MOLECULAR STUDIES OF PROTEIN CODING SEQUENCES
ON THE PLASTID-LIKE DNA OF PLASMODIUM FALCIPARUM

Andrea Michelle Whyte

PhD

University College London,
University of London

1996
To My Parents
ABSTRACT

In this laboratory, the discovery and sequencing of an extrachromosomal, 35 kilobase (35 kb) circular DNA from the apicomplexan parasites Plasmodium, Toxoplasma and Eimeria has led to the idea that this molecule is similar to plastid DNAs. However, at the onset of the project outlined here, only two thirds of the 35 kb molecule from P. falciparum had been sequenced.

A large portion of the reported data in this thesis was concerned with establishing the sequence of a 5.3 kb section of the 35 kb circle. These data led to the identification of the following genes encoding: two tRNAs (decoding serine and glycine), two open reading frames (ORFs) (one of which bears similarity to an ORF in the cyanelle of the alga Cyanophora paradoxa), the rpoC subunit of a prokaryote-like RNA polymerase (RNAP) and the ribosomal protein rps 2. In addition, the coding regions of the already partially sequenced rpoC and clpC genes were completed. An analysis of these genes and their chloroplast-encoded counterparts is included within this thesis.

Antibodies were made to the rpoB gene product encoded by the 35 kb circle. Two methods were used 1) generation in vitro of monoclonal antibodies to predicted synthetic peptides and 2) generation of rabbit polyclonal antibodies to a fusion protein. These antibodies successfully identified the β subunit of the 35 kb-encoded RNAP in parasite extracts, constituting the first direct evidence for functionality of the genome. In addition, by means of a polyclonal serum obtained against the E. coli RNAP, it was shown that the 35 kb-encoded RNAP shares epitopes with the corresponding polypeptides in E. coli.

In an attempt to ascertain the function of an ORF (ORF470) encoded on the plastid-like DNA of Plasmodium, I have established that homologues occur in a variety of bacteria including Mycobacterium, Corynebacter and Synechocystis spp. However, the function of this ORF remains unknown.

Finally, because of the anti-malarial activity of certain drugs directed against the elongation cycle of prokaryotic protein synthesis it became necessary to establish whether P. falciparum carries not only the elongation factor encoded by the 35 kb DNA but also a mitochondrial equivalent. Accordingly a search was made with the S. cerevisiae tufM gene, encoding the mitochondrial elongation factor Tu. Evidence for the presence of a nucleus encoded, P. falciparum tufM gene was determined in addition to the identification of the chromosome carrying it.
ACKNOWLEDGMENTS

I am grateful to Dr. R. J. M. Wilson for his constant support and guidance throughout this project and for his advice during the development of this thesis. I would also like to thank Dr. D. Williamson for valuable help, in addition to unlimited access to his computer.

Many thanks to Daphne Moore, Peter Moore, Paul Denny, Kate Roberts, Kaveri Rangachari, Peter Preiser, Irene Ling, Dr Mike Blackman, Dr Jose A. Guevara, Dr. William Stafford, Dr. Hilary Longhurst, Sola Ogun and Terry Scott-Finnigan for their immeasurable help. Thanks to Neil Cramphorn for his help in the preparation of the figures. Particular thanks to Malcolm Strath and Dr. Barbara Clough for their friendship and support.

The receipt of a post-graduate studentship from the Medical Research Council is gratefully acknowledged.

Thanks to Kerrin who has stuck with the ups and downs of this thesis. A special thankyou to my parents for their never ending support, encouragment and patience.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>alpha</td>
</tr>
<tr>
<td>ABTG</td>
<td>p-Amino-Benzyl, 1-thio-(\beta)-Galactopyranoside</td>
</tr>
<tr>
<td>Amp</td>
<td>ampicillin</td>
</tr>
<tr>
<td>Asn</td>
<td>asparagine</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>A_x</td>
<td>absorbance at x nm</td>
</tr>
<tr>
<td>β</td>
<td>beta</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>c.p.m</td>
<td>counts per minute</td>
</tr>
<tr>
<td>CIP</td>
<td>calf intestinal alkaline phosphatase</td>
</tr>
<tr>
<td>CTP</td>
<td>cytosine triphosphate</td>
</tr>
<tr>
<td>DAPI</td>
<td>6-diamino-2-phenylindole</td>
</tr>
<tr>
<td>DMSO</td>
<td>dimethyl sulphoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EM</td>
<td>electron microscopy</td>
</tr>
<tr>
<td>FCS</td>
<td>foetal calf serum</td>
</tr>
<tr>
<td>g</td>
<td>acceleration due to gravity</td>
</tr>
<tr>
<td>Glu</td>
<td>glutamate</td>
</tr>
<tr>
<td>GTP</td>
<td>guanosine triphosphate</td>
</tr>
<tr>
<td>h</td>
<td>hour</td>
</tr>
<tr>
<td>Ig</td>
<td>immunoglobulin</td>
</tr>
<tr>
<td>Ile</td>
<td>isoleucine</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl (\beta)-D-thiogalactopyranoside</td>
</tr>
<tr>
<td>IR</td>
<td>inverted repeat</td>
</tr>
<tr>
<td>J</td>
<td>Joule</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase</td>
</tr>
<tr>
<td>kDa</td>
<td>kilodaltons</td>
</tr>
<tr>
<td>LB</td>
<td>Luria broth</td>
</tr>
<tr>
<td>Leu</td>
<td>leucine</td>
</tr>
<tr>
<td>LrRNA</td>
<td>large ribosomal RNA</td>
</tr>
<tr>
<td>LSU</td>
<td>large subunit</td>
</tr>
<tr>
<td>Lys</td>
<td>lysine</td>
</tr>
<tr>
<td>M</td>
<td>molar</td>
</tr>
<tr>
<td>m</td>
<td>milli</td>
</tr>
<tr>
<td>μ</td>
<td>micro</td>
</tr>
<tr>
<td>min</td>
<td>minute</td>
</tr>
<tr>
<td>MOPS</td>
<td>3-(N-Morpholino) propanesulfonic acid</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>My</td>
<td>million years</td>
</tr>
<tr>
<td>n</td>
<td>nano</td>
</tr>
<tr>
<td>O.D.</td>
<td>optical density</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
</tbody>
</table>
ψ pseudouridylate
PAGE polyacrylamide gel electrophoresis
PBS phosphate buffered saline
PCR polymerase chain reaction
PFGE pulse field gradient gel electrophoresis
Phe phenylalanine
PIPS piperazine-N,N'-bis(ethanesulfonic acid)
PMSF phenylmethyl sulphonyl fluoride
RNA ribonucleic acid
RNAP RNA polymerase
rpm revolutions per minute
rRNA ribosomal RNA
RT-PCR Reverse Transcriptase Polymerase Chain Reaction
σ sigma
SDS sodium dodecyl sulphate
Ser serine
SrRNA small ribosomal RNA
SSC small single copy
SSU small subunit
TAE tris-acetate-EDTA
TBÉ tris-borate-EDTA
TEMED N,N,N,N tetramethyl ethylene diamine
Tris tris (hydroxymethyl) aminoethane
tRNA transer RNA
TTP thymine triphosphate
Tyr tyrosine
UV ultraviolet
v/v volume/volume
w/v weight/volume
w/w weight/weight
X-gal 5-bromo-4-chloro-3-indolyl, β-D-galactoside
INDEX OF CONTENTS

ABSTRACT i
ACKNOWLEDGMENTS ii
ABBREVIATIONS iii
INDEX OF CONTENTS v
INDEX OF FIGURES AND TABLES ix

CHAPTER 1: INTRODUCTION 1
1.0.0 The phylum Apicomplexa 2
1.1.0 Malaria - P. falciparum and the life-cycle 3
1.2.0 The parasite DNA content 5
1.3.0 The 35 kb molecule - its genetic content 7
1.4.0 Gene content and organisation 8
1.4.1 The rRNA genes 8
1.4.2 The IR_A arm 10
1.4.3 The IR_B arm 11
1.5.0 Transcription 15
1.6.0 Codon usage 16
1.7.0 Copy number and location 16
1.8.0 Inheritance 17
1.9.0 Comparison between the 35 kb DNA and chloroplast DNA 17
1.9.1 The 35 kb molecule - the smallest plastid DNA identified to date 18
1.10.0 Monophyly versus polyphyly 18
1.11.0 The common ancestor 25
1.12.0 An alternative model 27
1.13.0 More plant connections 28
1.14.0 Aims of this project 29

CHAPTER 2: MATERIALS AND METHODS 38
2.1.0 General buffers and media 39
2.2.0 Culture of parasites and immortalised cell lines 40
2.2.1 Culture of P. falciparum 40
2.2.2 Culture of myeloma Sp2/0Ag14 cells 41
2.3.0 SDS-PAGE and Western blotting 41
2.3.1 Removal of contaminating reactivity of Rb72 antibody with E. coli protein 42
2.3.2 Pre-absorption of culture supernatant 2E2.1 or anti-E. coli RNAP antibody (EcB) with pure E. coli RNAP 43
2.3.3 Antibody affinity select 43
2.3.4 Immunofluorescence 44
2.4.0 Preparation of parasite DNA and plasmids 44
2.4.1 Purification of parasite genomic DNA and plasmids 44
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.2</td>
<td>Restriction endonuclease digests and preparation of vector DNA</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Purification of DNA from agarose gels</td>
</tr>
<tr>
<td>2.4.4</td>
<td>Preparation of vector and insert DNA</td>
</tr>
<tr>
<td>2.4.5</td>
<td>Ligations and transformations</td>
</tr>
<tr>
<td>2.5.0</td>
<td>DNA sequencing of double stranded plasmids</td>
</tr>
<tr>
<td>2.6.0</td>
<td>Synthesis of oligonucleotides for Polymerase Chain Reaction (PCR)</td>
</tr>
<tr>
<td>2.7.0</td>
<td>PCR from 35 kb or yeast DNA</td>
</tr>
<tr>
<td>2.7.1</td>
<td>Sequences of oligonucleotides used and annealing conditions</td>
</tr>
<tr>
<td>2.7.2</td>
<td>PCR protocol</td>
</tr>
<tr>
<td>2.7.3</td>
<td>Cloning of PCR products into the TA vector</td>
</tr>
<tr>
<td>2.8.0</td>
<td>Southern blots</td>
</tr>
<tr>
<td>2.8.1</td>
<td>Southern transfer</td>
</tr>
<tr>
<td>2.8.2</td>
<td>Labelling of probes and hybridisation</td>
</tr>
<tr>
<td>2.8.3</td>
<td>Hybridisation temperature and washing conditions</td>
</tr>
<tr>
<td>2.8.4</td>
<td>Southern blot stripping</td>
</tr>
<tr>
<td>2.9.0</td>
<td>Screening the 35-kb DNA library</td>
</tr>
<tr>
<td>2.10.0</td>
<td>Analysis of positive clones and subcloning</td>
</tr>
<tr>
<td>2.11.0</td>
<td>RNA extraction</td>
</tr>
<tr>
<td>2.12.0</td>
<td>Northern Blots</td>
</tr>
<tr>
<td>2.13.0</td>
<td>Primer extension analysis</td>
</tr>
<tr>
<td>2.14.0</td>
<td>Reverse Transcriptase PCR (RT-PCR)</td>
</tr>
<tr>
<td>2.15.0</td>
<td>Pulsed Field Gradient Electrophoresis (PFGE)</td>
</tr>
<tr>
<td>2.16.0</td>
<td>In vitro production of monoclonal antibodies</td>
</tr>
<tr>
<td>2.16.1</td>
<td>ELISA assay</td>
</tr>
<tr>
<td>2.16.2</td>
<td>Immunotyping of culture supernatants</td>
</tr>
<tr>
<td>2.17.0</td>
<td>Fusion Protein Synthesis - cloning and expression of the pxy/9 construct</td>
</tr>
<tr>
<td>2.17.1</td>
<td>Purification of the β-galactosidase fusion protein</td>
</tr>
<tr>
<td>2.18.0</td>
<td>Sequence analysis</td>
</tr>
</tbody>
</table>

**CHAPTER 3: SEQUENCING AND ANALYSIS OF THE 35 kb-ENCODED rpoC GENES**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.0</td>
<td>Introduction</td>
</tr>
<tr>
<td>3.2.0</td>
<td>Results and Discussion</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Amplification and sequencing of rpoC₁/rpoC₂</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Analysis of the rpoC₁ gene</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Analysis of the rpoC₂ gene</td>
</tr>
<tr>
<td>3.3.0</td>
<td>Summary and future work</td>
</tr>
</tbody>
</table>
# CHAPTER 4: THE IDENTIFICATION OF \(rpo\) GENE TRANSLATION PRODUCTS

4.1.0 Introduction 108
4.2.0 Results and Discussion 109
4.2.1 The synthesis of monoclonal antibodies to predicted polypeptide sequences 109
4.2.2 Characterisation of 2E2.1 and the anti-\(E.\ coli\) RNAP antibody (EcB) 110
4.2.3 The determination of shared immunogenicity between the RNAPs of \(E.\ coli\) and \(P.\ falciparum\) 113
4.2.4 Fusion protein expression 116
4.2.5 Immunofluorescence studies 119
4.3.0 Summary and future work 120

# CHAPTER 5: THE SEQUENCE AND ANALYSIS OF THE 35 kb ENCODED \(rps2\) GENE

5.1.0 Introduction 133
5.2.0 Results and Discussion 135
5.2.1 Analysis of the \(rps2\) gene 135
5.2.2 Location of the \(rps2\) gene within plastid genomes 135
5.2.3 Transcription of the \(rps2\) gene 137
5.3.0 Summary and future work 137

# CHAPTER 6: DESCRIPTION OF THE clpC-LIKE MOLECULE ENCODED BY THE 35 kb DNA

6.1.0 Introduction 143
6.1.1 The ClpC gene - an introduction 143
6.2.0 Results and Discussion 146
6.2.1 The \(P.\ falciparum\) clpC-like gene 146
6.2.2 clpC gene expression 151
6.3.0 Summary and future work 151

# CHAPTER 7: A CONTINUATION OF THE DESCRIPTION OF GENES ENCODED BY THE 4 kb AMPLIFICATION PRODUCT

7.1.0 Introduction 162
7.1.1 tRNA general structure 162
7.2.0 Results and Discussion 163
7.2.1 Discussion of the tRNA\(^{Gly}\) and tRNA\(^{Ser}\) identified on the 4 kb amplification product 163
7.2.2 Identification of homologues to ORFs-78 and -105 164
7.3.0 Summary and future work 165

# CHAPTER 8: THE IDENTIFICATION OF ORF470 HOMOLOGUES

8.1.0 Introduction 174
8.2.0 Results and Discussion 177
8.2.1 The identification of ORF470 homologues 177
8.2.2 The isolation of \(Synechocystis\) PCC 6803 recombinant cosmids 179
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3.0 Summary and future work</td>
<td>181</td>
</tr>
<tr>
<td>CHAPTER 9: THE IDENTIFICATION OF A SECOND \textit{tuf}-GENE IN \textit{P. falciparum}</td>
<td>188</td>
</tr>
<tr>
<td>9.1.0 Introduction</td>
<td>189</td>
</tr>
<tr>
<td>9.1.1 Protein biosynthesis - an introduction</td>
<td>190</td>
</tr>
<tr>
<td>9.2.0 Results and Discussion</td>
<td>192</td>
</tr>
<tr>
<td>9.2.1 Evidence for the presence of a second \textit{tuf} gene in \textit{P. falciparum}</td>
<td>192</td>
</tr>
<tr>
<td>9.2.2 Chromosomal mapping of the EF-Tu homologue of mitochondria</td>
<td>195</td>
</tr>
<tr>
<td>9.2.3 The 35 kb-encoded \textit{tufA} transcript</td>
<td>195</td>
</tr>
<tr>
<td>9.3.0 Summary and future work</td>
<td>196</td>
</tr>
<tr>
<td>CHAPTER 10: CONCLUSION</td>
<td>205</td>
</tr>
<tr>
<td>10.1.0 Conclusion</td>
<td>206</td>
</tr>
<tr>
<td>10.2.0 Future work</td>
<td>210</td>
</tr>
<tr>
<td>CHAPTER 11: REFERENCES</td>
<td>212</td>
</tr>
</tbody>
</table>
## INDEX OF FIGURES AND TABLES

### Figures:

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A simplified phylogeny of the Apicomplexa, Dinozoa and Ciliophora.</td>
<td>31</td>
</tr>
<tr>
<td>1.2</td>
<td>The life cycle of <em>Plasmodium</em>.</td>
<td>32</td>
</tr>
<tr>
<td>1.3</td>
<td>Diagrammatic illustration of the <em>P. falciparum</em> 35 kb circular DNA molecule at the start of this project.</td>
<td>33</td>
</tr>
<tr>
<td>1.4</td>
<td>Diagrammatic illustration of the fully sequenced <em>P. falciparum</em> 35 kb circular DNA molecule.</td>
<td>34</td>
</tr>
<tr>
<td>1.5</td>
<td>The conservation of the ribosomal protein gene clusters in <em>E. coli</em> and chloroplast genomes.</td>
<td>35</td>
</tr>
<tr>
<td>1.6</td>
<td>A proposed evolutionary pathway towards the formation of cryptomonads.</td>
<td>36</td>
</tr>
<tr>
<td>1.7</td>
<td>A proposed evolutionary pathway from cryptomonads to photosynthetic dinoflagellates.</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic representation of the protocol used to amplify the 1.3 kb PCR product encoding part of the <em>rpoC</em>, <em>C</em> genes.</td>
<td>81</td>
</tr>
<tr>
<td>3.2</td>
<td>Schematic representation of the PCR protocol used to complete the sequence data of the 35 kb DNA.</td>
<td>82</td>
</tr>
<tr>
<td>3.3</td>
<td>Nucleotide and predicted amino acid sequence of the 35 kb encoded <em>rpoC</em> and <em>rpoC</em> genes.</td>
<td>83</td>
</tr>
<tr>
<td>3.4</td>
<td>Representation of the <em>rpo</em> genes and neighbouring ORFs.</td>
<td>89</td>
</tr>
<tr>
<td>3.5</td>
<td>Primer extension analysis of the <em>rpo</em> operon.</td>
<td>90</td>
</tr>
<tr>
<td>3.6</td>
<td>RT-PCR analysis of the <em>rpoB/C</em> junction.</td>
<td>91</td>
</tr>
<tr>
<td>3.7</td>
<td>PILEUP analysis of the 35 kb encoded <em>rpoC</em> gene.</td>
<td>92</td>
</tr>
<tr>
<td>3.8</td>
<td>PILEUP analysis of the 35 kb encoded <em>rpoC</em> gene.</td>
<td>96</td>
</tr>
<tr>
<td>4.1</td>
<td>Illustration of the positions of IW376 and IW227 in the 35 kb encoded <em>rpoB</em> gene.</td>
<td>122</td>
</tr>
<tr>
<td>4.2</td>
<td>The identification of the <em>P. falciparum</em> B RNAP polypeptide.</td>
<td>123</td>
</tr>
<tr>
<td>4.3</td>
<td>The identification of <em>P. falciparum</em> polypeptides which have shared antigenicity with the anti-<em>EcB</em> antibody.</td>
<td>124</td>
</tr>
<tr>
<td>4.4</td>
<td>The identification of <em>P. falciparum</em> polypeptides which have shared antigenicity with the anti-<em>EcB</em> antibody and anti-<em>E. coli</em> B/B' antibodies.</td>
<td>125</td>
</tr>
<tr>
<td>4.5</td>
<td>The identification of <em>P. falciparum</em> polypeptides which have shared antigenicity with anti-<em>E. coli</em> σ and α antibodies.</td>
<td>126</td>
</tr>
<tr>
<td>4.6</td>
<td>Identification of the pxy/9 fusion product.</td>
<td>127</td>
</tr>
<tr>
<td>4.7</td>
<td>Purification of the pxy/9 fusion product.</td>
<td>128</td>
</tr>
<tr>
<td>4.8</td>
<td>Identification of pxy/9 fusion product by anti-Rb72.</td>
<td>129</td>
</tr>
<tr>
<td>4.9</td>
<td>Detection of the 35 kb encoded β subunit by Western blot.</td>
<td>130</td>
</tr>
<tr>
<td>4.10</td>
<td>Detection of the 35 kb encoded RNAP by immunofluorescence.</td>
<td>131</td>
</tr>
<tr>
<td>5.1</td>
<td>A comparison of the <em>rpoB/C</em>, <em>rps</em> 2, <em>tsf</em>, and <em>atp</em> gene order in various plastid genomes.</td>
<td>138</td>
</tr>
<tr>
<td>5.2</td>
<td>Nucleotide and predicted amino acid sequence of the <em>P. falciparum</em> rps 2 gene.</td>
<td>139</td>
</tr>
<tr>
<td>5.3</td>
<td>PILEUP analysis of the 35 kb encoded rps 2 gene.</td>
<td>141</td>
</tr>
</tbody>
</table>
6.1 Box representation of a typical clp gene. 153
6.2 Identification of the 35 kb encoded clpC-like gene. 154
6.3 The nucleotide and predicted amino acid sequence of the 3' end of the clpC-like gene, amplified as a part of the 4 kb PCR product. 155
6.4 PILEUP analysis of the 35 kb encoded clpC gene. 157
7.1 A schematic representation of the genes identified on the 35 kb DNA by PCR. 167
7.2 The primary structure of yeast tRNA\textsuperscript{Ala}. 168
7.3 Sequence data of tRNA\textsuperscript{Gly}, tRNA\textsuperscript{Ser}, ORF 79 and ORF 105. 169
7.4 Predicted secondary structures of tRNA\textsuperscript{Ser} and tRNA\textsuperscript{Gly}. 171
7.5 PILEUP analysis of ORF105 and ycr3. 172
8.1 Box diagram of the \textit{M. leprae} ORF2. 182
8.2 The location of ORF 2 in \textit{M. leprae}. 183
8.3 Identification of ORF470 homologues in \textit{Mycobacterium}. 184
8.4 Identification of an ORF470 homologue in \textit{Corynebacter} spp. 185
8.5 Identification of ORF470 homologues in \textit{Synechocystis} PCC 6803. 186
8.6 Physical map of \textit{Synechocystis} PCC 6803. 187
9.1 Box representation of the \textit{S. cerevisiae} \textit{tuf}M polypeptide. 198
9.2 Identification of the \textit{P. falciparum} nucleus encoded \textit{tuf}M homologue. 199
9.3 Detection of the 35 kb encoded \textit{tuf}A gene. 200
9.4 Chromosomal mapping of the nucleus encoded \textit{P. falciparum} \textit{tuf}M homologue. 202
9.5 Ethidium bromide stained PFGE gel. 203
9.6 Northern blot transcript data of the \textit{tuf}A gene. 204

\textbf{Tables:}

2.1 PCR primers and annealing conditions. 50
9.1 Comparison of \textit{tuf} EcoRI fragments. 201
CHAPTER 1
CHAPTER 1

INTRODUCTION

1.0.0 The phylum Apicomplexa

This thesis is concerned with the genome of the malarial parasite *Plasmodium falciparum* but also considers aspects of the evolutionary origins of apicomplexans as a whole. All members of the phylum Apicomplexa are parasitic, many being obligate intracellular parasites (Levine, 1970). They all have an apical complex consisting of an arrangement of microtubules, rhoptries and organelles involved in attachment and invasion of the host cell. Members of this phylum include the causative agent of malaria *Plasmodium*, as well as the opportunistic pathogens observed in AIDS patients - *Toxoplasma* and *Cryptosporidium*, and various other pathogens such as *Eimeria*, and the piroplasms *Babesia* and *Theileria*. Phylogenetic analysis of small subunit ribosomal RNA (SrRNA) suggests that the dinoflagellate and ciliate clades are close to the ancestors of this entire group (Johnson et al., 1988; Gajadhar et al., 1991; Barta et al., 1991; Wolters et al., 1991; Gagnon et al., 1993).

More recent analysis of SrRNA genes by Escalante and colleagues (1995) suggest that the species of *Plasmodium* diverged from other members of the Apicomplexa several million years ago (see Figure 1.1.). These data further suggest that whereas the phylum Apicomplexa probably originated at around the same time as the plant, fungi and animal kingdoms more than one billion years ago, *Plasmodium* diverged from other members of the phylum in the pre-Cambrian period, several million years (My) ago. This divergence would have occurred at the beginning of the Palaeozoic era prior
to the evolution of vertebrates, after which the *Plasmodium* radiation gave rise to the human parasites *P. falciparum*, *P. malariae*, *P. vivax* some 129 My ago.

**1.1.0 Malaria - *P. falciparum* and the life cycle**

Malaria is still one of the mankind’s most prevalent diseases, infecting over one third of the population. Approximately 120 million clinical cases of malaria are reported annually of which an estimated 1% result in death (World Health Records, 1993). Regions of endemic transmission include Africa, the Indian Subcontinent, South America and South East Asia. As yet no effective vaccine has been produced to protect against the disease, and there is increasing resistance of the parasite to chemotherapy, in addition to acquired resistance of the mosquito vector to insecticides.

The disease is transmitted to humans via the female *Anopheline* mosquito, the parasites being transmitted from the salivary glands of the mosquito directly to human blood whilst feeding. Four species of the genus *Plasmodium* can infect humans causing malaria, infection by *Plasmodium falciparum* causing the most severe form of the disease which can result in coma and death within two weeks after infection.

The *Plasmodium* parasite life cycle oscillates between a sexual phase within the invertebrate mosquito host and a haploid asexual phase in a vertebrate host.
A diagrammatic representation of the life cycle of *Plasmodium* is shown in Figure 1.2 indicating both the haploid and diploid phases of the parasite. The intra-erythrocytic phase of the cycle is responsible for the clinical symptoms of malaria, this part of the life cycle being the main target for vaccine production and chemotherapy. Each multiplication cycle within the blood stream takes between 36 and 48 hours to complete, depending on species. During these cycles of asexual multiplication the morphological forms of the parasite can be categorised into three main stages: rings, trophozoites and schizonts.

The asexual cycle begins with the rupture of exoerythrocytic schizonts in the liver, releasing thousands of merozoites into the blood stream. Erythrocytes are invaded by these merozoites producing the earliest morphological form of the erythrocytic parasite - the ring stage, so called because of its appearance when viewed under the light microscope. As the parasite matures into trophozoites, its morphology changes and it occupies more space within the infected erythrocyte. In addition, pigment formed by degradation of haemoglobin to haemozoin accumulates in the parasite’s food vacuole. Multiple rounds of DNA replication occurs during the trophozoite stage giving rise to a multinucleated mature schizont, comprising up to 32 progeny known as merozoites. The schizont eventually ruptures, releasing the merozoites into the blood stream to initiate new rounds of infection.

This cyclical process can be maintained indefinitely in infected red cells *in vitro* (Trager and Jenson, 1976). Gametocytes develop from a proportion of the released
merozoites and upon ingestion by the *Anopheline* mosquito during a blood meal, develop into gametes.

### 1.2.0 The parasite DNA content

Study of the blood stages of the parasite has shown that the amount of nuclear DNA remains constant for approximately 28-30 hours after infection of the erythrocyte until the early trophozoite stage. DNA replication is then initiated and its amount increases exponentially until schizogony (Inselburg and Banyal., 1984; Preiser *et al.*, 1996).

The haploid nuclear genome of *Plasmodium* is estimated to be $2 - 2.5 \times 10^7$ base pairs (bp) in size and has been shown by kinetochore counting and pulse field gel analysis to have a karyotype consisting of 14 chromosomes (Kemp *et al.*, 1987).

The identification of two kinds of extrachromosomal DNA molecules complicates the DNA composition of the malarial parasite. These molecules were first observed during a search in the 1980's for the *Plasmodium* mitochondrial genome. In one study, a satellite DNA band comprising approximately 1% of the total genome was observed in caesium chloride gradients of DNA from the monkey parasite *Plasmodium knowlesi* (Williamson *et al.*, 1985). Electron microscopic studies of mitochondrial extracts from *Plasmodium lophurae* identified a low density DNA (of similar buoyant density to the satellite DNA previously observed) that contained circular molecules with a contour length of 10.3μm (Kilejian *et al.*, 1975). Similar molecules, hereafter termed the 35 kilobase (kb) DNA, were later detected in *Plasmodium berghei* organelle extracts (Dore *et al.*, 1983), *P. falciparum* (Gardner *et al.*, 1987).
al., 1988) and in *Toxoplasma gondii* (Borst et al., 1984). Cruciform structures were evident in the *Toxoplasma* preparations and predicted to be induced by an inverted ribosomal RNA (rRNA) repeat on the molecule (Borst et al., 1984). The cruciform structures were suggested by Borst to result from ‘snap-back’ along an inverted repeat (IR) encoding ribosomal RNAs.

Initial sequence data of the small subunit rRNA (SrRNA) encoded by the 35 kb DNA of *P. falciparum* (Gardner et al., 1991b) suggested that it was of mitochondrial origin, although its homology to chloroplast DNA rather than mitochondrial DNA was noted. At around this time, a linear DNA molecule consisting of tandemly repeated 6 kb units in *Plasmodium yoelii* and *P. falciparum* also was reported (Vaidya et al., 1989; Aldritt et al., 1989). The sequence of this molecule revealed the presence of genes for cytochrome oxidase c subunit I (*cox1*) and cytochrome b (*cob*), in addition to fragments of rRNA genes. This 6 kb DNA molecule was shown by Southern blot analysis to be present in all members of the *Plasmodium* family (Suplick et al., 1988; Joseph et al., 1989) and was suggested to be the parasite’s mitochondrial genome, as is now accepted. In accordance with previous suggestions, the rRNA genes within the 35 kb circle were shown to be present as an inverted repeat flanked, as it turned out, by transfer RNA (tRNA) genes in a similar fashion to that observed in plastid and chloroplast genomes (Gardner et al., 1991b). This evidence, in addition to a comparison between the rRNA genes of the 6 kb and 35 kb DNAs (Feagin et al., 1992), and a phylogenetic comparison between the large subunit rRNA genes in chloroplasts and the 35 kb DNA (Gardner et al., 1993), showed little relation between
the two malarial extrachromosomal DNA molecules but substantial similarity between plastid genomes and the circular molecule.

1.3.0 The 35 kb molecule - its genetic content

The 35 kb circular DNA has been identified in *T. gondii* and *Eimeria tenella* (Wilson *et al.*, 1993) and partly sequenced in both of these organisms (P. Denny, unpublished data). Although the entire *P. falciparum* 35 kb circle has now been sequenced (Wilson *et al.*, 1996), at the start of this project only about two-thirds of the genome were known - compare Figure 1.3 with Figure 1.4. One of the principle aims of this thesis was to complete the sequencing of the plastid-like DNA, as will be discussed later. However, at this juncture, a basic description will given of the genes already known at the start of the project or found during it by others, and a detailed discussion of those genes sequenced by myself will be given in the relevant Results sections.

Comparison between the malarial 35 kb molecule and its homologue in *T. gondii* has revealed that the gene order in these two molecules is substantially conserved (Beckers *et al.*, 1995; and P. Denny, personal communication). The nucleotide composition is extremely A+T-rich, leading to a bias in amino acid composition in favour of seven amino acids, Ile, Asn, Lys, Leu, Tyr, Phe and Ser. A gene map of the complete 35 kb circle of *P. falciparum* is shown in Figure 1.4. Most of the genes encoded by this molecule are involved in gene expression and in this sense the molecule has been likened to the remnant plastid DNA of the parasitic plant *Epifagus virginiana*, where only four genes not involved in the transcription / translation apparatus have been retained by the molecule (Wolfe *et al.*, 1992).
1.4.0 Gene content and organisation

The 35 kb molecule contains an IR of 10.5 kb, encoding duplicated genes for the SrRNA and the large ribosomal RNA (LrRNA) in addition to 9 tRNAs. Between the duplicated regions within the IR a small single copy (SSC) region is assumed to be present but due to difficulty encountered in sequencing the centre of the repeat region this has not been verified. The IR also includes the initial three codons of ORF470 on the IR\textsubscript{A} arm and the identical first three codons of rps 4 on the IR\textsubscript{B} arm. The IR was shown in ‘snap-back’ experiments to form a cruciform structure by allowing complimentary strands of endonuclease restricted, denatured, \textit{P. falciparum} plastid DNA to snap-back after renaturation (Wilson \textit{et al.}, 1993). This procedure permits the formation of hairpin-shaped molecules with increased mobility in agarose gel during electrophoresis. However, when these molecules were electrophoresed on denaturing alkaline gel the DNA migrated according to its predicted size.

1.4.1 The rRNA genes

The SrRNA gene contains a 3' terminal signature sequence found only in bacteria, mitochondria and chloroplasts (Gardner \textit{et al.}, 1991b.). Phylogenetic analysis of the 3' region of the LrRNA gene indicated that mitochondrial genes are distantly related to this gene with those of chloroplasts grouping nearer the malarial sequence (Gardner \textit{et al.}, 1993).

The order of the LrRNA and SrRNA genes within the IR is different from that observed in other known chloroplasts (Palmer \textit{et al.}, 1985) since the LrRNA genes within the circle are distal rather than proximal to the small single copy region.
Transcription of the LrRNA genes is away from the SSC rather than towards it, in contrast to that observed in plastids. P. Denny (personal communication) has observed that the gene order within this region is conserved between the plastid molecules of *P. falciparum* and *T. gondii*. With a fixed rRNA gene order in both of these apicomplexan plastid DNAs, one may predict that the event giving rise to this gene arrangement is likely to have occurred once prior to, or at the formation of the apicomplexans.

Land plant chloroplasts range in size from the 216 kb of *Pelargonium hortorum* (Palmer 1992) to 121 kb of *Marchantia polymorpha* (Ohyama et al., 1986; Ohyama et al., 1988). The difference is partly due to the length of the IR, which varies from 76 kb in *P. hortorum* to 20 kb in *M. polymorpha*. By contrast, members of the legume family have totally lost their IR, resulting in a single copy of the rRNA genes (Harris et al., 1994). IRs are not maintained in the red algal plastid genomes of *P. purpurea* and *Griffithsia pacifica* (Harris et al., 1994) although the red alga *Cyanidium caldarium* contains an IR consisting of rRNA genes only. The plastid DNAs of the golden-brown algae *Olisthodiscus luteus* and *Ochromonas dancia* (Delaney and Cattolico, 1989; Shivji et al., 1992) and of the brown alga *Dictyota dichotoma* (Kushel and Kowallic, 1987) also contain IRs encoding only rRNA genes.

The 4.5S rRNA gene in the plastids of land plants and some ferns is located downstream to the 3’ end of LrRNA gene. In contrast to this, cyanobacteria, the plastids in all algae and the early land plant *M. polymorpha* encode the equivalent
gene as part of the LrRNA (Kossel et al., 1991; Bowman et al., 1979; Harris et al., 1994). The 35 kb circle LrRNA gene, however, does not encode a 4.5S rRNA gene.

Additionally, all chloroplast genomes encode a 5S rRNA; these genes are remarkably conserved and frequently used in phylogenetic placement studies. Interestingly, this gene has not been identified in the 35 kb molecule and in this respect, it is unlike all other plastids DNAs. That this gene may not be essential for translation is indicated by the finding that mitochondria except those in land plants, also lack the 5S rRNA gene (Harris et al., 1994) and yet are still fully active in polypeptide synthesis. The absence of the 5S rRNA may be relevant to the debate about whether the 35 kb circle is of mitochondrial or plastid origin.

1.4.2 The IRa arm

ORF470, ORF101 and ORF51 are the first three open reading frames encountered downstream of the IR. The function of these coding regions is unassigned, although as will be discussed in Chapter 9, an ORF470 homologue has been identified in red algae. The presence of this coding sequence on the 35 kb circle was taken to point to a rhodophyte lineage for the molecule (Williamson et al., 1994) and has allowed this laboratory to postulate how the plastid molecule became embedded in Plasmodium and other members of the phylum Apicomplexa, (Wilson et al., 1994). This point will be discussed fully later.

A gap of 3 nucleotides separates ORF101 and the 5' ATG start codon of the rpoB gene encoding the β subunit, which is followed by a second gene encoding the β'
subunit, of a DNA-directed RNA polymerase (RNAP). This RNAP has significant similarity to those utilised in prokaryotic organisms and chloroplast/plastid organelles (Gardner et al., 1991a). The presence of rpo genes was a suggestive clue to the plastid ancestry of the \textit{P. falciparum} 35 kb molecule. On plastid DNAs, the genes \textit{rpoB}, \textit{rpoC}_1 and \textit{rpoC}_2 encode the \( \beta \), \( \beta' \) and \( \beta'' \) subunits respectively of a prokaryotic like RNAP and this gene order is conserved in nearly all chloroplast genomes. By contrast, the mitochondrial RNAP is encoded by a different gene and is always nucleus encoded (Masters et al., 1987). A phylogenetic analysis of the highly diverged \textit{Plasmodium rpoB} gene showed that the closest plastid sequence available was that of the protist \textit{Euglena} (Gardener et al., 1994; Howe et al., 1992). Complete sequencing on the \textit{Plasmodium} 35 kb molecule of the 35 kb DNA has failed to reveal an \textit{rpoA} gene encoding the \( \alpha \)-subunit and this is assumed to have been translocated to the nucleus.

At the onset of this project, the complete \textit{rpoB} gene and part of the \textit{rpoC}_1 gene had been identified. A large proportion of this project was designed not only to complete the sequencing of the latter genes but to raise antibodies to the \( \beta \) subunit to determine whether the gene product was present.

\textbf{1.4.3 The IRB arm}

Immediately following the inverted repeat lies the ribosomal protein gene \textit{rps 4} (a component of the alpha operon in \textit{Escherichia coli}). As mentioned before this gene includes the identical first three codons to those in ORF470. Similarity at the amino acid level between the malarial \textit{rps 4} gene and other \textit{rps 4} polypeptides is low, being
limited to the initial 20 amino acids and a central region of the polypeptide. rps 4 binds the 16S RNA and is involved in the assembly of the 30S ribosome. The alpha operon which contains this gene in *E. coli* is not conserved in the 35 kb molecule (Figure 1.5).

A cluster of 10 tRNAs has been identified downstream of rps 4. One of these genes tRNA^{Leu} includes an intron, the only one on the 35 kb circle (see Chapter 7). Twelve ribosomal protein genes interspersed with three unidentified open reading frames are located immediately downstream from the tRNA cluster. The order of these genes is reminiscent of the S10, *spc*, *alpha* and *str* operons (Figure 1.5), yet it appears that in *P. falciparum* these operons have been fused. Again, these genes share little similarity to their prokaryotic counterparts making identification difficult.

The order of ribosomal protein genes within the cluster is like that described in plastids (Figure 1.5) and has helped with their identification. The gene order appears more similar to that of the red algal *Porphyra purpurea* plastid DNA than higher plant chloroplasts. Both the 35 kb circle and the *P. purpurea* plastid DNAs have retained a truncated *str* operon and encode the *tufA* gene. Baldauf and colleagues (1990) have identified the *tufA* gene in the nucleus of higher plants where the gene product (EF-Tu) is subsequently imported into the chloroplast. An additional similarity between the 35 kb genome and the plastid DNA of *P. purpurea* is that, unlike higher plant chloroplasts, both continue to encode rps 5, rps 17 and rpl 4.
Most of the ribosomal protein genes located in the S10 operon are present on the circular DNA. rpl 4 normally, the third gene in the operon, starts the malarial counterpart. This gene in *E. coli* is under autogenous control, regulating the expression both of itself and other genes within the operon (Zengel and Lindhal *et al.*, 1994). rpl 23 is poorly conserved yet appears to be complete, whereas rpl 2 is truncated at both ends of the polypeptide. However, a conserved region at the C-terminus has been maintained. Both rps 19 and rps 3 are both poorly conserved, but the inferred gene order maintains that found in comparison to the prokaryotic S10 operon. The rpl 16 and rps 17 genes are separated by a frame shift and a one nucleotide overlap, rps 17 marking the end of the S10-like operon.

In *E. coli*, rpl 14 appears as the first gene within the spc operon and it also starts the malarial version. The rpl 14 gene is well conserved and is followed by rps 8 and rpl 6. In turn, these genes precede rps 5 which again is poorly conserved. However, the regions within this gene which confer sensitivity to spectinomycin and streptomycin in E. coli have been maintained. Homologues to rps 5 have been identified in the plastid genome of *P. purpurea* (Harris *et al.*, 1994) and the cyanelle in *Cyanophora paradoxa* (Loeffelhardt *et al.*, 1991; Michalowski *et al.*, 1990) but are not present in land plant chloroplasts. ORF91 separates this gene from rpl36 which, although A+T rich (85%) is well conserved at the amino acid level and completes this spc-like operon.

In the *P. falciparum* plastid DNA, the genes belonging to the alpha operon of *E. coli*, are only two in number and have been separated, the rps 4 gene being located at the
start site of transcription of the IR$_B$ arm whereas rps 11 lies proximal to rpl 36, preceding the beginning of the str operon.

rps 12 is the most conserved of the ribosomal protein gene homologues, maintaining those sites influencing sensitivity or tolerance to streptomycin in other organisms. This gene is located in the E. coli str operon and in this respect the gene order has been maintained because the next gene is for rps 7. The rps 7 gene is followed by a tufA gene thereby completing the malarial version of the operon.

The tufA gene encodes the polypeptide elongation factor EF-Tu. Similarity between this gene and that of E. coli at the amino acid level is only 45%. However, the regions involved in associating with GTP and defining the GDP binding domains appear to be conserved at the DNA level although the malarial polypeptide has not yet been identified. A more detailed description of this gene is presented in Chapter 9.

The tufA gene is not encoded in the chloroplasts of higher plants but has been identified in the cyanelle of Cyanophora, the plastids of the red alga P. purpurea and the thermophilic red alga Galderia sulphuraria and in the green algal genomes of the families Ulvophyceae, Chlorophyceae and Charophyceae (Baldauf et al., 1990). These last mentioned families are the presumed progenitors of green land plants. Significantly, transfer of the tuf gene to the nucleus probably occurred at around the period of land plant formation, since this gene’s presence is no longer observed in the chloroplast of M. polymorpha which represents the early land plant lineage. This indicates a similarity between the 35 kb plastid and green algae and red algae,
possibly suggesting that this gene order was established by a common ancestor to these organisms prior to the further divergence within each lineage (see below -The common ancestor (1.11.0)).

The EF-Ts gene utilised by prokaryotes in translation has been retained in the plastid genomes of both *P. purpurea* and *G. sulphuraria*, but it is not present on the malarial plastid DNA. This suggests that lateral transfer of this gene to the parasite nucleus has probably occurred.

An ORF of 78 amino acids and a cluster of 4 tRNAs (deciphering the anticodons Phe, Glu, Gly and Trp) followed by a second ORF of 129 amino acids separates the *tuf* gene from a clpC-like gene.

clpC occurs on the red algal plastid genome of *Porphyra* but has not been retained on the chloroplasts of higher plants. The retention of a clpC-like gene and ORF470 in both the plastid of *P. purpurea* and the 35 kb molecule implies that there may be an ancestral link between rhodophytes and the Apicomplexan plastid. A detailed summary of the function of clp genes and the relevance of a clp-homologue on the 35 kb molecule will be discussed in Chapter 6.

1.5.0 Transcription

Transcripts have been identified for all the tRNAs (Preiser *et al.*, 1995), the rRNAs (Gardner *et al.*, 1989 and 1991(b)) the rpo subunits (unpublished data and Gardner *et al.*, 1991(a)), ORF470, tufA, the clp gene and some of the ribosomal protein genes.
Evidence suggests that the tRNAs (Preiser et al., 1995), the S10-like operon and the rpo genes are transcribed polycistronically.

1.6.0 Codon usage

The codon usage of genes encoded by the 35 kb circle is different from that of the nuclear and mitochondrial encoded genes. *Plasmodium* nuclear DNA has a high A+T content (Musto et al., 1995) but the plastid DNA surpasses this bias with the plastid genome consisting of approximately 86% A+T nucleotides. The GC content of mitochondrial DNAs in insects and mammals also has diverged in the course of molecular evolution (Osawa and Jukes 1989) and a change in GC content in plastids has occurred during the evolution of the plastid. Most purple bacteria and cyanobacteria have a high GC content whereas plastid DNA has a low GC content. Reasons for this change in codon usage and codon bias include tRNA availability and codon-anticodon hydrogen (H)-bond strength (Ticher and Graur 1989). However, these suggestion do not account for the change towards AT richness. An AT bias appears to be a common feature of the organelle genomes of plants, animals and fungi (Aota et al., 1988). Plastid DNA replicates more rapidly than nuclear DNA and in order to be more energetically efficient during replication, a high AT content may be favoured due to the ease of breaking the H-bonds between AT base pairs than GC base pairs.

1.7.0 Copy number and location

The copy number of the 35 kb DNA has been approximated by Southern blot techniques to be 1-2 copies per cell at ring stage development (Preiser et al., 1996).
Genetic-crossing experiments indicate that the 35 kb molecule resides in an organelle not associated with the nucleus (Vaidya et al., 1993). This finding is in agreement with previous studies by Wilson (1992) which showed by Polymerase Chain Reaction (PCR) amplification of \textit{rpoB} and rRNA sequences on subcellular fractions of the organellar DNA of the rodent parasite \textit{P. yoelii} that the 35 kb molecule was not associated with either the nucleus or mitochondria. An organelle of unknown function has been described in \textit{Plasmodium} - the spherical body; \textit{Toxoplasma} - the Golgi-adjunct; and \textit{Eimeria} - the holzyylinder (Siddall et al., 1992). These multi-membranous organelles have been proposed to house the plastid DNA but no unequivocal information has been obtained in favour of this.

\textbf{1.8.0 Inheritance}

Genetic crossing using probes specific for the 6 kb and 35 kb malarial DNAs have confirmed that these molecules are inherited maternally like the mitochondria (Vaidya et al., 1993; Creasey et al., 1994). In fact, Vaidya (1993) showed by fluorescence studies that male \textit{Plasmodium} gametes do not contain any cytoplasmic organelles. This finding is in contrast to the male gametes of \textit{Eimeria} and \textit{Toxoplasma}, within which a mitochondrion has been identified (Scholtyseck et al., 1972).

\textbf{1.9.0 Comparison between the 35 kb DNA and chloroplast DNA}

The genetic composition and gene order within the \textit{Plasmodium} circular DNA suggest a relatedness between this molecule, plant chloroplasts, and algal plastids. The similarities observed between these molecules are discussed below.
1.9.1 The 35 kb molecule - the smallest plastid DNA identified to date

The Apicomplexan plastid is approximately half the size of the 70 kb plastid DNA found in the parasitic, non-photosynthetic plant *Epifagus virginiana* (Wolfe *et al.*, 1992) and the 73 kb plastid DNA in the non-photosynthetic euglenoid *Astasia longa* (Gockel *et al.*, 1994). In these cases, loss of photosynthesis correlates with a loss of genes involved in the photosynthetic process and in a concomitant reduction in genome size.

One can postulate, therefore, that retention of the plastid genome in these organisms must confer some other functional advantage on the organism, especially in the case of *Epifagus*, where a great number of the genes involved in gene expression, including the rpo genes, have also been lost (Wolfe *et al.*, 1992). This also is the case argued for the conservation of the 35 kb molecule in the malaria parasite, although far greater characterisation of this molecule is yet to be undertaken before a function can be determined. According to our model for the acquisition of the plastid in apicomplexans (see below), those genes involved in photosynthesis would have been lost as a result of or, resulting in, the inability to photosynthesize. How a plant-like molecule arrived in the parasite has been postulated by Wilson (1994) and is based upon the prediction that the 35 kb molecule is located in the spherical body.

1.10.0 Monophyly versus polyphyly

At this juncture it is imperative to underscore that every argument in favour of a polypyphtletic origination of plastid and mitochondrial organelles has been counterargued by a proposal for the opposite i.e. a monophyletic origin.
In this context, I have taken Gray's (1992) definition for monophyletic origin as being acquired from a single prokaryotic progenitor endosymbiosed by a eukaryote. By contrast, polyphyletic origins appertain to an organelle originating in different eukaryotic lineages due to independent endosymbiotic events with different prokaryotic progenitors.

An example of the type of argument found in the literature relating to monophyletic origins is that for chloroplasts. It is now generally accepted that these organelles were acquired by a single endosymbiotic event between cyanobacteria-like organisms and a phagocytic eukaryote. However, this theory is based upon phylogenetic trees utilising 16S rRNA sequences (Morden et al., 1992). rRNA genes, the “Rosetta stone” of phylogeny, have been relied on because of their ubiquitous presence in all organisms, their function remaining identical in all life forms. Other phylogenetic trees, for example, based on the PsbA locus supports the monophyletic hypothesis (Lockhart et al., 1992). The counter-arguments for the polyphyletic origins of chloroplasts are based on the presence of different pigments identified in various photosynthetic organisms, and phylogenetic analysis based on the rbcL genes as will now be discussed.

There are essentially three types of photosynthetic eukaryotes which are distinguished by their pigment composition (Christensen, 1964). Red algae (Rhodophyta) contain chlorophyll a and phycobilins. Green algae (Chlorophyta and Gamophyta) and euglenoid algae (Euglenophyta) contain chlorophylls a and b. Chromophytes utilise
chlorophylls \( a \) and \( c \). Two of the chromophyte phyla - Cryptophyta (cryptomonads) and Eustigmatophyta contain chlorophylls \( a \) and \( c \) and phycobilins.

Raven (1970) proposed that each of the three types of photosynthetic systems in algae arose as a result of at least three separate endosymbiotic events. Cyanobacteria were the only known prokaryotes to contain both phycobilins and chlorophyll \( a \) and were therefore the obvious choice for the progenitor of red algae. The remaining two prokaryotic endosymbionts were assumed to be free-living green or yellow organisms, no longer existent. The identification by Lewin (1976) of the prochlorophytes (prokaryotes containing chlorophyll \( a \) and \( b \) only) added credence to this hypothesis suggesting that these organisms were the progenitors to euglenoids and chlorophytes.

This proposal has since been rebuffed on the basis that 16S rRNA trees (Urbach et al., 1992) and trees of \( rpoC_1 \) fragments (Palenik and Haselkorn., 1992) from prochlorophytes indicate that these organisms originated from cyanobacteria. The alternative explanation is that the ability to harvest light using the chlorophyll \( b \) system has developed independently at least four times in the course of evolution; once in the ancestry of the plant lineage, and on three independent occasions within each of the three genera of the prochlorophytes. The differences between chlorophylls \( a \) and \( c \) and, \( a \) and \( b \) are biochemicaly simple and it has been proposed by Cavalier-Smith (1994) that the acquisition of the additional chlorophyll \( b \) or chlorophyll \( c \) occurred as natural events in red algae leading to the formation of separate algal lineages.
Chapter 1

The other line of argument in favour of a polyphyletic origin of chloroplasts is based on phylogenetic trees of the *rbcL* and *rbcS* genes, encoding the large and small subunits respectively of ribulose-1,5-bisphosphate carboxylase/oxidase (RUBISCO). In rhodophytes and chromophytes, these genes are plastid encoded and constitute a eubacterial-like operon (Valentin and Zetsche, 1990; Kostrzewa et al., 1990). In chlorophytes, euglenophytes and plants, *rbcS* is a nucleus encoded gene (Palmer, 1985). Phylogenetic trees of these two genes placed rhodophytes (containing chlorophylls *a/b* and phycobilins), and chromophytes (containing chlorophylls *a/c*) and cryptomonads (containing chlorophylls *a/c* and phycobilins) with β-purple bacteria and not with cyanobacteria (Morden and Golden, 1991). To accept the suggestion that rhodophyte plastids did not originate from cyanobacteria would be incompatible with the strong similarities to cyanobacteria with regard to pigmentation and structure of the photosynthetic membrane.

In order to maintain the monophyletic hypothesis for the origin of plastids, the explanation provided for the *rbcLS* phylogenetic trees is that the operon has been laterally transferred from a β-protop bacterium to a cyanobacterial ancestor (Martin et al., 1992; Morden et al., 1992; Douglas et al., 1990). If this were the case, the plastid genomes in these algae would be considered to be an evolutionary mosaic.

At this point it is obvious that the inferences drawn from phylogenetic trees depend on the gene being assessed. Incorrect trees are formed if the rates of evolution of a given gene are unequal in different lineages (Penny et al., 1991; Felsenstein 1978) or where two or more lineages have independently increased the GC or AT nucleotide content.
(Penny et al., 1987). The latter problem may be corrected by calculating phylogenetic trees based on the 1st and 2nd codon positions only. Nevertheless, it must be remembered that phylogenetic trees represent the evolution of a given gene and not necessarily of the organism.

The structure and organisation of the plastid DNAs of rhodophytes, chromophytes and green plants has been compared and contrasted by Shivji and colleagues (1992) in an attempt to bypass the quandary as to whether the evolution of a given gene is a true reflection of the evolution of the entire organism. Their conclusion was indecisive and suggested that although the overall similarities between plastid gene organisation supports monophyly of plastids, more data need to be assessed before any doubts relating to lateral transfer of genes into plastids could be ruled out. This is in addition to understanding where the differences in gene content within different plastids originated.

Additional comparisons, such as between the single-copy nuclear gene (GapA) and the plastid encoded glyceraldehyde-3-phosphate dehydrogenase (GAPDH) of the marine red algae Gracilaria verrucosa (Zhou and Ragan, 1994) and Chondrus crispus (Liaud et al., 1994) suggest that both algal and higher plants descended from a common progenitor.

Therefore the balance of the data indicates that the primary chloroplast was derived from the endosymbiosis of a cyanobacteria-like organism by a eukaryote, thereby forming the first red algae (Morden et al., 1992; Douglas et al., 1993). The
cyanobacteria-like organism would presumably have contained a double lipoprotein membrane encasing a peptidoglycan layer (refer to Figure 1.6). The plastids in red algae, chlorophytes and land plants are surrounded by a double membrane probably relating to the double lipoprotein membrane in cyanobacteria with the presumable loss of the peptidoglycan layer (Cavalier-Smith., 1987). This layer has been maintained in the membrane surrounding the cyanelle in the photoautotrophic protist *Cyanophora paradoxa* (Aitken and Stanier, 1979).

It is presumed that the original red algal endosymbiont also lost the engulfing phagosome membrane of the host cell leaving a double-membrane bound plastid. However, as depicted in Figure 1.6 the plastids in euglenophytes and most dinoflagellates have retained what may be the plasma membrane of the original host giving rise to plastids surrounded by three membranes. Secondary endosymbiotic events have been proposed to account for the four membrane bound plastids in chromophytes. Therefore in this case, the inner two membranes belong to the plastid, the third membrane is that of the red alga-like host plasma membrane and the fourth membrane may be that of the cryptomonad endoplasmic reticulum (this outermost membrane characteristically has ribosomes on its cytoplasmic surface).

A second eukaryote was proposed to have phagocytosed the primary endosymbiont giving rise to cryptomonads see Figure 1.6. Cryptomonads contain a double membrane bound nucleomorph located in the periplastidal space between the plastid and host nucleus (Greenwood *et al*., 1977). These organisms have been described as pivotal in the evolution of the plastid in dinoflagellates - see Figure 1.7, the
nucleomorph being the remnant of the primary eukaryotic endosymbiont nucleus. Sequence analysis of the rRNA genes of the nucleomorph of Cryptomonas Φ suggests that these genes are from a red algal source, supporting the idea that they originated from a red algal endosymbiont and that the periplastidal space represents the former algal cytosol (Wilcox and Wedemayer, 1984b; McFadden et al., 1994; Cavalier-Smith, 1994). Analysis of the nucleus genome of the cryptomonad Pyrenomonas salina has demonstrated that the original eukaryote organism was related to green algae (Maier, 1992). It has been suggested that the nucleomorph is responsible for the maintenance of the periplastidal space since the dinoflagellate Gymnodium acidotum no longer has a nucleomorph or periplastidal space (see below)(Wilcox and Wedemayer, 1984b).

Dinoflagellates are marine organisms and are often photosynthetic having acquired brown plastids by the endosymbiosis of chromophytic algae (containing chlorophylls a and c). Dinoflagellate plastids are surrounded by a triple-membrane. The photosynthetic dinoflagellates Amphidium wigrense and Gymnodium acidotum contain blue-green chloroplasts. G. acidotum contains an endosymbiotic cryptomonad (an example of a four membraned plastid) (Wilcox and Wedemayer., 1984a), and it has been proposed that upon further reduction of the cryptomonad genome, a partially degenerate endosymbiont similar to that found in G. acidotum evolved to produce the triple membrane bound plastid found in A. wigrense (Wilcox and Wedemayer. 1984b). In this example, the blue-green chloroplasts (containing both soluble phycobilins and chlorophylls a and c), identified in these two organisms were proposed to be derived from a cryptomonad endosymbiont. Wilcox (1984b) has
proposed that \textit{G. acidotum} is actually an evolutionary intermediate between cryptomonads and dinoflagellates; \textit{G. acidotum} still maintains the cryptomonad nucleomorph and periplastidal space which have been subsequently lost in \textit{A. wigrense}. Cavalier-Smith (1994) has proposed that the cryptomonad developed a chlorophyll \textit{c} system in addition to chlorophyll \textit{a} and was then acquired by another eukaryotic organism to form an algal chromobiote which was endosymbiosed by a dinoflagellate, giving rise to photosynthetic dinoflagellates. This hypothesis agrees with the work of Wilcox and colleagues (1984b). Thus, multiple rounds of endosymbiosis have been proposed to account for the multi-membranous plastid of dinoflagellates. This proposal is consistent with the other work that suggests that plastids \textit{per se} have a monophyletic origin.

\textbf{1.1.0 The common ancestor}

This laboratory, has argued that the 35 kb molecule might be a homologue of dinoflagellate plastid DNA since, as mentioned earlier in this Introduction, these organisms are the presumed progenitors of the apicomplexans. Approximately 800 million years ago certain dinoflagellates are proposed to have adopted a parasitic existence in polychaete worms, at which point a loss of photosynthesis is conjectured to have occurred. Until further analysis of dinoflagellate plastids has been performed, this proposal for the acquisition of the 35 kb molecule has little basis, but, it does succeed in interweaving the presence of the plastid in \textit{Plasmodium} spp., using the phylogenetic hypotheses for plastid acquisition and evolution of the phylum.
There are counter arguments to the hypothesis that the 35 kb plastid derives from dinoflagellates and is of red algal in origin. The SrRNA gene from *Toxoplasma gondii* 35 kb DNA trees with that of the euglenoids (Egea *et al.*, 1995) in agreement with an analysis by Gardner (1994) in this laboratory on the *rpoB* gene. This suggests that the *Plasmodium* and *Euglenoid rpoB* sequences diverged early from the other plastid sequences. *Euglenoid* plastids contain chlorophylls *a* and *b* (*a/b*) and are enveloped by three membranes. These algae are proposed to have evolved by one of two mechanisms. One hypothesis is that the euglenophyte plastids evolved by the selective uptake of isolated plastids from other chlorophyll *a/b* eukaryotes by a non-photosynthetic flagellate (Whatley *et al.*, 1979; Whatley, 1981). The second hypothesis postulates that the plastid was gained by a secondary endosymbiotic event involving an alga containing chlorophylls *a/b* (Gibbs, 1981). Gibbs’ hypothesis requires that the eukaryotic endosymbiont was reduced so that only the plastid and plasmalemma membranes remained. Phylogenetic trees of *rRNA* sequences (Markowicz *et al.*, 1988; Palmer, 1991) place euglenophytes together with *C. paradoxa/rhodophytes/chromophytes*. However, *E. gracilis* branches from the main eukaryotic line before all other algae including those containing chlorophylls *a* and *b*, namely rhodophytes. In order to maintain a monophyletic origin for plastids the primary endosymbiotic event between a member of the cyanobacterial radiation and a eukaryote must have occurred prior to the divergence of the euglenoids which acquired their plastids by a later endosymbiotic event.

On gene content alone, a red algal rather than euglenoid connection to the 35 kb molecule is stronger because the *tuf* gene, clpC and ORF470 are no longer maintained
on the plastids of *Astasia* and *Euglena* but they are encoded on the plastids of the rhodophyte alga *P. purpurea*. Unfortunately, phylogenetic analysis of the *Plasmodium* 35 kb molecule has been confounded because of its high AT nucleotide content. This has meant that the trees are biased towards other lineages containing AT-rich homologues (such as euglenoids) and corrections have had to be made in order to obtain true alignments (Gardner *et al.*, 1994).

One way to circumvent this problem has been to construct trees using only the nucleotides at the first and second codon positions, as described above. Other methods include the analysis of nucleotide sequences using the model of Tamura and Nei (1993) that allows for a transition/transversion bias where transitions occur between similar molecules i.e. purines (A and G) or pyrimidines (C and T). This model is therefore of use for the analysis of AT rich DNA such as that found in the 35 kb molecule. However, the extreme nucleotide composition of the 35 kb plastid DNA continues to be a problem and may cause discrepancies in phylogenetic analyses.

**1.12.0 An alternative model**

A “simpler” mechanism for the acquisition of the Apicomplexan plastid could be that a non-plastid containing ancestor of this phylum procured a prokaryote by a simple phagocytotic event. A transfer of genes from the prokaryote to the symbiont nucleus would therefore have to have occurred at an accelerated rate to form a highly derived molecule with a genetic composition resembling that of chloroplasts. This proposal also infers that these gene re-arrangements leading to the formation of the 35 kb circle (and for that matter, all other plastids) are in a sense pre-determined i.e. that the
general conformity in the gene order of the ribosomal proteins and *rpo* operon identified in plastids is not an evolutionary artefact but would have occurred in all of these molecules independently.

Until the location of the 35 kb molecule has been determined and dinoflagellate plastid DNA sequences determined, all phylogenetic theories appertaining to whether the 35 kb molecule is of a dinoflagellate origin are circumspect.

### 1.13.0 More plant connections

The similarities between apicomplexans and plants do not end at the presence of the 35 kb molecule. Other plant-like features have been identified in the parasite's genome including an enolase gene containing insertions and deletions (indels) similar to those observed in plants (Read *et al.*, 1994). A plant herbicide, the triazine derivative toltrazuril, has been shown to have an affect on *Eimeria* spp. (Melhorn *et al.*, 1984), in addition to the alga *Psaltriomonoas lanterna*. *P. lanterna* contains a multi-membranous organelle (the thylakosome) which in its mature stages shares similarities with the apicomplexan Holzylinder and in its immature forms resembles cyanobacteria, lacking phycobilisomes. Traces of chlorophyll *a* and the *psbA* gene encoding the D1 protein necessary for photosynthesis (Michel and Deisenhofer *et al.*, 1986; Hackstein *et al.*, 1995) have been identified (Hackstein *et al.*, 1994; Hackstein *et al.*, 1995). Hackstein (1995) additionally identified the presence of chlorophyll *a* and the *psbA* gene in *T. gondii*. Hackstein and colleagues (1995) have therefore suggested that apicomplexans should be sensitive to herbicides that affect the photoreactive centre and supports a phylogenetic relationship between apicomplexans
and cyanobacteria. Finally, the *Plasmodium* histone H2A gene has been shown to tree with the unicellular alga *Volvox* and the conifer *Picea*. These results indicate that if these organisms are not related, they diverged at close evolutionary times (Wilson *et al.*, 1994).

### 1.14.0 Aims of this project

1. As previously mentioned, a large part of this thesis was devoted to completing the sequence of the 35 kb molecule. The sequence for a 5.3 kb section is provided in the following sections of this thesis.

2. Antibodies were raised to the β subunit of the RNA polymerase (RNAP) in order to determine whether this protein was expressed and to determine the location of this polypeptide in the parasite.

3. As a step towards determining a function for ORF470, bacterial homologues have been identified. Chapter 8 presents data supporting the bacterial origins of this gene, with homologues determined in *Mycobacterium leprae*, *M. tuberculosis*, *M. smegmatis*, *Corynebacterium* and, *Synechocystis* PCC 6803 but not in higher organisms such as yeast.

4. Finally, the *tuf* gene on the 35 kb circle has become a subject of great interest for this laboratory. It was necessary to determine whether a similar *tuf* gene encoding a mitochondrial elongation factor was encoded in the nucleus. Accordingly, Southern
blot techniques were employed using a yeast mitochondrial *tuf* gene to search for a nucleus encoded homologue in the parasite and to map its chromosomal location.
**Figure 1.1**

A simplified phylogeny of the Apicomplexa, Dinozoa and Ciliophora.

This Figure has been adapted from Escalante and Ayala (1995).

The scale highlights the first fossilised appearance of the vertebrate groups. The uncertainty of the time of the appearance of the animal, plant and fungi kingdoms is indicated in parentheses.
Figure 1.2


Both the vector and human phases are represented.

**Phase 1. Fertilisation:** A female *Anopheles* sucks a blood meal from an infected person. Gametocytes become free male or female gametes. In a sudden exflagellation the male gamete produces up to 8 flagella and fertilises a female gamete to produce a zygote. The developing ookinete bores into the stomach wall to become an oocyst.

**Phase 2. Sporogony** - asexual development in the mosquito. The oocyst grows and divides to produce invasive sporozoites. The mature cyst ruptures releasing free sporozoites which migrate through the mosquito to the salivary glands.

**Phase 3. Hepatic schizogony** - asexual development in the liver. On feeding, the mosquito injects sporozoites into the blood which invade liver cells and form hepatic trophozoites. On growth, thousands of invasive merozoites are produced. The infected liver cells burst, releasing the merozoites into the blood.

**Phase 4. Erythrocytic schizogony** - asexual development in the blood. Invasion of erythrocytes by merozoites and the formation of erythrocytic trophozoites. The production of merozoites which on rupture of the red blood cell re-invade fresh erythrocytes. Some merozoites form gametocytes which develop further in the mosquito host.
Figure 1.3
Diagrammatic illustration of the *P. falciparum* 35 kb circular DNA molecule at the start of this project.

Genes are represented by filled boxes which are proportional to gene length. Partially filled box (*rpoC*) denotes incomplete sequence. Transfer RNA genes are identified by the single letter code for the cognate amino acid.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSU</td>
<td>small subunit</td>
</tr>
<tr>
<td>LSU</td>
<td>large subunit</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
</tbody>
</table>
**Figure 1.4**

Diagrammatic illustration of the fully sequenced *P. falciparum* 35 kb circular DNA molecule.

Genes are represented by filled boxes which are proportional to gene length. Genes on the outer circle are transcribed clockwise. Genes on the inner circle are transcribed counter-clockwise. Transfer RNA genes are identified by the single letter code for the cognate amino acid, with the anticodon in parentheses. Asterisk indicates the presence of an intron.

IR\textsubscript{A} and B - inverted repeat

SSU - small subunit

LSU - large subunit

ORF - Open Reading Frame.
P.falciparum pIDNA
(35kb)

IR_A
IR_B

LSU rRNA
SSU rRNA
LSU rRNA
SSU rRNA
Figure 1.5
The conservation of the ribosomal protein gene clusters in *E. coli* and chloroplast genomes (adapted from Harris *et al.*, 1994).

Shaded boxes in dark blue indicate genes that have been lost but identified elsewhere on the genome. Ovoid box in yellow represents *P. falciparum* ORF 91. In *Cyanophora paradoxa* and *Porphyra purpurea* arrows indicate that the operon starts with rpl 3 and ends with S12, S7 and *tufA* and S10 genes. In *P. falciparum* the operon begins with rpl 4 and ends with S12, S7 and *tufA*. 
Figure 1.6

A proposed evolutionary pathway towards the formation of cryptomonads.

Abbreviations: - N’ (blue) - nucleus of original eukaryotic phagotrophic host; P-plastid; N (yellow) - nucleus of second eukaryotic phagotrophic host; dn - nucleus of non-photosynthetic dinoflagellate.
photosynthetic prokaryote → phagotrophic eukaryote → photosynthetic eukaryote (red alga-like eukaryote)

peptidoglycan layer → primary endosymbiosis

primary endosymbiosis → loss of peptidoglycan layer (except in the cyanelle of C. paradoxa) and loss of phagocytic membrane

photosynthetic eukaryote (red alga-like eukaryote)

plasma membrane

cryptomonad

endoplasmic reticulum

nucleomorph

periplastidal space

secondary endosymbiosis

second phagotrophic eukaryote

plasma membrane
Figure 1.7
A proposed evolutionary pathway from cryptomonads to photosynthetic dinoflagellates.

Abbreviations:- N' (blue) - nucleus of original eukaryotic phagotrophic host; P-plastid; N (yellow) - nucleus of second eukaryotic phagotrophic host; dn - nucleus of non-photosynthetic dinoflagellate.
cryptomonad

nucleomorph

periplastidal space

cryptomonad

non-photosynthetic dinoflagellate

transfer of DNA to host nucleus and loss of membranes

Amphidinium wigrense

Gymnodinium acidotum

transfer of DNA to host nucleus and loss of membranes
CHAPTER 2
CHAPTER 2

MATERIALS AND METHODS

2.1.0 General buffers and media

<table>
<thead>
<tr>
<th>Buffer</th>
<th>Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB*</td>
<td>10.00 g Bacto-tryptone, 5.00 g Bacto-yeast extract, 10.00 g NaCl in 1 L, pH 7.0</td>
</tr>
<tr>
<td>NZY*</td>
<td>10.00 g NZ amine (casein hydrolysate), 5.00 g Bacto-yeast extract, 5.00 g NaCl, 1.84 g MgCl2.H2O in 1 L</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffered saline (140 mM NaCl, 3mM KCl, 3mM KH2PO4, 140mM Na2HPO4, pH 7.4)</td>
</tr>
<tr>
<td>SM</td>
<td>Sodium magnesium buffer (100 mM NaCl, 10 mM MgSO4, 50 mM tris-HCl pH 7.5, 0.01% w/v gelatin)</td>
</tr>
<tr>
<td>SOC</td>
<td>20.00 g Bacto-tryptone, 5.00 g Bacto-yeast extract, 0.50 g NaCl, 0.25 mM KCl, 10 mM MgCl2, 20 mM glucose, pH 7.0</td>
</tr>
<tr>
<td>SSC</td>
<td>Sodium chloride-sodium citrate buffer (1 × SSC is 0.15 M NaCl, 0.015 M tri-sodium citrate)</td>
</tr>
<tr>
<td>STE</td>
<td>Sodium tris EDTA buffer (10 mM tris-HCl pH 7.5, 1 mM EDTA, 150 mM NaCl)</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris borate EDTA (1 × TAE is 900mM tris-acetate, 2 mM EDTA pH 8.0)</td>
</tr>
</tbody>
</table>

* For LB or NZYagar add 15.00 g of Difco agar per liter.
TBE  Tris borate EDTA (1 x TBE is 900 mM tris-borate, 2 mM EDTA pH 8.0)

TBS  Tris buffered saline (20 mM tris-HCl pH 7.5, 150 mM NaCl)

2 x YT  16.00 g Bacto-tryptone, 10.00 g yeast extract, 5.00 g NaCl in 1 L, pH 7.0

2.2.0 Culture of parasites and immortalised cell lines

2.2.1 Culture of P. falciparum

*P. falciparum* was maintained in continuous culture according to the method of Trager and Jensen (1976). Clone C10 derived from a West African isolate BW was maintained in O⁺ erythrocytes at 1% haematocrit in RPMI medium (RPMI 1640 supplemented with 25 mM HEPES, 24 mM NaHCO₃, 0.2% w/v glucose, 25 mg ml⁻¹ gentamicin, 20 mg ml⁻¹ hypoxanthine) containing 0.5% w/v AlbuMAX™ (Gibco BRL). Cultures were gassed with 7% CO₂, 5% O₂ 88% N₂, incubated at 37°C and the medium was replaced daily. Parasites were synchronized by a combination of 5% sorbitol treatment (Lambros and Vanderberg, 1979), which leaves only rings intact. Parasitaemias were monitored by making thin smears of the cultures each day. These were air dried, fixed in 100% methanol for 1 min and stained for 10 min with 10% v/v Giemsa, (Gurr), in 10 mM Na₂HPO₄/KH₂PO₄, pH 7.2.
2.2.2 Culture of myeloma Sp2/0Ag14 cells

The myeloma Sp2/0Ag14 cells, was grown in RPMI 1640 (CELLect®, ICN), 10% v/v Foetal Calf Serum (Sera-Lab) at 37°C in a humidified 5% CO2 incubator. The medium was replaced when necessary and when cells had outgrown the medium (the pH<6.8, as determined by the indicator phenol red, changing from red to yellow). The exhausted culture supernatant was harvested by centrifugation at 700g, for 10 min.

2.3.0 SDS-PAGE and Western blotting

Samples solubilized in SDS sample cocktail (Biolabs) were heated at 100°C for 5 min and analysed by SDS-PAGE on homogenous 7.5% or 10% gels according to the method of Laemmli (1970). Electrophoresis was carried out at 20 mA for ~1 h using a Mighty Small II vertical slab gel unit (Hoefer Scientific Instruments). Molecular mass markers used were either prestained high molecular mass markers 14 400-200 000 M_r (Gibco BRL) or, prestained broad range molecular markers 6, 500-175 000 M_r (Biolabs). Following electrophoresis, SDS-PAGE gels were either:

(i) Stained with 0.1% w/v Coomassie Brilliant Blue R-250 (Sigma) in methanol:water:ethanoic acid (5:5:1) and destained with methanol:water:ethanoic acid (1:17:2).

(ii) Western blotted (Hoefer TE Series Transphor apparatus) at 75V, for 4 h at 6°C. For immunodetection, transfer to nitrocellulose (Schleicher and Schuell 0.45 mm pore size) was carried out in 25 mM tris, 150 mM glycine, 20% v/v methanol, by the method of Towbin et al. (1979).
Western blots to be probed with antibody were blocked with 5% w/v non-fat milk powder (Marvel) in PBS for 1 h at room temperature, washed in PBS, 0.05% v/v Tween-20 (Sigma) for 2 min and incubated with primary antibody diluted 1/100 in PBS, 0.05% v/v Tween-20 for 2 h at 37\textdegree{}C. Blots were then washed for 5 min with PBS, 0.05% v/v Tween-20 three times and incubated with Horseradish peroxidase-conjugated anti-rabbit IgG (Bio-rad), Horseradish peroxidase-conjugated anti-mouse IgG (Bio-radDako) or Horseradish peroxidases-conjugated IgM diluted 1/1000 in PBS, 0.05% v/v Tween-20, for 1 h at 37\textdegree{}C. Blots were then washed for 5 min in PBS, 0.05% v/v Tween-20 three times and once for 10 min in 20 mM tris-HCl pH 7.5 and then developed in 10 ml of ECL™ (Amersham) according to the manufacturers protocol. The blot was placed between Saran wrap (Dow) and exposed to XOMAT-AR (Kodak) film with intensifying screens (Dupont) for a minimum of 2 seconds.

2.3.1 Removal of contaminating reactivity of Rb72 antibody with \textit{E. coli} protein

To remove unwanted reactivity with \textit{E. coli} proteins, the rabbit serum (Rb72) was pre-absorbed with \textit{E. coli} whole cell lysate that had been separated by SDS/PAGE and immobilised on nitrocellulose. The nitrocellulose membrane was stained with Ponceau-S stain (0.2% in trichloroacetic acid) and the regions corresponding to 150-155 kDa were cut out so as not to preabsorb anti-\beta/\beta' antibodies from the polyclonal serum. Rb72 was incubated with the nitrocellulose membrane for 3 h at 37\textdegree{}C.
2.3.2 Pre-absorption of culture supernatant 2E2.1 or anti-\textit{E. coli} RNAP antibody (EcB) with pure \textit{E. coli} RNAP

15 ml of 2E2.1 culture supernatant or 1:500 dilution of EcB was incubated with 2\(\mu\)l \textit{E. coli} RNAP (Boehringer Mannheim) at 37°C for 1 hour with gentle shaking. Throughout this thesis, the only purified \textit{E. coli} RNAP referred to is that manufactured by Boehringer Mannheim. The antibody:RNAP complex was then used to probe Western blots as described in section 2.3.

2.3.3 Antibody affinity select

A preparative 10\% SDS-PAGE of \textit{E. coli} purified RNAP was Western blotted onto nitrocellulose (2.5.3) and, following staining with Ponceau S (0.2\% in trichloroacetic acid) the \(\beta/\beta'\), \(\sigma\) and \(\alpha\) subunit bands were excised, cut into 1 mm\(^2\) pieces and blocked in 5\% w/v non-fat milk powder (Marvel) in PBS for 1 h. This and all subsequent incubations and washes were carried out with vigorous agitation using a microfuge mixer (Eppendorf mixer 5432). Nitrocellulose pieces were then transferred to 2 ml microfuge tubes, washed twice in PBS over a period of 5 min and incubated with 500 ml of EcB antibody diluted 1/50 in PBS, 0.05\% v/v Tween-20 for 2 h at room temperature with constant agitation. Nitrocellulose pieces were then washed 5 times for 5 min and then 5 times for 10 min with 1.5 ml PBS, 0.05\% v/v Tween-20. A final wash for 10 min with 1.5 ml of PBS was carried out prior to elution with 500 ml 0.2 M glycine-HCl, 100 mM NaCl pH 2.5. Eluted antibody was neutralised with 40 ml 2 M tris base, BSA added to 2\% w/v and sodium azide added to 0.02\% w/v, final concentrations. The nitrocellulose pieces were neutralised by washing 3 times for 10 min in PBS and the antibody binding and elution cycle repeated four more.
times. Pooled eluted antibody was then used to probe \textit{P. falciparum} whole cell lysate, \textit{E. coli} whole cell lysate and pure \textit{E. coli} RNAP by the standard Western blotting procedure (section 2.5.3).

\textbf{2.3.4 Immunofluorescence}

\textit{Plasmodium}-infected erythrocytes (> 7\% parasitaemia), were washed twice in PBS, and resuspended in an equal volume of PBS. Thin smears of the suspension were made on glass slides, air-dried and fixed in ice-cold acetone:methanol (1:1) or methanol. All incubations were performed in a moist chamber at 37°C. 20\( \mu \)l of either a 1/100 dilution of \textit{EcB} or un-diluted affinity selected antibodies were placed on marked areas of the slides and incubated for 30 minutes. The slides were then washed three times in PBS and incubated for 30 minutes with a 1/1000 dilution of fluorescein-conjugated goat anti-rabbit IgG (Sigma). After three subsequent washes with PBS, the slides were incubated for 10 seconds with 6-diamino-2-phenylindole (DAPI) at 10\( \mu \)g/ml. The slides were washed in PBS a further three times and thoroughly air dried. After mounting in 90\% glycerol (in PBS v/v pH 8.6), a coverslip was added and the slides analysed under oil immersion at 1000x magnification using a Zeiss fluorescence microscope.

\textbf{2.4.0 Preparation of parasite DNA and plasmids}

\textbf{2.4.1 Purification of parasite genomic DNA and plasmids}

Parasite genomic DNA (clone C10) was prepared from asynchronous cultures. Parasitised erythrocytes were pelleted and washed once in ice cold PBS by centrifugation at 700g for 10 min. Erythrocytes were lysed on ice with two pellet
volumes of 1% v/v acetic acid in H₂O for 5 min and then diluted with 10 volumes of PBS. The parasites were collected by centrifugation at 2000g for 5 min at 4°C and washed five times with PBS. DNA was extracted by SDS lysis, RNase and Pronase digestion followed by phenol extraction, exactly as described by Snounou et al. (1988). The genomic DNA was resuspended in 10 mM tris-HCl, 10 mM EDTA pH 8.0 and stored at a concentration of 1 mg ml⁻¹ at 4°C. The quality of the genomic DNA was determined by the following criteria:

- An A₂₆₀nm/₂₈₀nm ≥ 1.7
- Being >100 kb and free of RNA as assessed by electrophoresis through a 0.4% w/v agarose-TBE gel and
- Digestion to completion by several common restriction endonucleases

To prepare plasmid or cosmid DNA, single recombinant colonies were inoculated into LB containing 100 mg ml⁻¹ ampicillin (Sigma) (or 50mg ml⁻¹ kanamycin - Sigma) and grown for 12-16 h at 37°C. For small scale preparations (mini-preps) 3-10 ml of culture was used, for large scale preparations (maxi-preps) 500 ml of culture was used. Plasmid DNA was prepared using Wizard™ minipreps or Wizard™ maxipreps DNA purification systems (Promega), according to the manufacturer’s instructions.

2.4.2 Restriction endonuclease digests and preparation of vector DNA

Restriction endonucleases and the reaction buffer used were those supplied by the manufacturer (Boehringer Mannheim, Gibco BRL or NBL). Analytical plasmid
digests (1-2 mg) were carried out in 30 ml reaction volumes for 1-2 h at 37°C using 10 U enzyme per mg of DNA. Preparative plasmid digests were scaled up versions (10-20 mg) of the analytical digests. Genomic DNA digests were carried out using 10 mg of DNA in 100 ml reaction volumes with 50 U of restriction endonuclease enzyme for 12 h. The DNA was then ethanol precipitated by addition of sodium acetate, pH 5.2, to 0.3 M (final concentration), followed by 2 volumes of 100% ethanol. After incubation at -70°C for 15 min, the DNA was pelleted by centrifugation at 12,000 g for 15 min at 4°C, washed in 80% ethanol, and the pellet air dried at room temperature for 10 min. The DNA was resuspended in 20 ml of double distilled H2O and resolved by agarose-TBE gel electrophoresis (Sambrook et al., 1989).

2.4.3 Purification of DNA from agarose gels

Gel purification of DNA was carried out using Glassmilk™ (BIO 101) according to the manufacturers’ protocol.

2.4.4 Preparation of vector and insert DNA

The vector and insert DNA were digested with the appropriate restriction endonuclease and purified from TAE agarose (section 2.7.3). Phosphatase treatment of gel purified vector, where appropriate, was carried out using Calf Intestinal alkaline Phosphatase (CIP) (Promega) in CIP buffer (50 mM tris-HCl, 1 mM MgCl2, 0.1 mM ZnCl2, 1 mM spermidine pH 9.3). To remove the phosphate from 5’ protruding

---

1 Unless specified otherwise by the manufacturer. e.g. SmaI (25°C).
termini, the DNA was incubated for 30 min at 37°C with 0.01 U of CIP per pmole of termini. To remove the phosphate from blunt termini, the DNA was incubated for 30 min at 37°C with 0.5 U of CIP per pmole of termini for 20 min at 37°C and after the addition of another equal aliquot of CIP, for 45 min at 55°C. The CIP was then inactivated in the presence of 5 mM of EDTA at 75°C for 15 min and the DNA phenol extracted and ethanol precipitated (section 2.7.2 and 2.7.3).

2.4.5 Ligations and transformations

Ligations were carried out in 10 ml reaction volumes using 100 ng of vector and 3:1, 1:1 and 1:3 molar ratios of vector:insert DNA. For ligations involving 3'- or 5'-protruding termini, incubation was at 25°C for 2-4 h with 0.5 U of Bacteriophage T4 DNA ligase (Gibco BRL) in 1 x ligase buffer (Gibco BRL). For blunt end ligations incubation was at 15°C for 12-24 h with 1 U of Bacteriophage T4 DNA ligase, in 25 mM tris-HCl pH 7.6, 5 mM MgCl₂, 0.5 mM DTT, 1 mM hexaminecobalt chloride, 1 mM ATP, 1 mM spermidine and 25% w/v polyethylene glycol-8000 (all reagents from Sigma).

Approximately 20-40 ng of the ligated DNA was used to transform E. coli. Electrocompetent cells (E. coli SURE™ or DH5α) were prepared by inoculating 1 L of LB with 1/100 volume of a fresh overnight culture. Cells were grown at 37°C with vigorous shaking until the OD₆₀₀nm = 0.6 and then chilled on ice for 30 min. Cells were pelleted at 3500g for 15 min at 4°C, and then washed sequentially, by
resuspension and centrifugation (3 500g for 15 min at 4°C), in 1 L of ice cold ddH2O, 0.5 L of ice cold ddH2O and 20 ml of 10% v/v glycerol. Cells were resuspended in a final volume of 2 ml (1-3 × 10^{10} cells ml^{-1}), frozen in 50 ml aliquots on dry ice and stored at -70°C. Electroporation was carried out by mixing ligated DNA with 40 ml of electrocompetent cells thawed on ice. The DNA-electrocompetent cell mix was placed in an ice-cold GenePulser Cuvette (0.1cm gap) and pulsed at 1.8 kV, 25 mF and 200W, using a GenePulser electroporator (Bio-rad). SOC medium (960 ml) was added immediately after the pulse and cells were grown at 37°C for 1 h with shaking at 225 rpm. Cells were plated onto LB-agar containing, 100 mg ml^{-1} ampicillin, or LB-agar containing 100 mg ml^{-1} ampicillin, 0.5 mM IPTG, 80 mg ml^{-1} XGAL (for blue-white colour selection). The genotypes of the E. coli strains used were:

DH5αF': F'/endA1, hsdR17(rK- mK+), gyrA96, relA1, thi-1, Δ(lacZYA-argF)U169 deo R (φ80dlacDM15)

SURE™: el4(mcrA), Δ(mcrCB-hsdSMR-mrr)171, sbcC, recB, recJ, umuC::Tn5 (kan'), uvrC, supE44, lac, gyrA96, relA1, thi-1, endA1 [F'proAB, lacI^{ZAM15}, Tn10, (tet')]

2.5.0 DNA sequencing of double stranded plasmids

Sequencing of plasmid DNA was carried by the dideoxy chain termination method (Sambrook et al., 1989) using Sequenase Version 2.0 (USB) in accordance with the manufacturer's protocol, except that heat denaturation was carried out (Mass and Roop., 1992). Plasmid DNA (2-3 pmol) and primer (6 pmol) was heated at 100°C, 5 min in the presence of 10% v/v dimethyl sulfoxide (Sigma), in a total volume of 8
ml, and incubated on ice for 5 min. Sequenase reaction buffer (2 ml) was added and the labelling and termination reactions carried out as suggested by the manufacturer. Compressions and 'ghost band' artefacts were resolved by the addition of terminal deoxynucleotidyltransferase (USB) as described by Fawcett and Bartlett, (1990). Sequencing reactions were run on 6% denaturing polyacrylamide gels (30 × 40 × 0.04 cm) in 1 × TBE buffer at 55 W using an S2 sequencing gel apparatus (Gibco BRL) as described by Sambrook et al. (1989). Gels were fixed for 20 min in 10% v/v acetic acid, 20% v/v methanol, dried under vacuum at 80°C (Model 1583, Bio-Rad) and exposed to BioMAX-MR film (Kodak) at room temperature for 14-24 h.

2.6.0 Synthesis of oligonucleotides for Polymerase Chain Reaction (PCR)

Oligonucleotides were synthesised at NIMR using an ABI 380B DNA synthesiser and supplied deprotected, at 0.1 mM in 35% ammonia. They were precipitated by addition of 1/10th volume of sodium acetate (pH 5.2) followed by 2.5 volumes 100% ethanol and centrifuged at 12 000g for 20 min. Pellets were washed twice in 80% ethanol, air dried and resuspended in 1.5 ml ddH2O. Following the determination of their concentration by UV spectrophotometry (1.0 at A260nm = 33 mg ml⁻¹) they were made to 100 pmol ml⁻¹.
2.7.0 PCR from 35 kb or yeast DNA

2.7.1 Sequences of oligonucleotides used and annealing conditions

Table 2.1 PCR primers and annealing conditions.

Sequences in bold are 5'-leaders containing restriction endonuclease sites.

<table>
<thead>
<tr>
<th>Gene to be amplified</th>
<th>Primers</th>
<th>Sequence 5' to 3'</th>
<th>Annealing temperature /° C</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpoC1/C2</td>
<td>DW3</td>
<td>CCATTAGGATGTACAAGTTTTAA TGCAG</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>SK</td>
<td>TCTAGAACTAGTGATAC</td>
<td></td>
</tr>
<tr>
<td>rpoC1/C2</td>
<td>dw11</td>
<td>CTTTTATTAACGTAGGTATTACC A</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>dw12</td>
<td>CGAATTGGAGCTCCACCG</td>
<td></td>
</tr>
<tr>
<td>clpC-&gt;</td>
<td>KR46</td>
<td>GGTATGTGTGTATTCAAGAGG AGG</td>
<td>52</td>
</tr>
<tr>
<td>rpoC2</td>
<td>rpoD</td>
<td>AGTTATATATAGGCTATAAA GTGAACC</td>
<td></td>
</tr>
<tr>
<td>rpoC2-&gt;</td>
<td>cl35R</td>
<td>CAAATTTTAATAAAATTTAATT TTAAAG</td>
<td>42</td>
</tr>
<tr>
<td>clpC</td>
<td>KR46</td>
<td>GGTATGTTGTTATCCAGAAGG AGG</td>
<td></td>
</tr>
<tr>
<td>rpoC2-&gt;</td>
<td>rpoD</td>
<td>AGTTATATATAGGCTATAAA GTGAACC</td>
<td>42</td>
</tr>
<tr>
<td>rps 2</td>
<td>cl35F</td>
<td>GCTATAATGAAAAATTTAATTT ACC</td>
<td></td>
</tr>
<tr>
<td>rpoB</td>
<td>A3</td>
<td>TCCCCCGGGGAAAACTTATTATAAAAGC</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>A4</td>
<td>TCCCCCGGGGATAAATTATATATTA ATAGGTAAC</td>
<td></td>
</tr>
<tr>
<td>rpoC1/C2</td>
<td>RT1</td>
<td>CCAAAATTTGGAATTCCTG</td>
<td>50</td>
</tr>
<tr>
<td>border</td>
<td>RW31</td>
<td>CTTTTCAAAGTTAATATAGTAC</td>
<td></td>
</tr>
<tr>
<td>EF-Tu</td>
<td>EFTU 5'</td>
<td>CATTAAATGCTAGCAGATATG</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>EFTU 3'</td>
<td>ATGAACATCGGTTCAGCATTAATG</td>
<td></td>
</tr>
</tbody>
</table>
2.7.2 PCR protocol

PCR was performed using 500 ng of each primer and either 2 ng of *P. falciparum* 35 kb DNA or 50 ng of *S. cerevisiae* genomic DNA (provided by D. Williamson, NIMR). The reaction mix contained 2.5 mM MgCl$_2$, 1x PCR buffer (Promega), 0.2 mM deoxy ribonucleic acids (dNTPs) (Promega) and 1 unit AmpliTaq (Cetus), made up to a volume of 100 µl with sterile water. The following PCR cycling conditions: 95°C, 1 min; (annealing temperature)°C, 1 min and 72°C, 1 min for 35 cycles unless otherwise described in the relevant Results sections.

2.7.3 Cloning of PCR products into the TA vector

PCR products were ligated into the TA Cloning kit, Invitrogen, according to the manufacturers protocol. Ligated DNA was used to transform competent DH5α cells and recombinants selected by blue-white colour selection (section 2.7.5). Plasmid DNA was sequenced with universal forward and reverse primers (section 2.6).
2.8.0 Southern blots

2.8.1 Southern transfer

Southern blotting was carried out by a modification of the method of Southern (1975).

After agarose gel electrophoresis, the DNA was “acid nicked” (where appropriate),
denatured and neutralized as follows:

0.2 M HCl, 20 min with two changes of buffer

Rinsed three times in dH2O

0.5 M NaOH, 1.5 M NaCl, 45 min with three changes of buffer

Rinsed twice in ddH2O

1.0 M tris-HCl, 1.5 M NaCl pH 8.0, 30 min with two changes of buffer

Each step was carried out by submerging the gel in 10 volumes of buffer with gentle
agitation on a rotary shaker. The “acid nicking” was omitted when DNA fragments to
be transferred were < 15 kb.

Capillary transfer to Hybond N⁺(Amersham) in 20 × SSC was carried out as
described in Sambrook et al. (1989) using a 500g mass and Quickdraw (Sigma) for 4-
12 h. Following transfer, the blot was washed in 10 × SSC for 5 min, placed on 3MM
Chr paper (Whatman) to air dry for 10 min and the DNA UV cross-linked for 45 s
(autocrosslink, UV Stratalinker, Stratagene). The blot was stored between Saranwrap
(Dow) at 4°C until used for hybridisation.

2.8.2 Labelling of probes and hybridisation

Double stranded probes were labelled with [α-32P] dATP ( 220TBq mmol⁻¹,
Amersham) using random hexamers (Prime-It™ II Random Primer Labelling kit,
Stratagene) either in solution or in low melting temperature agarose (Gibco BRL). Unincorporated label was removed using Sephadex G-50 column (Nick™ column, Pharmacia), according to the manufacturers instructions.

Southern blots were pre-hybridised for 1-2 h in the presence of denatured Salmon Sperm DNA (Stratagene); the Salmon Sperm DNA was denatured at 100°C and snap-cooled on ice for 5 minutes prior to the addition of the pre-hybridisation mixture to give 100 mg ml⁻¹ final concentration. Pre-hybridisation and hybridisation overnight was in 6 x SSC, 5 x Denhart's solution (1 x Denhart's is 0.02% w/v Ficoll, 0.02% w/v BSA fraction V, 0.02% w/v polyvinylpyrrolidone), 0.5% w/v SDS. Labelled, doubled stranded probe at a concentration of 1-3 ng ml⁻¹, using at least 1 x 10⁶ cpm ml⁻¹ hybridisation solution, and the probe’s specific activity >1 x 10⁸ cpm mg⁻¹ was boiled and snap-cooled on ice for 2 minutes before being added to the hybridisation mixture. Hybridisation was carried out in a rotory oven (Hybaid) using 0.05 ml of hybridisation solution per cm² of the blot.

2.8.3 Hybridisation temperature and washing conditions

Hybridisation temperatures were determined empirically:-

Low stringency hybridisation - 28 °C

High stringency hybridisation - 65 °C

Sequential washes with decreasing amounts of SSC and increasing washing temperature were carried out starting at room temperature, the most stringent wash being determined empirically. Each washing step was carried out for 15 min with two
changes of buffer in a shaking incubator. The blot was placed between Saran wrap (Dow) and exposed to XOMAT-AR (Kodak) film at -70°C with intensifying screens (Dupont) for 40-50 h.

2.8.4 Southern blot stripping

Radio-labelled probe was removed from Southern blots by immersing the blot in boiling ddH₂O/2%SDS in a flat tupperware container and allowing to cool to room temperature with constant shaking. This process was repeated 4 times after which the blot was placed between Saran wrap and exposed to autoradiography as above.

2.9.0 Screening the 35-kb DNA library

Bacterial lifts were taken in duplicate onto Hybond N⁺ (Amersham) by placing the nitrocellulose onto the bacteria for 1 minute. The bacteria on the master plate were allowed to recover by incubation for 1 hour at 37°C between consecutive lifts and the bacteria on the replica filters were allowed to grow for 4-6 h. Colonies were lysed and the DNA denatured, neutralised and fixed to the membrane by the following method:

The filter was sequentially placed, colonies uppermost, onto 3 sheets of 3MM Chr paper (Whatman) saturated with the following solutions:

10% w/v SDS pH 7.5 for 4 min
0.5 M NaOH, 1.5 M NaCl for 6 min
0.5 M tris-HCl, 1.5 M NaCl pH 7.5 for 5 min This step was repeated once
2 × SSC for 5 min
Excess liquid was removed from the filters between consecutive transfers by placing on dry 3MM Chr paper (Whatman) for 1 min between each step. The filters were allowed to air dry for 10 min and the DNA immobilised onto the membrane by UV cross linking for 45 seconds (autocrosslink, Stratalinker, Stratagene). Filters were floated on 5 × SSC, 0.5% w/v SDS, 1 mM EDTA and incubated at 42 °C with gentle agitation for 45 min, to aid the removal of bacterial debris. Prehybridisation, hybridisation and washing was carried out as described for Southern blots.

Secondary screening was carried out on positive signals which were observed on both the original and replica filters. Between 2 and 6 colonies corresponding to the positive signal were picked and placed in 1.5 ml of L-broth, 100 mg ml⁻¹ ampicillin and vortexed three times over a period of 10 min to resuspend the bacteria. A titre was carried out by diluting this suspension 1/1000 in L-broth, 100 mg ml⁻¹ ampicillin and plating 1 ml, 5 ml and 10 ml on 82mm diameter LB-agar, 100 mg ml⁻¹ ampicillin plates. Secondary screening of plates containing 200-500 colonies was carried out in the same way. Single colonies were picked and grown in L-broth, 100 mg ml⁻¹ ampicillin at 37°C for 16 h and plasmid DNA isolated by Wizard minipreps™.

2.10.0 Analysis of positive clones and subcloning

Miniprep DNA was digested with the appropriate restriction endonuclease, run on a 1% w/v agarose-TBE gel, visualised under UV in order to determine the size of inserts. Clones were then sequenced as described in section 2.6.
2.11.0 RNA extraction

Asynchronous parasites (> 20% parasitaemia) were acetic acid lysed as described in section 2.5.1. RNA was extracted using the RNaid™ Kit (Bio 101) according to the manufacturer's protocol. RNA concentration was determined by UV spectrophotometry at $A_{260}\text{nm}$ where $1OD_{260} = 40 \text{mg/ml}$.

2.12.0 Northern blots

RNA was resolved using 1% w/v agarose-formaldehyde gels electrophoresed in 3-(N-morpholino)-propanesulphonic acid (MOPS)/formaldehyde buffer (Ausubel et al., 1995), at a concentration of 30 $\mu$g/lane and transferred to Hybond-N+ (Amersham) by capillary transfer in 10X× SSC with no pre-treatment of the gel (Sambrook et al., 1989). Transferred RNA was immobilised by UV crosslinking for 45 s (autocrosslink, Stratalinker, Stratagene) and either stained with Methylene Blue (membrane placed in 5% v/v acetic acid for 5 min, followed by 0.04% Methylene Blue in 0.5 M sodium acetate pH 5.2 for 5 min) or probed with $[\alpha-^{32}\text{P}]$-radiolabelled probe. Northern blots were pre-hybridised at $65^\circ\text{C}$ for 1 h and hybridised at $65^\circ\text{C}$ for 12 h in 6 × SSC, 5 × Denhart's solution (1 × Denhart's is 0.02% w/v Ficoll, 0.02% w/v BSA fraction V, 0.02% w/v polyvinylpyrrolidone), 0.5% w/v SDS containing 2 ng probe ml⁻¹, and $\sim 2 \times 10^6$ cpm ml⁻¹ with the probe's specific activity $> 2 \times 10^8$ cpm mg⁻¹. Blots were washed for 1h at $65^\circ\text{C}$ in 6 × SSC, 0.2% w/v SDS with two changes of buffer and at $65^\circ\text{C}$ for 1.5 h in 0.1 × SSC, 0.2% w/v SDS and exposed to XOMAT-AR film (Kodak) for 15h at -70°C with intensifying screens (DuPont).
2.13.0 Primer extension analysis

Total parasite RNA was prepared (as described in section 2.11). 1 μg of primer APX1 5’(ATATAATATCTTAAATTATATATTATTCTG)3’ was added to 30 μg total RNA and the mixture precipitated by the addition of 0.1 volume sodium acetate pH 5.2, 2.5 volumes 96% ethanol and stored at -20 °C for 20 minutes. Following centrifugation at 12,000g for 10 minutes at 4 °C, the pellet was washed with 70% ethanol and then air dried.

The RNA:DNA pellet was redisolved in 30 μl of hybridisation buffer

- 40 mM PIPES pH 6.4
- 1 mM EDTA pH 8.0
- 0.4 M NaCl
- 80% formamide

and incubated at 85°C for 10 minutes. The hybridisation mixture was transferred to a beaker of water at 85°C and placed in a water bath at 37°C to ensure that re-annealing occurred slowly during cooling over 12 hours.

The annealed primer:RNA mixture was precipitated on ice for 1 hour after the addition of 170 μl ddH2O and 400 μl ethanol. The mixture was centrifuged at 12,000g for 15 minutes at 4°C, and washed with 500 μl 70% ethanol. The ethanol was subsequently aspirated and the pellet air dried.

The primer extension was performed for 2 hours at 37°C with 200 U MoMuLV reverse transcriptase (RNase H-; Promega) in a 20 μl reaction mixture containing the manufacturers recommended buffer with 2mM dNTP, 50 U RNase inhibitor, and 10 μCi [α-32P]-labelled dATP (220TTBq mmol⁻¹ Amersham). The reaction was terminated with 0.5 M EDTA, and free nucleic acids were extracted with
phenol/chloroform, ethanol precipitated (as described above) and the pellet redisolved in STOP buffer (from the USB sequenase™ kit). Electrophoresis was carried out on a 6% polyacrylamide denaturing gel (as described in section 2.6.1). Controls included no primer, and an RNA sample previously treated with RNase (Gibco-BRL). Transcript sizes were analysed by comparison with a sequence ladder of known sequence.

2.14.0 Reverse Transcriptase PCR (RT-PCR)

RT-PCR was carried out using total RNA (prepared as described 2.11). First strand synthesis was carried out by mixing 500 ng of primer RT1 5' (CCATTAAAATTGGTAATCCTG) 3' with 3 µg RNA and denaturing at 70°C for 10 min followed by quenching on ice for 3 min. Reverse transcription was carried out in a 20 ml reaction volume containing 20 mM tris-HCl pH 8.4, 50 mM KCl, 2.5 mM MgCl₂, 10 mM DTT, 500 mM each of dCTP, dATP, dGTP, dTTP (Pharmacia), and 200 U of Superscript II™ RNase H- Reverse transcriptase (Gibco BRL) at 42°C for 1 h. Reactions were heated at 70°C for 15 min, quenched on ice to inactivate the reverse transcriptase. Controls included no treatment with reverse transcriptase followed by PCR and, first strand synthesis on RNA pre-treated with RNase. Gene specific amplification of the rpoB/C₁ boundary was carried out using the first strand synthesis reaction as a template in 20 mM tris-HCl pH 8.4, 50 mM KCl, 1.5 mM MgCl₂, 200 mM each of dCTP, dATP, dGTP, dTTP (Pharmacia), 500 ng of primer RW31 and 2 U Taq polymerase (Amplitaq™, Perkin Elmer Cetus Instruments); in a
100 ml reaction volume. The PCR conditions were: 95°C, 1 min; 42°C, 1 min; and 72°C, 1 min for 35 cycles.

2.15.0 Pulsed Field Gradient Electrophoresis (PFGE)
Resolution of *Plasmodium falciparum* chromosomes by PFGE was carried out by P. Moore (NIMR) by the method of Hinterberg and Scherf (1994). After staining with ethidium bromide, the blot was hybridised to [α-32P]-radiolabelled probe and washed under high stringency as described (2.13.4).

2.16.0 In vitro production of monoclonal antibodies
The synthetic polypeptides IW376 and IW227 were coupled to keyhole limpet haemocyanin using gluteraldehyde (Harlow and Lane, 1988) by M. Strath. Monoclonal anti-mouse antibodies were generated to these KLH-polypeptides according to the manufacturers protocol (Immune Systems). Positive clones were identified by ELISA assay (see below).

Culture supernatants which recognised the polypeptide antigens were expanded initially into 24 well Costar plates, then 75 cm² flasks and finally 175 cm² flasks. Medium was changed as in 2.2.2. Supernatants were concentrated and stored at -20°C and the hybridomas were frozen in 10% DMSO and held in liquid N₂.

2.16.1 ELISA assay
Plastic plates (Nunc, 96 well) were incubated at room temperature for 1 hour with 50 µl of antigen (IW376 or IW227). The plates were then washed with PBS/BSA (0.5
mg/ml) three times. Supernatants from monoclonal cultures (50 μl) were incubated with the antigen for 1 h at room temperature. The plates were washed as before and incubated with 50 μl anti-mouse horse-radish-peroxidase (HRP) conjugated IgG antibody for 40 minutes at room temperature. The plates were next washed 4 times as above, prior to incubation with 50 μl HRP substrate (0.1% 1-chloro-4-naphthol, 0.2% H$_2$O$_2$) per well for 30 minutes. The reaction was stopped by the addition of 25 μl per well of 3 M NaOH. The wells were screened using an ELISA reader, at lambda 405nm.

2.16.2 Immunotyping of culture supernatants

Culture supernatants were immunotyped using the immunotype 1 kit manufactured by Sigma.

2.17.0 Fusion protein synthesis - cloning and expression of the pxy/9 construct

500 ng of the PCR amplification product (primers A3/A4) of the C-terminus region of the RNAP β subunit was digested with Smal and ligated into Smal cut, CIP treated expression vector pxy460. Insertion into the Smal site of this vector (provided by S. Ogun, NIMR), restores the reading frame to produce a β-galactosidase fusion protein (Holder et al., 1987). DH5α E. coli were transformed with the ligation mix and plated onto selective agar (containing 50μg/ml ampicillin) with the addition of 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal) to 40μg/ml. Blue colonies were picked into L-broth containing 50 μg/ml ampicillin and grown with constant shaking at 37°C overnight. IPTG induced bacteria lysate was screened by Western blotting using an anti-β galactosidase antibody (kindly provided by S. Ogun).
2.17.1 Purification of the β-galactosidase fusion protein

1 ml of a 6 h culture was inoculated into 100 ml of L-broth containing ampicillin (50 μg/ml) and grown overnight at 37°C. The culture was then added to 500 ml of L-broth containing 50 μg/ml ampicillin and IPTG to a final concentration of 0.3 mM. After induction at 37°C for 2 h, the cells were harvested by centrifugation at 6,000 rpm for 10 minutes in a Beckman J6B. The cell pellet was resuspended in 50 ml L-broth and re-harvested. The L-broth was aspirated and the bacterial pellet frozen at -20°C overnight.

The cell pellet was thawed into 20 ml of 25 mM Tris pH8.0/1 mM EDTA/0.2% NP40 containing 100 μl 100 mM PMSF. Upon resuspension, lysozyme was added to a final concentration of 1 mg/ml and incubated on ice for 2 h. 40 μl MgSO₄ and 40 μl 10 mg/ml DNase (Gibco/BRL) was added and incubation continued for a further 2 h. The material was centrifuged at 13,000 rpm for 10 minutes in a Sorvall SS-34 centrifuge and the supernatant volume was determined. Protein was precipitated by the addition of 231 mg ammonium sulphate per ml of supernatant and 1 ml 1 M NaOH per 10 g ammonium sulphate and incubation at 4°C overnight.

The precipitate was spun down at 9,000 rpm (Sorvall SS-34) and the supernatant discarded. The pellet was resuspended in a small volume of 20 mM Tris pH 7.4, 10 mM MgCl₂, 0.1 M NaCl, 10 mM β-mercaptoethanol. Dialysis against this buffer was performed for at least 4 h at 4°C. The sample was centrifuged at 10,000 rpm for 10 minutes (SS-34) and applied to a p-Amino Benzyl 1-thio-β-D-Galactopyranoside (ABTG)-agarose (Sigma), column pre-equilibrated in the above buffer at a flow rate
of 20 ml/hr. Fractions were collected every three minutes during elution with 0.1M sodium borate pH10, 10 mM β-mercaptoethanol at 4°C. The polypeptide concentration in aliquots of eluant were estimated by UV spectrophotometry, presuming that at a concentration of 1mg/ml, β-galactosidase has an $A_{280}=2.0$. 

2.18.0 Sequence analysis

Sequence data were analysed using the BLAST, GAP, TRANSLATE, MAP, REVERSE and PILEUP options of the Sequence Analysis Software Package (Genetics Computer Group, Madison Wisconsin)(GCG) (Devereaux et al., 1984).
CHAPTER 3
3.1.0 Introduction

Prokaryotic RNAPs are multimeric enzymes consisting of two \( \alpha \) subunits, a \( \beta \) subunit, a \( \beta' \) subunit and a regulatory \( \sigma \) subunit \( (\alpha_2\beta\beta'\sigma) \) encoded by the \( rpoA, rpoB, rpoC \) and \( rpoD \) subunits respectively. Chloroplasts encode a prokaryotic like RNAP which is utilised in the synthesis of mRNA and tRNA genes (Little and Hallick, 1988). A second RNAP which is presumed to be nucleus encoded has been demonstrated to be active in the transcription of pea chloroplast rRNA genes (Lakhani et al., 1992) and the chloroplast-encoded \( rpo \) genes although the gene encoding this polypeptide has not been determined (Greenberg et al., 1984; Greenberg et al., 1985; Hess et al., 1993; Pfannschmidt et al., 1994). The 35 kb encoded \( \beta \) subunit identified by Gardner (1991a.) has homology to the \( \beta \) subunits encoded by eubacteria and chloroplasts and suggested that the RNAP encoded by chloroplasts is also encoded by the \( P. falciparum \) plastid-like molecule.

The \( E. coli \) \( \beta \) subunit has a molecular mass 150,000 M, and the \( \beta' \) subunit one of 155,000 M. Whereas the eubacterial \( \beta' \) subunit is encoded by a single contiguous reading frame, chloroplasts differ in that the \( rpoC \) gene equivalent is present as a “split” gene, encoding two separate polypeptides. The \( rpoC_1 \) gene encodes a polypeptide designated \( \beta' \) which has homology to the amino-terminal region of the \( E. coli \) \( \beta' \) subunit, the \( rpoC_2 \) gene encoding a polypeptide homologous to the \( E. coli \)
carboxyl-terminus. The sequence data presented here indicate that as in the case of plastid RNAPs the *Plasmodium* plastome β' subunit is encoded by two separate genes. These genes both show homology to *E. coli* and plastid RNAPs and are therefore referred to as *rpoC₁* and *rpoC₂* in accordance with the terminology used for plastid RNAPs. The formation of a split *rpoC* gene in plastids is thought to have occurred early during plant evolution since this feature also occurs in cyanobacterium (Xie *et al.*, 1989). The occurrence of a split *rpoC* gene in *P. falciparum* is therefore consistent with other evidence that the RNAP and therefore the circle itself has a chloroplast or plastid lineage rather than a bacterial origin.

### 3.2.0 Results and Discussion

#### 3.2.1 Amplification and sequencing of *rpoC₁/rpoC₂*

Some sequence at the 5' end of the *rpoC* gene had been obtained previously from a clone isolated from a 35 kb circle library, however, there was no other sequence information downstream from this clone within 10 kb. The library previously used was known to contain many rearranged clones and was not thought to be the optimum way of cloning the rest of the *rpoC* gene. Instead, PCR was used to amplify the gene for sequencing. Due to the lack of sequence data within this region of the molecule no 3' sequence was available for use as a second primer. Accordingly, in order to amplify DNA in this region a known restriction site in the template DNA was modified by ligating an anchor DNA of known sequence which could be used as a second priming region in the PCR reaction.
There are few 6 base restriction sites in the 35 kb circle due to the high A+T base pair composition, the only one within this region being ScaI. Restriction digestion of the 35 kb DNA with ScaI produced four blunt ended fragments of 14.6 kb, 10.1 kb, 5.6 kb, and 4.2 kb respectively. The 10 kb fragment encodes the rpoB and part of the rpoC gene Figure 3.1a.

The anchor DNA used, was prepared from Bluescript™ plasmid by restriction digestion with EcoRV and XhoI to produce a linear 2.6 kb DNA with one blunt end and one sticky end. The SK primer site within this Bluescript™ fragment being located at the blunt end terminus. The gel-purified DNA was blunt end ligated to 35 kb DNA cut with ScaI to generate a pool of ligated molecules with the Bluescript™ DNA in a fixed orientation Figure 3.1b.

Using the PCR protocol described in Materials and Methods, and primers SK and DW3, at an annealing temperature of 39°C for 1 minute, an amplified product migrating at 1.3 kb was observed after separation on a 1% agarose gel. However, using a fresh template preparation and subsequent PCR reaction failed to produce more product, generating a smear of DNA ranging from the top to the bottom of a 1% agarose gel. Similar findings have been observed by Bell and DeMarini (1991) who explained this occurrence in the following way: Taq polymerase adds an extraneous adenosine at the 3' end of the PCR product producing a sticky end. When the amount of DNA template is limiting the PCR product concatamerises by virtue of these staggered ends and in turn becomes the new template thereby generating amplified amplified
products of varying sizes. It was found that by reducing the number of amplified cycles from 35 to 23, the 1.3 kb product (sk/d3) was reproduced.

The specificity of the sk/d3 product was confirmed by Southern blot analysis of *P. falciparum* genomic DNA (data not shown).

The sk/d3 product was amplified at a low yield and was therefore cloned via the TA cloning kit (Stratagene). Clones containing inserts were isolated using the X-gal blue/white assay. Inserts may be isolated from the vector DNA by restriction digestion with *EcoRI*. One clone out of 64 positive (white) clones contained an insert of the correct size. On sequencing, however, using the universal M13 forward and reverse primers positioned on either side of the insert within the TA vector, it was found that the product had rearranged, losing the DW3 primer site. This problem is common when manipulating 35 kb DNA, presumably due to the high A+T content.

The yield of sk/d3 product was therefore increased by nested PCR as follows. An initial PCR reaction was performed using primers dw11 and dw12, where dw11 is located 5' to the SK primer within the Bluescript™ vector and dw12 is positioned 3' to the DW3 primer within the *rpoC* gene. The resultant product was gel-purified and used as template DNA for the second PCR reaction using primers SK and DW3. The product yield was thereby increased from ca100 ng DNA/100μl to ca 300 ng DNA /100μl. The DNA was gel purified and run into low melting point agarose.

Sequencing reactions were carried out as described in Materials and Methods.
The sk/d3 product was found to code for the remainder of \textit{rpoC}_1 and the first 1,176 bp of \textit{rpoC}_2. By this time more of the 35 kb circle had been sequenced, yielding sequence information to within 4 kb of the SK terminus of the sk/d3 product. Using primers KR46 located at the 3’ region of the clp gene (used with permission from K. Rangachari), and \textit{rpoD} (located within the sk/d3 PCR product)(see Figure 3.2), the intervening 4 kb fragment was amplified using PCR as described in Materials and Methods with the following modifications: extension temperature (72 °C) for 3 minutes; cycling 40 times. As the product yield was low and could not, in this case, be boosted by nested PCR, TA cloning was attempted again. Repeated attempts at cloning this 4 kb PCR product failed; the only clones selected containing inserts of approximately 2 kb. The 4 kb PCR product was therefore used as a probe to screen the 35 kb circle Bluescript™ library mentioned above (with kind permission of K. Rangachari and D. Moore.). Forty positive clones were isolated and sequencing reaction were performed on DNA purified by mini-preparation using magic mini-preps™ (Promega).

The sequence data obtained from these clones were analysed using the GCG software package and open reading frames translated into putative protein sequence. Comparisons between these sequences and the \textit{rpoC}_2 sequences from both maize and spinach chloroplast were assessed using the GAP programme. One 535 bp clone (clone 35) had 27 % homology at the amino acid level to the carboxyl terminus of maize \textit{rpoC}_2 (nucleotide 3170-355 in Figure 3.3), and includes the ATG start codon of the succeeding \textit{rps} 2 gene.

68
Primers C135F and C135R were designed at the 5' and 3' termini of clone 35 to enable two smaller PCR products to be amplified (Figure 3.2). Four primers, C135F, C135R, KR46 and rpoD were used pairwise, specific products being expected from the C135F/rpo pair and the C135R/KR46 pair. Each primer pair generated predominant bands migrating on a 1% agarose gel at the following molecular weights; the C135F/rpo pair generated a major band of approximately 1.6 kb (designated rpo/F) and the CL35R/KR46 pair generated a major product at ca 2 kb (KR/R) - Figure 3.2b.

Six PCR reactions from each of the two primer pairs were pooled, purified in low melting point agarose gel, and subsequently were purified using Geneclean™ (Bio 101). In order to maximise yields of DNA the manufacturers protocol was modified in that DNA was eluted twice for thirty minutes at 50°C in a volume of water equal to the volume of resin used. These modifications increased the recovery yield from an observed 60% to 90% of the original amount of starting product (data not shown). The purified DNA was sequenced according to the Sequenase™ protocol with modifications described in Materials and Methods. The sequence data generated from these products were analysed as previously described.

Primers C135F and rpo generated sequence which had homology at the amino acid level to maize rpoC2. Primer C135R produced sequence which did not have significant homology with rpoC2. Analysing this sequence data with the BLASTP programme of the GCG package, did not provide information as to what this DNA encoded. Primer KR46 produced sequence which had homology to the clp gene thus extending the sequence data available for that gene. The specificity of the PCR products had
therefore been determined at the nucleotide level and the remainder of the two products were sequenced using primers synthesised according to the nucleotide sequence data. All sequence data were assembled and analysed using the Staden software programme and GCG package as described in Materials and Methods.

The remainder of this chapter contains a discussion of the sequence data obtained for the \textit{rpoC}_1 and \textit{rpoC}_2 genes (Figure 3.3). The additional sequence obtained of other genes in this region will be presented in the following chapter.

\textbf{3.2.2 Analysis of the \textit{rpoC}_1 gene}

The sequence determined from the sk/d3 PCR product added a further 117 amino acids to the C-terminal region of the partial polypeptide sequenced earlier (Gardener \textit{et al.}, 1991a). The entire, mature polypeptide is predicted to be 575 amino acids long.

Transcription data suggested that the \textit{Plasmodium} plastid \textit{rpoB/C} genes are transcribed polycistronically (Gardner \textit{et al.}, 1991a; Feagin \textit{et al.}, 1994) as are the maize \textit{rpoB/C}_1/\textit{C}_2 genes (Hudson \textit{et al.}, 1988). Primer extension analysis using an oligonucleotide (Apx1) located at 5' (n=104), 3' (n=73) in \textit{rpoB} (Figure 3.4) mapped the transcription start site or a processing site of the \textit{rpoB} operon 340 bp upstream to the \textit{rpoB} ATG start codon in the intergenic space between ORF 101 and ORF 51 - Figure 3.5. Two bands were obtained using this primer, the lower signal was however not always observed and was assumed to be an artefact, possibly caused by secondary structures causing premature termination of the extension product. This result was confirmed independently in a separate analysis by M. Gardner (Gardner \textit{et al.}, 1994).
Reverse transcriptase PCR (RT-PCR) of C10 DNase treated RNA (prepared as described in Materials and Methods), electrophoresed on a 1% agarose gel, was used to confirm that the \textit{rpoB} and \textit{rpoC} genes are co-transcribed (as described in Materials and Methods) - see Figure 3.6. Primers RW31 located at the 3' end of \textit{rpoB} and RW32 located at the 5' terminus of \textit{rpoC} were used to amplify a 463 bp region from purified 35 kb template DNA, spanning the 5 bp gap between the two reading frames. Two negative controls were included in this experiment:-1) PCR was performed on RNA pre-treated with RNase followed by reverse transcriptase; no amplified product could be detected thereby confirming that the positive RT-PCR result was not due to RNA contamination in the reverse transcriptase; 2) PCR was carried out on RNA which had not been treated with reverse transcriptase (data not shown), the absence of product confirmed that the RT-PCR signal had not been obtained because of DNA contamination of the RNA template.

Primer extension techniques were used to determine whether the \textit{rpoC} gene was co-transcribed with the \textit{rpoB} gene as previously determined by Northern blot techniques (Gardner \textit{et al.}, 1991a). No specific signals were visualised (data not shown), thereby suggesting that the \textit{rpoC} gene is transcribed as part of an operon with the \textit{rpoB} gene. The cyanobacterium \textit{N. commune rpo} gene cluster does not constitute an operon, in this case, the \textit{rpoC} and \textit{rpoC2} genes are transcribed separately from \textit{rpoB} (Xie \textit{et al.}, 1991), therefore, in this respect the \textit{P. falciparum rpo} operon resembles the chloroplast \textit{rpo} operon (Hudson \textit{et al.}, 1988), rather than cyanobacteria.
The *P. falciparum* β polypeptide is not highly conserved in comparison to other chloroplast homologues, especially due to the high degree of conservation between plastid *rpoC* genes (Figure 3.7). Homology has been maintained between *P. falciparum* and the eukaryotic homology blocks - labelled A-H (Jokerst *et al.*, 1989) in Figures 3.7 and 3.8. These consensus blocks are conserved amongst the largest subunits of eukaryotic RNAPs. The *E. coli* *rpoC* gene been split into two in relation to the chloroplast *rpoC* genes and *rpoC* coding sequences for the purpose of analysis of these genes.

Of particular interest, a region thought to form a zinc finger is conserved in the *rpoC* genes of chloroplasts and in *E. coli* (overlined in Figure 3.7). This zinc finger is analogous to a C4-type zinc finger which has been implicated in DNA binding by eukaryotic transcription factors (Struhl., 1989; Evans *et al.*, 1988). The chloroplast consensus motif C-X-C-X$_{15}$-C-X$_2$-C does not perfectly match the eukaryotic C$_4$-type consensus C-X$_2$-C-X$_{13}$-C-X$_2$-C. Sequence differences among zinc fingers, however, may allow diversity of function relating to a alteration in DNA sequence specificity (Freedman *et al.*, 1988; Igloi *et al.*, 1990). The *P. falciparum* *rpoC* gene sequence in this region has maintained all four cysteine residues unlike the cyanobacteria *N. commune* where the second pair of cysteines have been replaced by valine residues. Although it has not been demonstrated, the two serine residues within this region in *N. commune* may replace the function of the two cysteines (Berghofer *et al.*, 1988).

Three main blocks of homology are conserved throughout chloroplast *rpoC* genes as well as in the homologue from the malarial plastome, these correspond to the regions
K-R-V-D, R-A-P-T and F-N-A-D (corresponding to amino acids 376-392, 458-465 and 491-501 respectively in Figure 3.7. The F-N-A-D-F-D-G-D-Q motif is remarkably conserved between the *P. falciparum* β' subunit and those of the chloroplast subunits and the eukaryotic consensus block D. This region is universally present in all of the largest subunits of eukaryotic RNAPs (Kock *et al.*, 1988). Lerbs (1983) has shown that the spinach β' subunit is probably responsible for non-specific binding of the polypeptide to DNA. Considering the high degree of conservation within the regions cited above, in addition to the putative zinc finger motif, it has been suggested that these regions play an important role in the function of the molecule although the specific function has not, as yet, been determined.

All chloroplast *rpoC*₁ genes contain a group II type intron except those of the grass family namely rice and maize. The *P. falciparum* homologue does not contain an intron and in this respect is more similar to eubacterial *rpoC* genes and that of the gamma subunit of cyanobacteria than those of higher plants.

The intergenic space between the *rpoC*₁ and *rpoC*₂ genes is 10 bp which is small in comparison to the length of the intergenic regions reported in maize 201 bp (Igloi *et al* 1990) and spinach 151 bp (Hudson *et al*. 1988). This once again indicates the overriding feature of gene organisation within the 35 kb circle - a minimum wastage of space. There appear to be very few regions within the molecule which do not encode proteins or other apparatus necessary for transcription and translation and it is therefore easy to assume that as in the case of the two *rpoC* genes the distance separating them must be the smallest necessary.
3.2.3 Analysis of the \textit{rpoC}_2 gene

The \textit{Plasmodium rpoC}_2 gene is 2.88 kb in length, encoding a predicted polypeptide of 953 amino acids. However, an opal stop codon occurs at position 1962 in Figure 3.3 prematurely terminating the \textit{rpoC}_2 coding sequence. An initiation codon (ATG) in the next reading frame overlaps the last amino acid in \textit{rpoC}_2, see Figure 3.3. This is unprecedented in other eubacterial and chloroplasts \textit{rpoC}_2 genes characterised to date. A frameshift located in the plastid encoded rpl 23 gene in spinach (Thomas \textit{et al.}, 1988), and in the ndh gene encoded by the plastid of the holoparasitic flowering plant \textit{Cuscuta reflexa} (Haberhausen and Zetsche., 1994) are responsible for these genes being rendered redundant and leading to the formation of pseudogenes. To my knowledge, it has not been determined how (-1) frameshifts are translated, although there is evidence that, in \textit{Saccharomyces cerevisiae} (+1) frameshifts are translated due to translational pausing with the in-efficiently recognised stop codon at the ribosome P site followed by cognate binding of a special peptidyl-tRNA at the ribosome A site (Pande \textit{et al.}, 1995). By extrapolation, a similar process could occur in the case of the \textit{P. falciparum rpoC}_2 frameshift. The TGA codon where this frameshift occurs, can, in addition to encoding an opal stop codon, encodes tryptophan in non-plant mitochondria (Barrell \textit{et al.}, 1979), \textit{Mycoplasma capricolum} (Yamao \textit{et al.}, 1985) and tobacco rattle virus (Zerfass and Beier., 1992). In these examples, according to the codon capture theory, it is presumed that due to directional pressure towards an A+T bias, the tRNA^{Trp} mutated to a UCA anti-codon able to decode TGG and TGA, this would have been concomitant with the loss of release factor 2 (RF2), necessary for the termination of polypeptide synthesis at opal stops (Osawa \textit{et al.}, 1992). In the example of the tobacco rattle virus, the TGA codon is deciphered by a chloroplast encoded
tRNA\textsuperscript{Trp} carrying a CmCA anti-codon (Zerfass and Beier., 1992). In *P. falciparum* 5 other cases of plastid-encoded TGA stop codons have been found at the ends of the genes *rpl 23*, *rps 17*, *rps 11*, *rps 7* and ORF78. In fact Trp (TGG) itself is a rare codon occurring only 19 times within the 35 kb coding sequence. Therefore, in the case of this frameshift, the ribosome may slip back one nucleotide whilst pausing to decipher the rare TGA codon, this slippage would lead to a change of reading frame, thereby maintaining a contiguous polypeptide to be synthesised. Alternatively, the ribosome may use the opal stop TGA, or the amber stop TAA encoded 12 nucleotides downstream to the opal codon, to terminate translation and use the ATG codon to initiate the translation of a second polypeptide encoding the C-terminal region of *rpoC\textsubscript{2}*. Whether this frame shift invalidates the function of the *rpoC\textsubscript{2}* gene is yet to be addressed although RNase protection assays using a probe at the 5' terminus of the gene suggests that at least this part of the gene is expressed. RNase protection at the 3' terminus immediately downstream to the frameshift have yet to be performed in order to ascertain expression beyond this point. Figure 3.8 shows a PILEUP analysis of the *Plasmodium* sequence, which has been presented as a single polypeptide making the assumption (for convenience), that the frameshift does not render this gene a pseudogene and that the polypeptide is a single continuous protein which has not been fragmented into two. In fact, evidence from Western blots of *P. falciparum* whole cell lysate suggests that this may indeed be the case (see Chapter 4).

There is very little similarity or sequence conservation between the 35 kb *rpoC\textsubscript{2}* homologue and other known *rpoC\textsubscript{2}* genes. The most highly conserved region occurs at the N-terminal region of the malarial polypeptide, where three cysteine residues are
conserved within an 18 amino acid stretch (overlined in Figure 3.8). A histidine residue within this region (position 338 in Figure 3.8) is conserved in chloroplast sequences but not in Plasmodium, E. coli or Euglena where it has been replaced by tyrosine, arginine and cysteine, residues respectively. Hudson and colleagues (1988) have proposed that this domain may form a zinc finger - zinc is known to interact with transcription factors and may be a cofactor in some nucleotidyl transferases (reviewed by Vallee and Galdes, 1984) although no direct role has been established between metal co-factors and plastid RNAPs. This domain, however, in my opinion must play an important functional role within this subunit simply due to the degree of conservation exhibited between the Plasmodium rpoC₂ and the others discussed.

The predicted amino acid length of rpoC₂ genes varies from 1527 in maize and 1513 in rice to 1361, 1391 and 1386 in spinach, tobacco and liverwort, respectively. In bacteria, that of cyanobacterium is 717 and the related region in E. coli is 829 amino acids long.

All plastid rpoC₂ genes so far analysed contain an insert of approximately 600 amino acids long at position 369 in E. coli. The insert is approximately two thirds the size in the malarial homologue (392 amino acids); red algae (Porphyra) cyanobacteria (N. commune) and Euglena having insert lengths of 430, 356 and 267 amino acids. Once again this variation in insert chain length may have an evolutionary basis which will be discussed below.
The function of a chloroplast-specific insert is unknown, although interestingly the insert size has been further increased within plant chloroplasts. This situation is further complicated by an additional insertion within this region of up to 150 amino acids in the monocots (grasses rice and maize) although not in other monocotyledons e.g. lily (amino acids 676-830 in Figure 3.8). In fact this insert has been determined to be restricted to the Pocaceae, and has been suggested to have been formed by replication slippage (Cummings et al., 1994). Slipped strand mis-pairing occurs during replication whilst the DNA strands are denatured and re-annealed; the strands initially become displaced in relation to their original positions and during repair a resultant loss or gain of nucleotides occurs (reviewed by Levinson and Gutman 1987). It is just before this region that the frameshift in *P. falciparum* occurs (position 606 in Figure 3.8). Sequence patterns have been identified in the Pocaceae rpoC2 inserts which are often associated with slipped-strand mispairing - these features include 1) repeated tandem sequence motifs; 2) repeat units within a array differing by one or two nucleotides; 3) polypurine or polypyrimidine tracts. The *P. falciparum* sequence within this region does not fulfil all the features listed above. However, a nucleotide (or amino acid) repeat is evident in Figure 3.3. Nucleotides 1875-1880 and 1893-1904 are repeated as an array at positions 1923-1928 and 1941-1952. It is possible that during a reannealing event a nucleotide in the region n=1961 was lost leading to the formation of a frame shift, and an opal stop codon at n=1962 was formed.

The Pocace specific insert (amino acids 676-830 in Figure 3.8) is composed of amino acid heptameric repeats which have been suggested to form an alpha helix which can interact with the major groove of the DNA helix thus allowing a further interaction
between the protein and DNA. This is not the complete story since recent observations have shown that deletions within this additional insert may account for male sterility in sorghum (Chen et al., 1993). What is of particular interest is that this higher plant insert occurs directly upstream of a region which is maintained in the plastids of higher plants, the sequence motif R-G-S-G-I-V-K-F (with some conservative amino acid substitutions). This sequence has some homology to the consensus sequence of RNA binding proteins - (K,R)G(F,Y)(G,V)(F,Y)VX(F,Y) (Schwemmle et al., 1989). Thus the heptamer repeat in grass chloroplasts may function as in transcription factors, the activating domain being brought into contact with the DNA by virtue of leucine zipper or \( \alpha \)-helix motifs. This is an example of an integrated transcription factor, adding to function and thereby gaining an evolutionary advantage. Other chloroplasts, however, appear to contain the RNA binding protein sequence within the \( \beta'' \) sequence without the putative functional domain assigned to the presence of the grass specific insert. One implication is that grasses are evolutionary further advanced than other plants or have diverged and therefore acquired additional properties as a result of this divergence. Such domain insertions have been identified also in the extrachromosomal RNAP of the yeast \textit{Kluyveromyces lactis} (Wilson and Meacock, 1988).

The \( rpoC_2 \) of \textit{P. falciparum} does not contain either the grass-specific insert or the RNA binding protein motif. Here, the homology resembles that of red algae (\textit{Porphyra}) and cyanobacteria (\textit{Nostoc commune}). Phylogenetic trees constructed from nucleotide sequences of \( rpoB \) from eubacteria and plastids indicate that the \textit{Plasmodium} plastome sequence is closer to that of the only other protist of which \( rpo \)
sequence is available, namely *E. gracilis*. (Gardner *et al.*, 1994). Inferences drawn here simply on the basis of domain homology indicate that the *Plasmodium* plastome encodes for an RNAP more closely related to eubacterial and protist sequences than to higher plants.

### 3.3.0 Summary and future work

The sequence data presented in this chapter indicate that the *P. falciparum* 35 kb DNA encodes an *rpo* operon similar to that identified in plastid and chloroplast genomes. To this end, *rpoC*$_1$ and *rpoC*$_2$ genes have been identified that have homology to those encoded by plastid genomes. A frameshift has been identified in the *rpoC*$_2$ gene which may disrupt the polypeptide corresponding to the β" subunit. Northern blot analysis on the *tuf* gene (presented in Chapter 9) in addition to other transcript data produced in this laboratory, suggests that the 35 kb DNA is transcribed in polycistronic units, where each transcript may be initiated at a tRNA gene or cluster. If this is the case, it is possible that were the frameshift to disrupt the functionality of the β" subunit, a transcript would still be produced in order to express the rps 2 gene distal to the *rpoC*$_2$ coding region (Chapter 5).

Future work would presumably include the determination by RNase protection of transcripts within the region containing the frameshift. In addition, the generation of antibodies to regions of the *rpoC*$_2$ gene both upstream and downstream to the frameshift would enable the determination of whether these polypeptides are produced and if so, if the *rpoC*$_2$ translation product produced as a contiguous polypeptide or two proteins. It is possible that the frame shift is a result of a mutation within the C10
strain used for the production of sequence data. In order to determine whether this is the situation, it is necessary to amplify and sequence the corresponding region in other *Plasmodium* strains.
**Figure 3.1**

Schematic representation of the protocol used to amplify the 1.3 kb PCR product encoding part of the *rpoC₁/C₂* genes.

a) A linear representation of the 35 kb DNA. *ScaI* (S) sites are indicated with the size in kb of the fragments produced by digestion with *ScaI*.

b) Amplification scheme for the generation of a 1.3 kb PCR product encoding the 3’ region of *rpoC₁* and the initial 5’ 1.174 kb of *rpoC₂*. Figure (i) shows the 1.3 kb PCR product obtained using the SK and DW3 primers.
Linear map of 35kb molecule. Restriction digestion by Sca1 (S) produces the fragment sizes indicated (kb).

10.1 kb Sca1 fragment encodes rpoB and 5' region of rpoC, in addition to small ORFs. Shaded region denotes unsequenced DNA.

Ligation of Sca1 restricted 35kb DNA with pBluescript® SK restricted by EcoRV (R) and XhoI (X) produces a mixed pool of ligation products.

PCR amplification using SK primer (internal to pBluescript®) and DW3 primer (at the 3' terminus of rpoC,) generates an expected product of 1.3kb - figure (I)
Figure 3.2

Schematic representation of the PCR protocol used to complete the sequence data of the 35 kb DNA.

The protocol used to produce the 4 kb amplification product described in the text.
Fig. a) shows the 4 kb PCR product obtained with primers KR46 and rpoD.
Fig. b) shows the two PCR products obtained with primers KR46 (KR)/R and rpo/F.
1) Use α-32P labelled 4kb PCR product to screen cDNA library

2) 40 clones isolated and sequenced

1 clone (clone 35) has homology to chloroplast RNAP B subunit

3) Design primers to the 5' and 3' (F and R) termini of clone 35 to amplify remaining regions

2 PCR products amplified; KR/R and rpoF
Figure 3.3

Nucleotide and predicted amino acid sequence of the 35 kb encoded rpoC$_1$/C$_2$ genes.

The final 375 amino acids of the rpoC$_1$ sequence is presented in blue; rpoC$_2$ sequence in green; the frameshift in turquoise; the remaining rpoC$_2$ sequence in purple; non-coding sequence in black. Numbers correspond to nucleotide position.

Asterisks denote stop codons.
579  AAA AAT AAA ATT AAT AAT ATT TAT AAT AAT AAA TAT
      K N K I N I Y N K Y
615  TAC GAA ATT AAA AAT AAT TAT ATA AAT GTA TTT TTA
      Y E I K N N Y I N V F L
651  AAT AAT TAT TAT TAT TTA AAA GTT ATA AAT AAA ATT
      N N Y Y Y L K V I N K I
687  CAA GGT ATT TTA AAT AAT TTA TAT AAT AAA ATT
      Q G I L N N N L Y N K I
723  AAT CCT ATT TAT TCA AAT TTA TTT TTA TTT TTT AAT
      N P I Y S N L F L F N
759  AAT AAA ATA AAA ATA AAA TAT TCT CAA TTA CAA CAA
      N K I K I K Y S Q L Q Q
795  TTA ATA GGT TAT AAG GGT TAT ATT TCT AAT ATA AAA
      L I G Y K G Y I S N I K
831  GGA ATG ATT TAT GAA AAA CCT GTT ATA AAT AAT TAT
      G M I Y E K P V I N N Y
867  ATA AAT GAA TTA AAT ATA TAT GAA TAT ATT TTA TCT
      I N E L N I Y E Y I L S
903  TGT TAT GGA TCT AAA AAA GGT ATA ATT GAT ACA GCT
      C Y G S K K G I I D T A
939  TTA AAA ACA GCA GAT TCA GGA TAT TTA ACA AAA CGT
      L K T A D S G Y L T K R
975  TTA ATA AAT ATT ACA AGT AAT TTT ATA ATA AAA GAA
      L I N I T S N F I I K E
1011 TTA AAT TGT AAA TCA CCT TTT ATA TTA AAA TAT ATA
       L N C K S P F I L K Y I
1047 TTA AAT ATG GAT ATA TAT GGT AAT ATT ATA TTA CCT
       L N M D I Y G N I I L P
1083 TTA AAT ATA TTA AGA TTT AAA ATT TTA CAA AAT AAT
       L N I L R F K I L Q N N
1119 ATT TTA AAT TTA AAT AAT GGT ACT TTT ATT TAT ACA
       I L N L N N G T F I Y T
1731 AAA AAT ATC CAT AAT AAA TGG ATT CTA TAT AAT ATT K N I H N K W I L Y N I
1757 TAT ACA TAT TAT TTA TAT TAT CAT ATA AAG TTT Y T Y Y L Y Y Y H I K F
1803 TAT AAT TTA TAT AAT AAA GGT ATT ATT TTA AAT AAT Y N L Y N K G I I L N N
1839 AAT AAT AAT AAG TAT AAT GTT ATA TAT TTT TTA ATT N N N K Y N V I Y F L I
1875 AAT TAT TTT AAT TTA TTT TCA AAT TAT TAT TAT AAA N Y F N L F S N Y Y K
1911 ATT TAT AAT AAT AAT TAT AAT TTT ATT AAT TCA AAT I Y N N N Y N F I N S N
1947 TAT TAT TTT AAT AAATGAAT TTT ATA TTA AAA AAT Y Y F K K * M N F I L K N
1982 TTT AAT AAT ATA CAA ATT TTA AAT AAA TTA TTT TAT F N N I Q I L N K L F Y
2018 GTA AAT AAT ATA TTT ATA TAT TAT AAA TAT GAA AAA V N N I F I Y Y K Y E K
2054 AAA TTA TTT ATA TAT TTA AAT ATT ATA AAT AAT ATT K L F I Y L N I I N N I
2090 ATA ATT AAA AAA TAT TTA AAT TTT TAT AAG TAT ACT I I K K Y L N F Y K Y T
2126 TAT AAT AAA TTA TTT TTT ATA AAA AAA TAT AAT AAT AAT Y N K L F F I K K Y N N
2162 TTT TTA TAT TTA TAT GAA ATA TTT AAA TAT AAT TGG F L Y L Y E I F K Y N W
2198 TAT AAA TAT TTA TTG TTA AAT AAT AAA TAT TTA Y K Y L L L N N K Y N L
2234 TAT ATT ATA TAT AAT AAT TAT ATT AAA TAT TTA TAT Y I I Y N N Y I K Y L Y
2270 AAA TAT AAT ATA AAT ATT AAT TTA TAT TTT ATA AAA K Y N I N I N L Y F I K
2306  AAT TTA TTT TAT AAT AAT AAT TTT ATA CAT AAT
       N L F Y N N N N F I H N
2342  CAT ATT ATA TAT AAA AAT AAT TAT TAT ATT TAT AAT
       H I I Y K N N Y Y I Y N
2378  AAT AAT ATG AAT TTA TAT CAA TAT AAT AAA AAT ATT
       N N M N L Y Q Y N K N I
2414  TTA ATA AAT AAT AAT TTA TAT AAT AAA TTA TTT
       L I N N N L L Y N K L F
2450  TAT AAT TAT ATT AAT AAT ATT TAT AAT TTA TAT
       Y N Y I N N I Y N L Y
2486  TTA AAT GAT ATT ACT ATA GGT TTA CAA TCT ATA AAT
       L N D I T I G L Q S I N
2522  ATA ATA TTT GAA AAT AAA AAT ATA AAA GAT AAT ATT
       I I F E N K N I K D N I
2558  TTT TTT ATT TCA AAT AAT ATT TAT GTA ATA TTT TAT
       F F I S N N I Y V I F Y
2594  ATA AAA TAT TAT AAT TTA AAT AAT ATT ATA TAT
       I K Y Y N Y L N N I I Y
2630  ATA TAT AAT ATA TGT AAT AAA TAT AAT ATT AAT CAT
       I Y N I C N K Y N I N H
2666  TAT AAA TAT AAA TTA AAT TTT TAT TCT TAT ATT TTT
       Y K Y K L N F Y S Y I F
2702  GAA GAT ATT AGT TCT ATT TTA TAT AGT GGA TAT TCT
       E D I S S I L Y S G Y S
2738  TTA CAT ACA GAA TTT TAT TCA ATT AAT AAA AAT TTA
       L H T E F Y S I N K N L
2774  AAA TAT TAT TTT AGA TTT TTA AAA TCT ATA AAT
       K Y Y F R F L L K S I N
2810  ATT TAT CAA GCA ACT AAA AGT TCA TAT ATA TAT GTA
       I Y Q A T K S S Y I Y V
2846  TAT AAT ATT TTA ATT GAA TCT ATT TIG AAA CAA TAT
       Y N I L I E S I L K Q Y
2882  AGT TAT CAA AAT ATT TAT TTA CCT TCT ATT TAT TTT  
2898  SY Q N I Y L P S I Y F
2914  GAA CTT ATT ATA AAA AAA ATG TTA TCA TGT ATA AAG  
2930  E L I I K K M L S C I K
2946  ATA ATA TCT AAT AAT TTT AAA ATA TTT AAA TAT AAT  
2962  I I S N N F K I F K Y N
2978  GAT ATA ATA TCT TTA CAA TTA ATA AAT ATA ATT AAT  
3000  D I I S L Q L I N I I N
3021  TAT TCT TTA AAT TTA AAT AAA CAT TAT ATT TAT AAA  
3042  Y S L N L N K H Y I Y K
3063  TAT GAA CCA ATT ATT TTA GGT ATT ACA AAA TCA ATT  
3084  Y E P I I L G I T K S I
3105  TTA GCT AAT TCT GGA TTT TTA ACA AAT ATA AGT TTT  
3126  L A N S G F L T N I S F
3147  CAG AAT ACA TTT AAA ATA ATA AGT TTA AAT ATT TTA  
3168  Q N T F K I I S L N I L
3190  AAT AAT AAA ATT GAT TGG TTA ATA GAT ATA AAA TCT  
3212  N N K I D W L I D I K S
3234  AAA ATA ATA TTAG ACT GAT TTA TTA CTT GTT GGT AAT  
3256  K I I L T D L L P V G N
3278  GGA TGG TAT AGA TAT TTA GTA AAT TAATTTTTTAAAAT  
3300  G W Y R Y L V N *

Figure 3.4

Representation of the \textit{rpo} genes and neighbouring ORFs.

Diagram indicating the position of the \textit{rpo} genes and the neighbouring open reading frames (ORF). Primers RW31, RT1 and APX 1 are indicated.
Figure 3.5

Primer extension analysis of the \textit{rpo} operon.

Primer extension analysis of \textit{rpoB} using primer APX1. Two signals were typically obtained, the uppermost signal indicated by the shaded arrow relates to the presumed start site of transcription. The lower signal is presumed to be an artefact (see text). A nucleotide ladder of known sequence was electrophoresed alongside the primer extension product as a marker (GATC).
Figure 3.6

RT-PCR analysis of the $rpoB/C$ junction

Lane PCR    PCR product using primers RT1 and RW31
Lane RT-PCR Reverse transcriptase PCR across the $rpoB/C$ junction
Lane R      RNase treated template followed by RT-PCR
Figure 3.7

PILEUP analysis of the 35 kb encoded \textit{rpoC}_1 gene

The following genes are compared: \textit{P. falciparum} (Plasmodium) (Accession number: x95275); \textit{M. polymorpha} (P06273); \textit{N. tabacum} (P12116); Prochloroc DV1 (Urbach \textit{et al}., 1992); \textit{Prochloroc} LG (Urbach \textit{et al}., 1992); \textit{Prochlorothrix} (Urbach \textit{et al}., 1992); \textit{Nostoc commune} (nosco) (P14563); \textit{Cyanophora paradoxa} (Urbach \textit{et al}., 1992); \textit{Euglena gracilis} (P23580); \textit{Escherichia coli} (P00578).

The eukaryotic regions of homology A-E have been included (Jokerst \textit{et al}., 1989) at the appropriate positions.

Gene fragments or incomplete sequences are presented within brackets $\langle \rangle$.

The \textit{E. coli} \textit{rpoC} sequence has been split into two region bearing homology to the chloroplast \textit{rpoC}_1/C_2 as described by Hudson (1988).

Numbers correspond to the amino acid positions within the PILEUP.

Amino acids identical to the \textit{Plasmodium} sequences are shaded in yellow. Deletions inserted for optimal alignment are represented by dots (dashes in the eukaryotic sequences).
M. polymorpha  GPIEIQQKSF GSFQIYEHY QLRKNNQEI ISTYICTTAG
N. tabacum      TPIEVIHYESL GTFYEIYGHY LIVRSLLLKI LFIYIIRTIVG
N. commune  .PVKVTENED GTRTLYKFR RVRQDAKGNV LSQYIYTPG
E. gracilis  RMVKKL.... ........... ........VF FRTTIGRIFF

771
M. polymorpha  RILFNNQIEE AIQGTYKASL KQKFVFQKIE KNG
N. tabacum      HIALYREIEE AIQGFSRAYS SGT
N. commune      RVIYNNAIQE ALAS
E. gracilis    DDMIKEFL
Figure 3.8

PILEUP analysis of the rpoC₂ genes of *P. falciparum* (Accession number: x95275), *Z. mays* (maize) (accession number P16025); *Orysa* (rice)(P12093), *Spiola* (spinach)(P11704); *M. polymorpha* (P06274); *P. purpurea* (Reith, personal communication), *N. commune* (P14564); *E. gracilis* (P23581); *E. coli* (P00578). The eukaryotic consensus block F has also been included (Jokerst *et al.*, 1989).

Symbols used are as described in Figure 3.7.

The putative zinc finger discussed in the text is overlined. The frameshift in the *P. falciparum* rpoC₂ gene is indicated by an asterisk and the putative transcription factor domain described in the text is boxed.
<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. mays</em></td>
<td>IYSESDGEMH</td>
<td>WSTDVYHAPE</td>
<td>YQYGNRLRLRP</td>
<td>KTSHELWILS</td>
</tr>
<tr>
<td><em>Orysa</em></td>
<td>IYSESDGEMH</td>
<td>WSTDVYHAPE</td>
<td>YQYGNRLRLRP</td>
<td>KTSHELWILS</td>
</tr>
<tr>
<td><em>Spiola</em></td>
<td>IYSDSEGMHM</td>
<td>WSTDVYHAPE</td>
<td>FTYGNVHLLP</td>
<td>KTSHELWILS</td>
</tr>
<tr>
<td><em>M. polymorpha</em></td>
<td>YSNLEEGEMH</td>
<td>WSTKVRHA</td>
<td>YIHNSIHLIL</td>
<td>KTCHIWILS</td>
</tr>
<tr>
<td><em>P. pupurea</em></td>
<td>VLSDSLGSKIC</td>
<td>FSNLTVVE.E.T</td>
<td>DNNQYTRTRT</td>
<td>KTGLLWILS</td>
</tr>
<tr>
<td><em>&lt;N. commune&gt;</em></td>
<td>VASDIAESEQV</td>
<td>FAEVEPEQKT</td>
<td>DROGNTTTRT</td>
<td>RTRGLWILS</td>
</tr>
<tr>
<td><em>E. gracilis</em></td>
<td>VKAKISQGQIF</td>
<td>FPQI........</td>
<td>NSKQFGKIFW</td>
<td>IISSIMSF</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>VITEVSFGVR</td>
<td>FTMDI..DGQ</td>
<td>TITRQTDL</td>
<td>GLSSLVLDS</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>ITIKYLNVIK Y</td>
<td>..............</td>
<td>YNYSIQLLI</td>
<td>KNIHKWILY</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. mays</em></td>
<td>VSMC.RSSIA</td>
<td>SFSLHKDQDOQ</td>
<td>MNTYSPSVDG</td>
<td>RYIFDFSMAN</td>
</tr>
<tr>
<td><em>Orysa</em></td>
<td>VSMC.RSSIA</td>
<td>SFSLHKDQDOQ</td>
<td>MNTYSPSVDG</td>
<td>RYIFGLSMAD</td>
</tr>
<tr>
<td><em>Spiola</em></td>
<td>GRKY.RSSVV</td>
<td>PFSLSKDQDOQ</td>
<td>MNTYLSFEQ</td>
<td>IVISNPSVTN</td>
</tr>
<tr>
<td><em>M. polymorpha</em></td>
<td>GNHFHHKNDNL</td>
<td>SVLYFKNQDK</td>
<td>IDFPISLTKE</td>
<td>KNEFSFVKNK</td>
</tr>
<tr>
<td><em>P. pupurea</em></td>
<td>GEVYNIIPDA</td>
<td>NIPVNMKADQV</td>
<td>NGPIKLIL</td>
<td>TELINQYSK</td>
</tr>
<tr>
<td><em>&lt;N. commune&gt;</em></td>
<td>GEVYNLPPGA</td>
<td>ELVVKNGDAI</td>
<td>ASNGVLI..AE</td>
<td>TKLASLHGGV</td>
</tr>
<tr>
<td><em>E. gracilis</em></td>
<td>SLFFHHTKEF</td>
<td>SF...KNKI</td>
<td>CPTTNV..HK</td>
<td>KEL..FVNRK</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>AERTAGGKDL</td>
<td>RPALKIVDAQ</td>
<td>GNDVLIPGT</td>
<td>..........</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>N.............</td>
<td>IYYTYLYYYH</td>
<td>IKFYNLNYKG</td>
<td>..............</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. mays</em></td>
<td>DVQVSHRLL..</td>
<td>DTFGKKDREI</td>
<td>LDYLTIPDV</td>
<td>FNGHWNCYP</td>
</tr>
<tr>
<td><em>Orysa</em></td>
<td>DEVHRHRL..</td>
<td>DTFGKKDREI</td>
<td>LDYSTPDRIM</td>
<td>SNGHWNFVYP</td>
</tr>
<tr>
<td><em>Spiola</em></td>
<td>DQVKDKL..S</td>
<td>DFSKKKEDRI</td>
<td>TDYSEIRNRE</td>
<td>HCNL...IYP</td>
</tr>
<tr>
<td><em>M. polymorpha</em></td>
<td>TOYNLFHFHLF</td>
<td>YLY.KKNKIP</td>
<td>IKSQTLNTR</td>
<td>KNINSKKNYN</td>
</tr>
<tr>
<td><em>P. pupurea</em></td>
<td>VRIYQH........</td>
<td>TSS.SI</td>
<td>TNIQIITEEV</td>
<td>TVPACYRITD</td>
</tr>
<tr>
<td><em>&lt;N. commune&gt;</em></td>
<td>VRLPEA........</td>
<td>TPGKST</td>
<td>REIEIITASV</td>
<td>VLDQATVTVQ</td>
</tr>
<tr>
<td><em>E. gracilis</em></td>
<td>KELFPR........</td>
<td>KFGKLH........</td>
<td>KINRNVSLV</td>
<td>..............</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>................</td>
<td>................</td>
<td>................</td>
<td>..............</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>IILNANMNKY</td>
<td>NVTYFLINYP</td>
<td>NLPSNYYKI</td>
<td>YNNYNFINS</td>
</tr>
<tr>
<td>Start</td>
<td>End</td>
<td>Species</td>
<td>Sequence</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-----------------------</td>
<td>---------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>641</td>
<td>680</td>
<td>Z. mays</td>
<td>SMEIPFMGVL RRNTIFAYFD DPRYRKDKKG SGIVKFRYRT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orysa</td>
<td>SMEIPFMGVL RRNTIVAYFD DPRYKKDKKG SGIVKFRYRT</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiola</td>
<td>SMEIPINGIF RKNSIFAYFD DPRYRKRS SGITK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M. polymorpha</td>
<td>FLKIKKNGVL KNNEIFAILD DPSYKV KN SGILK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. pupurea</td>
<td>.FKSFPDQKI SDGHIVALEI SDTYK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;N. commune&gt;</td>
<td>.LRATPGTKQ QNGQVVAELI DDRYR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. gracilis</td>
<td>...........KK RSHHTICIL DQ</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. coli</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. falciparum</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td>681</td>
<td>720</td>
<td>Z. mays</td>
<td>LEEEYRTQEE EYRTREEEYR TREEDSEDEY ESPENKRYTR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orysa</td>
<td>L............. EDEYR TREKDSENYE GSPENKRTR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiola</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M. polymorpha</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. pupurea</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;N. commune&gt;</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. gracilis</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. coli</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. falciparum</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td>721</td>
<td>760</td>
<td>Z. mays</td>
<td>EGEKEYKILE DEYRTLEDEY ETLEDEYGIL EDEYRTLEKD</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orysa</td>
<td>.EDECKTL DEYRTREEEY ETLEDEYGIP ENEYETL</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiola</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M. polymorpha</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. pupurea</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;N. commune&gt;</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. gracilis</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. coli</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. falciparum</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td>761</td>
<td>800</td>
<td>Z. mays</td>
<td>SEELEYGSLEN KYRTREGEGE YEILEEESEE EYGSSSSED</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orysa</td>
<td>.EDEYGILED EYTR.... EEEESEY EYSPENKRY</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spiola</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M. polymorpha</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. pupurea</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;N. commune&gt;</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. gracilis</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>E. coli</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P. falciparum</td>
<td>..................................................................</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z. mays</td>
<td>Orysa</td>
<td>Spiola</td>
<td>M. polymorpha</td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>--------------</td>
</tr>
<tr>
<td>801</td>
<td>GSEKEYGTE</td>
<td>EDS....</td>
<td>EEDSEEDSED</td>
<td>EYGSPEEDSI</td>
</tr>
<tr>
<td>803</td>
<td>Z. mays</td>
<td></td>
<td></td>
<td>Orysa</td>
</tr>
<tr>
<td>805</td>
<td>Spiola</td>
<td></td>
<td></td>
<td>M. polymorpha</td>
</tr>
<tr>
<td>807</td>
<td>P. pupurea</td>
<td></td>
<td></td>
<td>&lt;N. commune&gt;</td>
</tr>
<tr>
<td>809</td>
<td>E. gracilis</td>
<td></td>
<td></td>
<td>E. coli</td>
</tr>
<tr>
<td>811</td>
<td>P. falciparum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>840</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>842</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>844</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>846</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>848</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>850</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>852</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>854</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>856</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>858</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>860</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>862</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>864</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>866</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>868</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>870</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>872</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>874</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>876</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>878</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>880</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>882</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>884</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>886</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>888</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>890</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>892</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>894</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>896</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Z. mays</td>
<td>Orysa</td>
<td>Spiola</td>
<td>M. polymorpha</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>961</td>
<td>KRKNWVVYQQR</td>
<td>KKFLKSKEKY</td>
<td>FVSVRPFAVAY</td>
<td>EMDEGINLAT</td>
</tr>
<tr>
<td>1000</td>
<td>KRKNWVVYQQR</td>
<td>KKFLKSKEKY</td>
<td>FVSVRPFAVAY</td>
<td>EMDEGINLAT</td>
</tr>
<tr>
<td>1001</td>
<td>LFPQDLLQEE</td>
<td>DNLQLRLVNVF</td>
<td>ISHENSKLTQ</td>
<td>RIYHTNSQFV</td>
</tr>
<tr>
<td>1040</td>
<td>LFPQDLLQEE</td>
<td>DNLQLRLVNVF</td>
<td>ISHENSKLTQ</td>
<td>RIYHTNSQFV</td>
</tr>
<tr>
<td>1081</td>
<td>LVKSTILYTR</td>
<td>RRYDRTSVGL</td>
<td>IPNRLDRNN</td>
<td>...TNSFYSK</td>
</tr>
<tr>
<td>1120</td>
<td>LVKSTILYTR</td>
<td>RRYDRTSVGL</td>
<td>IPNRLDRNN</td>
<td>...TNSFYSK</td>
</tr>
<tr>
<td>1120</td>
<td>YINNNNI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1281</td>
<td>1290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z. mays</td>
<td>MLEPFHLNWH</td>
<td>FLHHDSWEET</td>
<td>LAIIHLGQFI</td>
<td>CENLCLFKLH</td>
</tr>
<tr>
<td>Orysa</td>
<td>MLDPFQLNWH</td>
<td>FLPHDSWEET</td>
<td>SAKIHLGQFI</td>
<td>CENVCLFKSH</td>
</tr>
<tr>
<td>Spiola</td>
<td>IFNPFPKLNNV</td>
<td>FLHYHFCEET</td>
<td>STKIDLGQFV</td>
<td>CENVCITKKG</td>
</tr>
<tr>
<td>M. polymorpha</td>
<td>NYNLFPWCFIP</td>
<td>LFSSLKHKID</td>
<td>FQTIKLGQLL</td>
<td>FENFVISKYK</td>
</tr>
<tr>
<td>P. pupurea</td>
<td>TTTSDNKVPD</td>
<td>LHQQNTLQVV</td>
<td>GDWIRCGDIL</td>
<td>SESII......</td>
</tr>
<tr>
<td>&lt;N. commune&gt;</td>
<td>............</td>
<td>............</td>
<td>............</td>
<td>............</td>
</tr>
<tr>
<td>E. gracilis</td>
<td>............</td>
<td>............</td>
<td>............</td>
<td>............</td>
</tr>
<tr>
<td>E. coli</td>
<td>............</td>
<td>............</td>
<td>............</td>
<td>............</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>NNYKYNLNFY</td>
<td>............</td>
<td>............</td>
<td>............</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1321</th>
<th>1360</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. mays</td>
<td>IK......</td>
<td>KSG</td>
</tr>
<tr>
<td>Orysa</td>
<td>IK......</td>
<td>KSG</td>
</tr>
<tr>
<td>Spiola</td>
<td>TH......</td>
<td>LKG</td>
</tr>
<tr>
<td>M. polymorpha</td>
<td>TS......</td>
<td>YPSG</td>
</tr>
<tr>
<td>P. pupurea</td>
<td>.....</td>
<td>SLDG</td>
</tr>
<tr>
<td>E. gracilis</td>
<td>............</td>
<td>............</td>
</tr>
<tr>
<td>E. coli</td>
<td>............</td>
<td>............</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>............</td>
<td>............</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1361</th>
<th>1400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. mays</td>
<td>GKIYKGDRL</td>
<td>VTFIYEKRS</td>
</tr>
<tr>
<td>Orysa</td>
<td>GEIYKGDRL</td>
<td>VTFIYEKRS</td>
</tr>
<tr>
<td>Spiola</td>
<td>GIEIYEGDTL</td>
<td>VTFIYEKRS</td>
</tr>
<tr>
<td>M. polymorpha</td>
<td>GEIFKEGDTL</td>
<td>ITLYERLKS</td>
</tr>
<tr>
<td>P. pupurea</td>
<td>NALIRGGETT</td>
<td>AVLFVDAKRT</td>
</tr>
<tr>
<td>E. gracilis</td>
<td>KPFIRKNQLEL</td>
<td>YRVNFREKT</td>
</tr>
<tr>
<td>E. coli</td>
<td>EEMIPKWRQFL</td>
<td>NVFEGERVER</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>SSILYSYSL</td>
<td>HTEFYVSINK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1401</th>
<th>1440</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z. mays</td>
<td>DSLSPNLERR</td>
<td>IEDWNERIPR</td>
</tr>
<tr>
<td>Orysa</td>
<td>DSLSPNLERR</td>
<td>IEDWNERIPR</td>
</tr>
<tr>
<td>Spiola</td>
<td>DSLSPNLERR</td>
<td>IEDWNERIPR</td>
</tr>
<tr>
<td>M. polymorpha</td>
<td>NSVSINENGL</td>
<td>FEDWNNMIDK</td>
</tr>
<tr>
<td>P. pupurea</td>
<td>T......</td>
<td>DVLINP</td>
</tr>
<tr>
<td>E. gracilis</td>
<td>SSLKEIINNP</td>
<td>HDILTESFET</td>
</tr>
<tr>
<td>E. coli</td>
<td>HAVTRYI......</td>
<td>............</td>
</tr>
<tr>
<td>P. falciparum</td>
<td>............</td>
<td>YPR</td>
</tr>
</tbody>
</table>
Z. mays
Orysa
Spiola
M. polymorpha
P. pupurea
E. coli

Z. mays
Orysa
M. polymorpha

1601 1640

Z. mays KNLFAEMRD ILFLHTEL.V SSDSDVTNNF YETSETPFTP
Orysa KKLFAEMRD ILFLHTEL.V SSDSDVTNNF YETSESPFTP
Spiola QNLFEAGMD ILFYHREL.F ...ESCKSKN
M. polymorpha KEFFTKKINN VFYQDFFSI FPTTEIIHV N LKESIQMNK
P. pupurea PPSVRDLDI IILDDRTARN YFNNKTVD
E. coli SASLAEILNA GLGSDNE

1641

Z. mays IYTI
Orysa FI
M. polymorpha NNFSI
CHAPTER 4
CHAPTER 4

THE IDENTIFICATION OF \textit{rpo} GENE TRANSLATION PRODUCTS

4.1.0 Introduction

From the sequence of the 35 kb circle, three \textit{rpo} genes have been identified, as well as their transcripts. Comparisons between the \textit{rpoB} nucleotide sequences of \textit{E. coli} and \textit{P. falciparum} (Gardner et al., 1991b) shows that the \textit{P. falciparum rpoB} gene has 15\% nucleotide homology and 27\% amino acid identity with \textit{E. coli}. There are seven conserved domains between the two polypeptides including a zinc finger domain and the rifampicin binding domain. However, there are no reports on the polypeptides encoded by the \textit{P. falciparum rpo} genes.

The generation of antibodies to the \textit{rpoB} gene product would enable both the determination of gene expression and could be used to localise products of the 35 kb genome within the malarial parasite. Accordingly, two methods were employed to synthesise antibodies to the \textit{\beta} subunit: i) the production \textit{in vitro} of mouse monoclonal antibodies to synthetic polypeptides and ii) the production of rabbit polyclonal antibodies against fusion proteins. Additionally, a polyclonal antibody to \textit{E. coli} RNAP holoenzyme (\textit{EcB})(kindly provided by Dr R. Hayward, Edinburgh university) was used to determine whether, at the level of Western blot analysis, a degree of antigenic immunogenicity exists between the plastid RNAP subunits of \textit{P. falciparum} and those of \textit{E. coli}. 

108
4.2.0 Results and Discussion

4.2.1 The synthesis of monoclonal antibodies to predicted polypeptide sequences

Two peptide sequences IW376 and IW227 were synthesised at the NIMR. Both peptides were chosen on the basis of their amino acid composition and location within the protein; immunogenicity was predicted using the standard methods of Kyte and Doolittle (1982) and Hopp and Woods (1983). Figure 4.1 shows the location of the two peptides within the β subunit. IW227 is located in a region conserved between *P. falciparum* and *E. coli*, whereas the second peptide IW376 was selected from a non-homologous region. The antibody raised to IW227 was therefore predicted to be immunoreactive against the *E. coli* β subunit; this was thought to be useful in defining the specificity of the antibodies generated at the level of Western blot analysis.

The two polypeptides were independently covalently linked to Keyhole Limpet haemocyanin (KLH) and used to generate monoclonal antibodies using an *in vitro* immunisation protocol according to the manufacturers instructions (with the help of Malcolm Strath).

Hybridoma products in culture fluids were screened using ELISA assays and positive cultures were cloned by serial dilution. These clones were grown up in batch cultures and the culture supernatant used to probe Western blots of *P. falciparum* protein extract, *E. coli* whole cell lysate, as well as purified *E. coli* RNAP (all purified *E. coli* RNAP referred to throughout this thesis was supplied by Boehringer Mannheim). Only one culture supernatant, 2E2.1, recognised a discrete *Plasmodium* protein migrating at the molecular weight predicted for that of the β subunit (see below). The
other supernatants recognised several *Plasmodium* polypeptides or purified *E. coli* RNAP β and β' subunits only, and were therefore not characterised further. Antibody typing of 2E2.1 identified the immunoglobulin type as IgM. According to the manufacturers information only 10% of the antibodies produced should be IgM the remainder being IgG. There are technical difficulties associated both with purifying IgM immunoglobulins and using them to make an affinity select column, so for these reasons it was decided instead to raise antibodies to fusion proteins generated from the *rpoB* gene (see later).

### 4.2.2 Characterisation of 2E2.1 and the anti-*E. coli* RNAP antibody (*EcB*)

The clone expressing this antibody was unstable and having been stored in liquid nitrogen in the presence of DMSO it no longer produced antibody to the peptide. Accordingly, experiments performed using this antibody were limited due to the small amount of culture supernatant available.

The 2E2.1 culture supernatant was used to probe Western blots of malarial trophozoite whole cell lysate, and purified *E. coli* RNAP holoenzyme (Boehringer Mannheim). As can be seen in Figure 4.2 this antibody recognised an *E. coli* polypeptide doublet migrating at 150 kDa: The β and β' subunits of the *E. coli* holoenzyme migrate as a doublet (150 kDa and 155 kDa respectively) on 7.5% SDS-PAGE gels. On the other hand, the most predominant *P. falciparum* signal, was a band migrating at approximately 120 kDa (see arrow in Figure 4.2) that is assumed to be the β subunit:

The predicted molecular mass of the *rpoB* gene product determined from nucleotide sequence data is 122.5 kDa. In addition, a higher molecular mass polypeptide
migrating at approximately 200 kDa was recognised weakly by the 2E2.1 culture supernatant. High molecular mass polypeptides have been observed occasionally in other *P. falciparum* immunoblots probed with the anti *EcB* antibody and it has been assumed these reactions are non-specific.

Parallel blots were probed with the polyclonal anti-*E. coli* (*EcB*) antibody - Figure 4.3a. This antibody recognised a multiplicity of *E. coli* polypeptides in the track containing purified *E. coli* RNAP but the predominant signals observed correspond to the β' subunit (155Kd), the β subunit (150 kDa), the σ subunit (70 kDa) (indicated in Figure 4.3b.) and the α subunit (36 kDa) (-not indicated in Figure 4.3b). The other minor signals are presumably due to antibody recognition of RNAP degradation products. This antibody also recognised a *P. falciparum* protein migrating at approximately 120 kDa (arrow in Figure 4.3a), suggesting that there might be antigenic cross reactivity between the *E. coli* anti-β subunit antibody and the malarial β subunit. Since no other signals were apparent, the remaining subunits associated with the *P. falciparum* 35 kb encoded RNAP must be assumed to be unable to cross-hybridise with this particular anti-*E. coli* antibody.

Because of the apparent share of antigenicity between the β subunits of *P. falciparum* and *E. coli* an adsorption assay was carried out to ascertain the specificity of the signals observed: Assuming the *P. falciparum* 120 kDa band corresponds to the β subunit, this signal should be lost by pre-absorbing the antibody probe with an excess of *E. coli* β subunit. Polyclonal anti-*EcB* antibody at the same working dilution as used previously, and 2E2.1 culture supernatant, were independently incubated with an
excess of *E. coli* RNAP at 37°C for one hour. This mixture, containing RNAP and antibody complex, was used to probe Western blots identical to those described above. These results are summarised in Figure 4.3b. As predicted, the absorption mixture recognised the *E. coli* RNAP less intensely than the un-complexed antibody (compare Figure 4.3 a. with b.). In addition, the *P. falciparum* 120 kDa polypeptide was no longer detected. Therefore, the concentration of *E. coli* RNAP used was not great enough to pre-absorb all of the antibodies to the *E. coli* holoenzyme but was sufficient to compete out those antibodies recognising the *P. falciparum* β subunit. However, the signal-noise ratio, particularly in the track containing *P. falciparum* extract was increased. This might have been due to non-specific protein-protein interaction between the immobilised nitrocellulose-bound malarial extract and the RNAP-antibody complexes. An additional band migrating at approximately 140 kDa was observed (asterisked in Figure 4.3 b) that had not been detected in an identical track probed with anti-αE2 antibody alone (Figure 4.3a. (P.f)). No signals were observed when immunoblots were probed with the 2E2.1 culture supernatant pre-absorbed with *E. coli* RNAP (data not shown). This result confirmed that the malarial anti-β subunit antibody recognised the *E. coli* β subunit and was completely pre-absorbed by the RNAP.

These absorption experiments suggest that the polyclonal anti-αE2 antibody recognised a *P. falciparum* polypeptide of approximately 120 kDa and that this interaction was competed out in the presence of purified *E. coli* holoenzyme. It follows that there is antigenic cross reactivity between the *P. falciparum* β subunit and components of the *E. coli* holoenzyme. The complete absorption of those antibodies in
the polyclonal serum which recognised the parasite \( \beta \) subunit presumably relates to the small amount of polypeptide present in the \( P. falciparum \) lysate. In addition, recognition of the 120 kDa \( P. falciparum \) polypeptide and the 150-155 kDa \( E. coli \) proteins by antibodies in the 2E2.1 culture supernatant could be competed out by pre-absorption with an excess of \( E. coli \) holoenzyme, a further indication of cross reactivity.

### 4.2.3 The determination of shared immunogenicity between the RNAPs of \( E. coli \) and \( P. falciparum. \)

Having established with the polyclonal antibody that there is antigenic cross-immunoreactivity between \( E. coli \) and \( P. falciparum \) RNAP subunits, antibodies to each of the anti-\( E. coli \) RNAP subunits were affinity selected from the polyclonal serum. Figure 4.4a shows the typical pattern observed when purified \( E. coli \) RNAP was electrophoresed on a 7.5% SDS-PAGE gel and stained with Coomassie Brilliant Blue R-250. The affinity selected antibodies were used to probe immunoblots of \( P. falciparum \) whole cell lysates separated on SDS-PAGE, \( E. coli \) whole cell lysate, and \( E. coli \) RNAP (Boehringer Mannheim). Identical Western blots were probed, in parallel, with the polyclonal anti-\( E. coli \) RNAP serum. These results are shown in Figure 4.4b. and c. and Figure 4.5.

The affinity selected anti-\( E. coli \) \( \beta \) and anti-\( E. coli \) \( \beta' \) antibodies recognised a discrete \( P. falciparum \) polypeptide migrating at 120 kDa and a second band migrating at approximately 75 kDa, which would correspond to the predicted molecular weight of the \( \beta' \) subunit (70 000 \( M_r \)) - Figure 4.4c. As expected, the affinity selected anti-\( E. coli \)...
β/β' antibodies produced a significantly stronger signal against the *P. falciparum* 120 kDa and 75 kDa polypeptides than the anti-*EcB* polyclonal antibody serum (compare Figure 4.4b. and c.). Presumably this indicates an increase in concentration of the appropriate antibodies. A *P. falciparum* polypeptide migrating at approximately 105 kDa was weakly recognised by both the anti-*EcB* antibody and the affinity selected anti-*E. coli* β/β' antibodies (observed as a weak signal in Figure 4.4b and part of the diffuse signal in Figure 4.4c). This signal may correlate to the malarial β" subunit (predicted 117 000 M_0). If this is the case, I suggest that the frameshift within the *P. falciparum rpoC_2* gene, discussed in Chapter 3, has not disrupted the translation of the gene and a single polypeptide is produced. However, if the frameshift were to disrupt the coding sequence of this gene and assuming that epitopes were shared between both of the resultant polypeptides and the *E. coli* β' subunit, I would have expected to see two distinct polypeptides migrating at approximately 52 kDa and 64 kDa corresponding to the two coding regions within the split *rpoC_2* gene.

In addition, the other affinity select probes used failed to recognise any *P. falciparum* polypeptides corresponding to the α and σ subunits predicted to be utilised by the 35 kb-encoded RNAP. Troxler and colleagues (1994) were able to identify the σ subunit of maize, rice, *Chlamydomonas reinhardtii*, and *Cyanidium caldarium* using an antibody directed to the σ subunit of cyanobacterium *Anabaena* spp. PCC 7120. The gene encoding the σ subunit is not encoded by chloroplast genomes and has always been assumed to be nucleus encoded. This was the first evidence that the σ subunit is a component of chloroplast RNAP.
Unfortunately, the anti-*E. coli* antibody used here, has not allowed identification of any component of the 35 kb RNAP, other than the β subunit and possibly the β' subunit. Now that the entire 35 kb plastome has been sequenced (Wilson *et al.*, 1996), no genes corresponding to those encoding either the α or σ subunits have been identified. Thus, in line with the evidence that the chloroplast RNAP σ subunit is nuclear encoded, one may assume that both the α and σ *P. falciparum* genes are located in the nucleus.

The affinity select probes were not totally specific in recognising the corresponding *E. coli* polypeptides from which they had been purified. The co-purified anti-β and anti-β' antibodies also recognised smaller polypeptides in both the purified RNAP track and the *E. coli* whole cell lysate (wcl), corresponding to β/β' degradation products - Figure 4.4c. The anti-σ antibody recognised both β/β' subunits as well as the σ subunit, in addition to degradation products in each of the two bacterial protein preparations - Figure 4.5. The anti-α subunit antibody hybridised to all four components of the multimeric *E. coli* RNAP. This experiment has been repeated and the results verified.

To interpret this finding, I assume that during affinity selection, the SDS-PAGE separated subunits were still contaminated with degradation products of the higher molecular mass polypeptides. Therefore anti-β/β' antibody was co-purified with anti-σ antibody, and all four anti-subunit antibodies were purified with the lowest molecular mass subunit - the α subunit. The concentration of co-purified antibody is assumed to be low since cross-hybridisation with the *P. falciparum* β subunit was only observed in experiments using the anti-β/β' selected antibodies and not with the other preparations which are known to contain some of this antibody.
Due to lack of availability of both anti-\textit{E. coli} antibody and 2E2.1 culture supernatant, as well as the small amount of \textit{P. falciparum} protein, the \textit{P. falciparum} 120 kDa polypeptide was not immunoprecipitated for end-terminal sequencing. Accordingly amino acid analysis is still required to verify that this polypeptide is encoded by the \textit{rpoB} gene.

### 4.2.4 Fusion protein expression

Due to the relative lack of success in developing anti-peptide antibodies to the \textit{P. falciparum} \textit{B} subunit it was decided that antibodies would be raised to fusion proteins generated by expression of the \textit{rpoB} gene. The problems encountered with this technique are varied, ranging from no or minimal expression of the synthesised construct, to degradation of the expressed protein or its insolubility in an aqueous environment. The system chosen for expression also must take into account whether the protein of interest is post-translationally modified. In the case of the \textit{B} subunit, there are no such modifications and an \textit{E. coli} expression system was chosen.

Several problems were encountered with expression of the \textit{P. falciparum} gene and many constructs using different systems were tested before one system generated a fusion protein of the expected size. The expression systems tried included a trpE system using a pATH vector (Koerner \textit{et al.}, 1990) and a system which produces an N-terminal fused, maltose fusion protein (pMAL)\textsubscript{(NEB}). The fusion protein generated with the PATH system produced a protein of the predicted molecular mass but at a very low yield; the pMAL fusion protein, however, degraded on purification. Different regions of the \textit{rpoB} gene were used each time. The expression system
finally used - pxy460, generates an N-terminal fused β-gal fusion protein (Holder et al., 1987). A pxy460 construct was generated which included the final 286 amino acids of rpoB as described in Materials and Methods.

Upon induction of the construct pxy/9, with IPTG, a 170 kDa polypeptide was expressed. This could not be visualised by staining SDS-PAGE gels with Coomassie Brilliant Blue R-250 but was readily detected by immunoblotting with an anti-β-gal fusion antibody (kindly provided by I. Ling, NIMR). In order to achieve maximum expression, inductions were performed by 1) altering the IPTG induction concentration and 2) varying the time course of induction and detection by Western blotting techniques (see Figure 4.6). A pxy461 clone which upon IPTG induction gives a 118 kDa β-galactosidase product (provided by I. Ling) was induced for 3 hours with 0.3 mM IPTG as a positive control (Figure 4.6a). Maximum expression was achieved at 0.3 mM IPTG, inducing for 2.5 hours at 37 °C - Figure 4.6b. The fusion protein indicated by an arrow in Figure 4.7, was affinity selected over an ABTG agarose column. This resulted in only partial purification since, SDS-PAGE analysis on a 7.5% polyacrylamide gel and subsequent staining with Coomassie Brilliant Blue R-250, indicated that other polypeptides were able to bind ABTG were co-purified with the fusion protein (Figure 4.7 b). Immunoblots of fusion protein eluted from the affinity column and purified β-galactosidase (a kind gift from Dr. M. Blackman, Department of Parasitology, NIMR)(Figure 4.7a.), indicated that the faster migrating polypeptides were recognised by an anti-β-galactosidase antibody suggesting that these were fusion protein degradation products. The higher molecular weight bands migrating at approximately 200 kDa are presumed to be polypeptides
which bind non-specifically to the resin used to select for the fusion protein. This being the case, I reasoned that the fraction should not be further purified since antibodies to more epitopes could be generated from this fraction. Fusion protein was used to immunise each of two rabbits, each immunisation containing approximately 50 µg protein. Protein concentration was estimated by UV spectrophotometry at λ_{280}nm. The pre-immune sera from both rabbits were used to probe immuno-blots of purified *E. coli* RNAP. Rabbit pre-immune serum 71 recognised the *E. coli* β and β' subunits and was not further analysed. Serum from rabbit 72 (Rb72) was first checked in Western blots to ascertain whether antibodies within the serum could recognise the *E. coli* β subunit and fusion protein to which it was raised. Figure 4.8 shows weak reactivity of the pre-immune serum to the *E. coli* subunits and no reactivity to the fusion protein. An increased reactivity to the *E. coli* β/β' subunits was observed by the immune serum 72, in addition to reactivity to the pxy/9 fusion protein. The anti-β-galactosidase antibody showed cross-reactivity with the pxy/9 fusion protein only.

Using serum Rb72, immunoblots of early stage parasites (rings and early trophozoites), late stage parasites (late stage trophozoites and schizonts), *E. coli* RNAP, *E. coli* whole cell lysate, and fusion protein were probed with pre-immune serum, serum after the second immunisation (Rb72/June) and this same serum pre-absorbed with *E. coli* whole cell lysate protein extract. The results are summarised in Figure 4.9. The pre-immune serum did not recognise any polypeptides on this Western blot (data not shown). Both the Rb72 antibody and absorbed Rb72 antibody recognised a *P. falciparum* polypeptide (Figure 4.9) migrating at the molecular weight corresponding to that expected for the β subunit (122 kDa). No discrete *E. coli* polypeptide was recognised in the whole cell lysate. This indicates that although
antibody Rb72 recognised the *P. falciparum* β subunit it did not recognise the *E. coli* β subunit. The Western blot presented is one of a succession of immunoblots performed with this antibody. Unfortunately this antibody and the data provided by it were produced at the end of the project and further necessary experiments are yet to be performed - see discussion below.

### 4.2.5 Immunofluorescence studies

In order to determine the location of the plastid-RNAP within the parasite, smears of asynchronous parasites were incubated with either the polyclonal anti-*E. coli* antibody, the monoclonal antibody 2E2.1 culture supernatant, or Rb72(June) serum and visualised under immunofluorescence. In addition, the same smears were stained with DAPI in order to visualise the parasite nucleus under UV light. All three antibodies gave similar signals, with a stronger signal visualised from late trophozoites and schizonts. Fluorescent staining using 2E2.1 culture supernatant and the anti-*E. coli* antibody is presented in Figure 4.10. The signals produced were weak and the fluorescence faded rapidly. In order to determine whether this was due to a problem with either the probes or the smears, parallel, control experiments were performed. Parasite smears probed with an anti-MSPl antibody recognising only schizont parasites, and *E. coli* smears probed with the anti-*EcB*, both produced highly fluorescing signals (data not shown). The Rb72 serum produced a high background and the immunofluorescence pattern produced using this antibody have not been shown.
Schizonts binding either 2E2.1 or the anti-E. coli RNAP antibody probes produced similar results: The schizonts have a granular appearance which coincided with the individual, internalised merozoite nuclei, as determined by DAPI staining. Western blot data have indicated that the β subunit is more abundant in late stage parasites, but is also present in early stage parasites. An inability to observe ring/early trophozoites under immunofluorescence may be attributed to either a poor fluorescent signal or inaccessibility of antibody to antigen in early forms of the parasite. As mentioned in the introduction, the 35 kb genome may be located in the spherical body. Under EM, this vesicle is a multi-membranous organelle. Whether permeability of these membranes changes during maturation of the blood stage parasite is yet to be determined but may be responsible for antibody staining only of late stage parasites. However, the most probable reason for low fluorescence is the low level of rpo gene expression.

Other techniques, including in situ PCR hybridisation, are currently being used to localise the 35 kb molecule within the parasite (A.Roy, personal communication).

4.3.0 Summary and future work

The β subunit of the proposed multimeric P. falciparum RNAP was identified in Western blots using antibodies raised to both predicted oligonucleotides and rpoB fusion proteins.

The anti-E. coli anti-β and anti-β’ antibodies recognised a P. falciparum 120 kDa polypeptide presumed to be the plastid RNAP β subunit. The affinity selected anti-α
and anti-σ antibodies, did not recognise any \textit{P. falciparum} polypeptides, inferring that these epitopes are not shared. Lerbs and colleagues (1985) have demonstrated cross-immunoreactivity between the \textit{E. coli} β and β′ subunits and the β, β′ and β″ subunits of the spinach chloroplast multimeric RNAP using an antibody-linked polymerase assay (ALPA). This system involves incubating nitrocellulose bound SDS-PAGE electrophoresed RNAP, with an excess of polyclonal antibody raised to the same or homologous RNAP. A second binding reaction with native RNAP involves attachment of this enzyme with the unbound antibody FAB sites at the corresponding membrane-bound polypeptides. The resultant immunoglobulin complexed RNAP is then incubated with a transcription mixture containing $[\alpha^{-32}\text{P}]\text{dUTP}$. The nascent radio-labelled RNA transcript, once TCA precipitated may be visualised by autoradiography. Assuming the \textit{P. falciparum} plastome RNAP molecule could be partially purified, similar protocols would allow the independent polypeptides involved in transcription to be fully characterised.

It will be important in future work to immunoprecipitate of the malarial β subunit and determine its amino acid sequence in order to positively assign the \textit{rpoB} gene to the 120 kDa polypeptide recognised by the anti-\textit{E. coli} RNAP antibody and the 2E2.1 culture supernatant.
Figure 4.1

Illustration of the positions of IW376 and IW227 in the 35 kb encoded \textit{rpoB} gene.

Diagram showing the position and amino acid sequences (in green) of polypeptides IW376 and IW227 within the \textit{P. falciparum rpoB} gene.

The corresponding \textit{E. coli rpoB} amino acid sequence has been included to show the degree of conservation between each of the two polypeptides and \textit{E. coli}. 
Figure 4.2

The identification of the *P. falciparum* β RNAP polypeptide

Western blot of *Plasmodium falciparum* (*P.f*), and *E. coli* (*E.c*) whole cell lysates recognised by the 2E2.1 culture supernatant.

Arrow indicates the position of the *P. falciparum* signal corresponding to the β subunit.
Figure 4.3

The identification of \textit{P. falciparum} polypeptides which have shared antigenicity with the anti-\textit{EcB} antibody.

Western blot of \textit{Plasmodium falciparum} (\textit{P.f}), and \textit{E. coli} (\textit{E.c}) whole cell lysates recognised by a) the anti-\textit{E. coli} RNAP polyclonal antibody; b) anti-\textit{EcB} pre-absorbed with \textit{E. coli} RNAP.

The arrow indicates the \textit{P. falciparum} polypeptide migrating at that predicted for the \(\beta\) subunit which is no longer observed in panel b).

Asterisk indicates the new \textit{P. falciparum} polypeptide recognised by \textit{EcB} after pre-absorption with \textit{E. coli} RNAP.
Figure 4.4

The identification of *P. falciparum* polypeptides which have shared antigenicity with the anti-EcB antibody and anti-*E. coli* β/β' antibodies.

a) SDS-PAGE *E. coli* RNAP stained with Coomassie Brilliant Blue R-250. The polypeptides corresponding to the α, β/β' and σ subunits are indicated.

b and c) Identical Western blots of early, late and asynchronous (Asyn) stages of *P. falciparum* whole cell lysates, *E. coli* whole cell lysate and pure RNAP recognised by the polyclonal anti-*E. coli* antibody (part b.) and, affinity selected anti-*E. coli* β/β' antibodies (part c.).
Anti-\textit{E.coli} RNAP antibody

Anti-\textit{E.coli} $\beta\beta'$ subunit antibodies
Figure 4.5

The identification of *P. falciparum* polypeptides which have shared antigenicity with anti-*E. coli* σ and α antibodies.

Western blot of early, late and asynchronous (Asyn) stages of *P. falciparum* whole cell lysates, *E. coli* whole cell lysate and pure RNAP recognised by the affinity selected:-

a) anti-*E. coli* σ-antibodies

b) anti-*E. coli* α-antibodies
a) Anti-\textit{E. coli} \( \sigma \) subunit antibodies

b) Anti-\textit{E. coli} \( \alpha \) subunit antibodies
Figure 4.6

Identification of the pxy/9 fusion product.

a) Western blot of pxy/9 construct induced by varying concentrations of IPTG (5 mM, 1 mM, 0.3 mM, 0.1 mM, 0.03 mM), and pxy460 (Pxy) control construct recognised by an anti-β-galactosidase antibody.

Arrow indicates the fusion product recognised in all lanes except the pxy lane.

b) Western blot of pxy/9 induced at 0.3 mM IPTG for an increasing time (hours) recognised by an anti-β-galactosidase antibody.
a) IPTG Concentration (mM)

<table>
<thead>
<tr>
<th>Concentration (mM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0 1.0 0.3 0.1 0.03</td>
</tr>
</tbody>
</table>

b) Time in hours

<table>
<thead>
<tr>
<th>Time in hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

kDa

<table>
<thead>
<tr>
<th>Size (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>215 105 69.8</td>
</tr>
</tbody>
</table>
Figure 4.7

Purification of the pxy/9 fusion product.

a) Western blot of the elution profile obtained upon purification of pxy/9 using an anti-β galactosidase antibody. The pxy/9 fusion product is indicated by an arrow.

b) Coomassie Brilliant Blue R-250 stained SDS-PAGE gel of the purified pxy/9 fusion product.
Figure 4.8

Three identical Western blots of pxy/9 fusion product (Fusion protein) and *E. coli* RNAP recognised by:

**Pre**- the pre-immune serum of rabbit 72

**72**- the first post-immunisation polyclonal serum from rabbit 72.

**anti-β gal**- the anti-β galactosidase antibody

Arrows indicate the polypeptide corresponding to the pxy/9 fusion product.
Figure 4.9
Detection of the 35 kb encoded β subunit by Western blot.

Identical Western blots of early and late stage *P. falciparum* parasite whole cell lysate; *E. coli* whole cell lysate; purified pxy/9 fusion product (fr. 7) recognised by:-

a) rabbit 72 serum

b) rabbit 72 serum purified over *E. coli* whole cell lysate.
Figure 4.10

Detection of the 35 kb encoded RNAP by immunofluorescence.

Immunofluorescence of *P. falciparum* parasites recognised by anti-EcB (A and C) and 2E2.1 culture supernatant (E and G).

Corresponding DAPI stained parasites (B, D, F, and H)

The arrow in A and B points to the pattern of the same parasite under both immunofluorescence and DAPI.
CHAPTER 5
CHAPTER 5

THE SEQUENCE AND ANALYSIS OF THE 35 kb ENCODED rps 2 GENE

5.1.0 Introduction

The gene encoding the small ribosomal protein rps 2 is the final gene encoded by the IRₐ arm of the 35 kb circle and immediately follows the RNA polymerase subunit rpoC₁/C₂. This polypeptide has been shown to associate with the 30S ribosomal subunit (Harris et al., 1994) and may be involved in the association between the ribosome and tRNAs. The rps 2 polypeptide is homologous to the 68 kDa S. cerevisiae laminin receptor (Davis et al., 1992), which is involved in the adhesion of certain cell types to the extracellular matrix (Akiyama et al., 1990). An homologous polypeptide has also been identified in the S9 locus of the archeon organism Haloarcula marismortui (Ouzouis et al., 1995).

The rps 2 gene is located at 4' on the E. coli genome and is transcribed in an operon (the s2 operon) with the tsf gene which encodes the EF-Ts protein involved in protein translation - see Figure 5.1 (Zengel and Lindahl, 1994). Many of the proteins involved in ribosome activity have been demonstrated to be under autogenous control in order to regulate not only protein translation but also ribosome formation. In addition to the three rRNA molecules that constitute the ribosome, up to fifty proteins associate with these RNAs to form the functional ribosome complex. Up to 50% of the dry mass of a rapidly growing bacterium may be attributed to the ribosome and control of this ribosome complex is essential due to the extreme cellular energy required for its synthesis. There is little turnover of ribosomal constituents and the
control of ribosomal protein (r-protein) synthesis and rRNA transcription must be tightly regulated by the organism.

Most r-proteins bind naked RNA in vitro and are associated, therefore with the ribosome's initial assembly (Zengel et al., 1980; Nomura et al., 1980). The ability of these proteins to bind target regions on RNA is thought to be a feature of autogenous control since these proteins are also likely to be able to recognise specific sites in their own mRNA and thereby regulate RNA production (Nomura et al., 1980). This feature assumes that r-protein synthesis is related to the rate of RNA transcription (Zengel et al., 1980). An exception to this rule is apparent with the S2 operon since the rps 2 protein does not bind RNA yet it is under autogenous control (An et al., 1981).

Similar to the eubacterial protein expression system, chloroplasts consist of two ribosomal subunits, with between 22 to 31 r-proteins in the small subunit and between 32 to 36 r-proteins in the large subunit (Capel et al., 1982; Eneas et al., 1981; Schmidt et al., 1983). Methods of controlling expression of r-proteins in plastid genomes is complicated by the fact that the location of the r-protein genes is not restricted to the plastome but has been divided between the plastid genome and the nucleus (Harris et al., 1994). It has been assumed that after the primary endosymbiotic event, plastid formation involved partial transfer of ribosomal protein genes from the bacterium, a cyanobacterium, to the nucleus of the eukaryote host organism. This means that a system has evolved which co-ordinately regulates expression of r-proteins in both the plastid and host cell nucleus although the mechanism by which this occurs has not been elucidated.
The nucleotide sequence and predicted amino acid sequence of the *P. falciparum* plastid rps 2 homologue is presented in this chapter.

5.2.0 Results and Discussion

5.2.1 Analysis of the rps 2 gene

The rps 2 gene sequence encoded on the IR\textsubscript{A} arm of the 35 kb molecule encodes a 233 amino acid polypeptide (Figure 5.2).

The predicted amino acid sequence of the rps 2 gene has a low level of identity with homologues in *E. coli* and the *Euglena* chloroplast at 18% and 23% respectively. A PILEUP analysis of the amino acid sequences in Figure 5.3 indicates the poor conservation between the 35 kb encoded rps 2 gene and other chloroplast homologues (which share greater identity to each other). The principle point of interest of this gene is not its nucleotide or amino acid composition but its location within the 35 kb molecule.

5.2.2 Location of the rps 2 gene within plastid genomes

As in land plant chloroplasts, the rps 2 gene is located in the malarial plastid immediately downstream to the *rpo* B/C\textsubscript{1}/C\textsubscript{2} operon. Land plant chloroplasts, however, additionally encode a reduced form of the atp operon with which the rps 2 gene has been fused. Figure 5.1 shows a representation of the organisation of the *rpo*, atp and s2 operons in bacteria, higher-land plants, red algae, cyanobacteria and *P. falciparum*. It appears that the *Plasmodium* 35 kb plastid has an unusual gene arrangement since this molecule does not encode any of the genes comprising the atp
operon and is in this case dissimilar to other known plastid genomes. The gene order in the red algal *P. purpurea* plastid genome is of the type *rpoB/C1/C2-rps2-tsf-atpA/D/F/G/H/I* differing from the cryptomonad *C. paradoxa*, higher land plant chloroplasts, the plastid in the chromophytic algae *C. reinhardtii* and the 35 kb DNA in *P. falciparum* where the *tsf* gene is presumably nucleus encoded. The organisation differs slightly in *C. reinhardtii* where the *atp* operon has been disrupted and in the plastid in *Euglena gracilis* where the *rps2* gene and the *atp/I/H/F/A* genes have been translocated to a region encoding the ribosomal proteins (Hallick *et al.*, 1993). The *rps2* gene in this instance contains three group III introns in comparison to the 2 group III introns identified in the *rps2* gene encoded by the plastid of *Astasia longa* (Semeister *et al.*, 1990). Group III introns are believed to be specific to the plastids in *Euglena* and *Astasia* and may be evolutionarily related to group II introns (Christopher *et al.*, 1988; Christopher and Hallick 1989).

It has been suggested by Reith (1993) that the red algal *P. purpurea* gene organisation of these operons is ancestral to that found in chloroplasts where part of the *atp* operon and the *tsf* genes were later transferred to the nucleus. If this is the case, one must assume that the *P. falciparum* plastid gene order diverged after this gene arrangement had been established and the remaining *atp* genes were either transferred to the nucleus or lost. More information concerning the gene content and gene order in dinoflagellate plastids needs to be collected to determine whether these organisms have a gene organisation similar to that observed in the 35 kb molecule.
5.2.3 Transcription of the rps 2 gene

No transcription analysis has been performed on the 35 kb-encoded rps 2 gene and it is therefore unknown if this gene is transcribed as part of the upstream rpo operon. No sequences capable of forming hairpin structures have been identified downstream to the TAA stop codon, such features are commonly used in chloroplast genomes to terminate transcription (Hudson et al., 1988). It is unknown how transcription is terminated in the 35 kb molecule. However, by analogy to chloroplasts I would have expected to find a secondary structure feature such as a hairpin loop at the end of the rps 2 gene especially as this is the last coding sequence on the coding strand of the IR\textsubscript{A} arm.

5.3.0 Summary and future work

An rps 2 gene has been identified on the 35 kb molecule and is located 12 nucleotides downstream of the stop codon of the rpoC\textsubscript{2} gene. The typical plastid gene arrangement has been maintained, although the atp genes located immediately downstream to the rps 2 gene in other plastid DNAs has not been retained in the 35 kb DNA and have been presumably transferred to the nucleus or lost.

Future experiments should include the determination of rps 2 transcription and whether the rps 2 gene is co-transcribed with the rpo operon.
Figure 5.1

A comparison of the *rpoB/C*, *rps 2 (s2)*, *tsf*, and *atp* gene order in various plastid genomes.

Spaces between genes indicates that these coding regions are physically separated, genes drawn without any space indicates that they are located adjacent to each other. The question marks denotes those genes whose positions have not been identified.

This figure has been modified from Reith and Munholland (1993).
**E. coli**

- rpoB
- rpoC
- s2
- tsf

**Anabaena sp**

- rpoB
- rpoC1
- rpoC2
- s2
- tsf

**P. purpurea**

- rpoB
- rpoC1
- rpoC2
- s2
- tsf

**C. paradoxa**

- rpoB
- rpoC1
- rpoC2
- s2
- H
- G
- F
- D
- A

**Land plants**

- rpoB
- rpoC1
- rpoC2
- s2
- I
- H
- F
- A

**P. falciparum**

- rpoB
- rpoC1
- rpoC2
- s2

**C. reinhardtii**

- I
- F
- rpoB2
- rpoB1
- rpoC2
- A'
- H
- s2
Figure 5.2

Nucleotide and predicted amino acid sequence of the *P. falciparum* rps 2 gene.

Numbers indicate nucleotide positions.

Asterisk denotes stop codon.
ATG TTT ATT ACT TTT GAT AAT TTA TTA AAA TCA AAA
  M F I T F D N L L K S K

ATT TAT ATA GGA AAT ATT TAT AAG AAT ATT TAT TTT
  I Y I G N I Y K N I Y F

GAT AAT TAT AAA TAT ATA TAT AAA ATA AAA TTT AAT
  D N Y K Y I Y K I K F N

TAT TGT ATT TTA AAT TTT ACA TTT ATA GCT TTA TAT
  Y C I L N F T F I A L Y

TTA TAT AAG TTA TAT TTA TAT ATT TAT AAT ATA TCA
  L Y K L Y L Y I Y N I S

TTA ATG AAT AAT AAA ATT TTA TTT ATT AAT AAT AAT
  L M N N K I L F I N N N

AAT TTA ATA AAA AAT TTT ACA ATT AAA ATA TGT AAT
  N L I K N F T I K I C N

TTA ACA AAT AAT TTA TAT ATA AAT AAA TGG GTT TCT
  L T N N L Y I N K W V S

GGA TTA TTA ACT AAT TGG TTT GTA TTA AAA AAA AAA
  G L L T N W F V L K K K

ATT ATA ATA TAT ATT TGG ATA AGT AAA ATA TTA AAA
  I I I Y I W I S K I I K

AAT AAA TAT TTT AAT AAT ATA TTA TCT AAA AAA TGT
  N K Y F N N I L S K K C

ATA TAT AAT TTA AAT ATA ATT TAT ATT AAA TTA TAT
  I Y N L N I I Y I K L Y

AAT AAA TTT AAT GGT ATA AAA AAT ATG ATA AAT CTA
  N K F N G I K N M I N L

CCT AAA TAC ATA TTT TTA ACA AAT TTT AAT AAA AAT
  P K Y I F L T N F N K N

TTA ATT TTA AAA GAA ATT TTA AAA TTA AAA TTA ATT
  L I L K E I L K L K L I

TTA ATA AGT TTT ATA AAT TTA AGT TTA GAT TCA AGT
  L I S F I N L S L D S S
577  AAT ATA AAT ATA AAA ATT TTA GGA AAT TAT AAT AAT
    N I N I K I L G N Y N N
613  TAT AAA TCT TTA AAA TTA ATA TAT AAA ATA ATT TAT
    Y K S L K L I Y K I I Y
639  ACT TCA TTA ATT CAT AGT AAA ATT AAA AAT ATG TAA
    T S L I H S K I K N M *
Figure 5.3

PILEUP analysis of the 35 kb encoded rps 2 gene.

Amino acids identical to the *Plasmodium* sequences are shaded in yellow. Deletions inserted for optimal alignment are represented by dots.

*P. falciparum* (Accession number: x95275); *Epifagus virginiana* (P27068); *Nicotiana tabacum* (P06355); *Escherichia coli* (P02351); *Euglena gracilis* (P30389).
CHAPTER 6
CHAPTER 6

DESCRIPTION OF THE clpC-LIKE MOLECULE ENCODED BY THE
35 kb DNA.

6.1.0 Introduction

The sequence of the 4 kb amplification product described in section 3 filled in the last major unsequenced portion of the 35 kb circular molecule. This chapter and the one that follows completes the description of the genes encoded on the single copy IR_B arm. The PCR product also includes the point of cross over of transcription from the IR_B arm to the complimentary strand of IR_A.

The region of the amplification product encoding genes on the IR_B arm includes the sequence encoding the final 205 amino acids of the C-terminal region of a clp gene previously identified by K. Rangachari in this laboratory. Immediately downstream of the clp gene, two tRNAs, tRNA^{Gly} and tRNA^{Ser} were identified using the Staden algorithm and the Staden-Plus software and will be discussed in Chapter 7.

6.1.1 The clpC gene - an introduction

The clp family of proteins is both ubiquitous and diverse (Squires and Squires 1992). All members of the Clp family (except ClpA) are heat shock proteins (hsp) and are transcribed by the hsp σ^{32} RNAP subunit (Squires and Squires 1992). Hsps are involved in a multiplicity of functions, assisting both in normal growth and stress conditions, including molecular chaperoning. The majority of clp proteins are high molecular mass proteins of approximately 100,000 Mr, and contain two unrelated
nucleotide triphosphate (NTP) binding domains, N1 and N2, each of which is comprised of two conserved nucleotide binding motifs regions A and B encompassed by a leader and trailer region of varying lengths (Squires and Squires 1992) - see Figure 6.1. The N1 domain also contains a conserved B2 region of unknown function. There is a minimum of three sub-families in this group; clpA, -B and -C.

ClpA has been shown to function like DnaK and DnaJ as a molecular chaperone targeting abnormal proteins for degradation by the proteolytic subunit clpP. It has also been shown to protect luciferase from heat inactivation (Wickner et al., 1994). In *E. coli*, clpA associates with the non-homologous, heat-inducible protease, ClpP to form an ATP-dependent serine protease (protease Ti), (Katayama-Fujimura et al., 1987; Hwang et al., 1987 and Maurizi et al., 1990). Protease Ti is responsible for the proteolysis of polypeptides with abnormal amino termini, in a substrate dependent manner (Tobias et al., 1991). The clpP gene is encoded in the chloroplasts of higher plants and has been maintained in the vestigial genome of the plastid of *Epifagus virginiana* (Wolfe et al., 1992). This gene is one of four not involved in the translational apparatus of this plastid and on these grounds Wolfe proposed that the retention of this gene may explicate the maintenance of the vestigial plastid in the organism. In *E. coli*, ClpP is additionally regulated by clpX, which contains a single ATP-binding domain and with which it may be co-transcribed (Gottesman et al., 1993; Wojtkowiak et al 1993). Substrate specificity of the clpP protease is altered by association with either clpA or clpX. This complex contains twelve molecules of clpP combined with a hexameric ring of either clpA (reviewed by Maurizi et al., 1992; Thompson and Maurizi, 1994) or clpX. The clpP-A protease is able to degrade α-
casein and partially degrade lambda-O whereas the clpP-X complex is unable to
degrade \( \alpha \)-casein and fully degrades lambda-O (Wawrzynow et al., 1995). It would
appear that clpA and clpX act as substrate specifiers for proteolytic activity associated
with clpP. Whether substrate specificity of this proteosome may be further altered by
combination with other clp members is yet to be determined. ClpX has also been
demonstrated to perform molecular chaperone activities independent of clpP namely
dis-aggregating aggregates of lambda O with hydrolysis of ATP (Wawrzynow et al.,
1995). It has also been reported to disassemble the Mu transposase tetramer
(Levchenko et al., 1995).

ClpB has been identified in a variety of organisms including \textit{E. coli} (Squires et al.,
1991) and the parasite \textit{Leishmania} spp. (Hubel et al., 1995). ClpB corresponds to hsp
104 in \textit{S. cerevisiae}. No proteolytic activity in association with clpP has been
identified. However, hsp 104 has been shown to function as a molecular chaperone,
re-solubilising polypeptides which have aggregated during stress conditions rather
than participating in their proteolysis (Parsell et al., 1994a). Both nucleotide binding
sites in hsp 104 have been shown to be essential for thermotolerance (Parsell et al.,
1991). The function of clpB in \textit{Leishmania} is thought to be essential for parasite
viability and an increase in clpB mRNA has been observed in heat stressed
\textit{Leishmania} promastigotes (Brandau et al., 1995; Hubel et al., 1995).

ClpC has been identified in the gram positive bacteria \textit{Mycobacterium leprae} (Nath
and Laal., 1990); \textit{Streptococcus pneumoniae} (Pearce et al., 1993); \textit{Lactococcus lactis}
(Huang et al., 1993); \textit{Bacillus subtilis} (Msadek et al., 1994) as well as in the higher
plant *Lycopersicon esculentum* (tomato) (Squires and Squires 1992); and the red alga *Porphyra purpurea* (Reith-personal communication). In accordance with the chaperone activities attributed to clpB, clpC has been identified by Msadek (1994) as a pleitropic regulator of gene expression and growth at high temperature. Tomato ClpC is nucleus encoded and is imported into the chloroplast by means of a leader sequence which is subsequently cleaved (Squires and Squires., 1992). By contrast, the *P. purpurea* clpC gene is still plastid encoded and the function of this polypeptide within the red algal plastid or within chloroplasts in general can only be postulated. The presence of a clpP-clpC protease has not been demonstrated although degradation of ribulose 1,5-bisphosphate carboxylase in chloroplasts has been observed and could be due to a clpP proteosome (Gottesman *et al.*, 1990).

### 6.2.0 Results and Discussion

#### 6.2.1 The *P. falciparum* clpC-like gene

A clone encoding the 5' terminus of the 35 kb encoded clp gene was isolated from a 35 kb genomic library by K. Rangachari. In order to confirm that this gene originated from the 35 kb molecule, the recombinant clone was radio-labelled with [α-32P] and hybridised to a Southern blot of both *P. falciparum* genomic DNA and 35 kb DNA at high stringency conditions. Use of a unique *HaeIII* endonuclease restriction site at the 5' terminus derived from the sequence also verified the location of this clp gene on the 35 kb circle (data not shown). The nucleotide data which I subsequently obtained upon sequencing the 4 kb PCR product described in section 4 completed the 3' sequence of this gene. A PCR product (made with primers KR2 and RPOD-3) corresponding to a region of the N2 domain region of the clp gene was radiolabelled
with $\alpha^{32}\text{P}$ and hybridised to endonuclease restricted \textit{P. falciparum} total genomic DNA in a Southern blot. Hybridisation was performed at low stringency conditions and washed at both low (42°C) and high (68°C) stringency temperatures at 0.1X SSPE. Figure 6.2. At low stringency wash conditions (panel b), weak cross hybridisation signals with other DNA species, in addition to the expected primary signal, were observed suggesting that other clp homologues may be encoded by the nucleus. These signals were absent after washing at high stringency conditions (panel a). The signals observed at high stringency wash conditions relate to the signal expected from the nucleotide sequence available from the 35 kb circle.

The complete sequence of the \textit{P. falciparum} clp homologue encodes 765 amino acids, 92,000 M\(_r\) (Figure 6.3). The amino-terminus of this polypeptide is poorly conserved by comparison with other clp genes - Figure 6.4 At the amino acid level the 35 kb encoded clp gene shares more homology with clpC of \textit{L. esculentum}, \textit{Pisum sativum} (pea) and \textit{P. purpurea} (red algae) than with clpA or clpB in \textit{E. coli}. Members of the clpC family contain a conserved amino terminal leucine-rich repeat (overlined in Figure 6.4 which may be discernible in the \textit{P. falciparum} homologue as a short amino acid repeat consisting of leucine residues (asterisked in Figure 6.4. Homology at the first ATP binding box (N1) in the \textit{P. falciparum} homologue appears to be degenerate with low conservation of the consensus sequence Y-X\(_8\)-T-X\(_{13}\)-Y (Gottesman et al., 1990) preceding the B2 region. The spacer region (amino acids 511-635 in Figure 6.4 in the \textit{P. falciparum} clp is less than that of clpB proteins yet has little homology with that of clpC. However, conservation within the N2 box increases to approximately 60% identity with pea, with both the A and B sites in \textit{P. falciparum} being reasonably
well conserved. A lysine at amino acid 722 (Figure 6.4) is conserved in *P. falciparum* and has been demonstrated to be necessary for hexamer formation in hsp 104 (Parsell *et al.*, 1994). Hexamer formation is necessary for clpA function in both chaperone and clpP-A proteolytic activity and occurs in the presence of ATP (Wickner *et al.*, 1994).

Amino acid conservation between the malarial clpC and other versions is maintained towards the C-terminus, small regions of homology (between amino acids 881 - 952, Figure 6.4) being conserved amongst *P. falciparum* and all other identified clp polypeptides (Gottesman *et al.*, 1993). The function of this region of clp genes has not yet been determined, although it is likely that its conservation indicates a common function of clp proteins.

The *Plasmodium* clp gene has obviously diverged from the clp genes identified to date. Only the second of two ATP-binding domains has been conserved and is therefore assumed to be functional. However, since no biochemical studies have been performed on the ATPase activity of this polypeptide, it is possible that although the N1 domain appears to be degenerate it may have retained the ability to hydrolyse ATP.

It appears that at the amino acid level this gene is a degenerate form of clpC and may have a specialised function. Other single nucleotide domain clp proteins (i.e. clpX and clpY in *E. coli* and amiB in *Pseudomonas aeruginosa* (Wilson *et al.*, 1995)) are approximately 45 kDa - half the molecular weight of the di-nucleotide binding domain polypeptides clpA, clpB and clpC. The amiB protein is encoded by the amidase
operon and together with the amiS gene product is thought to form part of an ABC$_1$ transporter system. Here the amiB gene product represents a cytoplasmic ATPase which functions in conjunction with the integral membrane component amiS. ABC$_1$ transporters are involved in the transport of substrates across membranes accompanied by ATP hydrolysis. Functional ABC$_1$ transporters comprise a C-terminal ATP-binding domain of approximately 200 amino acids located on the cytoplasmic face of the membrane and an N-terminal hydrophobic integral membrane domain (usually consisting of six membrane spanning regions) which is presumed to recognise and translocate specific substrates across the cytoplasmic membrane. These two domains are often encoded as two distinct polypeptides in prokaryotes and a single polypeptide in eukaryotes (Hyde et al., 1990; reviewed by Doige and Ames 1993; Higgins et al., 1992). Two of each domain type are required for functional activity i.e. a functional ABC$_1$ transporter consists of 12 membrane spanning regions and two ATP-binding domains. The energy required for the transportation of substrates across the membrane is probably provided by ATP hydrolysis (Bishop et al. 1989; Mimmack et al., 1989), although it has also been suggested that the availability of ATP may regulate the activity of the transporter (Doige and Ames, 1993). Deletion of the amiB gene product had no affect on inducible amidase expression although the gene product has intrinsic ATPase activity. AmiB is another example of a clp homologue with an activity presumed to be independent of clpP associated proteolysis. By analogy with amiB, the divergent clpC gene encoded by the 35 kb circle might possibly contribute towards the import of nucleus encoded polypeptides into the plastome in a manner similar to that of ABC$_1$ transporters. However, no polypeptides have been identified on the $P$. 149
*falciparum* plastid that may act as trans-membrane proteins, although it may be that these components are nucleus encoded.

The import of nucleus encoded polypeptides into chloroplasts is mediated by amino-terminus transit peptides (reviewed by Theg and Scott, 1993) and in a situation analogous to the role of hsp 70 in mitochondria (Gambill *et al.*, 1993), it has been suggested that part of the import process in chloroplasts requires hsp 70 (Wu *et al.*, 1994). A mitochondrial clp homologue - hsp 78, is associated with membrane translocation when hsp 70 levels are limiting (Schmitt *et al.*, 1995). It has been proposed that hsp 70 is involved in ensuring unidirectional transit of polypeptides into mitochondria (Cyr *et al.*, 1993; Ungermann *et al.*, 1994; Schneider *et al.*, 1994), at the same time stabilising the polypeptide (Sanchez *et al.*, 1993), in an analogous role to that suggested for hsp 104 (Parsell *et al.*, 1994). In situations where hsp 70 is limiting, hsp 78 has been suggested to inhibit retrograde polypeptide movement but is unable to bind the polypeptide to prevent aggregation (Schmitt *et al.*, 1995). In an analogous situation, therefore, it is possible that the *P. falciparum* clpC polypeptide is involved in the import of nucleus encoded polypeptides into the putative plastome.

Another possible role for this gene product could include protecting polypeptides from heat inactivation. During the life cycle the parasite is subjected to multiple heat stresses, including elevated temperatures during fever in the mammalian host and cold shock in the vector host. The clpC gene product would understandably play an important chaperone activity in accordance with those activities like hsp 104 in yeast and clpB in *Leishmania*. 

150
The *E. coli* heat shock response protein clpB has been shown to be regulated by $\sigma^{32}$ (Squires *et al.*, 1991). No $\sigma^{32}$ heat shock promoter elements upstream to the initiator methionine codon of *P. falciparum* clpC have been identified. These regions are typically located within the -35 and -10 promoter regions immediately upstream to the start site of transcription and are of the type (-35)CCCTTGA(N)8,14TCCCCAT. The lack of such a sequence does not necessarily imply that the malarial clpC polypeptide is not regulated in response to heat shock. It may be that transcription of the 35 kb molecule as a whole is increased either in heat stress or in other stages of the parasite life cycle that haven’t been analysed as yet. The erythrocytic stages of *P. falciparum* are easily cultured *in vitro*, however no system has been developed to culture the sporozoite or gamete forms of the parasite. The ability to assess these other stages would provide a more complete analysis of expression of the 35 kb DNA.

### 6.2.2 clpC gene expression

Both Northern blot analysis and primer extension techniques failed to identify a specific *P. falciparum* clp transcript (data not shown) although RNase protection assays have confirmed clp gene expression (Preiser, unpublished data). Possible reasons for the inability to detect certain transcripts from the 35 kb molecule will be discussed in Chapter 9.

### 6.3.0 Summary and future work

The portion of the 35 kb sequence discussed in this chapter encodes a clp gene with homology to clp proteins containing a di-NTP binding domain. However, the predicted 35 kb encoded polypeptide consists of only a single NTP binding domain.
Further characterisation of the clp gene product would be desirable since the presence of this polypeptide in the organelle embodying the 35 kb molecule is probably associated with one of its housekeeping functions within the *Plasmodium* parasite. Initial experiments would involve determination of whether the clpC-like transcript is elevated in stress conditions in either the erythrocytic or sporozoite forms of the parasite. In *Leishmania* promastigotes clpB mRNA levels are increased during stress (heat shock) (Hubel *et al.*, 1995). Likewise it may be assumed that the clpC-like polypeptide may function in response to the numerous heat shock events subjected upon the *Plasmodium* parasite during its life cycle. It would be useful to generate antibodies to the clpC gene product for use in immunoprecipitation experiments, thereby confirming clp polypeptide synthesis. Expression of hsp 70 in *Leishmania* has been shown to be elevated at the translational level during stress conditions where the mRNA levels are steady throughout. Functional assays would include determination of an intrinsic *in vitro* (and ultimately *in vivo*) ATPase activity and whether this protein is able to associate with clpP to form a proteosome with a specific substrate specificity.
Figure 6.1

Box representation of the organisation of a typical clp gene.

The NTP binding domains are indicated (-1 and -2) in addition to the conserved regions A, B and B2. See text for a complete discussion.
Figure 6.2

Identification of the 35 kb encoded clpC-like gene.

Southern blot at low stringency of *P. falciparum* genomic DNA uncut (Lane 1), and endonuclease restricted with *Sca*I (lane 2); *Xba*I (lane 3); *Hind*III (lane 4) hybridised to a radio-labelled PCR product of the *Plasmodium* N2 domain.

a) signals obtained after the blot had been washed at high stringency.

b) signals observed after the blot had been washed at low stringency.
a) High stringency wash

b) Low stringency wash

White 17/2/92

N.I.M.R. London, NW7 1AA.
Figure 6.3

The nucleotide and predicted amino acid sequence of the 3' end of the clpC-like gene, amplified as a part of the 4 kb PCR product.

Numbers correspond to nucleotide positions.

Asterisk denotes stop codon.
CCT AAT TCA GTA ATA TTA TTT GAT GAA ATA GAA AAA
P N S V I L F D E I E K

GCA CAT CCT GAT ATA TAT AAT ATA ATG TTA CAA ATA
A H P D I Y N I M L Q I

TTA GAT GAA GGT AGA TTA ACA GAT TCT ACA GGT AAA
L D E G R L T D S T G K

TTA ATA GAT TTT ACA CAT ACA ATA ATT TTA TTA ACA
L I D F H T I I L L T

AGT AAT TTA GGT TGT CCA AAA AAT TAT GAT TTA TAT
S N L G C P K N Y D L Y

CTA AAA AAT AAA AAT TTT TTA TCA AAA TCG GAT TTA
L K N K N F L S K S D L

AAA GAA ATA GAA AAA AAT ATA AAA ATA AAT ATT AAT
K E I E K N I K I N I N

AAT TAT TTT AAA CCT GAA TTA TTA AAT AGA TTA ACT
N Y F K P E L L N R L T

AAT ATA TTA ATA TTT AAT CCT TTA AAT ATT AAT AAT
N I L I F N P L N I N N

TTA TTA TTT ATA TTT AAT AAA TTT ATA AAT GAA TTG
L L F I F N K F I N E L

AAA ATA AAA TTA TAT TTA AAT AAA TTA AAT ATT ATT
K I K L Y L N K L N I I

ATA CAT ATT AAT AAA GAA TTA AAA TAT TTT TTA GTT
I H I N K E L K Y F L V

AAA TTA ATG TAT AAT CCT TTA TAT GGA GCT CGT CCT
K L M Y N P L Y G A R P

TTA AAA AGA ATA TTA GAA TTA ATT TTT GAT AAA TCT
L K R I L E L I F D K S

ATA AGT GAT TTA TTA TTA ACT TAT AAT AAA CAT TAT
I S D L L L T Y N K H Y
541  TTT ATA AAA AAT AAA TAT ATT TTA TAT TAT TAT TTA
     F I K N K Y I L Y Y Y L

577  AAT AAA TAT TAT AAA TTA AAT TTT AAT ATA TAT TTA
     N K Y Y K L N F N I Y L

613  TTA TAA
     L *
Figure 6.4

PILEUP analysis of the 35 kb encoded clpC gene.

Amino acids identical to the *Plasmodium* sequences are shaded in yellow. Deletions inserted for optimal alignment are represented by dots.

*Plasmodium* (Accession number: x95276); t4a (tomato)(M32603); *P. sativum* (pea)(L09547); *E. coli* clpA (a)(P15716); *E. coli* clpB (b)(P03815); amiB (X77160).

The leucine-rich repeats observed in clpC coding sequences are overlined; the leucine repeat observed in the *Plasmodium* homologue is asterisked.

Both the N1 and N2 NTP binding domains are boxed with the corresponding A, B, B2 conserved sequences indicated.

Numbers on both the right and left hand side of the sequences relate to amino acid positions within the PILEUP.
921

P. sativum  VVDEGNP SMPY GARPLRRAIM LLEDGMAEKMLAREIKEGDS
t4a  VVDEGNP SMPY GARPLRRAIM LLEDGMAEKMLAGEIKEGDS
E. coli_b  LSENGYPYV GARPLKRAIQ QIENPLAQQI LSGELVPGKV
E. coli_a  LAEKGVDRAM GARPMARVIQ NLKKPLANEL LFGSLVGGQ
amiB  PEVLAKIARA GARALRRSVR HHLEVPLAEH LLDHHQPVDG
Plasmodium  LVKLMYNPLY GARPLKRILE IFDKSISSL LTYNKHYF IK

961

P. sativum  VIVDVS D.G KVIVLNGSG TP.ESLPEAT SI
t4a  VIVDVS D.G NVTVLNGTG APSDSAPEPV LV
E. coli_b  IRLEVNE..D RIVAVQ
E. coli_a  VTVALDKEKN ELTYGFQSAQ KHKAEAAH
amiB  NICTYLASLE HWEVRFVRR
Plasmodium  NKYLYYYYLN KYYKLNFN... ........I YLL
CHAPTER 7
CHAPTER 7

A CONTINUATION OF THE DESCRIPTION OF GENES ENCODED BY
THE 4 kb AMPLIFICATION PRODUCT

7.1.0 Introduction
This section describes the final two tRNAs, tRNA\textsuperscript{Gly} (UCC) and tRNA\textsuperscript{Ser}(UGA), to be identified on the 35 kb molecule. These tRNA genes are separated by an open reading frame (ORF) of 79 amino acids (ORF79) which has no assigned function. A second un-assigned 105 amino acid ORF (ORF105) downstream of tRNA\textsuperscript{Ser} completes the sequence of this strand, the 3' terminus of this ORF overlapping the 3' terminus of the rps 2 gene on the complimentary strand by 9 nucleotides. This marks the transition point of overlap to the complimentary strand on the IR\textsubscript{A} arm (Figure 7.1.)

7.1.1 tRNA general structure
tRNA molecules fold in solution into an L-shaped three-dimensional structure. The standard 2-dimensional representation resembles a cloverleaf (Figure 7.2). The four stems are stabilised by Watson-Crick base-pairing with three of the four stems ending in loops. Most tRNAs contain a four base sequence (UUCG)- the T\psi CG loop, which is modified as follows:- the first uridylate is methylated to form a thymidylate and the second uridylate is rearranged to form a pseudouridylate (ψ). This T\psi CG region is characteristic of tRNA molecules in addition to the di-nucleotide G motif in the D-loop and the anticodon loop. The conserved GG motif in the D-loop is involved in a
tertiary interaction with the conserved UC nucleotides in the TyCG loop (McCain, 1993.)

A total of 25 tRNAs (24 anticodons) and an initiator tRNA^Met have been identified on the 35 kb DNA of *P. falciparum* molecule and transcription of each of these molecules has been determined by both primer extension and 2-dimensional Northern blot techniques (Preiser et al., 1995). Six of these tRNAs (those encoding Leu, Glu, Gly(ACC), Cys, Gly(TCC) and Trp) have unusual features, that of tRNA^Gly^ will be discussed below. tRNA^Leu(UAA)^ contains a group I-like intron but this is only 130 nucleotides in length in comparison to the introns identified in both liverwort and tobacco chloroplast tRNA^Leu^ genes - 325 to 2526 nucleotides long (Ohyama et al., 1986; Shinozaki et al., 1986).

### 7.2.0 Results and Discussion

**7.2.1 Discussion of the tRNA^Gly^ and tRNA^Ser^ identified on the 4 kb amplification product**

The sequence of both tRNAs and the two ORFs are presented in Figure 7.3. Both the tRNA^Gly^ and tRNA^Ser^ can be folded with minimal free energy to the typical 2-dimensional clover leaf structure. tRNA Gly^(UCC)^ has an unusual feature in the T loop where the highly conserved GUUC has changed to AUUC. There is, however, another compensatory change from a C to a T thereby restoring the stem structure (Figure 7.4). Thus, it would appear that the necessary structure required for tRNA function has been maintained and, in addition to evidence of tRNA^Gly^ transcription, the possibility that this molecule is a pseudogene is unlikely. These nucleotide
substitutions are another indication of the extreme A-T pressure on the circle, in fact the A-T bias on this molecule is so extreme that only seven amino acids (Ile, Asn, Lys, Leu, Tyr, Phe, Ser) account for approximately 80% of the positions in 3000 codons and have an A or a T in the first position (Preiser et al., 1995). The wobble position is assumed to be utilised, allowing 57 codons to be recognised by the 24 anticodons identified on the plastome. Two separate anticodons for Gly (ACC, UCC) are used by the plastomes’ tRNAs to decode three of the four possible Gly codons. To explain this we should consider the evidence that the nucleotide at position 32 in tRNA\textsuperscript{Gly} in both \textit{E. coli} and mycoplasma significantly affects the decoding ability of this tRNA (Claesson et al., 1995). With a C at this position tRNA\textsuperscript{Gly} (ACC) can only decode GGT whereas with a U (UCC) it can utilise the wobble position decoding GGA and GGG. This situation is also found in the 35 kb molecule (where the GGC codon is not utilised), allowing the assumption that the same decoding process has been implemented. Assuming that the wobble position is used by the plastome translation machinery, 57 codons can be decoded by 24 tRNA anticodons making this molecule’s tRNA compliment the most minimal to date in chloroplasts.

### 7.2.2 Identification of homologues to ORFs -79 and -105

Searches of the SwissProt database with ORF 105 using the GCG BLASTP programme has identified a similar ORF of 120 amino acids in \textit{Cyanophora paradoxa} - ycr3 located in the intergenic region between rpl 3 and rpl 33 (Evrard et al., 1990). The two polypeptides are 56% similar and 32% identical see Figure 7.5. No function has been assigned to the ycr3 gene product but the retention of this gene in both organisms signifies that it may have functional importance. Various rRNA trees
suggest that the cyanelle of *C. paradoxa* either represents the earliest branching of the rhodophyte/chromophyte/euglenophyte cluster (Douglas and Turner, 1991) or, represents an early line, diverging from the plastid branch before the chlorophyll *a/b* plastids diversified (Turner et al., 1989). Therefore, the identification of an ORF in the 35 kb circle which has homology to an ORF in the cyanelle suggests a relatedness between these two genomes and once again indicates that an early rhodophyte is a strong candidate for the ancestor to the 35 kb DNA.

No protein homologue candidates have been identified for ORF 78.

### 7.3.0 Summary and future work

The identification of tRNA\(^{Gly}\) and tRNA\(^{Ser}\) completes the enumeration of tRNA molecules encoded by the 35 kb DNA. It has been proposed that the tRNAs identified on the circular molecule are sufficient for translation of 35 kb encoded genes (Preiser et al., 1995),

As yet, there is no reported sequence data in the EMBL database to allow a function to be assigned to the two ORFs identified in the 4 kb amplification product. Although a homologue to ORF 105 has been identified in *C. paradoxa*. One may assume that if these ORFs are an important feature of the plastid they will be maintained in the apicomplexan plastids identified in *T. gondii*, *Theileria* spp. and *Eimeria* spp. To determine whether this is the case, PCR amplification of these coding regions and subsequent sequencing of the resultant products must be performed. There are no transcript data of either ORF79 or ORF105 and although Northern blot analysis is the
obvious means to provide these data, due to the extreme difficulties encountered in RNA detection by both Northern blot and primer extension techniques (see Chapter 9), it would be preferable to utilise the more sensitive RNase protection assay.
Figure 7.1

A schematic representation of the genes identified on the 35 kb DNA by PCR.

The carboxy-terminus only of the clpC gene has been represented. The 23 nucleotide (n) overlap between tRNA^{Ser} and ORF 105 is depicted by arrows. The direction of transcription of these genes has been indicated.
Figure 7.2

The primary structure of yeast tRNA^Ala (Darnell et al 1986). See text for details.

Nucleotides are denoted by A; C; G; U. Modified nucleotides are shown in red.

(Abbreviations: D- dihydouridine; I - inosine; T- thymidine; ψ - pseudouridylate; m - methyl group). Both the anticodon and codon regions are indicated.
**Figure 7.3**

Sequence data of tRNA$^{\text{Gly}}$, tRNA$^{\text{Ser}}$, ORF 79 and ORF 105.

The tRNA genes are depicted in red, non-coding sequence in black, ORF 79 in green and ORF 105 in blue.

Asterisks denote stop codons.
AAA AGT AAA AAT TCA AAT AAT TAT ATA TAT AAT ATT
  K S K N S N N Y I Y N I

ATA AAT AAT AAA TAT AAA AAT ATA AAA TTA TTA TAT
  I N N K Y K N I K L L Y

ATT CTA TCA AAT AAT AAA TAT AAT TTA TTA TTA TTT
  I L S N N K Y N L L L F

AAA AAT ATT AAT TTA TGG AAT GTA TTA TTA AAT TAT
  K N I N L W N V L L N Y

AAT ATA ATA TTT AAT AAT ATA TAT ATA ATT AAA AAT
  N I I F N N I Y I I K N

ATA TTT ACA TAT TTT TAA
  I F T Y F *
Figure 7.4

Predicted secondary structures of tRNA\textsuperscript{Ser} and tRNA\textsuperscript{Gly}.

Nucleotides which are different from the usual conserved sequence are circled.
a.) Ser (tga)

b.) Gly (tcc)
Figure 7.5

PILEUP analysis of the 35 kb encoded ORF105 and ycr3.

Amino acids identical to the *Plasmodium* sequences are shaded in yellow. Deletions inserted for optimal alignment are represented by dots.

*Plasmodium* (Accession number: x95276) and the cyanelle encoded ORF- ycr3 (*C. paradoxa*) (P15811). Amino acid numbers are indicated.
Plasmodium  MGSNPFSFLC........ IIMIFN LYYLKLKNLL
C. paradoxa ...MNVLSYFTWFVHLSSVEL EWLNIYFLFLYTTLKKNLS

Plasmodium  LKKFKNIQINNNIKKIVYIKLFNILLKSKNSNNYIYNIIN
C. paradoxa  LKTF.................IFSFFISFCSA...LCACTL

Plasmodium  NKYNIKLLYILSNKNYNLLLFKNINLWNVLLNYNIIFNN
C. paradoxa  HFFNNQSFYYFLINLQSLTLFANITLY...FSILYK

Plasmodium  IYIKNIFTY
C. paradoxa  QQILKIQN.Y
CHAPTER 8
CHAPTER 8

THE IDENTIFICATION OF ORF470 HOMOLOGUES

8.1.0 Introduction

An un-identified reading frame ORF470 encoded by the malarial plastid bears 52% identity and 67% homology at the amino acid level to an open reading frame in the plastid DNA of the red alga *Antithamnion* spp. (ORF3A) (Williamson et al., 1994). Database searches have revealed similar ORFs in the unicellular algae *Cyanidium caldarium* and *P. purpurea* (Williamson et al., 1994). The presence of these homologous coding sequences within rhodophytes therefore suggested a possible lineage of the apicomplexan plastid as mentioned in Chapter 1. An homologous reading frame is encoded by 35 kb plastids in the Apicomplexans *E. tenella* and *T. gondii* strains RH and 18691 (P. Denny, personal communication). The *T. gondii* homologue, however encodes an opal TGA stop codon at position 285 which in other organisms is occupied by a conserved tryptophan (TGG), this situation is possibly an example of the codon capture theory where the TGG codon has moved by directional A+T pressure to TGA. This gene has been shown by RT-PCR to be expressed and is presumably translated in *T. gondii*.

Further characterisation of the *P. purpurea* homologue (ORF487) by Reith (personal communication, Institute for Marine Biosciences, Halifax, Nova Scotia) suggested that this gene product may perform a metabolic function since, the mRNA transcript levels were low in cells grown without light for 3 days, yet became abundant in the presence of light. This was in contrast to the mRNA levels of photosynthetic genes.
where the transcript level remained high after 3 days. Biosynthetic genes have exhibited similar transcript variations in levels (Apt et al., 1993) as those described for ORF487. ORF487 was also shown to be transcribed together with a neighbouring gene - ORF251 which is thought to be a member of the ABC$_1$ transporter family. This same gene order has been determined in both *Antithamnion* and *Cyanidium*.

The identification by Pietrokovski in 1994 of a coding sequence in *Mycobacterium leprae* (ORF2) encoding a putative protein of 483 amino acids, sharing 53% similarity and 32% identity with ORF470 has added to the significance of the retention of this ORF on the malarial plastid. An interesting feature of the mycobacterial ORF2 is that it contains an intervening protein sequence (intein) (Figure 8.1a). The intein is probably self excised (Davis et al., 1992) and encodes for a site-specific DNA endonuclease used for insertion of the intein coding sequence into other coding sequences (Gimble and Thorner, 1992; Perler et al., 1992). In this sense, inteins may be regarded as the protein equivalent of the self-splicing group I RNA intron. During this process the peptide C- and N-terminal regions (exteins) are spliced together via transpeptidation to form a second functional polypeptide. The presence of inteins is not common, having been observed to date only in the vacuolar proton pump ATPase subunit of the eukaryote *Saccharomyces cerevisiae* (Hirata et al., 1990; Kane et al., 1990) and *Candida tropicalis* (Gu et al., 1993), in the DNA polymerase of *Thermococcus littoralis* (Perler et al., 1992; Hodges et al., 1992), in RecA of *Mycobacterium tuberculosis* (Davis et al., 1992) and *Mycobacterium leprae* (Davis et al., 1994). There appears to be no phylogenetic relationship between the organisms containing inteins but it has been proposed that the presence of an intein within a gene
indicates that it has a functional significance either for the gene itself or the host organism (for a review see Colston and Davis, 1994).

The *M. leprae* ORF2 is believed to be co-transcribed within an operon comprising at least 6 ORFs. Figure 8.2 shows a diagrammatic representation of the gene order, no transcription data have been determined with regard to the transcription of these genes nor is it known if they are transcribed as an operon although Dr. J. Colston (NIMR) suggests that since ORFs 2-7 overlap by up to 20 nucleotides it is probable that they are transcribed as a polycistronic message. ORF3 has an un-assigned function yet bears homology to ORF2, the ORF3 coding sequence however does not contain an intein. Comparison with other polypeptides in the EMBL database predicts that ORF4 encodes an ABC transporter, ORF5 encodes a nifS polypeptide and ORF6 encodes nifU (E. Davis and D. Jenner, NIMR, personal communication).

It is possible that functional implications can be derived from the identification of an ORF470 homologue within this *M. leprae* operon because the corresponding gene in the rhodophytes *Antithamnion*, *Porphyra* and *Cyanidium* lies proximal to a member of the ABC transporter family (ORF251). That this ORF is probably co-transcribed with the *Porphyra* ORF487 gene suggests these genes are functionally related. Furthermore, the identification of the *M. leprae* ORF2 sequence, in addition to the presence of an intein within this gene, suggests that the ORF470 gene product has an important function and its presence alone on the 35 kb circle may be the purpose for the retention of this molecule in the parasite.
In agreement with this reasoning, data presented in this chapter indicate that ORF470 homologues are widespread in bacteria being encoded by *Mycobacterium smegmatis*, *Mycobacterium tuberculosis* (DNA provided by E. Davis, NIMR), *Lactobacillus*, *Corynebacter*, *B. subtilis* and *S. aureus* (DNA provided by M. Strath) and the cyanobacterium *Synechocystis* PCC 6803.

**8.2.0 Results and Discussion**

**8.2.1 The identification of ORF470 homologues**

Two PCR products (probes -1 and -3) were amplified from the *M. leprae* ORF2, the locations of these products within the ORF are indicated in Figure 8.1. Southern blots of endonuclease restricted genomic DNA from *M. leprae*, *M. tuberculosis* and *M. smegmatis* were hybridised with Probe-1 randomly labelled with [$\alpha$-$^{32}$P], the incubation conditions being 42°C overnight with washing at the same temperature at 0.2X SSC; 0.5X SDS. After autoradiography (Figure 8.3a), the blot was stripped and re-probed with [$\alpha$-$^{32}$P]-labelled Probe-3 in the same way (Figure 8.3b). Both probes hybridised with DNA fragments in each of the three organisms suggesting that homologous coding sequences to the *M. leprae* ORF2 are encoded by *M. smegmatis* and *M. tuberculosis*. Comparison of Figure 8.3 (parts a. and b.), however, suggests that an intein is present in *M. leprae* and possibly in *M. smegmatis*. This is because the signals observed with both probes were only identical for *M. tuberculosis* whereas additional signals could be observed with *M. leprae* and *M. smegmatis* DNA probed with Probe-1 rather than Probe-3 (indicated by arrows in Figure 8.3). These lower bands in *M. leprae* are consistent with sizes predicted from nucleotide data of the intein region.
The conservation of this ORF within mycobacteria indicates that the function of its expressed product is not one solely required for plastid maintenance but may fulfil a role in general cell maintenance or viability. Attempts were therefore made to find homologues in organisms more tractable to molecular manipulations for future characterisation of ORF470. Southern blots of endonuclease restricted genomic DNA from *Lactobacillus*, *Corynebacter* spp., *B. subtilis* and *S. aureus* (Oxford strain) were probed with \([\alpha-^{32}P]\) labelled Probe-3 as previously described. This probe was used since the aim of this experiment was not to detect inteins, the presence of intein coding sequence within Probe-1 possibly leading to mis-interpretation of data. The signals obtained for *Corynebacter* spp. only are presented in Figure 8.4, since the signals obtained with the other organisms were very weak. This result does suggest the presence of *M. leprae* ORF2 homologues in these organisms.

Attempts were made to identify this gene also in *E. coli*, the cyanobacterium *Anacystis nidulans* R2 and the filamentous thermophile *Phormidium laminosum* (genomic DNAs for the last two were provided by C. Howe Department of Biochemistry, Cambridge). Southern blots of genomic DNA probed with Probe-3 failed to identify potential homologues due to non-specific hybridisation of the probe (data not shown). The non-specific signals obtained were probably due to DNA degradation. However, having identified homologues of ORF470 in numerous bacterial organisms a further attempt was made to isolate clones from a *Synechocystis* PCC 6803 cosmid library (see below).
8.2.2 The isolation of *Synechocystis* PCC 6803 recombinant cosmids

A nylon membrane containing 768 *E. coli* clones of recombinant cosmids carrying *Synechocystis* PCC 6803 genomic DNA was provided by V. Shestopalov (Vavilov Institute of General Genetics, Russia) and hybridised with $[\alpha-^{32}P]$ labelled Probe-3. The grid corresponds, in configuration, to a microtiter plate and consist of 96 (8 x 12) clusters of spots. Each cluster contains 16 spots, which are 8 clones spotted in duplicate. The positions of the clones are: 1 2 3 4

5 6 7 8

8 7 6 5

4 3 2 1

allowing unambiguous recognition of positive clones.

Hybridisation was performed at 42°C overnight and washed at the same temperature at 0.2 X SSC; 0.5 X SDS prior to autoradiography. A unique probe is expected to recognise 8 clones from this representative library, each clone containing approximately 39 kb genomic DNA. 11 clones were recognised by Probe-3 (Figure 8.5) and upon request clones numbering:- A1#2, A1#4, A3#1, B6#4, D5#1, D6#3, D6#5, D11#5, E2#6, G12#4, H9#6, were kindly provided by V. Shestapolav. These clones were grown in overnight cultures in the presence of Kanamycin at 30μg/ml and cosmid DNA was purified using Magic Minipreps™ (Promega) according to the manufacturer’s protocol. A Southern blot of un-restricted recombinant cosmid DNA was probed with $[\alpha-^{32}P]$ labelled Probe-3 which hybridised to all the clones suggesting that they are “true” positives (data not shown). A physical map of *Synechocystis* PCC 6803 constructed by Y. Churin and T. Borner (provided by V.
Shestopalov) see Figure 8.6 indicates that the clones recognised by Probe-3 lie all in the region marked by an arrow.

Although these results were produced at the end of the project it was decided earlier to synthesise fusion proteins to ORF470. The pGEX expression system (Pharmacia) was used which produces an N-terminal Glutathione S-Transferase (GST) fusion protein upon induction with IPTG. An amplified region at the 3' terminus of ORF 470, in addition to the entire coding sequence, were independently cloned into pGEX vectors (Promega). Recombinant clones were isolated by endonuclease restriction digestion of plasmid DNA and the 5' and 3' termini of the insert/plasmid junctions determined by di-deoxy sequencing. The recombinant clones in addition to a positive control expression clone (provided by I. Ling NIMR) were induced and screened by immunoblotting with an anti-GST polyclonal antibody (provided by I. Ling). A signal from the control clone at the expected molecular weight was clearly observed but no signals from the ORF470 clones could be visualised (data not shown). In fact, in the cases of the fusion protein expressing the entire ORF470 gene product the induced bacteria died within 2 hours of induction. Reduction in the induction time did not yield any fusion products. This result suggests that over expression of ORF470 may be deleterious to the cell, indicating the potential importance of this gene product. Due to lack of time, no other expression system has been tried, although independent experiments by P. Preiser (NIMR, personal communication), with another ORF470 pGEX clone also failed to produce a fusion protein.
8.3.0 Summary and future work

The results presented in this chapter have identified homologues of ORF470 in several bacterial organisms. Genomic clones of *Synechocystis* PCC 6803 have been isolated from a Lorist6 cosmid library and it is hoped that sequencing of these clones in the future will provide information on the genes flanking this ORF.

Although attempts to produce fusion proteins of ORF470 failed, the lysis of bacteria upon induction suggested that overproduction of this gene product might be lethal to the host.

Future work on this ORF must include the generation of antibodies to this gene product. This may be achieved either by producing fusion proteins of ORF470 in other expression systems or by raising anti-peptide antibodies to predicted polypeptides. In addition to verifying ORF470 gene expression, antibodies to ORF470 would be used to determine whether antigenic cross reactivity exists between the 35 kb encoded gene product and those homologues identified in this study.
Figure 8.1

Box diagram of the *M. leprae* ORF2.

A) Representation of the *M. leprae* ORF2. The intein is indicated as a blue box within the extein regions. The primers used to amplify Probes -1 and -3 are indicated; the primer orientations (5'-3') indicated by arrows.

B) The regions of ORF2 encompassed by Probes -1 and -3.
a.)

5' extein

5' intein

3' extein

5' intein

ORF 5'

ORF 3'

b.)

<table>
<thead>
<tr>
<th>Primers used</th>
<th>Amplification product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probe 1: 5' intein/3'</td>
<td></td>
</tr>
<tr>
<td>Probe 3: 5'/3'</td>
<td></td>
</tr>
</tbody>
</table>
**Figure 8.2**

The location of ORF 2 in *M. leprae*. The direction of transcription is indicated by an arrow. The regions of overlap between coding regions are indicated by hemi-spheres.

Those ORFs which have been assigned functions are indicated.
Figure 8.3

Identification of ORF470 homologues in *Mycobacterium*.

A Southern blot of *M. leprae* (Lep), *M. tuberculosis* (Tb) and *M. smegmatis* (Sm) genomic DNA restriction endonuclease digested by 1 -HaeIII, 2 - PstI, 3 - EcoRI.

a) - hybridised with radio-labelled Probe-1. Arrows point to signals relating to intein sequence.

b) - hybridised with radio-labelled Probe-3.
Figure 8.4
Identification of an ORF470 homologue in *Corynebacter* spp.

Southern blot of *Corynebacter* genomic DNA 1- uncut, and endonuclease restricted by 2 - *Hae*III, 3 - *Pst*I, 4 - *Eco*RI hybridised to radio-labelled Probe-3.
Figure 8.5

Identification of ORF470 homologues in *Synechocystis* PCC 6803.

*Synechocystis* PCC 6803 cosmid DNA grid (a kind gift from Dr. V. Shestopalov, Vavilov Institute of General Genetics, Russia) hybridised to radio-labelled Probe-3.
Figure 8.6

Physical map of *Synechocystis* PCC 6803 (provided by Dr V. Shestopalov).

Arrow indicates the region which clones A1#2, A1#4, A3#1, B6#4, D5#1, D6#3, D6#5, D11#5, E2#6, G12#4, H9#6 have been mapped.
Synechocystis PCC 6803
3820 kb

psaAB
psbH
ndhH
ORF173
psbD2
psbAl
ndhCKJ
AA
3820/0 kb
3500
500
AB
dhD
AA
Synechocystis PCC 6803
000
3820 kb
ndhF
petH
psbB
500
2000
psbEF
psbK
0RF324
rriB
ctpA
rbcLS
ndhAIGE
rrnA
psbA2
psbA3
ORF184
psaC1
dfnA
tal
psaD
psbO
psbD1C
psbA1
ndhCKJ

NotI
CHAPTER 9
CHAPTER 9

THE IDENTIFICATION OF A SECOND \textit{tuf}-GENE IN \textit{P. falciparum}.

\subsection*{9.1.0 Introduction}

Most of the 35 \textit{kb} encoded genes are involved in gene expression. The identification of a \textit{tuf} gene encoding elongation factor (EF-Tu) on the 35 \textit{kb} DNA indicates that this molecule utilises self encoded genes not only for transcription but also translation processes. EF-Tu and EF-G co-ordinate polypeptide synthesis in bacterial organisms, the eukaryotic homologues EF-1\(\alpha\) and EF2 catalysing similar reactions within the eukaryotic cytosol. EF-1\(\alpha\) has been identified in \textit{P. falciparum} by D. Williamson and P. Ross MacDonald (PhD thesis 1989) but will not be discussed here.

Protein synthesis in the malarial mitochondrion has not been directly determined. However several lines of evidence indicate that it occurs. This evidence includes the observation of cytochrome oxidase activity in \textit{Plasmodium} spp. (Scheibel \textit{et al.}, 1988; Sherman, 1979) which indicates that the mitochondrion encoded cytochrome c oxidase subunits I and III (\textit{COI} and \textit{COIII}) gene products (necessary for enzyme activity) are present. Moreover the mitochondrion encoded apocytochrome \(b\) (CYb) protein was detected in Western blots (Feagin, 1994); and a peak for type \(b\) cytochrome appears in spectral assays of isolated mitochondria of \textit{P. falciparum} and \textit{P. yoelii} (Fry \textit{et al.}, 1991). Since evidence suggests that the malarial plastid 35 \textit{kb} molecule and mitochondria 6 \textit{kb} DNA do not inhabit the same organelle (Chapter 1, section 1.8), a nucleus encoded, mitochondria-associated EF-Tu molecule was
predicted to be present in the parasite. The gene encoding this polypeptide would be additional to the plastid encoded *tuf* gene and the eukaryotic EF-1α encoding gene.

RNase protection assays have been employed to show that the plastid *tuf*A gene is transcribed although presumably at levels below that necessary for detection using Northern blotting techniques (see discussion below).

Data presented within this chapter suggest that a second *tuf* gene is encoded in *P. falciparum* and the chromosome carrying this gene has been identified.

### 9.1.1 Protein biosynthesis - an introduction

EF-Tu and EF-G are prototypes of the G protein family in which the ‘active’ form of the enzyme binds GTP and the ‘inactive’ form GDP, hydrolysis of GTP taking place during the catalysis process.

Protein synthesis occurs on the ribosome but the number and sizes of the RNA molecules within the ribosomes of eukaryotes, prokaryotes and organelles differ. Bacterial and chloroplast ribosomes consist of a 50S subunit (containing a 5SrRNA, a 23S rRNA and large subunit ribosomal proteins) in addition to a 30S subunit (comprising 16S rRNA and small subunit ribosomal proteins). The association of the two subunits with the mRNA forms the active ribosome. Protein biosynthesis is a three stage process consisting of initiation, elongation and termination. This section will describe bacterial protein biosynthesis only. During initiation, messenger RNA (mRNA) is bound by the 30S subunit and recognised by an initiator transfer RNA
(tRNA). Assembly of the intact ribosome is then catalysed by initiation factors. Elongation is catalysed by three enzymes (Miller et al., 1977; Kaziro, 1978) and will be discussed in more detail. Termination occurs at the stop codon where release factors are utilised.

The A, P and E sites on the ribosome are available for association with tRNA molecules and during protein biosynthesis tRNAs are delivered to each of these sites respectively. The A site accommodates the aminoacyl-tRNA (\(\text{a}_\text{tRNA}\)) whilst the P site contains the peptidyl-tRNA complex. During a process catalysed by peptidyl transferase, the peptidyl-tRNA complex is transferred to the A site where a peptide bond is formed between the tRNA occupying this site and the growing polypeptide chain. This peptidyl-tRNA complex is translocated back to the P site and the deacylated tRNA to the E site while the ribosome progresses one codon along the mRNA. The ribosome is therefore thought to be in one of two states fluctuating between post- and pre- translocation. Elongation factors assist the switch between the two ribosome states by reducing the high activation energy barrier of approximately 90 kJ mol\(^{-1}\) (Schilling-Baretzko et al., 1992). EF-Tu drives the ribosome from the post- to the pre- translocational state whereas EF-G promotes the reverse reaction.

Two EF-Tu.GTP molecules participate in the recognition of \(\text{a}_\text{tRNA}\) and this pentameric complex delivers the \(\text{a}_\text{tRNA}\) to the ribosome A site. Occupation of the A site is associated with GTP hydrolysis to GDP by EF-Tu on the ribosome with the concomitant ejection of EF-Tu:GDP (Moazed et al., 1989). The EF-Tu molecules are dissociated from GDP by two EF-Ts molecules and upon dissociation of this complex
and re-association of EF-Tu with GTP another tRNA is recognised by EF-Tu (Kawashima et al., 1996).

The pre-translocational state of the ribosome is less energetically favoured than that of the post-state and transfer between the two is readily catalysed by EF-G. The intrinsic GTPase activity of this enzyme cleaves GTP causing the EF-G:GDP complex to fall off the ribosome. Unlike EF-Tu, EF-G readily dissociates from GDP to recombine with GTP to form an active EF-G molecule.

A yeast \(S.\ cerevisiae\) mitochondrial homologue of EF-Tu was identified by Nagata and colleagues (1983) by the hybridisation of Southern blots of endonuclease restricted genomic DNA, with a sequence corresponding to the \(E.\ coli\ tufB\) gene. Molecules migrating in the region of the observed signal were cloned and subsequently sequenced resulting in the identification of the yeast mitochondrial EF-Tu gene \(tuF/M\). This gene has 60% homology with the \(E.\ coli\ tufB\) gene at the nucleotide level, and 66% homology at the amino acid level.

9.2.0 Results and Discussion

9.2.1 Evidence for the presence of a second \textit{tuf} gene in \textit{P. falciparum}

Following the rationale of Nagata and colleagues, heterologous Southern blotting of \textit{P. falciparum} genomic DNA with a yeast probe was performed in order to ascertain whether a mitochondrial associated EF-Tu homologue additional to the \textit{tuf} gene on the 35 kb circle was encoded by the parasite. A region highly conserved between the \textit{E. coli tufA} and the yeast \textit{tuF/M} gene was amplified by PCR using primers \textit{ytuf5} and \textit{ytuf3}

192
This amplification product was randomly labelled with $\alpha^{-32}P$ and hybridised to Southern blots of P. falciparum and S. cerevisiae endonuclease restricted genomic DNA at low stringency. In parallel, an identical Southern blot was hybridised with a randomly labelled $\alpha^{-32}P$ PCR product (using primers EFTU 5' and EFTU 3') corresponding to the 3' end of the P. falciparum plastid tufA gene. Both blots were washed at 0.1X SSPE, 42° C.

These results are summarised in Figure 9.2 and Figure 9.3. The data show that following washing at low stringency conditions, the yeast and malarial probes recognised different malarial DNA molecules. Figure 9.2 indicates that, as expected, the yeast probe recognised yeast fragments with the same mobility as described by Nagata (1983). The presence of a band migrating >23 kb in Lane 1 suggests incomplete digestion by PstI. This probe also cross-hybridised with P. falciparum genomic DNA giving a single band in each digest. The data presented in Figure 9.3 suggests that although the malarial tufA probe readily detected the 35 kb tufA gene, no other P. falciparum DNA was recognised. Additionally, no yeast DNA was recognised with this probe, signals being absent after autoradiographs were exposed for several days (data not shown). Both of the parasite tuf genes and the identified yeast tufM gene are assumed to be single copy genes since only one reactive DNA fragment was recognised in various endonuclease digests. Two copies of the tuf gene are encoded by the human mitochondrion (Wells et al., 1995) and in Gram negative bacteria, whereas a single copy of this gene is present in Gram positive bacteria, except Mycobacterium vaccae where two copies are present (Sela et al., 1989; Dhandayuthapani et al., 1994).
The size of the *P. falciparum* DNA fragments recognised by the yeast *tufM* probe do not correlate with the known fragment sizes of the nuclear EF-1α gene (Williamson personal communication; P. B. Ross MacDonald - PhD thesis 1989.) or the 35 kb encoded *tufA* gene - see Table 9.1. The yeast *tuf* sequences compared to the *P. falciparum* plastid *tuf* sequence are more G+C rich and although malarial nuclear DNA has a high A+T content, I assume that the conserved mitochondrial *tuf* domains have retained a G+C composition sufficient for recognition by the yeast probe. The proposed second malarial *tuf* gene is assumed to be nucleus encoded since both the 35 kb circle and mitochondrial 6 kb element have been entirely sequenced. This finding adds further credence to the idea that gene products encoded by the 35 kb molecule do not participate in mitochondrial function. Moreover, the sequence of the plastid *tufA* gene did not reveal any peptide signal necessary for transport of the gene product into the mitochondrion. Peptide signal sequences have been identified in both the human and yeast mitochondrial *tuf* genes and, as expected, are absent in their bacterial counterparts (Wells *et al.*, 1995; Nagata *et al.*, 1983). Peptide signals have been shown to be utilised by most organisms for import and export of gene products between organelles, and are necessary for transport of proteins into the mitochondrion (Pfanner *et al.*, 1990). Therefore, assuming the two malarial extra chromosomal molecules are not co-localised, the sharing of gene products between the two malarial extra chromosomal elements would have to be mediated by inter-organelar peptide signals. Absence of such sequences in general, and particularly in this instance, implies that the 35 kb encoded *tufA* gene product could not be utilised by the mitochondrion. On the basis that protein synthesis in the mitochondrion is active, a nuclear encoded homologue must be present.
9.2.2 Chromosomal mapping of the EF-Tu homologue of mitochondria.

A PFGE blot (provided by P. Moore, NIMR) of *P. falciparum* and yeast chromosomes was hybridised with the yeast *tufl* probe. Hybridisation and washing conditions were identical to those used above. The signals obtained confirmed the localisation of the yeast *tufl* probe to yeast chromosome 15 (Amakasu *et al.*, 1993) and in addition, localised the *P. falciparum* homologue to chromosome 1 (see Figure 9.4). Figure 9.5 shows the ethidium bromide stained gel prior to blotting and Figure 9.4 shows the signals obtained after hybridisation with the yeast *tufl* probe.

9.2.3 The 35 kb - encoded *tufA* transcript

Northern blotting techniques were employed initially to determine whether the 35 kb encoded *tufA* gene was transcribed. Prior to probing the blot, a strip was stained with methylene blue in order to visualise the ribosomal RNA bands which were clearly discernible (indicated by arrows in Figure 9.6), suggesting that the RNA was intact. The Northern blot was divided into two and the strips hybridised with either the 35 kb *tuf* probe described above, or a region of the 35 kb encoded small ribosomal RNA (SrRNA) (clone 6b) which served as a positive control (see Figure 9.6). The SrRNA gene is highly expressed producing a transcript of 1.5 kb (Gardner *et al.*, 1991b). In contrast a series of bands but no unique signal was obtained using the *tufA* probe. RNase protection assays have since been used to show that the *tufA* transcript is present at low levels (Preiser, un-published data).

Identification of the malaria plastid transcripts has been hindered due to obtaining non-specific or non-repeatable signals whilst using Northern blotting techniques. The
RNase protection assay is more sensitive and has now been used to identify the presence of several 35 kb transcripts (P. Preiser - unpublished data). Whether the transcript abundance of 35 kb encoded genes changes in the non-erythrocytic stages of the parasite development has not been investigated due to the difficulty in obtaining non-erythrocytic forms of the parasite. This leads us to question whether the low level of plastid transcription during the erythrocytic forms of the parasite is due to lack of function of the 35 kb encoded genes at these stages. Alternatively the ‘smears’ observed whilst using Northern blotting and primer extension techniques may be due to polycistronic transcription of the 35 kb molecule similar to that observed in prokaryotic operons. Transcription of the 35 kb encoded rpo genes has been described as polycistronic (Feagin 1994; Gardner et al., 1991a) and it is possible that all of this plastid genome is expressed similarly. No promoter sites have been identified to date, part of the problem being the difficulty of identifying the A+T rich promoter elements typically observed in the prokaryotic Pribnow box consensus sequence (TATA(A/T)A(A/T)XG(X)₂G (Darnell et al., 1986).

9.3.0 Summary and future work.

A yeast mitochondrial tuf gene probe has allowed identification of a homologue in P. falciparum, additional to that encoded by the 35 kb circle. This gene is a single copy gene located on Chromosome 1. We propose that this might correspond to the malarial equivalent of tuf/M and would also predict that this gene has a leader sequence.
Obviously, further experiments need to be performed in order to verify that the DNA recognised by the yeast \textit{tufM} probe is, indeed, the mitochondrial \textit{tuf} gene. These experiments would either follow a path similar to that outlined above (Nagata \textit{et al.}, 1983) or would involve screening a \textit{P. falciparum} genomic YAC library of chromosome 1 with the same yeast probe previously described, followed by sequencing of the clones obtained.

Investigation of transcript production during the other stages of the parasite life cycle is required to address whether transcription of this gene, in addition to the other 35 kb encoded genes, is temporally regulated.
Figure 9.1

Box representation of the *S. cerevisiae* *tuf*M polypeptide.

This figure includes the amino acid sequence of the yeast *tuf*M PCR probe used in the detection of the *P. falciparum* nucleus encoded homologue. This sequence is presented in the form of a PILEUP analysis together with the corresponding region of the *E. coli* *tuf*A gene product (accession number P02990) and the *P. falciparum* 35 kb-encoded *tuf*A gene product (green).
**P. falciparum**

HIDCPGHSDY IKNMIGATQ MDAILVISI IDGIMPQTYE HLLLIKQIGI KNIIFNLKE DLCDDVELID FIKEVNEILL IKYNFDLNYI

**E. coli**

HVDCPGHADY VKNMITGAAQ MDGAIIVAA TDGQMPQTER HILLGRQVGV PYIIVFLNKC DMVDDEELLE LVEMEVEELL SQYDPGDRT

**S. cerevisiae tufM**

HVDCPGHADY IKNMITGAAQ MDGAIIVAA TDGQMPQTER HILLARQVGV QHIVVFVKNK DTIDDPEMLE LVEMEVEELL NEYGFDGDNA

---

**P. falciparum**

HILTGSALNV INIINQKNDY ELIKSNWQ ILNNLIQIID NII.IPTRKI NDYFLMSIED VFSITGRGTGTV VTGIKBEQCI NL

**E. coli**

PIVRGSAI.....KAL EGDAE..WAE KILELAGFID SYIPEPERAI DKPOLLPIED VFSISGRGTGTV VGTRVERGII KV

**S. cerevisiae tufM**

PIIMSAL..I.....CAL EGRQPEIGEIQ AIIKLDDAVD EYIPTPERDL KNKFLMPVED IFISGRGTGTV VGTR.......
Figure 9.2

Identification of the *P. falciparum* nucleus encoded *tufM* homologue.

Southern blot of *P. falciparum* and *S. cerevisiae* genomic DNA endonuclease restricted by *PstI* (lanes 1); *EcoRI* (lanes 2); *BglII* (lanes 3); unrestricted DNA (lane U). The blot was hybridised to the radio-labelled yeast *tufM* PCR product under low stringency conditions, as described in the text.
P. f  
1  2  3

S. cerevisiae  
1  2  3  U

kb
23.0 -
9.4 -
6.5 -
4.3 -
2.3 -
2.0 -
Figure 9.3

Detection of the 35 kb encoded *tufA* gene.

Southern blot of *P. falciparum* genomic DNA endonuclease restricted by *PstI* (lane 1); *EcoRI* (lane 2); *BglII* (lane 3); hybridised to radio-labelled *tufA* probe.
Table 9.1

Comparison of tuf EcoRI fragments

The EF-1α EcoRI fragment size was obtained from P. R. MacDonald’s -Ph.D thesis 1989.

The sizes of the DNA fragments presented in this table are approximate.
<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>GENE</th>
<th>SIZE / kb</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. falciparum</em></td>
<td>EF-1α</td>
<td>3.8</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>tufA</td>
<td>35</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>tufM(?)</td>
<td>16</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>tufM</td>
<td>10</td>
</tr>
</tbody>
</table>
Figure 9.4

Chromosomal mapping of the nucleus encoded *P. falciparum* *tufM* homologue.

Chromosome blot of the PFGE gel shown in Figure 9.5, hybridised with radio-labelled yeast *tufM* probe described in the text.
Decreasing amount of DNA on gel
Figure 9.5
Ethidium bromide stained PFGE gel

Decreasing amounts of DNA from left to right. *P. falciparum* (C10 strain) chromosome DNA (P.f), *S. cerevisiae* (strain YD148) chromosome DNA (Yeast).

Yeast chromosome 15 is indicated by the short arrow and *P. falciparum* (P.f) chromosome 1 is indicated by the long arrow.
Decreasing amount of DNA on gel

P. f.  Yeast  P. f.  P. f.  P. f.  Yeast  Yeast  Yeast
Figure 9.6

Northern blot transcript data of the \textit{tuf}A gene.

Northern blots of \textit{P. falciparum} total RNA hybridised with radio-labelled \textit{tuf}A or SrRNA (6b) probe as indicated.
CHAPTER 10
CHAPTER 10

10.1.0 Conclusion

The sequence data of the 5.3 kb amplified PCR product, whose analysis forms the bulk of this thesis, completes the sequencing of the 35 kb DNA molecule. The coding sequences identified include tRNA^{Ser}, tRNA^{Gly}, rpoC_2, and an ORF which bears similarity to an ORF encoded by the cyanelle in C. paradoxa. The sequence of the clpC-like gene (previously identified by K. Rangachari) was completed as well as that of the rpoC_1 gene (Gardner et al., 1991b).

Features of gene organisation and sequence similarity in this segment of the 35 kb DNA again supports the general similarity found between it and plastid DNA molecules. These features include the identification of a split rpoC gene, the rpoC_1/C_2/rps 2 gene order, and the ORF105 sequence. In fact, the identification of a similar ORF to ORF105 in the cyanelle of C. paradoxa, points to the ancestor of the 35 kb DNA being related to an early red algal-like organism (see Chapter 7).

Both monoclonal antibodies to a peptide based on the sequence of the predicted product of the rpoB gene, and polyclonal antibodies generated to an rpoB fusion protein detected a polypeptide migrating at approximately 120 kDa, consistent with what would be predicted for the P. falciparum β' subunit. Additional evidence for the translation of the rpoB/C_1/C_2 genes came from Western blot techniques using a polyclonal anti-E. coli RNAP antibody (EcB). My interpretation of these experiments is that the P. falciparum β, β', β'' polypeptides share epitopes with the E. coli RNAP β and β' subunits. The detection of a polypeptide migrating at approximately 105 kDa
in Western blots of *P. falciparum* whole cell lysate probed with affinity selected antibodies directed against *E. coli* β/β' subunits also suggests that the *P. falciparum rpoC*₂ gene product exists as a full length polypeptide (predicted 117,000 Mᵣ) and that the frameshift identified in this gene does not disrupt translation of the *rpoC*₂ gene.

The two tRNAs identified in the sequence data presented here were the last to be determined on the 35 kb DNA, bringing the total to 25 species. It has been suggested by Preiser and colleagues (1995) that this is the minimum number of tRNA molecules necessary for translation of the 35 kb-encoded genes.

In completing the complete clpC-like sequence it became clear that the N-terminal region of the predicted polypeptide is not highly conserved at the amino acid level whereas the C-terminus is highly conserved. By virtue of this degree of conservation it is likely that at least one of the ATP-binding domains (the N2 domain) is capable of ATP hydrolysis. Accordingly, this gene product although highly derived might still be associated with clp-like functions. Such activities include the protection of polypeptides from aggregation or heat in-activation during stress as with the yeast encoded chaperone hsp 104 (Parsell *et al*., 1994); association with a protease (i.e. clpP) to form a substrate specific protease - this would involve the import of the proteolytic subunit (clpP) encoded by the parasite nucleus; or thirdly, the 35 kb-encoded clpC-like polypeptide may be analogous to the amiB gene product in *P. aeruginosa* (Wilson *et al*., 1995), that interacts with the cytosolic domain of an ABC₁ transporter. Indeed, a similar role might be considered also for the ORF470 polypeptide, since this gene is co-transcribed with an ABC₁ transporter in the plastid
genomes of *P. purpurea*, *C. caldarium* and *Antithamnion* spp. Moreover, the ORF470 homologue in *M. leprae* is also thought to be transcribed in an operon that includes an ABC₁ transporter. Since prokaryotic encoded ABC₁ transporters are often transcribed in an operon together with their substrate, and if ORF470 is involved in a biosynthetic function, as suggested by the transcription data in *P. purpurea*, it is possible that an ABC₁ transporter is involved in export of this polypeptide from the plastome. The clpC-like polypeptide could associate with a membrane spanning polypeptide and provide the energy required for transport by ATP hydrolysis. In addition to the plastid encoded ABC₁ transporters in *P. purpurea*, *C. caldarium* and *Antithamnion* spp. the only other plastid encoded ABC₁ transporter identified to date is the MbpX gene, encoded by the chloroplast in *M. polymorpha* (Ohyama et al., 1988) although other chloroplasts encoded transporters have not been identified. No corresponding ABC₁ transporter gene has been identified in the 35 kb molecule although it is possible that this gene may have been transferred to the nucleus.

Attempts to express the complete ORF470 polypeptide in *E. coli* failed, because of lysis of the bacterial cells. In addition, attempts to express the C-terminal region of this polypeptide were unsuccessful (data not shown). These results suggest that overproduction of ORF470 might be deleterious to the cell, implying that expression or translation of this gene product is tightly regulated. The function of this polypeptide *in vivo* has not been determined although ORF470 homologues have been identified as mentioned above (see also Chapter 8). It is hoped that future analysis of the *Synechocystis* PCC 6803 clones identified here by cross-hybridisation, will be a step towards assigning a function to this gene product.
A nucleus encoded *P. falciparum* sequence that hybridised to a region of the yeast mitochondrial *tuf* gene was shown to be located on chromosome 1. I have suggested that this corresponds to the *P. falciparum tuf* gene utilised by the parasite’s mitochondria. Obviously, to confirm this, the gene must be cloned and sequenced. However, if it is found to correspond to a *tufM* gene, more credence would be added to the observations previously made that the 35 kb DNA and 6 kb mitochondrial DNAs do not function in a co-operative manner.

In seeming contradiction to this last statement, there has been one report that the 35 kb molecule and mitochondria might function in a co-operative fashion (Tomavo and Boothroyd., 1995). Differentiation of the *T. gondii* tachyzoite to the bradyzoite form has been suggested to be caused by inactivation or decrease in activity of the parasite mitochondrion. In the presence of drugs targeted to the mitochondrion, including atovaquone, the rapidly dividing, *T. gondii* tachyzoites involved in acute infection differentiated to the slowly dividing bradyzoites. However, two mutants to atovaquone were predisposed to differentiate to bradyzoites and these were found to be hyper-sensitive to clindamycin - clindamycin has been proposed by Pfefferkorn and Borotz (1994), to act on the plastid organelle in *T. gondii*. In correlating these observations, Tomavo and Boothroyd suggested that both the mitochondrial and plastid organelles contribute to the generation of a product “X”. In tachyzoites, “X” is produced equally by both organelles, however, in bradyzoites, the mitochondrial contribution to the production of “X” is substantially reduced whilst the plastid contribution remains unchanged. A decrease in the concentration of “X” would therefore lead to differentiation to a less metabolically active form of the parasite.
This hypothesis correlates the sensitivity of wild type *T. gondii* tachyzoites to both anti-mitochondrial and anti-plastid drugs; hyper-sensitivity of atovaquone resistant parasites to clindamycin; and the lesser degree of sensitivity of bradyzoites to atovaquone than tachyzoites. However to my knowledge, the sensitivity of bradyzoites to clindamycin has not been determined. If this hypothesis was correct, I would assume bradyzoites to be sensitive to clindamycin, presumably to the same degree as the atovaquone-resistant parasites. The identity of “X” is unknown but, it seems unlikely that both the mitochondria and the plastid produce a similar product. Comparison of the genes encoded by the completely sequenced 35 kb and 6 kb DNAs in *P. falciparum* has not identified similar gene products excepting those involved in gene expression.

The retention of the 35 kb molecule in *P. falciparum* and other apicomplexans suggests that it performs a function necessary or useful for the parasite. The majority of the genes encoded by the 35 kb molecule are involved in transcription or translation. It is therefore plausible to suggest that the key to identifying a role for this circular DNA lies in the retention of the clpC-like gene, ORF470, ORF105, ORF 79, ORF101, ORF51, ORF129, ORF78 and ORF91. Since the clpC gene and ORFs -470 and -105 have been identified in other plastid genomes these gene products may be the reason for the retention of this molecule.

10.2.0 Future work

1) With the exception of the tRNAs and rRNAs, the *P. falciparum* 35 kb encoded genes are transcribed at low levels. It is possible that the 35 kb molecule is not
functional at the intra-erythrocytic stages of the *Plasmodium* spp. parasite and the genes encoded by this DNA are therefore expressed at a low level. Expression of these plastid genes should be investigated at other stages in the life cycle (i.e. the sporozoite and gamete stages), although such experiments would be hindered due to the difficulties in culturing or obtaining these forms.

2) Purification of the RNAP and characterisation of this enzyme would enable the determination of which genes are transcribed by this molecule.

3) Determination of the function of ORF470 would contribute to assigning a function to the 35 kb molecule. To this end, analysis of the *Synechocystis* PCC 6803 clones have been identified and should help in assigning a role for this polypeptide.

4) The identification of reading frames bearing similarity to ORF105 in other chloroplasts and organisms should also aid the determination of a function for this plastid in the apicomplexans. Additionally, phylogenetic analysis of these homologues would add to the determination of the ancestor of the 35 kb molecule.
REFERENCES
REFERENCES


References


Hinterberg, K., and A. Scherf. 1994. PFGE: Improved conditions for rapid and high resolution separation of *Plasmodium falciparum* chromosomes. Parasitology Today, 6:225


References


Löffelhardt, W., C. Michalowski, M. Kraus, B. Pfanzagl, C. Neumann-Spallart, J. Jakowitzch, M. Brandter, H.J. Bohnert. 1991. rps10 and 6 other ribosomal protein genes from the S10/spc-operon not encountered on higher plant plastid DNA are located on the cyanelle genome of Cyanophora paradoxa. NATO ASI Ser Ser H., 55:155-165.


Msadek, T., F. Kunst and G. Rapoport. 1994. MecB of Bacillus subtilis, a member of the ClpC ATPase family, is a pleitropic regulator controlling competence gene


References


Siemeister, G., C. Buchholz, W. Hachtel. 1990. Genes for the plastid elongation
factor Tu and ribosomal protein S7 and six tRNA genes on the 73 kb DNA Astasia
longa that resembles the chloroplast DNA of Euglena. Mol. Gen. Genet., 220:425-
432.


1988. Cloning of the genomic fragment from Plasmodium chabaudi expressing a 105

Southern, E.M. 1975. Detection of specific sequences among DNA fragments

Struhl, K. 1989. Helix-turn-helix, zinc-finger and leucine zipper motifs for eukaryotic
transcriptional regulatory proteins. TIBS., 14:137-10.

sequence of a 5.8 kilobase pair repetitive DNA from Plasmodium falciparum. Mol.


Squires, C.L., S. Pedersen, B.M Ross and C. Squires. 1991. ClpB is the

Tamura, K. and M. Nei 1993. Estimation of the number of nucleotide substitutions
in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol.
Evol., 10(3):512-526

3:186-190.
References


AMENDMENTS TO REFERENCES


Phylogenetic analysis of the *rpoB* gene from the plastid-like DNA of *Plasmodium falciparum*

Malcolm J. Gardner \(^a\), Nick Goldman \(^b\), Phillip Barnett \(^a\), Peter W. Moore \(^a\),
Kaveri Rangachari \(^a\), Malcolm Strath \(^a\), Andrea Whyte \(^a\), Donald H. Williamson \(^a\),
Robert J.M. Wilson \(^a\)

\(^a\) Parasitology Division and \(^b\) Laboratory of Mathematical Biology, National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK

Received 28 January 1994; accepted 11 April 1994
Aims and Scope
The journal provides a medium for the rapid publication of investigations of the molecular biology, molecular immunology and biochemistry of parasitic protozoa and helminths and their interactions with both the definitive and intermediate host. The main subject areas covered are: chemical structure, biosynthesis, degradation, properties and function of small molecular weight substances, DNA, RNA, proteins, lipids and carbohydrates - intermediary metabolism and bioenergetics - molecular and biochemical studies on the mode of action of antiparasitic drugs - molecular and biochemical aspects of membrane structure and function - molecular and biochemical aspects of host-parasite relationships including analysis of parasitic escape mechanisms - characterisation of parasite antigen and parasite and host cell surface receptors - characterisation of genes by biophysical and biochemical methods, including recombinant DNA technology - analysis of gene structure, function and expression - mechanisms of genetic recombination.
Papers will only be accepted for publication if they fall within these areas, if they contain original work of high scientific quality, and if they are well presented.

Editors
C.C. Wang,
School of Pharmacy,
Department of Pharmaceutical Chemistry,
University of California,
San Francisco,
CA 94143-0446, USA
Tel.: (415) 476 1321
Fax: (415) 476 0688

Alan H. Fairlamb,
Department of Medical Parasitology,
London School of Hygiene and Tropical Medicine,
Keppel Street,
London WC1E 7HT, UK
Tel.: 071 927 2455
Fax: 071 636 8739

Tony Holder,
Division of Parasitology,
National Institute of Medical Research,
The Ridgeway, Mill Hill,
London NW7 1AA, UK
Tel.: 081 959 3665 ext. 2175
Fax: 081 913 8593

Rick Komuniecki,
Department of Biology,
University of Toledo,
Arts & Sciences,
2801 E. Bancroft St.,
Toledo,
OH 43606-3390, USA
Tel.: (419) 537 2065
Fax: (419) 537 7737

Editorial Board
N. Andrews, New Haven, CT, USA.
S.M. Beverley, Boston, MA, USA.
J.C. Boothroyd, Palo Alto, CA, USA.
P. Borst, Amsterdam, The Netherlands.
J.J. Cazzulo, Buenos Aires, Argentina.
D. Cioli, Rome, Italy.
C.G. Clark, Bethesda, MD, USA.
C.E. Clayton, Heidelberg, Germany.
W. Colli, São Paolo, Brazil.
G.H. Coombs, Glasgow, UK.
A.F. Cowman, Melbourne, Australia.
G.A.M. Cross, New York, NY, USA.
R. Docke, Urbana, IL, USA.
J.E. Donelson, Iowa City, IA, USA.
P.T. Engelund, Baltimore, MD, USA.
M.A.J. Ferguson, Dundee, UK.
W.E. Gutteridge, Beckenham, UK.
S. Hajduk, Birmingham, AL, USA.
K. Haldar, Palo Alto, CA, USA.
B.G. Harris, Fort Worth, TX, USA.
P.J. Johnson, Los Angeles, CA, USA.
D.J. Kemp, Casuarina, Australia.
P. Kühler, Zurich, Switzerland.
R.M. Maizels, London, UK.
J.H. McKeown, San Francisco, CA, USA.
D.P. McManus, Brisbane, Australia.
L.A. McReynolds, Beverly, MA, USA.
P.A.M. Michels, Brussels, Belgium.
M. Müller, New York, NY, USA.
T.W. Nilsen, Cleveland, OH, USA.
A.J. Parodi, Buenos Aires, Argentina.
M. Parsons, Seattle, WA, USA.
E.J. Pearce, Ithaca, NY, USA.
J.V. Ravetch, New York, NY, USA.
T. Seebeck, Bern, Switzerland.
M.E. Selkirk, London, UK.
D.F. Smith, London, UK.
E. Tannich, Hamburg, Germany.
B. Ullman, Portland, OR, USA.
L.H.T. van der Plouw, Rahway, NJ, USA.
R. Walter, Hamburg, Germany.
T. Wellemes, Bethesda, MD, USA.
D. Zilberstein, Haifa, Israel.
Phylogenetic analysis of the \textit{rpoB} gene from the plastid-like DNA of \textit{Plasmodium falciparum}

Malcolm J. Gardner \textsuperscript{a,1}, Nick Goldman \textsuperscript{b}, Phillip Barnett \textsuperscript{a}, Peter W. Moore \textsuperscript{a}, Kaveri Rangachari \textsuperscript{a}, Malcolm Strath \textsuperscript{a}, Andrea Whyte \textsuperscript{a}, Donald H. Williamson \textsuperscript{a}, Robert J.M. Wilson \textsuperscript{a}

\textsuperscript{a} Parasitology Division and \textsuperscript{b} Laboratory of Mathematical Biology, National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK

Received 28 January 1994; accepted 11 April 1994

Abstract

Malaria and other Apicomplexan parasites harbour two extrachromosomal DNAs. One is mitochondrial and the other is a 35-kb circle with some plastid-like features but whose provenance and function is unknown. In addition to genes for rRNAs, tRNAs and ribosomal proteins, the 35-kb circular DNA of \textit{Plasmodium falciparum} carries an \textit{rpoBC} operon which encodes subunits of a eubacteria-like RNA polymerase. The phylogenetic analysis of the complete \textit{rpoB} sequence presented here supports our inference that the 35-kb circle is the remnant of a plastid genome.

Key words: Malaria; Plastid; Phylogeny; \textit{Plasmodium falciparum}; RNA polymerase; \textit{rpoB}

1. Introduction

Protists of the phylum Apicomplexa, which includes malaria parasites (genus \textit{Plasmodium}), contain two extrachromosomal DNA molecules \cite{1,2}. One is a 6-kb sequence which encodes characteristic mitochondrial genes and is presumed to be the mitochondrial genome. The other is a 35-kb double-stranded circular DNA of unknown cellular location and provenance. Attention has been drawn to a resemblance between the 35-kb circular DNA of \textit{Plasmodium falciparum} and plastid DNAs, particularly the ‘plastome’ of parasitic plants such as \textit{Epifagus virginiana}, in that both genomes are highly reduced and comprised largely of genes involved with expression, any photosynthesis-related genes that may have been present having been deleted. Molecular phylogenetic comparisons based on nucleus-encoded small subunit rRNAs suggest that the Apicomplexa are related to the dinoflagellate-ciliate clade, and most probably are closer to the dinoflagellates \cite{3–6}. These studies concur with earlier systematic work that related the Apicomplexa to the dinoflagellates on the basis of their specialised apical organelles and alveolus-like membranous system beneath the plasmalemma \cite{7,8}. Present day photosyn-

---

**Abbreviations:** nt, nucleotide;

**Note:** Nucleotide sequence data reported in this paper have been submitted to the EMBL data base with the accession number X75544.

\textsuperscript{1} Present address: Malaria Program, Naval Medical Research Institute, 12300 Washington Avenue, Rockville MD 20852, USA

0166-6851/94/$07.00 © 1994 Elsevier Science B.V. All rights reserved

SSDI 0166-6851(94)00083-Y
thetic free-living dinoflagellates appear to have resulted from endosymbiotic events involving other photosynthetic eukaryotes [9]. In the postulated evolution of the Apicomplexa from a dinoflagellate-like ancestor, loss of photosynthesis was apparently an early event, possibly concomitant with adaptation to parasitism, such that common ancestry with a photosynthetic form has not hitherto been suspected.

We have previously reported the occurrence on the 35-kb circular DNA from *P. falciparum* of partial sequences of genes apparently encoding β and β' components of a eubacteria-like RNA polymerase [10]. The presence of these *rpoBC* genes on the circular DNA of *P. falciparum* offers strong circumstantial evidence that the circular DNA is derived from a plastid genome. The *rpoBC* operon is a universal feature of plastid DNAs and it could reasonably be expected that the malarial *rpoBC* genes would provide useful phylogenetic clues to the provenance of the 35-kb circle. As will be described below, such studies are complicated by the extreme divergence of the genes, their AT-richness, and the poverty of the database, problems encountered previously in a phylogenetic analysis of a short segment of the malarial *rpoC* gene by Howe [11]. Despite these difficulties, Howe concluded that the malarial circular DNA is of plastid origin.

Here we present a phylogenetic analysis of the complete *rpoB* gene from the 35-kb circular DNA of *P. falciparum*. This shows that the malarial *rpoB* gene groups with plastid rather than eubacterial homologues.

### 2. Materials and methods

**Shotgun library.** Partially purified 35-kb circular DNA was prepared by density gradient centrifugation of total *P. falciparum* DNA, clone C10 [12], as described [13–15]. Approximately 4 μg of circular DNA in 100 μl T4 DNA polymerase buffer (67 mM Tris pH 8.0/ 6.7 mM MgCl₂/ 10 mM mercaptoethanol/ 100 μM dNTPs) was sonicated to produce fragments of 200–1000 bp. After end repair by addition of 100 U of T4 DNA polymerase (Gibco), DNA fragments smaller than approximately 500 bp were removed by addition of an equal volume of 14% polyethylene glycol in 1.0 M NaCl, incubation overnight at 4°C and centrifugation. The precipitate was washed twice with 70% ethanol, air-dried and resuspended in 10 mM Tris HCl/ 0.5 mM EDTA, pH 8.0. The size-fractionated DNA was ligated into Smal-digested pBS KS II (+) and transformed into *E. coli*-SURE (Stratagene). A 32P-labelled PCR product (see below) was used to screen the shotgun library by colony hybridisation [16].

**Polymerase chain reaction amplification.** The 5' end of *rpoB* and its upstream region were amplified by PCR from total genomic DNA using two custom primers (G21 and G22) designed from the sequence of cloned circular DNA[10] and unpublished data: G22, 5' TTACTGAGGCAAATAAGCTGATACGTGGTAG; and G21, 5' ACAGAATTCCGATCCCTAATATTTCTAGGTAATT yielded a product of approx. 2 kb with terminal restriction sites (underlined; *PstI* and *EcoRI*, respectively). Amplification was for 35 cycles at 95°C for 0.5 min, 42°C for 1 min, and 72°C for 3 min in a Hybaid heating block (Model IHB2024). PCR products were separated by electrophoresis in a 1% agarose gel, excised from the gel and purified on a silica matrix using the GENE-CLEAN II kit (BIO 101 Inc., USA).

**Reverse transcription PCR.** Total parasite RNA [17] was pre-digested with DNase for 1 h at 37°C and recovered by phenol/chloroform extraction. An aliquot of the RNA was reverse transcribed at 42°C for 2 h with 200 U MoMuLV reverse transcriptase (Life Sciences) and the primer RT1, 5' CCATTCTAAAATGGAAGTAGAATTGGTAATCCTG, in the presence of the manufacturer’s recommended buffer, 10 mM dNTPs and 40 U RNAsin (Promega). Following heating at 95°C for 5 min, the cDNA was amplified by PCR, 500 ng of the primer RW31, 5' GGGCTTTAAGCTTGTTTGG being added in the recommended buffer containing 2.5 mM MgCl₂ (Promega). Amplification was carried out for 35 cycles with an annealing temperature of 50°C and extension at 72°C. The control contained no enzyme in the reverse transcriptase step.

**DNA sequencing.** DNA sequences were determined by the chain termination method using Sequenase Version 2 (US Biochemicals). Single strand plasmid DNAs were sequenced with forward and/or reverse
M13 primers and custom primers. PCR products separated in low melting point agarose were purified with GENECLEAN as described above, or were excised and annealed with the primer at 70°C in Sequenase reaction buffer and 0.05% (v/v) NP40. After boiling at 100°C for 5 min, samples were plunged to 30°C for 15 s then labelled at room temperature for 45 s. Termination was carried out at 50°C for 1 min. Sequences were compiled using the DBAUTO and DBUTIL programmes (Staden Plus, Amersham).

Primer extension. Total parasite RNA was prepared as described previously [17]. 10 pmol G56 primer (5' TTTCCTGAAATATAAATAATAAATTTTG) was hybridised to 2 μg RNA in 10 μl 80 mM KCl at 42°C for 1 h and cooled to room temperature. In a separate reaction with the oligonucleotide A01 (5' ATATAATACCTTAAATTATAATTTCTCTG), the hybridising mixture was cooled slowly at room temperature from 85°C and held at an annealing temperature of 37°C overnight. Primer extension was performed for 2 h at 37°C with 200 U MoMuLV reverse transcriptase (RNase H-;Promega) in a 20 μl reaction containing the manufacturer’s recommended buffer with 2 mM of each dNTP, 50 U RNase inhibitor, and 10 μCi 32P-labelled dATP (3000 Ci mmol -1 ; Amersham). The reaction was terminated with 0.5 M EDTA, and nucleic acids were extracted with phenol/chloroform, ethanol precipitated, and electrophoresed on a 6% polyacrylamide denaturing gel. Controls included absence of primer and an RNase-treated starting sample. Specific transcripts were analysed by size and by comparison with direct PCR DNA sequence; the specific region was amplified from 35 kb circle template DNA with primer A01 and a second primer G22 corresponding to sequence 900 nucleotides upstream of rpoB.

Phylogenetic analysis. Sequences for comparison with the malarial rpoB gene were obtained from the EMBL database. The following sequences were considered to be relevant and adequate for phylogenetic analysis: Escherichia coli (ecrpoi), Pseudomonas putida (ppropob), Zea mays plastid (czmrnpb), Spinacia oleracea plastid (chompob), Marchantia polymorpha plastid (chmpxx), Euglena gracilis plastid (chegrpo). The amino acid sequences were aligned using the GAP and PILEUP programs of the GCG package, version 7.1 [18], with minor modifications made by hand. Nucleotide sequences of regions considered suitable for phylogenetic analysis were then analyzed by maximum-likelihood methods using the DNA substitution models of Jukes and Cantor [19] and Hasegawa et al., [20], with and without the assumption of a molecular clock. To determine which model best fitted the data, the specific models were compared with each other by a likelihood ratio statistic and were also compared with an ‘unconstrained model’ which had no phylogenetic components: this allowed a measure of the overall adequacy of each specific model [21]. In addition, some results were verified using the neighbour-joining method [22].

3. Results

The rpoB gene and its transcription product. The malarial rpoBC genes are separated from the rRNA inverted repeat on the 35-kb circle by three contiguous, unidentified open reading frames, URF470, URF51, and URF101, as indicated schematically in Fig. 1, where the numerals indicate the number of amino acids in each predicted polypeptide. None of these malarial URFs corresponds to genes flanking rpoBC in other genomes, including chloroplasts (data not shown).

Hybridisation analysis of total RNA extracted from erythrocytic stages of P. falciparum using a
Fig. 2. Alignment of predicted amino acids of the RNP β subunit from representative eubacteria, plastids and P. falciparum. This alignment was used to identify homologous regions (blocks 1–10) for phylogenetic analysis of the nucleotide sequences. Symbols in the rifampicin locus (blocks 3, 4, 5, corresponding to clusters 1, 2 and 3 of Jon and Gross [43]) indicate amino acid identity (triangle), conservative substitution (square), or difference (circle) between P. falciparum and rifampicin-sensitive mutants of E. coli. The E. coli sequence used here is from a rifampicin-resistant strain (DS16V) [37].
probe which encompassed portions of *rpoBC* (the pfS35 probe shown in Fig. 1) revealed large (7-9.5 kb) transcripts [10]. The potential 5' end of the malarial *rpoB* transcript was mapped by two independent primer extension reactions with overlapping oligonucleotide primers (A01 and G56; Fig. 1) lo-
cated approx. 100 nt downstream from the presumptive 5' end of the malarial gene. The extension products were compared with the sequence of a cloned PCR product of the same region of the 35-kb circle. Although some minor extension products within URF101 differed between the two primers, both of them revealed the same major 5' end, which mapped to the spacer region between URF51 and URF101, about 310 nt 5' to the presumed \textit{rpoB} initiation codon (data not shown).

We had earlier predicted from both Northern blots and because of the short intergenic spacer (7 nt), that \textit{rpoB} and C are transcribed polycistronically [10]. This was confirmed by reverse transcription PCR across the junction zone, using primers shown in Fig. 1. cDNA was prepared from total RNA using a primer (RT1) 166 nt 3' to the predicted initiation codon of \textit{rpoC}, and subsequently amplified by PCR using an additional primer (RW31) 416 nt from the 3' end of \textit{rpoB}. A product of the anticipated size (423 nt) was obtained, indicating that at least some \textit{rpoB} transcripts include \textit{rpoC} sequences (data not shown).

\textbf{Analysis of the predicted malarial \textit{rpoB} subunit.} The predicted amino acid sequence of the malaria \textit{β} subunit was aligned with 6 other \textit{β} subunits to identify conserved domains (Fig. 2). Ten regions were conserved in all 7 sequences, and as observed previously [23], the plastid \textit{β} subunits have 8 deleted segments in comparison to \textit{E. coli}. The malarial \textit{β} subunit corresponded closely to the plastid sequences in the number and position of conserved and deleted segments.

A codon usage table for the predicted RNA polymerase (RNAP) \textit{β} subunit is shown in Table 1. Of the 1024 amino acids, 19\% consisted of isoleucine (codons ATA/ATT), 17\% asparagine (codon AAT), 12\% leucine (codon TTA), 9\% lysine (codon AAA), 9\% tyrosine (codon TAT) and 6\% phenylalanine (codon TTT). Taken together, these 6 over-represented amino acids with (A + T)-rich codons made up 72\% of the peptide and strongly biased the (A + T) composition of the open reading frame (88.3\%). The \textit{P. falciparum} \textit{rpoB} gene would encode a basic protein (pl 10.51) of 122.5 kDa resembling a plastid \textit{β} subunit (e.g. the maize subunit, which is 121.6 kDa with a pl of 9.35 [23], more closely than the \textit{β} subunit of \textit{E. coli}, which is 151 kDa with a pl of 4.93).

\textbf{Phylogenetic analysis of \textit{rpoB}.} Seven \textit{β} subunit sequences were aligned using the GAP and PILEUP programs of the GCG package [18]. After minor...
modification by hand, a total of 1053 nt sites were considered suitable for phylogenetic analysis, corresponding to the amino acid residues boxed in Fig. 2. In large regions outside the boxes there was no real evidence for homology across all of the sequences. Consequently most of the regions chosen for analysis correspond closely to the 'conserved' domains identified in various β subunits by earlier workers [23,24]. These sequences were analyzed using maximum likelihood analysis under the Jukes and Cantor (JC) or Hasegawa, Kishino, and Yano (HKY) models of DNA substitution (see Materials and methods). Comparison of the models by the statistical method of Goldman [21] (results not shown) indicated that the HKY substitution model without the assumption of a molecular clock best fitted the data. The tree

![Image](image-url)

Fig. 3. Phylogenetic trees constructed from the nucleotide sequences of rpoB from the eubacteria E. coli and P. putida, the plastids of Z. mays, S. oleracea, M. polymorpha and E. gracilis, and P. falciparum. (a) Preferred tree constructed using the HKY model of DNA substitutions without assuming a molecular clock. Circled numbers indicate relative content of (A + T) nucleotides. Branch lengths are indicated. (b) Rooted tree, assuming a molecular clock. Eubacteria, from which plastids are believed to have evolved, were chosen as the outgroup. Branch lengths are indicated. (c) Unrooted tree constructed from the second codon positions. Branch lengths are indicated. (d) Unrooted tree constructed from the amino acid sequences using the neighbour-joining method. Distances computed from $-\ln(1 - K)$; see text for further details.
inferred under this preferred model is shown in Fig. 3a. When a molecular clock was assumed, permitting a root to be inferred (as shown in Fig. 3b), this separated the bacterial sequences from all others, confirming the assumption that \textit{P. falciparum rpoB}, like \textit{Euglena} chloroplast \textit{rpoB}, diverged from the land plants after divergence from the bacteria. The eubacterial sequences were used as the outgroup to root the tree because it is generally accepted that plastids originated from eubacteria [25]. Although the HKY model without a molecular clock was found superior to any other model tested, it was rejected in Goldman’s method when compared with the ‘unconstrained model’, which has no components that depend on models of DNA substitution processes or evolutionary relationships. Hence inferences based on the HKY model must be treated with caution. The factors responsible for rejection of the DNA substitution models were not clear. However, it led us to evaluate the potential bias introduced by the different base compositions of the sequences. All available models assume that the composition of the sequences for comparison are similar, and also that the rates of nucleotide substitutions are uniform across all lineages, i.e. they are not directionally biased. Lockhart et al., [26] maintained that phylogenetic analyses ignoring violations of these assumptions may tend to group together unrelated sequences of similar \((A + T)\) content. In the present case, the \((A + T)\) contents ranged from 0.422 (\textit{P. putida}) to 0.803 (\textit{P. falciparum}). Nevertheless, inspection of Fig. 3a shows that the malarial sequence, with the highest \((A + T)\) content, was positioned between two groups with lower \((A + T)\) content. This suggests that the analysis was not dominated by simple \((A + T)\) bias. To confirm this interpretation, two additional analyses were performed, one using the second codon position only, and the second using the amino acid sequences themselves. It was noticeable that base compositional bias was much less at the second codon position where \((A + T)\) contents ranged from 0.621 (\textit{Z. mays}) to 0.738 (\textit{P. falciparum}). Accordingly, the 351 (1053/3) second codon positions available were analysed separately, giving the tree shown in Fig. 3c. Little difference from the earlier analyses resulted except that the \textit{P. falciparum rpoB} sequence was now more firmly placed amongst those of plastids. Thus we found no evidence that the original tree was strongly affected by variation in \((A + T)\) content or base substitution bias. Finally, the seven aligned sequences of 351 amino acids were analysed by the neighbour-joining method [22]. Two measures of the distances between sequences were used, either the proportion ‘\(K\)’ of different amino acids, and \(-\ln(1 - K)\) (see Saitou and Nei [22] and Iwabe et al., [24] for details). In both cases, the tree inferred (e.g. Fig. 3d) was very similar to the original tree (Fig. 3a) with minor differences in the relative branch lengths. Thus our examination of possible sources of bias due to the extreme \((A + T)\) content of the malarial sequence led us to have increased confidence in the results from the original best substitution model (HKY model without molecular clock), which suggested that the \textit{P. falciparum rpoB} sequence is more closely related to plastids than to eubacteria.

4. Discussion

In eubacteria, all classes of RNA are transcribed by a single RNA polymerase (RNAP), a multimeric enzyme with a subunit core composition of \(\alpha_2,\beta,\beta'\) [27]. Chloroplast DNA encodes homologous subunits \((\alpha_2,\beta,\beta' + \beta'')\) which are expressed to form a RNAP active in the transcription of chloroplast-encoded tRNAs and mRNAs [28,29]. There are strong indications that chloroplasts also contain a second RNAP that is nucleus-encoded [30,31]. This imported RNAP has been invoked to explain the transcription of plastid genes in the parasitic plant \textit{Epifagus virginiana}, whose truncated plastid genome has lost its \textit{rpo} genes [32]. In addition, plastids of mutant (white) barley plants, which do not contain plastid ribosomes and consequently are unable to express plastid-encoded polypeptides, nevertheless contain abundant \textit{rpo} transcripts, presumably produced by an imported RNAP [33]. The fact that the \textit{rpoBC} operon has been conserved in the malarial ‘plas­tome’, despite the latter’s extreme reduction in size, implies that it is functional. The \textit{rpoB} gene we have sequenced appears to be complete, and has a plastid-like organisation. However, the transcriptional activity of this presumed polymerase remains to be demonstrated.

A portion of the malarial \textit{rpoC} gene, correspond-
ing to rpoC1 of plastids, has also been sequenced [10]. The results of an analysis of this fragment, using a variety of methods, supported our suggestion that *P. falciparum* *rpo*BC originated from a photosynthetic lineage [11]. This analysis was based on a very limited data set but has been confirmed by our analysis of the complete *rpo*B gene. We add the proviso that current models of nucleotide substitution do not fit the data well in a likelihood ratio test [21], hence the results must be treated with caution until more satisfactory models are developed or better data sets are available. The *P. falciparum* *rpo*B sequence is remote from the only other protist plastid sequence available, that of *E. gracilis*, a protist that diverged early. Thus sequences from species nearer the inferred branching point for *Plasmodium* are required to break up the long unintersected branch on which it is presently placed. Little light has been shed on this phylogenetic problem by analysis of the small and large subunit rRNA genes on the 35-kb circular DNA of *P. falciparum* (Gardner et al., [17,34] and M.W. Gray, personal communication). With both of these rRNA genes any ‘apparent’ plastid link was rendered tenuous by the remoteness of the diverged malarial sequences. On the available evidence, the introduction of a photosynthetic lineage into the ancestry of the Apicomplexa, an ancient phylum now composed entirely of parasitic organisms, is most readily explained if they are viewed as a highly derived, non-photosynthetic group related to a photosynthetic dinoflagellate progenitor.

Investigation of the 35-kb circle has led to a completely new insight into the evolutionary history of the malaria parasite that may also have practical implications. Specifically, does this genome control a specialised organelle or metabolic pathway that can be exploited for chemotherapeutic purposes? It still is not known whether or not the antimalarial activity of rifampicin [35] is due to inhibition of an RNAP containing the *β* subunit encoded by the 35-kb circle, as we once proposed [10]. We have pointed out elsewhere that the rapid and profound inhibition of malarial nucleic acid and protein synthesis induced by rifampicin in vitro suggests a target other than the organellar RNAP in the parasitised cell [36]. It also has been noted that the cluster I region in *P. falciparum* *rpo*B has a substitution which confers rifampicin resistance (H526Q) in *E. coli* [37]. But this still leaves unanswered the possibility that the parasite RNAP is rifampicin-sensitive.

Our results suggest the *P. falciparum* *rpo*B gene is functional, for several reasons. It seems most unlikely that such a large open reading frame (1024 codons) would have been maintained intact without positive selection in the face of (a) a massive reduction in complexity of the organellar genome (from 150 kb in a ‘typical’ plastid genome to 35 kb in *Plasmodium*), and (b) the acquisition of extensive sequence divergence (only 44% amino acid identity with the conserved ‘core’ of *rpo*B in *Euglena gracilis*, its nearest neighbour in our analysis). Moreover, we have shown that the circular DNA of *Toxoplasma gondii* [38] has at least two of the basic features of the malarial circle, namely an inverted repeat comprising rRNA genes and a homologue of the *rpo*BC operon [39]. Yet phylogenetic analysis of nucleus-encoded small subunit rRNA sequences has shown that whilst *Plasmodium* and *Toxoplasma* are monophyletic they are no longer closely related [3,6,40].

We hope, ultimately, through molecular and biochemical approaches, to elucidate the metabolic function of the circular genome. As in the work of Palmer and his colleagues on the parasitic plant *E. virginiana* [41], we predict that the circular DNA in the Apicomplexa will encode at least one functional protein. Recently we have found a highly conserved plastid-like gene outside the *rpo* operon which is a candidate for such a role [42]. A phylogenetic analysis of this gene within the Apicomplexa would clearly assist in indicating whether or not it too is functional.

Acknowledgements

This work was funded in part by the UNDP/WORLD BANK/WHO Special Programme for Research in Tropical Diseases (TDR).

References


