of folic acid (FPGS) through the formation of an acyl phosphate intermediate. The polyglutamate tail length (n) varies between 1 and 10 and is usually shorter for prokaryotes than for eukaryotes (Bognar *et al.*, 1985). In *E. coli* the polyglutamate chain length is three (Osborne *et al.*, 1993), in *Leishmania* five (Santi *et al.*, 1987) and in *P. falciparum* it was observed that 5-methyl-tetrahydropteroyl pentaglutamate was the predominant cellular form of folates (Krungkrai *et al.*, 1989). Mammals have even larger chain lengths of up to nine glutamate residues (Osborne *et al*, 1993). Polyglutamylation is essential since it is required for the increased cellular retention of folates and affinity for folate-utilising enzymes as well as the accumulation of folates in the mitochondria where glycine synthesis takes place (Moran, 1999).

Due to the absence of DHFS in humans and expected differences between FPGS activities of humans and *P. falciparum* as shown above the possibility of designing selective drugs that are not detrimental to the human host is increased. This can, however, only be confirmed with functional assays for which significant amounts of active enzyme are required. The only crystal structure currently available for homology modelling is that of *Lactobacillus casei* FPGS (Sun *et al.*, 1998). Sufficient amounts of correctly folded, soluble DHFS-FPGS are thus needed to determine the kinetic properties of the protein and determination of the 3-dimensional structure of the enzyme.

Chapter 2 described the synthesis of a codon-optimised *dhfs-fpgs* for increased recombinant expression in *E. coli*. This chapter focuses on the optimisation of expression of the synthetic *P. falciparum dhfs-fpgs* gene and its functional verification to obtain sufficient amounts of soluble protein for future enzyme activity studies and structure determination.

The *E. coli* codon-adapted *dhfs-fpgs* obtained by PCR-mediated gene synthesis as described in Chapter 2 was cloned into several expression vectors and expressed in different *E. coli* cell lines to select the best system for expression of soluble protein. Gene complementation assays were performed to verify the functional activities of the gene products and to compare the complementation efficiency and thus activity of different tagged constructs to that of a tagless construct. Preliminary purification studies were also performed as a basis for future protein isolation strategies.

3.2 Materials and methods

3.2.1 Constructs, vectors and cell lines

Two systems were used for protein expression (Table 3.1): the IPTG inducible *T7lacUV5* promoter of the pET vector (Novagen, EMD Biosciences, Germany) and the anhydrotetracycline inducible *tetA* resistance gene promoter of the pASK vector (IBA, Germany). In both of these systems protein expression from the plasmid is tightly regulated to prevent the overproduction of protein, which might be toxic to cell growth.

Table 3.1: Vector systems used for the recombinant expression of synthetic *P. falciparum dhfs-fpgs*. Refer to Appendix E for the vector maps.

Vector	Antibiotic resistance	Tag	Induction	Proteolytic cleavage site
pET15b (Novagen)	Ampicillin	N-terminal hexahistidine peptide tag	T7*lacUV5* *promoter*, IPTG	Thrombin
pET22b (Novagen)	Ampicillin	C-terminal hexahistidine peptide tag	T7*lacUV5* *promoter*, IPTG	None
pASK-IBA3 (IBA)	Ampicillin	C-terminal StrepII tag (NH_2- WSHPQFEK - COOH)	*tetA* promoter, anhydrotetracycline	None

The pET vector systems (Novagen, EMD Biosciences, Germany) were chosen for His_6-tagged expression and affinity purification and the pASK vector (IBA, Germany) for Strep-tagged expression and affinity purification of recombinant proteins in *E. coli*. Two different synthetic *P. falciparum dhfs-fpgs* gene constructs were made, one with the stop codon for N-terminal His_6-tagged gene expression and another without the stop codon for readthrough and expression of a C-terminal His_6-tag or Strep-tag fused to the protein. The protein was also expressed without a tag. In this instance the construct was cloned with the stop codon preceding the C-terminal histidine tag of the pET22b vector (Novagen, EMD Biosciences, Germany) to terminate transcription before the tag. The outer primers (outF and outR or alt4R) were designed with *Nde*I and *BamH*I restriction enzyme sites for in-frame cloning into the pET vector system using ATG in the *Nde*I restriction site (CAT↓ATG) as the start codon (Table 2.1, Chapter 2). For cloning into the pASK-IBA3 system another set of primers (ib3F and ib3R) were designed for incorporation of two *Bsa*I sites at the 5'- and 3'-ends of the gene and subsequent directional cloning into the pASK-IBA3 vector (Table 3.2).

Table 3.2 Primers used for cloning of *P. falciparum dhfs-fpgs* into the pASK-IBA3 vector for C-terminal Strep-tagged expression. The highlighted parts of the sequence indicate the *Bsa*I restriction enzyme site (GGTCTCN ↓ NNNN).

Primer name	Sequence (5' to 3')	Length (nt)	T_m (°C)
ib3F	GCATCGGGTCTCGAATGGAGAAGAACCAGAACG	33	68
ib3R	GCATCGGGTCTCAGCGCTCACCAGGCTCGGTTCGTTC	37	75

Different *E. coli* cell lines were used to determine which system gave optimal protein expression (Table 3.3). Cells containing the pLysS plasmid show increased stability of toxic genes expressed by the system. The pLysS plasmid encodes a small amount of T7 lysozyme, which binds to T7 RNA polymerase thereby inhibiting transcription and preventing cell death (Huang *et al.*, 1999). BL21 Star (DE3) cells were chosen for their high level of gene expression and BL21 Gold (DE3) pLysS cells for ease of transformation (Invitrogen Technical Data, La Jolla, USA).

Table 3.3: *E. coli* strains used as hosts for protein expression.

Expression host	Genotype	Antibiotic resistance
BL21 (DE3) pLysS (Stratagene, La Jolla, USA)	B strain F⁻ *dcm ompT hsdS*$_B$ ($r_B^- m_B^-$) *gal* λ(DE3) [pLysS Cam^r]	Chloramphenicol
BL21 Star (DE3) (Invitrogen, La Jolla, USA)	B strain F⁻ *ompT hsdS* ($r_B^- m_B^-$) *gal dcm rne131* (DE3)	None
BL21 Gold (DE3) pLysS (Invitrogen La Jolla, USA)	B strain F⁻ *dcm ompT hsdS* ($r_B^- m_B^-$) dcm$^+$ Tet^r *gal* λ(DE3) *endA* Hte [pLysS Cam^r]	Chloramphenicol
BL21 Codon plus pRIL (Invitrogen La Jolla, USA)	B strain F⁻*ompT hsdS* ($r_B^- m_B^-$) *dcm*$^+$Tet^r*gal endA* Hte [argU ileY leuW Cam^r]	Chloramphenicol

3.2.2 Protein expression

$CaCl_2$ competent expression hosts (section 2.2.4.2) were freshly transformed with plasmid containing the different tagged gene constructs and plated onto LB plates with ampicillin (100 µg/ml) to select for the plasmid containing the gene, and supplemented with 100 µg/ml chloramphenicol for selection of the pLysS plasmid in BL21 (DE3) pLysS and BL21 Gold (DE3) pLysS cells. A portion (400 µl) of the transformed cells were not plated but diluted into 10ml LB liquid media containing 50µg/ml ampicillin and grown overnight at 30°C with

shaking for population expression (the total number of cells that received the plasmid). The overnight cultures were then diluted 20-fold in LB liquid media with ampicillin (50 µg/ml) and grown at 30°C with shaking until an optical density of 0.4 at 600nm was reached. Uninduced samples (2 ml each) were taken and a final concentration of 0.1 mM IPTG added to the cells for induction of protein expression from the pET system. For Strep-tagged expression a final concentration of 0.2 µg/ml anhydrotetracycline (of a 2mg/ml stock solution in dimethylformamide) was added to each culture for induction of protein expression. Expression was carried out at 30°C for 3.5 hours, after which cells were pelleted by centrifugation at 16000xg for 1 min or 5000xg for 15 min (for culture volumes >50ml) at 22°C. The pellets were dissolved into Bugbuster Protein extraction reagent (Novagen, EMD Biosciences, Germany), which is a zwitterionic detergent for gentle lysis of the cells. A volume of 5 ml "Bugbuster" per gram wet cell paste was used. For protein extraction the samples were left at 4°C overnight whereafter the induced samples were split into two fractions; one was used for total protein analysis and the other was centrifuged at 16000xg for 5min or 16000xg for 20 min (for culture volumes >50ml). The resulting pellet was analysed as the insoluble fraction and the supernatant as the soluble fraction.

3.2.3 Determination of the protein concentration

The Bradford method was used for protein concentration determination (Bradford, 1976). A stock solution of 10 mg/ml bovine serum albumin (BSA) was diluted to obtain a standard concentration range of 200, 100, 50, 25, 12.5 and 6.25 µg/ml protein. To 100 µl of each BSA standard or protein samples, 100 µl Coomassie Plus Assay Reagent (Pierce Biotechnology Inc., Rockford, USA) was added and the absorbancy measured at 595nm in 96 well microtitre plates (Bibby Sterilin Ltd., Stone, Staffordshire, UK). Protein concentrations were also estimated by comparison of the intensity of protein bands on an SDS-PAGE gel with that of the molecular marker, Broad Range Protein Molecular Mass Markers (Promega, Wisconsin, USA), which contains 0.3 µg/µl of the 50 kDa band and 0.1 µg/µl of the rest of the bands.

3.2.4 SDS-PAGE analysis

Polyacrylamide gel electrophoresis (PAGE) was performed with a 5% acrylamide-methylene-(bis) acrylamide stacking gel containing 0.5% SDS, 0.5% TEMED, 0.1% ammonium persulfate, 0.125 M Tris-Cl (pH 6.8) and an 8 or 12% acrylamide-methylene-(bis)acrylamide separating gel containing 0.535 M Tris-Cl (pH 8.8), 0.5% w/v SDS, 0.5% v/v TEMED and

0.1% w/v ammonium persulfate. Electrophoresis was performed in a 0.025 M Tris-0.2M glycine buffer (pH 8.3). Samples were diluted 1:1 in a 2X concentrated SDS sample buffer (125 M TrisCl, 0.02% w/v SDS, 0.5% v/v glycerol, 0.05% v/v β-mercaptoethanol and 0.01% w/v bromophenol blue) and denatured by boiling for 5 min. Gels were run at 70 Volts in the Minigel G-41 system (Biometra, Germany). Protein bands were visualised with Coomassie Blue G250 staining (0.1g Coomassie brilliant blue G250 in 40% methanol, 10% acetic acid, destained with 40% methanol, 10% acetic acid). Silver staining was used for visualisation of low protein concentrations. Gels were fixed for 30 min in 30% ethanol, 10% acetic acid and then treated for 30 min with 30% ethanol, 0.5M sodium acetate, 0.5% glutaraldehyde and 0.2% $Na_2S_2O_3$. Gels were then washed three times for 10 min with water and stained for 30 min in 0.1% $AgNO_3$, 0.02% formaldehyde and developed in 0.5% Na_2CO_3, 0.01% formaldehyde and developing stopped by the addition of 0.5 M EDTA (Merril *et al.*, 1981).

3.2.5 Functional complementation

Complementation of the *E. coli* K-12 mutant strain, Sf4 (F^- *folC strA recA tn10: srlC*) with the different synthetic constructs was performed to compare the complementation efficiency of their DHFS and FPGS activities (Bognar *et al.*, 1985). This specific strain has a single point mutation, A309T that reduces DHFS and FPGS activities to less than 3% of the wild type activities (Bognar *et al.*, 1985). Cells therefore produce basal folate levels to overcome thymine deficiency but are auxotrophic for methionine. Glycine is added to stimulate growth. Complementation was performed on minimal media plates containing 0.5% (w/v) glucose, 1.6% agar, 50 µg/ml streptomycin to select for the Sf4 mutant, 100 µg/ml ampicillin to select for the vectors containing the *dhfs-fpgs* gene in an M9 salts solution (1.3%(w/v) $Na_2HPO_4.7H_20$, 0.3%(w/v) KH_2PO_4, 0.05%(w/v) NaCl and 0.1%(w/v) NH_4Cl). Positive controls were supplemented with 40 µg/ml methionine and 20 µg/ml glycine. Constructs used for complementation were the vector alone (referred to further as U4) as negative control, the native *P. falciparum dhfs-fpgs* (referred to further as U5), the N-terminal histidine-tagged synthetic *dhfs-fpgs* (referred to in further as nh), the C-terminal histidine-tagged synthetic *dhfs-fpgs* (referred to in further as ch), the C-terminal Strep-tagged synthetic *dhfs-fpgs* (referred to in further as cs) and the tagless synthetic *dhfs-fpgs* (referred to in further as tl). The following titration method was used for complementation: Single colonies of Sf4 cells transformed with each of the above constructs were picked and grown in LB liquid medium at 30°C with shaking till growth was slightly visible (OD_{600nm} = 0.1). Of each of these cultures, 100 µl was placed in 5 ml LB liquid culture medium containing 50 µg/ml ampicillin and 50 µg/ml streptomycin,

grown overnight till saturation (OD_{600nm}> 1) and thus comparable cell density. A dilution series of 10-fold, 100-fold and 1000-fold of each culture volume were made and 5 µl of each dilution spotted onto the minimal media plates with and without supplementation (positive control). For complementation in minimal liquid media (0.5% w/v glucose, 50µg/ml streptomycin, 50µg/ml ampicillin and the M9 salts solution), 100 µl duplicate saturated overnight cultures of each of the constructs above, were placed in 5 ml culture medium and growth monitored at OD_{600nm} after 24 hours at 30°C.

3.2.6 Partial protein purification

3.2.6.1 Unfolding/refolding protocol

After protein extraction and separation of the soluble and insoluble fractions by centrifugation (refer to section 3.2.2), the inclusion bodies from BL21 Star (DE3) cells expressing the C-terminal His_6-tagged protein were further purified using the manufacturer's protocol for Bugbuster Protein Extraction Reagent (Novagen, EMD Biosciences, Germany). Cell pellets resulting from 50 ml of liquid culture were resuspended in 1.25 ml undiluted "Bugbuster" and mixed well by vortexing. Lysozyme (10mg/ml stock solution) was added to a final concentration of 200 µg/ml and the mixture vortexed for 5 min at 22°C. To this, 7.5 ml of a 10-fold dilution of "Bugbuster" was added, the mixture vortexed for 1 min and centrifuged at 16 000xg for 15 min at 4°C. The supernatant was removed and the pellet resuspended in 1 ml 10-fold diluted "Bugbuster". The sample was mixed well and centrifuged at 16 000xg for 15 min. This was repeated twice. Solubilisation of the purified inclusion bodies was done as follows (Sirawaraporn *et al.*, 1993): The remaining pellet was suspended in 1 ml of 20 mM potassium phosphate buffer (pH 7) containing 0.1 mM EDTA, 10 mM DTT, 0.2 M KCl and 6 M guanidine HCl and left at 4°C for 1 hour for solubilisation of the protein. The protein mixture was then added dropwise to achieve a 20 times dilution in a 20 mM potassium phosphate buffer (pH 7) containing 0.1 mM EDTA, 10 mM DTT, 0.2 M KCl and 20% glycerol at 4°C for refolding of the protein. This mixture was left overnight at 4°C and then centrifuged at 16000xg for 20 min to pellet all the remaining insoluble proteins. The soluble fraction was then used for affinity purification.

3.2.6.2 Affinity chromatography

Purification of histidine-tagged protein was conducted by immobilised metal ion affinity chromatography (IMAC) with HIS select HC Nickel affinity gel (Sigma-Aldrich, St.Louis, USA). This method of affinity purification relies on the electrostatic interaction between the

hexahistidine peptide tag of the target protein (which is negatively charged at pH 8) and the positively charged Ni^{2+} immobilised on the column resin. Unbound proteins are washed off and the target protein with the hexahistidine tag eluted by the addition of a high imidazole concentration, which binds competitively to the Ni^{2+} column resin. The batch purification method was used for purification of protein: 250 µl of the nickel affinity gel suspension was added to a microcentrifuge tube and centrifuged for 30 sec at 5000xg. The supernatant was removed and 200 µl of equilibration buffer (50 mM sodium phosphate (pH 8) and 0.3 M sodium chloride) was added. The solution was mixed well and centrifuged for 30 sec at 5000xg. The supernatant was discarded and 1 ml of the soluble fraction obtained after cell lysis as described in section 3.2.2 (adjusted to pH 8 with equilibration buffer) was added to the affinity gel matrix, mixed for 1 minute and centrifuged as before. For larger volumes of soluble fraction, the previous step was repeated a few times to load the whole fraction stepwise onto the matrix. The affinity gel was washed two times with 1 ml equilibration buffer. Each time the gel was mixed with the buffer, it was centrifuged at 5000xg for 30sec and the supernatant removed. The target protein was eluted in 250 µl of elution buffer (50 mM sodium phosphate, pH 8, 0.3 M sodium chloride and 250 mM imidazole). The gel and buffer were mixed well and centrifuged as before. This step was repeated with another 250 µl of elution buffer. Fractions (0.5 ml each) were taken for every wash and elution step and analysed with SDS-PAGE.

3.2.6.3 Size exclusion high performance liquid chromatography

Size exclusion chromatography was performed on a 300 x 7.8 mm Biosep-Sec-S3000 HPLC column (Phenomenex, California, USA) linked to the Beckman Gold HPLC system (Pump system 2). All buffers were filtered though a 0.22 µm membrane and degassed before use. Low Molecular weight markers (Amersham, Buckinghamshire, UK) made up to a concentration of 0.2 µg/µl in the size exclusion buffer (50 mM potassium phosphate (pH 7), 150 mM KCl) was used to calibrate the system. Soluble fractions of the uninduced and induced protein samples from BL21 (DE3) pLysS cells containing either C-terminal His_6-tagged- or tagless DHFS-FPGS were prepared as described in section 3.2.2. The soluble fractions were diluted 1:1 with the size exclusion buffer to adjust the pH to 7 and filtered through a 0.22 µm membrane. Samples were loaded onto the column and eluted with the size exclusion buffer at a constant flow rate of 0.5 ml/min for 40 min. The column was washed afterwards with the same buffer and stored in 0.05% sodium azide. Fractions (1ml each) collected from the size exclusion column were analysed with SDS-PAGE. Blank runs (consisting only of the size exclusion buffer) were performed before each sample run.

3.3 Results

3.3.1 Expression of a variety of *dhfs-fpgs* constructs

In all the different cell lines, the pET22b-*dhfs-fpgs* (C-terminal His_6-tag) construct was expressed to a greater extent than the pET15b-*dhfs-fpgs* (N-terminal His-tag) construct (Figure 3.1 A vs. B). The highest expression of the N- and C-terminal His-tagged fusion proteins was obtained from the BL21 Star (DE3) cell line (~20% of the total cell protein content) when the different *E. coli* cell lines BL21 (DE3) pLysS, BL21 Gold (DE3) pLysS and BL21 Star (DE3) were compared. The other cell lines showed only a slight increase in expression after induction by IPTG.

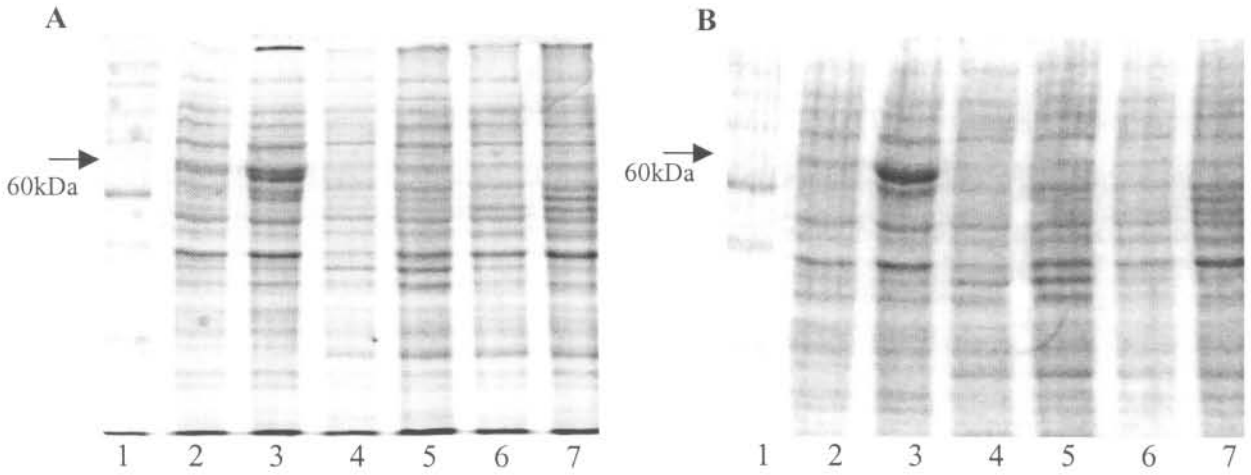


Figure 3.1: Expression of (A) pET15b-*dhfs-fpgs* (N-terminal His_6 tag) and (B) pET22b-*dhfs-fpgs* (C terminal His_6 tag) in various cell lines. Lane 1: Broad Range Protein Molecular Mass Markers, lane 2: BL21 Star (DE3) uninduced, lane 3: BL21 Star (DE3) induced, lane 4: BL21 Gold (DE3) pLysS uninduced, lane 5: BL21 Gold (DE3) pLysS induced, lane 6: BL21 (DE3) pLysS uninduced, lane 7: BL21 (DE3) pLysS induced. Arrows indicate the expected protein size (60kDa).

Comparison of the soluble and insoluble fractions of the N- (results not shown) and C-terminal His tagged constructs expressed by the BL21 Star (DE3) cells revealed however that the proteins were expressed as inclusion bodies (Figure 3.2 A). BL21 Gold (DE3) pLysS cells expressed a small amount of protein in the soluble fraction (Figure 3.2 B) and BL21 (DE3) pLysS cells expressed only soluble protein (Figure 3.2 C). Overall, the soluble expression of BL21 (DE3) pLysS cells was much less than the insoluble expression of BL21 Star (DE3) cells (Figure 3.2 A vs. C).

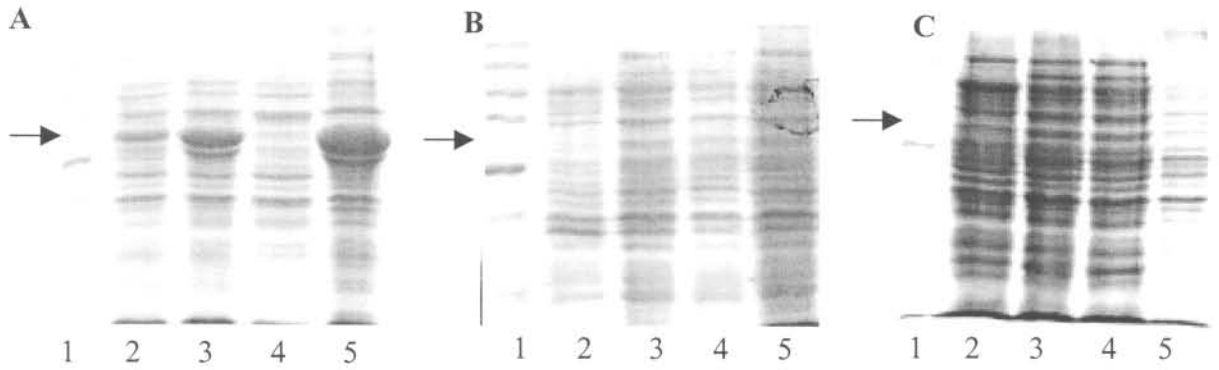


Figure 3.2: Solubility of C-terminal His-tagged DHFS-FPGS in A: BL21 Star (DE3), B: BL21 Gold (DE3) pLysS and C: BL21 (DE3) pLysS cell lines. Lane 1: Broad Range Protein Molecular Mass Markers, lane 2: uninduced cells, lane 3: induced total cell protein, lane 4: soluble protein fraction, lane 5: insoluble protein fraction. The arrows indicate the position of the expected 60kDa band.

Expression of the tagless *dhfs-fpgs* gene in BL21 (DE3) pLysS cells gave ~ 50% soluble protein (Figure 3.3 C, lane 3) but < 50% soluble in BL21 Gold (DE3) pLysS cells (Figure 3.3 B, lane 3) and even less soluble in BL21 Star (DE3) cells (Figure 3.3 A, lane 3). For all the cell lines total protein expression levels were less for the tagless construct than for the C-terminal Strep-tagged (Figure 3.4) or C-terminal His_6-tagged construct (Figure 3.2).

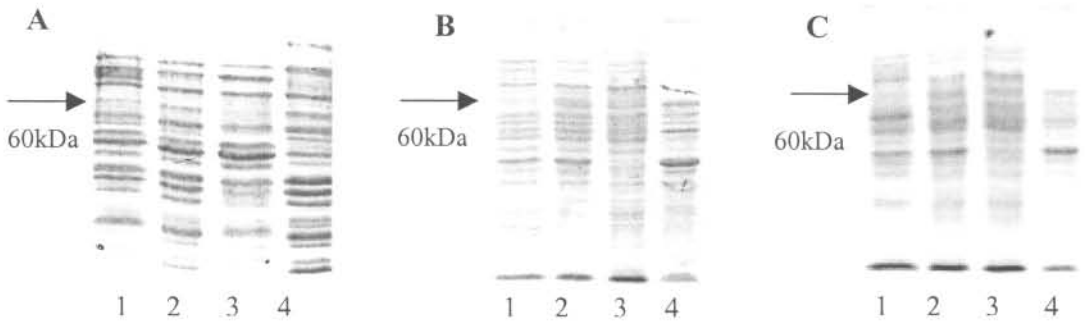


Figure 3.3: Solubility of tagless DHFS-FPGS in A: BL21 Star (DE3), B: BL21 Gold (DE3) pLysS and C: BL21 (DE3) pLysS cell lines. Lane 1: uninduced cells, lane 2: induced total cell protein, lane 3: soluble protein fraction, lane 4: insoluble protein fraction. The arrows indicate the position of the expected 60kDa band.

Expression of the C-terminal Strep-tagged DHFS in BL21 Star (DE3) cells resulted in inclusion bodies (Figure 3.4 A, lane 4) and soluble protein was expressed in very small quantities in BL21 Gold (DE3) pLysS cells and BL21 (DE3) pLysS cells (Figure 3.4 B and C lane 3).

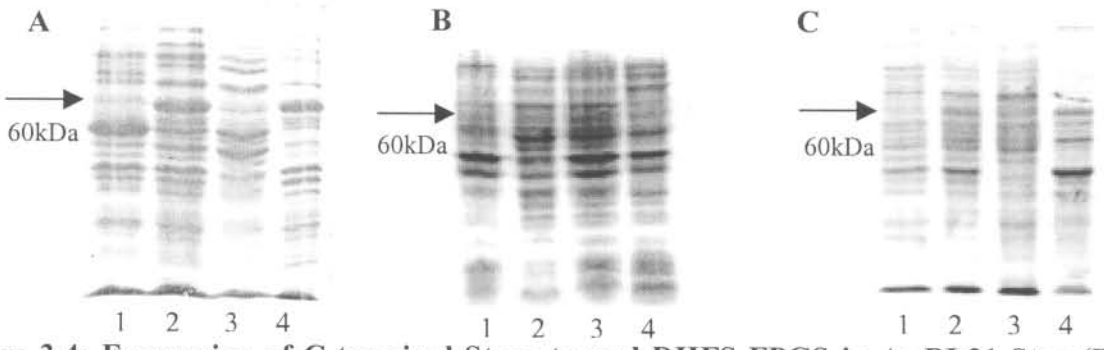


Figure 3.4: Expression of C-terminal Strep-tagged DHFS-FPGS in A: BL21 Star (DE3), B: BL21 Gold (DE3) pLysS and C: BL21 (DE3) pLysS cell lines. Lane 1: uninduced cells, lane 2: induced total cell protein, lane 3: soluble protein fraction, lane 4: insoluble protein fraction. The arrows indicate the position of the expected 60kDa band.

3.3.2 Functional complementation

The next step was to compare the ability of the different synthetic *dhfs-fpgs* constructs to complement a DHFS-FPGS deficient *E. coli* cell line, Sf4. Growth with methionine and glycine supplementation was observed regardless of their dilution on solid media, indicating the viability of the cells (Figure 3.5 B). On the minimal media plates however, growth was only observed for the lowest (10-fold) dilution of the C-terminal His tagged and tagless synthetic protein (Figure 3.5 A rows 4 and 5). Cells containing the N-terminal His_6-tagged and C-terminal Strep-tagged proteins formed pinpoint colonies on minimal media and the growth of cells transformed with native *P. falciparum dhfs-fpgs* (U5) or without a gene (negative control) was below the detection limit of the assay.

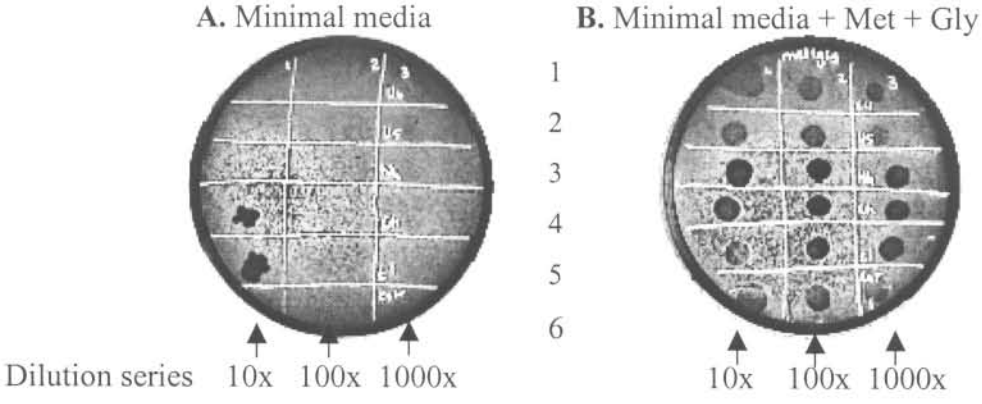


Figure 3.5: Complementation of DHFS-FPGS deficient *E. coli* (Sf4) by different synthetic *P. falciparum dhfs-fpgs* constructs on solid minimal media (plate A) or methionine and glycine supplemented minimal media (plate B). Sf4 cells contained either row 1: only vector (negative control), row 2: the native *P. falciparum dhfs-fpgs*, row 3: N-terminal His_6 tagged *dhfs-fpgs*, row 4: C-terminal His_6 tagged *dhfs-fpgs*, row 5: tagless *dhfs-fpgs* and row 6: C-terminal Strep tagged *dhfs-fpgs*.

The results obtained on solid media were verified by optical density growth measurements in liquid minimal media after 4 and 24 hours. The tagless construct (Figure 3.6, tl: cyan bar) showed the best growth after 24 hours. The Sf4 cells without any construct (negative control) showed no growth in minimal media and only slight growth in supplemented media (Figure 3.6, Sf4). Sf4 cells containing the C-terminal His_6-tagged construct had the second highest growth on minimal media after 24 hours (Figure 3.6, ch: cyan bar), followed by the C-terminal Strep-tagged and N-terminal His-tagged constructs. These results are consistent with the solid media results. *P. falciparum dhfs-fpgs* had higher growth levels than the negative control, but growth of both these transformants were much lower than for the cells containing the synthetic *dhfs-fpgs* constructs.

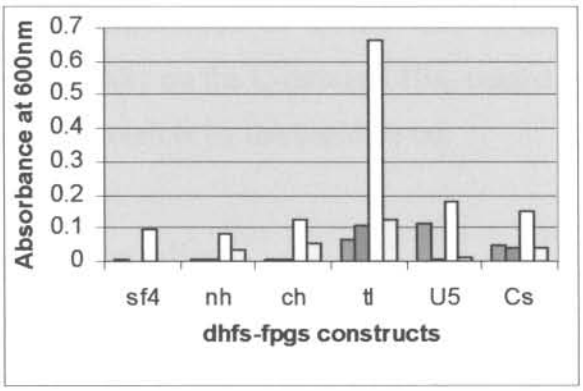


Figure 3.6: Growth of Sf4 *E. coli* containing different constructs in liquid media as measured by the optical density at 600 nm. Sf4=*E. coli* mutant cells without any construct (negative control), nh=N- terminal His_6 tagged *dhfs-fpgs*, ch= C-terminal His_6 tagged, tl= tagless *dhfs-fpgs*, U5= the *P. falciparum dhfs-fpgs* and cs= C-terminal Strep tagged *dhfs-fpgs*. Supplemented minimal media, 4 hours (■) and 24 hours (□), minimal media, 4 hours (■) and 24 hours (□).

Comparison of the growth curves over different time periods showed that the Sf4 cells containing the tagless construct grew at a much higher rate than for the Sf4 cells alone as indicated by the steep incline of the slopes (Figure 3.7). Growth for the Sf4 cell culture containing the tagless construct peaked at 8 hours and then declined over longer incubation periods (Figure 3.7: blue and yellow curves). At 8 hours there was a 18.75-fold increase in growth for tagless DHFS-FPGS in comparison with the negative control, Sf4, on supplemented media and a 9-fold increase on minimal media. Similarly, for C-terminal His_6-tagged DHFS-FPGS there was a 3-fold increase in growth on supplemented media and a 1.6-fold increase on minimal media when compared to the Sf4 cells alone (results not shown).

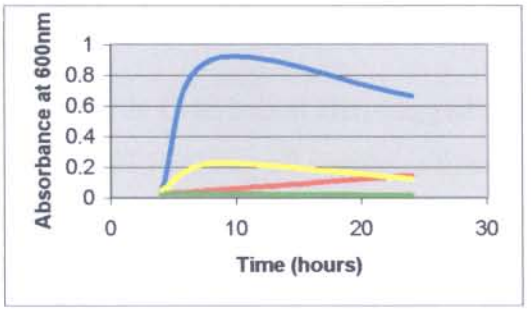


Figure 3.7 Growth curves over 24 hours for the tagless *dhfs-fpgs* construct in Sf4 cells (tl) versus the negative control, Sf4 cells without any construct (Sf4). (—) tl in supplemented media , (—) tl in minimal media, (—) Sf4 in supplemented media, (—) Sf4 in minimal media.

From these results it is thus evident that only the tagless and the C-terminal His_6-tagged constructs show significant complementation of DHFS and FPGS activity. Preliminary purification studies thus focused initially on the C-terminal His_6-tagged DHFS-FPGS construct due to the ease of purification made possible by the histidine tag.

3.3.3 Preliminary purification studies

3.3.3.1 Affinity purification of refolded C-terminal His_6-tagged DHFS-FPGS from inclusion bodies

Total protein was extracted from BL21 Star (DE3) cells expressing the C-terminal His-tagged DHFS-FPGS (Figure 3.8). The soluble and insoluble fractions (lanes 4 and 5 respectively) were separated and the insoluble fraction further purified to isolate the inclusion bodies (lane 6). After the denaturation-refolding protocol and affinity purification, the induced protein was obtained at an estimated concentration of 2 mg/l liquid culture (lane 9). Relatively pure protein of the expected size was obtained after affinity purification since the isolation of inclusion bodies removed most of the other proteins (compare with the affinity purification of the soluble fraction of BL21 (DE3) pLysS cells expressing the C-terminal His_6-tagged protein in Figure 3.9).

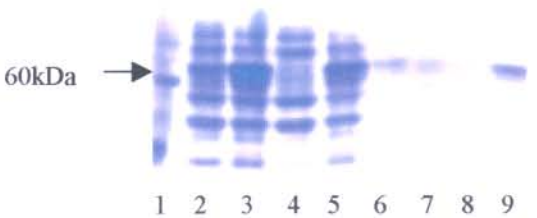


Figure 3.8: Affinity purification of resolubilised C-terminal His_6-tagged DHFS-FPGS obtained from BL21 Star (DE3) cells on a Ni^{2+} column. Lane 1: Broad Range Protein Molecular Mass Markers molecular marker, lane 2: uninduced cells, lane 3: total cell protein (induced cells), lane4: soluble fraction (induced cells), lane 5: insoluble fraction (induced cells), lane 6: purified refolded inclusion bodies, lanes 7 and 8: wash steps 1 and 2 and lane 9: eluted protein

3.3.3.2 Affinity purification of soluble C-terminal His_6-tagged DHFS-FPGS

The soluble fraction of the C-terminal His_6-tagged construct expressed in BL21 (DE3) pLysS cells was purified by affinity chromatography as described in section 3.2.6.2. It was evident that some of the C-terminal His-tagged DHFS-FPGS did not bind to the column and was removed during the wash steps (Figure 3.9, compare lanes 4 and 6). The protein of expected size was eluted along with other proteins (Figure 3.9, lane 6). Less of the pure C-terminal His_6-tagged protein was thus obtained from the soluble protein fraction expressed by BL21 (DE3) pLysS cells than for the refolded insoluble fraction expressed by BL21 Star (DE3) cells (Figure 3.8).

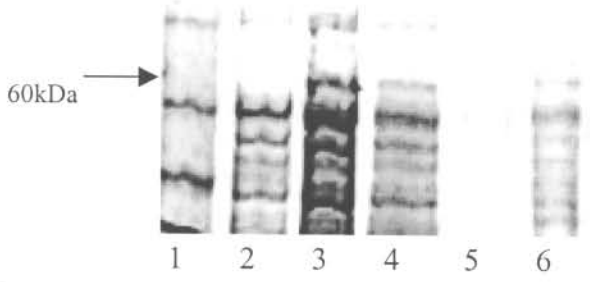


Figure 3.9: C-terminal His_6-tagged DHFS-FPGS affinity purified from the soluble fraction expressed by BL21 (DE3) pLysS cells. Lane 1: Broad Range Molecular Mass Marker, lane 2: uninduced BL21 (DE3) pLysS cells, soluble fraction, lane 3: induced BL21 (DE3) pLysS cells, soluble fraction, lane 4: unbound protein washed off the affinity column, lane 5: final wash fraction, lane 6: eluted protein.

3.3.3.3 Size exclusion high performance liquid chromatography

The size exclusion HPLC column was first calibrated with low molecular mass protein standards: 97 kDa, 66 kDa, 45 kDa, 30 kDa, 20.1 kDa and 14.4 kDa (Figure 3.10). The mass of the histidine tagged monomer was predicted as 62 kDa from the primary amino acid sequence (http://us.expasy.org/cgi_bin/pi_tool). Using the standard curve obtained for gel filtration of the molecular markers, it was predicted that the monomer would elute at a retention time of ~19.8 min and the dimer, if present, at ~16.3 min. The equation used for this prediction was derived from the molecular mass standard curve (Figure 3.10): log molecular weight= -0.0809 (retention time) + 3.395; r^2=0.9741.

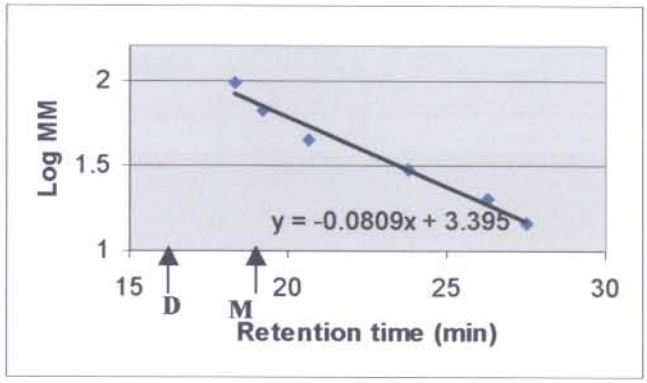


Figure 3.10: Retention times of low molecular mass protein standards. A logarithmic plot of the molecular mass vs. retention time. Data points correspond with 97kDa, 66kDa, 45 kDa, 30 kDa, 20.1 kDa and 14.4 kDa, respectively. Arrows D and M indicate the predicted retention times for the dimer and monomer, respectively.

Size exclusion HPLC was performed either on the crude cell extract (Figure 3.11) or on the fraction obtained after affinity chromatography (Figure 3.12). When affinity purification preceded size exclusion (Figure 3.12), there was a clear peak of induced protein in contrast to the slight shoulder observed by size exclusion of the total cell lysate (Figure 3.11). This result is most likely due to the removal of proteins in the first purification step.

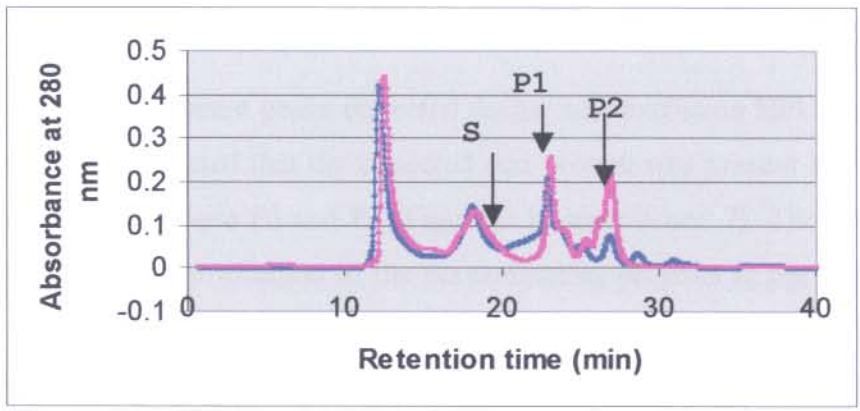


Figure 3.11: Size exclusion analysis of the expression of C-terminal His_6-tagged *dhfs-fpgs* from BL21 (DE3) pLysS cells. The uninduced cells (blue line) and induced cells (pink line) are superimposed. S represents a slightly induced shoulder at the expected retention time for the monomer. Induced peaks P1 and P2 correspond with molecular weights of 30 and 15 kDa respectively.

Comparison of the uninduced gel filtration profile with the induced profile (Figure 3.12) showed the induced protein of the expected monomer size (62kDa) at 19 min (P4). Due to peak broadening, it was predicted that fractions P2, P3 and P5 would also contain some of the induced protein. Other peaks were induced to a lesser extent and were found at 15 min (P1),

17min (P2), 33min and 37min. The 33 min and 37 min peaks corresponded to peaks observed from a blank buffer containing 250mM imidazole, used for elution of the affinity column bound proteins (results not shown). These peaks are absent in the size exclusion profile of the crude extract that contained no imidazole.

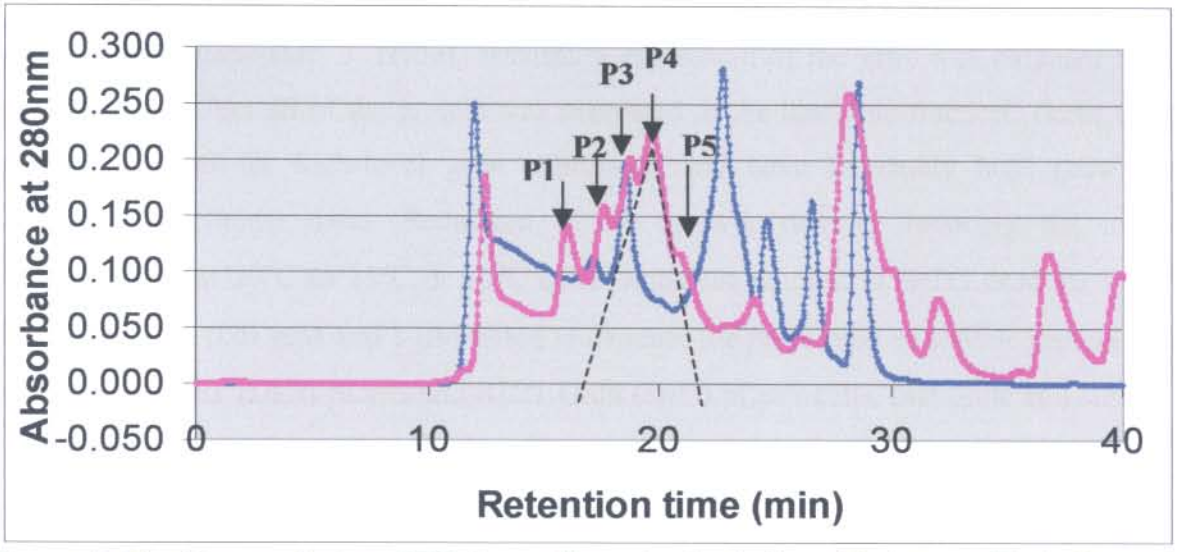


Figure 3.12: Size exclusion HPLC profiles obtained for affinity purified C-terminal histidine tagged DHFS-FPGS from BL21 (DE3) pLysS cells. Uninduced cells (blue line) and induced cells (pink line) are superimposed. P4 indicates the expected retention time (19.8 min) for a 62kDa protein. The dashed lines indicate the peak broadening towards the baseline, indicating the total time on the x axis in which the protein will elute from the column.

SDS-PAGE analysis of the different peaks collected during size exclusion HPLC after affinity purification (Figure 3.13) indicated that the expected size protein was present in fractions P2, P3, P4 and P5, but mainly between P3 and P4 (Figure 3.12 lanes 6 and 7). The size exclusion HPLC succeeded in the removal of some of the contaminating proteins in the mixture eluted from affinity chromatography.

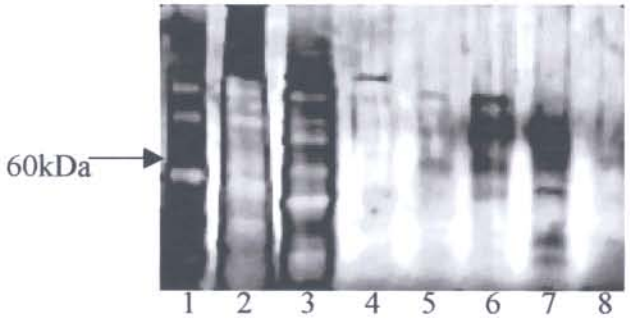


Figure 3.13: Silver-stained SDS PAGE of fractions obtained after size exclusion HPLC of the C-terminal His_6-tagged DHFS-FPGS. Lane 1: Broad Range Molecular Mass Markers, lane 2: uninduced protein, lane 3: induced protein, lane 4: fraction P1, lane 5: fraction P2, lane 6: fraction P3, lane 7: fraction P4, lane 8: fraction P5.

3.4 Discussion

3.4.1 Expression of various DHFS-FPGS constructs

The synthetic gene was expressed at about 45 times higher levels (0.225 g/l) than native *P. falciparum dhfs-fpgs* (<0.005 g/l), which was undetectable with Coomassie Blue staining (personal communication, J. Hyde). Maximum expression of the gene was obtained in BL21 Star (DE3) cells but all of the protein was expressed in the insoluble fraction. These cells are specifically used for high-level gene expression and have extremely high growth rates (Invitrogen Technical data). Reduction of the growth rate by reducing the induction temperature from 30°C to 23°C or 16°C or reducing the amount of IPTG used for induction from 0.1 mM to 0.01 mM and 1 μM failed to increase the proportion of soluble protein (results not shown). BL21 (DE3) pLysS and BL21 Gold (DE3) pLysS cells, that grew at a slower rate produced protein in the soluble fraction. The presence of the pLysS plasmid in these cell lines may also account for increased solubility through tighter control of gene expression and slower cell growth rates (Huang *et al.*, 1999). BL21 Gold (DE3) pLysS cells are normally used for more efficient transformation and sequencing of the cloned plasmids and approximately 30% of the protein expressed by these cells were soluble. BL21 (DE3) pLysS cells yielded only soluble protein and are normally the preferred cells to use for the expression of toxic proteins due to their high stringency of protein expression (Stratagene Technical Literature). The unusually high expression levels of this enzyme might be detrimental to the cells that normally only express native DHFS-FPGS (encoded by the *E. coli folC* gene) at low levels: approximately ~4% of the soluble cell protein (Bognar *et al.*, 1985). The protein expressed by BL21 (DE3) pLysS was much less than that of BL21 Star (DE3) cells and might thus account for the increased solubility. Although the synthetic gene was optimised for *E coli* codon usage, the amino acid composition of *P. falciparum* DHFS-FPGS was still biased in terms of positively charged amino acids: ~10% lysine which is double the lysine abundance in the *E. coli* genome ~4% (refer to appendix A). It is possible that this biased amino acid composition could account for the "toxicity" of the expressed protein as observed through the production of inclusion bodies.

In some instances, tags have been shown to cause insoluble expression (Ramage *et al.*, 2002). Other expression systems were also investigated to improve possible effects of the tag on soluble expression of the constructs. C-terminal Strep-tagged DHFS-FPGS was expressed mostly in the insoluble fraction by BL21 Star (DE3) cells, and in very low soluble levels in

BL21 Gold (DE3) pLysS and BL21 (DE3) pLysS cells, similar to the C-terminal His_6 tagged protein. Western blot analysis verified the large amount of insoluble protein as the Strep-tagged construct (results not shown). It was thus concluded that the altered chemical nature of the tag does not appear to influence the solubility of the expressed protein.

The tagless construct produced soluble protein mostly in BL21 (DE3) pLysS and BL21 Gold (DE3) pLysS cells, which supported the idea that the presence of a tag could be detrimental to the solubility of the protein. Alternatively, the solubility could also be attributed to the lower total expression levels of the tagless construct when compared to that of the C-terminal Strep and C-terminal His_6-tagged constructs. Since tags sometimes also increase expression levels (Ramage *et al.*, 2002), the absence of the tag might account for the overall lower observed expression and increased solubility.

3.4.2 Functional complementation

Functional complementation of the *E. coli* Sf4 mutant strain was done to determine the success of the synthetic *dhfs-fpgs* in producing a functional protein and to identify which of the different synthetic *dhfs-fpgs* constructs produced the most active protein. On minimal media plates, only the tagless- and C-terminal His_6-tagged constructs were capable of significantly restoring mutant cell growth from auxotrophy. Growth of the other cells was below the detection limit of the assay. Further investigations in liquid culture media at 4 and 24 hours showed the best growth for the tagless construct at both time intervals, with or without supplementation. It could thus be concluded that the synthetic gene encoded both DHFS and FPGS activities. It was evident from the growth curves that saturation of the Sf4 cell line was not achieved within 24 hours, because of the low nutrient content of the media. Furthermore, the decline in cell growth after 8 hours for Sf4 cells bearing the tagless construct might indicate possible "toxicity" of the expressed protein and/or products at that time. Taken together with the previous results, it was evident that the tags influenced not only the solubility but also possibly the enzyme activity of the expressed protein. Nevertheless, the tagged and untagged synthetic constructs seemed to complement the cells to a greater extent than the native *P. falciparum dhfs-fpgs* gene, possibly due to increased expression levels. It could thus be concluded at this stage that the synthetic *dhfs-fpgs* gene was successful in expressing higher levels of active protein, than obtained from the native *P. falciparum dhfs-fpgs* gene alone.

3.4.3 Partial purification

Due to the complexity of the enzyme assay, the expensiveness of the substrates and their sensitivity to oxidation it was deemed necessary to determine the enzyme activities of the crude extract and different purification steps within a single experiment to ensure reproducibility of the results. From the complementation results tagless DHFS-FPGS would be the preferred protein to purify for activity assays, but the absence of a tag makes affinity purification more difficult. Preliminary studies were therefore performed first with the C-terminal His_6-tagged DHFS-FPGS to establish the most suitable purification protocol, which is also applicable to tagless DHFS-FPGS, since there are no significant size or pI differences between these two enzymes. The purification of the His_6-tagged enzyme can thus be followed without determination of enzymatic activity with anti-His antibodies.

3.4.3.1 Unfolding and refolding of inclusion bodies

Solubilisation and refolding from inclusion bodies resulted in the successful affinity purification of relatively pure C-terminal His_6-tagged gene. The same method was used here as for the solubilisation of *P. falciparum* DHFR inclusion bodies obtained from synthetic gene expression (Sirawaraporn *et al.*, 1993). The refolding buffer contained a high K^+ concentration (0.2 M KCl) since it was shown that it stabilises the *Corynebacterium* DHFS-FPGS (Shane, 1980). Despite these guidelines, the correct refolding of the protein can only be verified by comparison of the enzyme activities of naturally soluble and the resolubilised protein obtained from inclusion bodies. Should activity assays show no difference in properties between the refolded C-His tagged- and the soluble tagless protein, this purification method could be used for large-scale protein isolation, since it has the highest protein yields.

3.4.3.2 Affinity purification of C-terminal His_6-tagged DHFS-FPGS

Since the resolubilisation of inclusion bodies could not assure active protein, preliminary purification of the soluble fraction rather than the insoluble fraction was attempted. Due to the ease of purification, the isolation of the C-terminal His_6-tagged construct expressed in the soluble form by BL21 (DE3) pLysS cells was investigated first. Affinity chromatography of the C-terminal His_6 tagged protein expressed in BL21 (DE3) pLysS cells did not yield high concentrations of purified protein in contrast to the larger amount of relatively pure protein obtained after denaturation and refolding of protein obtained from the inclusion bodies. However, it is expected that the inclusion bodies would yield more protein than the soluble

fraction as indicated by the SDS-PAGE expression analysis, section 3.3.3.1. Furthermore, some of the C-terminal His_6-tagged protein didn't initially bind to the column, implying that the availability of the histidine tag for affinity binding was less for the soluble DHFS-FPGS than refolded DHFS-FPGS, possibly due to variable flexibility of the soluble expressed protein. The affinity purification step was, however, essential since it improved the size exclusion HPLC profile. This result underlines the importance of the combination of both purification steps for the removal of proteins.

3.4.4 Future prospects

At the time of completion of the thesis, the anti-His_6 antibodies weren't available. Future experiments will thus include the detection of the C-terminal His_6-tagged *dhfs-fpgs* in the different purified fractions by these antibodies. This will guide the optimisation of the purification strategy and minimise the amount of samples required for activity assays of each purification step, namely size exclusion HPLC and anion exchange HPLC. Using the information gained from the C-terminal His_6-tagged expression and purification, the optimised experiments can be extrapolated to the tagless construct for size exclusion purification and anion exchange purification since there is no significant size or pI differences between the two proteins. Should a third purification step be necessary, a Cibacron Blue (AMP analogue) column for affinity purification of the tagless protein using the ATP-binding site (P-loop), will be used. Isolated tagless protein could then be used to raise polyspecific monoclonal antibodies for detection of the tagless protein in various isolated fractions, as with the C-terminal His_6-tagged protein and enzyme assays then performed on the identified fractions. It would even be possible to link these antibodies to a resin for affinity purification of the enzyme, which could be used for large-scale protein purification for kinetic studies, crystallisation and structure determination.

Chapter 4

In silico analysis of dihydrofolate synthase-folylpolyglutamate synthase (DHFS-FPGS)

4.1 Introduction

The role of folylpolyglutamate synthase (FPGS) in the cellular retention of reduced folate co-factors and inhibitors has made it an interesting target for cancer and bacterial chemotherapy (Gangjee *et al.*, 2002). Although FPGS occurs in all cells, only one crystal structure has been determined to date; namely that of the prokaryote *Lactobacillus casei* (Sun *et al.*, 1998). In contrast with FPGS, dihydrofolate synthase (DHFS) has been much less studied to date, since it only occurs in organisms capable of folate biosynthesis (section 1.5.1). In mammals (that import dietary folate) and even certain bacteria such as *L. casei* and *H. influenzeae*, DHFS activity is absent (Sun *et al.*, 1998). Eukaryotes such as *A. thaliana* and yeast have separate genes encoding DHFS and FPGS activity, but *P. falciparum* is the first eukaryote to date to contain both activities within a single gene product (Ravanel *et al.*, 2001) and (Salcedo, 2001)

.

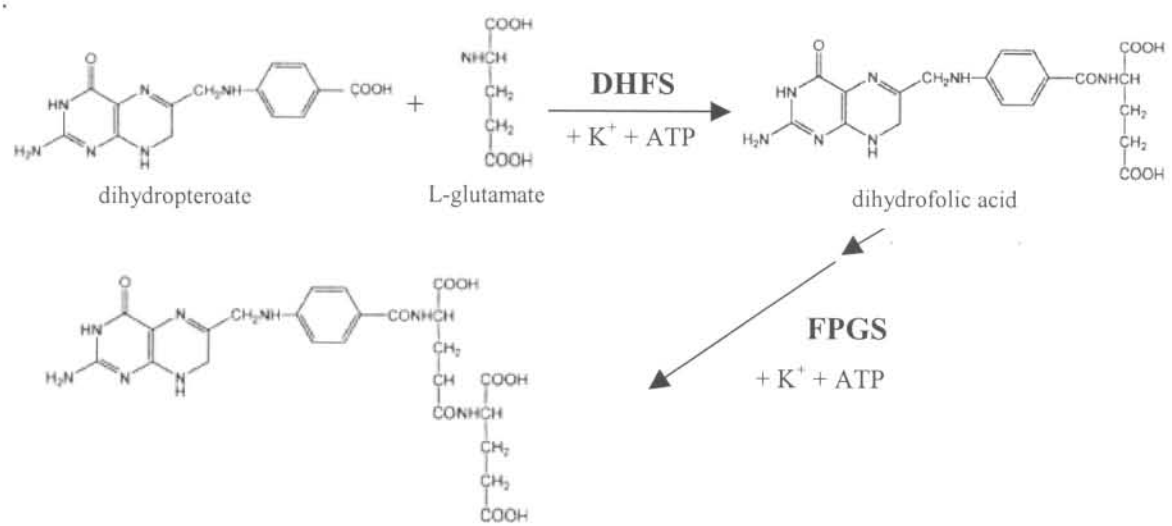


Figure 4.1: DHFS and FPGS enzyme reactions (adapted from Yuthavong, 2002).

Since DHFS and FPGS perform essentially the same enzyme reactions, namely the ATP dependant addition of a glutamate residue to a pteridine derivative, it is possible that the

bifunctional enzyme only contains one active site (Figure 4.1). This idea is supported by the fact that a single mutation abolishes both DHFS and FPGS activities in the bifunctional *E. coli* enzyme (Keshavjee *et al.*, 1991). The *L. casei* crystal structure revealed that the enzyme consisted of two domains (Figure 4.2): the N-terminal domain consisting of a mononucleotide binding fold (P-loop) similar to that found in proteins of the adenylate kinase family and a C-terminal domain, which is very similar to the folate-binding dihydrofolate reductase enzyme (Sun *et al.*, 1998). The active site is located between these two domains next to the P-loop and the folate-binding site is located in a C-terminal domain hydrophobic pocket between two alpha helices, A10 and A11 (Sun *et al.*, 1998). Despite the similarities between the FPGS C-terminal domain and folate-binding site of human DHFR, it is proposed that the orientation of the folate substrate, dihydrofolate (DHFR) and 5,10-methylene tetrahydrofolate (FPGS) might differ (Sun et al., 1998). In FPGS, the pteridine ring structure (fused white circles, Figure 4.2) will interact with the hydrophobic area between helices A10 and A11 and the growing chain of glutamate residues (dashed line, Figure 4.2) will project into the active site (dihydrofolate is in the opposite orientation). Another important feature is the 10-amino acid Ω-loop (cyan line, Figure 4.2), which has an important role in binding of the K^+ cations as well as interdomain stabilisation through hydrophobic interactions with the C-domain (Sun *et al.*, 1998). The linker region between the N- and C-terminal domains (green line, Figure 4.2) is important for domain movement and differences in the flexibility of this region might account for the different lengths of glutamate residues added in the various species (Sun *et al.*, 2001).

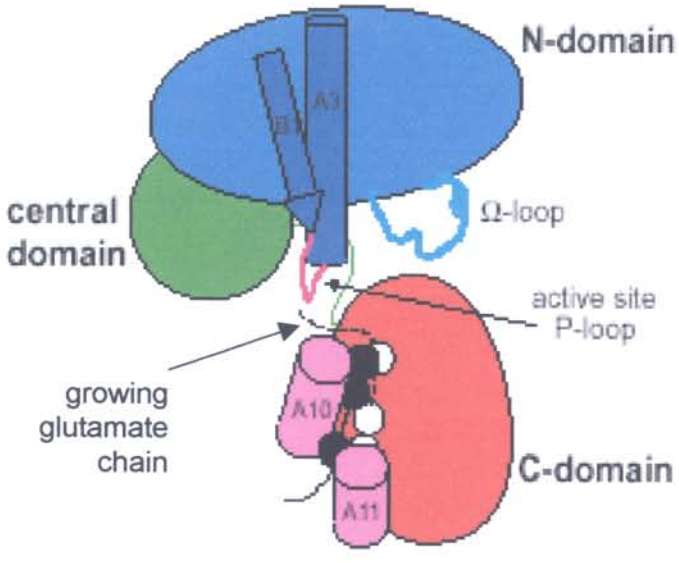


Figure 4.2: Schematic representation of *L. casei* FPGS. The two different binding modes of the folate substrates between helices A10 and A11 are compared for FPGS (white circles) and another folate-binding enzyme, DHFR (black circles) (Sun *et al.*, 1998).

To date, *P. falciparum* DHFS-FPGS has been characterized in terms of the DNA sequence, the predicted amino acid sequence and functional complementation. Comparison of these sequences with homologues indicated the presence of the ATP-binding P-loop, Ω loop as well as the FPGS signature sequence and selected residues required for catalytic activity (Salcedo *et al.*, 2001). Overall, alignments between the *P. falciparum* DHFS-FPGS and other homologues indicate low similarities (~30%), and even lower identities (~17%) (Lee *et al.*, 2001). Given this information, the possibility of designing selective drugs that are not detrimental to the human host, are increased. The predicted molecular mass of *P. falciparum* DHFS-FPGS is 56 kDa and the predicted pI 6.19 (http://us.expasy.org/cgi_bin/pi_tool). Apart from this limited information, nothing else of this enzyme is known. This chapter aims at gaining insight on certain predicted structural and genetic features of *P. falciparum* DHFS-FPGS to direct future mutagenesis experiments for determination of the enzymes' structure-function relationships.

4.2 Methods

4.2.1 Sequence alignments

DHFS and FPGS sequences were obtained for as many different organisms possible from the Swissprot and Swissall databases (http://srs.ebi.ac.uk). Monofunctional DHFS and FPGS as well as bifunctional DHFS-FPGS sequences were given in FASTA format for alignment by CLUSTAL X v 1.81 using the default gap penalty values (Thompson *et al.*, 1997). A phylogenetic tree based on this alignment was also obtained from CLUSTAL X. (ftp://ftp-igbmc.u-strasbg.fr/pub/**ClustalX**/). The sequences of various *Plasmodium* strains, the rodent parasite *Plasmodium yoelii*, as well as the less pathogenic human parasite *Plasmodium vivax*. were obtained by running a TBLASTN search (Altschul *et al.*, 1990) with the *P. falciparum* DHFS-FPGS sequence against *Plasmo*DB (http://PlasmoDB.org). The highest scoring hits ($E < 10^{-130}$) were aligned with the human host FPGS sequence and *P. falciparum* DHFS-FPGS sequences by CLUSTAL X to determine the interspecies sequence conservation as well as host-parasite differences.

4.2.2 Structure predictions

The secondary structure was predicted from the primary amino acid sequence with a range of independent computer programs: profile fed neural network systems (PHD) (Rost and Sander, 1993), GOR4 (Garnier *et al.*, 1996), hierarchical neural networks and SCRATCH Sspro (http://www.igb.uci.edu /tools/scratch/). A hydrophobicity profile for the primary amino acid sequence was obtained using Kyte and Doolittle parameters and a window size of 10 amino acids. An extensive search of the database PROSITE (Bairoch *et al.*, 1997) was run on the predicted primary amino acid sequence to identify possible amino acid sequence motifs involved in the activity or structure of the enzyme. Similarity searches also used the 3D-structure database PDB with the program SAM-T99 (Karplus *et al.*, 1998) (http://www.cse.ucsc.edu/research/compbio/HMM-apps/sam-t99) to identify proteins that could possibly share structural features with *P. falciparum* DHFS-FPGS. A Predict Protein search (http://www.embl-heidelberg.de/predictprotein/) was done, which is an extensive program searching various protein features such as the SEG low complexity sequences (Wootton and Federhen, 1996). A BLASTP search (Altschul *et al.*, 1990) of the ProDom database http://www.toulouse.inra.fr/prodom.html (Corpet *et al.*, 1999) was performed to

identify proteins with similar domains. A COILS version 2.2 search (Lupas, 1996) was performed to identify possible coiled-coil regions within the protein. To predict the solvent accessibility of amino acids and thus overall protein globularity, a GLOBE search was done (http://www.columbia.edu/~rost/Papers/98globe). Ten preliminary models of the *P. falciparum* DHFS-FPGS enzyme, based on the *L. casei* crystal structure was built with the program MODELER version 6v2 (Sali and Blundell, 1993) (Accelrys ®) using default parameters and visualised with the INSIGHT II program (Accelrys ®) (www.accelrys.com/insight/Modeler). Ramachandran plots for the 10 models were generated with PROCHECK (Morris *et al.*, 1992).

4.3 Results

4.3.1 Inter-species DHFS and FPGS alignments and phylogenetic analysis

Since no distinction between *P. falciparum* DHFS and FPGS domains can be made at the primary amino acid level, the predicted bifunctional sequence was aligned with homologous monofunctional DHFS or FPGS or bifunctional DHFS-FPGS enzymes (Figure 4.4). Sequences were obtained for prokaryotes: monofunctional *Streptococcus pneumoniae* DHFS, *Bacillus subtillus, Lactobacillus casei* and *Haemophilus influenzeae* FPGS and the bifunctional DHFS-FPGS of *Escherichia coli, Buchnera aphidicola* and *Buchnera biphidocola* as well as eukaryotes: yeast DHFS and FPGS, *Candida albicans*, *Arabidopsis thaliana*, human and mouse FPGS. A phylogenetic tree drawn from the alignment shows as expected that mammal (mouse and human) FPGS group together. Prokaryotic bifunctional DHFS-FPGS group together while prokaryotic monofunctional FPGS group together separately. *P. falciparum* DHFS-FPGS seems to be the closest related to prokaryotic bifunctional DHFS-FPGS. Yeast DHFS does not group with any other homologue and *Arabidopsis* and *Neurospora crassa* also group separately from the rest (Figure 4.3).

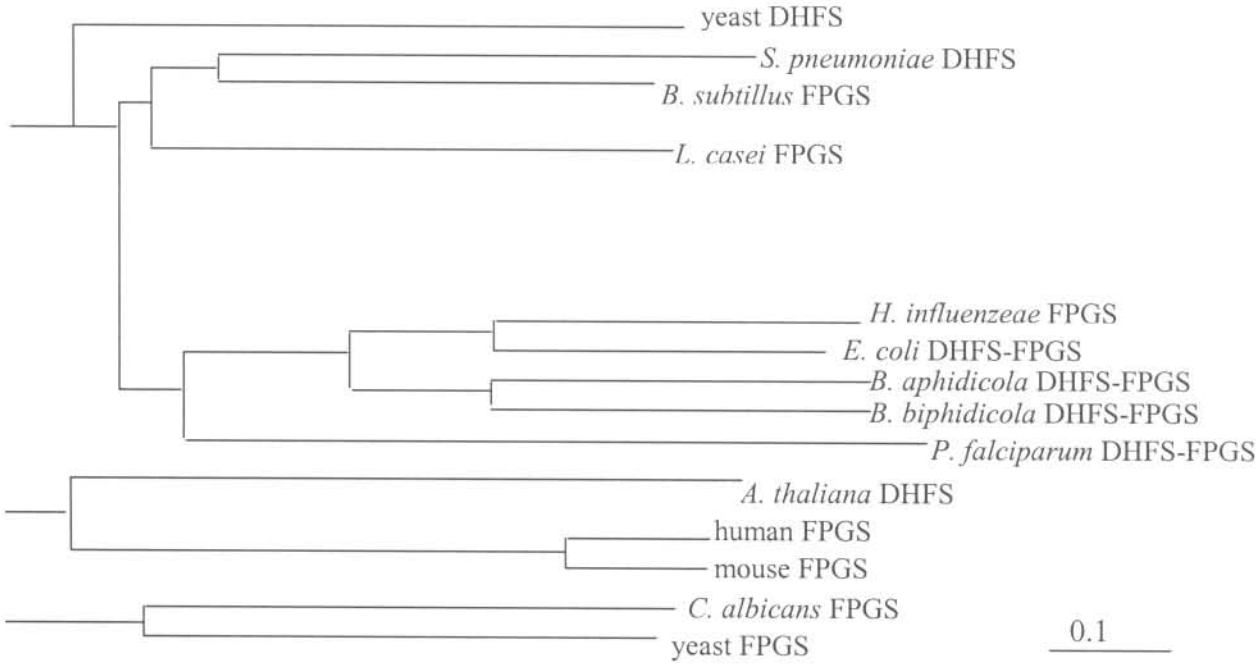
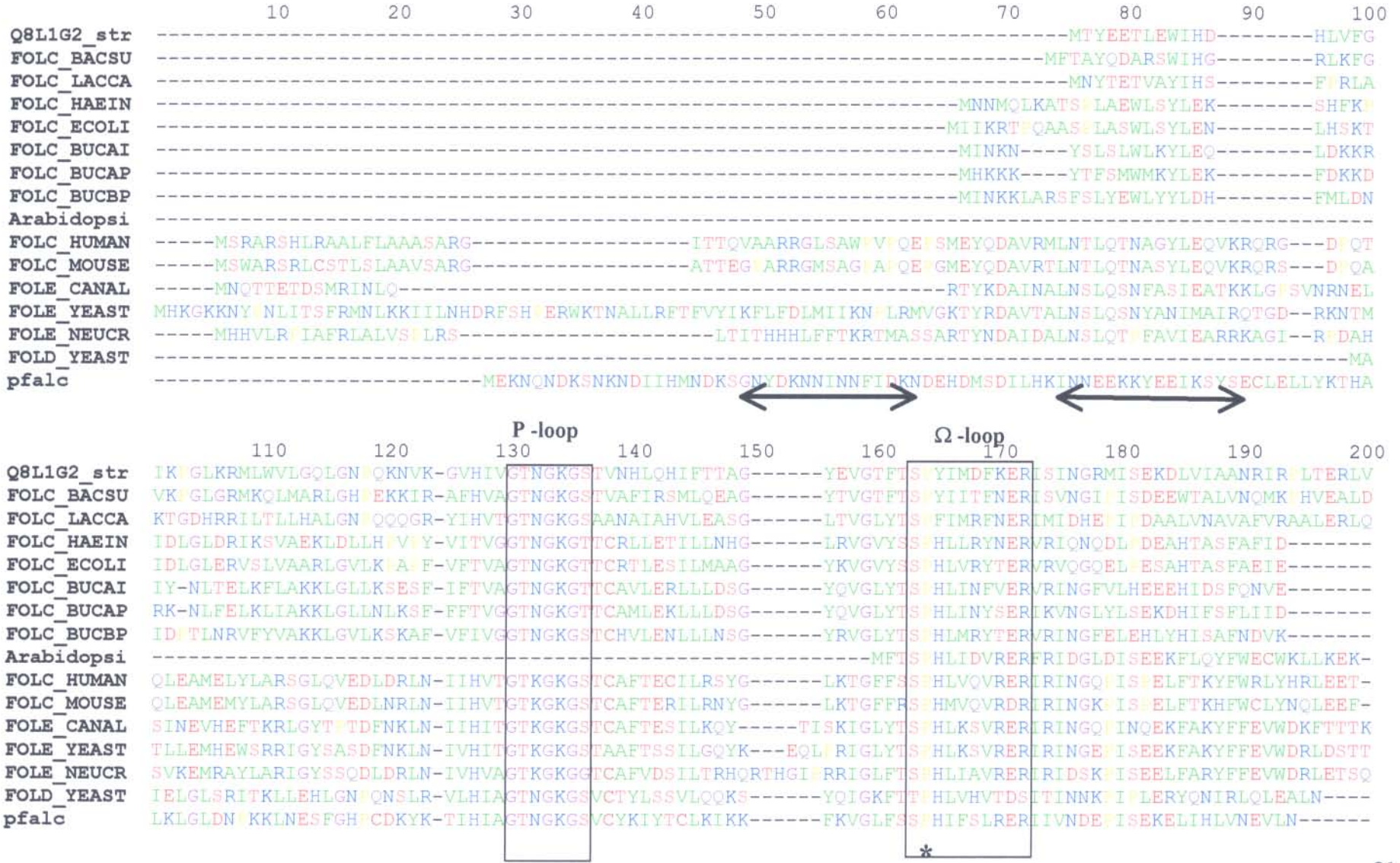


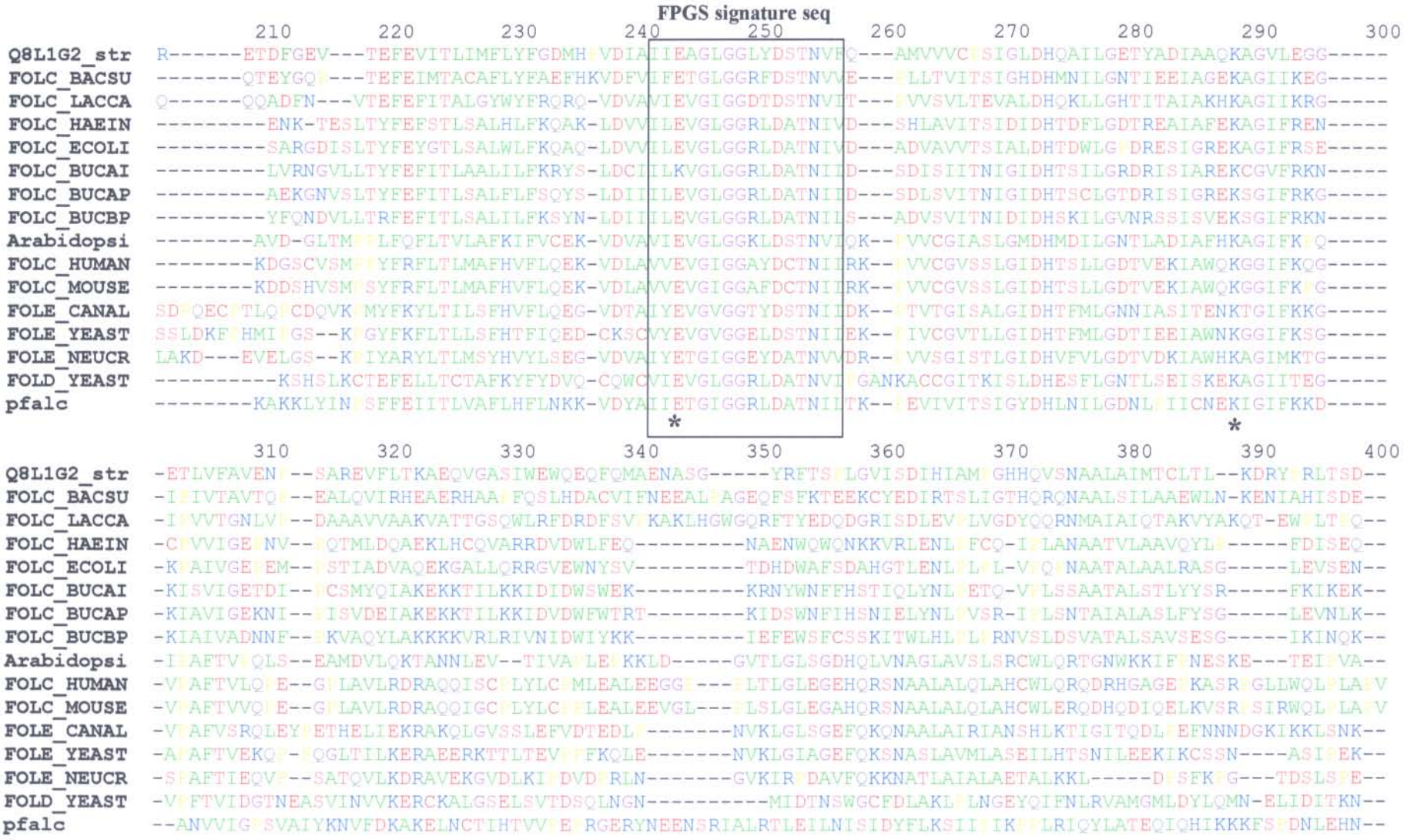
Figure 4.3: Phylogenetic analysis of the *P. falciparum* DHFS-FPGS protein based on its alignment with homologous proteins. The scale line indicates the evolutionary distance in arbitrary units to compare relative distances between proteins.

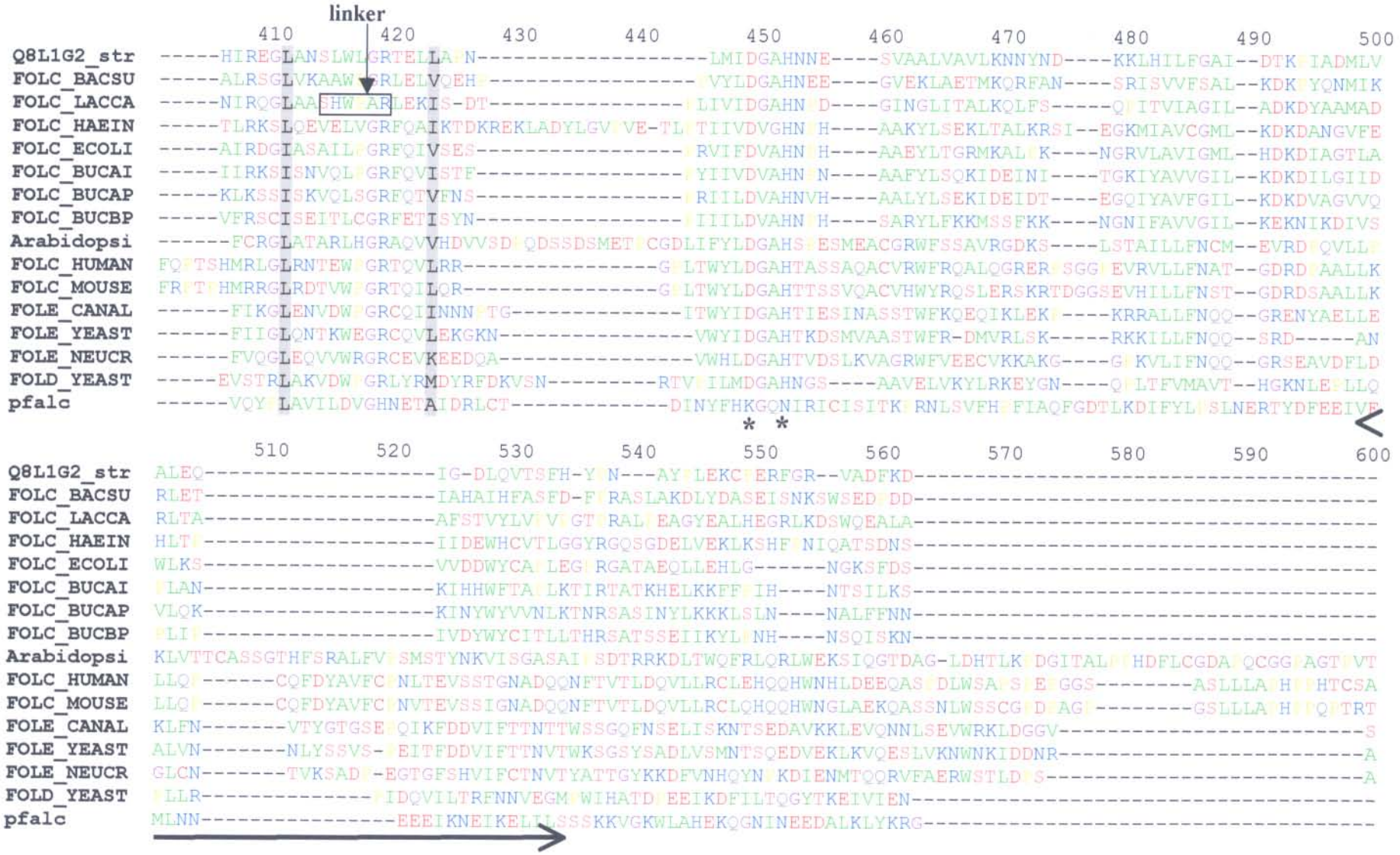
From the alignment, certain conserved features were verified within all the homologues, such as the ATP-binding P-loop "GTNGKGX" (Figure 4.4, 130-138), interdomain stabilising Ω-loop (Figure 4.4, 163-173) and FPGS signature sequence (Figure 4.4, 241-256). Interestingly, the FPGS signature sequence is highly conserved, even for monofunctional DHFS enzymes, such as that of *Streptococcus* and *A. thaliana.*

Overall, the N-terminal half of the sequence seems to be much more conserved between species than the C-terminal half as is verified by a superimposed homology model of *P. falciparum* DHFS-FPGS (red ribbons) on the *L. casei* FPGS crystal structure (green ribbons) (Figure 4.5). The yellow loop indicates the *P. falciparum* sequence that corresponds with the linker area in the alignment. N- and C-terminal halves are assigned according to the position of the *L. casei* linker hexapeptide region (Figure 4.4, 414-419) between the N-and C-terminal domains as determined by X-ray crystallography (Sun *et al.*, 1998).

The amino acid composition of the linker region itself appears not to be conserved, but two conserved hydrophobic residues are indicated by grey highlights on opposite sides of the linker region and the carbon atom backbone structure is conserved (Figure 4.5, yellow loop). The first 40 amino acids of the *P. falciparum* sequence does not align (Figure 4.4; Figure 4.5, red unstructured loop projecting from the N-terminal domain) with the homologous sequences and 20 of these amino acids also coincide with a low complexity region (Figure 4.4, black line at position 48-65). Stars throughout the alignment indicate other essential residues that cause a loss in function when mutated in *L. casei* FPGS (Figure 4.4) (Sheng *et al.*, 2000). These are well conserved between the different homologues and are also present in *P. falciparum,* except for D449 and H552 in the C-terminal region (Figure 4.4; Figure 4.5 yellow and blue ball and stick structures). The Ramachandran plot for the second homology model of ten (Figure 4.6) shows that only 4 residues (0.8% of the total), fall within disallowed areas.


10 20 30 40 50 60 70 80 90 100
Q8L1G2_str
FOLC_BACSU
FOLC_LACCA
FOLC_HAEIN
FOLC_ECOLI
FOLC_BUCAI
FOLC_BUCAP
FOLC_BUCBP
Arabidopsi
FOLC_HUMAN
FOLC_MOUSE
FOLE_CANAL
FOLE_YEAST
FOLE_NEUCR
FOLD_YEAST
pfalc
P -loop
Ω -loop
110 120 130 140 150 160 170 180 190 200
Q8L1G2_str
FOLC_BACSU
FOLC_LACCA
FOLC_HAEIN
FOLC_ECOLI
FOLC_BUCAI
FOLC_BUCAP
FOLC_BUCBP
Arabidopsi
FOLC_HUMAN
FOLC_MOUSE
FOLE_CANAL
FOLE_YEAST
FOLE_NEUCR
FOLD_YEAST
pfalc
*


FPGS signature seq
210 220 230 240 250 260 270 280 290 300
Q8L1G2_str R-------ETDFGEV---TEFEVITLIMFLYFGDMH VDIAIIEAGLGGLYDSTNVF Q---AMVVVC SIGLDHQAILGETYADIAAQKAGVLEGG-----
FOLC_BACSU -------QTEYGQ ---TEFEIMTACAFLYFAEFHKVDFV IFETGLGGRFDSTNVVE --- LLTVITSIGHDHMNILGNTIEEIAGEKAGIIKEG-----
FOLC_LACCA Q-------QQADFN---VTEFEFITALGYWYFRQRQ-VDVA VIEVGIGGDTDSTNVIT --- VVSVLTEVALDHQKLLGHTITAIAKHKAGIIKRG-----
FOLC_HAEIN ---------ENK-TESLTYFEFSTLSALHLFKQAK-LDVV ILEVGLGGRLDATNIVD ---SHLAVITSIDIDHTDFLGDTREAIAFEKAGIFREN-----
FOLC_ECOLI ---------SARGDISLTYFEYGTLSALWLFKQAQ-LDVV ILEVGLGGRLDATNIVD ---ADVAVVTSIALDHTDWLG DRESIGREKAGIFRSE-----
FOLC_BUCAI ---------LVRNGVLLTYFEFITLAALILFKRYS-LDCI ILKVGLGGRLDATNIID ---SDISIITNIGIDHTSILGRDRISIAREKCGVFRKN-----
FOLC_BUCAP ---------AEKGNVSLTYFEFITLSALFLFSQYS-LDII ILEVGLGGRLDATNIID ---SDLSVITNIGIDHTSCLGTDRISIGREKSGIFRKG-----
FOLC_BUCBP ---------YFQNDVLLTRFEFITLSALILFKSYN-LDII ILEVGLGGRLDATNILS ---ADVSVITNIDIDHSKILGVNRSSISVEKSGIFRKN-----
Arabidopsi --------AVD-GLTM LFQFLTVLAFKIFVCEK-VDVA VIEVGLGGKLDSTNVIQK-- VVCGIASLGMDHMDILGNTLADIAFHKAGIFK Q-----
FOLC_HUMAN --------KDGSCVSM YFRFLTLMAFHVFLQEK-VDLA VVEVGIGGAYDCTNIIRK-- VVCGVSSLGIDHTSLLGDTVEKIAWQKGGIFKQG-----
FOLC_MOUSE --------KDDSHVSM SYFRFLTLMAFHVFLQEK-VDLA VVEVGIGGAFDCTNIIRK-- VVCGVSSLGIDHTSLLGDTVEKIAWQKGGIFK G-----
FOLE_CANAL SD QEC TLQ CDQVK MYFKYLTILSFHVFLQEG-VDTA IYEVGVGGTYDSTNIIDK-- TVTGISALGIDHTFMLGNNIASITENKTGIFKKG-----
FOLE_YEAST SSLDKF HMI GS--K GYFKFLTLLSFHTFIQED-CKSC VYEVGVGGELDSTNIIEK-- IVCGVTLLGIDHTFMLGDTIEEIAWNKGGIFKSG-----
FOLE_NEUCR LAKD---EVELGS--K IYARYLTLMSYHVYLSEG-VDVA IYETGIGGEYDATNVVDR-- VVSGISTLGIDHVFVLGDTVDKIAWHKAGIMKTG-----
FOLD_YEAST ----------KSHSLKCTEFELLTCTAFKYFYDVQ-CQWC VIEVGLGGRLDATNVI GANKACCGITKISLDHESFLGNTLSEISKEKAGIITEG-----
pfalc --------KAKKLYIN SFFEIITLVAFLHFLNKK-VDYA IIETGIGGRLDATNIITK-- EVIVITSIGYDHLNILGDNL IICNEKIGIFKKD-----
310 320 330 340 350 360 370 380 390 400
Q8L1G2_str -ETLVFAVEN --SAREVFLTKAEQVGASIWEWQEQFQMAENASG-----YRFTS LGVISDIHIAM GHHQVSNAALAIMTCLTL--KDRY RLTSD--
FOLC_BACSU -I IVTAVTQ --EALQVIRHEAERHAA FQSLHDACVIFNEEAL AGEQFSFKTEEKCYEDIRTSLIGTHQRQNAALSILAAEWLN-KENIAHISDE--
FOLC_LACCA -I VVTGNLV --DAAAVVAAKVATTGSQWLRFDRDFSV KAKLHGWGQRFTYEDQDGRISDLEV LVGDYQQRNMAIAIQTAKVYAKQT-EW LT Q--
FOLC_HAEIN -C VVIGE NV-- QTMLDQAEKLHCQVARRDVDWLFEQ---------NAENWQWQNKKVRLENL FCQ-I LANAATVLAAVQYL -----FDISEQ--
FOLC_ECOLI -K AIVGE EM-- STIADVAQEKGALLQRRGVEWNYSV---------TDHDWAFSDAHGTLENL L L-V Q NAATALAALRASG-----LEVSEN--
FOLC_BUCAI -KISVIGETDI-- CSMYQIAKEKKTILKKIDIDWSWEK---------KRNYWNFFHSTIQLYNL ETQ-V LSSAATALSTLYYSR-----FKIKEK--
FOLC_BUCAP -KIAVIGEKNI-- ISVDEIAKEKKTILKKIDVDWFWTRT--------KIDSWNFIHSNIELYNL VSR-I LSNTAIALASLFYSG-----LEVNLK--
FOLC_BUCBP -KIAIVADNNF-- KVAQYLAKKKKVRLRIVNIDWIYKK---------IEFEWSFCSSKITWLHL L RNVSLDSVATALSAVSESG-----IKINQK--
Arabidopsi -I AFTV QLS--EAMDVLQKTANNLEV--TIVA LE KKLD-----GVTLGLSGDHQLVNAGLAVSLSRCWLQRTGNWKKIF NESKE---TEI VA--
FOLC_HUMAN -V AFTVLQ E--G LAVLRDRAQQISC LYLC MLEALEEGG --- LTLGLEGEHQRSNAALALQLAHCWLQRQDRHGAGE KASR GLLWQL LA V
FOLC_MOUSE -V AFTVVQ E--G LAVLRDRAQQIGC LYLC LEALEEVGL--- LSLGLEGAHQRSNAALALQLAHCWLERQDHQDIQELKVSR SIRWQL LA V
FOLE_CANAL -V AFVSRQLEY ETHELIEKRAKQLGVSSLEFVDTEDL -------NVKLGLSGEFQKQNAALAIRIANSHLKTIGITQDL EFNNNDGKIKKLSNK--
FOLE_YEAST -A AFTVEKQ - QGLTILKERAEERKTTLTEV FKQLE-------NVKLGIAGEFQKSNASLAVMLASEILHTSNILEEKIKCSSN----ASI EK--
FOLE_NEUCR -S AFTIEQV --SATQVLKDRAVEKGVDLKI DVD RLN-------GVKIR DAVFQKKNATLAIALAETALKKL-----D SFK G---TDSLS E--
FOLD_YEAST -V FTVIDGTNEASVINVVKERCKALGSELSVTDSQLNGN----------MIDTNSWGCFDLAKL LNGEYQIFNLRVAMGMLDYLQMN-ELIDITKN--
pfalc --ANVVIG SVAIYKNVFDKAKELNCTIHTVV E RGERYNEENSRIALRTLEILNISIDYFLKSII IK LRIQYLATEQIQHIKKKFS DNLEHN--


linker
410 420 430 440 450 460 470 480 490 500
Q8L1G2_str
FOLC_BACSU
FOLC_LACCA
FOLC_HAEIN
FOLC_ECOLI
FOLC_BUCAI
FOLC_BUCAP
FOLC_BUCBP
Arabidopsi
FOLC_HUMAN
FOLC_MOUSE
FOLE_CANAL
FOLE_YEAST
FOLE_NEUCR
FOLD_YEAST
pfalc
* *
510 520 530 540 550 560 570 580 590 600
Q8L1G2_str
FOLC_BACSU
FOLC_LACCA
FOLC_HAEIN
FOLC_ECOLI
FOLC_BUCAI
FOLC_BUCAP
FOLC_BUCBP
Arabidopsi
FOLC_HUMAN
FOLC_MOUSE
FOLE_CANAL
FOLE_YEAST
FOLE_NEUCR
FOLD_YEAST
pfalc

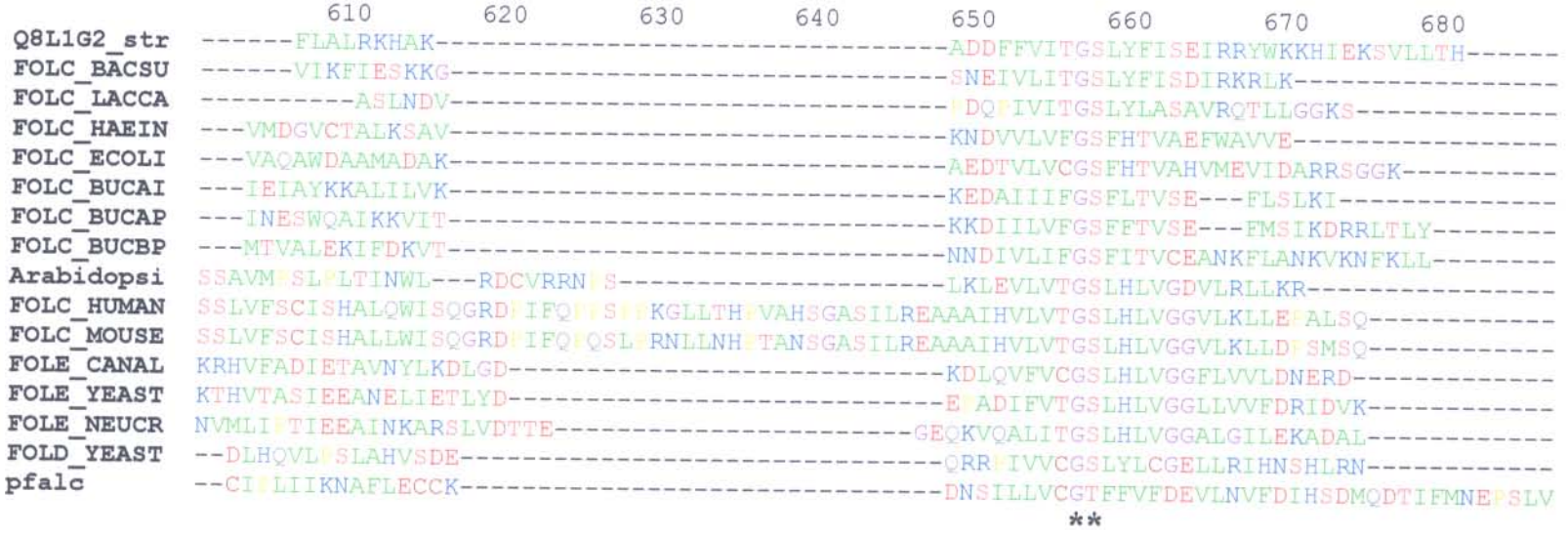


Figure 4.4: Alignment of *P. falciparum* DHFS-FPGS with homologous proteins. Q8L1G2_str: *Streptococcus* DHFS, FOLC_BACSU: *B. subtillus* DHFS-FPGS, FOLC_LACCA: *L. casei* FPGS, FOLC_HAEIN: *H. influenzeae* FPGS, FOLC_ ECOLI: *E. coli* DHFS-FPGS, FOLC_BUCAP: *B. aphidicola* DHFS-FPGS, FOLC_BUCBP: *B. biphidicola* DHFS-FPGS, Arabidopsi: *A. thaliana* FPGS, FOLC_HUMAN: human FPGS mitochondrial precursor, FOLC_MOUSE: mouse FPGS mitochondrial precursor, FOLE_CANAL: *C. albicans* FPGS, FOLE_YEAST: yeast FPGS, FOLE_NEUCR: *N. crassa* FPGS, FOLD_YEAST: yeast DHFS. p.falc: *P. falciparum* DHFS-FPGS. Thick black lines indicate areas of low complexity in the *P. falciparum* sequence. Stars indicate essential residues in the *L. casei* enzyme. Boxed areas indicate the P-loop, Ω-loop, FPGS signature sequence and *L. casei* linker features. The grey highlighted residues indicate conserved hydrophobic residues flanking the linker region of *L. casei.* Neutral amino acids are coloured green, negatively charged amino acids are coloured different shades of red according to pKa values, positively charged amino acids are coloured shades of blue according to pKa values, proline and glycine normally found in turns are coloured yellow and purple respectively.

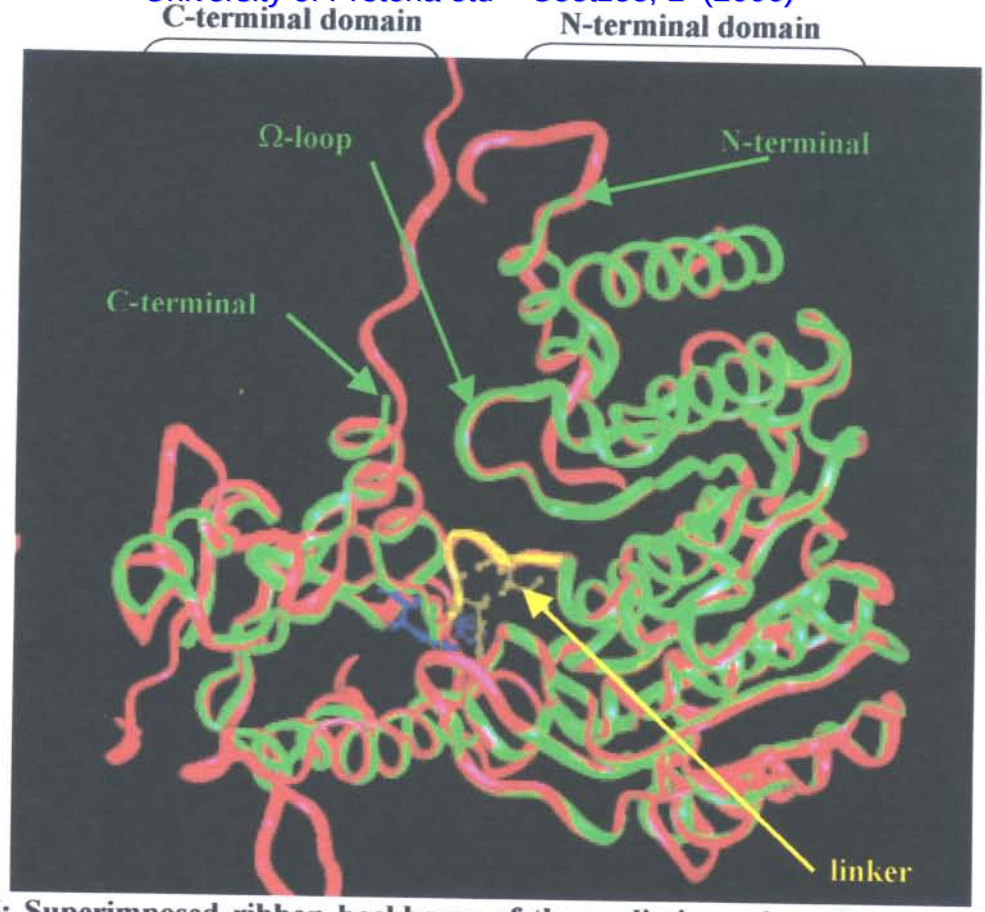


Figure 4.5: Superimposed ribbon backbones of the preliminary homology model of *P. falciparum* DHFS-FPGS (red) on the *L. casei* FPGS (green) crystal structure. The aligned Ω-loop, P-loop and linker area (yellow loop) are indicated with green arrows. Catalytic residues (D449 and H552) that do not align are indicated in blue ball and stick structures (*L. casei*) and yellow ball and stick structures (*P. falciparum*).

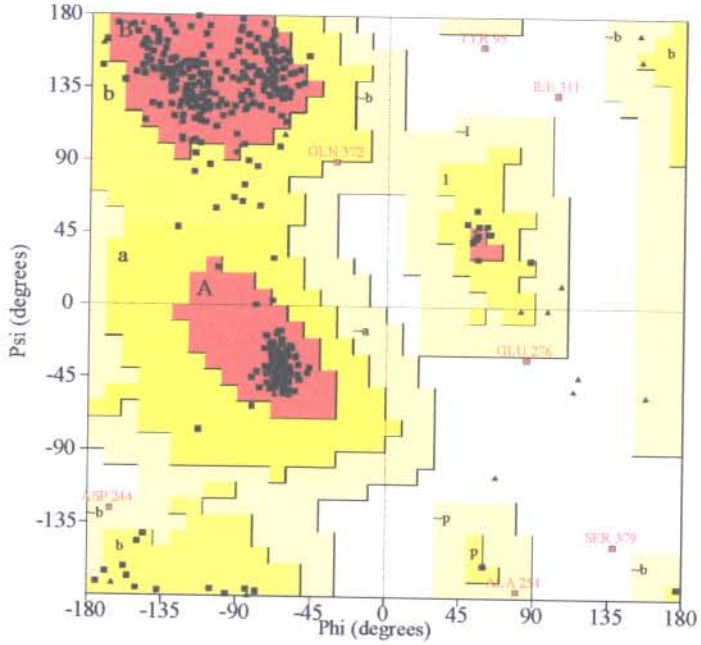


Figure 4.6: Ramachandran plot of the *P. falciparum* DHFS-FPGS homology model. Residues in most favoured regions [red: A, B, L], residues in additional allowed regions [yellow: a, b, l, p], residues in generously allowed regions [beige: ~a,~b,~l,~p] and residues in disallowed regions [white: Y95, E276, I311 and S379].

4.3.2 Conservation of DHFS-FPGS within the *Plasmodium* species and comparison with human FPGS.

The TBLASTN search of *Plasmo*DB using *P. falciparum* DHFS-FPGS as the query sequence, identified high scoring homologous gene sequences in *P. yoelii* ($E=4 \times 10^{-173}$) and *P. vivax* ($E=2.7 \times 10^{-130}$). Alignments of *P. falciparum* DHFS-FPGS with these *Plasmodium* sequences and human FPGS show ~10% identity and 24% similarity between the Plasmodium and human species (stars and :, Figure 4.7). The N-terminal extension also observed in Figure 4.5 and the alignment (Figure 4.4) seems to be specific to *P. falciparum*. An interesting feature is that the human enzyme has inserted sequences not found in the *Plasmodium* sequences (Figure 4.7, yellow highlighted sequences). Normally *Plasmodium* enzymes contain inserted sequences when compared to human homologues, which coincide with low complexity areas, (Pizzi and Frontali, 2001). This, however does not seem to be the case for DHFS-FPGS, where the human homologue, FPGS contains the majority of inserted sequences.

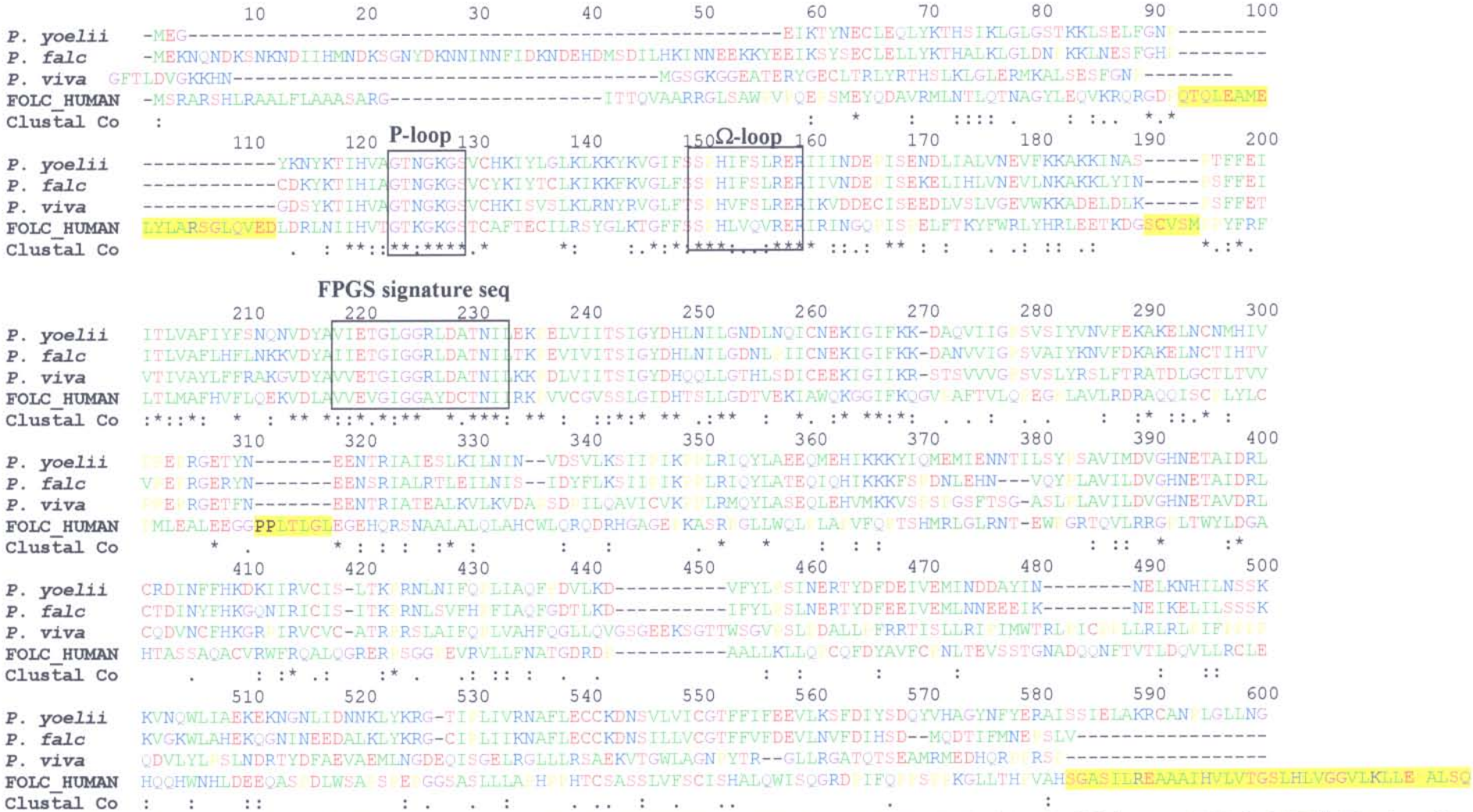


Figure 4.7: *Plasmodium* DHFS-FPGS vs Human FPGS alignment: *P. yoelii* DHFS-FPGS, *P. falc*: *P. falciparum* DHFS-FPGS, P. viva: *P. vivax* DHFS-FPGS, FOLC_HUMAN: human FPGS. Blocks indicate key enzyme features such as the P-loop, Ω-loop and FPGS signature sequence. Clustal Co= conserved sequence; stars indicate conserved residues (identities) and (:) similarities. Human inserted sequences are highlighted in yellow.

4.3.3 Secondary structure prediction of *P. falciparum* DHFS-FPGS

The primary amino acid sequence was analysed with a variety of programs for the prediction of secondary structure elements. According to the GOR4 program, most of the sequence consists of α-helices and random coils. α-Helices are predominantly found at positions 50-70, 150-180, 290-310 and 430-450 (Figure 4.8, blue lines). Random coils are predominant at positions 0-40, 280-290, 310-320, 380-390 and 410-420 (Figure 4.8, purple lines). The N-terminal extension as identified through alignments is part of a random coil as expected, but the other two low complexity areas coincide with α-helices.

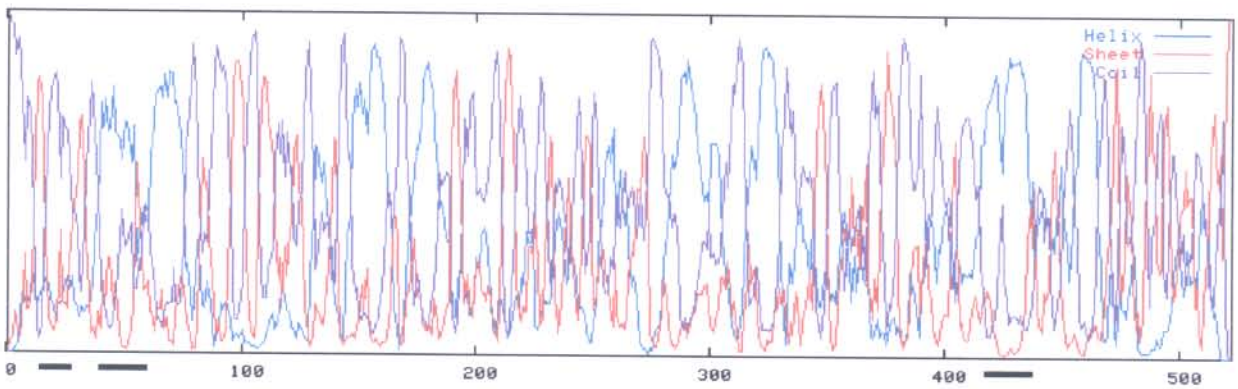


Figure 4.8: Secondary structure prediction of *P. falciparum* DHFS-FPGS with GOR4. A schematic representation indicating the possibility of each structural element (y axis) versus the position in the primary amino acid sequence (x axis). Black lines below the figure indicate the positions of low complexity sequence.

Alignments between the HNN, SCRATCH, GOR4 and PHD secondary structure predictions show mostly the same secondary structures (Figure 4.9). At a closer look it is observed that the two low complexity sequence areas consist of mixed helix-coil structures in contrast to the random coil observed for the N-terminal extension (Figure 4.9, black underlining). Taken on average of the different prediction methods the percentage of secondary structure is predicted to be 46% α-helices, 11.9% extended β sheets and 42.1 % random coils. Given that the requirements for the classification of proteins according to secondary structure are: all-alpha (%Helices>45 and % Extended beta sheets<5), all-beta (%Helices<5 and %Extended beta sheets>45) and alpha-beta (%Helices>30 and %Extended beta sheets > 20), DHFS-FPGS is predicted as a mixed class protein since the predicted secondary structures do not fall in the above categories. No coiled–coils were predicted by the program COIL and the overall topology was predicted to be compact and globular with 199 exposed residues by the program GLOBE.

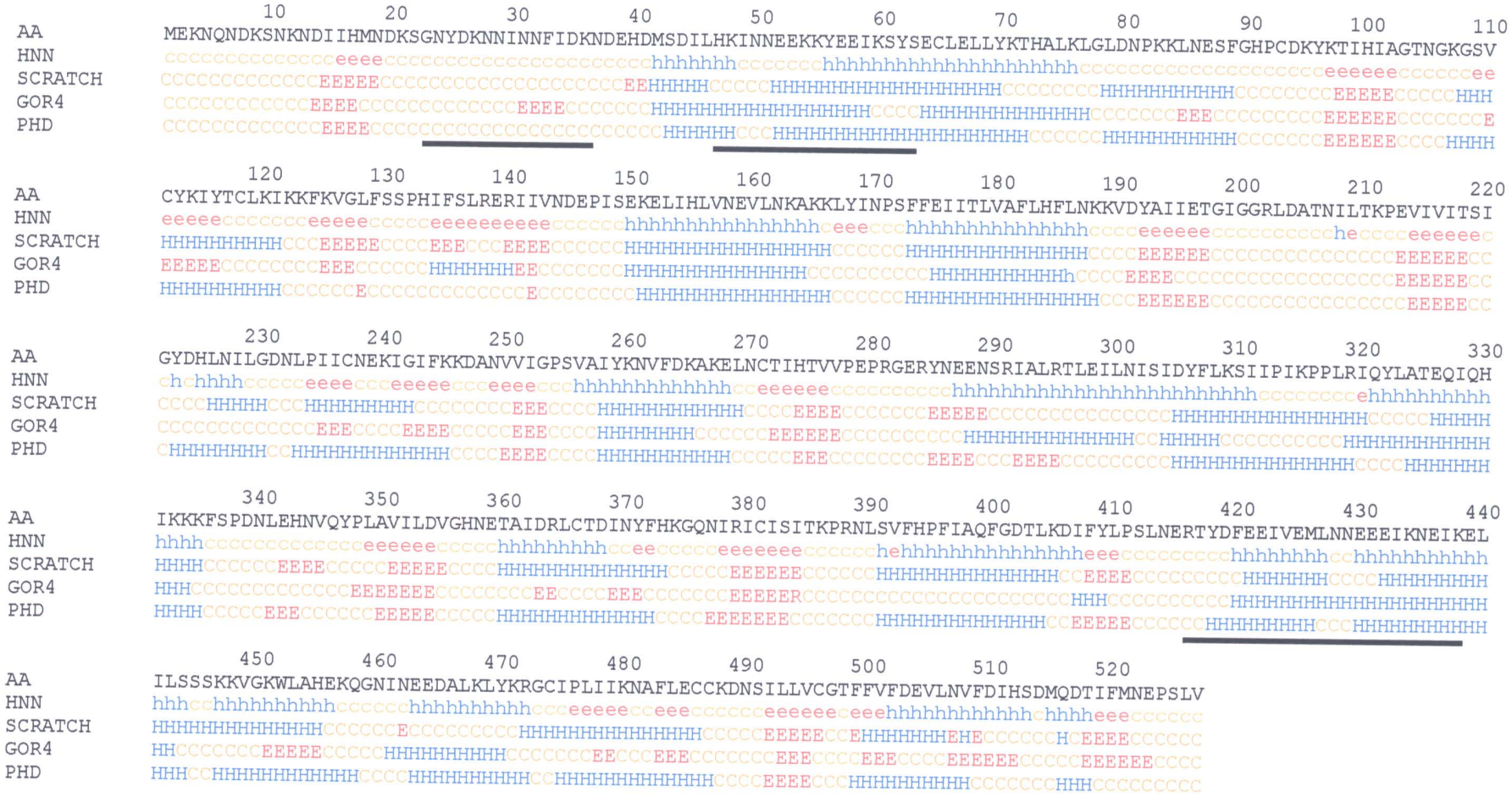


Figure 4.9: Alignment of independent secondary structure predictions of *P. falciparum* DHFS-FPGS from hierarchial neural networks (HNN), SCRATCH Sspro GOR4 and profile fed neural networks (PHD) based on the primary amino acid sequence (AA) indicated in black letters. Orange: random coil (c), red: extended β sheet (e) and blue letters: α-helix (H). The random coil observed in the first 40 amino acids corresponds with the inserted sequence that does not align with other DHFS of FPGS homologues. Black lines below the alignments indicate the positions of low complexity sequence

4.3.4 DHFS-FPGS hydrophobicity profile

The protein is predicted to be neither predominantly hydrophobic nor hydrophilic. The 40 amino acid N-terminal extension corresponds with the most hydrophilic area within the protein (Figure 4.10). Position 280-290, corresponding to a random coil is also very hydrophilic. Hydrophobic areas correspond mostly with isolated β-sheets. It is interesting that the omega loop, of which the primary function is interdomain stabilisation through hydrophobic interactions, does not occur in this plot at one of the hydrophobic maxima. In fact not one of the enzyme features seem to be hydrophobic, except for the conserved hydrophobic residues flanking the *L. casei* linker area (Figure 4.10). The other hydrophobic maxima are distributed between important catalytic sites or conserved motifs, rather than corresponding directly with these features and it may be that hydrophobic interactions between the hydrophobic maxima are required for shaping of the active site pockets or general stabilisation of enzyme structural features.

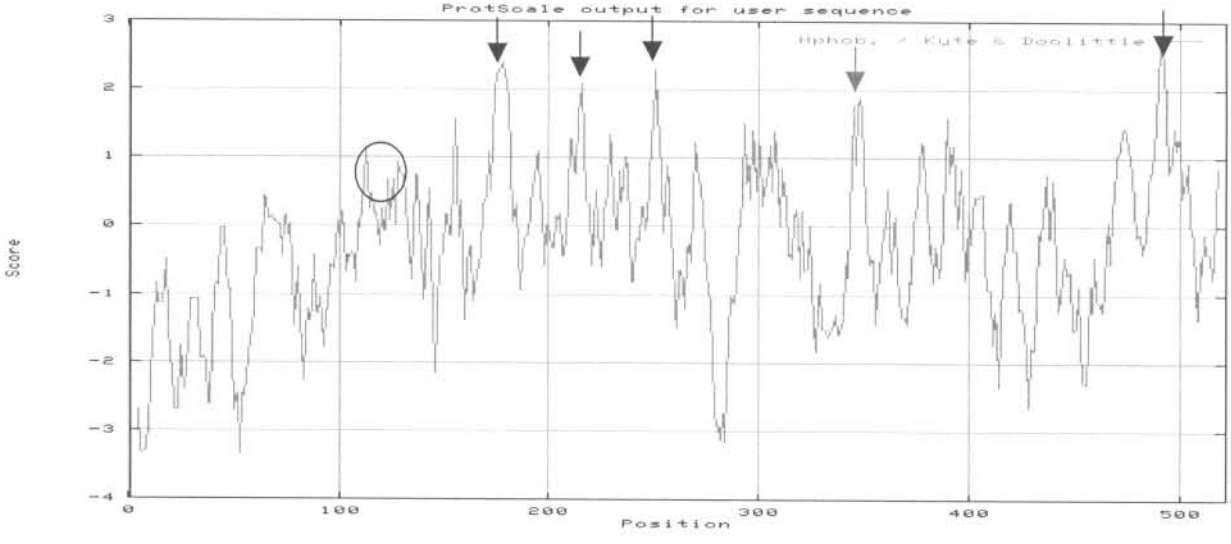


Figure 4.10: Hydrophobicity profile for the primary amino acid sequence of *P. falciparum* DHFS-FPGS using Kyte and Doolittle parameters. Hydrophobic maxima are indicated with arrows. The position of the omega loop is encircled

4.4 Discussion

4.4.1 Sequence conservation of DHFS-FPGS

The alignment of the P-loop, Ω-loop and FPGS signature sequence corresponds to that reported by (Salcedo *et al.*, 2001), yet our alignment did not show the conservation of C-terminal residues, D449 and H 452 (Figure 4.4) involved in FPGS activity as shown for *L. casei* (Sheng et al., 2000). It might be that the *P. falciparum* enzyme uses additional residues for activity, or has a slightly altered conformation of its active site due to its bifunctionality. Alignments of other *Plasmodium* strains also fail to align these two residues with the human homologue, which might indicate species-specific functional differences. Another alternative might be that the D416 and H419 residues of *P. falciparum* DHFS-FPGS should be aligned 33 residues to the right (Salcedo *et al.*, 2001) with the conserved D449 and H452 (Figure 4.4), but this would introduce a gap in the alignment around the linker area, which is an important structural feature and is shown by the homology model that the structure of the linker area remains conserved (Figure 4.5).

All the other catalytic residues align well and although the amino acid composition for the linker area is not conserved for *P. falciparum* DHFS-FPGS, it is structurally conserved in terms of its backbone structure and fold (Figure 4.5). The N-terminal is much more conserved than the C-terminal for all species, which is also shown through the superimposition of the *P. falciparum* homology model with the crystal structure. The *P. falciparum* enzyme, however, contains an N-terminal extension, which consists of a low complexity sequence and is a parasite-specific trait. Alignments with other *Plasmodium* strains indicate that this feature is unique to *P. falciparum* and might be implicated in the pathogenicity of the strain.

The phylogenetic tree deduced from the alignment shows as expected separate pro- and eukaryotic groups. This also confirms that the alignment is a good representation of the homologues. Interestingly, the *P. falciparum* sequence is closer related to the prokaryotic bifunctional enzymes than the eukaryotic enzymes. This is difficult to evaluate because no other bifunctional eukaryotic sequences are known at this stage and it may be that the bifunctionality forces the grouping of the *P. falciparum* enzyme towards other bifunctional enzymes. Alignment with the human homologue indicate low sequence similarity (24%) and

identity (10%) comparable to previous literature reports of 30% similarity and 15% identity (Lee *et al.*, 2001).

4.4.2 Predicted secondary structure

The overall topology of *P. falciparum* DHFS-FPGS is predicted to be globular, with almost half the residues exposed. The protein is classified as mixed class since there are almost equal numbers of α-helices and random coils predicted (~45%) and only a few extended β-sheets (~10%). The characteristic N-terminal extension coincides with a random coil as expected for low complexity sequences (Pizzi and Frontali, 2001), yet the other low complexity sequences correspond to mixed helix-random coil structures. The N-terminal extension furthermore coincides with the most hydrophilic portion of the enzyme as predicted by the Kyte and Doolittle hydrophobicity plot. Taken together with all the other information on this parasite-specific feature, the N-terminal extension as well as the divergent C-terminal domain might be interesting features for structural investigation.

Chapter 5
Concluding Discussion

5.1 'Discover, Develop, Deliver'

In the light of the decreasing effectivity of mosquito control and the absence of a safe and effective vaccine, antimalarial chemotherapy remains the first line of defence against malaria. Of the different antimalarial control strategies, the use of drugs is furthermore the only one that functions on two levels: prevention and cure. Over the past few decades, malaria has become predominantly an African problem due to neglected malaria control strategies and poor infrastructure, emphasising the dire need for cheap, effective and safe drugs that are readily available (www.rbm.int). The Medicines for Malaria Venture (MMV) slogan: '*discover, develop, deliver*' aptly describes the key factors involved in achieving this. The scope of this study lies within the 'discover' aspect: the evaluation of newly identified potential drug targets.

5.2 Antifolates: New targets, old pathway

In this study the 'discover' aspect is applied to the folate pathway, since it is a verified antimalarial drug target and current antifolates are losing their effectivity due to parasite resistance. Since the presence of folate cofactors is essential in all organisms, the antifolates are also widely used in cancer- and bacterial chemotherapy and thus forms a class of well-studied compounds (Moran, 1999). Antifolates also were among the first rationally designed chemotherapeutic compounds (Costi and Ferrari, 2001). The potential of this pathway for drug inhibition of *P. falciparum* has, however, been largely underexploited since only two enzymes of the pathway, DHPS and DHFR are proven drug targets. Resistance caused by mutations in both these drug targets greatly compromises the effectivity of current antifolates. It would be possible to rationally design novel antifolates against these well studied drug targets to improve the inhibitor efficiency, but it has been shown in the past that novel drugs and drug combinations against DHFR and DHPS only select for more dangerous and more

resistant strains carrying multiple mutations (Sibley *et al.*, 2001). Considering that certain proteins are more prone to obtaining resistance-causing mutations, the development of novel drugs targeted to other folate enzymes would circumvent this problem. The identification of three new *P. falciparum* genes encoding the folate biosynthetic proteins GTP-CH, SHMT and the bifunctional DHFS-FPGS, was thus an important event that laid the foundation for future drug target validation and drug discovery (Lee *et al.* 2001).

5.3 *P. falciparum* DHFS-FPGS: an attractive drug target

Of the different novel genes identified for folate-metabolism, the gene encoding DHFS-FPGS activity seems to be the most promising drug target. When compared to human FPGS, the DHFS-FPGS enzyme is an attractive drug target for selective inhibition due to the additional DHFS activity, low similarity and its bifunctionality, which is unique to the malaria parasite. Furthermore, these two enzymes function in different routes of folate metabolism; *de novo* folate synthesis (which can be bypassed in certain *P. falciparum* isolates by high serum folate concentrations) and folate salvage (refer to section 1.5.4). Furthermore it was shown in a functional complementation study that the current antifolates, which are structural analogs of the natural substrates don't select for resistant phenotypes of DHFS-FPGS as measured by mutations in the corresponding gene (Salcedo *et al.*, 2001). As shown in Chapter 4, the *P. falciparum* enzyme furthermore has characteristic features such as a parasite specific N-terminal extension, a divergent C-terminal domain and possibly different C-terminal catalytic residues, which might account for structural and functional differences and thus play an important role in the selective inhibition of the parasite enzyme. To properly evaluate the potential of DHFS-FPGS as a drug target, however, enzyme kinetics and structural studies need to be done, which require large amounts of active enzyme.

5.4 The aims of this study: obtaining sufficient amounts of active *P. falciparum* DHFS-FPGS

Malaria parasite genes are notorious for the problems involved in heterologous expression, brought about by the high A+T content and codon bias of the malaria genome, which results in low levels of gene expression (Baca and Hol, 2000). To address this problem, *P.*

falciparum dhfs-fpgs was synthesised as described in Chapter 2 with preferred *E. coli* codons and lower A+T content for high-level recombinant expression in *E. coli*. The method used for synthesis, the overlap-extension PCR of partially complementary oligonucleotides, was optimised and correct sequences could be obtained from as little as 30 total PCR cycles (assembly and amplification) of 1 pmole of partially overlapping oligonucleotides, in comparison with previous reports using 125 pmoles of complete overlapping oligonucleotides for a total of 50 PCR cycles (Carpenter *et al.*, 1999). The method used in this study therefore not only cut the costs of the gene synthesis with ~20% but also increased the efficiency of the PCR method. The optimisation revealed that crucial factors for obtaining the correct sequences were the quality of the oligonucleotides, the use of additional Mg^{2+} during the assembly step for efficient annealing of the overlaps, the use of a proofreading DNA polymerase as well as the use of a two-step PCR protocol; consisting of a limited number of assembly PCR cycles followed by dilution of the assembly mixture and primer addition for the specific amplification of the gene fragment. One of every three clones isolated contained the correct sequence. The full-length gene was also successfully constructed from ligation and further PCR of the correct subfragments.

This codon-optimised gene was then expressed in a variety of different vector systems and *E. coli* expression hosts to obtain sufficient amounts of active protein as described in Chapter 3. High-level gene expression in fast-growing BL21 Star (DE3) cells became toxic to the cells and the protein was produced in inclusion bodies. DHFS-FPGS is normally expressed in very low native levels, ~4% of the total proteins expressed by *E. coli* (Bognar *et al.*, 1985) and overexpression of human FPGS results in cancer (Osborne *et al.*, 1993), which might explain the insolubility observed with the high level expression of the synthetic gene. Lower expression levels in slower growing BL21 (DE3) pLysS cells produced lower amounts of total protein but more soluble protein. It was also found that the removal of a tag decreased total expression but increased soluble expression. By means of functional complementation it was shown that all the different synthetic DHFS-FPGS constructs contained DHFS and FPGS activity and complemented the deficient *E. coli* cell line to a greater extent than the native *P. falciparum* DHFS-FPGS.

The primary aims of the thesis was thus achieved by:

- the successful synthesis of the codon-optimised *P. falciparum dhfs-fpgs* and

- expression of active DHFS-FPGS from the synthetic gene at higher levels than the native gene.

Furthermore, functional complementation indicated that the tagless and C-terminal His_6-tagged DHFS-FPGS had the most DHFS and FPGS activity and preliminary purification consisting of affinity chromatography and size exclusion HPLC of the C-terminal His_6-tagged DHFS-FPGS proteins was performed as groundwork for future purification strategies.

The *in silico* analysis of DHFS-FPGS also identified possible parasite specific features, such as a 40 amino acid N-terminal extension consisting of low complexity sequence and a low conserved C-terminal domain. Mutagenesis strategies aimed at the cassette replacement of these domains could provide valuable structural information in future.

5.5 A look into the near and distant future

Due to the complexity and cost of the enzyme assay as well as the instability of the substrates, the purification of the DHFS-FPGS enzyme will first have to be optimised so that activity assays can be done on the enzyme obtained after different purification steps at the same time. Guided by the results from functional complementation, it seems that tagless DHFS-FPGS would be the preferred enzyme for activity assays. The purification of this enzyme however is complicated through the absence of a tag. Using the C-terminal His_6-tagged DHFS-FPGS, however for affinity purification, combined with detection of the tagged enzyme with anti-His antibodies; the size exclusion and anion exchange HPLC purification of the C-terminal His_6-tagged enzyme could be optimised and used to purify the tagless protein. Once sufficient amounts of protein are obtained, activity assays could then be performed. Polyclonal, monospecific antibodies raised against epitopes of tagless DHFS-FPGS could furthermore be used for the large-scale purification of active, correctly folded enzyme. This could then be used for the X-ray crystallography or NMR analysis of the enzyme and thus determination of its three dimensional structure. This information would be invaluable for the rational design and prediction of inhibitors as well as to provide insights on the role of DHFS-FPGS enzyme in the folate metabolism of the malaria parasite and thus provide new options for selective inhibition. In concluding the 'discover' aspect of antimalarial drug strategies, the development of novel antifolates targeted to this novel folate enzyme, which is potentially

less prone to resistance development could revolutionise the current antifolate status, by replacing drugs of declining efficiency with highly effective, novel inhibitors against which the parasite has no defence mechanism from previous exposure. Alternatively, the novel antifolates could be used in combination with current antifolates and thus control the parasite by means of synergistic inhibition of different points within the same pathway. This would mean that the 'develop' aspect has been successfully completed. If the cheap, effective drugs with short half-lives are widely distributed through large-scale control strategies, the 'deliver' aspect of antimalarial drug treatments could achieve stringent malaria control to relieve the burden caused by malaria on the developing countries of Africa.

References

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215:** 403-410.

Armstrong-Schellenberg, J.R.M. Hanson, K., Kikumbih, N., Mponda, H., Nathan, R., Lake, S., Mills, A., Tanner, M. and Lengeler, C. (2003). Cost-effectiveness of social marketing of insecticide-treated nets for malaria control in the United Republic of Tanzania. *Bulletins of the World Health Organisatio.* **81:** 269-276.

Baca, A.M. and Hol, W.G. (2000). Overcoming codon bias: a method for high-level overexpression of *Plasmodium* and other AT-rich parasite genes in *Eschericia coli*. *International Journal of Parasitology* **30:** 113-118.

Bairoch, A., Bucher, P. and Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucleic Acids Research* **25:** 217-221.

Bognar, A.L., Osborne, C., Shane, B., Singer, S.C. and Ferone, R. (1985). Folyl-poly-gamma-glutamate synthetase-dihydrofolate synthetase-cloning and high expression of the *Escherichia coli folC* gene and purification and properties of the gene product. *The Journal of Biological Chemistry* **260:** 5625–5630.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72:** 248-254.

Brooker, S., Guyatt, H., Omumbo, J., Shrett, R., Drake, L. and Ouma, J. (2000). Situation analysis of malaria in school-aged children in Kenya - what can be done? *Parasitology Today* **16:** 183-186.

Brooks, D.R., Wang, P., Read, M., Watkins, W.M., Sims, P.F.G. and Hyde, J.E. (1994). Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase:dihydropteroate synthase gene in lines of the human malaria parasite, *Plasmodium falciparum*, with differing resistance to sulfadoxine. *European Journal of Biochemistry* **224:** 397-405.

Bzik, D.J., Li, W., Horii, T. and Inselburg, J. (1987). Molecular cloning and sequence analysis of the *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthetase gene. *Proceedings of the National Academy of Science USA* **84:** 8360-8364.

Carlton, J.M. Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L. *et al.* (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii*. *Nature* **419:** 512-519.

Carpenter, E., Withers-Martinez, C., Hackett, F., Ely, B., Sajid, M., Grainger, M. and Blackman, M (1999). PCR based gene synthesis as an efficient approach for expression of the A+T rich malaria genome. *Protein Engineering* **12:** 1113-1120.

Cherest, H., Thomas, D. and Surdin-Kerjan, Y. (2000). Polyglutamylation of folate coenzymes is necessary for methionine biosynthesis and maintenance of an intact mitochondrial genome in *Saccharomyces cerevisiae*. *The Journal of Biological Chemistry* **275:** 14056-14063.

Clark, J.B., and Webb, R.B. (1955). A comparison of the crystal violet nuclear stain with other techniques. *Staining Technology* **30:** 73-78.

Cobb, B.D. and. Clarkson, J.M. (1994). A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Research* **22:** 3801-3805.

Coombs, G. Goldberg, D.E., Klemba, M., Berry, C., Kay, J. and Mottram, J.C. (2001). Aspartic proteases of *Plasmodium falciparum* and other parasitic protozoa as drug targets. *Trends in Parasitology* **17:** 532-537.

Corpet, F., Gouzy, J. and Kahn, D. (1999). Whole genome protein domain analysis using a new method for domain clustering. *Computational Chemistry* **23:** 333-340.

Costi, M.P. and Ferrari, S. (2001). Update on antifolate drug targets. *Current drug targets* **2:** 135-166.

Crabb B.S., Cooke, B.M., Reeder, J.C., Waller, R.F., Caruana, S.R., Davern, K.M., Wickham, M.E., Brown, G.V., Coppel, R.L. and Cowman, A.F. (1997). Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell* **89:** 287-96.

Desai, S., Bezrukov, S. and Zimmerberg, J. (2000). A voltage-dependent channel involved in nutrient uptake by red blood cells infected with the malaria parasite. *Nature* **406:** 1001-1005.

Djimde, A., Doumbo, O., Mahamadou, D., and Plowe, C.V. (2002). Host genetics and clearance of drug-resistant *Falciparum* malaria. 3rd Pan African Multilateral Initiative on Malaria (MIM) conference, Arusha, Tanzania.

Doolan, D. and Hoffman, S.L. (2001). DNA-based vaccines against malaria: status and promise of the multistage malaria DNA vaccine operation. *International Journal of Parasitology* **31:** 753-762.

Edelman M.J. and Gandara, D. R. (1996). Promising new agents in the treatment of non-small cell lung cancer. *Cancer Chemotherapy and Pharmacology* **37:** 385-393.

Eggleson, K. Duttin, K. and Goldberg, D. (1999). Identification and characterisation of Falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum. The Journal of Biological Chemistry* **274:** 32411-32417.

Fadili, A.E., Kundig, C. and Oullette, M. (2002). Characterization of the folylpolyglutamate synthetase gene and polyglutamylation of folates in the protozoan parasite *Leishmania. Molecular and Biochemical Parasitology* **124:** 63-71.

Ferone, R. (1977). Folate metabolism in malaria. *Bulletins of the World Health Organisation* **55:** 291-298.

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

Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F. and McFadden G.I. (2003). Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. *Science* **299:** 705-708.

Francis, S., Sullivan, D.J. and Goldberg, D. (1997). Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*. *Annual Review of Microbiology* **51:** 97-123.

Fry, M. and Beesly, J. (1991). Mitochondria of mammalian *Plasmodium* spp. *Parasitology Today* **102:** 17-26.

Fu, C., Smith, S., Simkins, S.G. and Agris, P.F. (2002). Identification and quantification of protecting groups remaining in commercial oligonucleotide products using monoclonal antibodies. *Analytical Biochemistry* **306:** 135-143.

Gallup, J.L. and Sachs, J.D. (2001). The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene.* **64:** 85-96.

Gangjee, A., Dubash, N.P., Zeng, Y. and McGuire J.J. (2002). Recent advances in the chemistry and biology of folypoly-gamma-glutamate synthetase substrates and inhibitors. *Current Medicinal Chemistry and Anti-Cancer Agents* **2:** 331-355.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M. *et al.* (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498-511.

Garnier, J., Gibrat, J.F. and Robson B (1996). GOR secondary structure prediction method version IV *Methods in Enzymology* R.F. Doolittle Ed. **266:** 540-553.

Gentile, G., Slotman, M., Ketmaier, V., Powell, J.R. and Caccone, A. (2001). Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae s.s. Insect Molecular Biology* **10:** 25-32.

Hoffman, S.L., Subramanian, G.M., Collins, F.H. and Venter, J.C. (2002). *Plasmodium*, human and *Anopheles* genomics and malaria. *Nature* **415:** 702-709.

Holding, P.A. and Snow, R.W. (2001). Impact of *Plasmodium falciparum* malaria on performance and learning: review of the evidence. *American Journal of Tropical Medicine and Hygiene.* **64:** 68-75.

Huang, J., Villemain, J., Padilla, R. and Sousa, R. (1999). Mechanisms by which T7 lysozyme specifically regulates T7 RNA polymerase during different phases of transcription. *Journal of Molecular Biology* **293:** 457-475.

Hyde, J.E. (1989). Point mutations and pyrimethamine resistance in *Plasmodium falciparum*. *Parasitology Today* **5:** 252-255.

James, A.A. (2003). Blocking malaria parasite invasion of mosquito salivary glands. *Journal of Experimantal Biology* **206:** 3817-3821.

Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846-856.

Keshavjee, K., Pyne, C., Kimlova, L.J., Huy, J., Beebakhee, G. and Bognar, A.L. (1991). Mutagenesis of the *folC* gene encoding folylpolyglutamate synthetase-dihydrofolate synthetase in *Escherichia coli*. *Archives of Biochemistry and Biophysics.* **284:** 9-16.

Kester, K.E., McKinney, D.A., Tornieporth, N., Ockenhouse, C.F., Heppner, D.G., Hall, T., Krzych, U., Delchambre, M., Voss, G., Dowler, M.G., Palensky, J., Wittes, J.and Cohen J, (2001). Efficacy of recombinant circumsporozoite protein vaccine regimens against experimental *Plasmodium falciparum* malaria. *Journal of Infectious Diseases* **183:** 640-647.

Koukourakis, M.I., Giatromanolaki, A. and Sivridis, E. (2003). Lactate dehydrogenase isoenzymes 1 and 5: Differential expression by neoplastic and stromal cells in non-small cell lung cancer and other epithelial malignant tumors. *Tumor Biology* **24:** 199-202.

Krishna, S., Eckstein-Ludwig, U., Webb R.J., Van Goethem, I.D., East, J.M., Lee, A.G., Kimura, M., O'Neill, P.M., Bray, P.G. and Ward, S.A. (2003). Artemisinins target the SERCA of *Plasmodium falciparum*. *Nature* **424:** 57-961.

Kristan, M., Fleischman, H., della Torre, A., Stich, A. and Curtis, C.F. (2003). Pyrethroid resistance/susceptibility and different urban/rural distribution of *Anopheles arabiensis* and *A. gambiae* ss. vectors in Nigeria and Ghana. *Medicinal and Veterinary Entomology* **17:**326-332.

Krungkrai, J., Webster, H. and Yuthavong, Y. (1989). *De novo* and salvage biosynthesis of pteroylpentaglutamates in the human malaria parasite, *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **32:**25-38.

Lee, C.S., Salcedo, E., Wang, Q., Wang, P., Sims, P. and Hyde, J. (2001). Characterisation of three genes encoding enzymes of the folate biosynthetic pathway in *Plasmodium falciparum*. *Parasitology* **122:**1-13.

Lell, B. Ruangweerayut, R., Wiesner, J., Missinou, M.A., Schindler, A., Baranek, T., Hintz, M., Hutchinson, D., Jomaa, H. and Kremsner, P.G. (2003). Fosmidomycin, a novel chemotherapeutic agent for malaria. *Antimicrobial Agents and Chemotherapy* **47:** 35-738.

Li, Y.C., Korol, A.B., Fahima, T., Beiles, A. and Nevo, E. (2002). Microsattelites: genomic disrtibution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11:** 453-2468.

Lupas, A. (1996). Coiled coils: new structures and new functions. *Trends in Biochemical Sciences* **21:** 375-382.

Manning, S.K., Woodrow, C., Zuniga, F.A., Iserovich, P., Fischbarg, J., Louw, A.I. and Krishna, S. (2002). Mutational analysis of the hexose transporter of *Plasmodium falciparum*

and development of a three-dimensional model. *The Journal of Biological Chemistry* **277:** 30942-30949.

Marsh, K. Forster, D., Waruiru, C., Mwangi, I., Winstanley, M., Marsh, V., Newton, C., Winstanley, P., Warn, P., Peshu, N., *et al.* (1995). Indicators of life-threatening malaria in African children. *New England Journal of Medicine* **332:** 1399-1404.

McKie, J.H., Douglas, K.T., Chan, C., Roser, S, Yates, R., Read, M., Hyde, J.E., Dascombe,M., Yuthavong, Y and Sirawaraporn ,W. (1998). Rational drug design approach for overcoming drug resistance: Application to pyrimethamine resistance in malaria. *Journal of Medicinal Chemistry* **41:** 1367-1370.

McGuire, J.J. and Bertino, J.R. (1981). Enzymatic synthesis and function of folylpolyglutamates. *Molecular and Cellular Biochemistry*. **38:** 1 -48.

Merril, C.R., Goldman, D., Sedman, S.A. and Ebert, M.H. (1981). Ultrasensitive stain for proteins in polyacrylamide gels shows regional variation in cerebrospinal fluid proteins. *Analytical Biochemistry* **72:** 248-254.

Miller, L.H. (1999). Evolution of the human genome under selective pressure from malaria: applications for control. *Parasitologia* **41:** 77-82.

Miller, L.H., Baruch, D.I., Marsh, K. and Doumbo, O.K. (2002). The pathogenic basis of malaria. *Nature* **415:** 673-679.

Moran, R.G. (1999). Role of folylpoly-gamma-glutamate synthetase in therapeutics with tetrahydrofolate antimetabolites: an overview. *Seminars in Oncology* **26:** 24-32.

Morris, A., McArthur, M., Hutchinson, E. and Thornton, J. (1992). Stereochemical quality of protein structure coordinates. *Proteins* **12:** 345-364.

Muentener, P. Schlagenhauf, P. and Steffen, R. (1999). Imported malaria (1985-95): trends and perspectives. *Bulletins of the World Health Organisation* **77:** 560-566.

Ohkanda, J., Lockman, J.W., Yokoyama, K., Gelb, M.H., Croft, S.L., Kendrick, H., Harrell, M.I., Feagin, J.E., Blaskovich, M.A., Sebti, S.M. and Hamilton AD. (2001). Peptidomimetic inhibitors of protein farnesyltransferases show antimalarial activity. *Bioorganic and Medicinal Chemistry Letters* **11:** 761-764.

Osborne, C., Lowe, KE and Shane, B. (1993). Regulation of folate and one-carbon metabolism in mammalian cells. Folate metabolism in Chinese hamster ovary cells expressing *Escherichia coli* or human folylpoly-gamma-glutamate synthetase activity. *Journal of Biological Chemistry* **268:** 21 657–21 664.

Pizzi, E and Frontali, C. (2001). Low-Complexity regions in *Plasmodium falciparum* proteins. *Genome Research* **11:** 218-229.

Prapunwattana, P., Sirawaraporn, W., Yuthavong, Y. and Santi, D.V. (1996). Chemical synthesis of the *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase gene. *Molecular and Biochemical Parasitology* **83:** 93-106.

Quigley, M. and Holmes, D.S. (1981). A rapid boiling method for the preparation of bacterial plasmids. *Analytical Biochemistry* **114:** 193-197.

Ramage, P., Hemming, R., Mathis, B., Cowan-Jacob, S.W., Rondeau, J.M., Kallen, J., Blommers, M.J.J., Zurini, M and Rudisser, S. (2002). Snags with tags: Some observations made with $(His)_6$ -tagged proteins *Life Science News (Amersham Biosciences)* **11:** 1-5.

Ravanel, S., Cherest, H., Jabrin, S., Grunwald, D., Surdin-Kerjan, Y., Douce, R. and Rebeille, F. (2001). Tetrahydrofolate biosynthesis in plants: Molecular and functional characterization of dihydrofolate synthetase and three isoforms of folylpolyglutamate synthetase in *Arabidopsis thaliana. Proceedings of the National Academy of Science USA* **98:** 15360–15365.

Richie, T.L. and Saul, A. (2002). Progress and challenges for malaria vaccines. *Nature* **415:** 694-709.

Ridley, R.G. (2002). Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature Insight* **415:** 686-693.

Rogers, D.J., Randolph, S.E., Snow, R.W. and Hay, S.I. (2002). Satellite imagery in the study and forecast of malaria. *Nature Insight* **415:** 710-715.

Romanos, M.A., Makoff, A., Fairweather, N., Beesley, K., Slater, D., Rayment, F., Payne, M and Clare, J. (1991). Expression of tetanus toxin fragment C in yeast-gene synthesis is required to eliminate fortuitous polyadenylation sites in AT rich DNA. *Nucleic Acids Research.* **19:** 1461-1467.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* **232:** 584-599.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.C. Richter, D.J., Schaffner, S.F., Gabriel, S.B. and Platko, J.V. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832-837.

Sachs, J. and Malaney, P. (2002). The economic and social burden of malaria. *Nature Insight* **415:** 680-685.

Salcedo, E., Cortese, J.F., Plowe, C.V., Sims, P.F.G. and Hyde, J.E. (2001). A bifunctional dihydrofolate synthetase– folylpolyglutamate synthetase in *Plasmodium falciparum* identified by functional complementation in yeast and bacteria. *Molecular and Biochemical Parasitology* **112:** 239-252.

Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* **234:** 779-815.

Sambrook, J., Fritsch, E.J. and Maniatis, T. (1989). In: *Molecular cloning: A laboratory manual* (2nd ed) Cold Spring Harbor, New York.

Santi, D., Nolan, P. and Shane, B. (1987). Folylpolyglutamates in *Leishmania major*. *Biochemical and Biophysical Research Communications* **146:** 1089-1092.

Saul, A. (1999). The role of variant surface antigens on malaria-infected red blood cells. *Parasitology Today* **15:** 455-457.

Shane, B. (1980). Pteroylpoly-γ–glutamate synthesis by *Corynebacterium* species. *The Journal of Biological Chemistry* **255:** 5649-5654.

Sheng, Y., Sun, X., Shen, Y., Bognar, A.L., Baker, E.N. and Smith, C.A. (2000). Structural and functional similarities in the AMP-forming amide bond ligase superfamily: implications for a substrate-induced conformational change in folylpolyglutamate synthetase. *Journal of Molecular Biology* **302:** 427-440.

Shenai, B., Sijwali, P., Singh, A. and Rosenthal, P. (2000). Characterisation of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *Journal of Biological Chemistry* **275:** 29000-29010.

Sibley, C.H., Hyde, J.E., Sims, P.F.G., Plowe, C.V., Kublin, J.G., Mberu, E.K., Cowman, A.F., Winstanley, P.A., Watkins, W.M. and Nzila, A.M. (2001). Pyrimethamine–sulfadoxine resistance in *Plasmodium falciparum*: what next? *Trends in Parasitology* **17:** 582-568.

Sirawaraporn, W., Prapunwattana, P., Sirawaraporn, R., Yuthavong, Y. and Santi, D.V. (1993). The dihydrofolate reductase domain of *Plasmodium falciparum* thymidylate synthase-dihydrofolate reductase. *The Journal of Biological Chemistry* **268:** 21637-21644.

Skerra, A. (1994). Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in *Escherichia coli*. *Gene* **151:** 131-135.

Stokstad, E.L.R. and Koch, J. (1967). Folic acid metabolism. *Physiology Reviews* **47:** 83-116.

Stowers, A.W., Cioce, V., Shimp, R.L., Lawson, M., Hui, G., Muratova, O., Kaslow, D.C., Robinson, R., Long, C.A. and Miller, L.H. (2001). Efficacy of two alternate vaccines based on *Plasmodium falciparum* merozoite surface protein 1 in an *Aotus* challenge trial. *Infections and Immunity* **69:** 1536-1546.

Sugiyama, T., Suzue, K., Okamoto, M., Inselburg, J., Tai, K. and Horii, T. (1996). Production of recombinant SERA proteins of *Plasmodium falciparum* in *Escherichia coli* by using synthetic genes. *Vaccine* **14:** 1069-1076.

Sun, X., Bognar, A., Baker, E and Smith, C. (1998). Structural homologies with ATP- and folate-binding enzymes in the crystal structure of folylpolyglutamate synthetase. *Biochemistry* **95:** 6647-6652.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmaugin, F. and Higgins, D.G. (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25:** 4876-4882.

Triglia, T., Wang, P., Sims, P.F.G., Hyde, J.E. and Cowman, A.F. (1998). Allelic exchange at the endogenous genomic locus in *Plasmodium falciparum* proves the role of dihydropteroate synthase in sulfadoxine-resistant malaria. *The EMBO Journal* **17:** 3807-3815.

Voet, D. and Voet, J.G. (1995). In: *Biochemistry* (3rd edidtion) John Wiley and Sons Inc, pp. 896-897.

Wang, P., Brobey, R.K.B, Horii, T., Sims, P.F.G. and Hyde, J.E. (1999). Utilization of exogenous folate in the human malaria parasite *Plasmodium falciparum* and its critical role in antifolate drug synergy. *Molecular Microbiology* **32:** 1254-1262.

Wang, P., Brooks, D.R., Sims, P.F.G. and Hyde, J.E. (1995). A mutation-specific PCR system to detect sequence variation in the dihydropteroate synthetase gene of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **71:** 115-125.

Wang, P., Read, M., Sims, P.F.G. and Hyde, J.E. (1997). Sulfadoxine resistance in the human malaria parasite *Plasmodium falciparum* is determined by mutations in dihydropteroate synthase and an additional factor associated with folate utilization. *Molecular Microbiology* **23:** 979-986.

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

Wootton, J.C. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology* **266:** 554-571.

Yuthavong, Y. (2002). Basis for antifolate action and resistance in malaria. *Microbes and Infection* **4:** 175-182.

Yuvaniyama, J., Chitnumsub, P., Kamchonwongpaisan, S., Vanichtanankul, J., Sirawaraporn, W, Taylor, P., lkinshaw, M.D. and Yuthavong, Y. (2003). Insights into antifolate resistance from malarial DHFR-TS structures. *Nature Structural Biology* **10:** 357-365.

Zangenberg, G., Saiki, R. and Reynolds, R. (1999). *Multiplex PCR: Optimisation Guidelines* In: PCR Applications: Protocols for Functional Genomics. Innis, M.A., Gelfand, D.H., and Sninsky, J.J. (eds.), Academic Press: San Diego, pp. 73–94.

Zhang, H., Howard, E.M. and Roepe, P.D. (2002). Analysis of the Antimalarial Drug Resistance Protein Pf*crt* Expressed in Yeast. *The Journal of Biological Chemistry* **277:** 49767-49775.

Appendices

Appendix A:
Codon Preferance tables (www.kazusa.or.jp/codon)

E.coli K12 strain codon preference table (1608122 codons)

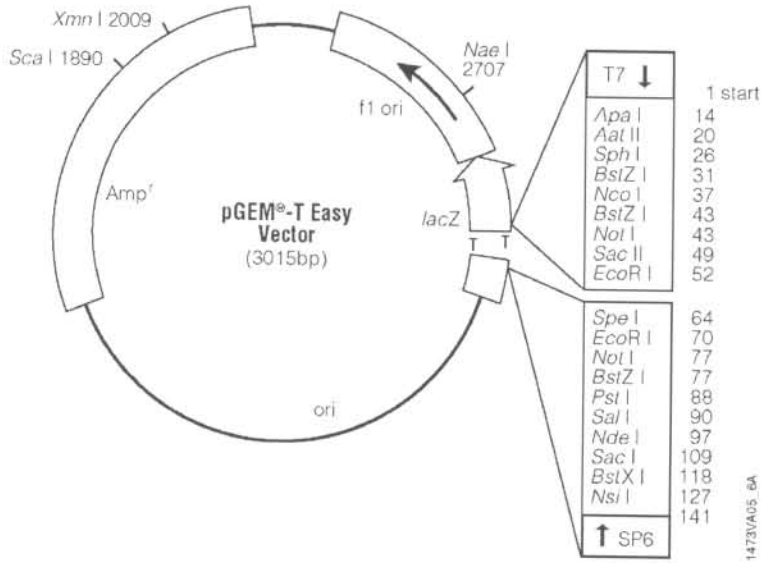
Phe	UUU	22.4(	35982)	Ser	UCU	8.5(	13687)	Tyr	UAU	16.3(	26266)
	UUC	16.6(	26678)		UCC	8.6(	13849)		UAC	12.3(	19728)
Leu	UUA	13.9(	22376)		UCA	7.2(	11511)	Stop	UAA	2.0(	3246)
	UUG	13.7(	22070)		UCG	8.9(	14379)		UAG	0.2(	378)
	CUU	11.0(	17754)	Pro	CCU	7.1(	11340)	His	CAU	12.9(	20728)
	CUC	11.0(	17723)		CCC	5.5(	8915)		CAC	9.7(	15595)
	CUA	3.9(	6212)		CCA	8.5(	13707)	Gln	CAA	15.4(	24835)
	CUG	52.7(	84673)		CCG	23.2(	37328)		CAG	28.8(	46319)
Ile	AUU	30.4(	48818)	Thr	ACU	9.0(	14397)	Asn	AAU	17.7(	28465)
	AUC	25.0(	40176)		ACC	23.4(	37624)		AAC	21.7(	34912)
	AUA	4.3(	6962)		ACA	7.1(	11366)	Lys	AAA	33.6(	54097)
Met	AUG	27.7(	44614)		ACG	14.4(	23124)		AAG	10.2(	16401)
Val	GUU	18.4(	29569)	Ala	GCU	15.4(	24719)	Asp	GAU	32.2(	51852)
	GUC	15.2(	24477)		GCC	25.5(	40993)		GAC	19.0(	30627)
	GUA	10.9(	17508)		GCA	20.3(	32666)	Glu	GAA	39.5(	63517)
	GUG	26.2(	42212)		GCG	33.6(	53988)		GAG	17.7(	28522)

Cys	UGU	5.2(	8340)								
	UGC	6.4(	10347)								
Stop	UGA	0.9(	1468)								
Trp	UGG	15.3(	24615)								
Arg	CGU	21.0(	33694)								
	CGC	22.0(	35306)								
	CGA	3.6(	5716)								
	CGG	5.4(	8684)								
Ser	AGU	8.8(	14092)								
	AGC	16.1(	25843)								
Arg	AGA	2.1(	3337)								
	AGG	1.2(	1987)								
Gly	GGU	24.9(	40019)								
	GGC	29.4(	47309)								
	GGA	7.9(	12776)								
	GGG	11.0(	17704)								

Plasmodium falciparum 3D7 strain codon preferance table (2776356 codons)

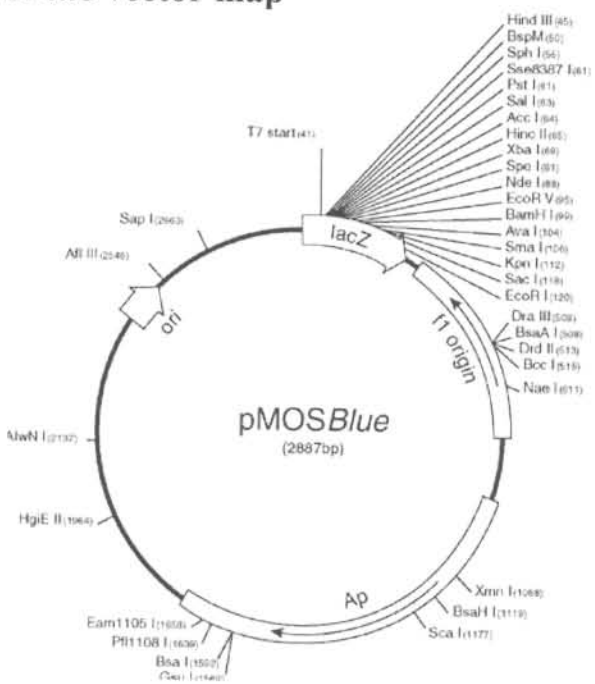
Phe	UUU	36.3(100866)	Ser	UCU	14.8(	41088)	Tyr	UAU	50.5(140208)	Cys	UGU 15.1(42007)
	UUC	7.1(19624)		UCC	5.1(	14067)		UAC	6.1(16953)		UGC 2.2(6245)
Leu	UUA	47.3(131340)		UCA	16.6(	46169)	Stop	UAA	0.9(2582)	Stop	UGA 0.3(810)
	UUG	10.3(28684)		UCG	3.0(	8225)		UAG	0.1(379)	Trp	UGG 4.9(13545)
	CUU	8.6(23932)	Pro	CCU	8.0(	22094)	His	CAU	20.8(57660)	Arg	CGU 3.0(8270)
	CUC	1.8(4959)		CCC	2.0(	5664)		CAC	3.5(9631)		CGC 0.4(1149)
	CUA	6.0(16735)		CCA	9.2(	25658)	Gln	CAA	24.2(67221)		CGA 2.4(6674)
	CUG	1.4(3983)		CCG	0.9(	2611)		CAG	3.7(10280)		CGG 0.3(746)
Ile	AUU	35.8(99371)	Thr	ACU	10.6(	29480)	Asn	AAU	122.6(340414)	Ser	AGU 20.3(56431)
	AUC	6.2(17289)		ACC	4.8(	13257)		AAC	19.8(54952)		AGC 3.9(10940)
	AUA	50.0(138720)		ACA	21.7(	60278)	Lys	AAA	95.6(265519)	Arg	AGA 16.0(44311)
Met	AUG	21.7(60328)		ACG	3.8(	10464)		AAG	21.4(59397)		AGG 4.3(11924)
Val	GUU	15.6(43447)	Ala	GCU	8.3(	23103)	Asp	GAU	55.6(154404)	Gly	GGU 12.0(33208)
	GUC	2.4(6713)		GCC	2.1(	5802)		GAC	8.6(23893)		GGC 1.3(3600)
	GUA	16.0(44474)		GCA	8.5(	23527)	Glu	GAA	62.3(172861)		GGA 12.7(35265)
	GUG	4.8(13274)		GCG	1.0(	2894)		GAG	10.4(29008)		GGG 2.8(7749)

Appendix B: pGEM T Easy Vector map



Position of the A-T cloning site is indicated in the multiple cloning cassette within the *lacZ* gene. The T7- and SP6 promoter regions indicate vector primer positions respectively. The ampicillin resistance gene (Amp^r) is also indicated.

Appendix C: pMOS*Blue* vector map



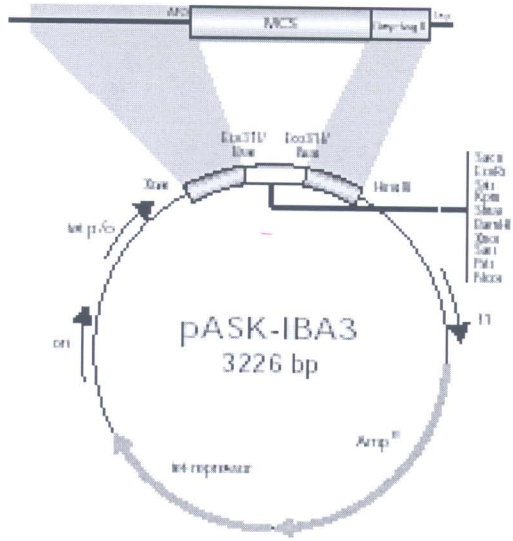
Cloning takes place at the linearised EcoRV site, which is also dephosphorylated. Phosphorylation of the insert during the pk reaction facilitates blunt ended cloning in the subsequent step. The vector also confers ampicillin resistance to identify transformed cells. Inactivation of the lacZ gene by the cloned insert also results in a white phenotype of recombinant clones. The plasmid furthermore contains lacZ ΔM15 and thus tetracycline resistance is observed in recombinant strains. This eliminates false white colonies that have lost the plasmid.

Appendix D: List of oligonucleotide sequences

Oligonucleotide name	Sequence (5′ to 3′ direction)
Fwd1	CGCGGACATATGGAGAAGAACCAGAACGATAAAAGCAACAAAAACGATATTATTCACATGAACGATAAAAGC
Fwd2	AACAACTTTATTGATAAGAACGATGAACATGATATGAGCGATATTCTGCATAAAATTAATAATGAGGAGAAGAAA
Fwd3	TGCCTGGAACTGCTGTATAAAACCCATGCGCTGAAACTGGGCCTGGATAACCCGAAGAAGCTGAACGAAAG
Fwd4	CCATCCATATTGCTGGCACCAACGGCAAAGGCAGCGTGTGCTATAAAATTTATACCTGCCTGAAGATCAA
Fwd5	AGCCCGCATATTTTTAGCCTGCGCGAACGCATTATTGTGAACGATGAACCGATTAGCGAAAAAGAACTGATTC
Fwd6	CAAAGCGAAGAAGCTGTATATTAACCCGAGCTTTTTTGAAATTATTACCCTGGTGGCGTTTCTGCATTTTTTAAAC
Fwd7	CGATTATTGAAACCGGCATTGGCGGCCGCCTGGATGCCACCAACATTCTGACCAAACCGGAAGTGATTGTGATTAC
Fwd8	CGATAACCTGCCGATTATTTGCAACGAAAAAATTGGCATTTTTAAAAAAGATGCGAACGTGGTGATTG
Fwd9	TGATAAAGCGAAAGAGCTCAACTGCACCATTCATACCGTTGTGCCGGAACCGCGCGGCGAACGCTATAACGAAGAA
Fwd10	CTGCGCACCCTGGAAATTCTGAACATTAGCATTGATTATTTTCTGAAAAGCATTATTCCGATTAAACCGCCGCT
Fwd11	GATTCAGCATATTAAGAAGAAGTTCAGCCCAGACAACCTGGAACATAACGTGCAGTATCCGCTGGCGGTGATTC
Fwd12	TGATCGCCTGTGCACCGATATCAACTATTTTCATAAAGGCCAGAACATTCGCATTTGCATTAGCATTAC
Fwd13	CATCCGTTTATTGCGCAGTTTGGCGATACCCTGAAAGATATTTTTTATCTGCCGAGCCTGAACGAACGCACCTA
Fwd14	TGAACAACGAAGAAGAAATTAAAAACGAAATTAAAGAACTGATTCTGAGCAGCAGCAAAAAAGTGGGCAA
Fwd15	ATTAACGAAGAAGATGCGCTGAAACTGTATAAACGCGGCTGCATTCCGCTGATTATTAAAAACGCGTTTCTG
Fwd16	GCTGGTTTGTGGTACCTTCTTCGTGTTTGATGAAGTGCTGAACGTGTTTGATATTCATAGCGATATGCAGGA
Rev1	GTTCATCGTTCTTATCAATAAAGTTGTTAATGTTGTTTTTATCATAGTTGCCGCTTTTATCGTTCATGTGAATAAT
Rev2	GTTTTATACAGCAGTTCCAGGCATTCGCTATAGCTTTTAATTTCTTCATATTTCTTCTCCTCATTATTAATTTTAT
Rev3	TGGTGCCAGCAATATGGATGGTTTTATATTTATCGCACGGATGGCCAAAGCTTTCGTTCAGCTTCTTCGGGT
Rev4	CAGGCTAAAAATATGCGGGCTGCTAAACAGTCCCACTTTAAATTTTTTGATCTTCAGGCAGGTATAAATTT
Rev5	GTTAATATACAGCTTCTTCGCTTTGTTCAGCACTTCGTTCACCAGATGAATCAGTTCTTTTTCGCTAATCG
Rev6	CCAATGCCGGTTTCAATAATCGCATAATCCACTTTTTTGTTTAAAAAATGCAGAAACGCCAC
Rev7	GCAAATAATCGGCAGGTTATCGCCCAGAATGTTCAGATGATCATAGCCAATGCTGGTAATCACAATCACTTCC
Rev8	GAGCTCTTTCGCTTTATCAAACACGTTTTTATAAATCGCCACGCTCGGGCCAATCACCACGTTCGCATCT
Rev9	TAATGTTCAGAATTTCCAGGGTGCGCAGCGCAATGCGGCTGTTTTCTTCGTTATAGCGTTCG
Rev10	GAACTTCTTCTTAATATGCTGAATCTGTTCGGTCGCCAGATACTGAATGCGCAGCGGCGGTTTAATCGGAAT
Rev11	TATCGGTGCACAGGCGATCAATCGCGGTTTCGTTATGGCCCACATCCAGAATCACCGCCAGCGGATAC
Rev12	CAAACTGCGCAATAAACGGATGAAACACGCTCAGGTTGCGCGGTTTGGTAATGCTAATGCAAATGCGAATG
Rev13	CGTTTTTAATTTCTTCTTCGTTGTTCAGCATTTCCACAATTTCTTCAAAATCATAGGTGCGTTCGTTCAGGCT
Rev14	TTCAGCGCATCTTCTTCGTTAATGTTGCCCTGTTTTTCATGCGCCAGCCATTTGCCCACTTTTTTGCTGCTG
Rev15	AGAAGGTACCACAAACCAGCAGAATGCTGTTATCTTTGCAACACTCCAGAAACGCGTTTTTAATAATCAGC

Oligonucleotide name	Sequence (5′ to 3′ direction)
Rev16a	CCGGATCCTTACACCAGGCTCGGTTCGTTCATAAAAATGGTATCCTGCATATCGCTATGAATATCA
Rev16b	CTCGGATAATGTTTCACCAGGCTCGGTTCGTTCATAAAAATGGTATCCTGCATATCGCTATGAATATCA
dhfs2f1	AATTTAAAGTGGGACTGTTTAGCAGCCCGCATATTTTTAGCCTGCGCGA
dhfs2r1	TGCCTAATCGGTTCATCGTTAACAATAATGGCTTCGCGCAGGCTAAAAATATGC
dhfs2f2	AACGATGAACCGATTAGCGAAAAAGAACTGATTCATCTGGTGAACGAAGTGCTG
dhfs2r2	AGCTCGGGTTAATATACAGCTTCTTCGCTTTGTTCAGCACTTCGTTCACCAGAT
dhfs2f3	GCTGTATATTAACCCGAGCTTTTTTGAAATTATTACCCTGGTGGCGTTTCTGCA
dhfs2r3	AATAATCGCATAATCGACTTTTTTGTTTAAAAAATGCAGAAACGCCACCAGGGT
dhfs2f4	CAAAAAAGTCGATTATGCGATTATTGAAACCGGCATTGGCGGCCGCCTGGATGCCA
dhfs2r4	GTAATCACAATCACTTCCGGTTTGGTCAGAATGTTGGTGGCATCCAGGCGGCCG
dhfs2f5	CCGGAAGTGATTGTGATTACCAGCATTGGCTATGATCATCTGAACATTCTGGGC
dhfs2r5	GTTGCAAATAATCGGCAGGTTATCGCCCAGAATGTTCAGATGAT
dhfs3f1	TGCGCACCCTGGAAATTCTGAACATTAGCATTGATTATTTTCTGAAAAGCATTATTCC
dhfs3r1	GGTCGCCAGATACTGAATGCGCAGCGGCGGTTTAATCGGAATAATGCTTTTCAGAAAA
dhfs3f2	GCGCATTCAGTATCTGGCGACCGAACAGATTCAGCATATTAAGAAGAAGTTCAGCCCA
dhfs3r2	CACCGCCAGCGGATACTGCACGTTATGTTCCAGGTTGTCTGGGCTGAACTTCTTCTTAA
Sequencing primer	Sequence (5′ to 3′ direction)
U19	AGGGTTTTCTCAGTCACGA
T7 promoter	TAATACGACTCACTATAGGG
Sp6 promoter	ATTTAGGTGACACTATA

- **C-terminal Strep-tagged expression: pASK-IBA3 vector (www.iba-go.com)**




```
                                                                  Eco31I PshAI
      RBS                                  XbaI       RBS          BsaI SacII BsmFI
5'-TAGAGAAAAGTGAAATGAATACTTCGACAAAAATCTAGATAACGAGGGCAAAAaatgGGAGACCCGCGGTCCC
          tetA: MetAsnSerSerThrLysIleEnd                          MetGlyAspArgGlyPro

                                                        PshAI Eco31I
 EcoRI   SstI   KpnI SmaI BamHI XhoI   SalI   PstI    BsmFI NcoI BsaI Eco47III
GAATTCGAGCTCGGTACCCGGGGATCCCTCGAGGTCGACCTGCAGGCGGACCATGGTCTCAgcgcTTGGAGCCAC
GluPheGluLeuGlyThrArgGlySerLeuGluValAspLeuGlnGlyAspHisGlyLeuSerAlaTrpSerHis
                                                          link

        BstBI        HindIII
CCGCAGTTCGAAAAATAATAAGCTTGACCTGTG-3'
ProGluPheGluLysEnd
Strep-tag II
```

Cloning primers:

Forward: 5'-NNNNNNGGTCTCNA ATG NNN NNN ($N_{n'}$)*

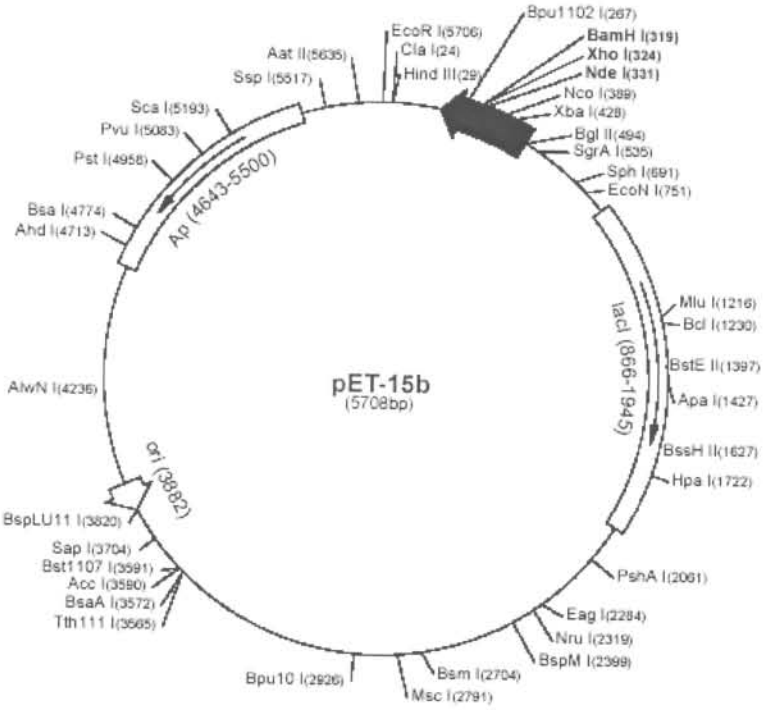
Reverse: 5'-NNNNNNGGTCTCNGC GCT NNN NNN... (N_n)

vector defined part | gene defined part

Appendix E: Vector systems used for recombinant expression

- **N-terminal His_6-tagged expression: pET15b vector (www.novagen.com)**

pET-15b sequence landmarks	
T7 promoter	463-479
T7 transcription start	452
His•Tag coding sequence	362-380
Multiple cloning sites	
(*Nde* I - *Bam*H I)	319-335
T7 terminator	213-259
lacI coding sequence	(866-1945)
pBR322 origin	3882
bla coding sequence	4643-5500



- **C-terminal His_6-tagged expression: pET22b vector (www.novagen.com)**

pET-22b(+) sequence landmarks	
T7 promoter	361-377
T7 transcription start	360
pelB coding sequence	224-289
Multiple cloning sites	
(*Nco* I - *Xho* I)	158-225
His•Tag coding sequence	140-157
T7 terminator	26-72
lacI coding sequence	764-1843
pBR322 origin	3277
bla coding sequence	4038-4895
f1 origin	5027-5482