# RESEARCH ARTICLES

## Evolutionary Pressures on Apicoplast Transit Peptides

*Stuart A. Ralph,*[*1] *Bernardo J. Foth,*[*2] *Neil Hall,*[†3] *and Geoffrey I. McFadden**

* Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville, Victoria, Australia; and
† The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

Malaria parasites (species of the genus *Plasmodium*) harbor a relict chloroplast (the apicoplast) that is the target of novel antimalarials. Numerous nuclear-encoded proteins are translocated into the apicoplast courtesy of a bipartite N-terminal extension. The first component of the bipartite leader resembles a standard signal peptide present at the N-terminus of secreted proteins that enter the endomembrane system. Analysis of the second portion of the bipartite leaders of *P. falciparum*, the so-called transit peptide, indicates similarities to plant transit peptides, although the amino acid composition of *P. falciparum* transit peptides shows a strong bias, which we rationalize by the extraordinarily high AT content of *P. falciparum* DNA. 786 plastid transit peptides were also examined from several other apicomplexan parasites, as well as from angiosperm plants. In each case, amino acid biases were correlated with nucleotide AT content. A comparison of a spectrum of organisms containing primary and secondary plastids also revealed features unique to secondary plastid transit peptides. These unusual features are explained in the context of secondary plastid trafficking via the endomembrane system.

### Introduction

The majority of proteins located in mitochondria and in the chloroplasts (plastids) of plants and algae are encoded in the nuclear genome and targeted posttranslationally to these organelles. Eukaryotic cells, thus, express hundreds of proteins in their cytosol that are translocated to either mitochondria or plastids courtesy of N-terminal sequence extensions called transit peptides. Transit peptides (both plastid and mitochondrial) are enriched for basic residues, but their primary sequence and length are highly diverse, and no consensus motifs exist. Mitochondrial transit peptides typically form amphipathic α-helices, but plastid transit peptides are not known to assume any consistent secondary structure (especially in aqueous environment), and their broad size range (from 20 to more than 100 amino acids) is indicative of a very plastic composition. Nevertheless, in plant cells, these two different classes of transit peptides direct the corresponding proteins from the cytoplasm into either the mitochondrion or the plastid with high fidelity.

Some algae and protists contain plastids with more than two membranes that derive from an evolutionary process referred to as *secondary* endosymbiosis. Secondary endosymbiosis results from a plastid-bearing eukaryote (an alga) being taken up by another, unrelated eukaryote. One such secondary endosymbiotic event is hypothesized to have given rise to the group now known as the Chromalveolata (Cavalier-Smith 1999; Fast et al. 2001; Harper and Keeling 2003). Extant examples include algae such as cryptomonads, diatoms, and dinoflagellates, as well as obligate parasites belonging to the phylum Apicomplexa, which are responsible for many diseases of great medical and veterinary importance such as malaria and toxoplasmosis. In all of these cases, the secondary plastid is enclosed by multiple membranes, the outermost of which is apparently a derivative of the original phagocytic vacuole. As a result, the first step of targeting proteins to secondary plastids is entry into the endomembrane system, which is achieved courtesy of an N-terminal signal peptide that precedes the transit peptide (Bhaya and Grossman 1991; Yung and Lang-Unnasch 1999; DeRocher et al. 2000; Ishida, Cavalier-Smith, and Green 2000; Waller 2000; Wastl and Maier 2000; van Dooren et al. 2001; Apt et al. 2002). In these organisms, plastid-targeting leader sequences, thus, consist of two separate parts (i.e., they are bipartite). Once inside the endomembrane system, the signal peptide is removed and the transit peptide then mediates a diversion from the default endomembrane secretion pathway into the plastid, crossing the two or three inner plastid membranes in the process (van Dooren et al. 2001).

Two unrelated groups of algae with secondarily derived plastids have one very intriguing characteristic in common. Both cryptophyte and chlorarachniophyte algae contain a remnant nucleus (nucleomorph) from the vestigial primary endosymbiont (Douglas et al. 1991; McFadden et al. 1994), and these nucleomorphs also encode plastid proteins (Douglas et al. 2001; Gilson and McFadden 2002). These algae, therefore, possess two genomes encoding proteins that have to be imported posttranslationally into the plastid: the nucleus encodes proteins with a bipartite leader (consisting of signal and transit peptide), whereas the nucleomorph encodes proteins with only a one-part (monopartite) leader sequence (representing a transit peptide).

The plastid of apicomplexan parasites (the "apicoplast") is of secondary endosymbiotic origin and is often referred to as a relict plastid because it contains no photosynthetic apparatus. However, the apicoplast has developed into somewhat of a model system for protein targeting

[1] Present address: Unité de Biologie des Interactions Hôte-Parasite, Institut Pasteur, Paris, France.

[2] Present address: Départment de Microbiologie et Médicine Moleculaire, CMU-Faculté de Médicine, Université de Genève, CH-1211 Genève 4, Switzerland.

[3] Present address: The Institute for Genomic Research, Rockville, Maryland.

Key words: *Plasmodium falciparum*, plastid, apicoplast, targeting, transit peptide, nucleotide bias.

E-mail: gim@unimelb.edu.au.

*Mol. Biol. Evol.* 21(12):2183–2194. 2004
doi:10.1093/molbev/msh233
Advance Access publication August 18, 2004

to secondary plastids for two reasons. First, *Toxoplasma gondii* and *Plasmodium falciparum* are amenable to transgenic experimentation, which has already resulted in several in vivo mutagenesis studies that directly addressed issues of organellar protein targeting (DeRocher et al. 2000; Waller et al. 2000; Yung, Unnasch, and Lang-Unnasch 2001; Foth et al. 2003). Second, EST and genome sequencing projects are underway for several apicomplexan parasites such as *T. gondii*, *Theileria parva*, and several species of the genus *Plasmodiu*m, which now provides us with the first large data sets of plastid transit peptides from nonplant organisms. Importantly, the complete nuclear sequence of both *P. falciparum* and *P. yoelii* have recently been published, providing by far the largest collection of secondary plastid proteins to date (Carlton et al. 2002; Gardner et al. 2002).

In this paper, we have taken advantage of the availability of hundreds of putative transit peptides from the malaria parasite *P. falciparum* and from the model plant *Arabidopsis thaliana* to conduct a comprehensive analysis of the amino acid composition of these targeting leader sequences. Smaller data sets of secondary and primary plastid proteins from other parasites, algae, and plants were also assembled. Our analysis shows that amino acid content of transit peptides differs markedly between these groups. However, rather than suggesting functional divergence, our analyses indicate that the amino acid differences are a function of the AT content of the corresponding DNA, which ranged from 42% to 77%. The notable exception of this overall trend was the content of serine and threonine, which may reflect a functional difference between the targeting mechanisms to plastids of primary and secondary origin. The analyses show that an excess of basic over acidic residues and a great variability in primary sequence are hallmarks of plastid transit peptides in general.

## Materials and Methods

### Data Sources—Bipartite Leaders

*P. falciparum* (90 Plastid-Targeted Proteins)

*P. falciparum* genomic sequence was obtained from the following sequencing centers: The NCBI *P. falciparum* Blast Database (http://www.ncbi.nlm.nih.gov/Malaria/blastindex.html), The Sanger Institute *P. falciparum* blast server (http://www.sanger.ac.uk/Projects/P_falciparum/blast_server.shtml), the TIGR parasites database (http://www.tigr.org/tdb/e2k1/pfa1/), Stanford Genome Technology center (http://sequence-www.stanford.edu/group/malaria/), and PlasmoDB (http://www. plasmodb.org). Gene predictions from Genefinder (P. Green, C. Wilson, L. P. Hilyer, http://ftp.genome.washington.edu/cgi-bin/Genefinder), Glimmer, (Salzberg et al. 1999), and Phat (Cawley et al. 2001) were inspected, as were predictions from bulk sequences.

*P. yoelii* (41 Plastid-Targeted Proteins)

Preliminary sequence data from the *Plasmodium yoelii* genome were obtained from The Institute for Genomic Research Web site (http://www.tigr.org/tdb/e2k1/pya1/). This sequencing program is carried on in collaboration with the Naval Medical Research Center and is supported by the U.S. Department of Defense.

*P. knowlesi* (28 Plastid-Targeted Proteins)

*P. knowlesi* sequence data were produced by the *P. knowlesi* Sequencing Group at the Sanger Institute and can be obtained from (ftp://ftp.sanger.ac.uk/pub/pathogens/P_knowlesi/).

*Theleria parva* (20 Plastid-Targeted Proteins)

*T. parva* preliminary sequence data was obtained from the Institute for Genomic Research Web site at (http://www.tigr.org/tdb/e2k1/tpa1/). Sequencing of *Theileria parva* was funded by the International Livestock Research Institute and the Institute for Genomic Research.

*Theileria annulata* (20 Plastid-Targeted Proteins)

*T. annulata* sequence data were produced by the *T. annulata* Sequencing Group at the Sanger Institute and can be obtained from (ftp://ftp.sanger.ac.uk/pub/pathogens/T_annulata/).

*Toxoplasma gondii* (14 Plastid-Targeted Proteins)

Preliminary genomic and cDNA sequence data was accessed via http://ToxoDB.org. Genomic data were provided by the Institute for Genomic Research (supported by the NIH grant #AI05093), and by the Sanger Center (Wellcome Trust). EST sequences were generated by Washington University (NIH grant #1R01AI045806–01A1).

*Guillardia theta* (Nuclear)

Four nuclear *G. theta* plastid proteins were obtained from GenBank. For all bipartite leaders, sequences were submitted to SignalP version 2 (NN) and the transit peptide was defined as commencing at the first amino acid after predicted signal peptide cleavage. The first 20 amino acids from this residue were considered for further amino acid analysis.

### Data Sources—Monopartite Leaders

From the more than 2,500 proteins predicted to be chloroplast targeted in the *Arabidopsis* genome (The Arabidopsis Genome Initiative 2000), a redundancy-reduced, high-confidence subset of 500 proteins was curated. High confidence ($> 0.95$) TargetP predictions were used to assemble similar data sets from SWISSPROT entries for *O. sativa* (150 proteins), *H. vulgare* (40 proteins), and *G. max* (57 proteins).

*Guillardia theta* (Nucleomorph)

Thirty two *G. theta* nucleomorph plastid proteins are those identified from the nucleomorph sequencing project (Douglas et al. 2001). For all monopartite transit peptides, the initial dipeptide was removed as per von Heijne and Nishikawa (1991), and the subsequent 20 amino acids were subjected to further analysis.

### Data Sources—Mature Proteins

Data sets of mature proteins were created by assembling the c-terminal 20 amino acids of each protein in the transit peptide data sets.

*Prediction of Apicoplast-Targeted Proteins*

Putative nuclear-encoded apicoplast protein precursors were inferred from similarity to known plastid proteins by pairwise sequence alignment using TBlastN version 2.1.3 (BLOSUM62; gap existence cost 11; per gap cost 1; Lambda ratio 0.85) (Altschul et al. 1997). Matches were subjected to further inspection generally if the EXPECT value was less than $10^{-5}$, but some apparently genuine matches did exceed this value. Care was taken to reduce false-positive identification of apicoplast proteins. Once a putative apicoplast protein was identified, the encoding sequence was inspected for likely intron/exon boundaries, and the resulting ORF was used to search the nonredundant GenBank translated Coding Sequence database. Proteins were only considered to be putatively apicoplast targeted if the top match was to another plastid protein or, occasionally, to a bacterial protein whose function was known to be found in other plastids. Proteins were excluded if they contained a mitochondrial transit peptide, as predicted by PSORT (Nakai and Kanehisa 1992) or TargetP (Emanuelsson et al. 2000), or if they contained no N-terminal extension when compared with any other predicted matching protein. To minimize false positives further, matches were also excluded if there was conflicting data regarding their localization or if they were members of a family for which contrary localization evidence existed. All matches where the coding sequence was radically different from the average nucleotide content of that species were discarded as possible contaminants of the sequencing project.

*Amino Acid Content Analysis*

Average amino acid compositions for sequence data sets were determined by calculating the amino acid compositions for the individual sequences first, and then averaging these values. This procedure was implemented to prevent longer sequences from disproportionately biasing the analysis. Statistical significance of differences between the average number of occurrences of different amino acids was tested for all data sets, except the small *G. theta* nuclear data set. Differences in the first 20 amino acids of transit peptides were tested for statistical significance using Poisson regression analysis with $\alpha = 0.01$ (type I error rate).

*Nucleotide Content Analysis*

For a subset of 50 of the 90 apicoplast transit peptides, the relevant coding nucleotide sequence was retrieved and analyzed for usage of each nucleotide. Average nucleotide content for sequence data sets was determined as for amino acid compositions above (calculations for individual sequences first, followed by averaging these values). Statistical significance was tested by performing Monte Carlo resampling: 1,000 pseudo data sets, each consisting of 50 sequences with the same length distribution as the 50 transit peptides, were created by randomly sampling codons from the sequence region corresponding to the mature protein of those same sequences. The nucleotide composition was calculated for every one of the 1,000 pseudo–data sets, and the mean and standard deviation for each of the four nucleotides was determined. Content of A,T,C, and G from the 50 full-length transit peptides was considered significantly ($p < 0.01$) different from that encountered in the mature protein regions of the same 50 sequences if it lay outside the 99% confidence interval defined as mean $\pm$ 2.576 standard deviation.

Transit Peptide Cost Estimation

The average cost for each amino acid in the transit peptide was estimated according to the metabolic costing (in high-energy P-bonds) if it were created from common precursor metabolites (given by Akashi and Gojobori [2002]). Cost is given as an average P-cost per amino acid in the transit peptide for each species.

*Sequence Logo*

To examine the positional distribution of amino acids along bipartite transit peptides, 20 transit peptide sequences from each chromalveolate species were gathered. Ten amino acids on each side of the predicted SP-cleavage site (SignalP version 2) were forcibly aligned with no gaps and a sequence logo generated from the total alignment using the Web server at http://www.cbs.dtu.dk/gorodkin/appl/plogo.html.

*Additional Software*

Amino acid compositions were determined using our own application written in Java (available upon request). Potential alpha helices were constructed and viewed using the helical wheel applet at http://cti.itc.Virginia.EDU/~cmg/Demo/wheel/wheelApp.html.

## Results and Discussion

To compare amino acid usage in transit peptides from different plastid-bearing organisms, data sets of putatively plastid targeted proteins were assembled for six apicomplexan parasites, two secondary endosymbiont nucleomorphs, two dicotyledon, and two monocotyledon land plants. Although the usage of some amino acids is similar between groups, others are hugely varied between the species (table 1). We wished to address the reasons for these dissimilarities. Systematic differences in amino acids can be caused by a number of factors, including nucleotide biases in the encoding DNA, divergence of function, different environmental availability of an amino acid or its precursors, or differential metabolic cost in synthesizing or employing certain amino acids. We first looked for correlations between the nucleotide and amino acid biases between the studied groups.

Genomic AT Content As the Major Driving Force for Amino Acid Bias in Plastid-Targeting Transit Peptides

The malaria parasite *P. falciparum* has the most AT-rich genome yet reported, with an overall AT content of 82% (Weber 1987, Gardner et al. 2002). Although the genome is enriched in both adenine and thymine, the coding strand is more enriched in adenine than in thymine. In addition to high AT content at the third codon position

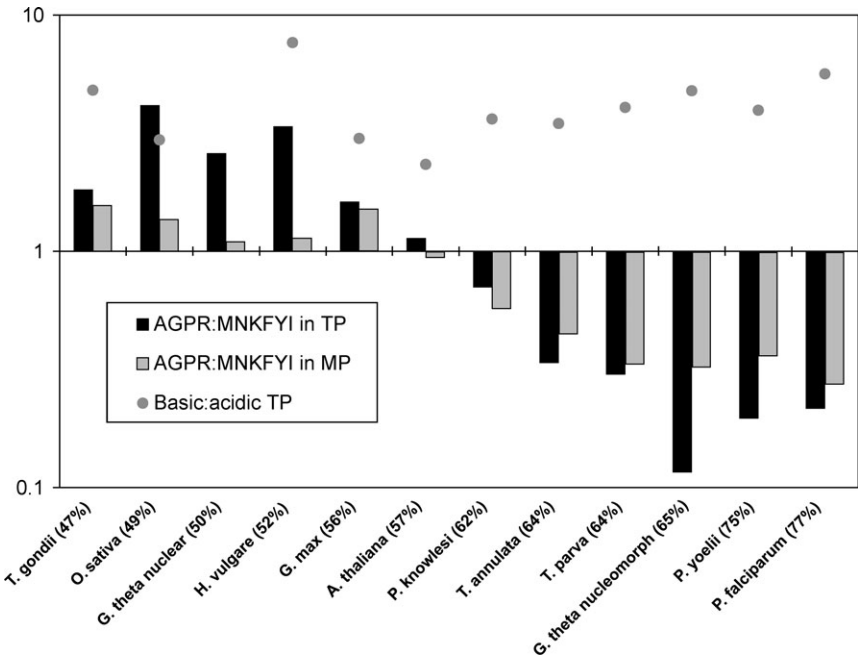


FIG. 1.—Correlation between genomic AT content and amino acid composition of plastid-targeted nuclear encoded genes. Average ratio (AGPR:MNKFYI) in transit peptide and mature proteins of amino acids encoded by GC-rich codons (Ala, Gly, Pro, and Arg) to amino acids encoded by AT-rich codons (Met, Asn, Lys, Phe, Tyr, and Ile). The ratio of basic (Lys, Arg, and His) to acidic (Asp and Glu) amino acids is shown by gray dots. Species are ranked by coding DNA AT content, with AT-poor genomes on the left and AT-rich genomes on the right. The average AT content in coding sequences is given in parentheses. Although overall amino acid charge is unresponsive to the AT bias, the AGPR:MNKFYI ratio falls progressively as the AT content of the coding DNA increases. Transit peptides are more sensitive to this nucleotide pressure than are mature proteins. Please note that no acidic residues are present in the nuclear-encoded *G. theta* transit peptides, so the basic:acidic ratio is undefined.

(Saul and Battistutta 1988), *P. falciparum* preferentially uses codons that contain A or T at any position (Musto, Rodriguez-Maseda, and Bernardi 1995). Accordingly, this extreme bias drives a genome-wide trend towards usage of those amino acids encoded by AT-rich codons (Met, Asn, Lys, Phe, Tyr, and Ile) and away from those requiring GC-rich codons (Ala, Gly, Pro, and Arg) (Verra and Hughes 1999; Singer and Hickey 2000). When the amino acid content of the apicoplast transit peptides from *P. falciparum* was compared with that of plastid transit peptides from *A. thaliana*, this bias was obvious: apicoplast transit peptides are significantly ($P < 0.05$, Poisson regression analysis [see *Materials and Methods*]) richer in Asn, Lys, Phe, Tyr, and Ile, and depleted in Ala, Gly, Pro, and Arg, as measured by the ratio of GC-favored to AT-favored amino acids (AGPR:MNKFYI; black bars in figure 1). The same trend was observed in other chromalveolates with AT-rich genomes; for example, in *P. yoelii* and nucleomorph-encoded genes of the cryptophyte *G. theta* (fig. 1). Figure 1 shows that the extent of this trend correlates well with AT-content: the more AT-rich the genome, the lower the AGPR:MNKFYI ratio. It is remarkable that this pattern was observed even within one genus (compare *P. falciparum*, *P. yoelii*, and *P. knowlesi*). Similarly, the same trend and correlation between amino acid composition and AT content was found for the transit-peptide sequences from the four plant species included in the analysis (*Oryza sativa* [rice], *Hordeum vulgar*e [barley], *Glycine max* [soybean], and *Arabidopsis thaliana* [cress] [fig. 1]). Indeed, dicotyledenous plants such as *A. thaliana* and soybean, *G. max*) commonly have a higher AT content than monocots (such as rice, *O. sativa*, and barley, *H. vulgare*) (Carels and Bernardi 2000; Initiative 2000; Yu et al. 2002), and the chloroplast transit peptides of these four plant species show the anticipated consequence for codon usage: transit peptides of rice and barley are enriched in amino acids encoded by GC-rich codons (Ala, Gly, Pro, and Arg) and depleted in those encoded by AT-rich codons (Met, Asn, Lys, Phe, Tyr, and Ile [fig. 1]) compared with those of *A. thaliana* and *G. max*. These comparisons were all conducted using the N-terminal 20 amino acids of each transit peptide. This region is both the most important for targeting and the region that can be most confidently predicted as being part of the transit peptide.

The same trends and correlations for AT content and amino acid composition were observed when amino acids were analyzed individually (tables 1 and 2, figure 2). Figure 2 presents the amino acid compositions for the transit peptides of the various plants and chromalveolates in comparison with those of *O. sativa*, which represented the most GC-rich plant species and the organism with the highest AGPR:MNKFYI ratio in the analysis (see figure 1). The figure shows that the expected trends—correlation of increasing AT content with depletion of alanine, glycine, proline, arginine and enrichment in methionine, asparagine, lysine, phenylalanine, tyrosine, isoleucine—were generally encountered in all of the AT-biased amino acids, except methionine (fig. 2). In contrast, amino acids encoded by codons that are, on the whole, AT neutral (Asp, Cys, Gln, Glu, His, Leu, Ser, Thr, Trp, and Val) generally did not exhibit strong trends. A comparison between *A. thaliana* and *P. falciparum* transit peptides

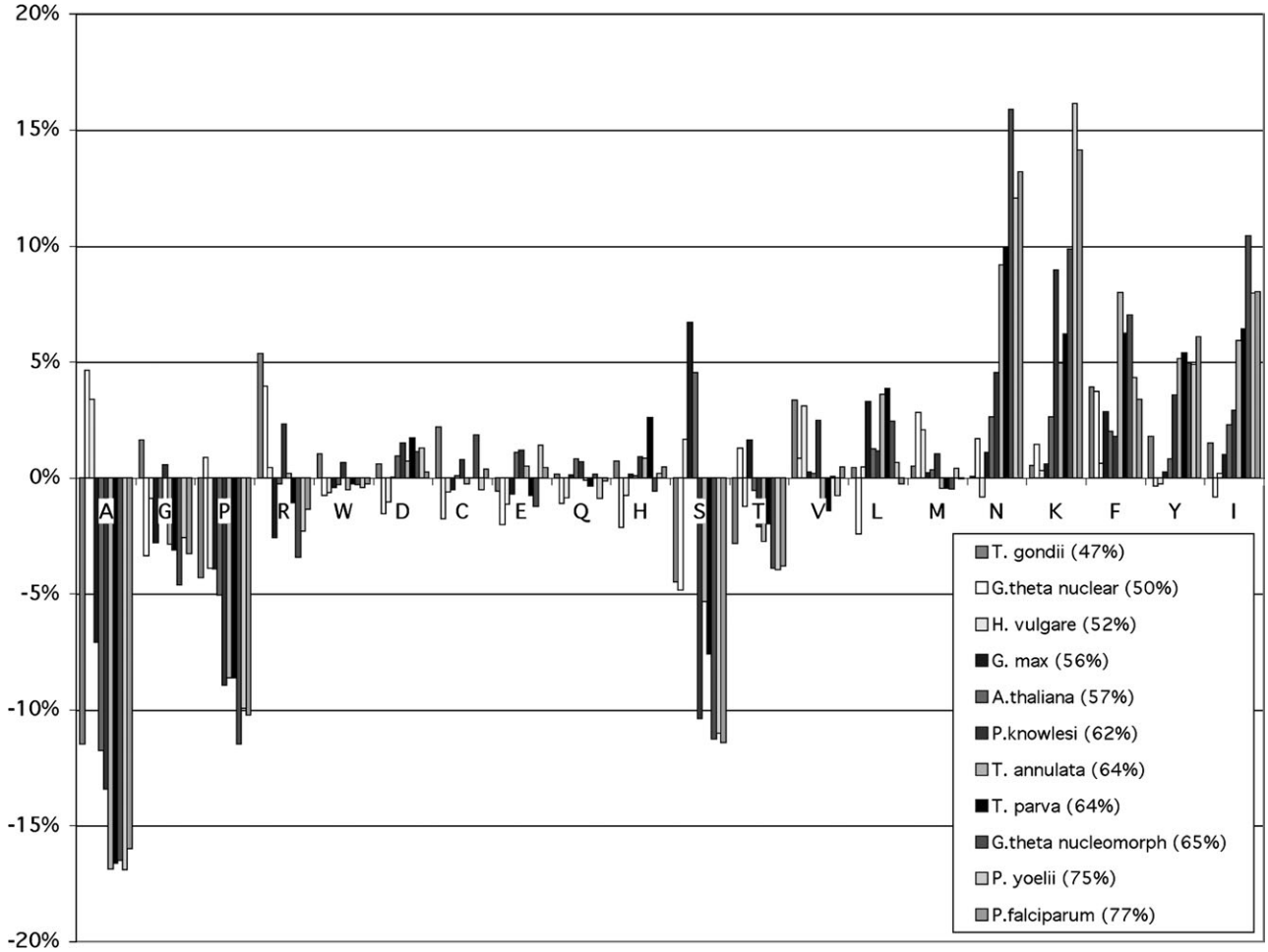


FIG. 2.—Relative abundance of amino acids in transit peptides. Amino acid content from the transit peptides of each species is compared with the amino acid composition of *O. sativa* transit peptides, the plant with the most GC-rich genome in this analysis. Amino acids are ranked from left to right in order of overall codon AT content. Within each amino acid, species are ranked from left to right in order of genomic AT content (denoted as percentage value next to species name). In general, amino acid usage correlates well with AT content, but serine and threonine are depleted in all chromalveolate transit peptides, regardless of nucleotide bias.

showed that only amino acids encoded by codons with a strong AT bias differed significantly ($P < 0.05$, Poisson regression analysis [see *Materials and Methods*]) in regard to their average abundance. Significant differences were not observed in most amino acids with AT-neutral codons. The notable exceptions were serine and threonine, which will be discussed below.

The connection between nucleotide and amino acid biases is also highlighted by comparisons between the transit peptide and mature-protein portions of genes (tables 1 and 2, figure 2). The correlation between amino acid composition and AT content depicted in figure 1 is generally more pronounced in the transit peptides (black bars) than in the corresponding mature-protein sequences (gray bars). This finding is consistent with transit peptides being generally less conserved and more variable than other protein sequences (von Heijne and Nishikawa 1991; Theg and Geske 1992; Bruce 2000, 2001). Whenever a global nucleotide bias is acting on a genome, for any given sequence a balance exists between compositional pressure towards that nucleotide bias and counteracting forces such as selection for protein functionality and/or three-dimensional structure. This means that where functional and structural constraints are lenient, nucleotide biases may be more exaggerated. Consequently, in a genome with strong nucleotide bias, the magnitude of the codon bias for an individual sequence is inversely correlated to its degree of conservation. As in other organisms, this tendency has been observed in some weakly conserved *P. falciparum* protein sequences (Singer and Hickey 2000), as well as in introns and intergenic regions (Gardner et al. 2002).

In the particular case of *P. falciparum* transit peptides, we examined a collection of 90 putative apicoplast-targeted proteins and found that usage of each nucleotide differed significantly between the sequences encoding the transit peptides and the sequences encoding the mature proteins (MPs), with the transit peptide–encoding sequences considerably more AT rich (78.0% versus 74.7%). In our analysis (see *Materials and Methods*), Monte Carlo resampling (1,000 pseudo–data sets) from the nucleotides coding for the mature protein never created a pseudo–data set with an average AT content exceeding 76%, demonstrating a significant

departure from the AT content of the transit peptides. Furthermore, in 90% of the apicoplast proteins examined, the DNA encoding the transit peptide was indeed more AT rich than that encoding the MP. This demonstrates an augmentation of nucleotide bias in transit peptides compared with mature proteins.

In general, codon usage in genes and, as a result, amino acid compositions of proteins are known to be susceptible to genome-wide nucleotide biases. For most genes of the extremely AT-rich malaria parasite *P. falciparum*, genome-wide compositional bias has been reported to be the major determinant of codon usage, although codons ending with C (cytosine) are more common in a few highly expressed genes (Musto, Rodriguez-Maseda, and Bernardi 1995; Musto et al. 1997, 1999). Despite detailed documentation of nucleotide bias in *Plasmodium* species (Musto, Rodriguez-Maseda, and Bernardi 1995; Musto et al. 1997, 1999; Verra and Hughes 1999; Weber 1987), the underlying cause of genomic AT bias is unclear. Differential abundance of redundant tRNAs has been evoked in some organisms as a driver of codon biases (Moriyama and Powell 1997), but the completion of the *P. falciparum* genome has revealed minimal tRNA redundancy (Gardner et al. 2002), discounting this explanation. Nucleotide biases can also be caused or modulated by gene-expression levels, DNA conformational structure or codon-anticodon interaction energies (Akashi 2001; Knight, Freeland, and Landweber 2001; Singer and Hickey 2000), but these forces have previously only been invoked to explain more modest biases. Singer and Hickey (2000) suggest that the high AT content may be the result of nonrandom mutational bias. It is not obvious why some *Plasmodium* species have strong AT biases and others weaker biases, nor is there a strict connection between AT bias and the relatedness of species (Perkins and Schall 2002; Rathore et al. 2001). If directional mutation is indeed the basis for AT bias, it could be caused by aberrant or atypical mutational repair mechanism. There is no obvious omission in DNA repair enzymes from the *P. falciparum* genome (Gardner et al. 2002), but DNA repair is only poorly characterized in any protist. Perhaps further study in this area will unearth mutational mechanisms that explain differential nucleotide biases.

## Are Transit Peptide Differences Between Organisms Caused by Functional Differences?

A possible explanation for the differential amino acid usage in transit peptides from different taxa is that they are functionally different. However, several lines of evidence strongly suggest that transit peptides from all groups are functionally equivalent. First, all organisms included in our analyses likely use the same protein import mechanism and a similar import apparatus to translocate proteins across the two (inner) plastid membranes because exchange experiments have demonstrated cross compatibility of transit peptides (DeRocher et al. 2000, Wastl and Maier 2000). For instance, the transit peptide portion of bipartite leaders from different chromalveolates such as diatoms, cryptomonads, and *Toxoplasma gondii* are sufficient to mediate import into plant chloroplasts in vitro, confirming the similarity of the protein import mechanisms of plant and chromalveolate plastids (Apt, Hoffman, and Grossman 1993; Lang, Apt, and Kroth 1998; Wastl and Maier 2000). Furthermore, the cryptomonad *G. theta* provides an interesting "internal control" because it has two sets of transit peptides, each with very different AT contents. In cryptomonads, plastid transit peptides are found both on proteins from the nuclear genome (where they are preceded by a signal peptide) and on proteins in the nucleomorph (where the targeted protein bears a simple monopartite transit peptide because the trafficking system is only required to cross two membranes (Wastl and Maier 2000)). The cryptomonad nucleus has no apparent AT bias, but the nucleomorph is highly AT rich, and the amino acid contents of the two sets of transit peptides show the anticipated consequences. Nucleomorph-encoded transit peptides are enriched for amino acids encoded by AT-rich codons, whereas nuclear-encoded transit peptides are enriched for amino acids encoded by GC-rich codons (fig. 1). These two highly divergent sets of transit peptides likely converge and interact with the same protein import machinery (for a possible caveat see Wastl and Maier [2000]), despite the fact that their amino acid content is markedly different. Thus, we hypothesize that the nucleomorph transit peptides have evolved to use a biased set of amino acids as a result of an increase in AT content in nucleomorph DNA.

It is worth noting here that the 32 nucleomorph-encoded transit peptides from *G. theta* (Douglas et al. 2001) are remarkably similar in amino acid composition to transit peptides of *Plasmodium falciparum* (Foth et al. 2003). Indeed, the *Plasmodium falciparum* transit peptide recognition tool PlasmoAP (Foth et al. 2003) gives high confidence scores for these cryptomonad transit peptides (24 of the 32 were recognized as secondary transit peptides), most probably because of convergence of amino acid content. In other words, the parallel bias for extraordinarily high AT content in these two genomes has resulted in similar amino acid usage profiles for their encoded transit peptides. Hopefully a *G. theta* EST project will provide a larger data set of nucleus-encoded (balanced AT content) transit peptides that can more robustly address this question in the future. Interestingly, the nonchromalveolate alga *Chlorarachnion* also retains a nucleomorph that encodes some plastid-targeted genes. Again, the nucleomorph has a higher AT content than the nuclear genome, and a comparison of transit peptides from the two genomes should prove interesting once more *Chlorarachnion* data are available.

If, as we contend, transit peptides can vary enormously in amino acid content, yet still perform the same function (sometimes even within a single organism such as *G. theta*), what then are the salient characteristics of plastid transit peptides? The only characteristic of transit peptides clearly demonstrated to be required for their functionality is an excess of basic over acidic residues (Bruce 2000; Foth et al. 2003). Accordingly we find that the ratio of basic to acidic amino acids in transit peptides is strikingly similar across all transit peptides examined, despite their vast differences in amino acid composition (circles in figure 1). Importantly, whereas the basic character of plant

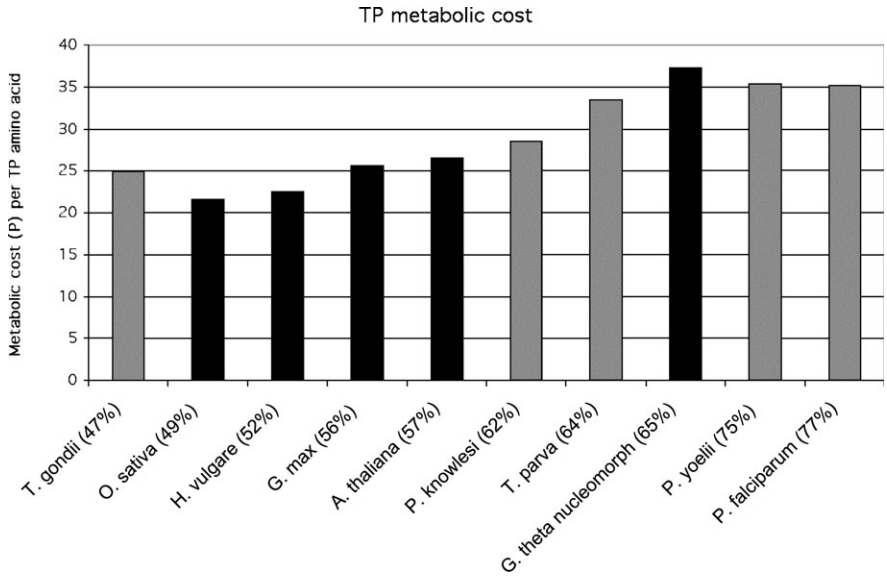


FIG. 3.—Average metabolic cost per transit peptide amino acid in parasitic and phototrophic organisms. The average cost for each amino acid in the transit peptide was estimated according to the metabolic costing (in high-energy P-bonds) if it were created from common precursor metabolites (given by Akashi and Gojobori [2002]). Chromalveolate transit peptides from parasitic species (gray) are longer than those from photosynthetic species (black), further inflating their total metabolic cost. Species are arranged in order of genomic AT content from left to right. The average AT content in coding DNA is shown in parentheses (in percent). Note that the trend to increased metabolic cost coincides with increased AT content.

chloroplast transit peptides is mostly from Arginine—an amino acid encoded by GC-rich codons—transit peptides from the more AT-rich genomes are highly enriched in the AT-favored basic amino acid lysine (fig. 2). These findings are consistent with the idea that the nucleotide bias effectively shapes the amino acid composition as long as transit peptide functionality is not compromised.

## Metabolic Cost and Nutrient Availability

Two other factors—metabolic cost and nutrient availability—have also been proposed to account for differential abundance of amino acids between organisms. Akashi and Gojobori (2002) showed that evolution of amino acid usage can be influenced by the varied metabolic cost of different amino acids. This, in turn, is dependent on the availability of energy, which is related to ecological and metabolic factors. The degree of energy limitation in plants remains controversial, but plants must synthesize all the amino acids they need. The malaria parasite *P. falciparum* on the other hand obtains its amino acids by scavenging from its host, so cost of amino acid anabolism should be much less of a factor in usage.

To investigate the role of amino acid metabolic cost, the average metabolic cost per amino acid was determined for transit peptides of several parasitic and photosynthetic organisms. The transit peptides of parasitic organisms do indeed contain more energetically costly amino acids (approximately 26 high-energy phosphate bonds per amino acid) than transit peptides of land plants (approximately 21 phosphate bonds [see figure 3]). For the parasitic chromalveolates, this could conceivably account for part of the relative trend away from some of the amino acids with lower metabolic costs. Although only a small number of chromalveolate transit peptide cleavage sites have been experimentally determined (Pancic and Strotmann 1993; Surolia and Surolia 2001; van Dooren et al. 2002), chromalveolate transit peptides also appear to be considerably longer than those of plants, further amplifying their total metabolic cost. In contrast, the transit peptides of *G. theta* resist the plant trend toward using metabolically inexpensive amino acids. Despite their phototrophic lifestyle, their transit peptides are more like those of the parasitic apicomplexans (longer and metabolically more expensive) than those of the phototrophic plants. This suggests that metabolic cost is not the primary force shaping amino acid usage in chromalveolate transit peptides. The trend towards metabolically costly amino acids is in any case difficult to extricate from an overlapping AT bias in many of the relevant codons. A complete budget of the metabolic cost of transit peptide synthesis would need to incorporate as yet unknown information, such as protein turnover dynamics and relative expression levels. The complexities of constructing such a budget and separating these factors from confounding influences, such as AT bias, makes it difficult to assess the impact of the metabolic cost of amino acid synthesis on transit peptide composition definitively.

Nutrient availability constitutes another significant pressure on the evolution of amino acid composition (Baudouin-Cornu et al. 2001). The parasitic lifestyle of apicomplexans presumably makes some nutrients more available than would be the case for either land plants or free-living algae. Such differences could potentially influence the abundance of amino acids with side chains that include either sulfur or nitrogen. Nitrogen limitation in plants could explain the use of the polar amino acids serine and threonine instead of nitrogen-containing asparagine as in *P. falciparum*. However sulfur-containing cysteine and methionine are similarly abundant in all the organisms studied, and no overall trend in relation to lifestyle is apparent in the abundance of nitrogen-containing side chains between the organisms studied. Compared with

plants, *Plasmodium* transit peptides are enriched in the nitrogen-containing asparagine and lysine but depleted for highly nitrogen-rich arginine. There are too many other complex factors to be able to distill out the relative importance of nutrient availability on transit peptide content between *Plasmodium* and plants. However, a more straightforward comparison between the transit peptides of the nitrogen-fixing dicot soybean *Glycine max* and the non–nitrogen fixing dicot *Arabidopsis thaliana* shows no nitrogen-skewed trend (fig. 2), discounting the likely significance of nutrient availability as a selective pressure on transit peptide amino acid content.

## Are Apicoplast Transit Peptides Phosphorylated?

When the content of individual amino acids in transit peptides from different sources was analyzed (see figure 2), the only significant trend that did not match the otherwise clear correlation between amino acid content and AT bias was the content of the hydroxylated amino acids serine and threonine (fig. 2). In the plant transit peptides, serine and threonine comprised approximately 25% of all amino acids and were considerably enriched also compared with the corresponding MPs (approximately 15% serine and threonine). In contrast, serine and threonine are massively depleted in the chromalveolate transit peptides compared with their plant counterparts (fig. 2). The average level of serine plus threonine in *A. thaliana* transit peptides was found to be significantly ($P \leq 0.05$) greater than that of the average of any chromalveolate transit peptide data set. Because serine and threonine are encoded by codons that are neutral with respect to AT bias, nucleotide bias does not account for this trend. So what could be driving this huge difference?

As discussed above, the lack of serine and threonine enrichment in the chromalveolate transit peptides is probably not caused by the metabolic cost of these amino acids. Because the serine and threonine side chains do not contain any potentially limiting elements such as nitrogen or sulfur, nutrient availability can also be discounted as a driving force for this trend. Thus, if serine and threonine depletion in chromalveolate transit peptides is not caused by ecological, metabolic, or nucleotide compositional factors, does this trend instead reflect a functional difference between plant and chromalveolate transit peptides? One of the proposed roles for serine and threonine in plant transit peptides may well offer the explanation. In plants, phosphorylation of serine and threonine in plastid transit peptides is thought to play an important part in recognition of transit peptides by 14-3-3-type chaperones (May and Soll 2000; Waegemann and Soll 1996). It has been suggested that phosphorylation may allow the trafficking system to differentiate between mitochondrial-targeting and chloroplast-targeting leaders (Waegemann and Soll 1996), thus, adding fidelity to a system that must discriminate between two otherwise similar classes of leader peptides. Even still, the system transports some proteins dually (accidentally or not) to both compartments (for examples see Chow et al. [1997] and Peeters and Small [2001]). Discrimination between mitochondrial and plastid transit peptides is unnecessary in organisms with secondary plastids (such as chromalveolate plastids) because of the two-step targeting process. Once the signal peptide inserts a plastid-bound protein into the endomembrane system, it cannot be misdirected to the mitochondria (van Dooren et al. 2001; Waller et al. 2000). Conversely, proteins genuinely destined for the mitochondria lack a signal peptide, precluding plastid targeting. This spatial separation of plastids and mitochondria in chromalveolates (and in all other known secondary endosymbionts) could have allowed the plastid transit peptides to dispense with transit peptide phosphorylation as a crucial feature that distinguishes between plastid-destined and mitochondrion-destined proteins.

Loss of plastid transit peptide phosphorylation in secondary plastids may also explain a hitherto puzzling experimental result. When the transit peptide of one apicoplast-targeted protein (*T. gondii* ribosomal protein S9) was attached to GFP without a signal peptide, it resulted in the reporter protein being targeted into the mitochondria courtesy of the apicoplast transit peptide (DeRocher et al. 2000). This suggests that apicomplexans have relaxed constraints for distinguishing between plastid and mitochondrial transit peptides, perhaps by abandoning phosphorylation of plastid transit peptides. To explore this possibility experimentally, Waller et al. (2000) deleted all serines and threonines from an apicoplast transit peptide and examined the effect on targeting. Normal apicoplast targeting of the reporter protein was observed, but it was noted that the transit peptide retained a tyrosine residue, which theoretically could have substituted for serine and threonine so that phosphorylation might still have occurred. Interestingly, all *P. falciparum* apicoplast transit peptides in our data set possess at least two hydroxylated residues, and although one transit peptide (ychB) lacks serine and threonine, it does contain two tyrosine residues. Further experimentation is required to define the role (if any) of transit peptide phosphorylation in apicomplexan plastids, and a mutated transit peptide that contains no hydroxylated amino acids at all is currently being prepared in our laboratory.

In the meantime, we note that the kinase responsible for chloroplast transit peptide phosphorylation in plants is located in the cytosol (Su et al. 2001; Waegemann and Soll 1996). If such a kinase existed for apicoplast transit peptides, it would presumably need to be located in the lumen of the endomembrane system, because apicoplast-targeted proteins are believed to be cotranslationally inserted into the endoplasmic reticulum (ER) lumen (van Dooren et al. 2002). Only one serine/threonine kinase possessing a signal peptide has been identified from the complete *P. falciparum* genome (Gardner et al. 2002), but this characterized kinase is a member of the Nek/NIMA family and has been shown to be involved in MAP signaling (Dorin et al. 2001) rather than transit peptide phosphorylation.

## Positional Sequence Information

To identify whether any positional information is present in apicoplast leaders, alignments were made from a combined data set of 118 chromalveolate bipartite leaders. Included in this data set were apicoplast-targeting leaders from each apicomplexan species, as well as the *G. theta*

nuclear (bipartite) sequences. The sequences were aligned around the predicted SignalP cleavage site, and a sequence logo generated from the 20 amino acids flanking the cleavage site (fig. 4). The sequence logo shows that the transit peptides from the organisms analyzed had a greatly elevated content of aromatic amino acids—particularly phenylalanine—at their first position (i.e., position +1 from the predicted SP cleavage site). This amino acid is potentially significant in targeting, as it becomes the N-terminus after signal peptidase removes the signal peptide within the lumen of the ER (van Dooren et al. 2002). The same trend was observed in a survey of secondary transit peptides from other chromalveolates (Deane et al. 2000). Therefore, the prevalence of phenylalanine in the first transit peptide position might be a functional feature of plastid trafficking via the endomembrane system along with the observed deficit of serine and threonine. Aromatic amino acids at or near the N-termini of peptides is important for binding to the ER-resident Hsp70 homolog BiP (Blond-Elguindi et al. 1993; Knarr et al. 1999) and binding of such chaperones likely plays an important role in apicoplast targeting (Foth et al. 2003; Yung, Unnasch, and Lang-Unnasch 2003).

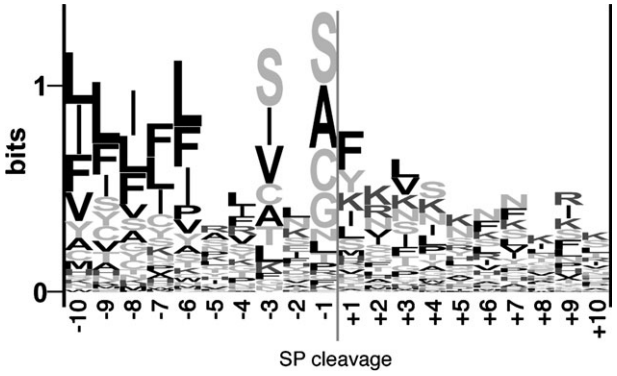


FIG. 4.—Positional sequence information in bipartite plastid-targeting leader sequences. This sequence logo represents an alignment of 118 plastid-targeting leaders from different chromalveolates (20 peptides each from *P. falciparum*, *P. yoelii*, *P. knowlesi*, *T. parva* and *T. annulata*, 14 peptides from *T. gondii*, and 4 peptides from *G. theta*). The sequences are aligned at the predicted signal peptidase (SP) cleavage site (http://www.cbs.dtu.dk/services/SignalP-2.0/). Sequence logos represent amino acids by their one-letter code. Although the height of each letter is proportional to the frequency of the respective amino acid at that position (called type 1 logo), the total height of each letter bar represents the Shannon information content (in bits) at that position, which indicates how biased the composition is at that position. The first five amino acids in the alignment (which correspond to the midregion of the signal peptide) are highly enriched in the hydrophobic amino acids isoleucine (I), leucine (L), valine (V), and phenylalanine (F). Spikes of information are also present at positions −1 and −3 relative to the SP cleavage site, consistent with known cleavage motifs (Nielsen et al. 1997). The first transit peptide position after the cleavage is enriched in the aromatic amino acids phenylalanine and tyrosine. No strong positional information is apparent for the remainder of the transit peptide.

## Transit Peptide Secondary Structure

Exactly what transit peptide feature the plastid import apparatus recognizes remains poorly understood, but some studies suggest that the unique lipid environment of the plant plastid membrane influences the conformation of the transit peptide (Pinnaduwage and Bruce 1996; Bruce 1998). Initial analyses of chloroplast transit peptides in aqueous environments found no pattern of structure, and it was proposed that transit peptides might be a perfect random coil (von Heijne and Nishikawa 1991). More recently, it has been realized that the outer membrane of plant chloroplasts, where the critical recognition event occurs, is enormously rich in galactolipids. Nuclear magnetic resonance (NMR) analyses of transit peptides in solvents that physiologically mimic galactolipids suggest that amphipathic α-helices may be

**Table 1**
**Average Amino Acid Content in Transit Peptides**

| | *T. gondii* (47%) | *O. sativa* (49%) | *G. theta* nuclear (50%) | *H. vulgare* (52%) | *G. max* (56%) | *A. thaliana* (57%) | *P. knowlesi* (62%) | *T. annulata* (64%) | *T. parva* (64%) | *G. theta* nucleomorph (65%) | *P. yoelii* (75%) | *P. falciparum* (77%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 6.43% | 17.89% | 22.50% | 21.25% | 10.79% | 6.12% | 4.46% | 1.00% | 1.25% | 1.41% | 0.98% | 1.89% |
| G | 7.50% | 5.87% | 2.50% | 5.00% | 3.07% | 4.12% | 6.43% | 3.00% | 2.75% | 1.25% | 3.29% | 2.61% |
| P | 8.57% | 12.89% | 13.75% | 9.00% | 8.95% | 7.83% | 3.93% | 4.25% | 4.25% | 1.41% | 2.93% | 2.67% |
| R | 11.43% | 6.07% | 10.00% | 6.50% | 3.51% | 5.81% | 8.39% | 6.25% | 5.00% | 2.66% | 3.78% | 4.72% |
| W | 1.79% | 0.77% | 0.00% | 0.13% | 0.35% | 0.47% | 1.43% | 0.25% | 0.50% | 0.47% | 0.37% | 0.50% |
| D | 2.14% | 1.54% | 0.00% | 0.50% | 1.58% | 2.47% | 3.04% | 2.25% | 3.25% | 2.66% | 2.80% | 1.78% |
| C | 3.93% | 1.74% | 0.00% | 1.13% | 1.23% | 1.85% | 2.50% | 1.50% | 1.75% | 3.59% | 1.22% | 2.11% |
| E | 1.43% | 2.01% | 0.00% | 0.88% | 1.32% | 3.11% | 3.21% | 2.50% | 1.25% | 0.78% | 3.41% | 2.44% |
| Q | 2.50% | 2.35% | 1.25% | 1.50% | 2.46% | 3.15% | 3.04% | 2.25% | 2.00% | 2.50% | 1.46% | 2.22% |
| H | 2.86% | 2.15% | 0.00% | 1.38% | 2.28% | 2.25% | 3.04% | 3.00% | 4.75% | 1.56% | 2.32% | 2.61% |
| S | 12.86% | 17.35% | 12.50% | 19.00% | 24.04% | 21.89% | 6.96% | 12.00% | 9.75% | 6.09% | 6.35% | 5.94% |
| T | 4.64% | 7.48% | 8.75% | 6.25% | 9.12% | 6.95% | 5.36% | 4.75% | 5.50% | 3.59% | 3.54% | 3.67% |
| V | 7.50% | 4.16% | 5.00% | 7.25% | 4.39% | 4.34% | 6.61% | 3.00% | 2.75% | 4.22% | 3.41% | 4.61% |
| L | 7.86% | 7.42% | 5.00% | 7.88% | 10.70% | 8.67% | 8.57% | 11.00% | 11.25% | 9.84% | 8.06% | 7.17% |
| M | 1.43% | 0.94% | 3.75% | 3.00% | 1.14% | 1.29% | 1.96% | 0.50% | 0.50% | 0.47% | 1.35% | 0.89% |
| N | 2.14% | 2.08% | 3.75% | 1.25% | 3.16% | 4.71% | 6.61% | 11.25% | 12.00% | 17.97% | 14.15% | 15.28% |
| K | 2.86% | 2.32% | 3.75% | 2.63% | 2.89% | 4.95% | 11.25% | 7.25% | 8.50% | 12.19% | 18.45% | 16.44% |
| F | 6.43% | 2.52% | 6.25% | 3.13% | 5.35% | 4.52% | 4.29% | 10.50% | 8.75% | 9.53% | 6.84% | 5.89% |
| Y | 2.14% | 0.37% | 0.00% | 0.13% | 0.61% | 1.16% | 3.93% | 5.50% | 5.75% | 5.31% | 5.24% | 6.44% |
| I | 3.57% | 2.08% | 1.25% | 2.25% | 3.07% | 4.36% | 5.00% | 8.00% | 8.50% | 12.50% | 10.03% | 10.11% |

NOTE.—The percentage amino acid composition of the initial 20 amino acids of each transit peptide was determined, then averaged for each species. Amino acids are ranked from top to bottom in increasing order of AT richness of corresponding codons. Source organisms are shown from left to right in order of exon AT content (shown in parentheses).

**Table 2**
**Average Amino Acid Content in Mature Peptides**

| | *T. gondii* (47%) | *O. sativa* (49%) | *G. theta* nuclear (50%) | *H. vulgare* (52%) | *G. max* (56%) | *A. thaliana* (57%) | *P. knowlesi* (62%) | *T. annulata* (64%) | *T. parva* (64%) | *G. theta* nucleomorph (65%) | *P. yoelii* (75%) | *P. falciparum* (77%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 8.25% | 9.73% | 13.75% | 11.63% | 7.11% | 6.65% | 5.18% | 3.75% | 2.50% | 3.44% | 4.15% | 3.89% |
| G | 8.27% | 7.55% | 7.50% | 8.13% | 5.53% | 6.33% | 4.82% | 4.75% | 3.25% | 4.84% | 3.41% | 3.17% |
| P | 4.66% | 5.87% | 3.75% | 5.63% | 8.51% | 4.48% | 4.46% | 2.50% | 4.00% | 2.03% | 3.17% | 2.33% |
| R | 12.63% | 6.07% | 2.50% | 3.63% | 7.28% | 5.75% | 3.57% | 3.75% | 3.77% | 3.28% | 4.15% | 2.67% |
| W | 0.36% | 1.61% | 0.00% | 1.63% | 0.44% | 1.30% | 0.71% | 0.50% | 0.50% | 0.63% | 0.73% | 0.67% |
| D | 3.97% | 5.03% | 3.75% | 4.50% | 3.68% | 5.72% | 3.75% | 4.50% | 3.77% | 4.06% | 4.02% | 5.56% |
| C | 0.36% | 1.98% | 3.75% | 1.50% | 1.93% | 1.66% | 3.57% | 1.25% | 0.75% | 1.09% | 1.83% | 1.00% |
| E | 7.86% | 6.98% | 1.25% | 6.50% | 6.05% | 6.75% | 6.07% | 6.75% | 4.52% | 4.84% | 6.10% | 6.39% |
| Q | 3.59% | 3.86% | 1.25% | 3.13% | 3.33% | 3.94% | 3.39% | 3.00% | 5.00% | 3.44% | 4.39% | 3.94% |
| H | 2.54% | 1.95% | 3.75% | 1.00% | 2.02% | 2.13% | 3.75% | 1.50% | 3.00% | 0.94% | 1.83% | 2.56% |
| S | 5.38% | 8.42% | 6.25% | 5.63% | 11.58% | 9.29% | 8.93% | 9.00% | 7.29% | 8.13% | 7.07% | 5.39% |
| T | 5.79% | 4.06% | 6.25% | 6.00% | 7.28% | 5.09% | 4.29% | 5.75% | 6.25% | 3.59% | 4.51% | 3.89% |
| V | 5.79% | 7.75% | 11.25% | 7.88% | 8.77% | 6.58% | 5.71% | 7.00% | 6.03% | 4.69% | 5.12% | 4.94% |
| L | 8.99% | 7.79% | 10.00% | 7.75% | 7.72% | 9.51% | 10.00% | 13.00% | 8.77% | 12.97% | 8.29% | 9.50% |
| M | 1.83% | 1.58% | 2.50% | 2.13% | 1.23% | 2.11% | 2.00% | 0.75% | 1.50% | 1.88% | 2.56% | 1.94% |
| N | 2.14% | 3.15% | 2.50% | 4.38% | 3.16% | 4.42% | 5.00% | 8.00% | 8.54% | 7.50% | 6.46% | 8.72% |
| K | 7.52% | 6.38% | 11.25% | 5.00% | 5.97% | 6.74% | 11.07% | 7.50% | 8.53% | 10.78% | 12.93% | 14.22% |
| F | 1.81% | 3.52% | 3.75% | 3.50% | 3.86% | 4.05% | 3.57% | 3.50% | 4.53% | 7.97% | 5.24% | 4.94% |
| Y | 2.88% | 2.65% | 0.00% | 3.50% | 1.58% | 2.39% | 4.64% | 4.75% | 5.02% | 2.34% | 5.12% | 5.11% |
| I | 5.41% | 4.06% | 5.00% | 7.00% | 2.98% | 5.03% | 5.71% | 8.50% | 12.53% | 11.56% | 8.90% | 9.17% |

NOTE.—The percentage amino acid composition of the final 20 amino acids of each protein was determined, then averaged for each species. Amino acids are ranked from top to bottom in increasing order of AT richness of corresponding codons. Source organisms are shown from left to right in order of exon AT content (shown in parentheses).

a common feature of these peptides, although only a handful of peptides have been analyzed thus far (Lancelin et al. 1994; Krimm et al. 1999; Wienk, Czisch, and Kruijff 1999; Wienk et al. 2000). Amphipathic helices consist of hydrophilic and polar residues on one face of the helix, and hydrophobic, nonpolar residues on the other face. In chloroplast transit peptides, the hydrophilic face is dominated by hydroxylated residues, which may interact via hydrogen bonds with galactolipids in the outer chloroplast envelope (reviewed in Bruce [2000]).

Apicoplast transit peptides from the *P. falciparum* data set were scanned for such features using a helical wheel projector. Although select sequences may be able to form amphipathic helices, the great majority of transit peptides surprisingly contained no such patterns. Even where potential amphipathic helices were observed, the amphipathicity was determined by basic, rather than hydroxylated residues—a pattern more reminiscent of mitochondrial presequences than chloroplast transit peptides (Bruce 2000). Again, this finding is consistent with a relaxation of the requirement for discrimination between chloroplast and mitochondrial transit peptides. Galactolipids similar to those found in plant chloroplast membranes have been identified in *P. falciparum* and *T. gondii* (Marechal et al. 2002), although their exact composition and localization remain to be established. If, as expected, these galactolipids reside in the two innermost membranes of the apicoplast (the equivalent of the two chloroplast membranes), they are unlikely to influence transit peptide conformation within the endomembrane system because of the so-called periplastid membrane, which separates the endomembrane space from the two inner plastid membranes (van Dooren et al. 2001). The composition of this membrane is not known and nor is there any information about the mechanism by which proteins are translocated across this additional membrane (van Dooren et al. 2001). The possible loss of this galactolipid interaction in transit peptide recognition might have had substantial impact on transit peptide secondary structure and may explain putative structural differences (such as the absence of amphipathic helices and the reduction in serine/threonine content) between apicoplast and plant transit peptides. NMR or circular dichroism analyses of apicoplast transit peptides in artificial micelles should prove very interesting in this context.

## Conclusions

Genome and expressed sequence tag (EST) sequencing projects of several chromalveolates have recently unveiled large numbers of proteins thought to be targeted to the secondary plastids of these organisms. All such putative plastid-targeted proteins identified to date possess a signal peptide that facilitates translocation into the endomembrane system, followed by a second, transit peptide–like segment responsible for plastid entry (van Dooren et al. 2001). At first glance, such chromalveolate transit peptides depart in many ways from long-familiar plant chloroplast transit peptides. However, the many differences regarding amino acid composition can be convincingly grouped into two categories. First, the more AT-rich genomes (especially of chromalveolate organisms) clearly favor those amino acids in transit peptides that are encoded by AT-rich codons and avoid those encoded by GC-rich codons. Consequently, certain amino acids appear to be replaced with biochemical equivalents encoded by more AT-rich codons: for example, basic lysine instead of arginine and aliphatic isoleucine instead of glycine or alanine. The huge scale of these changes suggests that the specific amino acids were not important in the first place, but rather their general biochemical features were.

The second change in chromalveolate transit peptides has been to dispose of a feature considered essential in plant

transit peptides: enrichment in the hydroxylated amino acids serine and threonine. The dispensability of this feature, and probably with it the propensity to be phosphorylated, supports the proposition that its function in plants—the avoidance of the mitochondrial targeting pathway—is redundant in chromalveolates, where the mitochondria and plastids are in topologically separate cell compartments. Ongoing improvements in transfection technology for various chromalveolates should make this an exciting question to address experimentally. Finally, our analyses confirm that plastid transit peptide function tolerates immense primary sequence variability. The surprisingly few global features that are shared between chromalveolate and plant transit peptides are, thus, likely to be those that have proved—over long evolutionary time scales—to be truly indispensable for transit peptide function. These are an excess of basic amino acids and a paucity of acidic amino acids.

## Acknowledgments

We thank Ian Woodrow and Ros Gleadow for helpful discussions. S.A.R. was supported by an MRS, B.J.F. was supported by an MIRS and MIFRS, GIM is an HHMI International Scholar and an ARC Professorial fellow. Support from the Australian Research Council and a Program Grant from the National Health and Medical Research Council are gratefully acknowledged. Sequencing of *Theileria parva* was funded by the International Livestock Research Institute and the Institute for Genomic Research. Preliminary *T. gondii* genomic sequence data were accessed via http://ToxoDB.org. We thank Ian Paulsen and David Roos for helpful discussions. Genomic data were provided by the Institute for Genomic Research (supported by the NIH grant #AI05093), and by the Sanger Center (Wellcome Trust).

## Literature Cited

Akashi, H. 2001. Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11**:660–666.

Akashi, H., T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA **99**:3695–3700.

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

Apt, K. E., N. Hoffman, A. R. Grossman. 1993. The γ subunit of R-phycoerythrin and its possible mode of transport into the plastid of red algae. J. Biol. Chem. **268**:16208–16215.

Apt, K. E., L. Zaslavkaia, J. C. Lippmeier, M. Lang, O. Kilian, R. Wetherbee, A. R. Grossman, P. G. Kroth. 2002. In vivo characterization of diatom multipartite plastid targeting signals. J. Cell. Sci. **115**:4061–4069.

The Arabidopsis Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Baudouin-Cornu, P., Y. Surdin-Kerjan, P. Marliere, D. Thomas. 2001. Molecular evolution of protein atomic composition. Science **293**:297–300.

Bhaya, D., A. Grossman. 1991. Targeting proteins to diatom plastids involves transport through an endoplasmic reticulum. Mol. Gen. Genet. **229**:400–404.

Blond-Elguindi, S., S. E. Cwirla, W. J. Dower, R. J. Lipshutz, S. R. Sprang, J. F. Sambrook, M. J. Gething. 1993. Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. Cell **75**:717–728.

Bruce, B. D. 1998. The role of lipids in plastid protein transport. Plant Mol. Biol. **38**:223–246.

———. 2000. Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol. **10**:440–447.

———. 2001. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. Biochim. Biophys. Acta **1541**:2–21.

Carels, N., G. Bernardi. 2000. Two classes of genes in plants. Genetics **154**:1819–1825.

Carlton, J. M., S. V. Angiuoli, B. B. Suh et al. (44 co-authors) 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii* yoelii. Nature **419**:512–519.

Cavalier-Smith, T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. **46**:347–366.

Cawley, S. E., A. I. Wirth, T. P. Speed. 2001. Phat—a gene finding program for *Plasmodium falciparum*. Mol. Biochem. Parasitol. **118**:167–174.

Chow, K. S., D. P. Singh, J. M. Roper, A. G. Smith. 1997. A single precursor protein for ferrochelatase-I from *Arabidopsis* is imported in vitro into both chloroplasts and mitochondria. J. Biol. Chem. **272**:27565–27571.

Deane, J. A., M. Fraunholz, V. Su, U. G. Maier, W. Martin, D. G. Durnford, G. I. McFadden. 2000. Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. Protist **151**:239–252.

DeRocher, A., C. B. Hagen, J. E. Froehlich, J. E. Feagin, M. Parsons. 2000. Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system. J. Cell Sci. **113**:3969–3977.

Dorin, D., K. Le Roch, P. Sallicandro, P. Alano, D. Parzy, P. Poullet, L. Meijer, C. Doerig. 2001. Pfnek-1, a NIMA-related kinase from the human malaria parasite *Plasmodium falciparum* Biochemical properties and possible involvement in MAPK regulation. Eur. J. Biochem. **268**:2600–2608.

Douglas, S., S. Zauner, M. Fraunholz, M. Beaton, S. Penny, L. T. Deng, X. Wu, M. Reith, T. Cavalier-Smith, U. G. Maier. 2001. The highly reduced genome of an enslaved algal nucleus. Nature **410**:1091–1096.

Douglas, S. E., C. A. Murphy, D. F. Spencer, M. W. Gray. 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. Nature **350**: 148–151.

Emanuelsson, O., H. Nielsen, S. Brunak, G. von Heijne. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. **300**:1005–1016.

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

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

Gardner, M. J., N. Hall, E. Fung et al. (45 co-authors). 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature **419**:498–511.

Gilson, P. R., G. I. McFadden. 2002. Jam packed genomes—a preliminary, comparative analysis of nucleomorphs. Genetica **115**:13–28.

Harb, O. S., B. Chatterjee, M. J. Fraunholz, M. J. Crawford, M. Nishi, D. S. Roos. 2004. Multiple functionally redundant signals mediate targeting to the apicoplast in the apicomplexan parasite *Toxoplasma gondii*. Eurkaryotic Cell **3**:663–674.

Harper, J. T., P. J. Keeling. 2003. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. Mol. Biol. Evol **20**:1730–1735.

Ishida, K., T. Cavalier-Smith, B. R. Green. 2000. Endomembrane structure and the chloroplast protein targeting pathway in *Heterosigma akashiwo* (Raphidophyceae, Chromista). J. Phycol. **36**:1135–1144.

Knarr, G., S. Modrow, A. Todd, M. J. Gething, J. Buchner. 1999. BiP-binding sequences in HIV gp160. Implications for the binding specificity of Bip. J. Biol. Chem. **274**:29850–29857.

Knight, R. D., S. J. Freeland, L. F. Landweber. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2: RESEARCH0010.

Krimm, I., P. Gans, J. F. Hernandez, G. J. Arlaud, J. M. Lancelin. 1999. A coil-helix instead of a helix-coil motif can be induced in a chloroplast transit peptide from *Chlamydomonas reinhardtii*. Eur. J. Biochem. **265**:171–180.

Lancelin, J. M., I. Bally, G. J. Arlaud, M. Blackledge, P. Gans, M. Stein, J. P. Jacquot. 1994. NMR structures of ferredoxin chloroplastic transit peptide from *Chlamydomonas reinhardtii* promoted by trifluoroethanol in aqueous solution. FEBS Lett. **343**:261–266.

Lang, M., K. E. Apt, P. G. Kroth. 1998. Protein transport into complex diatom plastids utilizes two different targeting signals. J. Biol. Chem. **273**:30973–30978.

Marechal, E., N. Azzouz, C. Santos de Macedo, M. A. Block, J. E. Feagin, R. T. Schwarz, J. Joyard. 2002. Synthesis of chloroplast galactolipids in apicomplexan parasites. Eukaryotic Cell **1**:653–656.

May, T., J. Soll. 2000. 14-3-3 proteins form a guidance complex with chloroplast precursor proteins in plants. Plant Cell **12**:53–64.

McFadden, G. I., P. R. Gilson, C. J. Hofmann, G. J. Adcock, U.-G. Maier. 1994. Evidence that an amoeba acquired a chloroplast by retaining part of an engulfed eukaryotic alga. Proc. Natl. Acad. Sci. USA **91**:3690–3694.

Moriyama, E. N., J. R. Powell. 1997. Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. **45**:514–523.

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

Musto, H., H. Rodriguez-Maseda, G. Bernardi. 1995. Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene **152**:127–132.

Musto, H., H. Romero, A. Zavala, K. Jabbari, G. Bernardi. 1999. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. J. Mol. Evol. **49**:27–35.

Nakai, K., M. Kaneshia. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics **14**:897–911.

Nielsen, H., J. Engelbrecht, S. Brunak, G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**:1–6.

Pancic, P. G., H. Strotmann. 1993. Structure of the nuclear encoded $\gamma$ subunit of CF0Fi of the diatom *Odontella sinensis* including its presequence. FEBS Lett. **320**:61–66.

Peeters, N., I. Small. 2001. Dual targeting to mitochondria and chloroplasts. Biochim. Biophys. Acta **1541**:54–63.

Perkins, S. L., J. J. Schall. 2002. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. J. Parasitol. **88**:972–978.

Pinnaduwage, P., B. D. Bruce. 1996. In vitro interaction between a chloroplast transit peptide and chloroplast outer envelope lipids is sequence-specific and lipid class- dependent. J. Biol. Chem. **271**:32907–32915.

Rathore, D., A. M. Wahl, M. Sullivan, T. F. McCutchan. 2001. A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of *Plasmodium* species. Mol. Biochem. Parasitol. **114**:89–94.

Saul, A., D. Battistutta. 1988. Codon usage in *Plasmodium falciparum*. Mol. Biochem. Parasitol. **27**:35–42.

Salzberg, S. L., M. Pertea, A. L. Delcher, M. J. Gardner, H. Tettelin. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics **59**:24–31.

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

Su, Q., K. Schmid, C. Schild, A. Boschetti. 2001. Effect of precursor protein phosphorylation on import into isolated chloroplasts from *Chlamydomonas*. FEBS Lett. **508**:165–169.

Surolia, N., A. Surolia. 2001. Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*. Nat. Med. **7**:167–173.

Theg, S. M., F. J. Geske. 1992. Biophysical characterization of a transit peptide directing chloroplast protein import. Biochemistry **31**:5053–5060.

van Dooren, G. G., S. D. Schwartzbach, T. Osafune, G. I. McFadden. 2001. Translocation of proteins across the multiple membranes of complex plastids. Biochim. Biophys. Acta **1541**:34–53.

van Dooren, G. G., V. Su, M. C. D'Ombrain, G. I. McFadden. 2002. Processing of an apicoplast leader sequence in *Plasmodium falciparum*, and the identification of a putative leader cleavage enzyme. J. Biol. Chem. **277**:23612–23619.

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

von Heijne, G., K. Nishikawa. 1991. Chloroplast transit peptides: the perfect random coil? FEBS Lett. **278**:1–3.

Waegemann, K., J. Soll. 1996. Phosphorylation of the transit sequence of chloroplast precursor proteins. J. Biol. Chem. **271**: 6545–6554.

Waller, R. F., M. B. Reed, A. F. Cowman, G. I. McFadden. 2000. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. EMBO J. **19**:1794–1802.

Wastl, J., U. G. Maier. 2000. Transport of proteins into cryptomonads complex plastids. J. Biol. Chem. **275**:23194–23198.

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

Wienk, H. L., R. W. Wechselberger, M. Czisch B. de Kruijff. 2000. Structure, dynamics, and insertion of a chloroplast targeting peptide in mixed micelles. Biochemistry **39**:8219–8227.

Wienk, H. L. J., M. Czisch B. de Kruijff 1999. The structural flexibility of the preferredoxin transit peptide. FEBS Lett **453**:318–326.

Yu, J., S. Hu, J. Wang et al. (86 co-authors). 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296**:79–92.

Yung, S., N. Lang-Unnasch. 1999. Targeting of a nuclear encoded protein to the apicoplast of *Toxoplasma gondii*. J. Eukaryot. Microbiol. **46**:79S–80S.

Yung, S. C., T. R. Unnasch, N. Lang-Unnasch. 2003. Cis and trans factors involved in apicoplast targeting in *Toxoplasma gondii*. J. Parasitol. **89**:767–776.

Martin Embley, Associate Editor

Accepted July 9, 2004