J Mol Evol (2004) 59:528–535
DOI: 10.1007/s00239-004-2645-4

JOURNAL OF MOLECULAR EVOLUTION

The Evolution of Amino Acid Repeat Arrays in *Plasmodium* and Other Organisms

Austin L. Hughes

Department of Biological Sciences, University of South Carolina, Coker Life Sciences Building, 700 Sumter Street, Columbia, SC 29208, USA

Received: 26 February 2004 / Accepted: 1 May 2004 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. Repeat arrays in protein-coding sequences were analyzed by a novel approach, based on analyzing the distribution of the pairwise proportion of nucleotide differences among units within a repeat array. The results showed that evidence of recent repeat array expansion was particularly characteristic of the repeat arrays of the malaria parasites (genus *Plasmodium*), supporting the hypothesis that *Plasmodium* is particularly prone to repeat array expansion by slipped-strand mispairing or a similar mechanism. Repeat arrays in *Plasmodium* asexual-stage antigens (which are exposed to the immune system of the vertebrate host) had unique characteristics with respect to the number of repeat units, as well as nucleotide and amino acid composition, suggesting that natural selection exerted by the host immune system has shaped features of these arrays.

Key words: Amino acid repeats — Immune evasion — *Plasmodium* — Slipped-strand mispairing

Introduction

In addition to the numerous well-documented repeating segments in noncoding regions (Wickstead et al. 2003), the coding regions of a great many organisms have been found to include repeating segments that encode repeated amino acid sequence motifs ranging from a single residue to repeated domains of 90 or more residues such as fibronectin type III domains. Evolutionary studies of longer repeated domains have applied phylogenetic methods to individual repeat units and have revealed that repeated domains may sometimes evolve in a concerted fashion (Hughes 1999, 2000; Thomas et al. 1997). In this concerted mode of evolution, repeat units duplicate independently within evolutionary lineages so that there are no orthologous relationships between corresponding repeat units in different lineages. Over evolutionary time, such repeat arrays have been found to expand and contract independently by duplication and deletion of individual repeat units.

In the case of shorter repeat units, phylogenetic analysis is not generally applicable because of the lack of sufficient information to reconstruct phylogenetic relationships. However, several studies have reported evidence of expansion and contraction of sort repeat arrays (e.g., Delhomme and Djian 2000; Van Rheede et al. 2003). Repeats of short amino acid motifs are particularly widespread among the surface proteins of the asexual stages of malaria parasites belonging to the genus *Plasmodium* (Protista: Apicomplexa) (Kemp et al. 1987). For example, the circumsporozoite protein (CSP), the major surface protein of the infective stage of the virulent human malaria parasite *Plasmodium falciparum*, includes about 40 repeats of the four-amino-acid motif Asn–Ala–Asn–Pro (NANP) or close variants. There is abundant allelic variation with respect to the number

Correspondence to: Austin L. Hughes; *email:* austin@biol.sc.edu

of repeat units, suggesting that expansion and contraction of these arrays are ongoing (Jongwutiwes et al. 1994).

Comparative studies of CSP sequences have revealed a number of unusual features. First, repeat array sequences differ markedly between different species of *Plasmodium* and even sometimes within a single species, as in the case of the monkey malaria parasite *P. cynomolgi*. This observation suggests that the repeat arrays evolve by some sort of concerted process. Furthermore, within *P. falciparum*, the NANP motif is encoded by only a small fraction of the possible codon combinations. Just 1 of the 64 possible ways of encoding NANP accounted for nearly 50% observed in a study of naturally occurring isolates (Jongwutiwes et al. 1994). This observation suggests that individual repeat units have spread recently by a rapid mechanism (Jongwutiwes et al. 1994) such as slipped-strand mispairing (Rich and Ayala 200).

It has been hypothesized that the presence of repeats in antigenic proteins of the asexual stages (which are found in the vertebrate host) is adaptive for the malaria parasite in evading immune recognition by the host (Kemp et al. 1987; Bickle et al. 1993; Reeder and Brown 1996). It has been hypothesized that the repeat arrays serve as a ''smoke-screen'' that elicits a strong but ineffective immune response on the part of the host (Kemp et al. 1987). Consistent with this hypothesis is the observation that, although there is a strong antibody response to the repeat regions of the *Plasmodium* CSP (Zavala et al. 1983), this response is not effective in eliminating infection (Nussenzweig and Nussenzweig 1984). Also consistent with this hypothesis is evidence that repeat arrays of *Plasmodium* antigens have a biased amino acid composition, including residues likely to form epitopes for host antibody (Verra and Hughes 1999).

The present paper uses a novel approach to study the evolutionary characteristics of repeat arrays in coding regions. The approach is based on the concept of a mismatch distribution (Sherry et al. 1994). The distribution of the pairwise proportion of nucleotide differences among individual repeat units is expected to reflect the mode of evolution of the repeat array. If repeat units have been duplicated recently by slipped-strand mispairing or a similar mechanism, it is expected that the mean proportion of nucleotide differences among repeats will, and the distribution will be skewed to the right (Fig. 1a). On the other hand, if the repeat units have an ancient origin, the mean proportion of nucleotide differences is expected to be higher, and the distribution relatively symmetrical (Fig. 1b). This approach was used to compare repeat arrays from CSP and other *Plasmodium* antigens with those of protein-coding genes of a wide variety of eukaryotes, prokaryotes, and viruses. In addition, because of previous evidence of biased amino acid composition in *Plasmodium* repeats (Verra and Hughes 1999), amino acid composition was compared among repeats from different groups of organisms.

Methods

Proteins including arrays of repeats of the same (or similar) short (3- to 25-residue) amino acid motif were identified in the GenBank database. Related sequences were located by BLASTP (Altschul et al. 1997) sequence homology search. Individual repeat units making up a given array were aligned by eye at the amino acid level; only cases with an amino acid sequence difference of 20% between repeat units were included. In the case of *Plasmodium falciparum*, for which a complete genome sequence is available, I chose only proteins whose repeats met the above criteria and for which some functional knowledge was available.

For each repeat array, all repeat units were compared pairwise at the DNA level. For a given array consisting of n repeat units, there were thus $(n^2 - n)/2$ pairwise comparisons among repeat units. The following quantities were computed for each repeat array: (1) *repeat units*, the number of repeat units in the array; (2) *residues per unit*, the mean number of residues per repeat unit; (3) p, the mean proportion of nucleotide difference for all pairwise comparisons among repeat units; (4) *skew*, the skewness of the distribution of p for all pairwise comparisons; (5) *prop. 0*, the proportion of all pairwise comparisons in which p was zero; (6) *prop.* >0.25, the proportion of all pairwise comparisons in which p was greater than 0.25; (7) *prop. hydrophilic*, the proportion of hydrophilic residues; (8) *prop. hydrophobic*, the proportion of hydrophobic residues; (9) pAT_1, the proportion A + T at the first codon position in repeats; (10) pAT_2, the proportion A + T at the second codon position in repeats; and (11) pAT_3, the proportion A + T at the third codon position in repeats.

The distributions of most of these variables were found to deviate significantly from normality (Kolmogorov–Smirnov test); therefore, nonparametric methods based on medians were used in hypothesis testing. The hypothesis of equality of medians across different categories of repeat array types was tested by the Kruskal–Wallis one-way nonparametric analysis of variance (Hollander and Wolfe 1973). All tests using this procedure were two-tailed. In preliminary analyses, when parametric methods were used, the results were essentially identical to those obtained with nonparametric methods.

The unit of analysis was the repeat array type ($N = 338$), defined as a set of repeats having homologous sequences. If more than one sequence was available with homologous sequences, values of the eight variables were averaged to give the value for the repeat array type. For example, there were 22 available allelic sequences for the *Csp* gene (encoding the circumsporozoite protein or CSP) of *Plasmodium falciparum*, all having repeats of the NANP motif (or some close variant). The CSP of *P. falciparum* constituted a single repeat array time, and values of the eight variables were averaged for all 22 alleles to constitute the entries for this repeat array time. Overall, 837 individual sequences were used in computing the values of the eight variables for the 338 repeat array types (see Supplementary Information).

Amino acid residues were classified as hydrophilic (D, E, K, R) or hydrophobic (C, F, I, L, M, V, W, and Y) following the hydrophobicity scale of Hopp and Woods (1981). Although numerous hydrophobicity scales are available in the literature, that of Hopp and Woods (1981) was reported to outperform others in predicting epitopes recognized by vertebrate immunoglobulins. Using this scale, Verra and Hughes (1999) found that repeats in

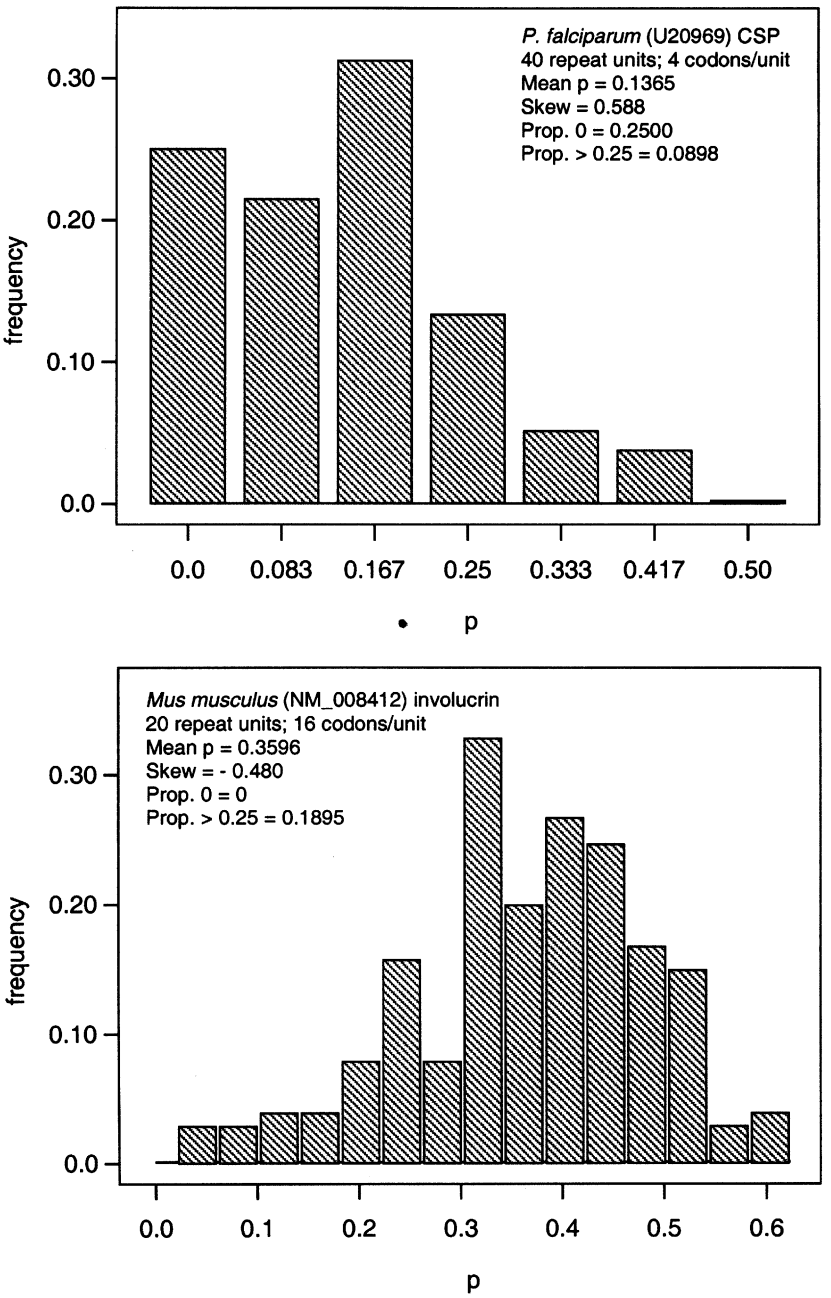

Fig. 1. Examples of the distribution of the proportion of nucleotide difference (p) in pairwise comparisons among repeat units. **A** Circumsporozoite protein (CSP) of *Plasmodium falciparum*, an example with a low mean p, high positive skew, a high proportion of comparisons in which $p = 0$, and a low proportion of comparisons in which $p > 0.25$. **B** Mouse involucrin, an example with a high mean p, negative skew, a low proportion of comparisons in which $p = 0$, and a high proportion of comparisons in which $p > 0.25$.

antigens of *Plasmodium falciparum* showed a tendency to avoid hydrophobic residues to a greater extent than expected based on nucleotide composition, and the avoidance of hydrophobic residues is a characteristic of epitopes recognized by immunoglobulins.

Results

Mismatch Distributions for Repeat Arrays

Table 1 summarizes medians and ranges of variables describing repeat arrays. The repeat arrays in *Plasmodium* proteins were characterized by low median *p*, by high positive median *skew*, by high median *prop. 0*, and by low median *prop. >0.25* (Table 1). Thus, in general, repeat arrays in *Plasmodium* antigens included a high proportion of repeat units identical or nearly so at the DNA level, indicating recent expansion of the repeat array. Certain other repeat array types showed similarity to those of *Plasmodium* antigens. Repeat arrays from antigens of parasitic protists other than *Plasmodium* shared the characteristics of low median *p* and high median *prop. 0* (Table 1). However, these other protist antigens differed from those of *Plasmodium* in having relatively low median *skew*. Repeat arrays of viruses, like those of *Plasmodium* antigens, had low median *p*, high median *skew*, and high median *prop. 0* (Table 1).

Table 1. Median (range) of variables describing repeat array characteristics

Data set	No. array types	Repeat units	Residues/ unit	*p*	*Skew*	*Prop. 0*	*Prop. >0.25*
Plasmodium							
CSP	28	17.0 (3, 65)	9.0 (4, 17)	0.10 (0.01, 0.25)	0.42 (−1.49, 1.36)	0.24 (0.00, 0.71)	0.03 (0.00, 0.67)
Other asexual-stage antigens	20	15.5 (3, 65)	8.0 (3, 20)	0.09 (0.00, 0.30)	0.27 (−1.51, 2.60)	0.18 (0.05, 0.89)	0.00 (0.00, 0.69)
Nonantigens	18	9.3 (4, 22)	7.0 (3, 10)	0.09 (0.02, 0.25)	0.21 (−2.12, 1.18)	0.17 (0.00, 0.61)	0.04 (0.00, 0.50)
Sexual-stage surface ptns.[a]	8	13.5 (5, 36)	4.5 (3, 10)	0.15 (0.0, 0.25)	−0.10 (−0.57, 1.19)	0.27 (0.03, 0.43)	0.11 (0.00, 0.38)
Sexual-stage nonsurface ptns.	14	8.0 (4, 18)	6.0 (3, 10)	0.15 (0.05, 0.25)	0.36 (−0.67, 0.62)	0.16 (0.05, 0.37)	0.14 (0.00, 0.50)
Parasitic protists (antigens)	13	8.3 (3, 30)	8.0 (4, 24)	0.11 (0.03, 0.55)	0.20 (−0.94, 1.40)	0.26 (0.00, 0.73)	0.09 (0.00, 1.00)
Dictyostelium	24	7.0 (3, 24)	6.0 (3, 23)	0.20 (0.01, 0.44)	0.34 (−1.73, 1.67)	0.09 (0.00, 0.60)	0.39 (0.00, 1.00)
Other nonparasitic protists	10	5.8 (4, 28)	5.0 (3, 23)	0.27 (0.14, 0.55)	−0.12 (−0.40, 0.81)	0.07 (0.00, 0.29)	0.54 (0.00, 1.00)
Bordatella							
Pertactin 5-residue	23	5.0 (3, 8)	5.0 (5, 5)	0.22 (0.16, 0.33)	−0.42 (−1.55, 0.48)	0.17 (0.00, 0.52)	0.53 (0.33, 0.67)
Pertactin 3-residue	13	5.0 (3, 8)	3.0 (3, 3)	0.10 (0.03, 0.23)	0.95 (0.00, 1.73)	0.57 (0.00, 0.73)	0.00 (0.00, 0.40)
Mycobacterium							
PPE 10-residue	17	26.0 (6, 38)	10.0 (10, 10)	0.42 (0.32, 0.52)	0.16 (−0.29, 0.60)	0.00 (0.00, 0.00)	0.94 (0.80, 1.00)
PPE 25-residue	6	5.0 (4, 10)	25.0 (23, 26)	0.26 (0.10, 0.35)	−0.26 (−1.29, 0.81)	0.00 (0.00, 0.33)	0.53 (0.10, 1.00)
Other bacteria							
Antigens	7	10.4 (4, 17)	8.0 (5, 12)	0.23 (0.13, 0.30)	0.10 (−0.36, 1.03)	0.04 (0.00, 0.07)	0.42 (0.05, 0.65)
Nonantigens	17	6.0 (3, 39)	7.0 (3, 20)	0.25 (0.03, 0.40)	0.09 (−0.50, 1.22)	0.01 (0.00, 0.75)	0.44 (0.00, 0.88)
Viruses	24	9.3 (3, 57)	6.0 (3, 16)	0.12 (0.00, 0.42)	0.38 (−0.59, 3.16)	0.25 (0.00, 1.00)	0.02 (0.00, 1.00)
Fungi	22	6.8 (3, 41)	7.5 (3, 26)	0.22 (0.09, 0.57)	−0.17 (−1.78, 0.50)	0.00 (0.00, 0.57)	0.34 (0.00, 1.00)
Plants	23	5.3 (3, 32)	10.0 (3, 22)	0.17 (0.00, 0.34)	0.00 (−1.08, 0.86)	0.00 (0.00, 0.42)	0.12 (0.00, 0.65)
Invertebrates	21	8.6 (4, 75)	7.3 (3, 19)	0.21 (0.04, 0.43)	−0.24 (−1.08, 0.81)	0.05 (0.00, 0.44)	0.31 (0.00, 1.00)
Vertebrates	30	6.6 (3, 55)	8.5 (3, 55)	0.31 (0.12, 0.55)	0.08 (−1.53, 1.32)	0.01 (0.00, 0.30)	0.58 (0.00, 1.00)

[a]ptns., proteins.

By contrast, the repeat arrays of nonparasitic organisms tended to have low or even negative *skew*, low *prop. 0*, and high *prop. >0.25* (Table 1). In most nonparasitic organisms, median *p* was relatively high, with the highest median value for nonparasites being that of vertebrates (Table 1). Interestingly, repeat arrays from bacteria, even the putatively antigenic proteins of parasitic bacteria, sometimes had characteristics more similar to those of nonparasitic organisms than those of parasitic protists. For example, the repeats of the PPE family of *Mycobacterium* had high median *p*, low median *skew*, very low median *prop. 0*, and very high median *prop. >0.25* (Table 1). An exception to this trend was seen in the three-residue repeats in pertactin of *Bordatella*. Unlike the five-residue repeats in the same protein, the three-residue pertactin repeats showed a pattern reminiscent of *Plasmodium* antigens: low median *p*, high median *skew*, high median *prop. 0*, and low median *prop. >0.25* (Table 1).

Tests of Hypotheses Regarding Plasmodium *Arrays*

Kruskal–Wallis nonparametric one-way analysis of variance was used to compare repeat array characteristics from *Plasmodium* asexual-stage antigens (ASSA), other *Plasmodium* proteins, and proteins from other organisms. The ASSA are the primary antigens exposed to the vertebrate immune system; thus, evidence that the repeats in these antigens share characteristics not seen in other *Plasmodium* repeats would support the hypothesis that host immune evasion has exerted selective pressure on these repeat arrays. For purposes of these analyses, the sexual-stage surface proteins were not grouped with ASSA because of the likelihood that the former are minimally exposed to the vertebrate immune system. These three categories showed significant differences with respect to all of the variables summarized in Table 1 (Fig. 2). The tests were significant at the 0.001 level or better for all variables except *residues/unit*, in which case the test was significant at the 0.05 level.

In individual comparisons, median values for non-*Plasmodium* repeat arrays were significantly different from those for *Plasmodium* ASSA in the case of every variable except *residues/unit* (Fig. 2). Non-*Plasmodium* repeat arrays had significantly lower median numbers of *repeat units* than did repeat arrays in *Plasmodium* ASSA (Fig. 2a), significantly higher median *p* than arrays in *Plasmodium* ASSA (Fig. 2c), significantly lower median *skew* than arrays in

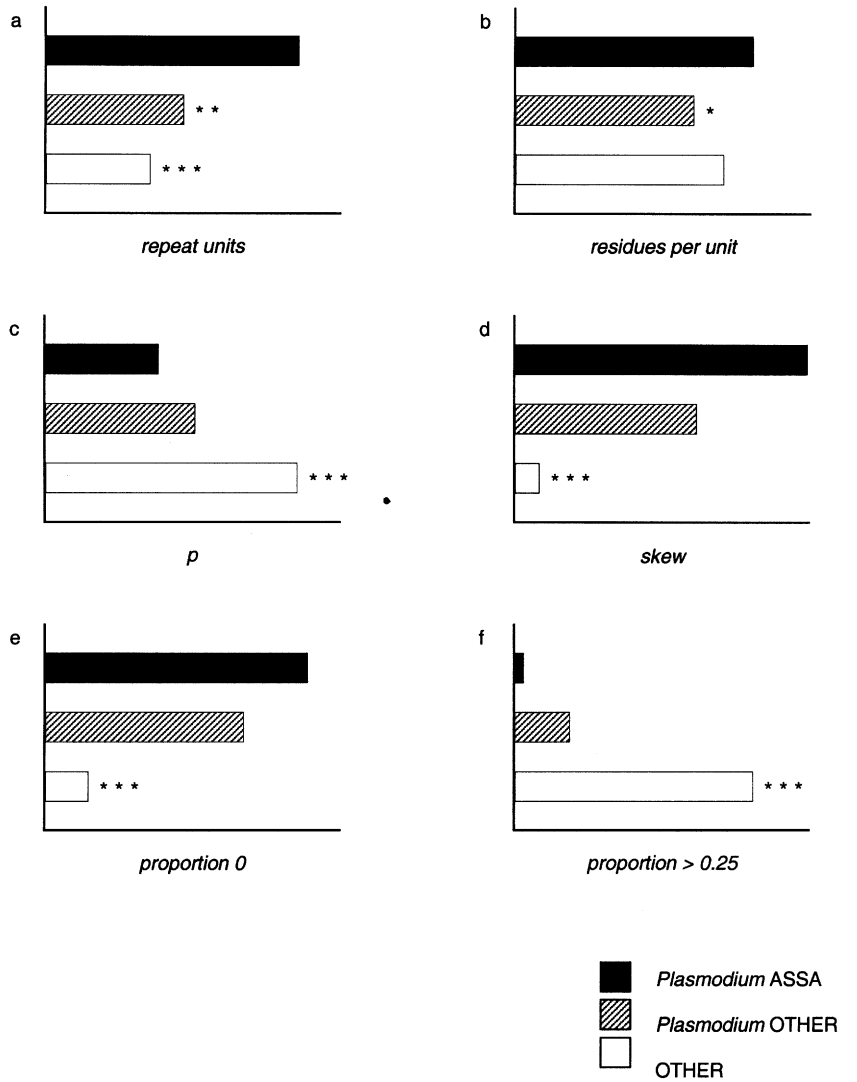

Fig. 2. Medians of variables describing repeat arrays in *Plasmodium* asexual-stage antigens (ASSA), other *Plasmodium* proteins, and other proteins. Kruskal–Wallis test, $p < 0.001$ for all variables except *residues per unit* (for which $p < 0.05$). Tests of the equality of individual group medians with that of *Plasmodium* ASSA: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

Plasmodium ASSA (Fig. 2d), significantly lower median *prop. 0* than arrays in *Plasmodium* ASSA (Fig. 2b), and significantly higher median *prop. >0.25* than arrays in *Plasmodium* ASSA (Fig. 2e). Individual comparisons revealed fewer significant differences between *Plasmodium* ASSA and other *Plasmodium* proteins. The latter had significantly lower median number of *repeat untis* (Fig. 2a) and lower median *residues per unit* (Fig. 2b) than *Plasmodium* ASSA.

Amino Acid and Nucleotide Composition

Verra and Hughes (1999) reported that the amino acid composition of the repeat arrays of *Plasmodium* antigens differed from nonrepeat regions of *Plasmodium* proteins, even after correcting for nucleotide content bias of the genes. The most striking difference between the repeats and other protein regions was an avoidance of hydrophobic residues in the former (Verra and Hughes 1999). Consistent with those findings, repeat arrays in *Plasmodium* ASSA showed significantly lower median *prop. hydrophobic* than arrays in either *Plasmodium* non-ASSA or non-*Plasmodium* proteins (Fig. 3a). *Plasmodium* ASSA showed a median *prop. hydrophilic* that was significantly higher than that of non-*Plasmodium* repeat arrays but was not significantly different from that of *Plasmodium* non-ASSA (Fig. 3b).

The genomes of *P. falciparum* and several other *Plasmodium* species are extraordinarily AT-rich (Weber 1988), although the sequence of a genomic region of *P. vivax* showed a considerably lower percentage A + T than the homologous region of *P. falciparum* (Tchavtchitch et al. 2001). In the present data, the median pAT_1 in repeat arrays in *Plasmodium* ASSA was actually significantly lower than that in *Plasmodium* non-ASSA or non-*Plasmodium* proteins (Fig. 3c). The median pAT_2 in repeat arrays in *Plasmodium* ASSA was significantly lower than that in *Plasmodium* non-ASSA, although not significantly different from that non-*Plasmodium* proteins

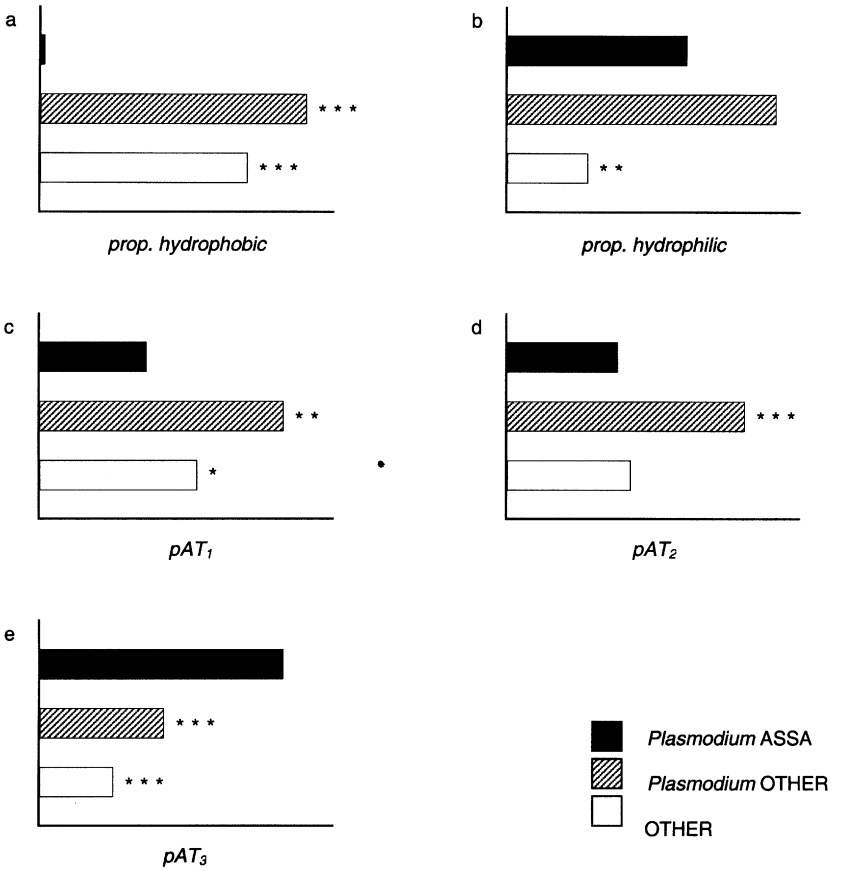

Fig. 3. Medians of variables describing repeat arrays in *Plasmodium* asexual-stage antigens (ASSA), other *Plasmodium* proteins, and other proteins, Kruskal–Wallis test, $p < 0.001$ for all variables. Tests of the equality of individual group medians with that of *Plasmodium* ASSA: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

(Fig. 3d). On the other hand, repeat arrays in *Plasmodium* ASSA had significantly higher median pAT_3 than either *Plasmodium* non-ASSA or non-*Plasmodium* proteins (Fig. 3e).

Discriminant Analysis

Linear discriminant analysis was used to examine the extent to which the variables measuring properties of the repeat array mismatch distribution, amino acid content, and nucleotide content provided sufficient information to distinguish *Plasmodium* repeat array types from those of other organisms and to distinguish *Plasmodium* ASSA from other *Plasmodium* proteins. Discriminant functions based on the variables summarized in Figs. 2 and 3 were constructed in order to classify repeat array types from *Plasmodium* versus those from other organisms. These functions were able to correctly classify repeat array types as belonging to *Plasmodium* or to other organisms 81.1% of the time (Table 2). The percentage of cases categorized correctly was uniformly high in most categories (Table 2). The lowest percentage correct was seen in *Dictyostelum* (54.2%) (Table 2). Misclassifications of *Dictyostelum* repeat array types probably occurred mainly because *Dictyostelium* repeats, like those of *Plasmodium*, had a relatively high A + T content, particularly at third positions.

An additional discriminant analysis was applied to repeat array types from *Plasmodium* ($N = 88$), with the goal of distinguishing ASSA, including CSP, from other *Plasmodium* proteins. As in previous analyses, the sexual-stage surface proteins were not grouped with ASSA. The overall rate of successful classification was 92.0% (Table 2). The category with the lowest success rate was sexual-stage surface proteins (75.0%) (Table 2).

Discussion

The present study used an approach based on the mismatch distribution (Sherry et al. 1994) to examine the extent to which repeat arrays in proteins have evolved in a concerted fashion by recent "expansion," i.e., recent duplication of repeat units. Repeat arrays from the malaria parasites of the genus *Plasmodium* were found to be characterized by a particularly high level of concerted evolution. The evidence for such concerted evolution was that, in comparison with repeat arrays from other organisms, the median proportion of nucleotide difference among repeat units was significantly reduced in

Table 2. Rates of successful classification by discriminant functions

Data set	Percentage successfully classified	
	Plasmodium vs. others	*Plasmodium* ASSA vs. other *Plasmodium* proteins
Plasmodium		
CSP	89.3	96.4
Other antigens	95.0	85.0
Nonantigens	83.3	100.0
Sexual-stage surface proteins	75.0	75.0
Sexual-stage nonsurface proteins	78.6	92.9
Parasitic protists (antigens)	76.9	—
Dictyostelium	54.2	—
Other protists (antigens)	38.5	—
Other nonparasitic protists	80.0	—
Bordatella		
Pertactin 5-residue	100.0	—
Pertactin 5-residue	100.0	—
Mycobacterium		
PPE 10-residue	100.0	—
PPE 25-residue	100.0	—
Other bacteria		
Antigens	71.4	—
Nonantigens	76.5	—
Viruses	75.0	—
Fungi	68.2	—
Plants	60.9	—
Invertebrates	71.4	—
Vertebrates	96.7	—
All	81.1	92.0

Plasmodium, and a significantly greater proportion of repeat units was identical at the DNA level (Fig. 2). As a result, the mismatch distribution showed a significantly greater positive skew in *Plasmodium* than in other organisms (Fig. 2d). The presence of a high proportion of identical or nearly identical repeat units is evidence of recent expansion of the repeat array by duplication of repeat units. Thus, our results show that such expansion occurs at an exceptionally high level in repeat arrays in *Plasmodium* proteins in comparison to repeat arrays in proteins of other organisms.

A pattern suggestive of recent array expansion was seen not only in the case of *Plasmodium* repeat arrays but also in the case of certain other repeat arrays. Interestingly, these also were arrays in proteins that may interact with the immune system of vertebrate hosts, namely, the repeat arrays of protist parasites other than *Plasmodium*, viral repeat arrays, and the three-residue repeat arrays of *Bordatella* pertactin (Table 1). Viral repeat arrays and the three-residue repeat arrays of *Bordatella* pertactin also shared with *Plasmodium* ASSA repeats a very low proportion of hydrophobic residues, which may be an adaptation for eliciting an immune response from the vertebrate host (Verra and Hughes 1999). On the other hand, certain bacterial proteins likely to interact with the vertebrate immune system showed no evidence of recent array expansion. These included the five-residue repeat arrays of *Bordatella* pertactin, PPE of *Mycobacterium*, and other bacterial antigens (Table 1).

In *Plasmodium* the ASSA are the proteins most exposed to the vertebrate immune system. Thus, if natural selection exerted by host immune surveillance has shaped the evolution of repeat arrays, it is expected that the effects of such selection will be most evident in asexual stage antigens. Repeats in ASSA were found to have significantly lower proportions of hydrophobic residues than other *Plasmodium* repeats (Fig. 3a). These results were consistent with earlier data on chemical properties of residues in *Plasmodium* repeats (Verra and Hughes 1999). Epitopes for antibody are known to include a substantial proportion of hydrophilic residues and a smaller proportion of hydrophobic residues (Hopp and Woods 1981; Geysen et al. 1988). Structurally, hydrophobic residues in linear antibody epitopes are often buried (Geysen et al. 1988). Thus the amino acid residue composition of repeats in ASSA is consistent with the hypothesis that these repeats are adapted to attract a T cell–independent antibody response by the vertebrate host (Kemp et al. 1987).

Repeat arrays in *Plasmodium* ASSA tended to include more repeat units than those in other *Plasmodium* proteins and those in other organisms (Fig. 2a). The increased length of ASSA arrays may also reflect their hypothesized role in attracting a host immune response.

Plasmodium ASSA were found to have an unusually low A + T content at the first and second positions of codons (Figs. 3c and d). This pattern evidently in part reflects the low frequency of hydrophobic residues in ASSA. It also reflects the frequent occurrence of ASSA repeat arrays of small residues (such as A, G, P, and S) that are neither strongly hydrophilic nor strongly hydrophobic. In addition, *Plasmodium* ASSA were distinguished by an unusually high A + T content at third codon positions (Fig. 3e); the latter difference is not easily accounted for by constraints at the amino acid level, since third position changes are largely synonymous.

Overall, there was no statistically significant evidence that ASSA are subject to a greater degree of concerted evolution than are other *Plasmodium* proteins (Fig. 2). Rather, the results suggest that concerted evolution of repeated arrays, by slipped-strand mispairing or a similar mechanism, is a genomewide feature of *Plasmodium*. Nonetheless, although all re-

peat arrays in *Plasmodium* species are subject to similar levels of concerted evolution, repeat arrays in ASSA have evolved unique features with respect to length and amino acid composition that are plausibly related to a role as "smoke-screens" that function to attract a T-cell–independent immune response (Kemp et al. 1987).

Discriminant analysis showed good success in separating *Plasmodium* ASSA from other *Plasmodium* proteins on the basis of variables measuring characteristics of repeat arrays. This suggests that measurement of these variables may be used to suggest the expression patterns of *Plasmodium* repeat-containing proteins of unknown function. Interestingly, the discriminant analysis showed the lowest success rate in the case of sexual-stage surface proteins (Table 2). Since the sexual stages are produced in the vertebrate host and are present in the bloodstream briefly, it may be that their surface proteins are subject to some of the same selective pressures as ASSA, resulting in a set of characteristics intermediate between those of ASSA and other *Plasmodium* proteins.

Acknowledgments. This research was supported by Grant GM043940 from the National Institutes of Health. I am grateful to Federica Verra for comments on the manuscript.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bickle Q, Anders RF, Day K, Coppel RL (1993) The S-antigen of *Plasmodium falciparum*: Repertoire and origin of diversity. Mol Biochem Parasitol 61:189–196

Delhomme B, Djian P (2000) Expansion of mouse involucrin by intra-allelic repeat addition. Gene 252:195–207

Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6:241–252

Geysen HM, Mason TJ, Rodda SJ (1988) Cognitive features of continuous antigenic determinants. J Mol Recognit 1:32–41

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley, New York

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828

Hughes AL (1999) Concerted evolution of exons and introns in the MHC-linked tenascin-X gene of mammals. Mol Biol Evol 16:1558–1567

Hughes AL (2000) Modes of evolution in the protease and kringle domains of the plasminogen-prothrombin family. Mol Phyl Evol 14:469–478

Jongwutiwes S, Tanabe K, Hughes MK, Kanbara H, Hughes AL (1994) Allelic variation in the circumsporozoite protein of *Plasmodium falciparum* from Thai filed isolates. Am J Trop Med Hyg 51:659–668

Kemp DJ, Coppel RL, Anders RF (1987) Repetitive genes and proteins of malaria. Annu Rev Microbiol 41:181–208

Nussenzweig RS, Nussenzweig V (1984) Development of sporozoite vaccines. Philos Trans R Soc Lond B 307:117–128

Reeder JC, Brown GV (1996) Allelic variation and immune evasion in *Plasmodium falciparum* malaria. Immunol Cell Biol 74:546–554

Rich SM, Ayala F (2000) Population structure and recent evolution of *Plasmodium falciparum*. Proc Natl Acad Sci USA 97:6994–2001

Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, Stoneking M (1994) Mismatch distributions of human mtDNA reveal recent human population expansions. Hum Biol 66:761–775

Tchavtchitch M, Fischer K, Huestis R, Saul A (2001) The sequence of a 200 kb portion of a *Plasmodium vivax* chromosome reveals a high degree of conservation with *Plasmodium falciparum* chromosome 3. Mol Biochem Parasitol 118:211–222

Thomas GH, Newbern EC, Korte CC, Bales MA, Muse SV, Clark AG, Kiehart DP (1997) Intragenic duplication and divergence in the spectrin superfamily of proteins. Mol Biol Evol 14:1285–1295

Van Rheede T, Smolenaars MMW, Madsden O, de Jong WW (2003) Molecular evolution of the mammalian prion protein. Mol Biol Evol 20:111–121

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

Weber JL (1988) Molecular biology of malaria parasites. Exp Parasitol 66:143–170

Wickstead B, Ersfeld K, Gull K (2003) Repetitive elements in genomes of parasitic protozoa. Microbiol Mol Biol Rev 76:360–375

Zavala F, Cochrane AH, Nardin EH, Nussenzweig RS, Nussenzweig V (1983) Circumsporozoite protein of malaria parasites contains a simple immunodominant region with two or more identical epitopes. J Exp Med 157:1947–1957