The relationship between protein domains and homopeptides in the *Plasmodium falciparum* proteome

Yue Wang 1, Hsin Jou Yang 1, Paul M Harrison 1

1 Dept. of Biology, McGill University, Montreal, QC, Canada

Corresponding Author: Paul M Harrison
Email address: paul.harrison@mcgill.ca

The proteome of the malaria parasite *Plasmodium falciparum* is notable for the pervasive occurrence of homopeptides or low-complexity regions (*i.e.*, regions that are made from a small subset of amino-acid residue types). The most prevalent of these are made from residues encoded by adenine/thymidine (AT)-rich codons, in particular asparagine. We examined homopeptide occurrences within protein domains in *P. falciparum*. Homopeptide enrichments occur for hydrophobic (*e.g.*, valine), or small residues (alanine or glycine) in short spans (<5 residues), but these enrichments disappear for longer lengths. We observe that short asparagine homopeptides (<10 residues long) have a dramatic relative depletion inside protein domains, indicating some selective constraint to keep them from forming. We surmise that this is possibly linked to co-translational protein folding, although there are specific protein domains that are enriched in longer asparagine homopeptides (≥10 residues) indicating a functional linkage for specific poly-asparagine tracts. Top gene ontology functional category enrichments for homopeptides associated with diverse protein domains include ‘vesicle-mediated transport’, and ‘DNA-directed 5'-3' RNA polymerase activity’, with various categories linked to ‘binding’ evidencing significant homopeptide depletions. Also, in general homopeptides are substantially enriched in the parts of protein domains that are near/in IDR. The implications of these findings are discussed.
The relationship between protein domains and homopeptides in the *Plasmodium falciparum* proteome

*Short title: Domains and homopeptides in *Plasmodium*

Yue Wang¹, Hsin-Jou Yang & Paul M. Harrison¹*

¹Department of Biology,
McGill University,
Montreal, QC,
Canada.

*Correspondence:
Paul M. Harrison

Email: paul.harrison@mcgill.ca
Abstract:
The proteome of the malaria parasite *Plasmodium falciparum* is notable for the pervasive occurrence of homopeptides or low-complexity regions (*i.e.*, regions that are made from a small subset of amino-acid residue types). The most prevalent of these are made from residues encoded by adenine/thymidine (AT)-rich codons, in particular asparagine. We examined homopeptide occurrences within protein domains in *P. falciparum*. Homopeptide enrichments occur for hydrophobic (*e.g.*, valine), or small residues (alanine or glycine) in short spans (<5 residues), but these enrichments disappear for longer lengths. We observe that short asparagine homopeptides (<10 residues long) have a dramatic relative depletion inside protein domains, indicating some selective constraint to keep them from forming. We surmise that this is possibly linked to co-translational protein folding, although there are specific protein domains that are enriched in longer asparagine homopeptides (≥10 residues) indicating a functional linkage for specific poly-asparagine tracts. Top gene ontology functional category enrichments for homopeptides associated with diverse protein domains include ‘vesicle-mediated transport’, and ‘DNA-directed 5’-3’ RNA polymerase activity’, with various categories linked to ‘binding’ evidencing significant homopeptide depletions. Also, in general homopeptides are substantially enriched in the parts of protein domains that are near/in IDRs. The implications of these findings are discussed.

Introduction

*Plasmodium falciparum (Pf)* is a single-celled protozoan that causes malaria in humans. Malaria causes hundreds of thousands of deaths every year, with ~405,000 in
2018 (Global Malaria Programme 2019). Treatment for malaria is confounded by its ability to adapt quickly to drugs and to the human immune system; its antigenic diversity is a major problem for vaccine development (Ferreira et al. 2004; Ferreira et al. 2003; Freitas-Junior et al. 2000). The complete genome sequence of Pf contains >5000 protein-coding genes (Gardner et al. 2002). Early analysis indicated that low-complexity regions (i.e., regions that consist mostly of a small subset of amino-acid types) or homopeptides (runs of single amino acids) are a prominent feature of the encoded proteins, with more than half of proteins being low-complexity over most of their sequences (Pizzi & Frontali 2001). Asparagine-rich regions are the most abundant (An & Harrison 2016; Pizzi & Frontali 2001). The low-complexity regions (LCRs) have been postulated to have a function primarily at the nucleotide level (Xue & Forsdyke 2003). Their abundance depends largely on genomic A+T or G+C content (DePristo et al. 2006; Xue & Forsdyke 2003), and they also acquire further low-complexity insertions and deletions according to a power-law rule: that is, longer LCRs acquire longer insertions/deletions (DePristo et al. 2006). Pf LCRs can be classified into three distinct types, including a high G+C type that is linked to recombination hotspots (Zilversmit et al. 2010). There is a pattern of enrichment of long intergenic poly(AT) tracts in Plasmodium species, some of which are immediately adjacent to genes and run into them (Russell et al. 2014). Although asparagine is preferred in LCRs of Pf, a different residue type with AT-rich codons (lysine) is more prominent in the CVK group, which is a set of four primate-infecting plasmodia (Chaudhry et al. 2018). As well as being sites of polymorphic variation themselves (Chaudhry et al. 2018), Pf LCRs are linked to increased single-nucleotide polymorphism in their vicinity (Haerty & Golding 2011).
Although homopeptides are sites of such polymorphic variation, earlier work showed that some homopeptides are deeply conserved across orthologs from bacteria and eukaryotes, suggesting ancient origin and functional essentiality (Faux et al. 2005). In general, homopeptides are more conserved in bacteria, than in archaea and eukaryotes, and there is a correlation between repeat length differences and species divergence (Uthayakumar et al. 2012). Homopeptides increase the functional versatility of proteins, and facilitate spatial organization of proteins in a repeat-dependent way (Chavali et al. 2017). They are also significantly linked to many human diseases (Lobanov et al. 2016).

Low-complexity regions rich in hydrophilic residues are significantly associated with protein intrinsic disorder (Delucchi et al. 2020; Romero et al. 2001). Previous surveys have shown that 10-40% of Pf LCR residues are predicted as intrinsically disordered in tracts \( \geq 40 \) residues long, and that >60% of sequences have such a tract (Feng et al. 2006; Mohan et al. 2008). Such annotated disordered regions in Pf are significantly depleted of predicted MHC-binding peptides, which has implications for vaccine development, since many vaccine target proteins are intrinsically disordered (Guy et al. 2015).

Low-complexity regions rich in asparagine (and/or glutamine) are common in domains that form prions (i.e., self-propagating amyloid particles) (Harbi & Harrison 2014a; Harbi & Harrison 2014b; Harbi et al. 2012; Harrison 2017; Su & Harrison 2019). In budding yeast (Saccharomyces cerevisiae), propagation of these particles can be sustained during budding, mating and laboratory protocols (Harbi & Harrison 2014c). Predicted prions have been detected in all the domains of life (Espinosa Angarica et al. 2013), including thousands in viruses and phages (Tetz & Tetz 2017; Tetz & Tetz 2018),
and tens of thousands in bacteria (Harrison 2019). \textit{Pf} has prion-like domains (that arise in asparagine-rich LCRs) in 10-24\% of its proteins (Singh et al. 2004) (An & Harrison 2016; Pallares et al. 2018). Just like \textit{Pf}, there are \textit{Saccharomycetes} fungi that have high proportions of prion-like proteins with poly-asparagine in them (An et al. 2016). There is some evidence that asparagine-rich LCRs act as ‘tRNA sponges’ that slow down the translation rate of proteins to aid in co-translational folding (Filisetti et al. 2013; Frugier et al. 2010). This is a type of parallel function at the DNA/RNA level for which there is increasing evidence for intrinsically disordered regions in proteins (Pancsa & Tompa 2016).

Here, we investigate the relationship between homopeptides and both defined protein domains and intrinsic disorder in \textit{Pf} proteins. We observe significant depletions in homopeptides for specific types of amino acid, in particular asparagine and aspartate. Homopeptides are substantially enriched in the parts of protein domains that are near or in intrinsically disordered regions (IDRs).

### Methods

#### Source Data

The UniProt (Boeckmann et al. 2003) reference proteome for \textit{Plasmodium falciparum strain 3D7} was downloaded from www.uniprot.org in January 2019. Protein domain annotations for \textit{Pf} were taken from the Pfam database (El-Gebali et al. 2019). For comparative analysis, three further reference proteomes were downloaded from the same source for the following: another \textit{P. falciparum} strain (FCH/4), \textit{P. yoelii} (strain 17XNL), and \textit{P. vivax} (strain Salvador I).
Annotation

Homopeptides were defined as repetitions of one amino-acid type with a minimum length of 3 residues (Figure 1).

Proteins were annotated for intrinsic disorder using the DISOPRED3 and IUPRED2a programs (Dosztanyi et al. 2005; Huntley et al. 2015; Ward et al. 2004). IUPRED2a operates on inputted single sequences, and predicts intrinsic disorder by estimating inter-residue interaction energies (Erdos & Dosztanyi 2020). It was the best performing single-sequence method for intrinsic disorder annotation, with an AUC (area-under-curve) value of 0.83 for the ROC curve in a recent assessment (Nielsen & Mulder 2019); DISOPRED3 also had a value of AUC=0.83 in this assessment, and was one of the best performing methods that use evolutionary information as input. Only regions of predicted disorder ≥30 residues long were considered. A 30-residue length cut-off was used since this is a common threshold or boundary value used in characterizing intrinsically-disordered regions, or in training algorithms for prediction of intrinsic disorder (Atkins et al. 2015). Thus, we used the default ‘long’ parameter choice for the IUPRED2a program. Also, the DISOPRED3 program was run with a 2% expected false positive rate for the algorithm training set, which is in is the author’s recommended parameter range (Jones & Cozzetto 2015). The results from either intrinsic disorder annotation program were considered separately.

Enrichments & depletions
Because we wish to examine the effects of protein domain structure and intrinsic disorder on the occurrence of homopeptides in Pf, we checked whether there is any deviation from random placement for homopeptides within protein domains and annotated intrinsic disorder. These can be either enrichments relative to background populations or depletions. The background populations were either the whole proteome or the set of protein domain annotations as described below. Enrichments and depletions for homopeptides in protein domains were calculated as depicted (Figure 1). These were determined for individual amino-acid types in homopeptides using equation (1) below for hypergeometric probability, for sampling with replacement:

\[
P(k) = \frac{{K \choose k} {N-K \choose n-k}}{{N \choose n}} \quad (1)
\]

with the sample counts given by:

\[k = \text{number of residues in homopeptides in domains of one amino-acid type}\]
\[n = \text{number of residues in homopeptides in domains}\]

and the background counts given by the quantities:

\[K = \text{number of residues in homopeptides in the proteome of one amino-acid type}\]
\[N = \text{number of residues in homopeptides in the proteome}\]

Enrichments/depletions for the amount of homopeptides in specific protein domain types were also calculated with the sample counts given by:
and the background counts given by the quantities:

\[ K = \text{number of residues in one domain type} \]
\[ N = \text{number of residues in all domain-types} \]

Enrichments and depletions for protein domains overlapping or near annotated intrinsically disordered regions (IDRs) were also calculated with the sample counts given by:

\[ k = \text{number of residues in one domain type which near or inside disordered regions} \]
\[ n = \text{number of residues in all domain types which near or inside disordered regions} \]

and the background counts given by the quantities:

\[ K = \text{number of residues in one domain type} \]
\[ N = \text{number of residues in all domain-types} \]

Proximity to IDRs was determined using a 10-residue buffer at either end of the annotated IDRs (Figure 1).

Gene ontology (GO) category enrichments were also analysed for specific types of homopeptide enrichment (Huntley et al. 2015). These were calculated by mapping the protein domains onto GO categories, and re-totalling the numbers of residues per GO category rather than per domain.
All enrichments/depletions were calculated using hypergeometric probability with appropriate Bonferroni corrections for multiple hypothesis testing. For example, the Bonferroni correction for enrichments/depletions of homopeptides of individual amino acids in protein domains was $P = 0.05/20 = 0.0025$, since there are 20 different amino acids being sampled from the same background population.

**Propensity**

A propensity for homopeptides of a specific amino-acid type to occur in protein domains ($P_{\text{dom}}$) was calculated as:

$$P_{\text{dom}} = \log_{10} \left[ \frac{(k/n)}{(K/N)} \right] \quad (2)$$

The values of $k$, $n$, $K$ and $N$ are as listed above just below equation (1). This was calculated for the homopeptide threshold $\geq 3$ residues.

**Results**

**Enrichments and depletions of homopeptides in protein domains**

Homopeptides are abundant and pervasive in the Pf proteome, yet it is not clear from a structural perspective which homopeptides are more tolerated in protein domains. We analysed the preferences of homopeptides of each specific amino-acid type for insertion into protein domains, using three different length thresholds for homopeptides (Table 1). The statistical enrichments/depletions are listed, as well as the fraction of the homopeptide populations for each amino acid, and the propensity ($P_{\text{dom}}$) of homopeptides of each amino acid for protein domains, calculated as described in *Methods*. For the minimum homopeptide threshold of $\geq 3$ residues length, the most enriched amino acids
include the major aliphatic hydrophobic residues valine, isoleucine and leucine, which is
to be expected because of the extensive hydrophobic cores of protein domains. Also, the
small residues alanine and glycine exhibit highly significant enrichments. For these
amino-acid types, the enrichments are only for short homopeptides (of size 3 or 4
residues), since the enrichments disappear for longer homopeptide thresholds (Table 1).
Positively-charged homopeptides and other hydrophilic homopeptides are also generally
enriched (lysine, arginine, serine, threonine), while negatively-charged homopeptides are
significantly depleted or show no preferences. Lysine homopeptides are the second most
abundant in the proteome and are made from AT-rich codons; their relationships with
specific protein domains are discussed below. Some amino acids show an enrichment,
with a comparable propensity for structural domains (P_{dom}) as for other amino acids, but
these are not significant. Most strikingly though, short asparagine homopeptides are
highly significantly depleted within protein domains.

Histograms of homopeptide length also indicate that within protein domains,
homopeptides generally lack the longer homopeptide lengths (≥10 residues) that make
up the majority of homopeptides outside of protein domains (Figure 2).

For the longest homopeptide threshold (≥10 residues length), the number of
amino-acid types which are comparatively tolerated in protein domains dramatically
decreases to one (asparagine; Table 1). Keeping in mind that the enrichment calculations
are effectively based on the comparison of different amino-acid types, there should
always be at least one enriched amino-acid type unless there are completely no
homopeptides at all at a certain threshold. Poly-asparagine homopeptides are depleted
in domains until the threshold is extended to 10, which leaves it as the only one existing
in domains. The enrichment observed for longer polyasparagine tracts (≥10 residues, Table 1) arises from a small number of specific protein domains that may have a functional linkage for these polyasparagine tracts (Table S1), for example for specific protein interactions.

The enrichment/depletion results for individual amino-acid types are little affected by the boundary definition of protein domains (i.e., chopping off 3, 5 or 7 residues from the ends of the domains, Table S3), with just some enrichments for glutamine and glutamate becoming significant for these shortened domains. This indicates that these homopeptides significantly occur near the ends of protein domains.

Since lysine and arginine homopeptides are in general significantly enriched in protein domains, and asparagine and aspartate significantly depleted, we examined which individual protein domains are linked to these trends (Table S1). Despite the substantial general depletion of asparagine homopeptides within protein domains, there are 87 individual domains with significant enrichment of asparagine homopeptides, including the low-copy-number Sin-like region and the Sacl homology domain; these also stand out when we restrict the analysis to polyasparagine tracts ≥10 residues long (Table S1). The most prominent lysine homopeptide enrichments are for Rifin and PfEMP DBL domains. Specific domains are also linked to hydrophobic or small-residue homopeptides, such as glycine homopeptides arising for ribosomal proteins (Table S1).

*Gene Ontology enrichments*
We examined the enrichments and depletions of Gene Ontology functional categories associated with homopeptides in protein domains (Table S2). Some top enrichments of Gene Ontology (GO) functional categories for homopeptides in protein domains include: GO:0003899 (DNA-directed 5'-3' RNA polymerase activity), an enrichment caused by seven different protein domains, and GO:0042578 (phosphoric ester hydrolase activity), which is unique to the SacI homology domain (which is involved in clathrin-mediated endocytosis; 3 copies in Pf) (Table S2). Inspection of other GO category enrichments indicate that they are also caused by diverse protein domains, e.g., GO:0016192 (vesicle-mediated transport) which is linked to homopeptide enrichments in 11 different protein domains, pointing to specific functional significance for homopeptides in the interaction of these proteins. Nonetheless, in general ‘protein binding’ is significantly depleted in the list (Table S2), as are the other high-level ‘RNA-binding’ and ‘GTP-binding’ terms.

The relationship between specific protein domain homopeptides and intrinsic disorder

IDRs tend to have homopeptides and low-complexity sequences in them (Romero et al. 2001). We surmised that the relationship of different protein domains with homopeptides might be caused by their proximity to or overlap with intrinsically disordered regions (IDRs) of proteins. In general, homopeptides are enriched in the parts of protein domains that are near or in IDRs (Table 2; results for either the DISOPRED3 or IUPred2A program are shown). Also, there is only one individual protein domain type that is
significantly depleted in homopeptides near/in IDRs, with the remainder of significant deviations being enrichments (Table 2).

**Comparison of trends in other Plasmodia**

The trends observed for *Pf* strain 3D7 were validated by analysis of another *Pf* strain (FCH/4) that was picked from the UniProt reference proteome list (Boeckmann et al. 2003) (Table S4). There is just one small change with enrichments of glutamate homopeptides in protein domains becoming significant (Table S4). Comparisons were also made with proteomes of *P. yoelii*, a malaria parasite of rodents, and *P. vivax*, a member of the CVK group of primate-infecting plasmodia (Chaudhry et al. 2018). *P. yoelii* has an overall approximately even predominance of N and K homopeptides, and *P. vivax* has predominance of K homopeptides rather than of N homopeptides (Table S4). The depletion of N homopeptides in protein domains is maintained in *P. yoelii*, but there is no significant depletion/enrichment in *P. vivax*. K homopeptides also become significantly depleted within protein domains in *P. yoelii*, despite their similar overall levels to *Pf* (29% in *P.yoelii* versus 23% in *Pf*). The results for homopeptide enrichments in parts of protein domains overlapping IDRs (Table 2) also remain highly significant for these three other *Plasmodia* proteomes (P-values ~0.0).

**Discussion**

**Homopeptide trends**
Homopeptide enrichments within protein domains, such as for hydrophobic (L, I or V) or small (A or G) residues, disappear at longer lengths (≥5 residues). This indicates a limit to their toleration within protein domain cores, for example because they are not so easily accommodated in regular secondary structures.

We observed a substantial significant relative depletion of short asparagine runs (<10 residues long) in protein domains. Plasmodia have acquired great amounts of N homopeptide tracts during evolution, but statistically these have not been appearing or ‘landing’ within domains. The lack of short intra-domain asparagine runs may be because they interfere with protein folding in some way. For example, they may slow down co-translational protein folding due to a lack of asparaginyl-tRNAs, since levels of asparaginyl-tRNAs in Pf are normal despite the high amounts of asparagine in their coding sequences (Filisetti et al. 2013; Frugier et al. 2010). Thus asparagine homopeptides may be ‘tRNA sponges’ that soak up tRNAs and slow down translation and co-translational folding (Filisetti et al. 2013; Frugier et al. 2010). It is possible that homopeptides, and in particular poly-asparagine homopeptides may make protein domains more prone to misfolding. Although generally slower translation is thought to aid in correct co-translational folding (Waudby et al. 2019), sometimes faster translation is more desirable through segments that are prone to misfolding (O’Brien et al. 2014), or for translational efficiency at buried residue sites or sites that are vulnerable to structurally disruptive mutations (Wang et al. 2015; Zhou et al. 2009). However, experiments with Pf chaperone Hsp110c, have shown that Pf has cellular mechanisms that are designed to prevent aggregation linked to asparagine tracts (Muralidharan & Goldberg 2013). A few specific protein domains have enrichment of long polyasparagine tracts. Such tracts may
have a specific functional role in these proteins, perhaps for protein or nucleic-acid interaction. Another possibility might be that correct folding of these specific domains is not affected by slow rates of translation, thus N homopeptides can arise in them because of their general abundance in the proteome. Interestingly, the significant depletion of N homopeptides is maintained in the rodent malaria pathogen *P. yoelii*, but there is an absence of significant depletion/enrichment in the more distantly related *P. vivax* malaria pathogen from the CVK group. This may indicate that the N homopeptides in the *P. vivax* protein domains are reduced mainly to those that have functional roles.

**Gene Ontology enrichments**

In the gene ontology enrichments/depletion analysis we see a general trend for depletion of functional categories associated with ‘binding’ (protein binding; RNA binding; GTP binding; ion binding). This suggests that homopeptides may be selected against in structured interaction interfaces, perhaps since they introduce a lack of interaction specificity, or increase the likelihood of off-target binding. Also, the gene ontology results indicate that homopeptide occurrences may be useful information for the discrimination of protein function from the analysis of sequences (Huntley et al. 2014; Huntley et al. 2015; Jiang et al. 2016; Le et al. 2019a; Le et al. 2019b; Mutowo-Meullenet et al. 2013).

**Intrinsic disorder**

We surmise that homopeptides in protein domains are enriched in the parts of the domains near or in IDRs because IDRs generally have more tolerance for
insertions/deletions, and are the main determinants of changes in protein length over evolution (Light et al. 2013). Also, our results indicate that the parts of protein domains that can become intrinsically disordered are enriched in homopeptides relative to other domain parts. This may be an important part of encrypting their ability to transition structurally (Narasumani & Harrison 2015).

Conclusions

The most pervasive homopeptide in *Plasmodium falciparum*, poly-asparagine, is substantially depleted within protein domains, whereas other homopeptides that we might expect in the hydrophobic core of domains, such as poly-leucine or -valine or -isoleucine, and other generally abundant homopeptides (such as lysine) are enriched at the shorter homopeptide lengths studied. We hypothesize that generally poly-asparagine formation is repressed inside protein domains because its occurrence may slow co-translational folding (Filisetti et al. 2013; Frugier et al. 2010), which might be problematic within a protein domain that has only partially been translated. (Scenarios in which both transient fast and slow folding may be problematic for co-translational protein folding are possible (O'Brien et al. 2014)). Further experimental work is needed to investigate these hypotheses. In general, homopeptides are depleted for functional categories associated with diverse types of binding, indicating that they may interfere with specificity in structured interfaces. However, the parts of protein domains that can become intrinsically disordered have homopeptide enrichment relative to other parts of domains. Since some domains fold upon binding to other proteins, and the parts of protein domains that overlap
intrinsic disorder have such homopeptide enrichment, these results suggest that protein homopeptides may be useful in effecting such structural transitions.

References

An L, Fitzpatrick D, and Harrison PM. 2016. Emergence and evolution of yeast prion and prion-like proteins. *BMC Evol Biol* 16:24. 10.1186/s12862-016-0594-3

An L, and Harrison PM. 2016. The evolutionary scope and neurological disease linkage of yeast-prion-like proteins in humans. *Biol Direct* 11:32. 10.1186/s13062-016-0134-5

Atkins JD, Boateng SY, Sorensen T, and McGuffin LJ. 2015. Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. *Int J Mol Sci* 16:19040-19054. 10.3390/ijms160819040

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31:365-370.

Chaudhry SR, Lwin N, Phelan D, Escalante AA, and Battistuzzi FU. 2018. Comparative analysis of low complexity regions in Plasmodia. *Sci Rep* 8:335. 10.1038/s41598-017-18695-y

Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, Latsheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, and Babu MM. 2017. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nat Struct Mol Biol* 24:765-777. 10.1038/nsmb.3441

Delucchi M, Schaper E, Sachenkova O, Elofsson A, and Anisimova M. 2020. A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder. *Genes (Basel)* 11. 10.3390/genes11040407

DePristo MA, Zilversmit MM, and Hartl DL. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* 378:19-30. 10.1016/j.gene.2006.03.023

Dosztanyi Z, Csizmok V, Tompa P, and Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433-3434. 10.1093/bioinformatics/bti541

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, and Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* 47:D427-D432. 10.1093/nar/gky995

Erdos G, and Dosztanyi Z. 2020. Analyzing Protein Disorder with IUPred2A. *Curr Protoc Bioinformatics* 70:e99. 10.1002/cpbi.99

Espinosa Angarica V, Ventura S, and Sancho J. 2013. Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains. *BMC Genomics* 14:316. 10.1186/1471-2164-14-316

Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, and Whisstock JC. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* 15:537-551. 10.1101/gr.3096505
397 Feng ZP, Zhang X, Han P, Arora N, Anders RF, and Norton RS. 2006. Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. *Mol Biochem Parasitol* 150:256-267. 10.1016/j.molbiopara.2006.08.011

400 Ferreira MU, da Silva Nunes M, and Wunderlich G. 2004. Antigenic diversity and immune evasion by malaria parasites. *Clin Diagn Lab Immunol* 11:987-995.

402 10.1128/CDLI.11.6.987-995.2004

403 Ferreira MU, Ribeiro WL, Tonon AP, Kawamoto F, and Rich SM. 2003. Sequence diversity and evolution of the malaria vaccine candidate merozoite surface protein-1 (MSP-1) of *Plasmodium falciparum*. *Gene* 304:65-75. 10.1016/s0378-1119(02)01180-0

406 Filisetti D, Theobald-Dietrich A, Mahmoudi N, Rudinger-Thirion J, Candolfi E, and Frugier M. 2013. Aminoacylation of *Plasmodium falciparum* tRNA(Asn) and insights in the synthesis of asparagine repeats. *J Biol Chem* 288:36361-36371.

409 10.1074/jbc.M113.522896

410 Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrass U, Wellemes TE, and Scherf A. 2000. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* 407:1018-1022.

413 10.1038/35039531

414 Frugier M, Bour T, Ayach M, Santos MA, Rudinger-Thirion J, Theobald-Dietrich A, and Pizzi E. 2010. Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. *FEBS Lett* 584:448-454. 10.1016/j.febslet.2009.11.004

417 Gardner MJ, Hall N, Fung E, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shalom SJ, Suh B, Peterson J, Angiuoli S, Pertain A, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, and Barrell B. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419:498-511. 10.1038/nature01097

425 Global Malaria Programme WG. 2019. World malaria report 2019.

426 Guy AJ, Irani V, MacRairld CA, Anders RF, Norton RS, Beeson JG, Richards JS, and Ramsland PA. 2015. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *PLoS One* 10:e0141729.

429 10.1371/journal.pone.0141729

430 Haerty W, and Golding GB. 2011. Increased polymorphism near low-complexity sequences across the genomes of Plasmodium falciparum isolates. *Genome Biol Evol* 3:539-550.

432 10.1093/gbe/evr045

433 Harbi D, and Harrison PM. 2014a. Classifying prion and prion-like phenomena. *Prion* 8:161-165. 10.4161/priv.27960

435 Harbi D, and Harrison PM. 2014b. Classifying prion and prion-like phenomena. *Prion* 8.

436 Harbi D, and Harrison PM. 2014c. Interaction networks of prion, prionogenic and prion-like proteins in budding yeast, and their role in gene regulation. *PLoS One* 9:e100615.

438 10.1371/journal.pone.0100615

439 Harbi D, Parthiban M, Gendoo DM, Ehsani S, Kumar M, Schmitt-Ulms G, Sowdhamini R, and Harrison PM. 2012. PrionHome: a database of prions and other sequences relevant to prion phenomena. *PLoS One* 7:e31785.
Harrison PM. 2017. fLPS: Fast discovery of compositional biases for the protein universe. *BMC Bioinformatics* 18:476. 10.1186/s12859-017-1906-3

Harrison PM. 2019. Evolutionary behaviour of bacterial prion-like proteins. *PLoS One* 14:e0213030. 10.1371/journal.pone.0213030

Huntley RP, Harris MA, Alam-Faruque Y, Blake JA, Carbon S, Dietze H, Dimmer EC, Foulger RE, Hill DP, Khodiyar VK, Lock A, Lomax J, LovingRC, Mutowo-Meullenet P, Sawford T, Van Auken K, Wood V, and Mungall CJ. 2014. A method for increasing expressivity of Gene Ontology annotations using a compositional approach. *BMC Bioinformatics* 15:155. 10.1186/1471-2105-15-155

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, and O'Donovan C. 2015. The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* 43:D1057-1063. 10.1093/nar/gku1113

Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraei SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci P, Jones DT, Chapman S, Bke D, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasisidharan R, Yang H, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcket-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giolli M, Piovesan D, Tosatto SC, Del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk AD, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida ESDC, Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orenco CA, Rost B, Greene CS, Mooney SD, Friedberg I, and Radiivojac P. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 17:184. 10.1186/s13059-016-1037-6

Jones DT, and Cozzetto D. 2015. DISOPRED3: precise ordered region predictions with annotated protein-binding activity. *Bioinformatics* 31:857-863. 10.1093/bioinformatics/btu744

Le NQK, Yapp EKY, Nagasundaram N, Chua MCH, and Yeh HY. 2019a. Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture. *Comput Struct Biotechnol J* 17:1245-1254. 10.1016/j.csbj.2019.09.005

Le NQK, Yapp EKY, and Yeh HY. 2019b. ET-GRU: using multi-layer gated recurrent units to identify electron transport proteins. *BMC Bioinformatics* 20:377. 10.1186/s12859-019-2972-5

Light S, Sagit R, Sachenkova O, Ekman D, and Elofsson A. 2013. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol* 30:2645-2653. 10.1093/molbev/mst157
Lobanov MY, Klus P, Sokolovsky IV, Tartaglia GG, and Galzitskaya OV. 2016. Non-random distribution of homo-repeats: links with biological functions and human diseases. *Sci Rep* 6:26941. 10.1038/srep26941

Mohan A, Sullivan WJ, Jr., Radivoja P, Dunker AK, and Uversky VN. 2008. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol Biosyst* 4:328-340. 10.1039/b719168e

Muralidharan V, and Goldberg DE. 2013. Asparagine repeats in Plasmodium falciparum proteins: good for nothing? *PLoS Pathog* 9:e1003488. 10.1371/journal.ppat.1003488

Mutowo-Meullenet P, Huntley RP, Dimmer EC, Alam-Faruque Y, Sawford T, Jesus Martin M, O'Donovan C, and Apweiler R. 2013. Use of Gene Ontology Annotation to understand the peroxisome proteome in humans. *Database (Oxford)* 2013:bas062. 10.1093/database/bas062

Narasumani M, and Harrison PM. 2015. Bioinformatical parsing of folding-on-binding proteins reveals their compositional and evolutionary sequence design. *Sci Rep* 5:18586. 10.1038/srep18586

Nielsen JT, and Mulder FAA. 2019. Quality and bias of protein disorder predictors. *Sci Rep* 9:5137. 10.1038/s41598-019-41644-w

O'Brien EP, Vendruscolo M, and Dobson CM. 2014. Kinetic modelling indicates that fast-translating codons can coordinate cotranslational protein folding by avoiding misfolded intermediates. *Nat Commun* 5:2988. 10.1038/ncomms3988

Pallares I, de Groot NS, Iglesias V, Sant'Anna R, Biosca A, Fernandez-Busquets X, and Ventura S. 2018. Discovering Putative Prion-Like Proteins in Plasmodium falciparum: A Computational and Experimental Analysis. *Front Microbiol* 9:1737. 10.3389/fmicb.2018.01737

Pancsa R, and Tompa P. 2016. Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. *Trends Biochem Sci* 41:898-906. 10.1016/j.tibs.2016.08.009

Pizzi E, and Frontali C. 2001. Low-complexity regions in Plasmodium falciparum proteins. *Genome Res* 11:218-229. 10.1101/gr.gr-1522r

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, and Dunker AK. 2001. Sequence complexity of disordered protein. *Proteins* 42:38-48.

Russell K, Cheng CH, Bizzaro JW, Ponts N, Emes RD, Le Roch K, Marx KA, and Horrocks P. 2014. Homopolymer tract organization in the human malarial parasite Plasmodium falciparum and related Apicomplexan parasites. *BMC Genomics* 15:848. 10.1186/1471-2164-15-848

Singh GP, Chandra BR, Bhattacharya A, Akhouri RR, Singh SK, and Sharma A. 2004. Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum. *Mol Biochem Parasitol* 137:307-319. 10.1016/j.molbiopara.2004.05.016

Su TY, and Harrison PM. 2019. Conservation of Prion-Like Composition and Sequence in Prion-Formers and Prion-Like Proteins of Saccharomyces cerevisiae. *Front Mol Biosci* 6:54. 10.3389/fmolb.2019.00054

Tetz G, and Tetz V. 2017. Prion-Like Domains in Phagobiota. *Front Microbiol* 8:2239. 10.3389/fmicb.2017.02239

Tetz G, and Tetz V. 2018. Prion-like Domains in Eukaryotic Viruses. *Sci Rep* 8:8931. 10.1038/s41598-018-27256-w
Uthayakumar M, Benazir B, Patra S, Vaishnavi MK, Gurusaran M, Sureka K, Jeyakanthan J, and Sekar K. 2012. Homepeptide repeats: implications for protein structure, function and evolution. *Genomics Proteomics Bioinformatics* 10:217-225. 10.1016/j.gpb.2012.04.001

Wang E, Wang J, Chen C, and Xiao Y. 2015. Computational evidence that fast translation speed can increase the probability of cotranslational protein folding. *Sci Rep* 5:15316.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, and Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337:635-645. 10.1016/j.jmb.2004.02.002

Waudby CA, Dobson CM, and Christodoulou J. 2019. Nature and Regulation of Protein Folding on the Ribosome. *Trends Biochem Sci.* 10.1016/j.tibs.2019.06.008

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

Zhou T, Weems M, and Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol* 26:1571-1580. 10.1093/molbev/msp070

Zilversmit MM, Volkman SK, DePristo MA, Wirth DF, Awadalla P, and Hartl DL. 2010. Low-complexity regions in Plasmodium falciparum: missing links in the evolution of an extreme genome. *Mol Biol Evol* 27:2198-2209. 10.1093/molbev/msq108
Table 1 (on next page)

Homopeptides enriched/depleted amino-acid types in protein domains sorted by p-values
# Table 1: Homopeptide enriched/depleted amino-acid types in protein domains

| Amino Acid (one letter code) | Homopeptide amount (≥3) in pfam domain (n=190 52) | Fraction in domains | $P_{dom}$ *** | Homopeptide amount (≥3) in proteome (N=209 236) | Homopeptide amount (≥5) in pfam domain (n=2121) | Homopeptide amount (≥5) in proteome (N=613 33) | P-value * | Homopeptide amount (≥10) in pfam domain (n=349) | Homopeptide amount (≥10) in proteome (N=177 24) | P-value |
|-----------------------------|-----------------------------------------------------|---------------------|---------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|----------|-----------------------------------------------|-----------------------------------------------|----------|
| A                           | 723                                                 | 0.76                | +0.92         | 954                                           | 0                                             | 0                                             | 1        | 0                                             | 0                                             | 1        |
| L                           | 1676                                                | 0.20                | +0.34         | 8250                                          | 2.44e-222                                    | 0                                             | 0       | 0                                             | 0                                             | 1        |
| I                           | 1527                                                | 0.21                | +0.36         | 7315                                          | 4.69e-214                                    | 0                                             | 0       | 0                                             | 0                                             | 1        |
| G                           | 579                                                 | 0.38                | +0.62         | 1527                                          | 1.30e-206                                    | 0                                             | 125     | 0                                             | 0                                             | 1        |
| V                           | 378                                                 | 0.34                | +0.57         | 1108                                          | 4.13e-118                                    | 0                                             | 0       | 1                                             | 0                                             | 1        |
| R                           | 454                                                 | 0.28                | +0.49         | 1606                                          | 1.72e-108                                    | 0                                             | 121     | 0                                             | 0                                             | 1        |
| T                           | 417                                                 | 0.18                | +0.30         | 2341                                          | 2.67e-40                                     | 15                                            | 318     | 0                                             | 0                                             | 1        |
| K                           | 5036                                                | 0.10                | +0.04         | 48223                                         | 1.81e-31                                     | 509                                           | 10312   | 1.95e-18                                     | 0                                             | 1        |
| S                           | 1059                                                | 0.11                | +0.08         | 9511                                          | 9.57e-13                                     | 75                                            | 1167    | 1.51e-07                                     | 0                                             | 146      | 0.05417 |
| P                           | 131                                                 | 0.17                | +0.27         | 774                                           | 2.45e-12                                     | 0                                             | 129     | 0                                             | 0                                             | 1        |
| F                           | 481                                                 | 0.12                | +0.12         | 3913                                          | 4.00e-12                                     | 0                                             | 144     | 0                                             | 0                                             | 1        |
| Y                           | 393                                                 | 0.11                | +0.08         | 3582                                          | 1.54e-05                                     | 0                                             | 190     | 0.00123                                     | 0                                             | 0        | 1        |
| C                           | 27                                                  | 0.18                | +0.30         | 148                                           | 0.00022                                     | 0                                             | 5       | 0.83864                                     | 0                                             | 0        | 1        |
| Q                           | 121                                                 | 0.11                | +0.08         | 1098                                          | 0.00384                                     | 15                                            | 256     | 0.01668                                     | 0                                             | 107      | 0.11832 |
| Amino Acid | E Value | Z Value | # Homopeptides | P-value | E Value | Z Value | # Homopeptides | P-value |
|------------|---------|---------|----------------|---------|---------|---------|----------------|---------|
| E          | 0.09    | -0.01   | 12572          | 0.00914 | 65      | 0       | 1907           | 0.05070 |
| W          | 0.33    | +0.56   | 9              | 0.03576 | 0       | 1       | 0              | 0       |
| M          | 0.12    | +0.12   | 189            | 0.04567 | 0       | 0       | 1              | 0       |
| N **       | 0.04    | -0.36   | 92722          | 0       | 1363    | 43632   | 5.11e-13       | 349     |
| D          | 0.07    | -0.11   | 12636          | 7.83e-20| 79      | 2642    | 0.01802        | 0       |
| H          | 0.04    | -0.36   | 758            | 2.31e-07| 0       | 124     | 0.01267        | 0       |

* P-value threshold = 0.0025 (with a Bonferroni correction accounting for tests on the twenty amino acids). P-values of 0.0 are infinitesimally small beyond the precision of the computation.

** Significant enrichments or depletions are in bold. Underlined ones are homopeptide-depleted amino acids.

*** $P_{dom}$ is the propensity of homopeptides of a specific amino-acid type to occur in protein domains. It is calculated as described in Methods.
Table 2 (on next page)

Enrichment of homopeptides within protein domains that are near or overlapping IDRs
**Table 2: Enrichment of homopeptides within protein domains that are near or overlapping IDRs**

* These P-value results are not affected by chopping off 3, 5 or 7 residues from the ends of the protein domains, as for Table S3.

** Total number of individual protein domains that have enrichment of homopeptides within their parts that are near or overlapping IDRs.

*** Total number of individual protein domains that have depletion of homopeptides within their parts that are near or overlapping IDRs.

| Intrinsic disorder annotator | Total number of domain residues in/near IDRs | Homopeptide residues in/near IDRs | Total number of domain residues | Total number of homopeptides in domains | P-value* | ** | *** |
|-----------------------------|---------------------------------------------|-----------------------------------|---------------------------------|----------------------------------------|---------|----|-----|
| IUPred2A                    | 38940                                       | 2845                              | 808565                         | 19052                                  | 0.0     | 49 | 1   |
| DisoPred3                   | 15928                                       | 1381                              | 808565                         | 19052                                  | 0.0     | 29 | 1   |
Figure 1

Schematic of the analysis.

Homopeptides were defined as ≥3 consecutive amino acids of the same type in a sequence. Protein domains near intrinsically disordered regions (IDRs) were determined using a 10-residues buffer. Also, if the IDR is within a protein domain or otherwise overlaps it, a 10-residue buffer is considered on either side of the IDR as shown.
Figure 2

Distribution of homopeptide length inside and outside of protein domains.

(a) The distribution of homopeptide length for all residues both inside and outside protein domains. The natural log of the total number of homopeptides for a given length is used. Parts (b), (c), (d) are the same distributions but for K-, N-, and D-homopeptides respectively.
