Protein Engineering vol.12 no.12 pp.1113–1120, 1999

PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome

Chrislaine Withers-Martinez[1,2], Elisabeth P.Carpenter[2], Fiona Hackett[1], Barry Ely[3], Mohammed Sajid[1], Muni Grainger[1] and Michael J. Blackman[1,4]

[1]Division of Parasitology, [2]Division of Protein Structure and [3]Division of Virology, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK

[4]To whom correspondence should be addressed.
E-mail: mblackm@nimr.mrc.ac.uk

The A+T-rich genome of the human malaria parasite *Plasmodium falciparum* encodes genes of biological importance that cannot be expressed efficiently in heterologous eukaryotic systems, owing to an extremely biased codon usage and the presence of numerous cryptic polyadenylation sites. In this work we have optimized an assembly polymerase chain reaction (PCR) method for the fast and extremely accurate synthesis of a 2.1 kb *Plasmodium falciparum* gene (*pfsub-1*) encoding a subtilisin-like protease. A total of 104 oligonucleotides, designed with the aid of dedicated computer software, were assembled in a single-step PCR. The assembly was then further amplified by PCR to produce a synthetic gene which has been cloned and successfully expressed in both *Pichia pastoris* and recombinant baculovirus-infected High Five^{TM} cells. We believe this strategy to be of special interest as it is simple, accessible and has no limitation with respect to the size of the gene to be synthesized. Used as a systematic approach for the malarial genome or any other A + T-rich organism, the method allows the rapid synthesis of a nucleotide sequence optimized for expression in the system of choice and production of sufficiently large amounts of biological material for complete molecular and structural characterization.

Keywords: gene synthesis/*Pichia pastoris*/*Plasmodium falciparum*/protein expression/subtilisin-like protease

Introduction

The *Plasmodium falciparum* genome is probably the most A+T-rich (~82% overall) of any organism known and this has posed a number of considerable technical hurdles to the current genome sequencing effort (Triglia and Kemp, 1991; de Bruin *et al.*, 1992; Gardner *et al.*, 1998). The high A+T content of *P. falciparum* coding sequences [~76% for nuclear genes and even higher for the malarial plastid genome (Wilson *et al.*, 1996)] also presents significant difficulties in the analysis of protein structure and function, since A+T-rich DNA can be extremely difficult to express in commonly used prokaryotic and eukaryotic expression systems. In both *Escherichia coli* and yeast, the level of expression of a gene has been shown to correlate with the use of certain preferred codons; 'incorrect' codon usage results in low-level expression (Grantham *et al.*, 1980; Bennetzen and Hall, 1982; Makoff *et al.*, 1989; Romanos *et al.*, 1991; Hernan *et al.*, 1992; Martin *et al.*, 1995; Sreekrishna *et al.*, 1997; Hale and Thompson, 1998; Schmidt-Dannert *et al.*, 1998). Perhaps more critically, in *Saccharomyces cerevisiae* or the methylotrophic yeast *Pichia pastoris*, certain A+T-rich stretches of sequence can act as polyadenylation or transcription termination signals, resulting in low-level or truncated mRNA (Romanos *et al.*, 1991; Sreekrishna *et al.*, 1997). In some cases this problem can be resolved by identification and appropriate mutagenesis of the offending sequence domains, but in more extreme cases this is not possible. For example, attempts to express fragment C of the *Clostridium tetani* tetanus toxin in *E.coli* or yeast (Makoff *et al.*, 1989; Romanos *et al.*, 1991) was initially thwarted owing to the A+T-rich DNA of this bacterium. Efficient expression of the protein was only accomplished by complete gene synthesis to obtain coding sequences with an appropriate codon usage and increased G+C content.

Novel strategies are urgently needed to combat malaria. Over two-thirds of the world's population reside in malaria endemic regions and there are thought to be up to 500 million clinical cases of the disease annually. Serious complications following infection with *P. falciparum* are frequent and falciparum malaria is estimated to cause between 1.5 and 2.7 million deaths per year (Anon., 1994). There is no widely available vaccine against malaria and drug-resistant *P. falciparum* is a widespread and growing problem. Clinical malaria is caused by growth of the parasite in circulating red blood cells. The invasive merozoite stage of the parasite enters red blood cells, replicates within the cell, then is released to invade new red cells and repeat the cycle. Red cell invasion is known to require the activity of parasite serine proteases (Blackman *et al.*, 1993; McKerrow *et al.*, 1993). Work in this laboratory has recently identified a *P. falciparum* gene (*pfsub-1*) encoding a member of the subtilisin-like serine protease superfamily (Blackman *et al.*, 1998). The primary gene product is processed in two consecutive steps during transport through the parasite secretory system and the putative mature protease is concentrated in secretory organelles within merozoites, indicating that it may play a role in red cell invasion (Blackman *et al.*, 1998). Inhibition of the enzyme may therefore block invasion. We have been interested in achieving high-level heterologous expression of catalytically active PfSUB-1 for structural and enzymological studies. Attempts to express the entire *pfsub-1* gene, or domains of it, in *E.coli* resulted in extremely low levels of expression of insoluble protein (M.Sajid and M.J.Blackman, unpublished data); expression of the malarial gene in *P. pastoris* or baculovirus was equally unsuccessful, probably owing to the unfavourable codon bias of the 72% A+T-rich *pfsub-1* coding sequence. To solve this problem, it was decided to synthesize the *pfsub-1* gene, adapting the codon usage for optimum expression in *P. pastoris*. Recent advances in gene synthesis technology have led to an increased number of available gene synthesis methods (Grantham *et al.*, 1980; Prapunwattana *et al.*, 1996; Mehta *et al.*, 1997; Au *et al.*, 1998), the most attractive ones relying on the use of polymerase chain reaction (PCR) (Prodromou and Pearl, 1992; Graham *et al.*, 1993; Stemmer *et al.*, 1995; Casimiro *et al.*, 1997; Brocca *et al.*, 1998). Here we have adapted and optimized in terms of error rate the assembly PCR method of Stemmer *et al.* (1995). In this procedure, a DNA polymerase is used in a primary PCR (called the assembly process) to build increasingly long DNA fragments from a pool of overlapping oligonucleotides (oligos). A second PCR is then performed to amplify specifically the previously assembled synthetic product. Any point mutations in the final amplified product are corrected by subsequent subcloning steps. We present new PCR parameter settings which, combined with exclusive use of the proofreading *Pfu* DNA polymerase, have allowed us to synthesize rapidly the 2.1 kb *pfsub-1* gene in the absence of a functional screen and with unprecedented accuracy. The final synthetic *pfsub-1* gene has been successfully expressed not only in *P. pastoris*, achieving levels of recombinant protein of 0.2–0.5 g/l, but also in recombinant baculovirus-infected insect cells. We invite researchers working on organisms with A+T-rich genomes to consider gene synthesis as a feasible, systematic approach to high-level protein expression for protein engineering and structure–function studies.

Materials and methods

Materials

The oligos, one 53-mer and 103 40-mers, were synthesized on a 40 nmol scale with no extra purification and dissolved in water to final concentration of 25 μM each (Oswell DNA Service, UK). All restriction enzymes were obtained from Boerhinger Mannheim or from Gibco/BRL Life Technologies. Cloned *Pyrococcus furiosus* (*Pfu*) DNA polymerase was purchased from Stratagene (La Jolla, CA, USA). The *pMos*Blue blunt-ended cloning kit was supplied by Amersham Life Science and T4 DNA ligase by New England Biolabs, UK. DH5α^{TM} *E.coli* competent cells were obtained from Gibco/BRL Life Technologies. Plasmid DNA purification was performed with MiniSnap (Invitrogen, San Diego, CA, USA). Automated DNA sequencing was done on a Perkin-Elmer ABI prism 377 DNA sequencer, using dye terminator cycle sequencing. Sequencing analysis was performed using the AutoAssembler (Factura) package. Nucleotide sequence comparisons were performed using Lasergene DNAStar software. The *P. pastoris* expression kit including the pPIC9K vector and media were obtained from Invitrogen. The pVL1393 transfer vector, AcMNPV linear baculovirus DNA and Sf9 and High Five^{TM} insect cells were supplied by Invitrogen and the insect cell media by Gibco/BRL Life Technologies and Expression Systems, LLC (Woodland, CA, USA).

Gene design

The sequence of the synthetic *pfsub-1* gene was designed according to *P. pastoris* codon usage (Bennetzen and Hall, 1982; Sreekrishna *et al.*, 1993) with the aid of the CODOP program (available from EPC, l-carpen@nimr.mrc.ac.uk). CODOP is a Unix perl script which provides a number of molecular biology functions, including codon optimization with host organism preference as proposed by Hale and Thompson (1998). CODOP reads a codon usage table, assessing the frequency of each codon per 1000 codons (F). It then calculates the codon preference N:

$$N = F_{\mathrm{aa1}}n/(F_{\mathrm{aa1}} + F_{\mathrm{aa2}} + F_{\mathrm{aa3}} \ldots + F_{\mathrm{aa}n})$$

where n is the number of synonymous codons and F_{aa1} to $F_{\mathrm{aa}n}$ are the proportions per 1000 codons of each synonymous codon. All codons with N below a user-defined cut-off, in this case 0.6, are rejected. The gene to be analysed is translated and then back-translated with all low-abundance codons replaced with codons exhibiting a value of N above the cut-off value. The abundance of the remaining codons with N above the cut-off are normalized so that the sum of the abundances of synonymous codons is equal to 1. The codons are then assigned with the aid of a random number generator which gives a random number between 0 and 1. Restarting the program, with a different input to the random number generator will therefore give a different set of oligos. Several sets of oligos can be produced in this way and compared for restriction enzyme sites and melting temperatures. The program allowed the insertion of preferred restriction sites and the generation of oligos 40 nucleotides in length from both strands of the gene. The melting temperature of each of the 20 nt overlaps was adjusted to around 60°C. The resulting set of 104 oligos was subjected to the Genetics Computer Group software package (version 8-Unix) using the options repeat, stemloop and overlap.

Gene assembly and amplification

Gene assembly.

Equal volumes of solutions of each oligo (25 μM each) were combined and the mixture was diluted 10-fold in 50 μl of a PCR mixture [20 mM Tris–HCl, pH 8.8, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 3 mM $MgSO_4$, 0.1% Triton X-100, 0.1 mg/ml BSA, 0.2 mM each dNTP, 2.5 U of *Pfu* polymerase]. The PCR program consisted of one denaturation step at 94°C for 60 s, followed by 25 cycles at 94°C for 30 s, 52°C for 30 s and 72°C for 2 min.

Gene amplification.

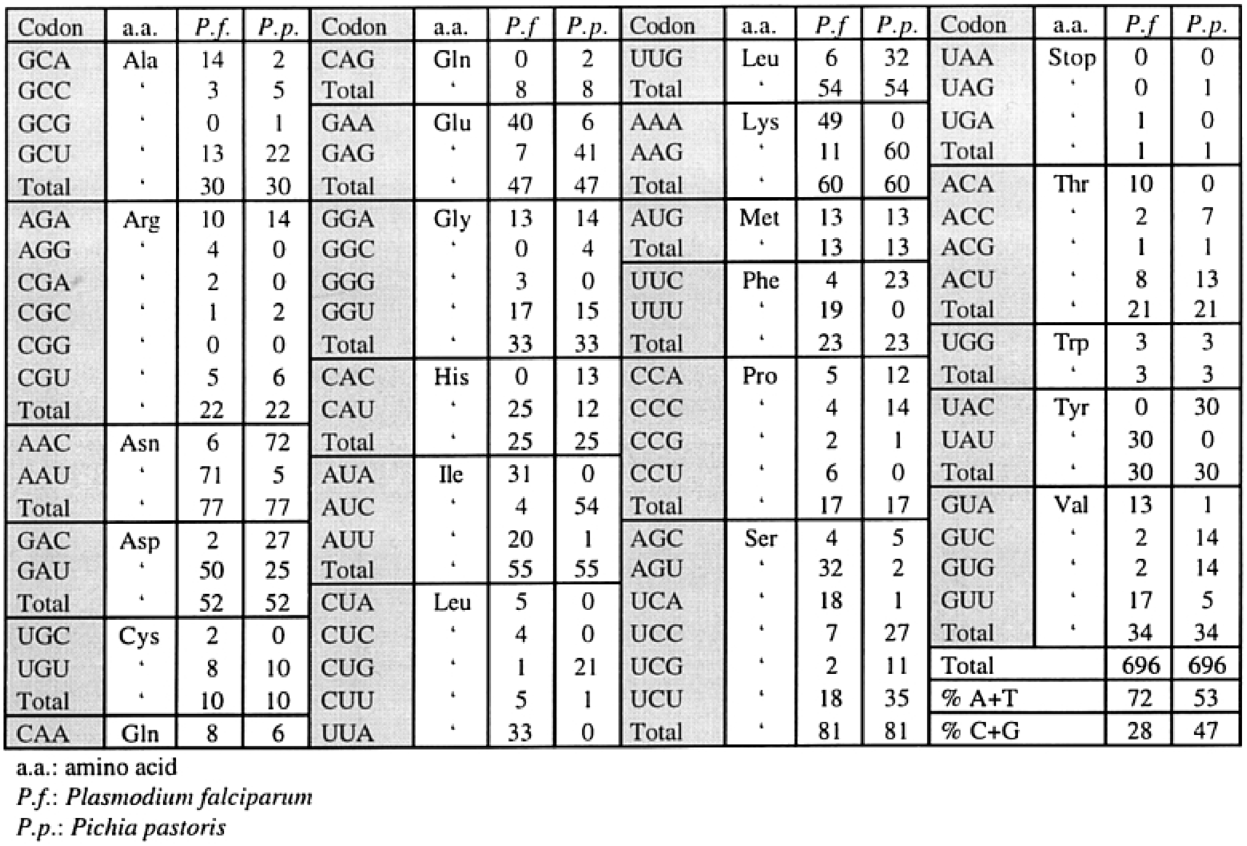
An aliquot of the gene assembly mixture (5 μl) was diluted 10-fold in 50 μl of PCR mixture [20 mM Tris–HCl, pH 8.8, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 3 mM $MgSO_4$, 0.1% Triton X-100, 0.1 mg/ml BSA, 0.2 mM each dNTP, 2.5 U of *Pfu* polymerase and the two outermost primers at 1 μM each]. The outer primers were the same as the two external ones used in the assembly process. The PCR program consisted of a denaturation step cycle at 94°C for 60 s, then 25 cycles at 94°C for 45 s, 68°C for 45 s, 72°C for 5 min and a final incubation cycle at 72°C for 10 min. The PCR product was desalted using a PCR purification kit (Qiagen).

The best results in terms of mutation frequency were obtained when the *pfsub-1* gene was constructed and cloned in two separate sections of 1.1 and 1 kb. Assembly PCR of the two sections was achieved using the same conditions as described above, with 56 oligos being assembled in the first section and 50 oligos in the second section. For the amplification PCR step, the extension time was reduced to 2.5 min.

Cloning steps and DNA sequencing

PCR blunt-ended products were cloned into *pMos*Blue using the *Eco*RV site. The ligation products were used to transform DH5α^{TM} *E.coli* competent cells and selection performed on L agar supplemented with 50 μg/ml ampicillin and 15 μg/ml tetracycline. Plasmids isolated from white colonies were screened for the presence of insert by restriction analysis. For sequencing inserts, primers P6, P15, P24, P33, P42, P58, P67, P76, P85 and P94 were selected from those used in the gene synthesis. Subcloning steps to correct point errors introduced during gene synthesis were performed in

Table I. Codon composition of the native and synthetic *pfsub-1* gene



Codon	a.a.	*P.f.*	*P.p.*	Codon	a.a.	*P.f*	*P.p.*	Codon	a.a.	*P.f*	*P.p.*	Codon	a.a.	*P.f*	*P.p.*
GCA	Ala	14	2	CAG	Gln	0	2	UUG	Leu	6	32	UAA	Stop	0	0
GCC	‘	3	5	Total	‘	8	8	Total	‘	54	54	UAG	‘	0	1
GCG	‘	0	1	GAA	Glu	40	6	AAA	Lys	49	0	UGA	‘	1	0
GCU	‘	13	22	GAG	‘	7	41	AAG	‘	11	60	Total	‘	1	1
Total	‘	30	30	Total	‘	47	47	Total	‘	60	60	ACA	Thr	10	0
AGA	Arg	10	14	GGA	Gly	13	14	AUG	Met	13	13	ACC	‘	2	7
AGG	‘	4	0	GGC	‘	0	4	Total	‘	13	13	ACG	‘	1	1
CGA	‘	2	0	GGG	‘	3	0	UUC	Phe	4	23	ACU	‘	8	13
CGC	‘	1	2	GGU	‘	17	15	UUU	‘	19	0	Total	‘	21	21
CGG	‘	0	0	Total	‘	33	33	Total	‘	23	23	UGG	Trp	3	3
CGU	‘	5	6	CAC	His	0	13	CCA	Pro	5	12	Total	‘	3	3
Total	‘	22	22	CAU	‘	25	12	CCC	‘	4	14	UAC	Tyr	0	30
AAC	Asn	6	72	Total	‘	25	25	CCG	‘	2	1	UAU	‘	30	0
AAU	‘	71	5	AUA	Ile	31	0	CCU	‘	6	0	Total	‘	30	30
Total	‘	77	77	AUC	‘	4	54	Total	‘	17	17	GUA	Val	13	1
GAC	Asp	2	27	AUU	‘	20	1	AGC	Ser	4	5	GUC	‘	2	14
GAU	‘	50	25	Total	‘	55	55	AGU	‘	32	2	GUG	‘	2	14
Total	‘	52	52	CUA	Leu	5	0	UCA	‘	18	1	GUU	‘	17	5
UGC	Cys	2	0	CUC	‘	4	0	UCC	‘	7	27	Total	‘	34	34
UGU	‘	8	10	CUG	‘	1	21	UCG	‘	2	11	Total		696	696
Total	‘	10	10	CUU	‘	5	1	UCU	‘	18	35	% A+T		72	53
CAA	Gln	8	6	UUA	‘	33	0	Total	‘	81	81	% C+G		28	47

a.a.: amino acid
P.f.*: *Plasmodium falciparum
P.p.*: *Pichia pastoris

Gene synthesis was designed according to *P. pastoris* codon usage. Amino acids are referred to using the standard three-letter code.

*pMos*Blue using unique restriction sites within the gene and vector.

Expression in P. pastoris

The synthetic gene was cloned into the *P. pastoris* pPIC9K vector using the *Sna*BI and *Eco*RI sites of the polylinker region. The construct was linearized prior to transformation of the *P. pastoris* GS115 (his4) strain by electroporation (pulse conditions were 1.5 kV at 400 Ω and 25 μF). Several colonies resistant to 1 mg/ml G418 were selected for expression trials in 250 ml shake flasks, containing 50 ml of rich medium (BMGY). The best candidate was grown in a 4 l fermenter, adding 10 μg/ml tunicamycin when switching to induction with BMMY medium. Three days following induction, 20 g of cells were lysed using a cell disrupter and resuspended in a denaturing buffer (8 M urea, 20 mM imidazole, 0.1% v/v Nonidet P40, 0.1 M NaH_2PO_4, 10 mM Tris–HCl, pH 8.2). The recombinant protein was then purified by metal chelate chromatography using Ni–NTA agarose (Qiagen). Bound protein was washed on the column with a 8–0 M gradient of decreasing concentration of urea and finally eluted in a fully soluble form with 50 mM EDTA. Eluted protein was visualized on 10% SDS–PAGE gels by Coomassie Brilliant Blue staining. The band corresponding to PfSUB-1 was confirmed by Western blotting using an antiserum raised against an *E.coli*-derived recombinant PfSUB-1 fragment (Blackman *et al.*, 1998).

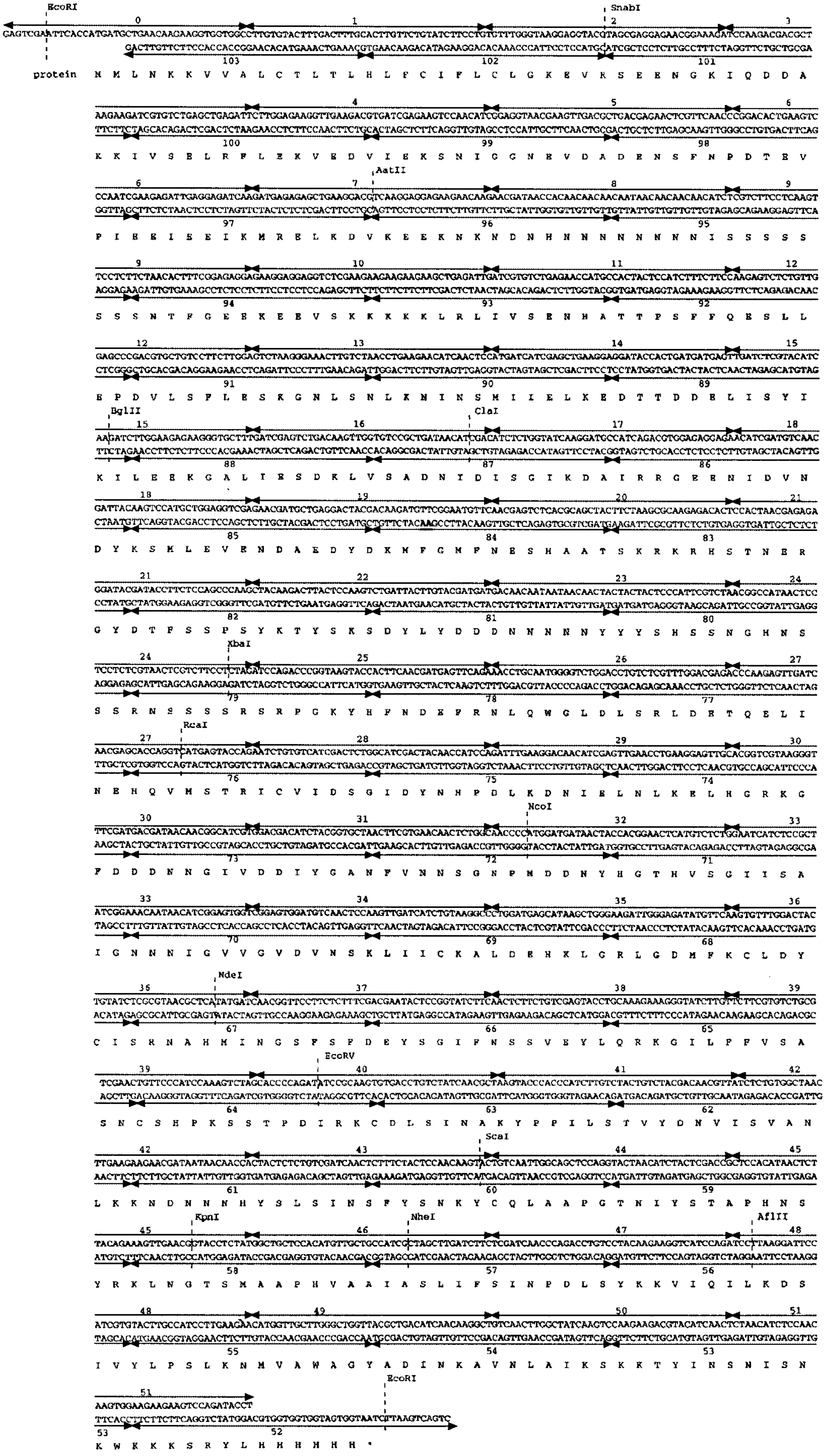
Expression in baculovirus-infected High Five^{TM} insect cells

The full-length synthetic *pfsub-1* gene was subcloned into the *Eco*RI site of the baculovirus transfer vector pVL1393 downstream of the polyhedrin promoter. Sf9 cells were cultured at 27°C in complete TC100 medium and co-transfected with the plasmid construct and the viral DNA according to the Invitrogen guidelines. Recombinant baculoviruses were plaque-purified and single plaques were picked for amplification. High Five^{TM} cells were cultured in ESF 921 protein-free medium. Cells were grown in roller bottles at a density of 10^6 cells/ml and then infected with recombinant baculovirus at various multiplicities of infection. Tunicamycin was added to the High Five^{TM} cells to a final concentration of 0.5 μg/ml at the time of infection. The medium was harvested 72 h post-infection, clarified by centrifugation at 5000 *g* for 30 min and filtered through a 0.22 μm filter (Whatman). The secreted protein was detected in the culture supernatant by Western blot.

Results

Design and assembly of the synthetic gene

The design of the oligos used for synthesis of the 2.1 kb *pfsub-1* gene necessitated great attention to detail, owing to the requirement for a large number to be mixed in one PCR. The nucleotide sequence of the gene was designed according to the *P. pastoris* codon usage preference (Bennetzen and Hall, 1982; Sreekrishna *et al.*, 1993). In addition, the panel of oligos was rigorously screened and matched in order to meet the following criteria: (i) a decrease in the overall A+T content with the elimination of potential transcription termination signals; (ii) elimination of palindromic sequences conducive to stable intramolecular hairpins; (iii) minimization of tandem


EcoRI
SnabI
AatII
BglII
ClaI
XbaI
RcaI
NcoI
NdeI
EcoRV
ScaI
KpnI
NheI
AflII
EcoRI
protein
M M L N K K V V A L C T L T L H L F C I F L C L G K E V R S E E N G K I Q D D A
K K I V S E L R F L E K V E D V I E K S N I G G N E V D A D E N S F N P D T E V
P I E E I S E I K M R E L K D V K E E K N K N D N H N N N N N N N N I S S S S S
S S S N T F G E E K E E V S K K K K K L R L I V S E N H A T T P S F F Q E S L L
E P D V L S F L E S K G N L S N L K N I N S M I I E L K E D T T D D E L I S Y I
K I L E E K G A L I E S D K L V S A D N I D I S G I K D A I R R G E E N I D V N
D Y K S M L E V E N D A E D Y D K M F G M F N E S H A A T S K R K R H S T N E R
G Y D T F S S P S Y K T Y S K S D Y L Y D D D N N N N N Y Y Y S H S S N G H N S
S S R N S S S S R S R P G K Y H F N D E F R N L Q W G L D L S R L D E T Q E L I
N E H Q V M S T R I C V I D S G I D Y N H P D L K D N I E L N L K E L H G R K G
F D D D N N G I V D D I Y G A N F V N N S G N P M D D N Y H G T H V S G I I S A
I G N N N I G V V G V D V N S K L I I C K A L D E H K L G R L G D M F K C L D Y
C I S R N A H M I N G S F S F D E Y S G I F N S S V E Y L Q R K G I L F F V S A
S N C S H P K S S T P D I R K C D L S I N A K Y P P I L S T V Y D N V I S V A N
L K K N D N N N H Y S L S I N S F Y S N K Y C Q L A A P G T N I Y S T A P H N S
Y R K L N G T S M A A P H V A A I A S L I F S I N P D L S Y K K V I Q I L K D S
I V Y L P S L K N M V A W A G Y A D I N K A V N L A I K S K K T Y I N S N I S N
K W K K K S R Y L H H H H H H *

or inverted repeats (<10 bp in length) which are likely to give rise to non-specific priming; and (iv) optimization of the 20 nucleotide overlap between each 40-mer primer, to give a melting temperature in the range 58–62°C, in order to allow subsequent use of the primers for DNA sequencing. A Kozak concensus translation initiation sequence was incorporated in the extreme 5′ oligo for efficient expression of the gene in *P. pastoris* and an additional five histidine codons were introduced just prior to the stop codon in the extreme 3′ oligo. A number of unique restriction sites were introduced at strategic positions throughout the synthetic gene to facilitate subsequent gene manipulation and mutagenesis. Oligo design was performed with the aid of the Unix codon optimization program CODOP (see Materials and methods). This program translates a given DNA sequence into a protein sequence and then, using a user-defined codon usage table, back-translates the protein sequence with an improved codon usage. The program rejects codons with abundances below a cut-off value, then assigns a high-abundance codon to each residue in the protein sequence, using high abundance codons in proportion to their use in the codon usage table. Both strands of the sequence are then divided into overlapping oligos of 40 bases in length, melting temperatures are calculated for all the overlaps and restriction sites generated along the sequence are displayed. The resulting panel of oligos was then analysed using the Genetics Computer Group software package (GCG Version 8-Unix) for the presence of undesirable repeats, inverted repeats, stemloop structures and regions of complementarity which could potentially lead to non-specific intermolecular hybridization. In most cases these sequences were readily eliminated whilst maintaining the codon preference. Non-optimum codons were resorted to only if required to create unique restriction sites or at repetitive sequences. Systematic, reiterative use of these two programs resulted in the final selection of 104 unique oligos for gene synthesis. Table I shows a comparison of the codon composition of the synthetic gene with that of the wild-type *P. falciparum* gene. Codons not present in highly expressed yeast genes have been drastically decreased in frequency and a number of very rare codons eliminated. For example, 31 ATA (Ile) codons and 49 AAA (Lys) codons present in the native gene have been completely removed. The overall A+T composition has been reduced from 72% in the native gene to 53% in the synthetic product. The final, codon-optimized sequence of the synthetic *pfsub-1* sequence and the relative positions of the 104 oligos is shown in Figure 1 together with the predicted amino acid sequence.

The initial assembly reaction (Figure 2) involved the construction of the full-length gene from a stoichiometric mixture of the 104 oligos. An aliquot of this assembly reaction mixture was then used as a template for the amplification process, in which only the two outermost primers of the assembly were added, at a concentration of 1 μM each. Optimum yields of the PCR products using *Pfu* DNA polymerase were obtained with 3 mM $MgSO_4$ in the PCR. Analysis of the two PCRs on 1% agarose gels revealed the presence of the 2.1 kb expected product (Figure 3A). In an alternative approach, the 5′ and 3′ 'halves' of the gene were synthesized separately, in the form of two DNA fragments of 1.1 and 1 kb, respectively (Figure 3B). The PCR conditions for the assembly reaction remained unchanged, although the number of the oligos in each assembly decreased to 56 and 50, respectively, for each reaction. The synthetic DNA products were blunt-end ligated into *pMos*Blue for cloning and sequencing. For sequencing reactions, primers hybridizing to sites ~400 bp apart within the gene were chosen from the panel used in the synthesis, allowing coverage of both strands of the entire *pfsub-1* sequence with consistent overlaps. Complete sequence analysis of three of the 2.1 kb clones and five of each of the smaller clones identified an average of only 3.5 nucleotide substitution errors per kb in the 2.1 kb PCR product and an average of only 1.5 error per kb in each of the two 1.1 and 1 kb products. These mutations were distributed randomly, suggesting that the oligos were not the source of the errors. Since *Pfu* polymerase, which exhibits a 3′ → 5′ proofreading exonuclease activity, was used for all amplification steps, we assume that these errors were most likely introduced during the assembly process. The reduced mutation frequency observed in the smaller products was probably a direct result of the reduction in the number of oligos mixed together during the assembly reaction.

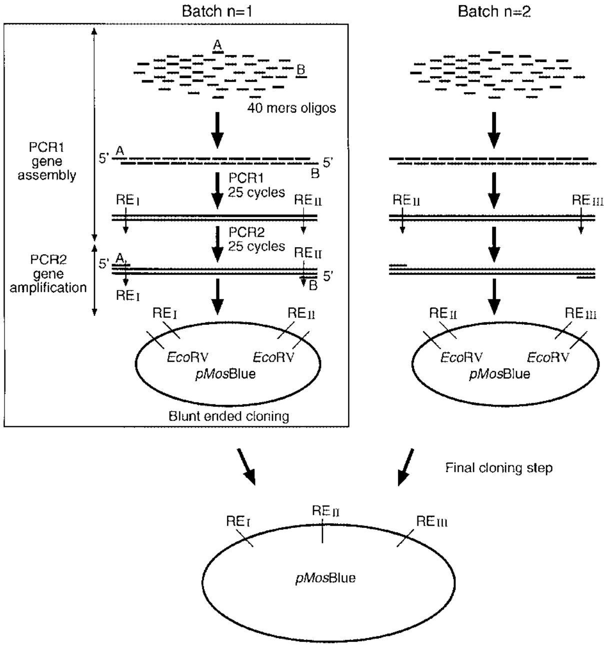


Fig. 2. Schematic illustration of the two-step PCR-based gene synthesis. The synthetic gene is assembled by DNA polymerization from a set of overlapping complementary oligonucleotides in a first PCR (gene assembly). The assembled product is then amplified using the two outermost primers (labelled A and B) in a second PCR (gene amplification). The blunt-ended PCR product is finally cloned into the *Eco*RV site of a suitable cloning vector (*pMos*Blue) for sequencing and further subcloning steps. Synthesis of the complete desired sequence may be achieved in a single batch (shown boxed, labelled Batch n=1) or alternatively portions of the gene may be synthesized separately, e.g. in two batches as shown here or in more batches if required. The gene fragments can then finally be combined by subcloning, making use of an internal common restriction site (RE_{II}). This site and additional restriction sites (RE_I and RE_{III}) incorporated into the outermost primers are used during insertion of the synthetic sequence into expression vectors.

Fig. 1. Nucleotide and deduced amino acid sequence of the synthetic *pfsub-1* gene. Important restriction sites are shown together with the designation and positions of the individual oligonucleotides used for gene synthesis (arrowed).

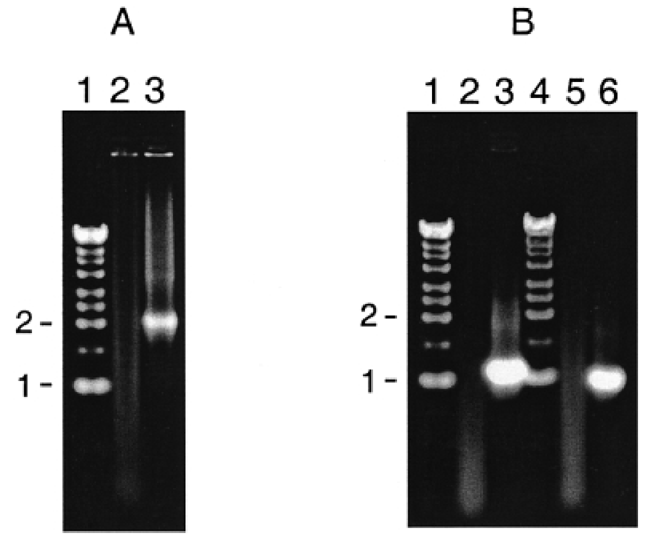


Fig. 3. Agarose gel electrophoresis of synthetic gene products. (A) Analysis on a 1% gel of the PCR products from the assembly (lane 2) and amplification (lane 3) steps of a single batch *pfsub-1* gene synthesis. (B) Analysis on a 1% gel of the PCR products from a dual batch *pfsub-1* gene synthesis. Lanes 2 and 3 contain, respectively, the products of the assembly and amplification steps of synthesis of the 1.1 kb 5′ 'half' of the gene, whilst lanes 5 and 6 contain, respectively, the products of the assembly and amplification steps of the remaining 1 kb 3′ portion of the gene. Remaining tracks contain DNA molecular weight markers (KiloBase DNA markers, Pharmacia). The positions of migration of the 1 and 2 kb marker fragments are indicated.

Protein expression

Expression from the synthetic *pfsub-1* gene was initially assessed in *P. pastoris*. The PfSUB-1 signal sequence was replaced by the pre-pro domain of the *S. cerevisiae* α-mating factor by cloning the gene into *Sna*BI /*Eco*RI-digested pPIC9K and the linearized vector used to transform *P. pastoris*. Transformants containing multiple chromosomal insertions of the recombinant vector were selected and, following preliminary inductions to select the highest producing clones, a single clone was induced in a 4 l fermenter. Since N-glycosylation of blood-stage *P. falciparum* proteins is rare (Gowda *et al.*, 1997), induction was performed in the presence of tunicamycin. No recombinant protein was secreted by the clone. Examination of total cell extracts by Western blotting showed that the induced recombinant product accumulated intracellularly in an insoluble form. Taking advantage of the C-terminal hexahistidine tag, the recombinant protein was purified under denaturing conditions from extracts of the induced clone by nickel chelate chromatography (Holzinger *et al.*, 1996) (Figure 4). From these purification data, the expression level of the recombinant PfSUB-1 was estimated at 0.2–0.5 g/l. N-terminal amino acid sequencing of the purified protein showed that the α-factor N-terminal secretory signal sequence had been removed whereas the α-factor pro domain was still present, suggesting that the protein had undergone translocation into the yeast ER but not been further processed.

The baculovirus system was next considered as an alternative expression system which might better support proper folding and post-translational processing of PfSUB-1. The codon usage of *Autographa californica* nuclear polyhedrosis virus (AcMNPV) is less stringent than that of *P. pastoris (*Ranjan and Hasnain, 1994), so it was considered that our synthetic gene should also be well expressed in the baculovirus system. Infection of High Five^{TM} insect cells in the presence of tunicamycin at 0.5 μg/ml resulted in secretion of readily detectable levels of apparently correctly processed PfSUB-1 (Figure 5). Preliminary purification runs with the secreted product indicated that expression levels of the recombinant protein were of the order of 2–5 mg/l (not shown).

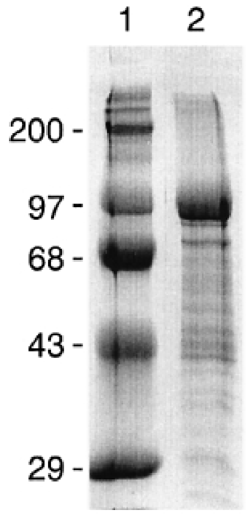


Fig. 4. Expression of the synthetic *pfsub-1* gene in *P. pastoris*. A *P. pastoris* clone containing multiple genomic copies of the synthetic *pfsub-1* gene was induced and the recombinant gene product purified in a single step from the insoluble fraction of the cell pellet by metal chelate chromatography on Ni–NTA agarose. Purified protein was subjected to SDS–PAGE on a 10% gel and visualized by Coomassie Brilliant Blue staining. The positions and sizes in kDa of molecular weight marker proteins (lane 1) are indicated.

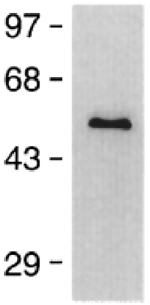


Fig. 5. Expression of the synthetic *pfsub-1* gene in baculovirus-infected insect cells. High Five^{TM} insect cells were infected with recombinant baculovirus carrying the synthetic *pfsub-1* gene under the control of the polyhedrin promoter. Culture supernatant harvested 72 h post infection was subjected to electrophoresis on a 10% SDS–polyacrylamide gel. Fractionated proteins were transferred to nitrocellulose and the blot was probed with an anti-PfSUB-1 antiserum. The secreted recombinant product is secreted as a 54 kDa protein, which is the same size as the primary processing product of *pfsub-1* expression in the malaria parasite (Blackman *et al.*, 1998). Positions and sizes (in kDa) of marker proteins are indicated.

Discussion

We have described a PCR-based gene synthesis method for the fast and accurate construction of the 2.1 kb *P. falciparum pfsub-1* gene. The flexibility of the method, in which complementary oligos are mixed together to generate a synthetic product in a single reaction, is impressive. There is no practical limit on the number of oligos which may be mixed together, suggesting that genes much bigger than *pfsub-1* could be successfully synthesized. Although the synthesis itself is guaranteed by careful and systematic oligo design, the introduction of point errors into the synthetic sequence is difficult to avoid.

Since the most time-consuming step of the synthesis is the removal of such mutations by subsequent subcloning steps, the crucial parameter for overall success of the method is determined by the fidelity of the polymerase used in the reaction. This becomes even more critical when no functional assay is available for the facile identification and elimination of clones bearing mutations in sequences encoding functionally critical residues. Stemmer *et al.* (1995) used a combination of *Taq* polymerase and *Pfu* polymerase to ensure both processivity and proofreading activities for assembly PCR synthesis of the 0.9 kb β-lactamase gene; following selection and cloning of the synthetic gene in a plasmid bearing a tetracycline resistance gene, ampicillin resistance assays indicated that only 76% of the clones expressed functionally active β-lactamase. Sequencing of one of these clones indicated the presence of three point mutations. Using similar PCR conditions for *pfsub-1* synthesis, we observed an average error frequency of nine point substitutions per kb in our final cloned product (C.Withers-Martinez, unpublished data). Accordingly, we set out to modify the PCR conditions so as to allow the exclusive use of *Pfu* polymerase, as this enzyme exhibits the highest fidelity of any thermostable DNA polymerase, with an average error rate of 1.3×10^{-6} per bp duplicated (Cline *et al.*, 1996) Under our optimized conditions, the overall error frequency for our 2.1 kb synthetic product was only 3.5 point substitutions per kb. By dividing the gene synthesis into sections of about 1 kb to be subsequently combined by cloning steps, we were able to reduce the mutation frequency further to only 1.5 nucleotide substitution error per kb. This error frequency is at least twofold lower than that reported by Stemmer *et al.* (1995) and is sufficiently low that a functional screen is unnecessary and screening of clones by direct sequence analysis becomes feasible and cost-effective. Since the frequency of PCR-derived errors increases with increasing number of amplification cycles, it might be possible to reduce the error rate even further by using fewer cycles in the reaction; we did not explore this possibility in the present study, but the large amount of PCR product obtained suggests that it may be possible to reduce the number of cycles significantly whilst still obtaining acceptable yields. The oligos used in the present method were relatively short compared with some other reported PCR-based gene synthesis methods, where oligos larger than 80 nucleotides were used (Prodromou and Pearl, 1992; Casimiro *et al.*, 1997). The use of large oligos has certain advantages; gene design is simplified and a smaller number of oligos can be used in a single reaction, reducing the number of oligo ends and potentially reducing the frequency of PCR-derived errors. Also, the synthesis of larger oligos may be more cost-effective, particularly in the case of large genes. On the other hand, shorter oligos have the important advantage of being less likely to form secondary structures in solution and are less likely to contain errors introduced during synthesis. The use of short oligos also has advantages if we consider the long-term study of the synthesized molecule. First, since they have been designed to possess similar melting temperatures, they can be used again as primers for sequencing of the newly synthesized gene. Second, one or several oligos can be substituted by oligos containing mutations and subsequently assembled in a fragment flanked by unique restriction sites at each end, simplifying and speeding protein mutagenesis experiments.

The complete gene synthesis process, including assembly and amplification, allowed the production of the final synthetic product in one day. Including the subsequent steps of DNA sequencing and subcloning, the constructs used in the *P. pastoris* expression experiments were completed within a matter of weeks. Although expensive in terms of initial capital outlay (the total cost of the oligos used here, at £0.6 per nucleotide, was of the order of £2500), the simplicity and accuracy of the PCR-based gene synthesis described here render it feasible to consider the routine, complete synthesis of malarial genes of interest prior to attempting expression in any heterologous system. Our preliminary expression results are extremely encouraging; PfSUB-1 is expressed intracellularly in *P. pastoris* in the range 0.2–0.5 g/l, providing enough material for extensive refolding assays. The protein is also expressed and secreted in baculovirus-infected High Five^{TM} cells in a correctly processed form. The significant amount of protein produced will readily allow enzymological and structural studies.

New approaches to malaria control will require an improved understanding of mechanisms of drug action and resistance, the identification of new drug targets, improved diagnostic tools and the development of an effective vaccine. All of these aims will be facilitated by the ongoing malaria genome project, which was initiated in 1996 and is progressing rapidly, as evidenced by the recent completion of the entire sequence of chromosome 2 of the 14 chromosome, ~30 megabase haploid *P. falciparum* nuclear genome (Gardner *et al.*, 1998). The genetic information generated by this project will provide access to the complete array of *P. falciparum* open reading frames, allowing researchers readily to identify and study potential chemotherapeutic targets and vaccine candidate antigens. There is little doubt that the rate-limiting step in fully realizing the potential of these advances in genomics will be that of heterologous expression of correctly folded, functionally active parasite gene products for characterization at the molecular structural level; the A+T bias of the parasite genome will constitute a permanent problem in this regard. The systematic PCR-based approach to gene redesign described here will bridge the existing technological gap between identification of putative targets at the nucleotide level and their expression for structure–function studies. In this Institute, the gene synthesis method described here has already been applied successfully to the synthesis of two other malarial genes (C.Withers-Martinez, unpublished data).

Additional data

Sequence data from this study have been deposited with the EMBL/GenBank Data Libraries under Accession Number AJ242589.

Acknowledgements

We thank Dr Anthony Holder for critical reading of the manuscript and for many helpful discussions. We thank Professor Guy Dodson for his enthusiastic support of this project. C.W.-M. was on a sabbatical scientific leave from the CNRS-UPR9039 (AFMB), IBSM, 31 Chemin Joseph-Aiguier, 13402 Marseille cedex 20, France, and was supported by an EMBO long-term fellowship.

References

Anon. (1994) *Weekly Epidemiol. Rec.*, **69**, 309–314.

Au,L., Yang,F., Yang,W., Lo,S. and Kao,C. (1998) *Biochem. Biophys. Res. Commun.*, **248**, 200–203.

Bennetzen,J.L. and Hall,B.D. (1982) *J. Biol. Chem.*, **257**, 3026–3031.

Blackman,M.J., Chappel,J.A., Shai,S. and Holder. A.A. (1993) *Mol. Biochem. Parasitol.*, **62**, 103–114.

Blackman,M.J., Fujioka,H., Stafford,W.H.L., Sajid,M., Clough,B., Fleck,S.L., Aikawa,M., Grainger,M. and Hackett,F. (1998) *J. Biol. Chem.*, **273**, 23398–23409.

Brocca,S., Schmidt-Dannert,C., Lotti,M., Alberghina,L. and Schmid,R.D. (1998) *Protein Sci.*, **7**, 1415–1422.

Casimiro,D.R., Wright,P.E. and Dyson,H.J. (1997) *Structure*, **5**, 1407–1412.

Cline,J., Braman,J.C. and Hogrefe H.H. (1996) *Nucleic Acids Res.*, **24**, 3546–3551.

de Bruin,D., Lanzer,M. and Ravetch,J.V. (1992) *Genomics*, **14**, 332–339.

Gardner,M.J. *et al.* (1998) *Science*, **282**, 1126–1132.

Gowda,D.C., Gupta,P. and Davidson,E.A. (1997) *J. Biol. Chem.*, **272**, 6428–6439.

Graham,R.W., Atkinson,T., Kilburn,D.G. Miller,R.C.,Jr and Warren R.A.J. (1993) *Nucleic Acids Res.*, **21**, 4923–4928.

Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) *Nucleic Acids Res.*, **8**, r49–r62.

Hale,R.S. and Thompson,G. (1998) *Protein Express. Purif.*, **12**, 185–188.

Hernan,R.A., Hui,H.L., Andracki,M.E., Noble,R.W., Sligar,S.G., Walder,J.A. and Walder,R.Y. (1992) *Biochemistry*, **31**, 8619–8628.

Holzinger,A., Phillips,K.S. and Weaver,T.E. (1996) *BioTechniques*, **20**, 804.

Makoff,A.J., Oxer,M.D., Romanos,M.A., Fairweather,N.F. and Ballantine,S. (1989) *Nucleic Acids Res.*, **17**, 10191–10202.

Martin,S.L., Vrhovski,B. and Weiss,A.S. (1995) *Gene*, **154**, 159–166.

McKerrow,J.H., Sun,E., Rosenthal,P.J. and Bouvier,J. (1993) *Annu. Rev. Microbiol.*, **47**, 821–53.

Mehta,D.V., DiGate,R.J., Banville,D.L. and Guiles,R.D. (1997) *Protein Express. Purif.*, **11**, 86–94.

Prapunwattana,P., Sirawaraporn,W., Yuthavong,Y. and Santi,D.V. (1996) *Mol. Biochem. Parasitol.*, **83**, 93–106.

Prodromou,C. and Pearl,L.H. (1992) *Protein Engng*, **5**, 827–829.

Ranjan,A. and Hasnain,S.E. (1994) *Virus Genes*, **9**, 149–153.

Romanos,M.A., Makoff,A., Fairweather,N.F., Beesley,K.M., Slater,D.E., Rayment,F.B., Payne,M.M. and Clare,J.J. (1991) *Nucleic Acids Res.*, **19**, 1461–1467.

Schmidt-Dannert,C., Pleiss,J. and Schmidt R.D. (1998) *Ann. N. Y. Acad. Sci.*, **864**, 14–22.

Sreekrishna,K. (1993) In Baltz,R.H., Hegeman,G.D. and Skatrud,P.L. (eds) *Industrial Microorganisms: Basic and Applied Molecular Genetics*. American Society of Microbiology, Washington, DC. Chapter 16, pp. 119–126.

Sreekrishna,K., Brankamp,R.G., Kropp,K.E., Blankenship,D.T., Tsay,J.T., Smith,P.L., Wierschke,J.D., Subramaniam,A. and Birkenberger,L.A. (1997) *Gene*, **190**, 55–62.

Stemmer,W.P.C., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) *Gene*,**164**, 49–53.

Triglia,T. and Kemp,D.J. (1991) *Mol. Biochem. Parasitol.*, **44**, 207–212.

Wilson,R.J. *et al.* (1996) *J. Mol. Biol.*, **261**, 155–172.

Received July 7, 1999; revised September 1, 1999; accepted September 15, 1999