LATERAL GENE TRANSFER FROM BACTERIA TO PROTISTS

by

AUDREY PATRICIA DE KONING

B.Sc., The University of Northern British Columbia, 1999

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Genetics Graduate Programme

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 2002

© Audrey Patricia de Koning, 2002
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.
ABSTRACT

The advent of genome-level sequencing has brought a wealth of genetic information that can be used to trace the evolution of organisms. Phylogenies based on homologous proteins often do not agree on implied species relationships, indicating that the evolutionary path taken by some of these genes differs from the evolution of the organisms. One explanation for this observed pattern is lateral gene transfer, the movement of genes across species boundaries. There are mechanisms to facilitate such movement in prokaryotic organisms. In eukaryotes, it is less clear what mechanisms might facilitate incorporation of foreign genes into a species' genome. Multicellular eukaryotes are presumably more resistant to lateral gene transfer, because the majority of their cells do not have access to genes from other organisms. Moreover, germ and somatic cell lineages are often separate, which greatly reduces the probability that any gene that gains access to a foreign cell will be heritable and subsequently become part of its host species' genome. Phagotrophic protists, however, represent one type of eukaryote where these restrictions do not exist. Such organisms commonly ingest bacteria as food, and chance events may lead to the incorporation and use of bacterial genes. Lateral gene transfer from organelles of symbiotic origin (the chloroplast and mitochondrion) to the nucleus of the host organism represents a specialized case of lateral gene transfer resulting from phagotrophy. Lateral gene transfer in eukaryotes, therefore, is expected to involve bacterial genes incorporated into the genomes of phagocytic species (or those that have evolved from them). Accordingly, genes of bacterial origin were sought among the genetic information available from protists. Over 2000 predicted genes from the genome-sequencing project for the malarial parasite, Plasmodium falciparum, were
screened to find those that are highly similar to genes from bacterial species. Phylogenetic reconstruction methods were used to elucidate the evolution of ‘bacteria-like’ genes. Thirteen laterally transferred genes were identified on the basis of protein phylogeny, of which nine are probably transferred from symbiotic organelles and three appear to involve transfers with bacteria. The conclusion is that genes introduced to the genome through lateral gene transfer are not nearly as prevalent in this eukaryote as they are in bacteria.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
<tr>
<td>PREFACE</td>
</tr>
<tr>
<td>CHAPTER 1</td>
</tr>
<tr>
<td>1.1</td>
</tr>
<tr>
<td>1.2</td>
</tr>
<tr>
<td>1.2.1</td>
</tr>
<tr>
<td>1.2.2</td>
</tr>
<tr>
<td>CHAPTER 2</td>
</tr>
<tr>
<td>2.1</td>
</tr>
<tr>
<td>2.1.1</td>
</tr>
<tr>
<td>2.1.2</td>
</tr>
<tr>
<td>2.2</td>
</tr>
<tr>
<td>2.2.1</td>
</tr>
<tr>
<td>2.2.2</td>
</tr>
<tr>
<td>2.2.3</td>
</tr>
<tr>
<td>2.2.4</td>
</tr>
<tr>
<td>2.2.5</td>
</tr>
<tr>
<td>2.2.6</td>
</tr>
<tr>
<td>2.3</td>
</tr>
<tr>
<td>2.3.1</td>
</tr>
</tbody>
</table>
2.3.2 Phlyogenetic analysis of individual 'bacteria-like' genes ........31

2.3.2.1 Type II topoisomerase ...........................................33
2.3.2.2 Adenylosuccinate lyase ...........................................40
2.3.2.3 Elongation factor G ...............................................46
2.3.2.4 Ribosome release factor .........................................52
2.3.2.5 Proteasome subunit HslV ........................................54
2.3.2.6 Pseudouridine synthase .........................................58
2.3.2.7 Ribosomal RNA methyltransferase ............................61
2.3.2.8 RNA 3' terminal phosphate cyclase ............................65
2.3.2.9 Ribosomal RNA adenine dimethylase ..........................68
2.3.2.10 S-adenosyl methionine-dependant methyltransferase ....71
2.3.2.11 GTP-binding protein .............................................74

2.3.3 Conclusion: Lateral gene transfer in P. falciparum ..........78

BIBLIOGRAPHY ..................................................................80
LIST OF TABLES

Table 2.1 Protein sets used for BAEwatch comparison.......................... 21
Table 2.2 Complete genomes queried for homologues of P. falciparum genes .... 24
Table 2.3 Summary of BLAST searches with P. falciparum sequence sets. ....... 30
Table 2.4 List of P. falciparum genes evaluated by phylogenetic analyses........ 32
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Directional transmission of genetic information</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>Example of a divergent region in an alignment</td>
<td>27</td>
</tr>
<tr>
<td>2.2</td>
<td>Architecture of type II topoisomerase</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>Phylogeny of the B subunit of type II topoisomerase</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Phylogeny of the A subunit of type II topoisomerase</td>
<td>37</td>
</tr>
<tr>
<td>2.5</td>
<td>Phylogeny of type II topoisomerase</td>
<td>39</td>
</tr>
<tr>
<td>2.6</td>
<td>Phylogeny of adenylosuccinate lyase</td>
<td>42</td>
</tr>
<tr>
<td>2.7</td>
<td>Multiple sequence alignment of adenylosuccinate lyase</td>
<td>43</td>
</tr>
<tr>
<td>2.8</td>
<td>Phylogeny of adenylosuccinate lyase 'mixed' clade</td>
<td>45</td>
</tr>
<tr>
<td>2.9</td>
<td>Phylogeny of elongation factor G</td>
<td>48</td>
</tr>
<tr>
<td>2.10</td>
<td>Multiple sequence alignment of elongation factor G</td>
<td>50</td>
</tr>
<tr>
<td>2.11</td>
<td>Phylogeny of ribosome release factor</td>
<td>53</td>
</tr>
<tr>
<td>2.12</td>
<td>Multiple sequence alignment of protease subunit HslV</td>
<td>56</td>
</tr>
<tr>
<td>2.13</td>
<td>Phylogeny of protease subunit HslV</td>
<td>57</td>
</tr>
<tr>
<td>2.14</td>
<td>Phylogeny of RsuA pseudouridine synthase</td>
<td>59</td>
</tr>
<tr>
<td>2.15</td>
<td>Phylogeny of RNA m^5^C methyltransferase</td>
<td>63</td>
</tr>
<tr>
<td>2.16</td>
<td>Phylogeny of RNA 3' terminal phosphate cyclase</td>
<td>66</td>
</tr>
<tr>
<td>2.17</td>
<td>Schematic alignment of rRNA adenine dimethylase sequences</td>
<td>69</td>
</tr>
<tr>
<td>2.18</td>
<td>Phylogeny of ribosomal RNA adenine dimethylase</td>
<td>70</td>
</tr>
<tr>
<td>2.19</td>
<td>Schematic alignment of SAM-dependant methyltransferase</td>
<td>72</td>
</tr>
<tr>
<td>2.20</td>
<td>Phylogeny of SAM-dependant methyltransferase sequences</td>
<td>73</td>
</tr>
<tr>
<td>2.21</td>
<td>Schematic alignment of GTP-binding protein sequences</td>
<td>76</td>
</tr>
<tr>
<td>2.22</td>
<td>Phylogeny of GTP-binding proteins</td>
<td>77</td>
</tr>
</tbody>
</table>
Lateral gene transfer has of late, become a much-discussed topic in evolutionary genetics. The nature of these dialogues is, more so than most other biological debates, often hotly adversarial. I think this is because evidence of lateral gene transfer is almost always equivocal, and arguments supporting or refuting cases of lateral transfer inevitably invoke parsimony and rely on an opinion of how 'likely' a proposed series of evolutionary events is. For example, in response to the claim of massive lateral transfer from Archaea to the eubacterium *Aquifex*, Kyrpides and Olsen [1] state "...vertical inheritance via a thermophilic lineage from the archaeal-bacterial common ancestor will be a more parsimonious explanation than independent lateral transfers as suggested by Aravind et al.." Aravind et al. [2] then reply, "...one is forced to postulate that repeated loss of archaeal genes occurs during the evolution of Bacteria. This is not impossible but does make the hypothesis of Krypides and Olsen less parsimonious.” Unfortunately for those of us interested in lateral transfer, differences in a 'gut-feeling' opinion are hard to reconcile.

Over the last 10 years, examples have accumulated of genes (and even larger genetic units) whose evolution may be explained by lateral transfer. Many researchers have celebrated this somewhat surprising revelation, and are making some rather broad claims of the importance of lateral gene transfer. According to de la Cruz and Davies [2], “genes have flowed through the biosphere, as in a global organism”, which leads to the claim by some that phylogenetic classification may not be a useful concept because of rampant lateral gene transfer [3-5]. Gene transfer has been hypothesized to drive the formation of bacterial operons [6] and to maintain the universality of the genetic code [7]. Several researchers have put forward the idea
that lateral gene transfer has played an integral role in bacterial speciation [8-10]. It has even been suggested that lateral transfer is more important than mutation in the driving the evolution of novel functions [11]. In response to this seemingly blanket acceptance of the importance and contribution of lateral gene transfer to evolution, researchers who were not quite so convinced of its ubiquity began to present dissenting views and advocate that caution be used both in inferring cases of lateral transfer and in interpreting the evolutionary implications of lateral transfer [12-14]. Finally, in the same way that an interesting new song on the radio becomes irritating when overplayed, the tide of opinion is turning against gene transfer. Advocates of lateral gene transfer are now often greeted with cynicism, as illustrated in a recent quip from C.G. Kurland [15]; “Nothing in science is more self-aggrandizing than the claim that ‘all that went before me is wrong’ except perhaps the clam that ‘I can explain everything’. Proponents of global HGT (horizontal gene transfer) have it both ways.” This thesis attempts to provide a balance view of lateral gene transfer by keeping an open mind to all evidence, but still challenging all conclusions with scepticism.

-AD
1.1 Defining Lateral Gene Transfer

The fundamental meaning of lateral gene transfer is the movement of genes between genomes in a direction other than vertical - i.e. genes that are not acquired through inheritance from an immediate ancestor. Such lateral gene movement uncouples the acquisition of genetic information from reproduction. This has several consequences. First, genetic information may move between genomes with widely disparate gene content. Second, while reproduction involves the transfer of an entire genome from parent to offspring, lateral transfer can move much smaller units of information between organisms. Thus an organism that has incorporated laterally transferred information into its genome is largely unchanged, because that organism has a copy of all or nearly all of its parents genes, and is much less similar to the donor of the laterally transferred information. Finally, vertical movement of genetic information must occur with the creation of every new organism, and so most members of a species participate every generation. Lateral movement does not involve 'everybody, all the time,' and may occur sporadically in only a small subset of organisms.

Genetic information moves purely vertically in clonal reproduction (figure 1A). Prokaryotes, asexual single-celled eukaryotes, and the somatic cells of multicellular eukaryotes reproduce clonally through binary fission. Each newly created cell is genetically identical to its progenitor, barring any variation caused by mutation. In the prokaryotic world, parasexual processes such as conjugation, transduction, and transformation, also generate genetic variation (figure 1B). Unlike clonal
reproduction, parasexual processes cause lateral gene movement, which occurs with only some genes, between only some individuals, some of the time. In eukaryotes, variation can be generated by recombination facilitated through sexual reproduction, which involves the vertical transmission of genetic information from two parents (figure 1C). In contrast with clonal reproduction, which creates assemblages of distinct genetic lineages, sexual reproduction mixes parental genetic information in the offspring and creates a gene pool. There are boundaries to the gene pool however, and sexual reproduction generally involves mechanisms that prevent more distantly related individuals from contributing. These gene pool boundaries define what we typically call a species. This ‘sex-centric’ concept of species is characterized by wholesale chromosomal recombination and is more difficult to apply to the localized recombination caused by parasexual processes in the prokaryotic world. There are some parallels, however, as recombinatorial exchange in prokaryotes also has some limits based on degree of relatedness. Recombination is mediated by the RecA enzyme whose efficiency drops sharply as sequence identity decreases. Thus, recombination is more likely and more frequent among more closely related individuals [16]. As in sex-generated gene pools, it is probably these limits to recombination between distantly related prokaryotes that preserves the associations of certain microbiological characters (e.g. appearance, presence/absence of metabolic pathways, etc.) that traditionally define prokaryotic species.

Occasionally genetic information crosses species boundaries. If it moves vertically by sexual reproduction, we call the process “hybridization” (figure 1D). Typically only closely related (recently diverged) species will hybridize, and so this
Figure 1.1 Directional transmission of genetic information. In the top panel, dots depict individuals, while in the bottom panel dots represent species. (A) Genomic information moves vertically through clonal reproduction. (B) Parasexual processes move genetic elements laterally. (C) In sexual reproduction, each parent vertically transmits a copy of its genome. (D) Hybridization moves information vertically, outside of species boundaries. (E) Lateral gene transfer moves information between organisms, independent of inherited information.
sort of genetic transmission does not strictly cross species boundaries but rather blurs them. In both prokaryotes and eukaryotes, genes may move laterally between species (figure 1E), either introducing novel genetic information to a species (akin to mutation in an individual) or replacing existing information (similar to recombination). In eukaryotic organisms, this sort of lateral transmission does not blur the species boundary, because vertical transmission of information through clonal and/or sexual reproduction happens with more information (the whole genome), more often (every generation), and with less divergent information (from within the species). In prokaryotes, lateral gene transfer between species occurs right alongside the lateral transmission that contributes to within-species recombination. As such, lateral transfer in prokaryotes probably exists along a continuum: frequent lateral transfer occurs between closely related individuals and causes variation through recombination within the species, less frequent transfer between members of closely related strains and (by analogy to hybridization) blurs species boundaries, and occasional lateral gene transfer occurs between more distantly related species.

Although in prokaryotic systems parasexual processes move genes laterally within a species, the term Lateral Gene Transfer (LGT) will be used specifically to describe the transfer of genetic information outside of normal species boundaries. Generally, LGT is used to describe lateral movement of any unit of genetic material including small segments [17], genes [18], operons [19], and even whole genomes [20].
1.2 Mechanisms that cause lateral gene transfer in eukaryotes

The following sections detail what we know about possible mechanisms for lateral gene transfer in extant eukaryotic organisms. It must be remembered that conditions may have been very different in the past. For example, while viruses that can infect organisms from different kingdoms are unknown today, they may have existed in the past. Going further back in time, Woese and co-workers have put forth the idea that some of the evidence of lateral gene transfer that we see today results from gene movement at the time of the last universal common ancestor. They note that simple organisms would have had very few barriers to gene exchange, and would represent a 'lateral exchange gene pool' [21, 22].

1.2.1 Acquisition of foreign genetic material through endocytosis

DNA cannot simply diffuse across cell walls and membranes, and so genetic material must have some route of access into a new host. In eukaryotes, genetic material can be imported into cells by endocytosis, in which invaginations of the cell membrane pinch off to form internal vacuoles [23]. Once the phagocytic vacuole is internalized, it generally fuses with a lysosome and the incoming microorganism is digested [24]. Heterotrophic unicellular eukaryotes do this for a source of nutrients, and bacteria-eaters are common in the protist world. Multicellular organisms have a variety of specialized cells (e.g. monocytes, macrophages and neutrophils) that phagocytose microorganisms, but these are generally components of an immune response that targets foreign cells.

Several microorganisms have ways to escape degradation once inside the cell, and this can lead to longer associations with eukaryotic cells. For instance, some
mycobacteria, *Legionella pneumophila*, and the protist *Toxoplasma gondii* can inhibit fusion of the phagosome with the lysosome. *Listeria monocytogenes*, *Shigella flexneri* (bacteria) and *Trypanosoma cruzi* (eukaryote) can lyse the phagosome and escape into the cytosol. *Leishmania spp.* (a protist) and *Salmonella typhimurium* (bacterium) can simply inhabit the phagolysosome and seem to resist digestion [25-27]. There are numerous instances of intracellular microorganisms that can replicate in the vacuole or cytosol and inhabit eukaryotic cells; such associations may be either transient or stable [28]. Both parasitic and commensal relationships exist; *Chlamydia*, *Rickettsia*, and *Legionella* are intracellular bacteria that cause disease in humans, but live mutualistically with Amoebae [29]. The most striking example of a long-term association is that of the eukaryotic organelles of symbiotic origin. These include plastids evolved from a cyanobacteria-like ancestor and mitochondria derived from an alpha-proteobacterial progenitor [30]. It has been estimated that these obligate associations began over 1.5 billion years ago for the mitochondria [31] and only slightly later for plastids. Some protist groups contain secondary plastids. These organelles are derived from endosymbiosis of plastid-containing eukaryotes (red or green algae) [30]. The existence of these secondary symbionts shows that uptake of organisms in eukaryotes is not limited to bacteria, and that internalization through endocytosis can import a source of eukaryotic genes into eukaryotic cells.

Microorganism-containing vacuoles are eventually digested in the host eukaryote, and it is at this point that DNA from the lysed microorganism may escape and gain access to the host nucleus. Organelles, too, are digested by lysosomes (in this case by specialized 'self'-targeting lysosomes), and some studies have revealed that liberated organellar DNA can end up in the nucleus of the host [32]. Escape of
intact DNA from lytic vacuoles probably occurs accidentally and so may not be a frequent source of foreign DNA for lateral gene transfer. However, over evolutionary time, it may have occurred repeatedly. In the case of bacteria ingested for food, repeated uptake and digestion provide a constant source of DNA-containing vacuoles. For organelles, mutualists, and parasites, continued presence in the eukaryotic host (with accompanying deaths and reproduction) ensures that opportunity after opportunity occurs for exposure to foreign DNA[33].

There are wide ranges of both prokaryotic and eukaryotic organisms that can be the donors of DNA acquired by phagocytosis, and the only real limitation on who can be internalized is size - the host cell is generally larger than its prey. As well, since DNA acquisition depends on contact between two cells, we would expect that lateral gene transfer events involving this mechanism would only occur between organisms that occupy a similar milieu. The eukaryotes that can acquire genes in this manner are largely restricted to single-celled organisms. Among the multicellular eukaryotes, cell specialization leaves only restricted set of cells with the ability to phagocytose. In animals, these are the assorted macrophages that function in the immune responses [34]. In plants, only specialized nodule cells in the root hairs of legumes are known to bring in organisms through phagocytosis, and they only internalize their own species-specific rhizobial symbiont to facilitate nitrogen fixation [35]. Neither of these are relevant to lateral gene transfer, as these cells are somatic, and only transfers to germ-line cells will be inherited.
1.2.2 Introduction of foreign genes through viral infection

A virus is an infectious element that, when independent of a host cell, exists as a nucleic acid genome enveloped in a capsid coat. They represent a potential conduit for lateral gene transfer because they are specialists at slipping in and out of cells, and can move between organisms. Moreover, viruses replicate intracellularly, often in intimate contact with their host genome. If genetic material from the host is incorporated into the viral chromosome, the virus will transfer genes from its last host into its next host.

Unlike endocytotic mechanisms for lateral gene transfer, viral transduction obviates the need for donor and recipient cells to be in close proximity. The capsid protects the viral genome from degradation and allows the virion to be widely dispersed [36]. Consequently, viruses can facilitate the dissemination of genes between organisms separated in space or time.

There is some evidence that eukaryotic viruses can incorporate host genes into their own genomes. The sequencing of several viral genomes has identified ORFs that have high similarity to genes in these viruses’ hosts, and this is proposed to be evidence of common origin [37]. Tidona and Darai [38], noting that the viral copies of eukaryote genes do not contain introns, suggested that DNA viruses might capture host genes through either reverse transcription of host mRNAs followed by recombination or by recombination of viral and host DNA and subsequent deletion of introns. Becker [37] proposed that single-stranded RNA viruses, which replicate in the host cytoplasm, could pick up host genes by recombination with host RNA. Unfortunately, the means by which host genes might end up in viral genomes have not been investigated for most eukaryotic viruses.
Mechanisms of host gene capture have been studied the most in the retroviruses. Free retrovirions contain diploid RNA genomes that are reverse-transcribed into haploid DNA upon infection. The viral DNA inserts itself into the host genome, with the help of a virus-encoded enzyme, integrase. The virus then uses the host cell machinery to transcribe retroviral mRNAs and to process, splice and translate the transcripts. Two full mRNA transcripts (and some viral proteins) are packaged into viral particles that bud from the eukaryotic host to regenerate infective free virions [39]. Errors in viral transcription, reverse-transcription, and/or packaging may lead to production of viruses with host genes. A mutant retrovirus whose genome lacks a packaging recognition site was shown (in a xenologous cell line) to pack host RNAs instead, leading to the production of pseudogenes in the genome of subsequently infected cell host [40-42].

There is also some evidence that random host mRNAs are occasionally packaged along with viral RNA to create the diploid RNA of the free virion. Upon reinfection, reverse-transcriptase (RT) can jump between the templates at short stretches of homology in otherwise unrelated RNAs, causing recombination with the host gene between the regions important for viral infectivity. This forms proviral DNA that has integrated a cellular gene into its genome [43, 44]. Viral/host fusion mRNAs are created by accidental read-through transcription [45]. The viral/host fusion mRNAs are packaged into virions along with a normal viral genomic RNA transcript and as described above, RT mediated recombination regenerates the infective virus [46].

While retroviruses-mediated gene transfer has been observed (e.g. [47-49]), these transfers are between cells of the same species. There are not yet any known examples of cross-species lateral transfer by retroviruses [50].
To detect virus-mediated lateral gene transfer in nature, the new sequence must cause some observable change in the recipient. The best evidence for retroviral gene transfer is the existence of viruses that can induce tumours upon infection because they carry oncogenes derived from cellular host genes [47-49]. Most oncogenic retroviruses are replication defective because they lack essential portions of the viral genome [51, 52], and thus only replicate if wild-type viruses are present in the same cell to provide viral products in trans. Oncogenic retroviruses do not induce significant numbers of naturally occurring tumours; this is presumably because oncogene capture is a rare event that produces viruses with a defective phenotype [53].

Among higher eukaryotes, viral infection can only lead to lateral gene transfer if genes enter germ cells. There is some evidence that this can happen. Viral infections of the reproductive tract are observed in mammals [54]. Moreover, endogenous retroviruses (ERVs) are found in the genomes of all higher eukaryotes in which they have been sought [55]. ERVs are integrated retroviruses that are vertically inherited by hosts in a Mendelian fashion [50]. Exogenous retroviruses and ERVs are very similar in sequence and genome organization and thus are thought to have originated from rare infections of germ cells [50, 55, 56].

For viruses to cause lateral gene transfer through movement of genes across species boundaries, they must be able to infect more than one species. Infection begins with an interaction of the free virion with a host cell surface receptor. A great many viruses are highly specific for their receptors and infect only a narrow range of hosts [57]. In general, viruses do not cross generic boundaries [36], and host gene movement primarily leads to horizontal transfer between close relatives. However,
some broad-host-range viruses have been found. The bacteriophage PRD1 can infect and unusually broad range of hosts, including multiple genera of both gram-positive and proteobacterial groups [58]. There are also some broad-host-range eukaryotic viruses; for example the Borna Disease virus occurs in at least all warm-blooded animals [59]. Furthermore, a study by Jensen et al. [57] suggests that the perception that phages are extremely host-specific may be an experimental artefact. The standard method of virus isolation involves serial rounds of viral growth on one host, which may select for those viruses with the highest host specificity. When Jensen et al. [57] incubated environmental samples on a mixture of two potential hosts, they isolated several phages (bacterial viruses) capable of infecting multiple host species. However, no viruses are known that can infect both bacteria and eukaryotes.
CHAPTER 2  LATERAL GENE TRANSFER IN A PROTIST GENOME

2.1 Introduction

2.1.1 Genomic evidence of lateral gene transfer

As detailed in section 1.2, studying the mechanisms of lateral gene transfer can give us some indication of the likelihood of transfer events. However, such events introduce an allele or gene that may only have a transient association with the recipient organism and some of its descendants. To elucidate the contribution of lateral gene transfer to the evolution of a species, it is more relevant to focus on those genes that become a long-term component of the genetic repertoire of a species. Several recent studies have attempted to do this by measuring what proportion of a species' genome has arisen through the lateral transfer of genes, as opposed to de novo creation of new genes through mutational evolution (e.g. [60-63]. These studies have given widely varying estimates, which may in part reflect differences in methodology used for detection (see [64-66] for discussions of anomalies caused by method). Moreover, there is no reason to suppose that all organisms will have been affected by lateral transfer to the same degree. Differences in the mechanisms that cause lateral transfer in different organisms will certainly contribute to observed inequalities. As well, the ecology and lifestyle of a particular species will impact the amount of foreign DNA to which an organism is exposed. For example, we would expect the genomes of organisms that live in interdependent microbial communities, such as those found in a rumen, to have a larger percentage of laterally transferred genes than pelagic marine microbes. Indeed, some trends are
emerging in genome-based studies of lateral transfer among eubacteria. While 10-15% of the genomes of some bacteria, such as *E. coli* and *Synechocystis*, are estimated to be laterally acquired, host-dependant endosymbionts and endoparasites, such as *Buchnera* [67] and *Rickettsia* [9]), show no evidence of transfer. This is presumably because the intracellular lifestyle of these organisms precludes contact with other bacteria [67]. It remains to be seen whether this particular trend continues to hold as more genomes are analyzed.

Genomic-based studies of lateral gene transfer have focused mostly on transfers between prokaryotes, because the large majority of sequenced genomes are from prokaryotes. Much less is known about the contribution of lateral transfer to the evolution of eukaryotic species. Only seven eukaryotes have been fully sequenced to date: a vertebrate (*Homo sapiens*), three fungi (*Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Encephalitozoon cuniculi*), two non-vertebrate animals (*Caenorhabditis elegans* and *Drosophila melanogaster*) and a plant (*Arabidopsis thaliana*). The Arabidopsis Genome Initiative [20] carried out a search for all *Arabidopsis* nuclear genes with highest pairwise similarity (based on BLAST scores) to a bacterial gene. The vast majority of the genes identified in this way were similar to cyanobacteria and probably reflect transfer of genes from the chloroplast. The International Human Genome Sequencing Consortium [68] identified 113 human genes as being of probable bacterial origin. Subsequently, more rigorous analyses [69-71] showed that this was a rather large overestimate of the proportion of laterally transferred genes in the human genome. Even so, the original estimate represents only 0.4% of the human genome, which is much less than estimates in many prokaryotic genomes. The paucity of laterally transferred genes is hardly surprising.
As discussed in section 1.2, several biological features of higher eukaryotes provide extra hurdles to the establishment of a laterally transferred gene in a species' genome. Animals, plants, and fungi comprise only a small subset of the diversity of eukaryotes, and so the current sample of sequenced genomes is not representative enough to evaluate the contribution of lateral transfer to eukaryotic evolution.

The objective of this study was to determine the extent of laterally transferred genes in a protistan eukaryote. I chose to analyze the available sequence data from a genome-sequencing project currently underway: that of the nuclear genome of an alveolate, *Plasmodium falciparum*. This genome is being sequenced by a random shotgun method, and so although it is not complete, the set of genes that are available are a more-or-less unbiased sample. I will begin by reviewing salient facts about *P. falciparum* and its genome.

### 2.1.2 *Plasmodium falciparum*

*Plasmodium falciparum* is one of four *Plasmodium* species that cause human malaria [72] a debilitating and often deadly disease that is endemic in over 100 countries worldwide [73]. Of the four species, *P. falciparum* is the most virulent and is responsible for the almost all of the one million deaths caused by malaria annually [73].

The *P. falciparum* parasite lives intracellularly in human liver and red blood cells and within mosquito midgut lumen epithelial cells. Free parasites also move through the liver and blood of humans and through the gut, hemolymph and saliva of mosquitoes. *P. falciparum* is exposed to bacteria in at least some stages of its lifecycle. An examination of mosquito midgut contents has demonstrated that
malarial oocytes often coexist with a variety of bacteria [74, 75]. However, phagocytosis of bacterial cells has not been documented for this parasite, and it is doubtful that *P. falciparum* has any access to bacterial DNA through cell ingestion in its free-living stages. Once inside the human erythrocyte, endocytosis does occur and is a route for the parasite to bring in nutrients from both the host cell cytosol and from the parasitophorous vacuole in which it resides [76]. Nevertheless, these compartments are devoid of foreign DNA, and so any laterally transferred genes would have to come from free DNA in the extracellular environment. Studies investigating the antiplasmodial effect of antisense oligonucleotides have suggested that larger molecules can enter the parasite from the extracellular milieu [77]. More investigation is needed to determine how and which macromolecules may enter malarial parasites from the external surroundings [78] and whether a route exists whereby free foreign DNA can enter *P. falciparum*.

The *P. falciparum* genome consists of 14 chromosomes with a total size of about 26 Mb [79]. Overall, the genome has an unusually high AT content of 82% [80], representing both very AT-rich intergenic sequences and codon usage that is AT-biased [81]. Fully annotated sequences of both chromosome two [82] and chromosome three [83] have been published. An examination of sequences from these two chromosomes illuminates some notable features of *P. falciparum* genes.

Only 42% of genes on chromosome 2 [82] and 33% of genes on chromosome 3 [83] have identifiable homologues in other species. This is less than half that is found on sequenced chromosomes from other organisms [82]. The paucity of conserved proteins may be an indication that *P. falciparum* evolves at a faster rate than other
organisms, or it may indicate that the divergence time is longer between Plasmodium and the organisms prevalent in sequences databases.

On chromosomes two and three, about 90% of the predicted proteins contain regions characterized by high compositional bias and low complexity [82, 83]. Twenty percent of the low complexity regions are comprised of tandem arrays of short peptide motifs while the other eighty percent represent non-repetitive regions containing runs of a single amino acid [84]. Such regions involve more than 60% of the total peptide in about half of all the proteins [80]. Although regions of low complexity have been found in other organisms, the prevalence and extent of such sequences is much higher in *P. falciparum* than in other eukaryotes, such as *S. cerevisiae*, *C. elegans* [82], or *Dictyostelium discoideum* [80]. The amino acid frequencies within low complexity regions correlate with A-richness in codons, and this bias results in an overall hydrophilic quality [80]. When low-complexity regions occur within globular proteins, it is assumed that these hydrophilic segments 'loop out' of the main core of the folded protein. These regions do not appear to be evolutionarily conserved, because they manifest as *P. falciparum*-specific insertions in alignments with homologues of other species [80, 81] and because their size is often polymorphic among *P. falciparum* strains [83].

About 45% of all genes on chromosomes two and three have introns [82, 83]. Although the majority of these genes have only one intron [83], some are quite complex and may contain up to 15 introns [83, 85]. Alternative splicing occurs in at least some *P. falciparum* genes. Leonard van Lin and his colleagues [86] examined a high gene density region of chromosome ten. Six genes were found in only 13.6 kb; three of these contained multiple introns. In two of these genes, sequencing of
edited transcripts showed that alternative splicing and multiple polyadenylation sites led to several gene products. The presence of multi-exon genes makes it difficult to determine the correct set of open reading frames (ORFs) with automated gene finding programs. The researchers who sequenced chromosomes two and three used different computer algorithms and the resulting set of ORFs predicted by these programs do not agree [87]. Recently, R. Huestis and K. Fischer [88] performed a re-analysis of ORFs in the chromosome 2 sequence. They used *P. falciparum*-specific sequence features to create a computer algorithm that identified likely introns and exons. All identified genes were then subject to an extensive manual analysis and when the identified gene structure varied from the original chromosome 2 annotation, cDNA sequences were used to confirm the position of introns and exons. This analysis added 135 introns to chromosome 2 and put introns into at least 40 heretofore intronless genes. The new exon structure of some genes resulted in significant structural changes to the predicted proteins. The work of Huestis and Fischer [88] underscores the need for experimental confirmation of the predicted protein sequences that are being produced in the *P. falciparum* genome-sequencing project.

Two extrachromosomal genetic elements are also present in *P. falciparum*. A tandemly repeated 5.9 kb linear molecule corresponds to the mitochondrial genome [89], while the apicoplast genome, which is derived from a chloroplast genome, is a 35 kb circle [90]. Both these organellar genomes are reduced when compared with counterparts in other eukaryotes. The mitochondrial genome only encodes rRNAs and a few respiratory chain proteins and is the smallest mitochondrial genome known [91]. Although the apicoplast genome is homologous to the typical plastid found in plants and algae, it lacks all genes associated with photosynthesis and is much smaller.
Unlike plant and algal chloroplasts that are bound by two membranes, four membranes surround the apicoplast. These extra two membranes reflect the secondary endosymbiotic origins of the apicoplast, in which a cell containing a double-membrane bound chloroplast was engulfed by another eukaryotic cell [93].

Reduction of the *P. falciparum* organellar genomes has been accompanied by lateral gene transfer from the organelles to the nuclear genome [31, 92]. Nuclear-encoded genes of organellar origin can be identified through phylogenetic comparison with homologues. Extant plastids and mitochondria each form a monophyletic group whose nearest neighbour lies within a specific bacterial clade; mitochondria are derived from an alpha-proteobacterium [31], and plastids descended from a cyanobacteria-like ancestor [94]. Nuclear genes of organellar origin reflect these phyletic associations.

Many nuclear-encoded genes of organellar origin produce proteins that still function in the organelle. These proteins are transcribed in the cytoplasm and then directed to the appropriate organelle by cell machinery that recognizes leader sequences at the N-terminus of the transcribed protein. These leader sequences are subsequently cleaved from the protein upon import into the organelle [95]. In *P. falciparum*, nuclear-encoded, mitochondrion-targeted proteins (e.g. [96, 97]) possess leader sequences with properties typical of mitochondrion-targeting peptides identified in other eukaryotes [98, 99]. Proteins targeted to the *P. falciparum* apicoplast use a bipartite leader peptide [100]. The first portion of the leader sequence is a classic signalling peptide that targets secreted proteins to the endoplasmic reticulum (ER) in eukaryotes [99]. Because secondary endosymbiotic plastids are topologically outside of their host cell, the signal peptide facilitates
transport to the apicoplast. The second portion of the leader sequence guides the protein into the apicoplast [100]. This part of the leader sequence resembles the transit peptides used to direct proteins into the primary endosymbiotic chloroplasts of plants and algae and has a typical net positive charge [92]. However, unlike plant transit peptides, which are enriched in serine and threonine residues [101], *P. falciparum* transit sequences are rich in lysine and asparagine. This is probably a manifestation of the genome’s AT bias [92]. Based on the number of plastid-targeted genes found in the *A. thaliana* genome and taking into account the absence of many plastid pathways in the apicoplast, Waller and his colleagues [100] estimated that between 6 and 17% of all proteins in the *P. falciparum* genome may be targeted to the apicoplast.

2.2 Methods

2.2.1 *P. falciparum* sequence data source

Deduced protein sequences for *Plasmodium falciparum* genes were obtained from the Malaria Genetics and Genomics website [102] at the National Centre for Biotechnology and Information (NCBI). Several sequencing centres provided the data made available at this site. Sequence data for *P. falciparum* chromosomes 1,3,4,5,6,7,8,9, and 13 were obtained from the Sanger Centre [103]. Sequence data for chromosome 12 were obtained from the Stanford DNA Sequencing and Technology Centre [104]. Preliminary sequence data for *P. falciparum* chromosomes 2,10,11, and 14 were obtained from the Institute for Genomic Research [105]. On Jan 20th, 2001, 2073 *P. falciparum* protein sequences were available at NCBI and were downloaded for investigation. At the time of this analysis, only small amounts of sequence were
available for most chromosomes with the exceptions of chromosomes 2 and 3, which have been published, and chromosome 12, which was completely sequenced although not annotated.

2.2.2 Identification of 'bacteria-like' genes using BAEwatch

The BAEwatch automated sequence analysis system [106] was used to compare the set of *P. falciparum* proteins with a large set of proteins from other organisms and to identify those proteins that showed a higher similarity to bacterial proteins than to proteins from other eukaryotes. The 'comparison' data set was comprised of the SWISSPROT and TrEMBL databases plus a number of unfinished bacterial genomes, as detailed in table 2.1. For each *P. falciparum* protein, BAEwatch used gapped BLAST version 2.0 [107] and MSPcrunch [108] algorithms to generate a list of top matching hits with the 'comparison' data set. The results were placed in a database that was then queried for those proteins whose top hits were bacterial. BAEwatch allows the user to ignore top hits of a particular taxonomic level; for example, if the level chosen is "Family" no top hits from organisms that belong to the same family as the query organism will be considered. For this analysis, all hits from within the phylum apicomplexa were ignored and so the identified 'bacteria-like' genes were those for which the *P. falciparum* gene was more similar to bacterial genes than it was to any other non-apicomplexan sequence. The phylum level was chosen to allow detection of lateral gene transfer that occurred before the divergence of the apicomplexans, a group of parasitic protists with similar lifestyles.

The 'bacteria-like' *P. falciparum* genes were assigned a score that essentially quantified the difference in similarity between the most similar bacterial hit and the
Table 2.1 Protein sets used for BAEwatch comparison. All protein sets were those available in March of 2001. Genomes listed as unfinished were still in progress in March 2001.

<table>
<thead>
<tr>
<th>Published bacterial genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Borrelia burgdorferi</em> B31</td>
</tr>
<tr>
<td><em>Campylobacter jejuni</em> NCTC11168</td>
</tr>
<tr>
<td><em>Chlamydia trachomatis</em> D</td>
</tr>
<tr>
<td><em>Chlamydia muridarum</em> MoPn</td>
</tr>
<tr>
<td><em>Chlamydophila pneumoniae</em> AR39</td>
</tr>
<tr>
<td><em>Chlamydophila pneumoniae</em> CWL029</td>
</tr>
<tr>
<td><em>Chlamydophila pneumoniae</em> J138</td>
</tr>
<tr>
<td><em>Deinococcus radiodurans</em> R1</td>
</tr>
<tr>
<td><em>Escherichia coli</em> O157</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em> Rd</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em> 26695</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em> J99</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em> CSU93</td>
</tr>
<tr>
<td><em>Mycoplasma genitalium</em> G37</td>
</tr>
<tr>
<td><em>Mycoplasma pneumoniae</em> M129</td>
</tr>
<tr>
<td><em>Neisseria meningitidis</em> MC58</td>
</tr>
<tr>
<td><em>Neisseria meningitidis</em> Z2491</td>
</tr>
<tr>
<td><em>Pasteurella multocida</em> PM70</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em> PAO1</td>
</tr>
<tr>
<td><em>Rickettsia prowazekii</em> MadridE</td>
</tr>
<tr>
<td><em>Treponema pallidum</em> Nichols</td>
</tr>
<tr>
<td><em>Ureaplasma urealyticum</em> serovar3</td>
</tr>
<tr>
<td><em>Vibrio cholerae</em> N16961</td>
</tr>
<tr>
<td><em>Xylella fastidiosa</em> NA</td>
</tr>
</tbody>
</table>

| Translations of all coding genes in GenBank SWISSPROT and TREMBL databases [109] |

<table>
<thead>
<tr>
<th>Published eukaryotic genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Arabidopsis thaliana</em></td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
<tr>
<td><em>Homo sapiens</em> - ENSEMBL March 2001 data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unfinished bacterial genomes [110]</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aquifex aeolicus</em> VF5</td>
</tr>
<tr>
<td><em>Bacillus halodurans</em> C-125</td>
</tr>
<tr>
<td><em>Bacillus subtilis</em> 168</td>
</tr>
<tr>
<td><em>Buchnera</em> sp APS</td>
</tr>
<tr>
<td><em>Escherichia coli</em> K12</td>
</tr>
<tr>
<td><em>Lactococcus lactis</em> subsp lactis IL1403</td>
</tr>
<tr>
<td><em>Synechocystis</em> sp PCC6803</td>
</tr>
<tr>
<td><em>Thermotoga maritima</em> MSB8</td>
</tr>
</tbody>
</table>
most similar eukaryotic hit. Genes with low scores (below 10) were discarded from further consideration, because they probably reflect those genes that are highly conserved in all organisms. Generally-conserved proteins pose a problem for analyses that rely on output lists from BLAST, because the small differences in BLAST bit scores do not sufficiently discriminate between levels of similarity [106]. The output from BAEwatch included a list of the subset of \textit{P. falciparum} genes that have a noteworthy similarity to bacterial genes. BAEwatch was thus used as a first-pass tool to identify putative laterally transferred genes.

2.2.3 Identification of features of \textit{P. falciparum} genes.

All 'bacteria-like' \textit{P. falciparum} sequences were evaluated to detect the presence of N-terminal leader sequences. TargetP [99] was used to identify probable signal peptides and mitochondrial targeting peptides. Sequences were subsequently run through the PATS prediction system [111] to detect apicoplast-targeting leader peptides. While TargetP is trained to identify plant transit peptides, it does not recognize the bipartite leaders used by \textit{P. falciparum} to target proteins to the apicoplast [111]. Indeed, in this analysis TargetP recognized only the signal peptide portion of the bipartite leaders and thus invariably misclassified apicoplast-targeted proteins (as identified by PATS) as being targeted for secretion.

'Bacteria-like' genes were used to query the Interpro database [112]. Interpro detects sequence similarity to protein signatures derived by a number of methods. Signatures usually represent functional domains that are found in a variety of proteins. The functional domains identified in \textit{P. falciparum} genes were considered as clues to the possible function of the protein.
2.2.4 Retrieval and identification of putative homologues

Each bacteria-like gene was used to query the non-redundant protein database at the National Centre for Biotechnology Information (NCBI [113]) in October of 2001, using gapped BLAST 2.0 [107]. At this time, the NCBI protein database included a number of completely sequenced genomes from species listed in table 2.2. The set of proteins with an expectation (E) value below 0.001 formed the basis for selection of homologues. The E value is a measure of the expected number of chance alignments that will be as good as the hit, given the size of the protein database [114]. This rather high (non-stringent) exclusion threshold was chosen because E values vary with the length of the query protein. For example, *P. falciparum* proteins with leaders and internal low complexity regions are much longer than bacterial homologues (which lack these features), and thus BLAST queries with such proteins result in high expect values even when the regions outside of the *P. falciparum*-specific segments are very similar.

Putative homologue sets were manually culled according to the following *ad hoc* rules: 1) Because a high expect score cutoff was used, some BLAST hits represented sequences with matches to only functional protein domains that are small and common in a wide variety of proteins. Such proteins may have evolved through domain shuffling and are not generally suitable for phylogenetic reconstruction. To remove such hits, genes with regions of similarity that extended over less that 25% of the hit’s total peptide were excluded. In a few cases, this eliminated all BLAST search results as possible homologues of the *P. falciparum* gene. Furthermore, the schematic visual representation of search results that was provided with NCBI’s BLAST output was inspected. Genes whose regions of similarity were
Table 2.2 Complete genomes queried for homologues of *P. falciparum* genes

<table>
<thead>
<tr>
<th>EUBACTERIA</th>
<th>Aquificales</th>
<th>Archaea</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Firmicutes</strong></td>
<td><strong>Proteobacteria</strong></td>
<td><strong>Aquiferales</strong></td>
<td><strong>Archaea</strong></td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>Alpha</td>
<td>Aquifex aeolicus</td>
<td>Archaeoglobus fulgidus</td>
</tr>
<tr>
<td>Corynebacterium glutamicum</td>
<td>Agrobacterium tumefaciens</td>
<td></td>
<td>Halobacterium sp.</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>Caulobacter crescentus</td>
<td></td>
<td>Methanothermobacter thermoautotrophicus</td>
</tr>
<tr>
<td>Mycobacterium leprae</td>
<td>Mesorhizobium loti</td>
<td></td>
<td>Methanocaldococcus janaschii</td>
</tr>
<tr>
<td>Low GC group</td>
<td>Sinorhizobium meliloti</td>
<td></td>
<td>Pyrococcus abyssi</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>Rickettsia conorii</td>
<td></td>
<td>Pyrococcus horikoshii</td>
</tr>
<tr>
<td>Bacillus halodurans</td>
<td>Rickettsia prowazekii</td>
<td></td>
<td>Thermoplasma acidophilum</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td></td>
<td></td>
<td>Thermoplasma volcanium</td>
</tr>
<tr>
<td>Clostridium acetobutylicum</td>
<td>Beta</td>
<td>Deinococcus/Thermus</td>
<td>Archaeoglobus fulgidus</td>
</tr>
<tr>
<td>Lactococcus lactis</td>
<td>Neisseria meningitidis</td>
<td>Deinococcus radiodurans</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>Campylobacter jejuni</td>
<td>Spirochaetes</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma pulmonis</td>
<td>Helicobacter pylori</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycoplasma pneumoniae</td>
<td>Gamma</td>
<td>Cyanobacteria</td>
<td></td>
</tr>
<tr>
<td>Streptococcus pneumoniae</td>
<td>Buchnera sp.</td>
<td></td>
<td>Haemophilus influenzae</td>
</tr>
<tr>
<td>Ureaplasma urealyticum</td>
<td>Escherichia coli</td>
<td></td>
<td>Pasteurella multocida</td>
</tr>
<tr>
<td></td>
<td>Haemophilus influenzae</td>
<td>Chlamydiales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pasteurella multocida</td>
<td></td>
<td>Chlamydia pneumoniae</td>
</tr>
<tr>
<td></td>
<td>Pseudomonas aeruginosa</td>
<td></td>
<td>Chlamydia trachomatis</td>
</tr>
<tr>
<td></td>
<td>Vibrio cholerae</td>
<td></td>
<td>Chlamydia muridarum</td>
</tr>
<tr>
<td></td>
<td>Xylella fastidiosa</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yersinia pestis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

24
limited to less than half of the regions of similarity of the majority of other search results were excluded. Hits excluded from the homologue set because of the above rules always had high expectation scores. 2) If a particular species had more than one hit to the query, they were considered paralogous genes, and only the gene with the lowest E score was retained. 3) If all (or most) species contained more than one hit and the multiple sequences formed clear similarity groups (i.e. a copy in each species with the same E value), then the multiple hits were treated as orthologous gene families and were retained. Gene function annotations were sometimes helpful in identifying orthologous gene families. 4) To produce phylogenies that are useful for looking at lateral gene transfer, broad taxonomic sampling is needed. Therefore, the species with the lowest E value was chosen as a representative for its taxonomic family. 5) The final set of orthologues for each *P. falciparum* gene was limited to one hundred sequences, in order to facilitate timely phylogenetic reconstruction. If, following removal of sequences as described above, the set of orthologues exceeded 100 sequences, only those with the 100 lowest expect scores were retained.

2.2.5 Sequence alignments

Amino acid sequences of *P. falciparum* genes and their sets of homologues were aligned using the CLUSTALX sequence alignment program [115]. Analysis was conducted at the protein level because amino acid data only reflects replacement substitutions in the DNA (i.e. those substitutions that change the amino acid specified by the codon). Silent substitutions often have species-specific base biases, indicating that a different mutation or selective force is operating in different lineages. This sort of differential selection can lead to sequence convergence, which causes errors
in inferred phylogenetic relationships [116]. This is especially a concern for the AT-biased *Plasmodium falciparum*. Analyses of amino acid instead of DNA sequences should thus reduce the confounding effect of sequence convergence on phylogenetic analysis.

Multiple sequence alignments were visually inspected. The 'show low-scoring segments' option in CLUSTALX was invoked, which highlights regions of poor alignment quality. These regions represent segments of the peptide that are too divergent to be reliably aligned [115]. Sequences were removed whose low-scoring segments comprised more than one-third of the total sites in the alignment. When a putative homologue was removed because of a high proportion of low quality segments, the multiple alignment was re-calculated with the remaining homologues. In some cases, the *P. falciparum* sequence was very divergent and the alignment revealed low-scoring segments for over one third of the protein. In this case, the *P. falciparum* gene was dropped from further analyses.

Alignments were edited to remove aligned sequence positions (sites) that contained gaps, from all species. Internal gaps represent insertions or deletions (indels) in the evolution of a gene. Because phylogenetic inference methods are based on models of amino acid substitution and indels are formed through quite different processes than substitution, gap sites must be excluded from phylogenetic analyses [116]. The proteins used in the alignments were of variable length, so terminal gaps were also present. The ends of sequences were cropped from the alignment to exclude all sites with terminal gaps. Very short sequences with long terminal gaps were excluded entirely from the homologous protein set, as removal of their terminal gaps left a sample size of sites that was too small for phylogenetic analyses. In addition, sites from all species were excluded from the alignment if
positional homology of residues was uncertain. These were visibly apparent as positions where little residue similarity was evident (figure 2.1) and corresponded to sites where low-scoring segments were present in all or most sequences. Site-exclusion editing of alignments was done using GeneDoc [117] version 2.6.002.

Figure 2.1 Example of a divergent region in an alignment. Sites, such as those in the box, are excluded from consideration in distance calculations because the positional homology of the amino acids cannot be assumed. For instance, the circled leucine residue could be placed anywhere along the gap in front of it, without worsening the alignment.
2.2.6  Phylogenetic reconstruction

Phylogenetic inferences were made by constructing trees based on pairwise distances between sequences. Although maximum likelihood methods can outperform distance-based methods in providing the correct tree, maximum likelihood methods require much more computation time than distance methods [116]. The time required increases exponentially with each additional taxon, and maximum likelihood often becomes intractable when more than 15 taxa are to be evaluated. Consequently, distance-based methods were more suitable for phylogenetic analyses of the *P. falciparum* genes and their large sets of homologues.

Maximum likelihood distance estimates were calculated with Tree-Puzzle [118] version 5.0, using the WAG substitution model [119] and amino acid frequencies estimated from the data. Distances were corrected for among-site rate heterogeneity with a "mixed" model, described by a discrete gamma-distribution with eight rate categories plus an additional category for invariant sites. Tree-Puzzle calculated both the alpha shape parameter of the gamma distribution (a) and the fraction of invariant sites (i). One hundred bootstrap datasets were generated using the Seqboot program of Phylip [120]. Distances for bootstrapped sequence sets were calculated by Tree-Puzzle using the same parameters as above, except that to reduce computation time values for the a and i parameters were set at the values previously calculated. The potential biases introduced by this simplification were not explored. The shell script Puzzleboot [121], was used to feed bootstrap datasets to Tree-Puzzle and to produce a concatenated file of distance matrices for all bootstrap replicates. Unrooted phylogenies were constructed from distance matrices using a neighbour-joining
algorithm as implemented in BioNJ [122]. The Consense program in Phylip [120] was used to calculate bootstrap values.

2.3 Results and Discussion

This section presents results of the BAEwatch analysis and the phylogenetic analyses of individual genes. Extensive discussion is included with the results in each section.

2.3.1 BAEwatch

BAEwatch was used to evaluate 2073 *P. falciparum* ORFs. About 65% of these ORFs had some similarity to existing sequences in the protein database. The sequencing projects for chromosomes 2 and 3 [82, 83] performed a similar BLAST analysis on the set of ORFs identified on those chromosomes, but reported much lower match percentages (table 2.3). The higher value found by the BAEwatch analysis is probably not an indication of some intrinsic property of the set of 2073 *P. falciparum* ORFs (which includes the sequences from chromosomes two and three). Rather, this high value occurs because the protein databases searched by BAEwatch were substantially larger than those searched by the chromosome two and three sequencing projects two years earlier, and thus the likelihood of finding matching gene segments was increased. BAEwatch identified 1.4% of *P. falciparum* ORFs as being most similar to a bacterial gene. This value was lower than that reported for chromosome 2 (table 2.3). This may be because the scoring system implemented by BAEwatch considers generally-conserved proteins as not significant, while the values reported for chromosome two were based purely on BLAST hits.
Table 2.3 Summary of BLAST searches with *P. falciparum* sequence sets.

<table>
<thead>
<tr>
<th>Description</th>
<th>Chr 2 [82]</th>
<th>Chr 3 [83]</th>
<th>BAEwatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ORFs evaluated</td>
<td>210</td>
<td>217</td>
<td>2073</td>
</tr>
<tr>
<td>ORFs with similarity to sequences</td>
<td>120</td>
<td>94</td>
<td>1349</td>
</tr>
<tr>
<td>in protein databases</td>
<td>(57%)</td>
<td>(44%)</td>
<td>(65%)</td>
</tr>
<tr>
<td>ORFs most similar to bacterial sequences</td>
<td>16</td>
<td>not reported</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>(7.6%)</td>
<td></td>
<td>(1.4%)</td>
</tr>
</tbody>
</table>

BAEwatch identified thirty *P. falciparum* ORFs as 'bacteria-like'. Two ORFs were discarded as false positives, because the query of the protein database at NCBI in October 2001 (see methods) revealed that the most similar proteins were from eukaryotes. Another fifteen ORFs were excluded from phylogenetic analysis for the following reasons: 1) One *P. falciparum* ORF was an exact copy of a short segment of another *P. falciparum* gene that was identified by BAEwatch. This short ORF was located on a different chromosome than the longer sequence. The BLAST lists of similar sequences were identical for the short and long ORFs, although BLAST hits had lower expect values with the longer sequence because regions of similarity existed that were not present in the shorter sequence. The shorter sequence was discarded from further analysis, as it was clearly a recent duplication. 2) Two ORFs were very short (<80 amino acids), and were discarded because they did not contain enough informative sites for phylogenetic analysis. 3) Five *P. falciparum* sequences were
discarded because no significant potential homologues were found. In these genes, the regions of similarity that were identified by BLAST were limited to small common motifs that are found in a large number of proteins. Notably, in four of these five cases, the *P. falciparum* gene was unusually long (over 1800 amino acids). 4) Seven sequences were too divergent to give reasonable alignments with putative homologues. This left 13 sequences (named in table 2.4), which were subject to phylogenetic analysis. It should be emphasized that the approach taken in this study does not guarantee that all lateral gene transfers in the *P. falciparum* dataset were found. BAEwatch is used as a starting point for likely candidates, and genes that were deemed not amenable to phylogenetic reconstruction may be good targets for study by other methods.

2.3.2 Phylogenetic analysis of individual 'bacteria-like' genes

The following eleven sections present the phylogenetic analyses conducted on the individual 'bacteria-like' genes identified by BAEwatch. Two sections present the analyses of two proteins as the proteins are related; type II topoisomerase B subunit and A subunit are evaluated together in section 2.3.2.1 and GTP-binding proteins 1 and 2 are discussed together in section 2.3.2.11. For each gene, sequence features (e.g. the presence of targeting peptides) and phylogenetic relationships are examined to determine if a lateral gene transfer event has taken place, and if so, what the probable origin of the laterally transferred gene is.
Table 2.4 List of *P. falciparum* genes evaluated by phylogenetic analyses

<table>
<thead>
<tr>
<th>Putative function</th>
<th>Chromosome Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type II topoisomerase: B subunit</td>
<td>12</td>
</tr>
<tr>
<td>Type II topoisomerase: A subunit</td>
<td>12</td>
</tr>
<tr>
<td>Adenylosuccinate lyase</td>
<td>2</td>
</tr>
<tr>
<td>Elongation factor G</td>
<td>12</td>
</tr>
<tr>
<td>Ribosome release factor</td>
<td>2</td>
</tr>
<tr>
<td>Proteasome subunit HslV</td>
<td>12</td>
</tr>
<tr>
<td>Pseudouridine synthase</td>
<td>2</td>
</tr>
<tr>
<td>rRNA methyltransferase</td>
<td>12</td>
</tr>
<tr>
<td>rRNA 3' terminal phosphate cyclase</td>
<td>12</td>
</tr>
<tr>
<td>rRNA adenine dimethylase</td>
<td>12</td>
</tr>
<tr>
<td>S-adenosyl methionine-dependant methyltransferase</td>
<td>12</td>
</tr>
<tr>
<td>GTP-binding protein 1</td>
<td>12</td>
</tr>
<tr>
<td>GTP-binding protein 2</td>
<td>3</td>
</tr>
</tbody>
</table>
2.3.2.1 Type II topoisomerase

Two bacteria-like genes in *P. falciparum* encode type II DNA topoisomerases. These enzymes alter the topological state of DNA in the cell by creating transient double-stranded breaks. Type II topoisomerases are essential to all cells for replication, separation of replicated chromosomes, and maintenance of superhelicity [123]. Both eukaryotes and prokaryotes possess type II topoisomerases, although in different forms. The eukaryotic topoisomerase II is a homodimer of a single polypeptide of approximately 1500 amino acids. Eubacteria typically possess two type II topoisomerases: DNA gyrase and topoisomerase IV. Both of these enzymes are heterotetramers, comprised of two different subunits designated A and B for gyrase, and ParC (or A) and ParE (or B) for topoisomerase IV [124].

These type II topoisomerases form a homologous family. The N-terminal half of the eukaryotic gene (topoisomerase II) is homologous to the B subunits from gyrase and topoisomerase IV, and the C-terminal half is homologous to the A subunits (figure 2.2). A number of individual bacterial species have only one type II topoisomerase (see figures 2.3, 2.4). Among mesophiles, these species are exceptions and probably reflect lineage-specific loss of one of the two enzymes. The eubacterial hyperthermophiles have only one type II topoisomerase. Interestingly, the archaea possess a type II topoisomerase known as topoisomerase VI that, while comprised of the same two-subunit heterotetramer structure as eubacterial proteins, is non-homologous to the eukaryotic and bacterial types [124]. However, some of the completely sequenced archaeal genomes also possess A and B subunits of gyrase. *P. falciparum* has a eukaryotic topoisomerase II [125]. *P. falciparum* also appears to
Figure 2.2 Architecture of Type II Topoisomerases. The four main forms of type II topoisomerase are shown: Eukaryotic topoisomerase II (top panel), Archaeal topoisomerase, Eubacterial gyrase, and Eubacterial topoisomerase IV. The Archaeal and Eubacterial proteins contain a B subunit (middle panel) and an A subunit (bottom panel). Species or taxonomic groups that differ from the main form are shown beneath that form. Insertions are depicted below sequences, and homologous insertions are shaded alike. Putative targeting peptides are cross hatched. MT, mitochondrion-localized; P. falciparum, Plasmodium falciparum; G. theta, Guillardia theta nucleomorph; A. thaliana, Arabidopsis thaliana; D. discoideum, Dictyostelium discoideum.
have the bacterial form of type II topoisomerase. The set of homologues for two 'bacterial-like' proteins identified by BAEwatch on chromosome 12 consists of A and B subunits of gyrase and topoisomerase IV from bacterial species.

The alignment of the B subunit of *P. falciparum* and its set of homologues provided 510 unambiguous positions that were used to build a phylogenetic tree. Gyrase and topoisomerase IV each form separate, well-supported clades (figure 2.3) among the low GC firmicutes and proteobacteria, suggesting that these genes arose from a duplication within these groups of eubacteria. *P. falciparum*, like many of the bacteria, has only one B subunit type II topoisomerase. The sequence groups with topoisomerase IV from the chlamydiales and *Borrelia burgdorferi* (a spirochaete). This placement has some bootstrap support (68%), but caution is warranted in placing too much confidence in the clade as it consists of relatively divergent sequences with long branches and could be an artefact of long branch attraction.

The A subunits of Topoisomerase IV are only well conserved in the N-terminal half of the peptide and so only 262 unambiguously aligned positions were used to build the phylogeny (figure 2.4). As in the B subunit tree, gyrase and topoisomerase IV fall into separate clades, and higher order bacterial taxa form well-supported groups. Again, the *P. falciparum* sequence aligns with the chlamydiales and *B. burgdorferi*, albeit with low bootstrap support (50%) and a long branch.

In addition to *P. falciparum*, two other eukaryotes have B and A subunits of a type II topoisomerase: *Arabidopsis thaliana*, a plant, and *Guillardia theta* (the subunits are found in the nucleomorph genome, the remnant nucleus of a secondary symbiosis). These organisms have plastids and their sequences group convincingly with the cyanobacteria in both the B subunit (figure 2.3) and A subunit (figure 2.4)
Figure 2.3 Phylogeny of the B subunit of Type II topoisomerases. The P. falciparum sequence is boxed for reference. The phylogeny is an unrooted BioNJ tree based on rate-corrected maximum-likelihood distances calculated with TreePuzzle. Bootstrap values over 50% are shown at nodes. Taxonomic groups are shown to the right of species' names for clades supported by bootstrap values. The branch length scale represents 0.1 amino acid changes. G, gyrase; t, topoisomerase IV. Where no suffix letter is present for species names, only one type II topoisomerase form has been identified.
Figure 2.4 Phylogeny of the A subunit of Type II topoisomerases. See figure 2.3 for details.
trees. The gene architecture for the B and A subunits clearly shows that *P. falciparum*, *G. theta* and *A. thaliana* all have N-terminal extensions reminiscent of leader sequences (figure 2.2). All three of these B subunit extensions are predicted to be plastid targeting, but of the A subunits, curiously only the *A. thaliana* protein is predicted to be localized to the plastid. The *G. theta* leader is predicted to be a signal peptide (indicating that the A subunit is exported), and the *P. falciparum* leader has no predicted function. As type II topoisomerase subunits do not have any function on their own [124], it is likely that B and A subunits are localized to the same place in a cell if they are functional. It is possible that A subunits from *G. theta* and *P. falciparum* have some as yet unidentified way of ending up in the plastid alongside their B subunit counterparts; these sequences may use a targeting peptide that is not recognized by the PATS prediction algorithm, or perhaps the A subunit 'hitches a ride' with the B subunit.

The combination of localization leaders and specific phylogenetic affiliation with the cyanobacteria, leads to the conclusion that 'bacteria-like' type II topoisomerase subunits originated in the plastid and were transferred to the nucleus in the case of *A. thaliana* and to the nucleomorph (the remnant nucleus of the secondarily symbiotic plastid) in the case of *G. theta*. Similarly for *P. falciparum*, the 'bacteria-like' topoisomerase subunits most probably function in the apicoplast, but the origin of the gene is less certain as the apparent relationship to the *B. burgdorferi/Chlamydiales* clade is inconsistent with apicoplast inheritance.

To explore the relationship between prokaryotic and eukaryotic type II topoisomerases, B and A subunits for prokaryotic gyrase and topoisomerase IV as well as for the 'bacteria-like' eukaryotic sequences (*P. falciparum*, *A. thaliana*, *G. theta*)
Figure 2.5 Phylogeny of type II topoisomerase. The phylogeny is based on an alignment of eukaryotic topoisomerases with concatenated B and A subunits of bacterial topoisomerases. For clarity, only major nodes of large bacterial clades are labelled with bootstrap values. mt, mitochondrially localized; G, Gyrase; T, Topoisomerase IV; E, Topoisomerase II; the protein of species with no suffix letter is of bacterial form (with separate A and B subunits). See figure 2.3 for details.
were concatenated and aligned with eukaryotic topoisomerase II, providing 659 unequivocally aligned sites for phylogenetic analysis. Eukaryotic sequences form a well-defined clade separate from the prokaryotic sequences (figure 2.5), as we would expect from the architecture of type II topoisomerases (figure 2.2). Additionally, the phylogeny shows that in both A. thaliana and P. falciparum the typical eukaryotic topoisomerase II is not related to the bacterial subunit forms. In the concatenated tree, the eukaryotic clade forms the longest branch. P. falciparum has been displaced as the sister group to the Chlamydiales/B. burgdorferi clade by the long eukaryotic clade branch, and now P. falciparum falls in a relatively well-supported clade comprised of the cyanobacterial and the plastid-targeted eukaryotic genes. The observation that the two longest branches unite in all three trees, gives credence to the conjecture that the P. falciparum sequence-placement in the B and A subunit trees (figures 2.3 and 2.4) is spurious. The placement of P. falciparum with the cyanobacteria in figure 2.5 lends support to a plastid origin of the bacterial form of type II topoisomerase in P. falciparum.

2.3.2.2 Adenylosuccinate lyase

The adenylosuccinate lyase gene in P. falciparum is located on chromosome 2. Adenylosuccinate lyase is a bifunctional enzyme that catalyzes the first step in de novo purine biosynthesis; it can also form AMP from adenylosuccinate as part of a purine salvage pathway. P. falciparum is incapable of purine biosynthesis, so its enzyme only functions in purine salvage [126].

The adenylosuccinate lyase gene from P. falciparum is 471 amino acids in length. It is easily aligned with its group of homologues and is relatively conserved,
allowing 331 sites to be considered in the phylogenetic analysis. The tree inferred from this alignment nests the *P. falciparum* sequence within the gamma group of proteobacteria (figure 2.6). The tree recovers some major taxonomic groups; specifically, the animal and fungal eukaryotes, low GC firmicutes, and cyanobacteria all form highly supported clades. Only the alpha and epsilon divisions of proteobacteria are monophyletic. The gamma and beta proteobacteria fall in a highly supported mixed group with three eukaryotes and an archaeon, *Halobacterium sp.* The archaea, with the exception of *Halobacterium sp.*, form a well-supported group with internal branches leading to the animal/fungal eukaryotic clade and the mixed group. Both the fungal/animal eukaryotic clade and the mixed clade form distinct groups from the rest of the tree. This observation is born out by the presence of group defining indels in the multiple sequence alignment of adenylosuccinate lyase homologues (figure 2.7).

The relationship of *P. falciparum* to the gamma proteobacteria is intriguing. Two other eukaryotic sequences group within the gamma proteobacteria. Neither the *A. thaliana, Leishmania major,* nor the *P. falciparum* sequences possess leader peptides. Moreover, the phylogenetic position of these eukaryotic sequences excludes both the cyanobacteria and the alpha proteobacteria, precluding an organellar origin for this gene. As an aside, the sequencing project of *P. falciparum* chromosome 2 [82] identified adenylosuccinate lyase as an organellar gene, presumably based on an apparent targeting peptide. There is no biochemical indication that this is so, and I found no evolutionary evidence for their conclusion. I analysed the adenylosuccinate lyase protein with the plastid-targeting peptide.
Figure 2.6 Phylogeny of adenylosuccinate lyase. The 'mixed' clade contains gamma and beta proteobacteria, an archaeon, and eukaryotes. Details as in figure 2.3.
Figure 2.7. Multiple sequence alignment of Adenylosuccinate lyase. A portion (about 20%) of the *P. falciparum* adenylsuccinate lyase gene is shown aligned with homologues from other organisms. Amino acid numbering for each sequence is shown to the right of the sequence line. The top number line specifies sites positions in the overall multiple alignment. Gaps are indicated by dashes. Sites that are over 80% percent conserved in all sequences are shaded dark grey, and sites between 60 and 80 percent conserved are shaded light grey. Groups of sequences that share unique motifs and/or indels are separated into blocks. Unique indels are boxed.
prediction software PATS [111], which is a neural network prediction tool that has specifically been trained on confirmed *P. falciparum* plastid-targeting sequences. PATS did not identify a plastid-targeting sequence.

The three eukaryotic sequences that fall in the mixed clade have in common their exclusion from the other eukaryotic sequences. The eukaryotic clade (figure 2.6) is composed of animals and fungi only, which are generally considered more closely related to each other than either is to the plants. Because the branching order within the mixed clade is not resolved in this tree, a new tree was constructed using sequences from just this clade. Alignment of only this clade allowed for an additional 22 informative sites to be included in the analysis. Because *Halobacterium sp.* is excluded from the other sequences in the mixed clade with a bootstrap value of 92% (see figure 2.6), it was used to root the mixed clade sub-tree (figure 2.8). This tree has slightly better resolution than the mixed clade in the large tree. The branching topology resulting from rooting with an archaeal species makes taxonomic sense; the trypanosome *L. major* and the apicomplexan *P. falciparum* are the most basal eukaryotes, with *A. thaliana* branching next, and presumed anomalous gamma/beta bacteria form a single clade.

The overall phylogeny (figure 2.6) supports four overall distinct clades: the bacteria (excluding alpha and gamma proteobacteria), the animal and fungal eukaryotes, the archaea (excluding *Halobacterium sp.*), and the mixed clade. The bacterial clade is distinct from the animal and fungal eukaryotes, the archaea (excluding *Halobacterium sp.*), and the mixed clade, but the branching order between these last three clades is not resolved and all are equally likely to be related to the
Figure 2.8 Phylogeny of adenylosuccinate lyase 'mixed' clade. Only sequences belonging to the mixed clade identified in figure 2.6 were used to build the phylogeny. The phylogeny is a BioNJ tree made with rate-corrected maximum likelihood distances, and rooted on the *Halobacterium sp.* sequence. Bootstrap values over 50% are shown at nodes.
others. The simplest hypothesis to explain the phylogeny of adenylosuccinate lyase (figure 2.6) involves at least two lateral transfers. If we assume that the mixed clade branching order is as shown in figure 2.8, then a lateral transfer event from an early eukaryote-like organism to an ancestor of the extant gamma and beta proteobacteria, would account for the mixed clade. Additionally, the animal/fungal clade would have had to acquire its gene from an archaea, through lateral transfer, to account for its position outside *Halobacterium sp.* More adenylosuccinate lyase sequences are needed to identify if these hypotheses are correct. Specifically, sequences from lower eukaryotes and plants are needed to break up the long branches connecting the animal/fungal eukaryotic clade with the mixed clade.

2.3.2.3 Elongation factor G

A 'bacteria-like' gene for elongation factor G was identified by BAEwatch in chromosome 12 of *P. falciparum*. Elongation factor G (EF-G) is a one of a family of ubiquitous translation elongation factors that are found in all organisms [127]. EF-G is believed to have arisen from a very early duplication event, because a paralogous gene is found in organisms from all three domains. Consequently, EF-G has been extensively used to study various questions concerning evolutionary events at the base of the tree of life [128-131].

Because homologues were available from many organisms, the set of homologues of the *P. falciparum* sequence was chosen by taking the top 50 (lowest expect scores) sequences from the BLAST search of the protein database at NCBI. These sequences provided 640 well-aligned sites to build the phylogeny (figure 2.9). Upon discovering that all the eukaryotic sequences that happened to be included in the set of
homologues were likely of mitochondrial origin (figure 2.9), the phylogeny was regenerated with the inclusion of two archaeal and two nuclear eukaryotic EF-G (also known as EF-2) genes. This analysis (not shown) resulted in a less well-resolved phylogeny of the same overall topology and with a longer branch joining the eukaryotic and archaeal sequences to the backbone of the tree. In this additional phylogeny, the *P. falciparum* sequence grouped within a moderately supported (72 percent bootstrap) clade that included spirochaetes, cyanobacteria, organellar sequences, and alpha proteobacteria. The exclusion of the eukaryote sequences from this clade was taken as evidence that the *P. falciparum* sequence was 'bacteria-like' and not closely related to eukaryotic nuclear EF-G.

The EF-G phylogeny presented in figure 2.9 supports several accepted taxonomic relationships, but with some important differences. The low GC firmicutes are monophyletic (ignoring the undefined placement of *Clostridium acetobutylicum*), as are the actinobacteria and the alpha, beta, gamma, and epsilon subgroups of proteobacteria. The proteobacteria as a group, however, are polyphyletic. The mitochondrion- and plastid-targeted sequences are each monophyletic. Unexpectedly, the mitochondrion-targeted eukaryotic sequences do not branch with the alpha proteobacteria. Rather, the alpha proteobacteria and the plastid targeted sequences form sister clades. The cyanobacteria appear to be polyphyletic with three representatives forming a well-supported clade and one representative, *Synechocystis sp.*, related to *Vibrio cholerae* (a gamma proteobacteria). A strongly supported branch groups the *P. falciparum* sequence with the spirochaetes and mitochondrion-targeted eukaryotic clades.
Figure 2.9 Phylogeny of Elongation Factor G. The *P. falciparum* sequence is boxed for reference. Bootstrap-supported higher-order taxonomic clades are labelled to the right of the species names. Bootstrap values over 50 are shown at nodes. MT: mitochondria-targeted, CP: chloroplast-targeted, as annotated in sequence databases. Other details as in figure 2.3.
The *P. falciparum* sequence has an N-terminal leader of 89 amino acids, which is similar in size to the leaders of the chloroplast-targeted *A. thaliana* and *Glycine max* sequences (figure 2.10). Moreover, this leader is predicted to be an apicoplast transit peptide. However, the phylogeny of EF-G shows that the *P. falciparum* sequence is much more closely related to the mitochondrion-targeted and spirochaete clade, than to the cyanobacteria (the presumed progenitors of the plastid genes). The *P. falciparum*/spirochaetes/mitochondrion-targeted clade is strongly supported by bootstrapping (99 percent) and is also united by a unique indel (figure 2.10). The unusual organellar-bacterial relationships in the EF-G phylogeny complicate the interpretation of the origin of the *P. falciparum* gene.

Any interpretation of the EF-G phylogeny as it is presented in figure 2.9 would involve multiple lateral transfer events, with some more biologically plausible than others. For example, the clade that unites alpha proteobacteria and plant chloroplast targeted sequences, could be explained as follows: First, an EF-G gene was transferred from an ancestor of the alpha proteobacteria to the nucleus of an ancestor of plants. Such an event may not be all that unlikely, as an extant alpha bacterium has the ability to insert tumour-inducing genes into the nucleus of the plant cells it infects [132]. Subsequently, the alpha proteobacterial gene acquired a transit peptide, and functionally replaced the cyanobacteria-like EF-G in the plastid. The *P. falciparum*/ spirochaete/mitochondrion-targeted clade is much harder to reconcile. The highly-supported grouping of the *P. falciparum* gene with other eukaryote mitochondrion-targeted genes suggests that the *P. falciparum* gene was transferred from the mitochondrion to the nucleus where it picked up a transit
Figure 2.10 Multiple sequence alignment of elongation factor G. An N-terminal segment of elongation factor G from *P. falciparum* is shown aligned with homologues from other species. Shading indicates the level of residue conservation in each site column: light grey, over 60% conserved; dark grey, over 80% conserved; black, 100% conserved. Numbers preceding the eukaryotic sequences indicate the length of the N-terminal leader sequence. MT: mitochondrion-targeted, CP: chloroplast-targeted, as labelled in gene annotation.
peptide and functionally replaced the cyanobacteria-like EF-G in the plastid. Under this hypothesis, however, we would expect the *P. falciparum* gene and all eukaryote mitochondrion-targeted genes to branch within the alpha proteobacteria/plant chloroplast clade, and they are excluded with a bootstrap value of 80 (figure 2.9). We also must suppose that the spirochaete (eubacteria) genes have a plastid origin, which represents a rather inexplicable lateral gene transfer.

It is, of course, quite possible that the EF-G phylogeny is an artefact of the method used to align the sequences and reconstruct the phylogeny, caused by using models of evolution that do not accurately describe the real processes. Lopez *et al.* [131] examined EF-G sequences and concluded that for this gene, variation in rates of change between sites are not best described by the gamma distribution (as used in the phylogenetic analysis I have presented here), but rather that rates co-vary between sites and between species. However, the phylogeny they presented for EF-G, built taking the covarion model into account, showed the same curious features as the phylogeny in figure 2.9: proteobacteria were not monophyletic, mitochondrion-targeted sequences were monophyletic, and chloroplast-targeted sequences clustered with alpha proteobacteria, far from the cyanobacteria.

Because the *P. falciparum* sequence is specifically related to the bacterial/eukaryotic organelle group rather than to the eukaryotic group of EF-G, the gene has been laterally transferred from a eubacterium or organelle to the *P. falciparum* nucleus. Both my phylogenetic analysis and that of Lopez *et al.* [131] suggests that the EF-G gene has experienced a complex evolutionary history and thus phylogenetic evidence for identity of the donor of this transfer remains elusive.
2.3.2.4 Ribosome release factor

Chromosome two of *P. falciparum* contains a gene resembling ribosome release factor, a protein that releases ribosomes from mRNA at the termination of protein biosynthesis [133]. The *P. falciparum* gene is 498 amino acids in length, while its set of homologues include eubacterial sequences of about 180 amino acids, and eukaryote organellar sequences of about 250 amino acids. The *P. falciparum* protein is much longer because of a 320 amino acid N-terminal extension. About half of this leader peptide is comprised of low complexity sequence that contains 10 repeats of a heptamer. The *P. falciparum* sequence is predicted to be targeted to the apicoplast. This small protein is quite well conserved, and 140 sites were used for phylogenetic analysis.

The phylogeny of ribosome release factor is shown in figure 2.11. The tree recovers several expected phylogenetic groupings, although the branching order between these groups is not resolved. Groups consisting of chloroplast- and mitochondrion-targeted eukaryotic genes were also identified and these clades were sister to cyanobacteria and alpha proteobacteria, respectively. Unfortunately, the placement of the *P. falciparum* gene is not resolved, and it is very divergent. In the absence of any phylogenetic evidence, the presence of an apicoplast-targeting leader sequence is taken as an indication of the gene’s plastid origin.
Figure 2.11 Phylogeny of Ribosome Release Factor. The sequence from *P. falciparum* is boxed for reference. MT: mitochondrion-localized, CP: chloroplast-localized, as identified in protein database annotation. Other figure details given in figure 2.3.
2.3.2.5 Protasome subunit HslV

A gene for the proteasome subunit HslV was identified on chromosome 12 in P. falciparum. Proteasomes are cylindrical assemblies of protein-degrading enzymes. Three types of proteasomes are known. Eukaryotes and archaea possess a 20S proteasome. This 20S proteasome is curiously also found in one sole group of bacteria, the actinomycetes [134], and it has been suggested that this represents a lateral gene transfer to the actinomycetes from a non-eubacterial source [135]. A second type of proteasome, ClpAP/ClpXP, has been identified from almost all groups of eubacteria, although it is notably absent from the mycoplasmas (low GC firmicutes lacking a cell wall). These proteasomes are also present in some chloroplasts and in the human mitochondrion [134]. The protease subunit of ClpAP/ClpXP proteasomes is ClpP and this gene is unrelated to the protease subunit of 20S. A third type of proteasome, the HslVU, has been identified in bacteria, but its distribution among bacterial groups is sporadic. It is not present in the cyanobacteria, actinomycetes, chlamydiales or mycoplasmas. All bacterial groups that have HslVU proteasomes also have the ClpAP/ClpXP proteasomes [134]. HslV is the protease subunit of HslVU type proteasomes. It shares some sequence features with the 20S protease subunit, and may be a distant homologue [136].

The main proteasome used in eukaryotic cells is the 20S. The ClpAP/ClpXP proteasomes are also found in eukaryotes and are associated with organelles. ClpP subunits are encoded on plastid genomes (e.g. [30, 137]) and plastid-targeted ClpP genes occur in nuclear genomes (e.g. [138, 139]). ClpP subunits have been shown to localize to the mitochondrion as well [140]. In P. falciparum, a 20S proteasome [96, 141] and an apicoplast-targeted, nuclear encoded ClpP [83] have been identified. In
contrast to the widespread occurrence of the ClpP proteasome subunit, genes of the HslV proteasome are unknown in eukaryotes with the exception of the HslV from *P. falciparum* used in the HslV phylogeny presented here. None of the five completely sequenced eukaryote genomes queried in this analysis (see table 2.2) possess an HslV homologue.

The HslV gene from *P. falciparum* is 187 amino acids long and shows no substantial length variation from its homologues. These genes are easily alignable, although the *P. falciparum* sequence is somewhat divergent at the N-terminal end of the peptide (figure 2.12). One hundred thirty four amino acid site positions were deemed co-linear and used to construct the tree shown in figure 2.13. The phylogeny of HslV supports the major groups of bacteria fairly well, but does not resolve relationships between them. The HslV gene from *P. falciparum* groups within a clade comprised of alpha proteobacteria (figure 2.13). In *P. falciparum*, the HslV has no identifiable leader sequence and is not predicted to be targeted to an organelle. The specific grouping of the *P. falciparum* sequence with those from alpha Proteobacteria, however, indicates that this gene was likely transferred from the mitochondrion to the nucleus. Whether it is targeted to the mitochondrion by some unknown mechanism, or has found some function in the cytosol, remains to be seen. This gene is not located in any extant mitochondrial genomes. The absence of this gene in both the nucleus and mitochondrion of other eukaryotes may mean that while this gene was transferred from the mitochondrion to the nucleus in a *P. falciparum* ancestor, it was lost in all other eukaryote lineages and was a gene of dispensable function in mitochondria. This hypothesis is consistent with the absence of a
<table>
<thead>
<tr>
<th>Organism</th>
<th>Alignment Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Campylobacter jejuni</strong></td>
<td>120</td>
</tr>
<tr>
<td><strong>Helicobacter pylori</strong></td>
<td>140</td>
</tr>
<tr>
<td><strong>Borrelia burgdorferi</strong></td>
<td>160</td>
</tr>
<tr>
<td><strong>Bacillus subtilis</strong></td>
<td>180</td>
</tr>
<tr>
<td><strong>Listeria innocua</strong></td>
<td>200</td>
</tr>
<tr>
<td><strong>Staphylococcus aureus</strong></td>
<td>220</td>
</tr>
<tr>
<td><strong>Lactococcus lactis</strong></td>
<td>240</td>
</tr>
<tr>
<td><strong>Thermotoga maritima</strong></td>
<td>260</td>
</tr>
<tr>
<td><strong>Aquifex aeolicus</strong></td>
<td>280</td>
</tr>
<tr>
<td><strong>Escherichia coli</strong></td>
<td>300</td>
</tr>
<tr>
<td><strong>Vibrio cholerae</strong></td>
<td>320</td>
</tr>
<tr>
<td><strong>Haemophilus influenza</strong></td>
<td>340</td>
</tr>
<tr>
<td><strong>Pseudomonas aeruginosa</strong></td>
<td>360</td>
</tr>
<tr>
<td><strong>Buchnera sp.</strong></td>
<td>380</td>
</tr>
<tr>
<td><strong>Xylella fastidiosa</strong></td>
<td>400</td>
</tr>
<tr>
<td><strong>Ralstonia solanacearum</strong></td>
<td>420</td>
</tr>
<tr>
<td><strong>Mesorhizobium loti</strong></td>
<td>440</td>
</tr>
<tr>
<td><strong>Sinorhizobium mellotii</strong></td>
<td>460</td>
</tr>
<tr>
<td><strong>Caulobacter crescentus</strong></td>
<td>480</td>
</tr>
<tr>
<td><strong>Rickettsia prowazekii</strong></td>
<td>500</td>
</tr>
<tr>
<td><strong>Plasmodium falciparum</strong></td>
<td>520</td>
</tr>
</tbody>
</table>

**Figure 2.12** Multiple sequence alignment of the Protease subunit HsIV. The gene from Plasmodium falciparum is shown aligned with homologues from other organisms. Numbers at the end of lines correspond to the amino acid position in that particular sequence. Numbers across the top of the block indicate the overall site position in the alignment. Shading indicates residue conservation per site: Black, 99-100%; dark grey, 80-99%; light grey, 60-79%.
Figure 2.13 Phylogeny of protease subunit HslV. The *P. falciparum* sequence is boxed for reference. Supported clades are identified with taxonomic affiliation. Details are as in figure 2.3.
mitochondrion-targeting peptide in *P. falciparum*. If the HslV subunit is still functional in the cytosol of *P. falciparum*, it may have a unique function, an interesting possibility that deserves further study.

2.3.2.6 Pseudouridine synthase

*P. falciparum* contains a gene located on chromosome 2 that encodes a pseudouridine synthase. This gene was sequenced and identified by the *P. falciparum* chromosome 2 sequencing project [82]. Pseudouridine synthases form the modified base pseudouridine from uridines in various RNAs [142]. The set of homologues for *P. falciparum* sequence consists of eubacterial RsuA pseudouridine synthases. These enzymes modify a specific uracil in the rRNA of the small ribosomal subunit. Previously studies have noted that among the many types of pseudouridine synthases, the RsuA pseudouridine synthases appear to form a homologous group [142, 143].

The *P. falciparum* gene is 338 amino acids long. RsuA homologues from different bacteria vary in length and are relatively divergent. The conserved regions in the multiple sequence alignment are separated by stretches with little or no similarity. Consequently, only 113 amino acid sites were used for phylogenetic reconstruction. Unfortunately, the tree that resulted from this analysis is not well resolved, and has generally low bootstrap values (figure 2.14).

Many of the bacterial species have multiple RsuA homologues. The phylogeny of RsuA (figure 2.14) groups the multiple homologues into four clades for the proteobacteria, two clades for the low GC Firmicutes, and two clades for the cyanobacteria. Although bootstrap support is low for these clades, all bacterial species with multiple homologues show the same pattern; a particular homologue is
Figure 2.14 Phylogeny of RsuA pseudouridine synthase. The *P. falciparum* sequence is boxed for reference. Taxonomic affiliation of species is labelled. Species suffix numbers represent general group associations, and are independent of the number of homologues in any one species. (e.g. *S. typhimurium* has three homologues named 1, 3 and 4).
more closely related to sequences from close-relatives (defined *a priori* - see table 2.2) than either to a sequence from a more distant relative or to another homologue from its own species. This lends confidence to the delineation of homologous groups. There is only one highly supported larger clade in the tree, which unites group one of cyanobacteria with the proteobacteria group four (figure 2.14). This association is supported in the multiple sequence alignment, by the presence of a region at the N-terminal end of the peptide that is conserved in these sequences alone. The region was identified at the Interpro database [144] as a common RNA binding domain. The occurrence of this particular RNA binding domain in some homologues of RsuA pseudouridine synthases has been found by other researchers [145].

A well-supported clade unites *A. thaliana* with the chlamydiales. It has previously been noted that plants have a number of genes that are related to chlamydial genes. This relationship can appear when lineage-specific gene loss occurs in the cyanobacteria. Because chalmydiales and cyanobacteria appear to be sister-groups, plant genes of plastid origin are most closely related to chlamydial homologues in the absence of a cyanobacterial homologue [106]. Such an evolutionary history seems likely in the RsuA tree, as lineage specific loss of homologues is common. This means that the *A. thaliana* and cyanobacterial RsuA sequences are paralogues, and that the chlamydial sequences are the closest orthologues for *A. thaliana*.

Because of the length heterogeneity among RsuA genes, no obvious N-terminal extension can be seen in the *P. falciparum* or *A. thaliana* sequence. However, both of these genes are predicted to have plastid-targeting peptides. Unfortunately, there is no support for the *P. falciparum* branch in the RsuA tree (figure 2.14) and its
placement is deemed unresolved. The presence of a targeting peptide coupled with the phylogenetic position of A. thaliana, argues for a plastid origin of the RsuA gene in A. thaliana. The targeting peptide and fact that the only other eukaryote with a RsuA has one of plastid origin, leads to the tentative conclusion that that P. falciparum RsuA pseudouridine synthase was laterally transferred from the apicoplast progenitor, to the P. falciparum nuclear genome.

2.3.2.7 Ribosomal RNA methyltransferase

BAEwatch identified an intron-containing gene on chromosome 12 in P. falciparum as 'bacteria-like'. The homologues of this gene in other species are all predicted to be ribosomal RNA methyltransferases. In E. coli, this gene was recently cloned and the enzyme was shown to methylate a C residue at a specific position in the 16S RNA [146]. Based on conservation of the identified E. coli functional sites in sequences from other species, it was suggested that these m^5C rRNA methyltransferases comprise a family of proteins with representatives in all three domains of life [146]. Very few functional studies have been done on enzymes of this family, although some of the eukaryotic m^5C RNA methyltransferases have been shown to localize to the nucleolus during cell proliferation. Reid et al. [147] grouped the m^5C rRNA methyltransferases into six categories based on the length of variable C and N-terminal regions and on small sequence motifs. Anantharaman et al. [145] further grouped m^5C rRNA methyltransferases into four homologous groups, according to the presence of specific RNA binding domains.

The P. falciparum gene is 376 amino acids in length. The multiple sequence alignment for the gene reveals small regions of conserved sequence (about 5-20 amino
acids) separated by larger non-conserved regions. The N and C-terminal regions of this protein show considerable size variation even among peptides from closely related species. Consequently only 142 sites were used were used for phylogenetic reconstruction.

The first pattern evident from the phylogeny of m^5C rRNA methyltransferases (figure 2.15) is that a long branch divides the tree in two. The difference between the members of these two groups is not obvious from the multiple sequence alignment. However, a search with these sequences in the Interpro database reveals that sequences in the first group possess a NusB RNA binding domain at the N-terminal region of the peptide. The grouping of sequences by the presence of the NusB domain is consistent with one of the homologous groups identified by Anantharaman et al. [145]. The NusB-containing group is eubacterial, with the P. falciparum sequence present as the only eukaryote.

Eight well-supported clades that encompass accepted related species are seen in the NusB-containing portion of the tree (fig 2.15). These describe a chlamydiales branch, an actinobacteria branch, a low GC firmicutes branch, a gamma proteobacteria branch, two alpha proteobacteria branches, and two beta proteobacteria branches. Synechocystis sp. and Thermotoga maritima, as sole representatives of their respective cyanobacteria and thermotogales groups, are not included within the larger groups, and are considered to define their own clades. The relationship between most of these groups is not resolved. However, a strongly supported branch unites the type two alpha proteobacteria, the chlamydiales, and the type two beta proteobacteria, and defines the sequences in these groups as orthologues.
Figure 2.15 Phylogeny of m^5C rRNA methyltransferase. The *P. falciparum* sequence is boxed for reference. Suffix numbering of species denotes that multiple homologues exist. The dotted line divides the tree into two groups, one of which contains sequences that possess an N-terminal NusB RNA binding domain (see text). Details as in figure 2.3
A highly supported branch links the *P. falciparum* sequence with the chamydial branch. The gene in *P. falciparum* has undoubtedly been transferred from a bacteria, as it represents a purely eubacterial form of m$^5$C rRNA methyltransferases. The question then becomes: is an organelle the source of the gene, or was the gene transferred from a free-living bacterium? As the *P. falciparum* sequence is not most closely related to the alpha bacterial orthologues in the phylogeny, a mitochondrial source is unlikely. It is less easy to evaluate the relationship of the *P. falciparum* sequence with the cyanobacteria. If the event that led to paralogous genes in alpha and beta proteobacteria was gene duplication in the lineage that led to these two groups, then the *Synechocystis sp.* sequence is an orthologue of the *P. falciparum* sequence and the distance between them argues against a plastid origin. However, if a duplication occurred early in the divergence of Eubacteria, then lineage-specific loss has produced the singular homologues of the actinobacteria, low GC firmicutes, gamma proteobacteria, chlamydiales, thermotogales, and cyanobacteria. In this case, the cyanobacterial sequence may be more distantly related to the *P. falciparum* sequence because they are paralogous. If the *P. falciparum* gene has a plastid origin, the loss of the orthologue in the cyanobacterial lineage then leads to the apparent association between the *P. falciparum* and its next closest relatives, the chlamydiales [106].

It is unlikely that the *P. falciparum* m$^5$C rRNA methyltransferase is localized to the apicoplast. The protein sequence does not possess an identifiable N-terminal extension, and no targeting peptide was identified. Moreover, no plastid-targeted m$^5$C rRNA methyltransferases have been identified in any other organisms. If m$^5$C rRNA methyltransferase does have a plastid origin in *P. falciparum*, the protein has
probably been co-opted for use by the cell rather than the organelle. Unfortunately, the patterns left by the evolution of m\textsuperscript{5}C RNA methyltransferases, are not informative for the determination of the origin of the m\textsuperscript{5}C RNA methyltransferase gene transfer in 
\textit{P. falciparum}.

2.3.2.8 RNA 3' terminal phosphate cyclase

A homologue of RNA 3' terminal phosphate cyclase was identified on chromosome 12 of \textit{P. falciparum}. This family of RNA-binding enzymes makes a cyclic phosphodiester at the end of RNA molecules [148]. The authors who originally described the \textit{E. coli} and human RNA 3' terminal phosphate cyclase noted at the time that similar proteins appear in all three domains of life [149, 150].

The \textit{P. falciparum} protein is 569 amino acids long, and has an 80 amino acid leader sequence that is identified as plastid targeting. A 90 amino acid insert is also found in the middle of the \textit{P. falciparum} protein. Two hundred thirty nine homologous positions were selected from the multiple sequence alignment and used to make the RNA 3' terminal phosphate cyclase phylogeny (figure 2.16).

A well-supported branch (bootstrap of 94 percent) divides the member species in two. The upper clade encompasses the bacteria and the \textit{P. falciparum} sequence, and the lower clade is comprised of the archaea, the eukaryota, and a lone species of actinobacteria. The \textit{P. falciparum} sequence branches well in the interior of the upper bacterial clade and is clearly more related to the bacteria than to the eukaryotes or archaea (figure 2.16). The bacterial clade is fully resolved and well supported, although a number of long branches are present. The long branches unite sequences that are not expected to group together by criteria of accepted phylogenetic
Figure 2.16 Phylogeny of RNA 3’ terminal phosphate cyclase. The P. falciparum sequence is boxed for reference. Higher order taxonomic groupings are shaded. See figure 2.3 for details.
relationships (see table 2.2) based on other genetic and morphological characters. *Bacillus halodurans* (a low GC firmicute) and *Deinococcus radiodurans* (a deinococcus) are grouped together by long branches with 98 percent bootstrap support. A 98 percent bootstrap separates *Listeria monocytogenes* from its fellow low GC firmicute, *Clostridium acetobutylicum*. A bootstrap value of 100 divides *L. monocytogenes* and *B. halodurans*, which are related at the family level. A 99 percent bootstrap value groups the *P. falciparum* sequence with the sequence from *L. monocytogenes*. It should be noted, however, that no sequence architecture features unite *P. falciparum* and *L. monocytogenes*.

The plastid localization signal in the *P. falciparum* sequence indicates that RNA 3' terminal phosphate cyclase functions in the apicoplast. However, the phylogeny shows a specific relationship of this gene with a low GC firmicute. Two explanations are possible for this conflicting evidence. 1) The gene may have been laterally transferred from an ancestor of the low GC Firmicutes into the nucleus of a progenitor of *P. falciparum*, and then subsequently acquired a targeting peptide that allowed it to function in the plastid. 2) The relationship of the *P. falciparum* and *L. monocytogenes* genes is artefactual, caused by long branch attraction, and the RNA 3' terminal phosphate cyclase functions in the compartment where it evolved. I favour this second hypothesis as being more biologically plausible, and I tentatively conclude that RNA 3’ terminal phosphate cyclase is a laterally transferred gene of plastid origin.
2.3.2.9 Ribosomal RNA adenine dimethylase

Ribosomal RNA adenine dimethylases are RNA modification enzymes that methylate two adjacent adenosines in rRNA. A gene was identified on chromosome 12 in *P. falciparum* that is similar to rRNA adenine dimethylases in other organisms. The group of homologues that were identified for this *P. falciparum* gene included archaeal, eukaryote and eubacterial sequences.

Both the *P. falciparum* and the *A. thaliana* genes possess N-terminal leader sequences (figure 2.17) that are predicted to be chloroplast-targeting peptides. The *P. falciparum* target peptide is long (180 amino acids) and contains a region of low complexity made of 10 copies of a hexamer repeat. A second region of low complexity, characterized by of long tracts of lysine and asparagine, comprises about half of the 160 amino acid insert in the *P. falciparum* gene (figure 2.17). In addition to the indels depicted in figure 2.17, the multiple alignment of this gene contains many small gaps and consists of blocks of well-conserved sites separated by divergent and gap-laden sequence. These divergent sequence sites were discounted and 177 sites remained to be used to calculate phylogenetic distances.

The distance tree for rRNA adenine dimethylase (figure 2.18) resolves clades that define major taxonomic groups such as the divisions of proteobacteria, the actinobacteria, the low GC Firmicutes, the spirochaetes, the chlamydiales, the cyanobacteria and the eukaryotes. There is no support, however, for the branches that connect these groups. The position of the plastid-bearing organisms, *A. thaliana* and *P. falciparum*, are poorly supported. In the absence of conflicting indications from the phylogeny, the presence of a plastid-targeting sequence is considered
Figure 2.17 Schematic alignment of rRNA adenine dimethylase sequences. The general architecture of the protein is shown for the three domains of life. Members of these domains whose proteins deviate from the general form are also shown. Open boxes represent residues and lines represent indels. The number of amino acids are given to the right of specific proteins. Shaded regions indicated low sequence complexity.
Figure 2.18 Phylogeny of Ribosomal RNA adenine dimethylase. The *P. falciparum* sequence is boxed for reference. Plastid-bearing organisms are indicated in bold. mt; mitochondrion-targeted. General notation as in figure 2.3.
evidence that the rRNA adenine dimethylase gene was transferred from the plastid to the nucleus of *P. falciparum*.

2.3.2.10 S-adenosyl methionine-dependant methyltransferase

A gene on chromosome 12 in *P. falciparum* was identified by BAEwatch as being 'bacteria-like'. This gene has a set of homologues for which function has not yet been determined. An S-adenosyl methionine (SAM) binding domain was identified in all the homologous proteins, which identifies them as a type of methyltransferase [151]. This is a very common domain found in almost 3000 proteins in the Interpro database [112]. The *E.coli* homologue of the *P. falciparum* gene has been shown to be active on membrane localized substrates [152].

The homologues for the *P. falciparum* sequence included three other eukaryotes amongst a mostly eubacterial protein set. *P. falciparum* and two of the other eukaryotes have sequences with N-terminal extensions (fig 2.19). In *P. falciparum*, this leader is predicted to be an apicoplast-targeting peptide. Two hundred thirty four positionally conserved residues were selected and used to construct a phylogeny (figure 2.20). The main taxonomic groupings of the eubacteria are resolved in the SAM methyltransferase tree, including proteobacterial groups alpha through epsilon, spirochaetes, low GC firmicutes (with the exception of *Mycoplasma genitalium*), chamydiales, and the cyanobacteria. The two animal sequences group together as sister to the alpha proteobacteria, and probably represent sequences transferred from the mitochondrion. The position of *A. thaliana* is unresolved.
Figure 2.19 Schematic alignment of SAM-dependant methyltransferase sequences. The top line represents the general form of all bacterial homologues. The eukaryotic sequences are shown underneath. Putative organelle-targeting peptides are cross hatched. The amino acid length of the protein is given to the right of the specific sequences.
Figure 2.20 Phylogeny of SAM-dependant methyltransferase. The sequence from *P. falciparum* is boxed for reference. Eukaryotic sequences are named in bold. Taxonomic affiliations are shown to the right of supported clades. See figure 2.3 for general notation.
The sequence from *P. falciparum* is paired with the cell wall-less, low GC firmicute, *Mycoplasma genitalium*, with 94 percent bootstrap support (figure 2.20). At face value, this relationship signifies that the methyltransferase gene from *P. falciparum* shared a common ancestor with the gene from *M. genitalium*. However, the apicoplast-localization peptide argues for a plastid origin for this gene. The branch lengths that join these sequences are long. A systematic error in phylogenetic analysis, such as results from applying an incorrect model of sequence evolution, will compound in long branches [116]. Bootstrap values address random sampling error, so while the high bootstrap value on the branch that unites *P. falciparum* and *M. genitalium* is compelling, it may have no significance if the model of phylogenetic inference used is incorrect. The conspicuous absence of the other two fully sequenced mycoplasma species may indicate that the function of this gene is not essential for the parasitic mycoplasmas. A loss of functional constraint in the *M. genitalium* sequence can lead to mutational saturation, and thus the apparent relationship to *P. falciparum*. For these reasons, a plastid origin is favoured for this SAM methyltransferase, although it deserves further study.

2.3.2.11 GTP-binding protein

Two of the thirteen 'bacteria-like' genes identified by BAEwatch had matching sets of homologues. These two proteins are not identical. They show marked differences in architecture (figure 2.21) and do not represent a recent duplication. The function of the proteins considered here is unknown. The sequences possess two copies of a characteristic GTP-binding motif, but little is known beyond that, as no functional studies of these proteins have been carried out. As part of a much larger

74
study examining the evolution of GTPases, Liepe et al. [127] noticed the similarity between these proteins and named the group EngA.

The homologues for this gene are primarily bacterial, but four other eukaryotic sequences were also identified. Three of the six eukaryotic sequences had N-terminal extensions; these were *P. falciparum* 1, *A. thaliana* 1, and *G. theta* (fig 2.21). The leader peptide of *P. falciparum* 1 was very long (381 amino acids). Except for the first 50 amino acids, the leader is comprised of low complexity sequence, containing 13 tandem repeats of a heptamer and several poly-lysine stretches. Nevertheless, this leader was identified as a plastid targeting peptide, as were the leaders of *A. thaliana* 1 and *G. theta*. Of the other three eukaryotic sequences, *P. falciparum* 2 and *Oryza sativa* did not contain N-terminal extensions, and no targeting peptides were identified. The *A. thaliana* 2 sequence was predicted to have a mitochondrion-targeting peptide of length 118. One hundred and eighteen amino acids is more than a fifth of the total protein. Moreover, the first 118 amino acids share considerable similarity to the sequences of the other homologues, including bacteria (figure 2.21). If the *A. thaliana* 2 sequence does have a 118 amino acid mitochondrion-localizing peptide, it is highly atypical.

The phylogeny constructed with this group of proteins shows that eukaryotic sequences have one of three associations (figure 2.22). First, the *O. sativa* and *A. thaliana* 2 sequences are well supported as a sister group to the alpha proteobacteria. These genes were most likely transferred from the mitochondrion to the nucleus. Second, the *P. falciparum* 1, *G. theta* and *A. thaliana* 1 sequences form a sister group to the cyanobacteria. The evidence presented here suggests that these genes were
Figure 2.21 Schematic alignment of GTP binding protein sequences. The general architecture of this primarily eubacterial protein is shown in the first line. Open boxes represent residues and lines represent indels. The length of the peptide is given at the end of the protein; the value for Eubacteria is an average for all eubacterial homologues. The values given in front of the sequences are the lengths of N-terminal leaders. *: cleavage site of the predicted mitochondrion-targeting peptide of A. thaliana 2 (see text).
Figure 2.22 Phylogeny of GTP-binding protein. The sequences from *P. falciparum* are boxed for reference. Eukaryotes are shown in bold. Details as in figure 2.3.
laterally-transferred from the plastid to the nucleus, in the case of *P. falciparum* and *A. thaliana*, and to the nucleomorph in the case of *G. theta*. Third, the *P. falciparum* 2 sequence groups with *Treponema pallidium*, a spirochete. This grouping is not well supported, and the very long branch of the *P. falciparum* 2 sequence casts suspicion on this placement. The *P. falciparum* 2 sequence has undoubtedly been transferred from a bacterium or organelle (because no archaeal or eukaryotic homologues exist), but there is not enough information to determine the likely donor of the sequence.

With the possible exception of the *P. falciparum* 2 sequence, the taxonomic distribution of the protein (figure 2.22) is limited to the eubacteria and those eukaryotes that have acquired the gene from an organelle. The *O. sativa* and *A. thaliana* sequences are the sole representatives of a mitochondrial homologue for this protein. The absence of other fully sequenced eukaryotes in the tree is interesting as it may mean that the mitochondrial gene transfer event only took place in plant lineage. Certainly, the animal and fungal groups with fully sequenced genomes do not encode a representative of the gene. We await the sequencing of additional eukaryotes from outside the higher eukaryotes, to determine when this transfer occurred.

### 2.3.3 Conclusion: Lateral gene transfer in *P. falciparum*

In total, thirteen *P. falciparum* genes were analysed for evidence of lateral gene transfer. Of these, eight (type II topoisomerase B subunit, type II topoisomerase A subunit, ribosome release factor, pseudouridine synthase, RNA 3’ terminal phosphate cyclase, rRNA adenine dimethylase, SAM dependant methyltransferase, and GTP-binding protein 1) probably represent transfers from the plastid to the nucleus of
an ancestor of *P. falciparum*. Two additional genes show relationships that are more complex and while the evidence for plastid origin is weak, the data do not exclude a hypothesis of plastid origin (rRNA methyltransferase, elongation factor G). One gene is likely to have been transferred from the mitochondrion to the nucleus in an ancestor of *P. falciparum* (proteasome subunit HslV). A further one gene has equivocal evidence for a mitochondrion origin, but such an interpretation of the phylogenetic data is not ruled out (GTP-binding protein 2). The remaining one gene of thirteen (adenylosuccinate lyase) provides evidence of other lateral gene transfers that do not include *P. falciparum*. I conclude that the majority of laterally transferred genes in the *P. falciparum* genome are of organellar origin. Based on this study of a subset of 2073 *P. falciparum* genes, I do not predict that many non-organellar gene transfers will be observed in the complete genome when it has been fully sequenced.

The paucity of laterally transferred genes may be a consequence of *P. falciparum*’s parasitic lifestyle. The reduction of genome size that often accompanies transition to parasitism and an intracellular lifestyle [153, 154] may preclude lateral gene transfer to *P. falciparum*, unless the acquired gene replaces an essential gene already present.


100. Waller RF, Reed MB, Cowman AF, McFadden GI: Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *Embo J* 2000, 19:1794-802.


103. [www.sanger.ac.uk/Projects/P_falciparum/](http://www.sanger.ac.uk/Projects/P_falciparum/) Sequencing of *P. falciparum* chromosomes 1,3,4,5,6,7,8,9, and 13 was accomplished as part of the malaria Genome Project with support by the Wellcome Trust.

104. [www-sequence.Stanford.edu/group/malaria](http://www-sequence.Stanford.edu/group/malaria) Sequencing of *P.falciparum* chromosome 12 was accomplished as part of the Malaria Genome Project with support by the Burroughs Wellcome fund.

105. [www.tigr.org](http://www.tigr.org) Sequencing of chromosomes 2, 10, 11, 14 was part of the International Malaria Genome Sequencing Project and was supported by awards from the U.S. Department of Defense.


109. SWISSPROT and TrEMBL databases: http://www.ebi.ac.uk/swissprot/.


132. Zupan J, Muth TR, Draper O, Zambryski P: The transfer of DNA from
agrobacterium tumefaciens into plants: a feast of fundamental insights.

133. Ohnishi M, Janosi L, Shuda M, Matsumoto H, Hayashi T, Terawaki Y, Kaji A:
Molecular cloning, sequencing, purification, and characterization of
Pseudomonas aeruginosa ribosome recycling factor. *J Bacteriol* 1999,
181:1281-91.

134. De Mot R, Nagy I, Walz J, Baumeister W: Proteasomes and other self-
compartmentalizing proteases in prokaryotes. *Trends Microbiol* 1999, 7:88-
92.

R, Baumeister W: The first characterization of a eubacterial proteasome: the


Chlamydomonas chloroplast clpP gene contains translated large insertion

138. Schaller A, Ryan CA: Molecular cloning of a tomato leaf cDNA encoding an
aspartic protease, a systemic wound response protein. *Plant Mol Biol* 1996,
31:1073-7.

Cavalier-Smith T, Maier UG: The highly reduced genome of an enslaved algal

140. de Sagarra MR, Mayo I, Marco S, Rodriguez-Vilarino S, Oliva J, Carrascosa JL,
Castan JG: Mitochondrial localization and oligomeric structure of HClpP, the

141. Zou CB, Nakajima-Shimada J, Nara T, Aoki T: Cloning and functional
expression of Rpn1, a regulatory-particle non-ATPase subunit 1, of
proteasome from Trypanosoma cruzi. *Mol Biochem Parasitol* 2000, 110:323-
31.

modifying enzymes by iterative genome search using known modifying


