BMC Bioinformatics

BioMed Central

Research article

The *MB2* gene family of *Plasmodium* species has a unique combination of S1 and GTP-binding domains

Lisa C Romero[1], Thanh V Nguyen[2], Benoit Deville[3], Oluwasanmi Ogunjumo[1] and Anthony A James*[1,4]

Address: [1]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA 92697-3900, USA, [2]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA, [3]Biomérieux, 69280 Marcy L'Etoile, FRANCE and [4]Department of Microbiology and Molecular Genetics, University of California, Irvine, CA 92697 USA

Email: Lisa C Romero - lopie@uci.edu; Thanh V Nguyen - thanhnguyen@burnham.org; Benoit Deville - benoit.deville@eu.biomerieux.com; Oluwasanmi Ogunjumo - oogunjum@uci.edu; Anthony A James* - aajames@uci.edu

* Corresponding author

Published: 28 June 2004

BMC Bioinformatics 2004, **5**:83 doi:10.1186/1471-2105-5-83

Received: 27 January 2004
Accepted: 28 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/83

Abstract

Background: Identification and characterization of novel *Plasmodium* gene families is necessary for developing new anti-malarial therapeutics. The products of the *Plasmodium falciparum* gene, *MB2*, were shown previously to have a stage-specific pattern of subcellular localization and proteolytic processing.

Results: Genes homologous to *MB2* were identified in five additional parasite species, *P. knowlesi*, *P. gallinaceum*, *P. berghei*, *P. yoelii*, and *P. chabaudi*. Sequence comparisons among the *MB2* gene products reveal amino acid conservation of structural features, including putative S1 and GTP-binding domains, and putative signal peptides and nuclear localization signals.

Conclusions: The combination of domains is unique to this gene family and indicates that *MB2* genes comprise a novel family and therefore may be a good target for drug development.

Background

Malaria causes 300–500 million clinical cases and more than one million deaths world-wide each year [1]. Efforts to reduce disease that rely on chemotherapeutics and insecticides are undermined by an increase in drug and insecticide resistance. Development of new control mechanisms is facilitated by an understanding of basic *Plasmodium* biology, and the identification of unique and vulnerable properties that can be exploited by therapeutics, vaccines or other control strategies. However, in spite of more than a century of malaria research, much is still unknown about the biology of *Plasmodium* parasites due in large part to the extraordinary complexity of their life cycles. These parasites are able to infect vertebrate and invertebrate hosts, survive in intracellular and extracellular environments, invade multiple types of cells, and evade the immune responses of both hosts. At each stage of the life cycle, new sets or combinations of proteins are expressed, therefore the biology of the parasite is continually changing. The uniqueness and complexity of this parasite is underscored by the revelation that approximately two-thirds of the *P. falciparum* open reading frames (ORFs) predicted from the recently completed genome sequence do not match proteins in existing databases [2]. This finding presents a unique opportunity to identify and characterize novel *Plasmodium* proteins that may be exploited for therapeutic benefits.

The *P. falciparum* gene, *MB2*, is a single copy gene on chromosome five, and is expressed as a single mRNA species in blood-stage parasites [3]. The conceptual translation product is 1610 amino acids (aa) in length and has a predicted mass of 187 kDa. The gene product was divided into three domains based on primary sequence properties: an amino-terminal basic domain (B), a central acidic domain (A), and a carboxyl-terminal domain (G) with high sequence similarity to GTP-binding domains of prokaryotic IF-2. The *MB2* gene product is expressed in sporozoites, asexual blood stages (ABS) and gametocytes, however, it displays a pattern of stage-dependent cellular localization and apparent proteolytic processing throughout the parasite life cycle.

Genes homologous to the *P. falciparum MB2 (PfMB2)* have been identified in five *Plasmodium* species. Alignment of the conceptual translation products of all *MB2* genes reveals a distinctive primary amino acid structure characterizing this gene family. Four conserved sequences of amino acids conceivably important for function define this structure. Two of the four sequences share homology to known functional domains: 1) an S1 domain found in RNA-binding proteins is located in the amino-terminal B domain; and 2) the previously-mentioned putative GTP-binding domain located in the carboxyl-terminus. Two additional sequences in the A domain have no known similarities or functional correlations with any known proteins in existing databases. In addition to these conserved sequences, three types of predicted signaling motifs are present in all *MB2* gene products identified. These include signal peptides, nuclear localization sequences (NLSs) and putative cell-surface retention signals (CRSs).

Results

Distribution of MB2 gene in apicomplexa

Primary characterizations of *MB2* genes were carried out by heterologous screening of libraries and database searches of Plasmodium species and other apicomplexa (*Babesia bovis*, *Eimeria tenella*, *Theileria annulata*, *T. parva*, *Cryptosporidium parvum* and *Toxoplasma gondii*). Corresponding genes with open reading frames (ORFs) were found only in the malaria parasites.

Determination of ORFs in the MB2 gene family

The *PfMB2* B domain (aa 1 to 489) was used to search the Sanger Institute Pathogen Sequencing Unit *P. knowlesi* database. A >20 kilobase (kb) genomic contig, c000500602, was identified (Table 1). A 4128 bp ORF (nt 12317 to 16447) within this contig was designated *PkMB2* based on similarity to *PfMB2* and the other *MB2* homologous genes. The predicted translation product is 1376 aa in length.

Seven contiguous cDNA and genomic DNA fragments identified by library screening and gene amplification techniques were assembled to construct the *PgMB2* contig. (Table 1). A genomic clone spans 4011 bases of the 4638 bp ORF and ~450 bp of the 3'-end were determined by amplification from genomic DNA, therefore, only ~200 bases were from cDNA sequence alone and not confirmed from genomic DNA. The contig incorporates more than 5600 bp and includes 5'- and 3'-end non-coding sequences. The complete *PgMB2* ORF encodes a predicted translation product 1546 aa in length.

Table 1: Properties of *Plasmodium MB2* gene homologues.

Gene Name[a]	Species	Vertebrate Host	Gene size[b] (bp)	Translation product (aa)
PfMB2	*P. falciparum*	Human	4830	1610
PkMB2	*P. knowlesi*	Primate	4128	1376
PgMB2	*P. gallinaceum*	Avian	4638	1546
PbMB2	*P. berghei*	Rodent	3972	1324
PyMB2	*P. yoelii*	Rodent	4140	1380
PcMB2	*P. chabaudi*	Rodent	3921	1307

a. Abbreviations: Pf, *Plasmodium falciparum*; Pk, *P. knowlesi*; Pg, *Plasmodium gallinaceum*; Py, *P. yoelii*; Pb, *P. berghei*; Pc, *P. chabaudi*. b. Gene sizes are based on reconstructions of the open reading frames from the following sources: *PfMB2*, Genbank accession nos. **AF378132**, **AF378133**, **AF378134**, **AF378135**, **AF378136**, **AF378137**, **AF378138**; *PkMB2*, Sanger center contig c000500602; *PgMB2*, Genbank accession nos. **AY328135**, **AY328136**, **AY328137**, **AY328138**, **AY328139**, **AY328140**, **AY328141**; *PbMB2*, Genbank accession nos. **AY330346**, **AZ527175**, and Sanger Center contigs berg-524g06.q1c, berg-172a06.q1c, berg-5f11.p1c, berg-524g01.q1c, berg-514c07.q1c; *PyMB2*, Genbank accession nos. **AY330344**, **AY330345**, **AABL01000943**, and Sanger Center contig ChrPyI_c25655; *PcMB2* accession nos. **AY330347**, **AY330348**, and Sanger Center contigs Pch0968c12.p1c, Pch0781b11.q1c, Pch0290h07.q1c, Pch0915e09.q1c, Pch0264d01.q1c, Pch222b04.p1c, Pch0781b11.p1c, Pch0264d01.p1c, Pch0290h07.p1c, Pch0915e09.p1c, Pch222b04.q1c, Pch0865d07.q1c, Pch0092g12.p1c, Pch1109c10.p1c, Pch1109c10.q1c, Pch0092g12.q1c, Pch257f05.p1c, Pch0272f01.p1c, Pch1029f05.p1c, Pch0282c03.p1c, Pch322a07.p1c, Pch335b11.p1c, and Pch317g10.p1c.

Table 2: Primers used for amplification of *MB2* gene sequences.

Species[a]	Primer	Primer sequence[b]
PgMB2	PgMB2-1	5'-GGNCAYATHAAYCAYGGNAARACN-3'
	PgMB2-2	5'-RAANGCYTCRTGNCCNGGNGTRTC-3'
	PgMB2-3	5'-CGTCCTTATTTGATTACATATG-3'
	PgMB2-4	5'-CAATAAAAGTAAAAGTGTATTCATC-3'
	PgMB2-5	5'-CTCCAGCTTTGGGTAACGAATTC-3'
	PgMB2-6	5'-GCAGTAAGCTTTAATGCGATTATAATTGG-3'
	PgMB2-7	5'-GTAAGCTTTAATGCGATTATAATTGG-3'
	PgMB2-8	5'-TACATAAATATACTATGAAAAGTC-3'
PyMB2	PyMB2-1	5'-GCTGCTCCTTCATTAACAGAACATAC-3'
	PyMB2-2	5'-GCATTACCTTCTTGATTAATACTACACG-3'
	PyMB2-3	5'-GGGAACAAAAAAATAAAGAAAAGTGTGTC-3'
	PyMB2-4	5'-GCTATTCTGTATTATATTTCTATTGG-3'
PbMB2	PbMB2-1	5'-GAATGGTAGTGTAGGTTATTTGCATAG-3'
	PbMB2-2	5'-GTTCAGAAATTTCTTTTGCAATTGTTTC-3'
PcMB2	PcMB2-1	5'-CAGCATAGCTCAGTATTGCCAGCC-3'
	PcMB2-2	5'-GTTCAGAAATTTCTTTTGCAATTGTTTC-3'
	PcMB2-3	5'-GTCGGGGTTGTATACACATCGTC-3'
	PcMB2-4	5'-GCCCACATATTTTGATTTCCGTC-3'

a. This column indicates the species from which the primers were derived. Abbreviations as in Table 1. b. R = A+G, Y = C+T, N = A+C+G+T, H = A+T+C

Five genomic shotgun sequences were identified by BLAST search [4] of the Sanger Institute Pathogen Sequencing Unit *P. berghei* (ANKA strain) database using sequences derived from conserved regions of the B and G domains of *MB2* ORFs (Table 1). In addition, a genomic contig, accession no. **AZ2527175**, was identified from the PlasmoDB database. These six contigs represent the first 1350 bp and final 987 bp of the *PbMB2* ORF. Primers derived from these regions (PbMB2-1 and PbMB2-2, Table 2) were used to amplify the internal missing sequence from *P. berghei* genomic DNA. The complete ORF of *PbMB2* consists of 3972 nucleotides and the predicted translation product spans 1324 aa. A region (nt 33 to 49) in the 5'-end of this gene contains 15 (thymine) residues interrupted by a guanine (nt 36) and a cytosine (nt 41) and results in a predicted translation initiation codon 60 nts upstream of the predicted translation initiation codons of *PyMB2* and *PcMB2*.

A 10 kb genomic contig, MALPY00946, identified by Stefan Kappe (Seattle Biomedical Research Institute) from a BLAST search of the PlasmoDB database, contains sequences from the *P. yoelii MB2* gene (Table 1). 3352 bp of this contig (accession no. **AABL01000943**) represent a partial ORF with similarity to *PfMB2* extending to the 3'-end of the gene. The complete 5'-end sequence was obtained with gene amplification techniques using a cDNA library and genomic DNA as templates. The complete *PyMB2* ORF consists of 4140 nucleotides and encodes a predicted protein 1380 aa in length. There is a region in the 5'-end (nt -26 to -12, where +1 is the first nucleotide in the translation initiation codon) that contains a long stretch of thymine (T) bases (15 or 16 Ts). Removal of one thymine (15 Ts) results in a truncation of the 5'-end. Therefore, to confirm the number of bases in this poly-T tract (15 Ts), an independent gene amplification reaction was performed with *Pfx* polymerase and the resulting product was sequenced.

A total of 22 genomic shotgun sequences identified from BLAST searches of the *P. chabaudi* database were used to construct three non-overlapping contigs representing regions of *PcMB2* (Table 1). The sequences of the missing regions were determined by gene amplification from a *P. chabaudi* cDNA library using primers (PcMB2-1, PcMB2-2, PcMB2-3 and PcMB2-4, Table 2) derived from the three contigs. There is a poly-T tract in the 5'-end of the gene (nt -26 to -12). The numbers of thymine bases (14 Ts interrupted by one cytosine (nt-22)) were re-confirmed by independent gene amplification with *Pfx* polymerase and sequencing. The complete *PcMB2* ORF consists of 3921 nucleotides, which are translated into 1307 aa.

Conserved regions

Alignment of the amino acid sequences of the various *MB2* gene products reveals four conserved domains (Figure 1). While the ordering of the domains is conserved, the overall amino acid identity calculated by the Clustal method among the predicted products is low, only 38–44%, except among rodent parasites, which are much higher (72–79%) (Table 3). This higher level of identity is consistent with phylogenetic analyses inferred from previous *Plasmodium* sequence comparisons [5,6].

Table 3: Amino acid identities of *MB2* gene products from six *Plasmodium* species.[a,b]

	PkMB2	*PgMB2*	*PbMB2*	*PyMB2*	*PcMB2*
PfMB2	39.4	41.4	43.1	43.3	42.2
PkMB2		38.3	38.9	38.4	37.5
PgMB2			43.8	43.9	44.0
PbMB2				76.0	79.4
PyMB2					71.7

a. Identities are expressed as percentages b. Abbreviations are the same as in Table 1

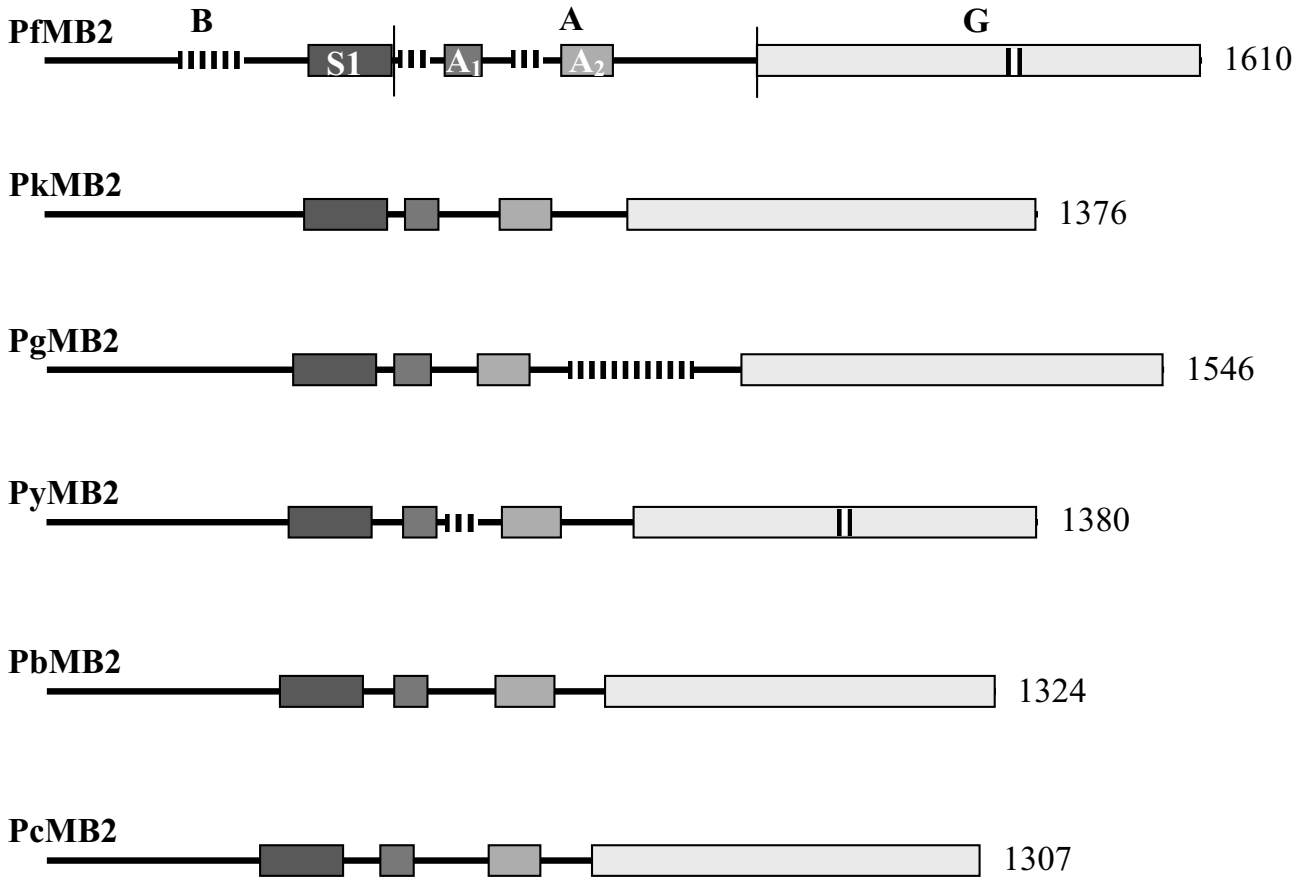

Figure 1
Schematic diagram of the predicted translation products of six *MB2* genes. The letters B, A and G label the respective domains of PfMB2. The boundaries of these domains are delimited by thin vertical lines. Shaded bars represent conserved regions, three of which, S1, A_1, and A_2 are labeled in PfMB2. Vertical lines interrupting the horizontal lines or contained within the boxes indicate regions containing repeated amino acid sequences. Numbers on the right indicate the number of amino acids in each product. Abbreviations: Pf, *Plasmodium falciparum*; Pk, *P. knowlesi*; Pg, *P. gallinaceum*; Py, *P. yoelii*; Pb, *P. berghei* and Pc, *P. chabaudi*.

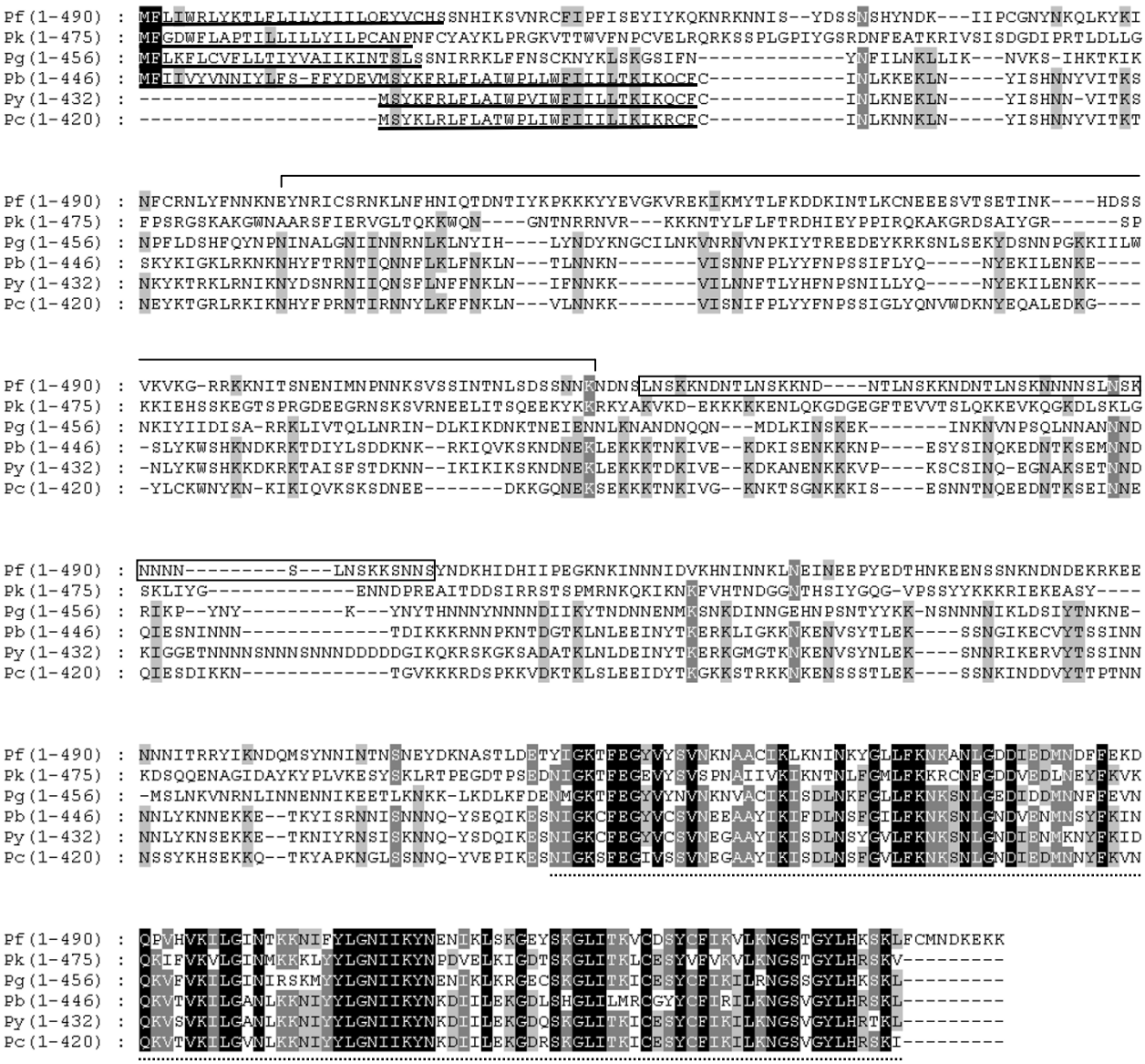

Figure 2
Alignment of the predicted amino acids in the PfMB2 B domain with the amino-terminal domains of MB2 homologue translation products. Amino acids highlighted in black are conserved in all genes, those highlighted in gray are conserved in the majority of genes (at least four). A solid line underlines predicted signal peptides. A solid bracket above the Pf sequence delimits the antigenic region of PfMB2 [3]. A dotted line underlines the predicted S1 domain. A repeat region of PfMB2 is boxed. Alignment generated using the Clustal method (DNAStar) and manual alignment. Numbers in parenthesis indicate represented amino acids. Abbreviations are the same as in Figure 1.

Of all the conserved domains, an amino terminal region of approximately 120 residues has the highest level of amino acid similarity (an average of 68% identity, Figure 2). The PROSITE (Swiss Institute of Bioinformatics, [7]) and SMART [8,9] computer programs identified amino acid similarities of this region in each of the gene products with S1 domains. A search of the GenBank and EMBL databases using the BLAST program [4] and this conserved

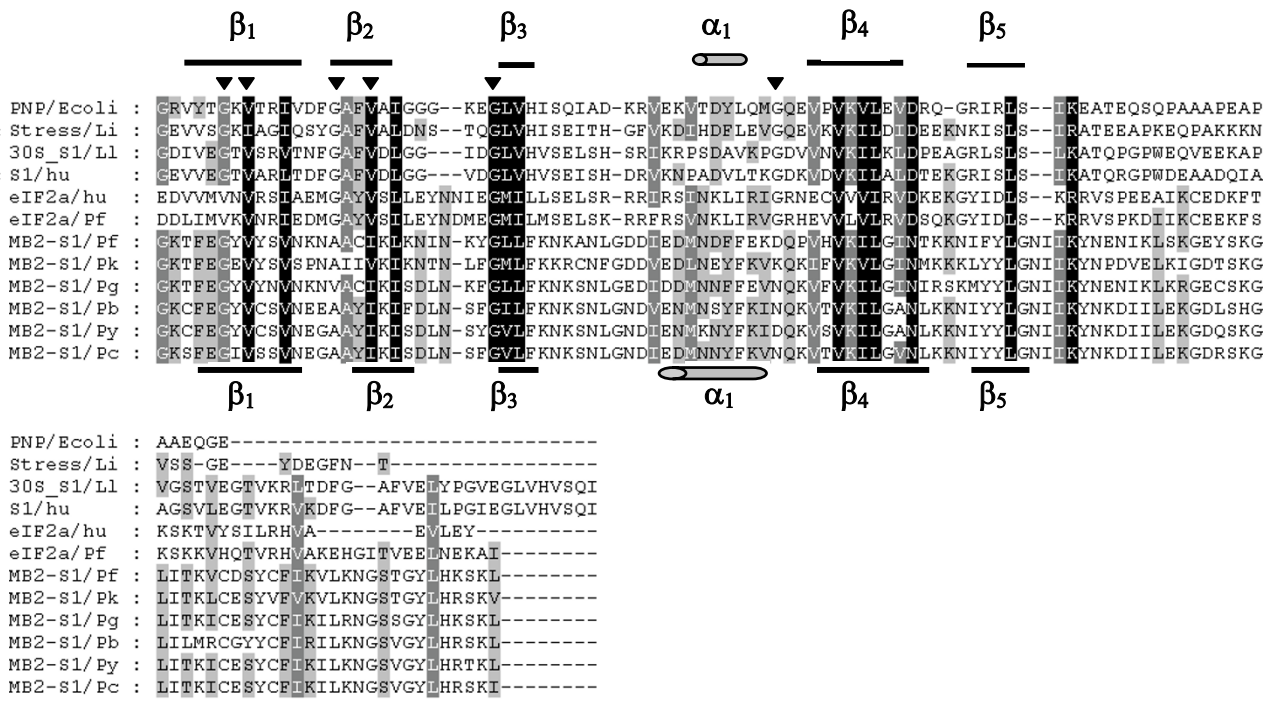

Figure 3

Alignment of MB2 S1 domains with representative S1 domain sequences. Representative sequences containing an S1 domain are PNP/Ecoli: polyribonucleotide phosphorylase (PNPase), *Escherichia coli* [**P05055**/622–711], Stress/Li: protein similar to *B. subtilis* general stress protein 13, *Listeria innocua* [**NP_471798**/7–103], 30S_S1/Ll: 30S ribosomal protein S1, *Lactococcus lactis lactis* [**NP_266994**/195–312], S1/hu: ribosomal S1 protein, human [**AAA97575**/154–223], eIF2a/hu: eukaryotic initiation factor 2 alpha subunit, human [**P05198**/17–121], eIF2a/Pf: Putative eukaryotic initiation factor 2 alpha, *Plasmodium falciparum* [PF07_0017/20–138], identified from PlasmoDB database. MB2 S1 domains are included from six species: Pf = *P. falciparum*, Pk = *P. knowlesi*, Pg = *P. gallinaceum*, Pb = *P. berghei*, Py = *P. yoelii*, Pc = *P. chabaudi*. The secondary structure determined for PNPase [10] is represented above the alignment with a α helix indicated by a cylinder and β sheets indicated by thick lines. The structure predicted for PfMB2 is represented below the alignment. Black triangles indicate highly conserved residues identified by Bycroft *et al*., [10]. Asterisks indicate sequences identified by BLAST search with MB2 sequences. Alignment generated using the Clustal method (DNAStar). Similar amino acid residues are shaded.

region as a query also revealed similarity to S1 domains from several proteins including a human ribosomal S1 protein, other ribosomal S1 protein homologues, and prokaryotic stress proteins containing S1 domains.

Alignment of the MB2 S1 domain with other known S1 domains reveals amino acid sequence conservation at several of the highly conserved residues (four of six) reported for known S1 domains [10] (Figure 3). Furthermore, secondary structure predictions of the putative MB2 S1 domain performed by PSIpred [11] and PROF predictions [12] predicted a βββαββ structure, similar to that solved for the *E. coli* polynucleotide phosphorylase [10].

Two of the four conserved domains were identified in regions corresponding to the central A domain of *PfMB2* and these were designated A_1 and A_2 (Figure 4). These domains have an average of 61% and 49% identity with *PfMB2*, respectively. Neither region generated significant similarity results when GenBank, EMBL, and PlasmoDB databases were searched.

The G domain, previously identified as having similarity to the GTP-binding domain of prokaryotic IF-2 [3], was conserved highly among the homologous gene products (an average of 65% identity with the sequences encoded by *PfMB2*). Dever *et al*. [13] identified three motifs, GXXXXGK, DXXG, and NKXD, each separated by 40–80 aa, that are conserved among different classes of GTP-binding proteins. All three motifs are present in the *MB2* gene products (Figure 5). The first and second motifs are involved in interactions with the phosphate portion of the

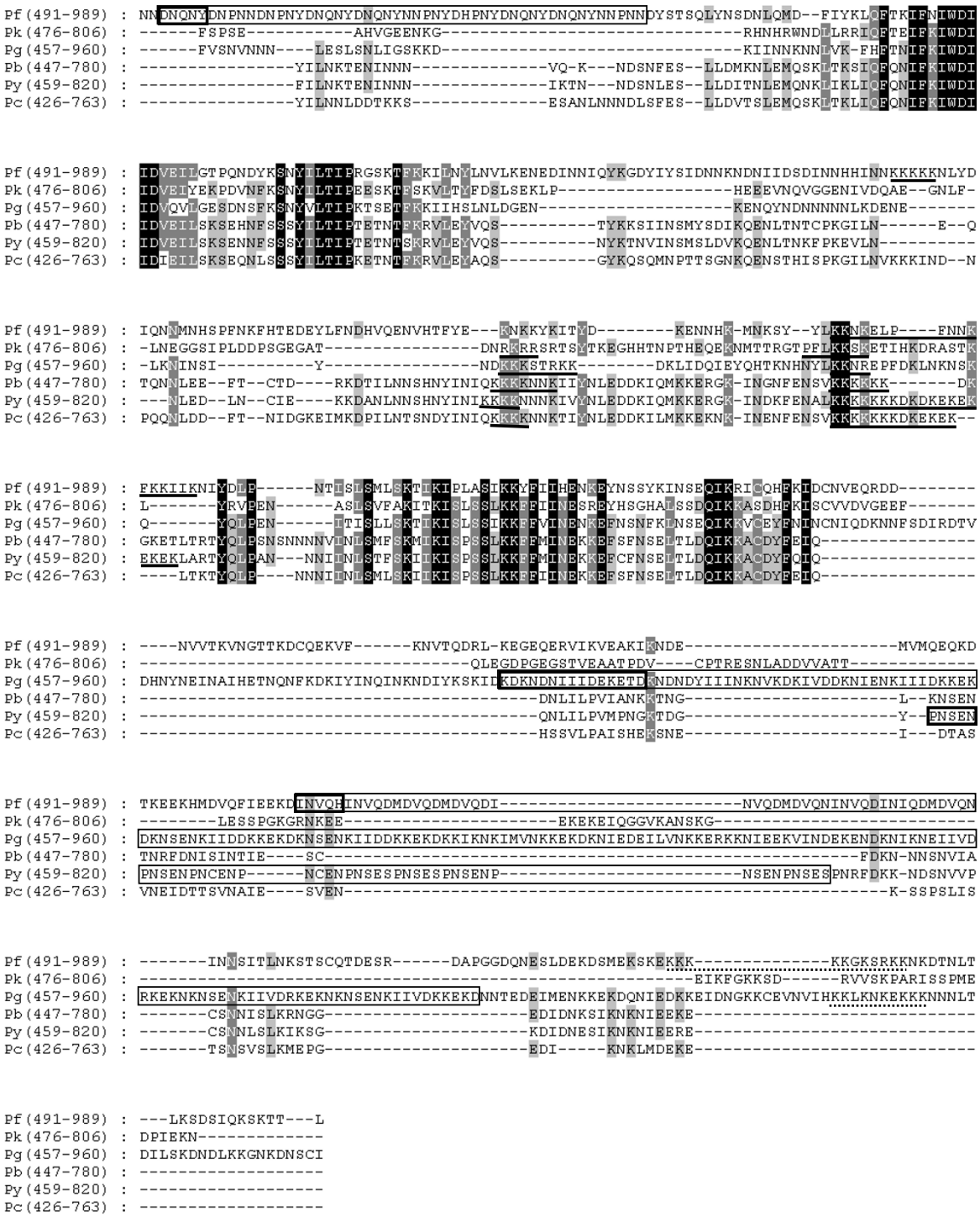

Figure 4
Alignment of the PfMB2 A domain with the central domains of homologous *MB2* translation products. Repeated amino acid sequences are boxed with a single repeat unit outlined with a dark box. Putative nuclear localization sequences are underlined with a solid line. Putative cell-surface retention sequences are underlined with a broken line. Alignment generated using the Clustal method (DNAStar) and manual alignment. Numbers in parenthesis indicate represented amino acids. Abbreviations are the same as in Figure 2.

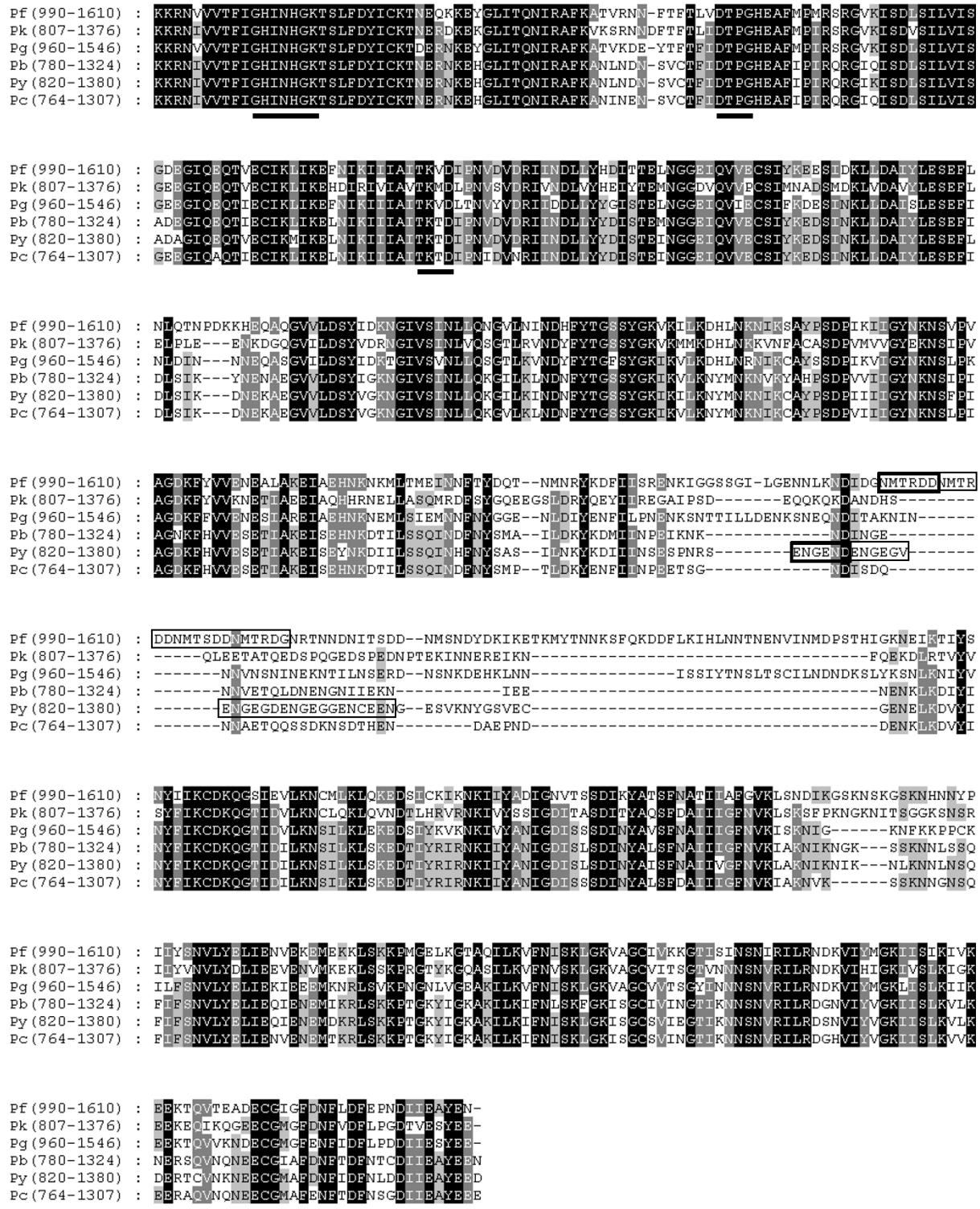

Figure 5
Alignment of carboxyl-terminal G domains of homologous MB2 translation products. The three G domain consensus sequences are underlined. Repeated motifs are boxed with a single motif outlined by a dark box. Generated by Clustal method (DNAStar). Numbers in parenthesis indicate represented amino acids. Abbreviations are the same as in Figure 2.

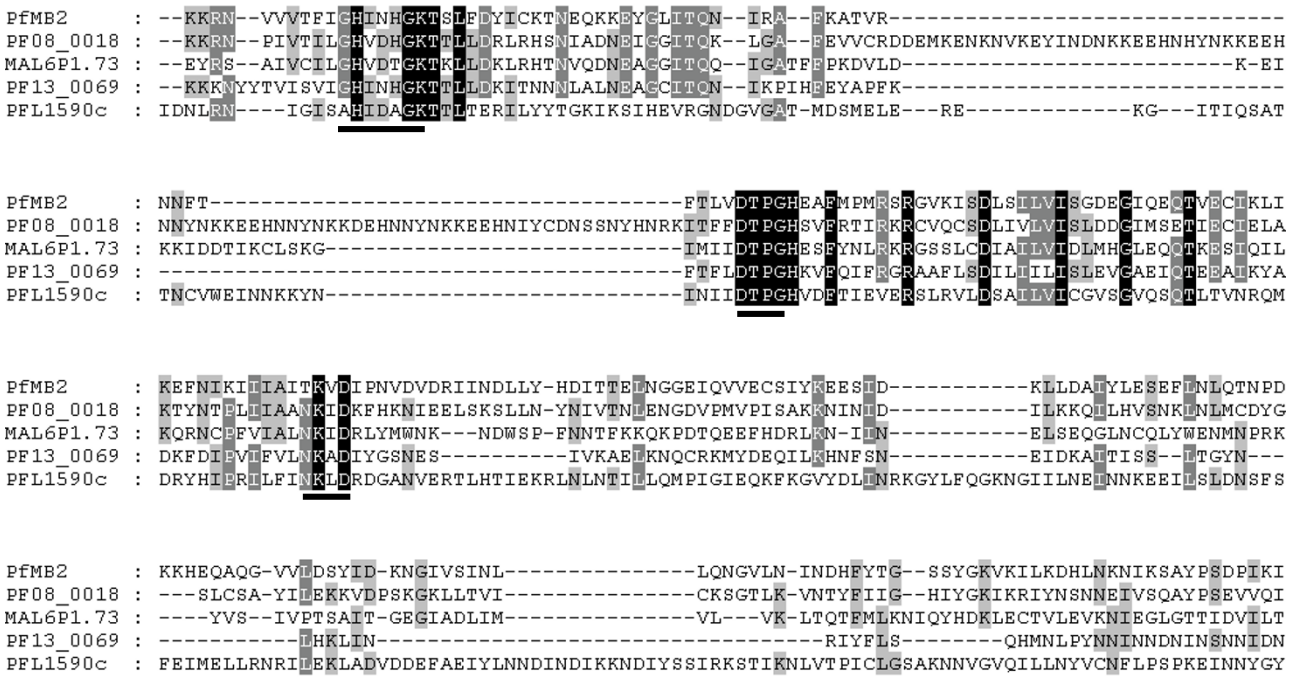

Figure 6
Alignment of *Plasmodium falciparum* G domains. Sequences were obtained from the PlasmoDB database. Amino acids included in alignment are PfMB2 [989–1232] and four annotated *P. falciparum* contigs from the PlasmoDB database: PFO8_0018 [650–959], translation initiation factor-like protein; MAL61.73 [390–638], putative translation initiation factor, IF-2; PF13_0069 [207–403], putative translation initiation factor, IF-2; PFL1590c [41–340], putative elongation factor G. The three G domain consensus motifs are underlined. Generated by Clustal method (DNAStar) and manual alignment. Abbreviations are the same as in Figure 2.

GTP molecule [13]. The sequences in MB2 proteins, GHINHGK and DTPG, respectively, adhere strictly to these motifs and are conserved absolutely among *MB2* gene products. The last motif is involved in nucleotide specificity and the *MB2* gene products vary from the consensus sequence. In all *MB2* gene products, the consensus asparagine (N) is replaced by a threonine (T) (TKXD; with the X being valine (V), methionine (M) or T [rodent genes]). Replacement of the polar residue, asparagine (N), with a different polar residue, tyrosine (Y), or a basic residue, lysine (K), abolished the GTP-binding activity of ras21 mutant protein [14]. It is unknown what effect, if any, the threonine replacement observed in *MB2* gene products has on nucleotide binding or specificity. Other putative *P. falciparum* G domains have the conserved NKXD motif (Figure 6) suggesting that this divergence from the third motif is not a feature of *Plasmodium* GTP-binding proteins in general, rather that it may be a definitive characteristic of *MB2* gene products.

The conservation of the G domain among the *MB2* gene family and the similarities to elongation factors could indicate that these parasite products are functional analogues of the translation factors. However, *P. falciparum* has a number of genes whose products are characterized as elongation factors and none of these have S1 domains (Figure 6).

Predicted signal sequences

Signal peptide predictions by SignalP [15], PSORT and PSORT II [16] computer programs for all five homologous gene products produced similar, conflicting results as reported for *PfMB2* [3]. SignalP predicted signal peptides and cleavage sites (Figure 2) while PSORT and PSORT II gave low scores, predicting no cleavage sites for all *MB2* gene products, except PyMB2 (PSORT II). Putative transmembrane domains were identified by the SMART computer program [8,9] in only two MB2 translation products, PgMB2 and PbMB2, and placed at aa 5–24 in PgMB2 and aa 20–42 in PbMB2, overlapping the

predicted signal peptides. However, using the TMpred program [17], transmembrane domains were identified in all six *MB2* gene products in regions overlapping the signal peptides. As was proposed for PfMB2 [3], it is possible that the *MB2* gene products are attached to the surface of sporozoites by an uncleavable signal anchor. The PyMB2 and PcMB2 signal sequences are truncated due to a frame shift caused by a deletion of a thymine residue in a poly-(T) tract. Without this deletion the amino-terminal sequences are highly conserved among the rodent parasite signal sequences.

PSORT II and PROSITE programs predicted multiple nuclear localization signals (NLSs) in B and A domains for all proposed translation products. The occurrence of multiple NLSs in a protein is not unusual, and it is thought that weak NLSs may be additive or synergistic [18]. All ORFs encode predicted NLSs in the same two regions of their central domains (Figure 4).

Cell surface retention sequences (CRS) are polybasic domains implicated in membrane localization [19,20]. They appear to be homologous to NLSs and may function as such when appropriate. There is evidence that adjacent sequences may be important in targeting a protein to the plasma membrane or nucleus such that the context of polybasic motifs may determine their function. All translation products have potential CRS similar to that described for PfMB2 [3].

Examination of the amino acid composition of the homologous gene products reveals that *PkMB2* and *PgMB2* encode regions with a predicted negative charge preceding the G domain, but these regions are shorter (322 and 254 aa, respectively) than the 496 aa A domain encoded by *PfMB2*. The rodent malaria *MB2* gene products each have much shorter acidic domains of approximately 80–130 aa immediately preceding the G domain. Therefore, the N-terminal basic and central acidic domains assigned to PfMB2, based on amino acid composition, are not conserved in structure or size, but an acidic domain immediately preceding the G domain is present in all *MB2* gene products. This domain, however, varies widely in size and composition among *MB2* gene products.

Repeat sequences in MB2 gene products

Imperfect, tandemly-repeated sequences are encoded in the non-conserved regions of three *MB2* genes: *PfMB2*, *PgMB2* and *PyMB2*. In addition to the repeated sequence described by [3], an additional repeat region was identified in the G domain of *PfMB2*, from aa 1315 to 1338 (Figure 5). This six-residue sequence, N, M, T, R, D, D, is repeated four times. A highly-charged, imperfect repeat is encoded in the central domain of *PgMB2* (Figure 4). This 15-residue repeat with consensus K, N, I/S, E, N/D, K, I, I, V/I, D, K, K, E, K, D, is repeated eleven times with a five-residue non-repeat occurring between the second and third motifs. *PyMB2* encodes a five-residue motif, P, N, S/C, E, N/S, that is tandemly repeated nine times in the central domain (Figure 4) and a six-residue motif, E, N, G, E, N/G, D, repeated four times in the G domain (Figure 5).

Discussion

Alignment of *MB2* conceptual translation products from the six *Plasmodium* species reveals a conservation of overall structure that in the absence of other similar genes allows us to conclude that these are orthologous members of a gene family. Of four conserved regions, two have similarity to known protein domains, specifically the putative S1 domain and GTP-binding domain (G domain). The S1 domain was identified first in the *E. coli* ribosomal S1 protein [10]. This protein consists of six copies of an ~70 aa motif, and has an essential role in facilitating translation initiation by interacting with the ribosome and mRNA. S1 domains were subsequently found in a variety of proteins including eukaryotic initiation of translation factor 2 alpha, eIF2α, bacterial and chloroplast translation initiation factor-1, IF-1, a bacterial exonuclease polynucleotide phosphorylase, PNPase, and RNA-helicases [10]. Common themes among these proteins are associations with RNA and ribosomes, and an involvement in initiation of translation and mRNA turnover. S1 domains also have been found in DNA-binding proteins including a human S1-like protein that demonstrates preferential DNA-binding *in vitro* [21].

Structural analysis of a representative S1 domain from *E. coli* PNPase revealed a five-stranded antiparallel β-barrel [10]. This structure belongs to an ancient superfamily of RNA-binding proteins called OB fold proteins. Members of this DNA and RNA-binding protein family feature cold-shock domains (CSD) found in bacterial cold-shock proteins, eukaryotic Y-box proteins, and S1 domains. From cellular localization studies of *PfMB2* products, we know that the putative S1 domain of MB2 is localized in the nucleus in asexual blood stages of this parasite [3]. If this domain is a functional nucleic acid-binding domain, a role for MB2 in the nucleus could include binding DNA to regulate gene expression in a stage-specific way, perhaps as a transcription factor. Two lines of evidence indicate that gene regulatory components of *Plasmodium*, including promoters and transcription factors, are unique to these organisms. 1) Plasmodium promoters do not function in transfection experiments in mammalian cells [22]; and 2) SV40 and other viral promoters that are active in eukaryotes fail to drive gene expression in *P. falciparum* [22,23]. Furthermore, with the exception of a TATA-box binding protein, no transcription factors have yet been cloned [22,23]. Therefore, if MB2 is a transcription factor,

the unique structure of this gene is not surprising. Alternatively, if MB2 binds mRNA, it could be involved in the post-transcriptional regulation of pre-mRNA in the nucleus of ABSs.

The MB2 G domains are most similar to the family of GTP-binding domains involved in protein synthesis, specifically those of prokaryotic IF-2 and elongation factors. Initiation of translation factors, IF-2, are essential for initiation of protein synthesis. They recruit the initiator tRNA and conduct it as Met-tRNA$_f$·eIF2·GTP ternary complex. Eukaryotic IF-2 is composed of three polypeptide chains complexed throughout the initiation cycle (reviewed in [24]). Elongation factors containing this domain promote GTP-dependent binding of aminoacyl tRNA to the A site of ribosomes and catalyze the translocation of synthesized protein from A to P sites.

The S1 and G domains identified in predicted *MB2* gene products are commonly found in proteins involved in the protein synthesis pathway, and this could be an indication that this is the pathway in which MB2 is involved. However, word search of the Pubmed and ProDom databases for "S1 and GTP-binding" yielded no proteins that contain both of these domains. Furthermore, as we have shown, the *P. falciparum* genome contains other genes that are better fits in terms of amino acid conservation for the roles in protein synthesis. It is interesting that both of these domains are found in two separate proteins subunits of the initiation factor, eIF2 [4,24].

With the exception of *MB2* gene products from rodent species, the intervening sequences of amino acids between the conserved domains are not similar. These regions may be divergent due to a lack of selective pressure acting on them to maintain similarity. This suggests these domains are not important for function, but rather may have a structural role that is not strictly sequence-dependent. Alternatively, this divergence may reflect the potential dissimilarity of ligands in the respective hosts [25]. However, the putative antigenic regions of the *P. falciparum* MB2 B region [3] are not conserved in the homologous genes at the primary sequence level. Also, there are no repeat domains in the B regions of the homologous genes.

The non-conserved regions of *P. falciparum*, *P. gallinaceum* and *P. yoelii* contain repeated amino acid motifs. Many malarial antigens, including CS, TRAP, and merozoite surface antigens (MSA), contain short, tandemly repeated motifs [26]. Much of the antibody response to malaria parasites is directed towards these repeat regions. Repeat regions in surface proteins of several different parasites, which elicit ineffective T-cell-independent antibody responses, have been hypothesized to divert immune-responses away from more critical epitopes [26,27]. The purpose of the repeat regions in *MB2* gene products is unknown, but their presence support the hypothesis that homologous proteins are on the surface of some stage of the parasite and exposed to the immune system.

Conclusions

We have identified six orthologous *MB2* genes and comparisons of the conceptual translation products of these genes have highlighted domains that may be important for function, specifically two domains found in proteins involved in protein synthesis. Furthermore, analysis of predicted protein characteristics suggests that localization signals are conserved. Therefore, the structural characteristics that define the *MB2* gene family consist of the capacity to encode four conserved domains, including a S1 and putative GTP-binding domain, a signal sequence and one or more NLSs.

In this post-genomics era, comparative genomics between different species of malaria parasites is now a possibility. Consequent to this will be the rapid discovery of homologues of less well-conserved genes. Identification of homologous genes enables researchers to identify functional domains that may not be obvious in the analysis of single genes. As 60% of predicted *P. falciparum* genes identified in the genome sequencing project do not match existing database sequences [2], the importance of utilizing this method of comparative genomics to derive functional clues about these novel genes is evident. This large number of novel genes presents a unique opportunity to identify and characterize proteins that can be targeted by anti-malarial therapeutic agents, vaccines and other methods of controls [28,29].

Methods

Identification of MB2 genes

Computer-assisted alignment of the amino acid sequence in the predicted *PfMB2* G-domain with other IF-2 G domains revealed three conserved sequence motifs found in known GTP-binding proteins [13]. Two of these motifs, GHINHGK and DTPGHEAF, were used to design degenerate oligonucleotide primers, PgMB2-1 and PgMB2-2 (Table 2), that were used to amplify a product from a *P. gallinaceum* genomic DNA (strain 8A, gift of Ken Vernick, University of Minnesota). The resulting 165 base-pair (bp) gene fragment was used to amplify an 118 bp homologous probe using the specific primers, PgMB2-3 and PgMB2-4 (Table 2), to screen a number of *P. gallinaceum* libraries. A near full-length *PgMB2* gene sequence was compiled using information derived from seven clones isolated from three libraries: a genomic library (gift of David Kaslow, Vical, San Diego), a salivary gland sporozoite cDNA library and an oocyst sporozoite cDNA library. The latter two libraries were made from parasites isolated from infected mosquitoes (D. Fidock, B.

Beerntsen, and A.A. James, unpublished). Missing 3'-end sequence was obtained by inverse gene amplification techniques using *P. gallinaceum* genomic DNA (gift of Joseph Vinetz, University of California, San Diego) digested with the *TaqI* restriction enzyme and the gene-specific primers, PgMB2-5 and PgMB2-6 (Table 2) derived from the known sequence. The sequence of this amplification product was confirmed by independent gene amplification with a high-fidelity polymerase, *Pfx* (Invitrogen) and the gene specific primers, PgMB2-7 and PgMB2-8 (Table 2).

A partial ORF with homology to *P. falciparum MB2* was identified from the *P. yoelii* genome sequence project (Table 1). The complete sequence of the ORF was determined by amplifying a product from a *P. yoelii* blood stage cDNA library (strain 17X, ATCC: MR4) using a primer, PyMB2-1 (Table 2), derived from the original genomic contig (Table 1), and an anchored vector primer, T7. Inverse gene amplification was performed using *P. yoelii* genomic DNA (gift of Andy Waters, Leiden University Medical Center), digested with *TaqI* and amplified using the specific primers, PyMB2-2 and PyMB2-3 (Table 2), or digested with *Mbo I* and amplified using specific primers, PyMB2-3 and PyMB2-4 (Table 2). All primers were derived from the known sequence.

Full-length or partial *MB2* ORFs from *P. knowlesi*, *P. berghei* and *P. chabaudi* were identified in the PlasmoDB http://www.plasmodb.org or Sanger Institute Pathogen Sequencing Unit http://www.sanger.ac.uk databases by searching with conserved regions of *MB2* identified by amino acid alignments of *PfMB2*, *PgMB2*, and *PyMB2* (Table 1). Missing nucleotide sequences for *PbMB2* were amplified from *P. berghei* genomic DNA (gift of Andy Waters, Leiden University Medical Center) using primers, PbMB2-1 and PbMB2-2 (Table 2), derived from the retrieved sequences. The missing segments of *PcMB2* were isolated by gene amplification from a *P. chabaudi* cDNA library (ATCC: MR4) using the specific primers, PcMB2-1 and PcMB2-2, or PcMB2-3 and PcMB2-4 (Table 2), derived from the three separate contigs (Table 1).

When the high fidelity DNA polymerase, *Pfx*, was used for gene amplification, two or three clones from one product were sequenced. When *Taq* polymerase was used, three products amplified independently were sequenced. Amplification conditions for reactions using Taq polymerase were 94° for 5 min, followed by 30–35 cycles of 94° for 30–60s, 50–60° for 30–60s, and 72° for 1–2 min. Amplifications for reactions using *Pfx* polymerase were 94° for 5 min followed by 30 cycles of 94° for 15s, 45–61°C for 30s, 60–72°C for 60s.

Protein alignments

Predicted translation products were aligned by the Clustal method using the program Megalign (DNASTAR). Refinements were made by eye within the Megalign program. The alignments were exported as MSF files and figures were produced using the GENEDOC program.

List of abbreviations

aa, amino acid(s); ABS, asexual blood stages; bp, base-pair; CSD, cold shock domains; CRS, cell surface retention signal; NLS, nuclear localization signal; nt, nucleotides, OB, oligosaccharide /oligonucleotide binding domains; ORF, open reading frame

Authors' contributions

LCR carried out the isolation and primary characterization of DNA fragments and conducted the sequence analyses to define all of the genes. TVN completed the analysis of the *P. falciparum* gene and provided reagents. BD isolated the initial *P. gallinaceum* clones. OO assisted in characterizing the *P. yoelii* ORF. AAJ conceived of the study and participated in its design, analysis and coordination.

Acknowledgements

The authors thank Lynn Olson for help in typing the manuscript, and the Burroughs-Wellcome fund for support. AAJ was a Burroughs-Wellcome scholar in Molecular Parasitology and a recipient of the New Initiatives in Malaria Research Award. Support of the National Institutes of Health (AI23697) is gratefully acknowledged.

References

1. **Malaria at a glance; http://mosquito.who.int/cmc_upload/0/000/014/813/Malaria_at_a_glance1.htm.** World Bank Report; 2001.
2. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419:**498-511.
3. Nguyen TV, Fujioka H, Kang AS, Rogers WO, Fidock DA, James AA: **Stage-dependent localization of a novel gene product of the malaria parasite, Plasmodium falciparum.** *J Biol Chem* 2001, **276:**26724-26731.
4. Kedzierski L, Escalante AA, Isea R, Black CG, Bernwell JW, Copper RL: **Phylogenetic analysis of the genus Plasmodium based on the gene encoding adenylosuccinate lyase.** *Infection, Genetics and Evolution* 2002, **1:**297-301.
5. Perkins SL, Schall JJ: **A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences.** *J Parasitol* 2002, **88:**972-978.
6. Appel RD, Bairoch A, Hochstrasser DF: **A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server.** *Trends Biochem Sci* 1994, **19:**258-260.
7. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
8. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains.** *Proc Natl Acad Sci U S A* 1998, **95:**5857-5864.

9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
10. Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG: **The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold.** *Cell* 1997, **88:**235-242.
11. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16:**404-405.
12. Ouali M, King RD: **Cascaded multiple classifiers for secondary structure prediction.** *Protein Sci* 2000, **9:**1162-1176.
13. Dever TE, Glynias MJ, Merrick WC: **GTP-binding domain: three consensus sequence elements with distinct spacing.** *Proc Natl Acad Sci U S A* 1987, **84:**1814-1818.
14. Clanton DJ, Hattori S, Shih TY: **Mutations of the ras gene product p21 that abolish guanine nucleotide binding.** *Proc Natl Acad Sci U S A* 1986, **83:**5076-5080.
15. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6.
16. Nakai K, Kanehisa M: **A knowledge base for predicting protein localization sites in eukaryotic cells.** *Genomics* 1992, **14:**897-911.
17. Hofmann K, Stoffel W: **TMbase - A database of membrane spanning proteins segments.** *Biol Chem Hoppe-Seyler* 1993, **374:**166.
18. Garcia-Bustos J, Heitman J, Hall MN: **Nuclear protein localization.** *Biochim Biophys Acta* 1991, **1071:**83-101.
19. Hancock JF, Paterson H, Marshall CJ: **A polybasic domain or palmitoylation is required in addition to the CAAX motif to localize p21ras to the plasma membrane.** *Cell* 1990, **63:**133-139.
20. Boensch C, Kuo MD, Connolly DT, Huang SS, Huang JS: **Identification, purification, and characterization of cell-surface retention sequence-binding proteins from human SK-Hep cells and bovine liver plasma membranes.** *J Biol Chem* 1995, **270:**1807-1816.
21. Eklund EA, Lee SW, Skalnik DG: **Cloning of a cDNA encoding a human DNA-binding protein similar to ribosomal protein S1.** *Gene* 1995, **155:**231-235.
22. Dechering KJ, Kaan AM, Mbacham W, Wirth DF, Eling W, Konings RN, Stunnenberg HG: **Isolation and functional characterization of two distinct sexual-stage-specific promoters of the human malaria parasite Plasmodium falciparum.** *Mol Cell Biol* 1999, **19:**967-978.
23. Horrocks P, Dechering K, Lanzer M: **Control of gene expression in Plasmodium falciparum.** *Mol Biochem Parasitol* 1998, **95:**171-181.
24. Pain VM: **Initiation of protein synthesis in eukaryotic cells.** *Eur J Biochem* 1996, **236:**747-771.
25. Templeton TJ, Kaslow DC: **Cloning and cross-species comparison of the thrombospondin-related anonymous protein (TRAP) gene from Plasmodium knowlesi, Plasmodium vivax and Plasmodium gallinaceum.** *Mol Biochem Parasitol* 1997, **84:**13-24.
26. Verra F, Hughes AL: **Biased amino acid composition in repeat regions of Plasmodium antigens.** *Mol Biol Evol* 1999, **16:**627-633.
27. Wrightsman RA, Dawson BD, Fouts DL, Manning JE: **Identification of immunodominant epitopes in Trypanosoma cruzi trypomastigote surface antigen-1 protein that mask protective epitopes.** *J Immunol* 1994, **153:**3148-3154.
28. Foth BJ, *et al.*: **Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*.** *Science* 2003, **299:**705-708.
29. Lell B, *et al.*: **Fosmidomycin, a novel hemotherapeutic agent for malaria.** *Antimicrob. Agents Chemotherapy* 2003, **47:**735-738.