Line space 1.5 in text and 1 in references for reading, final one should be in double.

Abundance of Intrinsically Unstructured Proteins (IUPs) in *P. falciparum* and other Apicomplexan Parasite Proteomes

Zhi-Ping Feng[a*], Xiuzhen Zhang[b], Pengfei Han[b], Neeraj Arora[b],
Robin F. Anders[c], and Raymond S. Norton[a]

[a] *The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria, 3050, Australia*

[b] *School of Computer Science and IT, RMIT University, Melbourne, 3001, Australia*

[c] *Cooperative Research Centre for Vaccine Technology, Department of Biochemistry, La Trobe University, Bundoora, Victoria, 3083, Australia*

Keywords: malaria; intrinsically unstructured proteins; predictors; *Plasmodium falciparum*; repeats

Abbreviations

IUPs, Intrinsically Unstructured Proteins; RADAR, Rapid Automatic Detection and Alignment of Repeats in protein sequences; EDR, extremely long disordered regions of ≥100 consecutive residues; GO, Gene Ontology; DBL domain, Duffy-binding-like domain.

* Corresponding author: Phone: +61 3 9345 2695, Fax: +61 3 9347 0852
E-mail: feng@wehi.edu.au. (Z.-P. Feng)

Abstract

Preliminary sequence analysis of *P. falciparum* has shown that the proteome of this organism is enriched in intrinsically unstructured proteins (IUPs), which are either completely disordered or contain large disordered regions. IUPs have been characterized as a unique class of proteins that plays an important role in biology and disease. In this study, the IUP contents in some proteomes of apicomplexian parasites, especially the proteome of *P. falciparum* and its various life cycles were evaluated with DisEMBL-1.4. Compared with other proteomes, apicomplexian species are extremely abundant in proteins containing long disordered regions, and the IUP contents in mammalian *Plasmodium* species are higher than in most apicomplexian parasite species. Furthermore, the sporozoite proteome of *P. falciparum* appeared to be distinct from the other stages in having an even higher content of disordered proteins. The extremely long disordered regions with over 100 consecutive residues in *P. falciparum* proteins were found to be correlated with repetitive sequences. The presence of repeats in IUPs and the abundance of homo-peptide repeats in this species are presumably related to their faster evolution and the complexity of the organism. As IUPs have a greater potential to interact with multiple partners, the structural plasticity of IUPs, especially poly-Asn containing IUPs, in *P. falciparum* presumably plays an important role in contributing to immune evasion, host-parasite interactions and the mortality of malaria.

1. Introduction

Malaria caused by infection with the apicomplexian parasite *Plasmodium falciparum* causes more than one million deaths among children under the age of five each year [1]. The complete sequencing of the *P. falciparum* genome and a number of proteomic studies have resulted in a dramatic increase in knowledge of the protein antigens of this pathogen. This information will hopefully lead to the development of a highly effective vaccine or new drugs that will help limit the impact of malaria. The genomes of malaria parasites are unusual in having a high percentage of genes that lack apparent orthologues in other organisms. Thus, it has not been possible to ascribe structure and/or function to a large number of *Plasmodium* proteins from knowledge of homologous proteins. Although NMR and crystallographic approaches have both provided insights into the structure of some malaria vaccine candidates, structure determination is a time-consuming process and there is little prospect that these methods will yield structural information on the majority of novel *Plasmodium* proteins in the near future.

In recent years it has been realized that many proteins contain large segments that lack ordered structure under physiological conditions and in some cases the entire protein may be disordered. Many functions have been ascribed to these intrinsically unstructured proteins (IUPs), which have been reported to be more common in higher organisms [2-3]. IUPs are often involved in key biological processes such as transcriptional and translational regulation, membrane fusion and transport, cell-signal transduction, protein phosphorylation, the storage of small molecules and the regulation of self-assembly of large multiprotein complexes [4-9]. The disordered state in IUPs creates larger intermolecular interfaces [7], which increases the speed of interaction with potential binding partners even in the absence of tight binding, and provides flexibility in binding diverse ligands [8,10].

Disordered proteins contain extended segments of low sequence complexity and biased composition, with a preference for highly charged residues and a relative lack of hydrophobic residues [6,11]. These observations imply that IUPs are predictable, and at least 15 online services have been established to provide such predictions (see details from Table S1 of *Supplementary Materials*). However, the overall accuracies of these predictors are hard to compare because they rely on specific algorithms and particular parameter sets, which generally depend on the dataset from which they were derived.

DisEMBL [12] is one service widely used for predicting IUPs. It can be freely downloaded and installed locally for large-scale sequence analysis. DisEMBL comprises three different predictors trained on three definitions of disordered chain. Instead of a two-state classification of order and disorder, it partitions residues into three flexibility groups. The Loops/coils predictor defines residues assigned as α helix (H), 3_{10}-helix (G) or strand (E) as ordered and all other states (T, S, B, I, and L) as loops (also known as coils) by DSSP [13]. The training set for the Hot-loops predictor constitutes a refined subset of the above extracted by DSSPcont [14], namely those loops with a high degree of mobility as determined from Cα temperature factors (B factors). REMARK465 predictor defines residues as disordered if their coordinates are missing from X-ray structure entries in Protein Data Bank (PDB) [15].

In this paper, we have used DisEMBL-1.4 to analyze the IUP content in 28 eukaryota, 20 archaea, 22 bacteria, and the proteins expressed at four stages of the life-cycle of *P. falciparum*. Although IUPs encoded by genes on chromosomes 2 and 3 of *P. falciparum* were analyzed before the whole genome was completed [2], systematic analysis of the whole malarial proteome, comparison with other proteomes, and evaluation of IUPs in the proteomes of various stages of the *P. falciparum* life-cycle have not been undertaken. Our analyses based on 28 eukaryota and 42 prokaryota show that apicomplexan species are unusual among eukaryotes in their contents of proteins containing longer disordered regions. Furthermore, IUPs are more abundant in *Plasmodium* species than in most other apicomplexian parasites, with the possible exception of *Toxoplasma gondii*. Therefore we present more detailed analysis on IUPs in *Plasmodium* species, especially in *P. falciparum*, which include general properties of their IUPs and their possible biological implications in this fatal malaria species.

2. Material and Methods

2.1. Proteins of whole genomes

The protein sequences of six species of *Plasmodium* were extracted from PlasmoDB (http://plasmodb.org/). All the small fragments or peptides containing residues fewer than 70, as well as those sequences that did not start with methionine were then removed in ordered to avoid those with poor annotations and to focus on longer disordered regions in proteins. The protein sequences of *P. vivax* were also obtained from TIGR (http://www.tigr.org/). Seven sets of proteins with the total numbers and average lengths listed in Table 1 were produced. The length distributions of the sequences in these seven sets of *Plasmodium* species are shown in Fig. S1 of *Supplementary Materials*. The proteomes of the complete genomes of 20 eukaryota (including 5252 *P. falciparum* proteins and 7756 *P. yoelii* proteins without removing small fragments), 20 archaea and 22 bacteria were obtained from the Integr8 (release 22) at http://www.ebi.ac.uk/genomes. The proteome of *Toxiplasma gondii* was based on GlimmerHMM predictions from ToxoDB (http://www.toxodb.org/) (Release 3.0). The proteomes of the apicomplexan parasites *Cryptosporidium parvum* and *Cryptosporidium hominis* were extracted from CryptoDB (http://www.cryptodb.org/). The apicomplexan proteome of *Theileria parvum* was extracted from TIGR (http://www.tigr.org/).

Table 1 here

2.2. P. falciparum *proteins expressed in different stages of the parasite life-cycle*

1831 protein sequences for which there was evidence of expression in a specific stage of the *P. falciparum* life-cycle were also extracted from PlasmoDB (http://plasmodb.org/). In all cases, evidence of expression had been obtained by mass spectrometry. Using keyword searches, we extracted 409 proteins expressed in the gametocyte (sexual stages), 323 in the merozoite, 263 in the sporozoite, and 417 in the sexual stage of trophozoite. The proteomes expressed in these four stages were described in 2002 by Florens et al. [16] and Lasonder et al. [17].

2.3. The use of online predictors

Initially we used the following different predictors for trial and comparison: DISOPRED2 [18], FoldIndex [19], GlobPlot 2.1 [20], IUPred [21], PreLink [22] , RONN [23], DRIPPRED [24] , NORSp [25] and the three predictors from two versions of DisEMBL [12]. All online predictions and those based on the DisEMBL pipeline package were performed using their default settings. Since the three DisEMBL predictors are based on the same algorithm but different training sets, the comparison of outputs from these different predictors may extend our knowledge of protein disordered regions and possibly facilitate the development of more effective algorithms. Although the numbers of disordered residues are probably under-estimated by REMARK465 and over-estimated by Coils, their predictive results, especially with Hot-Loops and REMARK465 of DisEMBL-1.4 can nonetheless provide useful information on IUP content of *P. falciparum* compared with other species.

2.4 Repeat sequence search

Homo-peptide repeat sequences were searched for with in-house software, which is available on request. We counted all homo-peptides longer than 6 residues, a cut-off chosen because of its significantly low probability (≤1%) of occurrence by chance [26]. Complicated repeats were identified and aligned with RADAR (Rapid Automatic Detection and Alignment of Repeats in protein sequences) [27] from http://www.ebi.ac.uk/Radar/.

3. Results

3.1. IUPs in proteomes of completed genomes

Figures 1 to 3 here

The average proportions of IUPs that contain disordered regions of various lengths and the proportions of disordered residues within these regions predicted by Hot-Loops are shown in Figs. 1-3 for the proteomes of 63 organisms. The proportion of IUPs was defined as the ratio of the number of proteins that contain at least one disordered region of a given length to the total number of proteins investigated. The proportion of IUPs provides no information on sequence length or whether two or more disordered regions appeared in one sequence. Hence we calculated the proportion of disordered residues, defined as the ratio of the number of residues within a given length of disordered regions to the total number of residues investigated. For eukaryotic proteomes, the proportions of IUPs and the proportions of disordered residues are more diverse in unicellular eukaryotic species than in higher organisms (Fig. 1). They are much lower in *Guillardia theta* and *Paramecium tetraurelia*, but much higher in apicomplexian parasite species. The *T. gondii* and *P. falciparum* proteomes are the most abundant in the proportions of IUPs (Fig. 1(A)) and disordered residues (Fig. 1(B)). Although the proportion of IUPs in *P. yoelii* is just above the average in Fig. 1(A), the proportion of disordered residues shown in Fig. 1(B) is very close to that of *P. falciparum.* For 20 archaea proteomes, the organism most abundant in IUPs is *Methanopyrus kandleri* (Fig. 2), which can grow at 80-110^{o}C. Similarly, the organism with the highest IUP content among the 22 bacteria is *Aquifex aeolicus* (Fig. 3), which is also a hyperthermophilic organism that can grow at 95 ^{o}C.

The proportions of IUPs and disordered residues are usually much higher in the proteomes of eukaryota than bacteria, with archaea lying between these two. The average content of IUPs containing long disordered regions (≥ 40 consecutive residues) [2] and the average content of disordered residues within these regions are 34.2 and 7.1% for 21 eukaryota (Fig. 1(A) and (B)), 16.5 and 3.5% for 20 archaea ((Fig. 2(A) and (B)), and 11.2 and 2.1% for 22 bacteria (Fig. 3(A) and (B)). As shown in Fig. 4, apicomplexian parasite species form a unique class of organisms with respect to their protein structural properties; they have the highest average content of IUPs with long disordered regions and the highest standard deviation compared with the proteomes of fungi, higher organisms, archaea and bacteria.

Figure 4 here

The observation of higher IUP content in eukaryota, lower IUP content in prokaryota, and higher IUP content in *P. falciparum* are qualitatively consistent with previous analyses for a few species using other predictors [2,3]. As expected from our initial trial and comparison, the contents of IUPs and disordered residues predicted by REMARK465 (Table S2 of *Supplementary Materials*) are much lower than those predicted by Hot-Loops, but the differences vary markedly among the different species, which suggests there are species-specific differences in the properties of IUPs.

3.2 IUPs in various Plasmodium *species*

Figure 5 here

In 2000 Dunker *et al*. [2] estimated that about 35% of the proteins encoded by genes on *P. falciparum* chromosomes 2 and 3 (422 proteins) contained long disordered regions. In 2003 they indicated that this was an under-estimate, and was more likely to be 52-67% [28]. In the analyses carried out here long disordered regions were predicted by Hot-Loops and REMARK465 to be 48.7% and 10.8%, respectively, of the 5188 proteins encoded by the *P. falciparum* genome. The proportions of IUPs with disordered regions of various lengths and the proportions of disordered residues within these regions predicted by Hot-Loops for the proteomes of the six species of *Plasmodium* parasites are summarized in Fig 5 (A) and (B). The proportions of IUPs in *P. knowlesi*, *P. vivax* from TIGR and *P. falciparum* were predicted to be higher than in other *Plasmodium* proteomes (Fig 5 (A)). The numbers in *P. vivax* from PlasmoDB were obviously lower because of the larger number of small fragments in this sequence set. Fragments of fewer than 100 residues account for 26% in *P. vivax* from PlasmoDB, in contrast to 3% in *P. falciparum* and *P. vivax* from TIGR sequences. Furthermore, Fig. 5(B) shows that the proportions of disordered residues in the two *P. vivax* proteomes are very close, and higher than in *P. falciparum.* The proportions of disordered residues in *P. knowlesi*, *P. vivax* and *P. falciparum* are higher than in other *Plasmodium* species. The enrichment of IUP in *P. falciparum* and the relatively lower content of disordered residues are a consequence of many long protein sequences; 10.6% of the *P. falciparum* proteins, 8.2% of the *P. knowlesi* proteins, and 10.9% of the *P. vivax* (TIGR) proteins contain EDRs, in contrast to the 6.6% found in *P. yoelii*, 3.6% in *P. berghei* and 3.3% in *P. chabaudi.* IUPs in mammalian *Plasmodium* species are obviously more abundant than in rodent *Plasmodium* species.

The content of EDR-containing proteins was estimated to be much smaller by REMARK465 (1.9 % for *P. falciparum,* 1.7 % for *P. knowlesi*, 4.6 % for *P. vivax* (TIGR*)*, 1.4% for *P. vivax* (PlasmoDB), 1.2 % for *P. yoelii*, 0.7 % for *P. chabaudi* and 0.5 % *P. berghei*), but mammalian *Plasmodium* proteins are still more abundant in EDRs content than rodent *Plasmodium* proteins according to this predictor.

3.3. IUPs in different stages of the P. falciparum *life-cycle*

Figure 6 here

An analysis using Hot-Loops of the proteomes of four different *P. falciparum* life-cycle stages shows that sporozoite proteins are the most enriched in IUPs (Fig. 6(A) and (B)). As with the difference between *P. falciparum* and the other *Plasmodium* species, the finding of a higher IUP content in the sporozoite proteins reflects the existence of many longer protein sequences in this life-cycle stage; there are an average of 1775 residues in sporozoite proteins compared with 723, 726 and 1009 residues in merozoite, trophozoite and gametocyte proteins, respectively. In addition, the contents of EDR-containing proteins in the sporozoite (22.8%), trophozoite (16.1%) and gametocyte proteomes (13.9%) were over-represented compared with 10.6% in the whole *P. falciparum* proteome, whereas they were under-represented in the merozoite proteome (9.9%).

3.4. Top IUPs in P. falciparum *proteome estimated by REMARK465*

Tables 2 to 4 here

As REMARK465 gives the most conservative estimate of IUP content among the predictors tested, the top *P. falciparum* IUPs identified by this predictor were investigated further. Of the 5252 *P. falciparum* sequences, 101 were predicted by REMARK465 to contain EDRs. All except one of these EDRs (residues 51-279 in the hypothetical protein PFD0080c) were confirmed by the FoldIndex predictor [19], a method based on the average residue hydrophobicity and net charge of the sequence [5], with 94% of the EDRs in these 101 IUPs being predicted as the highest scoring disordered regions. Among these 101 proteins, 39 have detailed annotations in the PlasmoDB or Swiss-Prot/TrEMBL databases (Table 2), but the other 62 are listed as hypothetical proteins (although many have apparent orthologues in other organisms). The structural features, stage of expression and Gene Ontology (GO) assignments of these 101 IUPs are summarized in Tables 3 and 4. Further details are given in Table S3 of *Supplementary Materials*.

Trans-membrane proteins account for about 20% of these 101 IUPs, which is close to the content of membrane proteins found in the Swiss-Prot/TrEMBL databases (22%), and the number of single trans-membrane proteins is over double the number of multi-trans-membrane proteins. In contrast to the analysis of IUPs above, the expression of these 101 top IUPs is not associated with one particular life-cycle stage. Among these 101 IUPs there are 25% orphan proteins [29] (proteins without orthologues in *Plasmodium* species), 62% proteins without orthologues in apicomplexian species and 54% proteins without orthologues in other eukaryotic species.

The GO assignments indicate that 71% of the 39 IUPs with assigned molecular functions are involved in molecular interactions (protein/protein; protein/DNA and protein/nucleotide). Of the 29 proteins with an assignment to a cellular component, 50% are membrane associated. The biological processes of the known top IUPs involved are quite diverse but the most common is transcriptional and translational initiation, which accounts for about 18% of the 39 proteins with assigned processes. In addition to some well-known biological processes of IUPs involved, about 5% of the 39 proteins with assigned processes are involved in cell invasion or cell communication, and 3% are involved in evasion of host immune response.

3.5 Repeat sequences in IUPs

In order to identify further sequence features of the disordered regions, especially EDRs, we performed a repetitive sequence search in the EDRs within the top 101 IUPs, and found that all of them contain repeats. The details of these repeats and alignments according to RADAR [27] are given in Table S4 of *Supplementary Materials*. The amino acid compositions of the 101 proteins, the repeats in EDRs and the whole proteome of 5188 proteins are summarised in Fig. 7. Obviously, the residues in the 101 proteins are biased towards those with higher flexibility indices [30], and in the repeats this bias is even stronger. For example, the frequencies of E, D, S and R in EDRs are 11, 7, 4 and 3% higher, respectively, than in the whole proteome. However, the frequency of N in EDR is unexpectedly 3.3% lower.

Figure 7 here

We also searched for homo-peptide repeats or single amino acid repeats longer than six residues in six *Plasmodium* species and in four stages of the *P. falciparum* life-cycle. The frequencies of these repeats are shown in Fig. 8 (A) and (B), respectively. It is clear that the proteins in *P. falciparum* have different features compared with other *Plasmodium* species and that the homo-peptide repeats in the sporozoite have different features compared with other stages of the *P. falciparum* life cycle. The extreme example is poly-N repeats, which are about 68% more abundant in *P. falciparum* than in other species of *Plasmodium*, and the most enriched in the sporozoite stage. These poly-N repeats have the lowest complexity and are predicted to be disordered in structure. However, the repeats in EDRs of the top 101 IUPs are relatively deficient in Asn, suggesting that Asn prefers to follow itself, rather than participates in more complex repeats. Similarly, the frequency of poly-F is higher in *P. knowlesi* (Fig. 8(A)), although the overall content of Phe in this species is about 2% lower than in *P. falciparum*.

Figure 8 here

4. Discussion

We analysed IUPs in the proteomes of 28 eukaryota and 42 prokaryota, as well as proteins expressed in various stages of the *P. falciparum* life-cycle. The results show that apicomplexian species form a unique protein structural family compared with other eukaryotic species. The proteomes of mammalian *Plasmodium* are the most abundant in IUPs among the apicomplexian species analysed (with an exception of *T. gondii*), and proteins expressed in the sporozoite stage of the life-cycle of *P. falciparum* are distinct from those in other stages. IUPs in *P. falciparum* are also rich in repetitive sequences.

4.1. Abundance of IUPs in apicomplexian parasitological proteomes

Our analyses show that apicomplexian species are unusual in the abundance of IUPs encoded in their genomes. As expected from the results of previous studies, IUPs are more common in apicomplexian species than in bacteria and archaea. IUPs are also more common in apicomplexian species than in other unicellular eukaryotes and in higher organisms.

Apicomplexan species have a common apical complex and common organelles associated with host cell attachment and invasion, as well as establishment of an intracellular "parasitophorous vacuole". The abundance of IUPs in this group of organisms suggests that IUPs may play a role in parasite attachment and/or host cell invasion and survival. The four apicomplexian organisms with the most abundant IUPs are *T. gondii*, *P. vivax, P. knowlesi and P. falciparum. T. gondii* is one of the most common parasitic diseases of animals and humans. It has very low host specificity, and probably infects almost any mammal. *P. vivax* is the most widely distributed human malaria parasite, although seldom fatal. *P. knowlesi* is a primate malaria parasite that causes monkey malaria but may also be acquired by humans either naturally or artificially. *P. falciparum* accounts for 80% of all human malarial infections and 90% of deaths. More systematic studies are clearly needed for each of these organisms in order to gain a better understanding of the general roles of IUPs, but for the detailed analyses carried out in this study we have focused on *Plasmodium* species because of the wealth of genome and proteome data available from different species of *Plasmodium* and from the different stages of the *P. falciparum* life-cycle.

4.2. Abundance of IUPs in P. falciparum *proteome and unusual distribution in the sporozoite*

The proportion of *P. falciparum* proteins predicted to be IUPs (proteins containing at least one disordered region 40 residues or more in length) by the Hot-Loops predictor of DisEMBL (48.7%) was consistent with a previous prediction of 52 – 67% by PONDR for proteins encoded by genes on chromosomes 2 and 3. However, these predictions for IUPs were both lower than the 88.2% and 94% of ORFs on chromosomes 2 and 3, respectively, predicted to contain low complexity regions by the SEG program [31-33]. This is not surprising given that some of the low complexity sequences identified by SEG are relatively hydrophobic or are involved in ordered but non-globular structures such as coiled coils.

IUPs were similarly abundant in *P. vivax*, the other human parasite studied, and in the simian parasite *P. knowlesi*, but less abundant in the three rodent parasites studied. In part this may reflect a higher abundance of proteins with low complexity regions and the greater average length of proteins in *P. falciparum.* Others have noted that *Plasmodium* proteins can be much longer than orthologous proteins in other species as a consequence of the insertion of stretches of sequence predicted to form non-globular structures [34-35]. However, this does not explain the differences between, for example, *P. knowlesi* and *P. yoelii* where the average lengths of the proteins analysed were very similar, at 494 and 518 residues, respectively.

The first of the *Plasmodium* genomes to be sequenced was that of *P. falciparum* and a striking finding was the large number of genes (around two-thirds) with no apparent orthologous gene in other organisms. Furthermore, for 736 of 5268 genes identified in the *P. falciparum* genome, no orthologue was identified in the genomes of the three rodent parasites, *P. yoelii, P. berghei* and *P. chabaudi* [29]. Many of the proteins encoded by genes that are unique to *P. falciparum* or other *Plasmodium* species are relatively large and contain extensive sequence repeats or low complexity segments that are predicted to be disordered. This bias towards IUPs amongst these ‘orphan” proteins was confirmed by analysing a set of 117 proteins unique to *P. falciparum* and for which evidence of expression was obtained by mass spectrometric analyses (see details from Table S5 and Fig. S2 of *Supplementary Material*). These proteins included many of the most unstructured proteins identified by re-analysis of the coding sequences in the *P.*

falciparum with REMARK465, the most conservative of the three DisEMBL predictors (see below).

Pandey *et al.* [36] found using the FoldIndex predictor [19] that increasing structural disorder was associated with the evolution of non-housekeeping proteins in eukaryotes. The results of our analyses of IUPs in *Plasmodium* species are consistent with this observation despite the unusual insertions of disordered sequences in many housekeeping proteins, for example, the GTPases [35] and the subtilisin-like protease-1 [37]. The genes in *Plasmodium* species that lack orthologues in other organisms (by definition non-housekeeping) tend to be located in the subtelomeric regions of the chromosomes. Hall et al [29] identified 12 distinct subtelomeric gene families to which *P. falciparum*-specific genes could be assigned and only five of these had identifiable gene families in all *Plasmodium* species with sequenced genomes. Although the function of many of the corresponding proteins is unknown it is clear that many have roles in attachment and invasion and have evolved rapidly to generate large families of highly variable proteins, many of which have been noted previously to contain extensive regions of low complexity sequence, a characteristic of IUPs [38]. The best characterized of these gene families in *P. falciparum* are the *var, stevor* and *rifin* gene families and multiple representatives of the corresponding proteins have been identified by Hot-Loops as containing significant regions of disorder.

It is important to distinguish between two types of proteins that the analyses carried out here have identified as IUPs. In one class of protein low complexity sequences characterized by stretches of Asn residues or other hydrophilic amino acids are found inserted into otherwise globular domains or as extensions to otherwise highly ordered proteins. Such sequences, which may be relatively short, are found in a large percentage of *Plasmodium* proteins, including those that have orthologues in other organisms. Although the function of these sequences is largely unknown, detailed studies of three enzymes have shown the inserts to be important for activity [37, 39-40]. MSP3 and related proteins have highly acidic C-terminal regions that are predicted to be unstructured and these may be important for providing a negative charge to a cell surface that lacks sialoglycoproteins. The merozoite surface antigen apical membrane antigen 1 (AMA1) has flexible loops protruding from the globular domains stabilized by disulphide bonds [41-44] and it has been suggested that one or more of these flexible loops may protect a hydrophobic binding cleft from antibody binding [44].

At the other end of the spectrum are proteins that have very long segments of disordered sequence that in some cases account for a large proportion of the total polypeptide chain, for example MSP2 and the S antigen. To focus on proteins of this second class the conservative predictor REMARK465 was used to identify proteins with extended regions of disorder and a set of 101 IUPs was identified, each containing a region of more than 100 residues predicted to be disordered by REMARK465. All of these proteins contain extensive sequence repeats. In a few cases the repeats are long and highly degenerate but most of these IUPs contain extensive arrays of tandemly repeated short sequences. Only 39 of these 101 proteins were identified in the annotation, with the remaining 62 proteins listed as hypothetical. Of the 39 identified proteins relatively few have recognisable orthologues in non-apicomplexian species. These 39 proteins include only a single *var* gene product (*Pf*EMP1, PF08_0140) and no *rifins* or *stevors*, which is not surprising given the presence of domains containing multiple intramolecular disulphide bonds in these proteins. The long disordered region identified in one *var* gene product is a highly acidic sequence of 119 residues linking two Duffy binding-like (DBL) domains EBL-1 and SH3 kinase. The SEG program [31] has identified sequences of low complexity in these interdomain regions in all *Pf*EMP1s (shown in PlasmoDB) but the sequence identified by REMARK465 in PF08_0140 is exceptionally long. Presumably these flexible interdomain sequences in *Pf*EMP1 facilitate the protein-protein interactions between DBL domains and host proteins that lead to cytoadherence [45].

Among these top IUPs are numerous proteins that have been studied extensively for many years because they are antigens and of interest as possible targets of protective immune responses. These include MSP2, Ag332, MESA, glutamate-rich protein and CS protein (although the CS protein was not identified by REMARK465 or Hot-Loops, but was predicted to contain EDR of more than 300 residues by most of the other predictors such as Coils of DisEMBL, FoldIndex and NORSp). The repeats in many of these molecules are highly immunogenic both when used to immunise animals or when individuals are exposed to natural infections. Indeed, some of them were among the first

Plasmodium proteins to be cloned because the strength of their reactivity with antibodies allowed them to be readily selected from recombinant protein expression libraries. The CS protein and MSP2, two top IUPs with extensive repeat sequences, are both under development as potential malaria vaccines [46, 47]. Although the B-cell epitopes encoded by the repeats in these two antigens are immunodominant it has been difficult to establish that antibodies to the repeats mediate protection against the parasite. Given the capacity of intrinsically unstructured regions of proteins to adopt more different ordered structures when interacting with different target ligands [5-11, 48] antibodies induced by these regions of a protein may recognize a variety of antigen conformers and for this reason they may react poorly if at all with the antigen *in situ* on the parasite surface.

Sporozoites are an invasive stage and possess the apical complex machinery involved in host cell invasion. The sporozoite proteome appeared markedly different from other stages of the parasite life cycle, with a higher proportion (49%) of unique proteins, while trophozoites, merozoites and gametocytes only had between 20 and 33%. The common proteins to all four stages (6%) were mostly housekeeping proteins such as ribosomal proteins, transcription factors, histones and cytoskeletal proteins. The antigenically variant surface proteins, some markers of the sporozoite stage, and many proteins that have a single trans-membrane domain associated with rhoptries, micronemes and dense granules, were largely expressed in sporozoites [17]. There is accumulating evidence for the presence of IUPs in these types of proteins, such as AMA1, CS protein, and sporozoite surface protein 2 (SSP2; also known as TRAP). The abundance of IUPs in the sporozoite stage of the malaria life cycle, especially at the sporozoite surface, may facilitate responses to external stimuli, stress and the phases of the cell cycle, as well as increasing sensitivity to proteolysis. The high proportion of IUPs in sporozoites may induce many antibodies to provide an immunological smokescreen to the parasite in order to evade the human immune system. They may also reflect stronger antigenic variation and multiple protein interactions than recognised so far.

4.3. Repeats, low complexity and disordered fragments in P. falciparum *proteomes: evolution and potential biological implication*

Marcotte *et al.* [49] found that proteins in eukaryotes have three times more repetitive sequences than those of prokaryotes, with archaea falling in between. This observation is qualitatively consistent with our observation of the distribution of IUP contents in the three kingdoms. Repeat regions are often hyper-mutable, rapidly gaining and losing residues during the course of evolution [50]. This property is also consistent with the flexible properties of IUPs, and further suggests that protein repeats often correlate with IUPs. Marcotte *et al.* also found that the kingdoms have very few repeats in common and proposed that error-prone repeat expansion allows repeat-containing proteins to evolve more quickly than non-repeat-containing proteins [49]. The correlations between repeats and IUPs suggest that the abundance of IUPs leads to the fast evolution of *P. falciparum* and contributes to a larger evolutionary distance between this species and other organisms.

The extreme abundance of low complexity regions and homo-repeats in *P. falciparum* has been noticed since the first two chromosome of *P. falciparum* were sequenced [32-34, 51-52]. The compositional analysis of hydrophilic non-repetitive low-complexity segments of *Plasmodium* reveals they are rich in acidic residues (Glu and Asp) but prefer Lys and, even more so, Asn [51]. Our homo-peptides search results show that Asn prefers to expand to longer poly-N (for example the 52-N in putative FHA domain protein, and 60-N in hypothetical protein PF13_0139), but is relative not likely to be involved in other hydrophilic repeats. The extreme asymmetry selection of Asn compared to Lys may be related to an active role of these elements in the production of immunodominant epitopes that could provide a smokescreen against the host immunogenic response [34].

It is well established that more than 40% of the homo-repeat-containing proteins in the human proteome are associated with disease [26]. In particular, poly-Q is associated with a number of neurological disordered such as Huntington's, Alzheimer's and Parkinson's diseases [53]. Singh et al [26] found that ~1300 proteins in *P. falciparum* contain prion-like domains, which accounts for 35.7% of its proteins, and is 22.5 and 27.2% higher than in *P. yoelii* and *C. parvum,* respectively. These prion-like domains are likely to participate in formation of amyloid structures due to their exceptional Asn-richess. Our homo-peptides searches show that poly-N in *P. falciparum* is about 68%

more abundant than in *P. yoelii* and 73% more abundant than in *P. knowlesi* and *P. vivax*. The high abundance of poly-N in *P. falciparum* and roughly equal abundance of IUPs in all mammalian *Plasmodium* species (or even higher in *T. gondii*) suggest that IUPs are probably responsible for disrupting host immune systems and may thus contribute to the wide distribution of associated diseases, while poly-N repeats containing IUPs correlate more closely with fatal malaria in humans.

Given the limitations of available proteomic information (such as the large difference between *P. vivax* proteins sequences from PlasmoDB and TIGR) and the variable accuracy of current IUP predictions (such as inconsistent predictions on protein repeats that are either higher complexity or have larger proportions of hydrophobic residues, and the lack of sufficient species-related information during the development of all predictors), further systematic studies on IUPs in the *P. falciparum* and related genomes are needed. However, the abundance of IUPs in mammalian *Plasmodium* species and the remarkably high abundance of poly-N homo-repeats in *P. falciparum* are noteworthy features that definitely play some important roles in the disease process. Whether these IUPs add an extra dimension to the pattern of cross-reactivities between the antigens or they lead to a significant proportion of the antibody response being directed towards irrelevant epitopes, further detailed biological work is clearly needed.

Acknowledgements

The authors thank Rune Linding and Bob MacCallum for advice on the DisEMBL and DRIP-PRED predictors, Jan Slapeta, Toby Sargeant and Terry Speed for useful discussions. Z.P. Feng is supported by an APD award from the Australian Research Council.

References

[1] Gardner MJ, Hall N, Fung E et al. Genome sequence of the human malaria parasite. *Plasmodium falciparum*. Nature 2002;419:498 - 511. spaces between numbers?

[2] Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 2000;11:161-71.

[3] Ward JJ SJ, McGuffin LJ, Buxton BF and Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337: 635-45.

[4] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573-82.

[5] Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415-427.

[6] Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527-33.

[7] Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R. Extended disordered proteins: targeting function with less scaffold. Trends Biochem Sci 2003;28:81-5.

[8] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology 2005;6:197-208.

[9] Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells 2001;6:1-12.

[10] Tompa P, Szasz C, Buday L. Structural disorder throws new light on moonlighting. Trends Biochem Sci 2005;30:484-89.

[11] Dunker AK, Lawson JD, Brown CJ et al. Intrinsically disordered protein. J Mol Graph Model 2001;19:26-59.

[12] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure (Camb) 2003;11:1453-9.

[13] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-637.

[14] Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure (Camb) 2002;10:175-84.

[15] Berman HM, Westbrook, J., Feng, Z et al. The Protein Data Bank. Nucleic Acids Research 2000;28:235-42.

[16] Florens L, Washburn MP, Raine JD et al. A proteomic view of the *Plasmodium falciparum* life cycle. Nature 2002;419:520-6.

[17] Lasonder E, Ishihama Y, Andersen JS et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature 2002;419:537-42.

[18] Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53 Suppl 6:573-8.

[19] Prilusky J, Felder CE, Zeev-Ben-Mordehai T et al. FoldIndex(C): a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 2005;21:3435-8.

[20] Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 2003;31:3701-8.

[21] Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827-39.

[22] Coeytaux K, Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 2005;21:1891-900.

[23] Thomson R, Esnouf R. Prediction of natively disordered regions in proteins using a bio-basis function neural network. Lecture Notes in Computer Science 2004;3177:108-16.

[24] MacCallum RM. Striped sheets and protein contact prediction. Bioinformatics 2004;20 Suppl 1:I224-I231.

[25] Liu J, Rost B. NORSp: Predictions of long regions without regular secondary structure. Nucleic Acids Res 2003;31:3833-5.

[26] Singh GP, Chandra BR, Bhattacharya A, Akhouri RR, Singh SK, Sharma A. Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in *Plasmodium falciparum.* Mol Biochem Parasitol 2004;137:307-19.

[27] Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. Proteins 2000;41:224-37.

[28] Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573-84.

[29] Hall N, Karras, M, Raine JD et al. A comprehensive survey of the *Plasmodium* life-cycle by genomic, transcriptomic, and proteomic analyses. Science 2005;307:82-86.

[30] Vihinen M, Torkkila E, Riikonen P, Accuracy of protein flexibility predictions. Proteins 1994;19:141-9.

[31] Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem.1993;17: 149-63.

[32] Gardner MJ, Tettelin H, Carucci DJ et al. Chromosome 2 Sequence of the Human Malaria Parasite *Plasmodium falciparum.* Science 1998;282:1126-32.

[33] Bowman S, Lawson D, Basham D et al. The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature 1999;400: 506-7.

[34] Pizzi E, Frontali C. Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res 2001;11:218-29.

[35] Aravind L, Iyer LM, Wellems TE, Miller LH. *Plasmodium* biology: genomic gleanings Cell 2003;115:771-85.

[36] Pandey N, Ganapathi M, Kumar K, Dasgupta D, Das Sutar SK, Dash D. Comparative analysis of protein unfoldedness in human housekeeping and non-housekeeping proteins. Bioinformatics 2004;20:2904-10.

[37] Jean L, Withers-Martinez C, Hackett F, Blackman MJ. Unique insertions within *Plasmodium falciparum* subtilisin-like protease-1 are crucial for enzyme maturation and activity. Mol Biochem Parasitol 2005;144:187-97.

[38] Fischer K, Chavchich M, Huestis R, Wilson DW, Kemp DJ, Saul A. Ten families of variant genes encoded in subtelomeric regions of multiple chromosomes of *Plasmodium chabaudi*, a malaria species that undergoes antigenic variation in the laboratory mouse. Mol Microbiol. 2003;48:1209-23.

[39] Sarma GN, Savvides SN, Becker K, Schirmer M, Schirmer RH, Karplus PA. Glutathione reductase of the malarial parasite *Plasmodium falciparum*: crystal structure and inhibitor development. J Mol Biol. 2003;328:893-907.

[40] Clarke JL, Sodeinde O, Mason PJ. A unique insertion in *Plasmodium berghei* glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase: evolutionary and functional studies.Mol Biochem Parasitol. 2003;127:1-8.

[41] Pizarro JC, Normand BV, Chesne-Seck ML et al. Crystal structure of the malaria vaccine candidate apical membrane antigen 1. Science 2005;308:408-11.

[42] Feng ZP, Keizer DW, Stevenson RA et al. Structure and inter-domain interactions of domain II from the blood-stage malarial protein, Apical Membrane Antigen 1. J Mol Biol 2005;350:641-56.

[43] Nair M, Hinds MG, Coley AM et al. Structure of domain III of the blood-stage malaria vaccine candidate, *Plasmodium falciparum* apical membrane antigen 1 (AMA1). J Mol Biol 2002;322:741-53.

[44] Bai T, Becker M, Gupta A et al. Structure of AMA1 from *Plasmodium falciparum* reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. Proc Natl Acad Sci U S A. 2005;102:12736-41.

[45] Singh SK, Hora R, Belrhali H, Chitnis CE, Sharma A. Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. Nature. 2006 439:741-4.

[46] Anders RF, Coppel RL, Brown GV, Kemp DJ. Antigens with repeated amino acid sequences from the asexual blood stages of *Plasmodium falciparum*. Prog Allergy. 1988;41:148-72.

[47] Genton B, Al-Yaman F, Betuela I et al. Safety and immunogenicity of a three-component blood-stage malaria vaccine (MSP1, MSP2, RESA) against *Plasmodium falciparum* in Papua New Guinean children. Vaccine. 2003 22:30-41.

[48] Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18: 343-84.

[49] Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J Mol Biol 1999;293:151-60.

[50] Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 1998;95:10774-8.

[51] Brocchieri L. Function low-complexity regions in *Plasmodium* proteins: In search of a function. Genome Res 2001;11:195-7.

[52] Xue HY, Forsdyke DR. Low-complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. Mol Biochem Parasitol 2003;128:21-32.

[53] Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations Proc Natl Acad Sci U S A. 2002;99:333-8.

Table 1: *Plasmodium* proteomes used in this study

Species	# Protein	Average length (aa)	Sequence ≥100 aa (%)
P. falciparum [a]	5188	767	97
P. vivax [a]	10737	334	74
P. vivax [b]	5393	694	97
P. yoelii [a]	5708	518	95
P. knowlesi [a]	6087	494	91
P. chabaudi [a]	4761	283	86
P. berghei [a]	4769	345	83

[a] Species extracted from PlasmoDB (http://plasmodb.org/). [b] Species extracted from TIGR ((http://www.tigr.org/). All the small fragments containing few than 70 residues, as well as those sequences not starting with methionine, were removed from the species extracted from PlasmoDB.

Table 2. Top 39 IUPs in *P. falciparum* proteome identified by REMARK465. The remaining 62 proteins are hypothetical. Further details of these 101 proteins are listed in Table S3.

ID[a]	Protein	DR(s)[b] (length in aa)
O96239	DNA helicase, putative	119-236 (1997)
Q6LFJ6	Translation initiation factor IF-2, putative	28-136 (977)
Q8I0U6	Ring-infected erythrocyte surface antigen	897-1085 (1085)
Q8I2B4	Ebl-1 like protein, putative	834-974 (1567)
Q8I2D8	*P. falciparum* RESA-like protein with DnaJ domain	265-823, 832-976 (1451)
Q8I3A0	E1-E2_ATPase/hydrolase, putative	1231-1387 (2563)
Q8I486	Interspersed repeat antigen, putative	1555-1720 (1720)
Q8I492	Mature parasite-infected erythrocyte surface antigen (MESA) or PfEMP2	204-304 (1434)
Q8I540	Cell cycle control protein, putative	32-113, 214-316 (967)
Q8I5N5	GTP-binding protein, putative	92-206 (874)
Q8I5S6	Eukaryotic translation initiation factor 3 subunit 10, putative	1042-1253 (1377)
Q8I6U6	Gene 11-1 protein	169-290, 10266-10589 (10589)
Q8IAK3	Erythrocyte membrane protein 1 (PfEMP1)	1591-1709 (2980)
Q8IAN4	DNA repair protein rad54, putative	935-1039 (1239)
Q8IC04	Heat shock protein 86 family protein	443-542, 551-666 (912)
Q8IFL9	*P. falciparum* trophozoite antigen r45-like protein	385-591, 601-759, 769-931 (1222)
Q8IHN3	Antigen 332, putative	1345-1458 (5507)
Q8IHP3	MAEBL, putative	1248-1530 (2055)
Q8IID4	Dynein heavy chain, putative	4118-4250 (5251)
Q8IIQ7	Asparagine-rich antigen	1081-1282 (1483)
Q8IJ57	S-antigen	92-574 (585)
Q8IJI0	Pre-mRNA splicing factor, putative	201-448 (538)
Q8IJ56	Glutamate-rich protein	542-709 (1233)
Q8IJX0	RhoGAP protein	453-575 (653)
Q8IKY8	Chloroquine resistance marker protein	2793-2907 (3704)
Q9TY97	Merozoite surface protein 2	43-232 (272)
Q9TY99	Knob associated histidine-rich protein	351-520 (654)
Q9U0N1	Glutamic acid-rich protein (Garp)	540-673 (673)
O96166	Cysteine protease, putative	15-116 (1096)
Q8I1T4	*P. falciparum* protein kinase, putative	1200-1339 (1339)
Q8I410	RNA polymerase I	2390-2512 (2914)
Q8I5Y3	Eukaryotic translation initiation factor 3 subunit 8, putative	174-287 (984)
Q76NN6	Ran binding protein 1 (EC 3.6.1.47)	164-280 (280)
Q7KQK4	Zinc finger transcription factor (Krox1)	620-731, 760-917 (1461)
Q8IDX7	*P. falciparum* reticulocyte binding protein 2 homolog b	537-648 (1115)
Q8IJP9	ADA2-like protein	1309-1410 (2578)
Q8IJS7	QF122 antigen	1027-1139 (1139)
Q8IE99	Splicing factor, putative	234-413 (864)
Q8IDX6	Reticulocyte binding protein 2 homolog a (Normocyte binding protein 2a)	2676-2834 (3130)

[a] UniProt ID; [b] Residue numbers from start to end of the EDRs predicted by REMARK465 of DisEMBL-1.4 [12].

Table 3. Properties of the top 101 IUPs from Table S3[*].

Features	**Number of proteins**	**Life-cycle stage**	**Number of proteins**
Signal peptide	20	Sporozoite	43
Single-trans-membrane-helix	14	Gametocyte	52
Multi-trans-membrane-helices	6	Merozoite	48
GPI–anchored	2	Ring/Schizont	36
Othologues in other *Plasmodia*	72		
Othologues in apicomplexa	36	Unknown	27
Othologues in other eukaryota	45		

[*] The information of othologues is based on the annotations from PlasmoDB.

Table 4. Summary of the GO assignments of the top 101 IUPs fromTable S3[a].

Molecular Function	Cellular Component	Biological Process
Heat shock protein binding (3)	Membrane (11)	Transcription, translational initiation (7)
Nucleotide binding (12)	Cytoplasm, nucleus (7)	Catalytic activity (1)
ATP, GTP, metal ion binding (11)	Intracellular membrane-bound organelle (4)	Protein biosynthesis, protein folding (5)
Glycosaminoglycan binding (1)	Extracellular (3)	Nucleic acid metabolism (1)
Kinase activity (2)	Unknown (72)	Cellular process (1)
Receptor activity (2)		Cell invasion, cell communication (2)
Microtubule motor activity (2)		Amino acid phosphorylation (2)
Unknown (63)		Pathogenesis, response to stimulus, evasion of host immune response (1)
		Unknow (63)

[a] Protein numbers are shown in the brackets.

Figure Legends

Fig. 1. Distributions of predicted averaged proportions of IUPs with disordered regions of various lengths (A) and proportions of disordered residues within these regions (B) predicted by Hot-loops of DisEMBL-1.4 for the proteomes of 20 completed genomes of eukaryota, and uncompleted genome of *T. gondii.* The results indicate that apicomplexan parasite proteins form a unique structural family in eukaryotic organisms.

Fig. 2. Distributions of predicted average proportions of IUPs with disordered regions of various lengths (A) and proportions of disordered residues within these regions (B) predicted by Hot-loops of DisEMBL-1.4 for the proteomes of 20 completed archaea. *M. kandleri* has the most abundant IUPs and can grow under 80~110^o C.

Fig. 3. Distributions of predicted average proportions of IUPs with disordered regions of various lengths (A) and proportions of disordered residues within these regions (B) predicted by Hot-loops of DisEMBL-1.4 for the proteomes of 22 completed bacteria. *A. aeolicus* and *T. maritima* are the two most abundant IUPs organisms and both of them can grow at 95^o C.

Fig. 4. Comparison of different groups of organisms according to the average percentage of proteins that contain disordered regions ≥ 40 residues (A) and the average percentage of disordered residues in such disordered regions (B). The apicomplexan species include: *P. falciparum*, *P. yoelii*, *T. gondii*, *C. parvum*, *C. hominis*, *T. parvum*. The high organisms include: *A. thaliana*, *B. rerio*, *R. norvegicus*, *C. elegans*, *D. melanogaster*, *M. musculus*, *H. sapiens*. The species included in other groups are indicated in Table S2.

Fig. 5. Distributions of predicted proportions of IUPs for the proteomes of six species of *Plasmodium* with disordered regions of various lengths (A) and proportions of disordered residues within these regions (B) predicted by Hot-loops of DisEMBL-1.4. IUPs are more abundant in mammalian *Plasmodium* than in rodent *Plasmodium.*

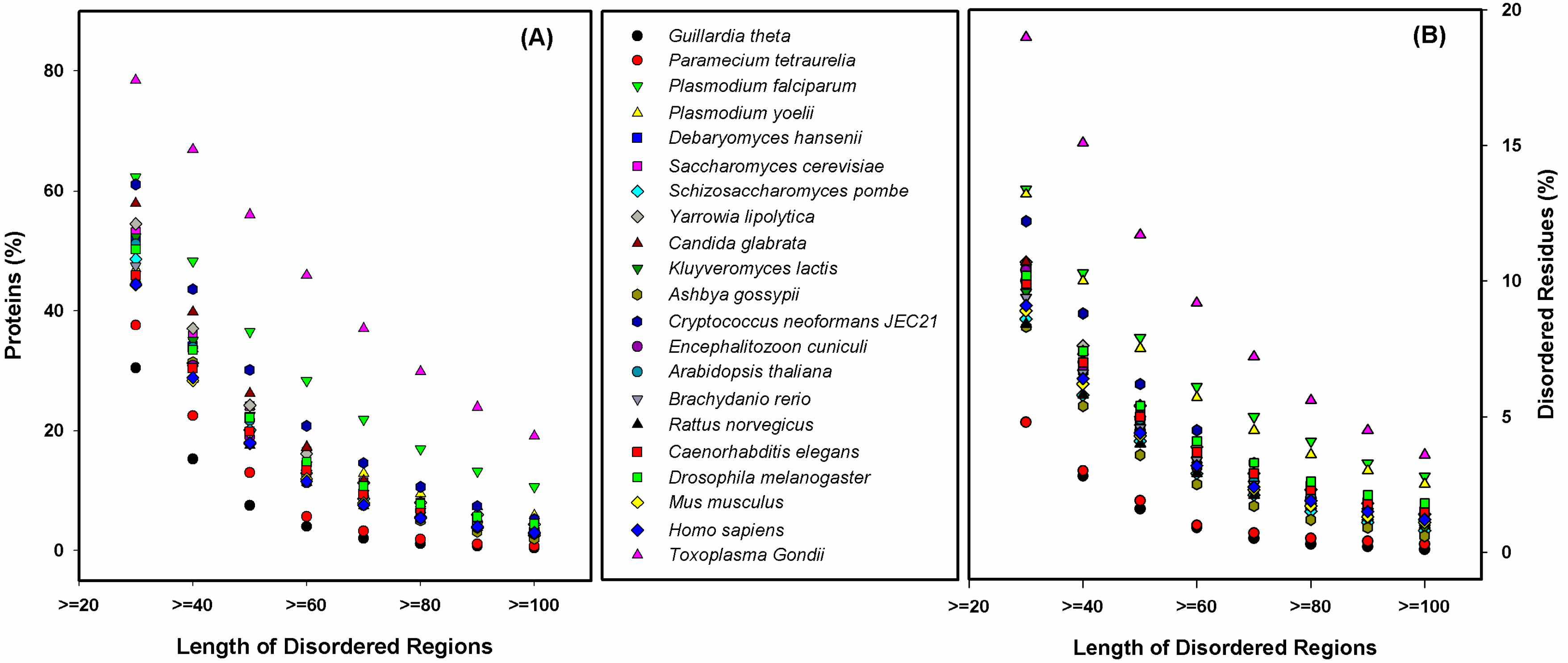
Fig. 6. Distributions of predicted proportions of IUPs for the proteomes of four stages of the *P. falciparum* life-cycle with disordered regions of various lengths (A) and proportions of disordered residues within these regions (B) predicted by Hot-loops of DisEMBL-1.4. The

results indicate that the proteins expressed on the sporozoite are abundant in long disordered regions compared with those expressed in other stages of the malaria life-cycle.

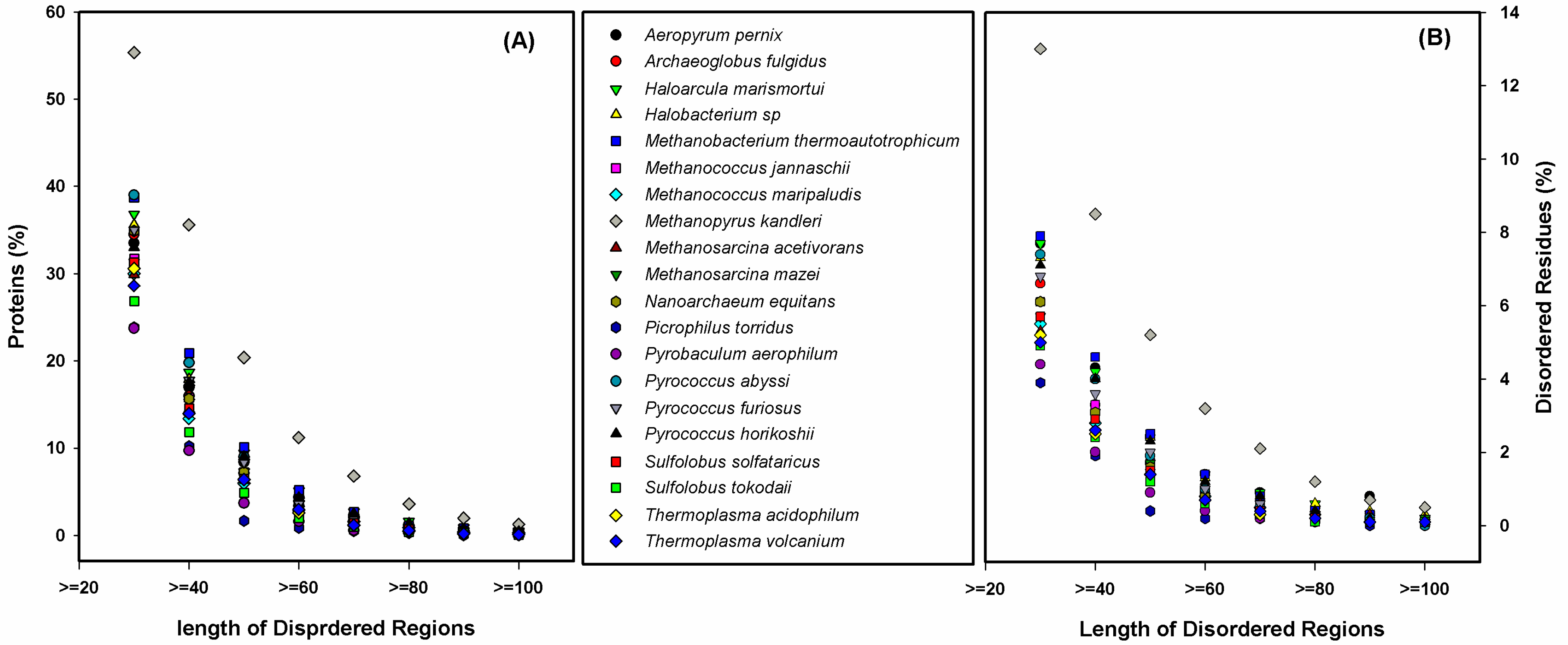
Fig. 7. Amino acid compositions of the whole *P. falciparum*, top 101 IUPs and the repeats in EDRs of the top 101 IUPs. The results show that the amino acid composition in the repeats is biased towards residues that have a higher tendency to be exposed on the protein surface [32] than in the top 101 IUPs and in the whole *P. falciparum* proteome, with the exception of Asn, which has a very high potential to appear in homo-peptide repeats.

Fig. 8. Frequencies of the 20 homo-peptide repeats in the proteomes of (A) six species of *Plasmodium* and (B) four stages of malaria life cycle. The result of (A) indicates the frequency of poly-N repeats of longer than six residues in *P. falciparum* is over 68% higher than in other species of *Plasmodium*. The proteins expressed at the sporozoite stage have higher frequencies of poly-N, poly-K than at all other stages of the malaria life cycle.

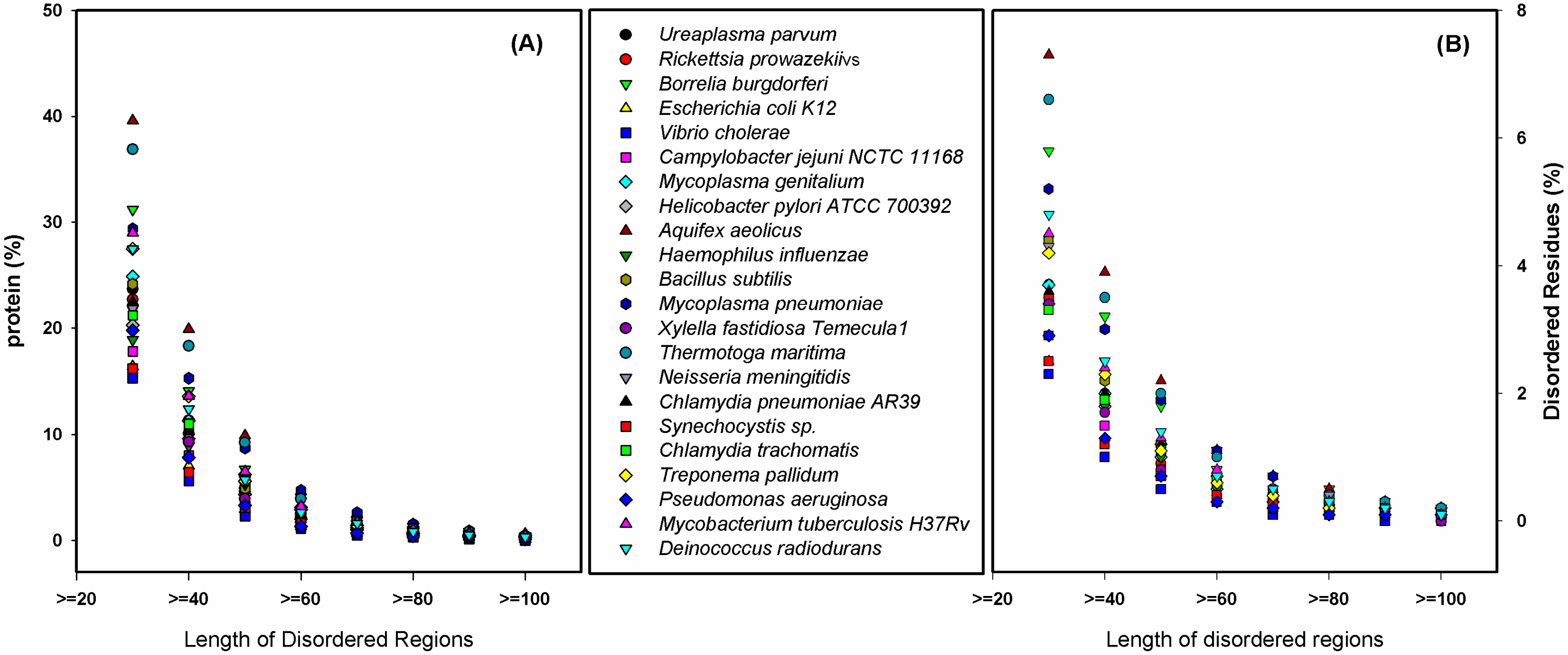
Feng et al., Fig. 1



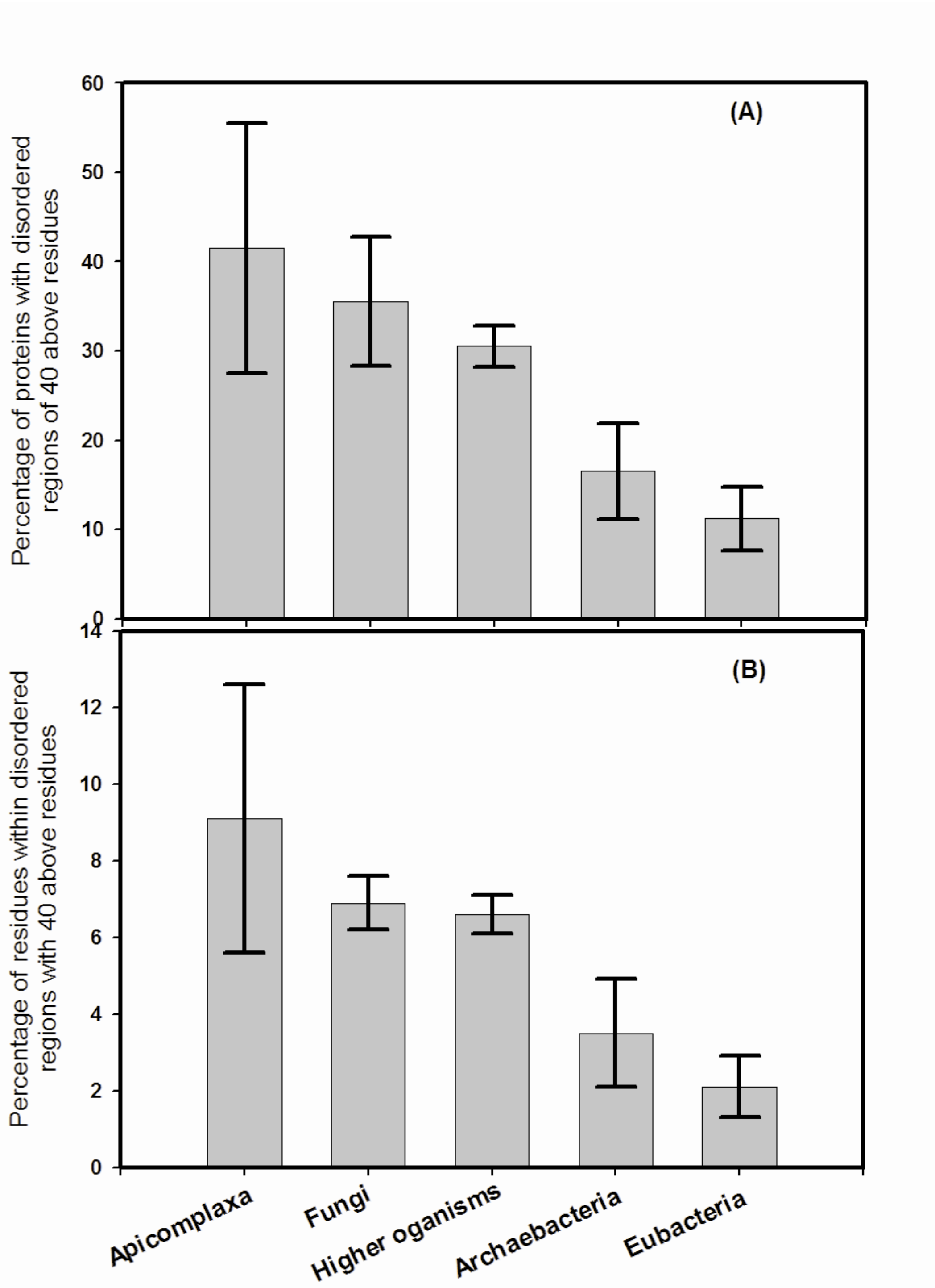
Feng et al., Fig. 2



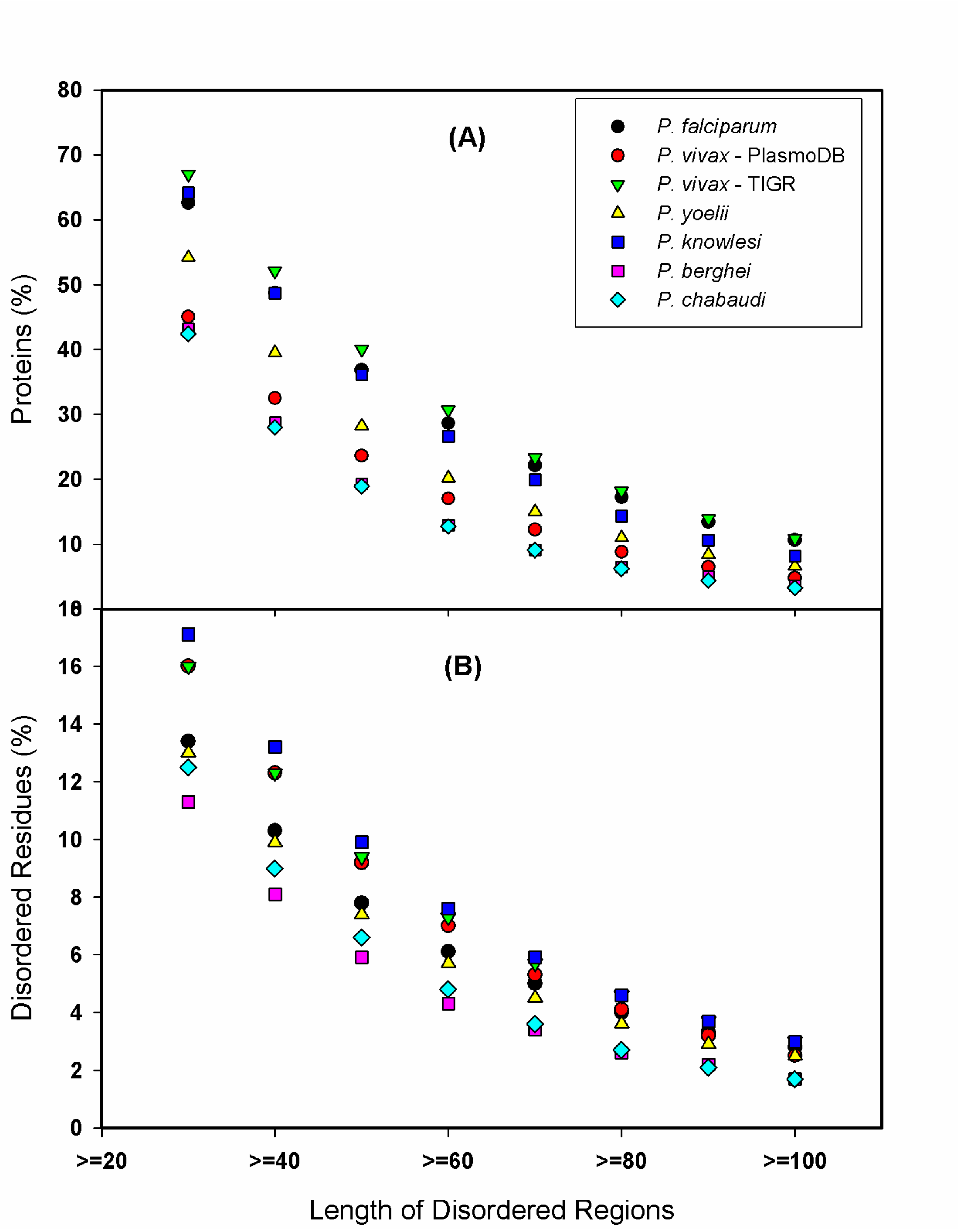
Feng et al., Fig. 3



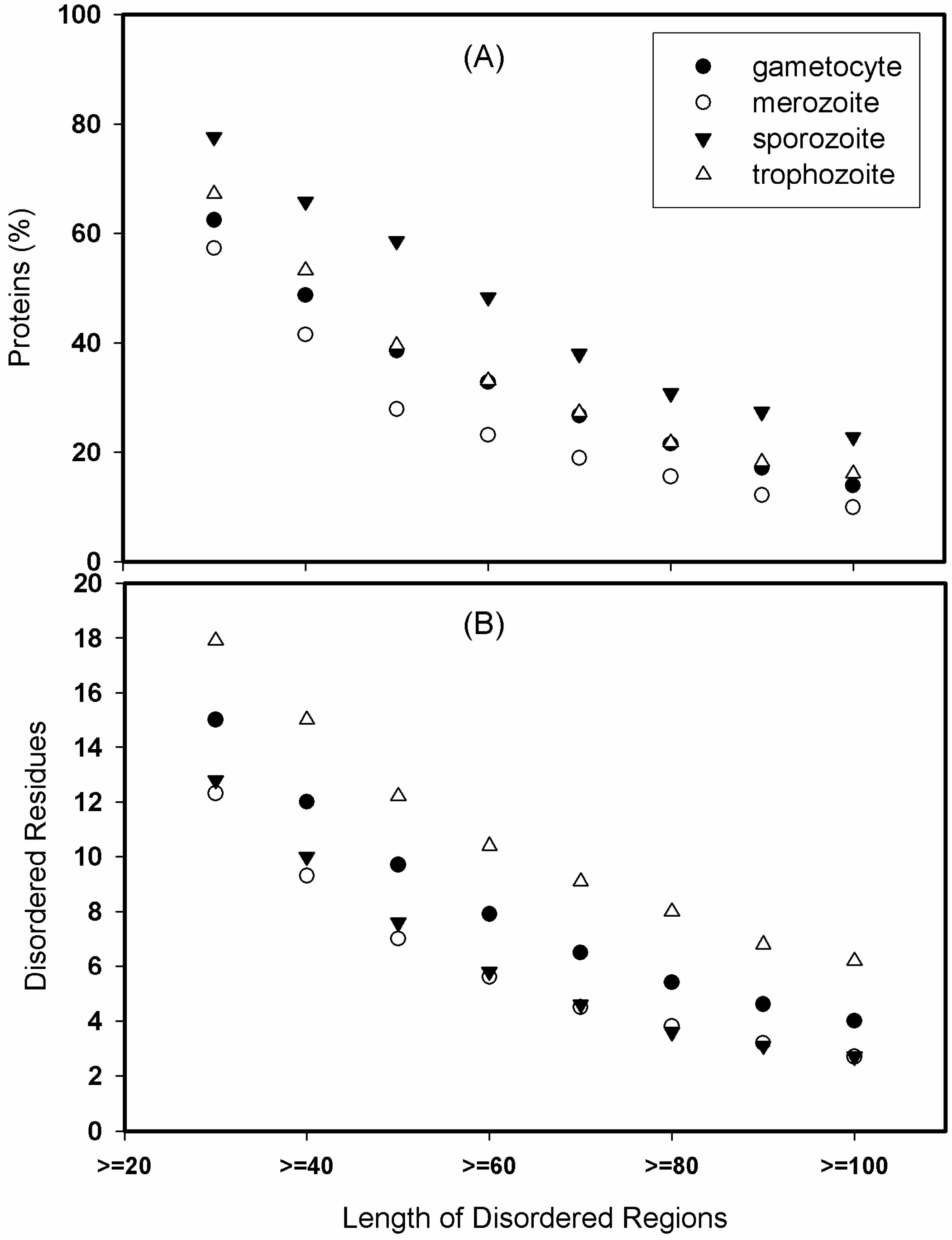
Feng et al., Fig. 4



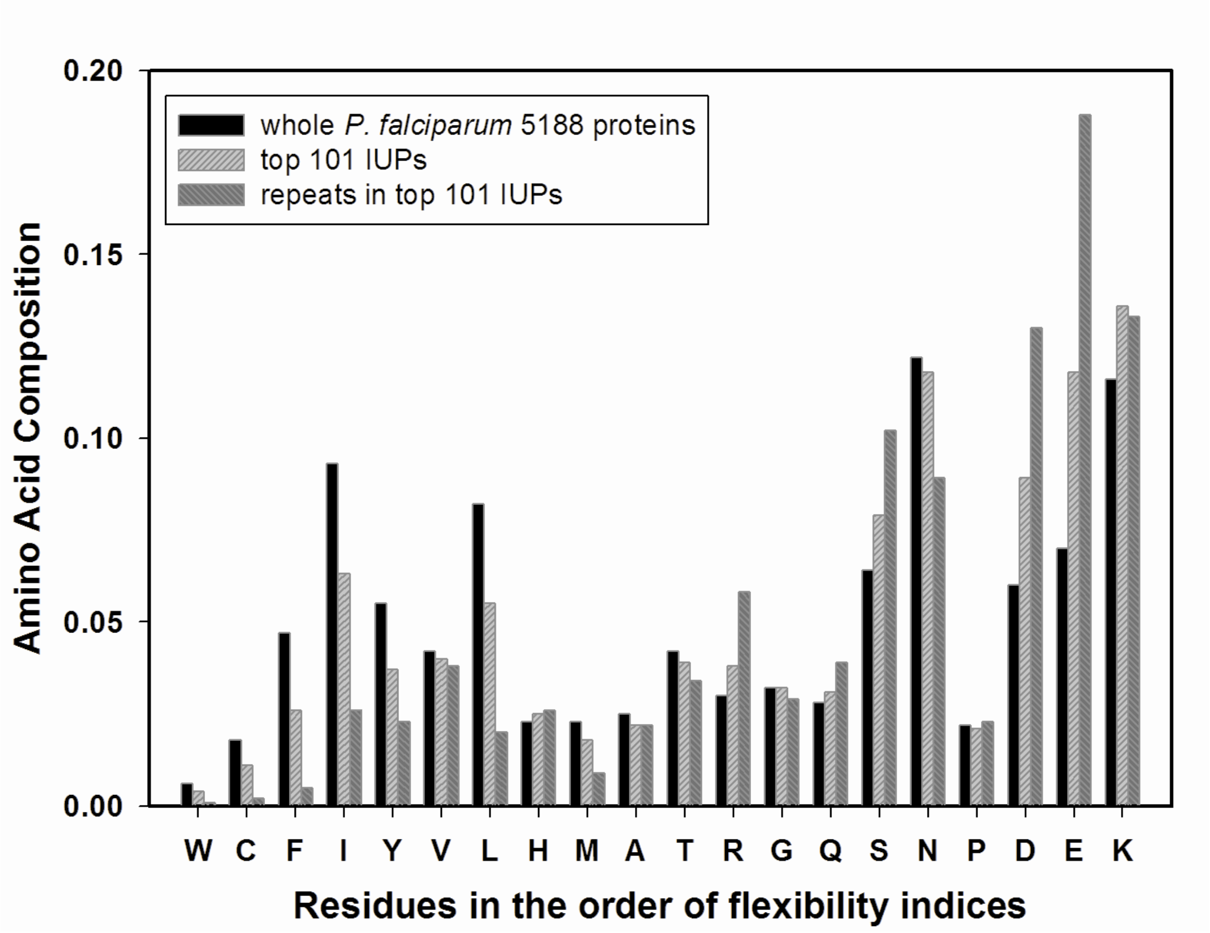
Feng et al., Fig. 5



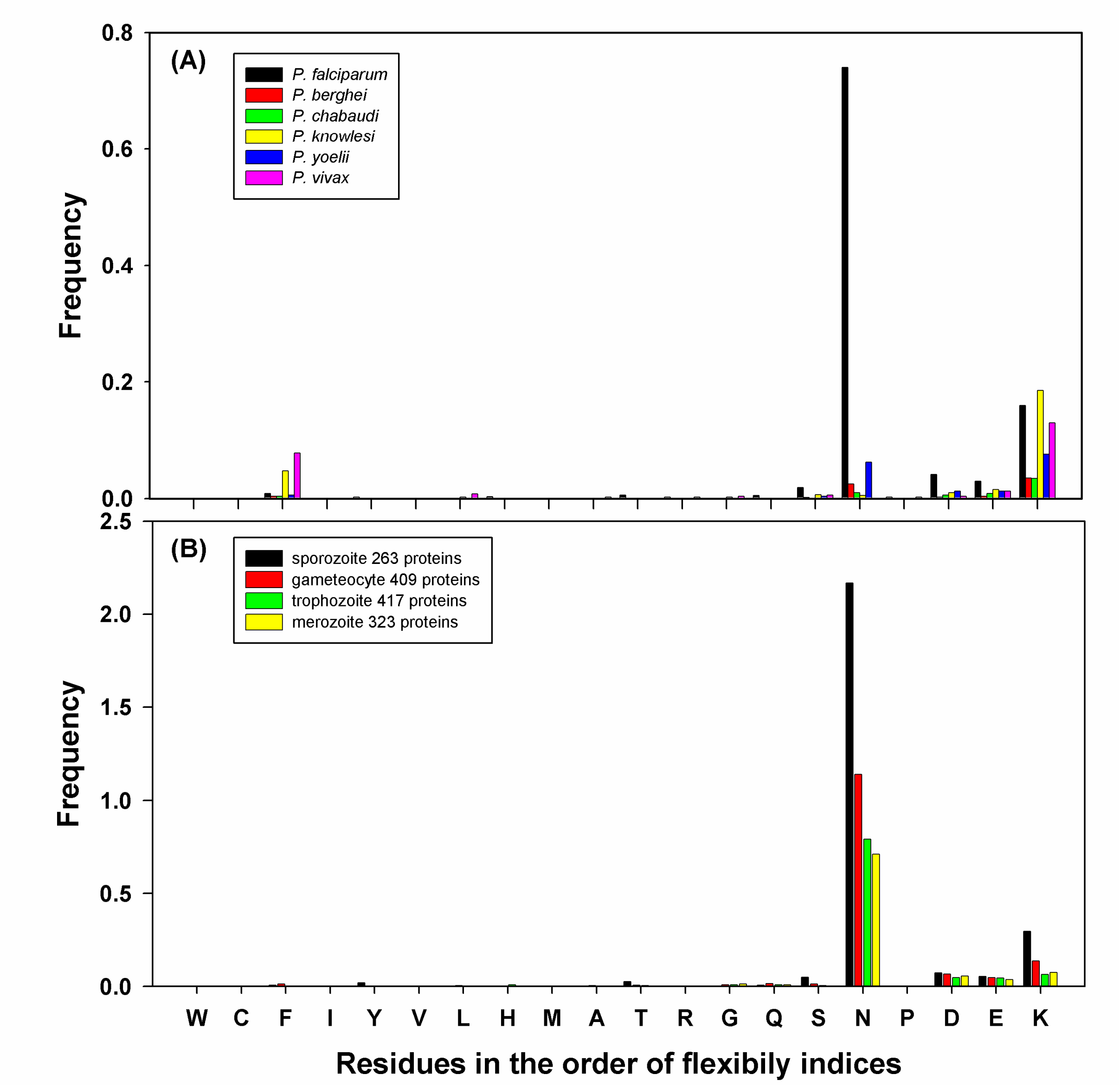
Feng et al., Fig. 6



Feng et al., Fig. 7



Feng et al., Fig.8


(A)
P. falciparum
P. berghei
P. chabaudi
P. knowlesi
P. yoelii
P. vivax
Frequency
(B)
sporozoite 263 proteins
gameteocyte 409 proteins
trophozoite 417 proteins
merozoite 323 proteins
Frequency
W C F I Y V L H M A T R G Q S N P D E K
Residues in the order of flexibily indices