Title page

Article title:
New Evolutionary and Structural Insights In Plasmodium Proteins

Authors:
1*Andrea Cappannini , 1* Sergio Forcelloni and 1 , 2 Andrea Giansanti

Affiliation:

Sapienza University of Rome, Department of Physics, P.le A. Moro 5, 00185 Roma, Italy.
2
Istituto Nazionale di Fisica Nucleare, INFN, Roma1 section. 00185, Roma, Italy.

Corresponding Author:
Andrea Cappannini
Sapienza University of Rome, Department of Physics, P.le A. Moro 5, 00185 Roma, Italy.
E-mail:
andreacappannini@gmail.com.it
Phone:
+393403365228

Abstract

In this work, we perform a systematic analysis of evolutionary forces (i.e. mutational bias and natural selection) acting in plasmode proteomes. The main focus is to clarify the origin of low complexity strands in plasmodium

proteins. For this purpose, we consider P.falciparum as the principal case of study and we take into account the other parasites as a suitable benchmark with which compare the low complexity enrichment, GC content and evolutionary pressures acting on codon bias. Using different bioinformatic tools and a new index that we call here *selective pressure index*, we find evidences that low complexity insertions are inserted in proteins where mutational bias acts as the primarily evolutionary force. Protein length has revealed to be an important evolutionary and fitness characteristic for proteins (Wang et al.,2004;Li,1997;Akashi, 2003;Moriyama and Powell, 1998;Pall et al. 2001). Interested, we lead a further study on protein length finding a preferential LC insertion in longer proteins, which are , in turn, less subject to natural selection. Additionally, dividing the proteins between LCPs (Low complexity containing proteins) and nLCPs (the counterpart), we observe that LCPs are significantly longer than nLCP. Overall, we find evidence that this class of proteins containing long low-complexity segments could be a signal of evolutionary and adapting path for these parasites.

1 Introduction

Low complexity regions (LCRs) or non-random regions of few amino acids are abundantly present in proteins (Kumari et al.,2005). Although the evolutionary history and functional roles of LCRs remain unclear (Battistuzzi et al., 2016), they are evolutionary conserved, being present in all kingdoms of life, from Bacteria, to Archaea, to Eucarya (Wotton & Federhen, 1994).

The replication slippage, a mutational process that expands or contracts microsatellites tracts of repetitive DNA motifs, was proposed as the major factor in determining the genetic origin of LCRs (Guy-Franck et al. 2008, Ellegren,2004). The expansion or contraction of these regions, whenever it happens in protein-coding regions, can greatly enlarge the space of protein sequences, generating novel cellular functions with minimal genetic load (Kashi & King, 2006; Toll-Riera et al. 2012; Green & Wang, 2004). In line with that, several lines of evidence have shown that evolutionary recently emerged proteins tend to have more low complexity insertions (LCI) than older ones, thus suggesting a crucial evolutionary and adaptive role of these regions (Toll-Riera et al.,2012). Despite their repetitive aminoacidic structure

biological functions for LC are variegated : although LCRs are frequently involved in neurodegenerative diseases (Gatchel Zoghbi,2005) and developmental diseases (Brown Brown,2004), they also play a central role in protein-protein interactions (Xiao Jeang, 1998), protein-nucleic acid interactions (Shen et al., 2004), and protein subcellular localization (Salichs et al., 2009). However, it is understudied their evolutionary history and functional roles (Battistuzzi et al., 2016).

P.falciparum, the etiological agent of the most virulent form of malaria (Pizzi & Frontali, 2001), has been found filled in low complexity (LC) being these residues strongly enriched in asparagine (N), (Pizzi & Frontali,2001). Several attempts has been made to understand the origin of these strands and their function (Pizzi Frontali,2001; Filisetti et al.2013; Frugier et al. 2010). It has been proposed how LCR may acts as an immune-evasive mechanism (Pizzi Frontali,2001) or as tRNA sponges (Frugier et al., 2010; Filisetti et al., 2013).Considering P.falcipatum it has been proposed how these strands could lead to the formation of novel proteins domains endorsed by the action of heat shock proteins preventing theme to form prion domains (Muralidharan et al .2012;Cobb et al.,2017) . Controversies are found in establishing which evolutionary force can be the main protagonist in shaping these regions. Pizzi and Frontali (Pizzi Frontali, 2001) suggest, how the selection could be the main protagonist in shaping these asparagine regions while others (Chaundry et al, 2017) suggest how a selection toward asparagine would drive the insertion of an initial seed enlarged by replication slippage in a multi-layered action of both mutational bias and selective pressure. Trying to furtherly understand this context , We started our analysis by dividing each CDS in two main cluster i.e. LCP, Low complexity containing proteins and nLCP, i.e. their counterpart, applying some classical bioinformatics tool as ENC (Effective Numebr of Codons) considering the improved versio of Sun and collegues (Sun et al. 2012) and introducing a new index that we call SPI (Selective Pressure Index) to compare CDS distance from the Wright expected curve making possible to compare proteins with the same ENC but different GC content in

the third codon position. Since protein length has been found to be a rather important factor in protein evolution, we also lead an analysis about this feature crossing our results with the above tools. Eventually, we find how mutational pressure acts as the priority body in shaping codon usage bias on LCRs containing proteins suggesting how LCRs are preferentially inserted where selective pressure is less important in structuring the Codon Usage Bias.

2 Methods

2.1 Data Sources and Filtering

The whole set of coding DNA sequences (CDS) and the corresponding proteins for each plasmodia were retrieved from PlasmoDB (https://plasmodb.org) and NCBI GenBank (ftp://ftp.ncbi.nih.gov), respectively. We consider only CDSs that start with the start codon (AUG), ending with a stop codon (UAG, UAA, or UGA), and having a multiple length of three. Incomplete and duplicated genes, with internal gaps or not recognized nucleotides were excluded.

2.2 Prediction of Low Complexity regions

To find LCRs we used SEG with standard parameters (Wotton Federhen,1994).

2.3 GC-content

The GC-content of a gene is the percentage of guanine and cytosine bases with respect to the total length of that gene. Similarly, it is possible to define the GC-content in the first (GC1), second (GC2), and third (GC3) codon positions, as follows:

$$GC_{1,2,3} - content = \frac{C_{1,2,3} + G_{1,2,3}}{A_{1,2,3} + T_{1,2,3} + C_{1,2,3} + G_{1,2,3}}$$

2.4 ENC, Effective Number of Codons

We calculated the effective number of codon (ENC) to estimate the extent of the codon usage bias of plasmode genes. The values of ENC range from 20 (when just one codon is used for each amino acid) to 61 (when all synonymous codons are equally used for each amino acid) (Wright, 1990). Thus, the smaller is the ENC value the larger is the extent of codon preference in a gene. The ENC values of all individual plamsode genes (>300 bp) were calculated by using the improved implementation by Sun et al. (2012) . The codon table is separated in synonymous codon families. The 6-fold codon families (Leu, Ser, and Arg) were divided into 2-fold and 4-fold codon families. For each coding sequence, we quantify Fα, defined for each synonymous codon family α as:

$$F_\alpha = \sum_{k=1}^{m_\alpha} \left(\frac{n_{k_\alpha}}{n_\alpha}\right)^2$$

where m_α is the number of codons in the codon family α, $n_{i\alpha}$ with $i = 1, 2, \dots, m_\alpha$ is the number of occurrences of codon *i* of the codon family α in the coding sequence, and $n_\alpha = \sum_{k=1}^{m_\alpha} n_{k_\alpha}$. Finally, the gene-specific ENC is defined as:

$$ENC = N_s + \frac{K_2 \sum_{\alpha=1}^{K_2} N_\alpha}{\sum_{\alpha=1}^{K_2}(n_\alpha F_\alpha)} + \frac{K_3 \sum_{\alpha=1}^{K_3} N_\alpha}{\sum_{\alpha=1}^{K_3}(n_\alpha F_\alpha)} + \frac{K_4 \sum_{\alpha=1}^{K_4} N_\alpha}{\sum_{\alpha=1}^{K_4}(n_\alpha F_\alpha)}$$

where N_s is the number of codon families with a single codon (i.e., Met and Trp codon families) and K_m is the number of families with degeneracy m. The ENC-plot is a plot in which the ENC is the ordinate and the GC3 is the abscissa (Forcelloni & Giansanti, 2019). Depending on the action of mutational bias and natural selection, different cases are discernible. If a gene is not subject to selection, a clear relationship is expected between ENC and GC3:

$$\mathrm{ENC} = 2 + \mathrm{GC}_3 + \frac{29}{\mathrm{GC}_3^2 + (1 - \mathrm{GC}_3^2)^2}$$

Genes for which the codon choice is only constrained by mutational bias are expected to lie on or just below the Wright's theoretical curve. So, points that fall below the bell-shaped line suggest deviations from a null model of no-codon preferences (Novembre, 2002). Alternatively, if a particular gene is subject to selection, it will fall below the Wright's theoretical curve. In this case, the vertical distance between the point and the theoretical curve provides an estimation of the relative extent to which natural selection and mutational bias affect CUB.

2.5 Selective Pressure Index (SPI)

The vertical distance between the point and the Wright's theoretical curve provides an estimation of the relative extent to which natural selection and mutational bias affect CUB. The farther the distance, the stronger is the action of natural selection. Therefore, we propose the Selective Pressure Index (SPI), which is formally defined as follows:

$$\mathrm{SPI}(\mathrm{GC}_3) = \frac{\mathrm{ENC_w}(\mathrm{GC}_3)\text{-}\mathrm{ENC_p}(\mathrm{GC}_3)}{\mathrm{ENC_w}(\mathrm{GC}_3)\text{-}20}$$

ENCw is the ENC value of the Wright's theoretical curve for a determined GC3 value and ENCp is the ENC score for the protein taken in consideration for that GC3 content. The variation in the use of a codon can represent a more or less important shift of the balance between natural selection and mutational pressure depending on the content in GC. The SPI helps to compare the distances under the Wright's theoretical curve being a normalized index in the range of [0, 1] A SPI score of 0 represents a gene whose CUB will not be subject to selection, in a compatible way with its GC content. On the other hand, a SPI of 1 represents a gene where ther is a strong selection toward a very narrow pattern of codons. Proteins with the same CUB, but different content in GC, can be uniquely identified through their distance from the curve which will represent the shift from a situation where selective pressure acts as the priority body in determining codon patterns.

2.5 LCPs and nLCPs

To have a reference frame to compare evolutionary pressures we divided CDS sets in two branches, i.e. LCPs, proteins with a t least one LCR inside their chain and nLCPs that have none.

3 Results

3.1 ENC-plot analyses of plasmodia

An ENC-plot analysis was performed to estimate the action of natural selection and mutational bias on the codon usage of LCPs and nLCPs. In Fig. 1, we show the ENC-plots obtained for LCPs and nLCPs, together with best-fit curves obtained through a nonlinear fit of Wright's shape,

From Figure 1, it can be observed the the distribution corresponding to LCPs are is closer to Wright's theoretical curve. Similar observations are also found in the other case studies relating to the

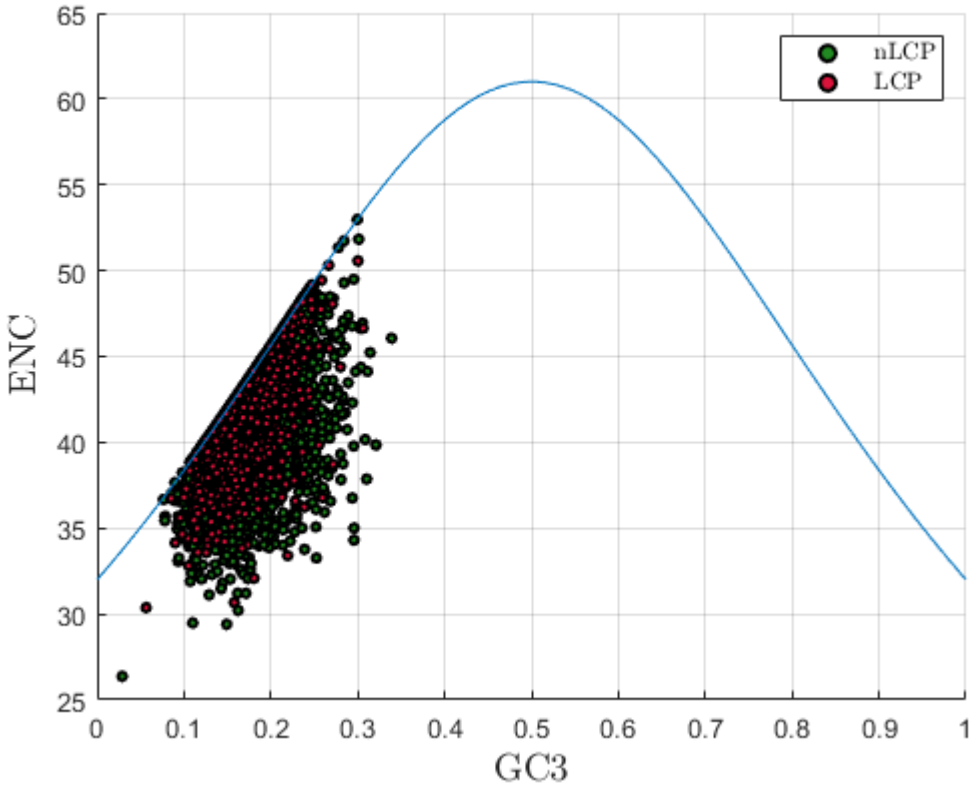


Figure 1 ENC plot forP.falciprum. In red Low Complexity Proteins (LCP). In green no tLow Complexity Proteins (nLCP). Each point represents a protein

remaining plasmodes. For the sake of simplicity, we then compared the ENC distributions of LCP and nLCP for each plasmode (welch t-test). The variety of Guanine and Cytosine Content among the plasmodium species allows to observe the evolutionary diversifications of these parasites. Looking at the boxplots in Fig.2 we observed a rather common and repetitive trait i.e. that LCP distributions tend to have higher ENC values than nLCPs. LCPs tend to be expressed with a greater number of codons, a first hint that mutational pressure acts as the main force in shaping codon patterns. Despite most of the pairwise comparisons are significant ($p < <0.001$) some of the plasmodes show controversies. P.chabaudi, P.yoeli and P.petteri comparisons are not significant ($p>0.05$). However, looking at their ENC plots LCPs tend always to stay closer to the Wright Expected curve so the non-

significancy could be traced back to the reduced number of low complexity domains spread into their CDS.

3.3 SPI

The ENC (see methods) introduces a masking of the associated data to the mRNA transcripts as two of these, based on a first analysis implemented through ENC, although having the same ENC, could see their codon usage patterns originate from a very different action of the evolutionary pressures, since the content in guanine and cytosine is different. The SPI solves this problem because it weighs the action of selective pressure and mutational pressure based on the content of guanine and cytosine.

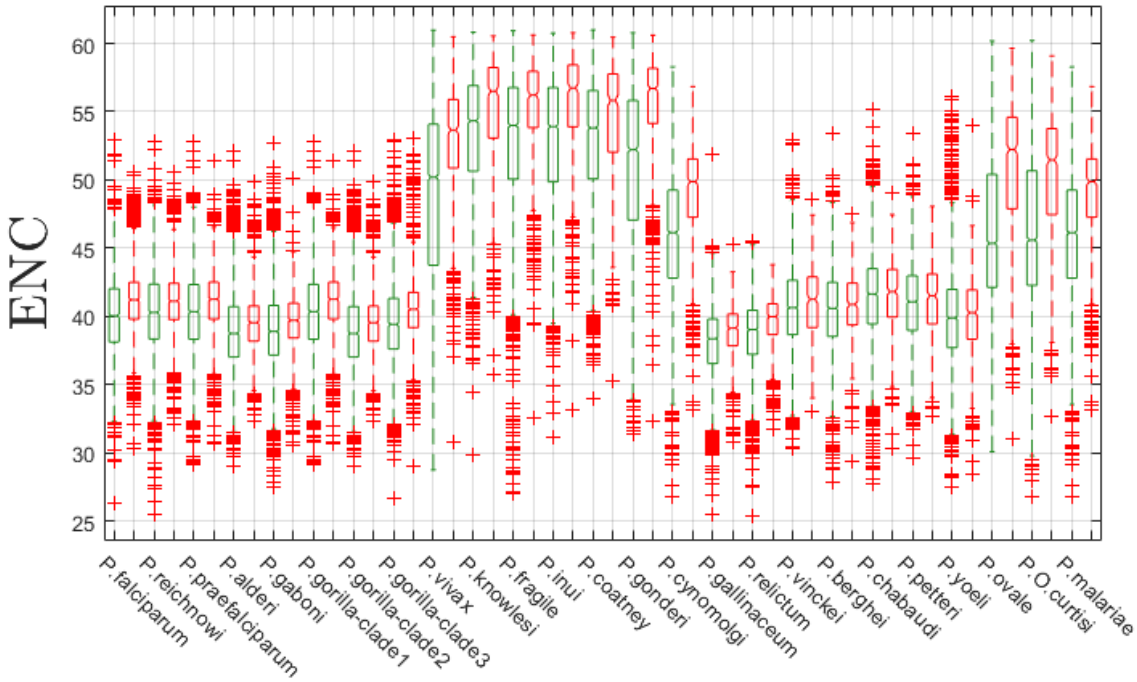


Figure 2 ENC plot forP.falciprum. In red Low Complexity Proteins (LCP). In greennot Low Complexity Proteins (nLCP).Pairwise comparisons do not differ significantly inP.berghei,P.chabaudi,P.petteriandP.yoeli.Matlab notch function has been activated:whenthe notches in the box plot do not overlap, you can conclude, with 95% confidence, that the true medians do differ. https://it.mathworks.com/help/stats/boxplot.html8

In figure 3 we carried out different pairwise comparisons (t-tests). LCP transcripts appear to be closer to Wright's theoretical curve which suggests a greater weight of the mutational pressure in determining codon usage patterns. All comparisons are significant ($p \ll 0.001$). Turning to the correlations between protein length (measured in nucleotides) with respect to the SPI, we consider Fig. 4 a-b, representative of P.falciparum. In both sets the SPI tends to decrease with increasing protein length. This observation is of extreme interest because a value of the SPI that tends to zero corresponds to a codon bias where the codon choice occurs indiscriminately in a compatible manner with the GC content of the transcript. This hooks up well with some studies(Wang et al.,2004;Li,1997;Akashi, 2003;Moriyama and Powell, 1998;Pall et al. 2001) that asserts how longer proteins are evolutionary newer than shorter ones. We find, as it can be seen later, that LCPs are the longest proteins in these organisms. So it is tempting to assert that longer proteins would accept more mutations as they are more nuanced from an evolutionary point of view. This lack of selective

constraint would permit to accept LCRs. So, if it is true that LCPs are evolutionary newer LCRs would serve as a "blank sheet" where evolution can conceive and shape new functions. As we can see from Tab 2 a similar situation also appear in the other plasmodes reinforcing the foregoing.

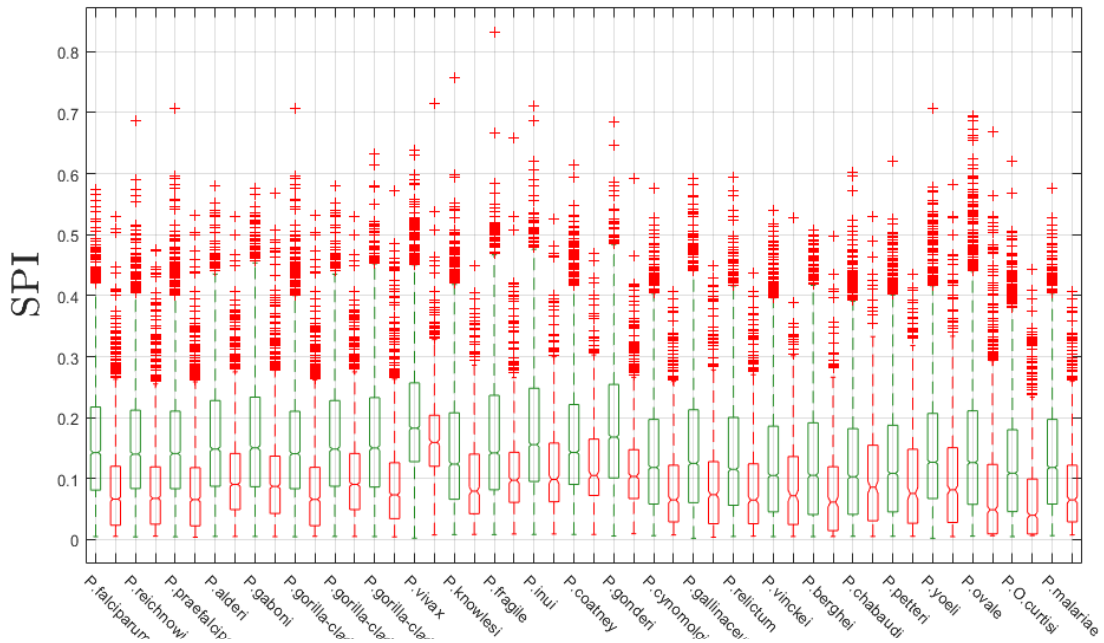


Figure 3 SPI for LCPs vs nLCP for each plasmodium. In red Low Complexity Proteins(LCP). In green not Low Complexity Proteins (nLCP)

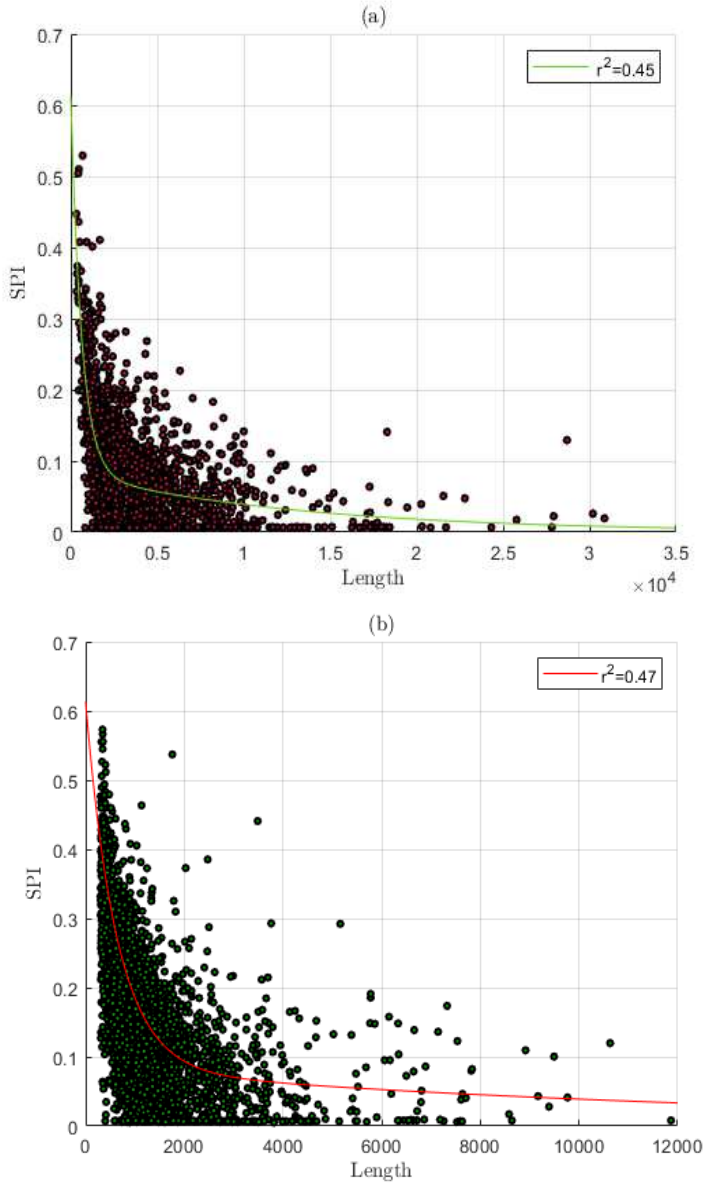


*Figure 4 SPI vs Length in LCPs in **P.falciparum**(b) SPI vs Length in nLCPs in P.falciparum. Length is measured by nucleotide count. The curve represents the best fit with a decreasing exponential function. Each point rep-resents a protein. The curves for the other plasmodes show similar trends, of which see the correlation coefficients in Table 1*

Table 1

Plasmodia	**nLCP r-square**	**LCP r-square**
P.falciparum	0.47	0.45
P.reichnowi	0.50	0.44
P.praefalcipaum	0.48	0.43
P.alderi	0.48	0.39
P.gaboni	0.49	0.44
P. gorilla clade 1	0.48	0.43
P.gorilla clade 2	0.48	0.39
P.gorilla clade 3	0.50	0.44
P.vivax	0.49	0.47
P.knowlesi	0.72	0.67
P.fragile	0.72	0.75

P.inui	0.73	0.74
P.coatney	0.69	0.63
P.cynomolgi	0.63	0.66
P.gonderi	0.61	0.49
P.gallinaceum	0.51	0.45
P.relictum	0.52	0.50
P.vinckei	0.61	0.54
P.berghei	0.59	0.52
P.chabaudi	0.60	0.53
P.petteri	0.62	0.63
P.yoeli	0.48	0.47
P.ovale	0.64	0.67
P.o.curtisi	0.60	0.52
P.malariae	0.73	0.72

3.1 LCP are longer than nLCP

We eventually analyzed the protein length distributions of the two groups of proteins: LCP and nLCP (Fig. 5). We compared LCPs and nLCPs length distributions (welch t-test). Each comparison was significant ($p < 0.0001$)

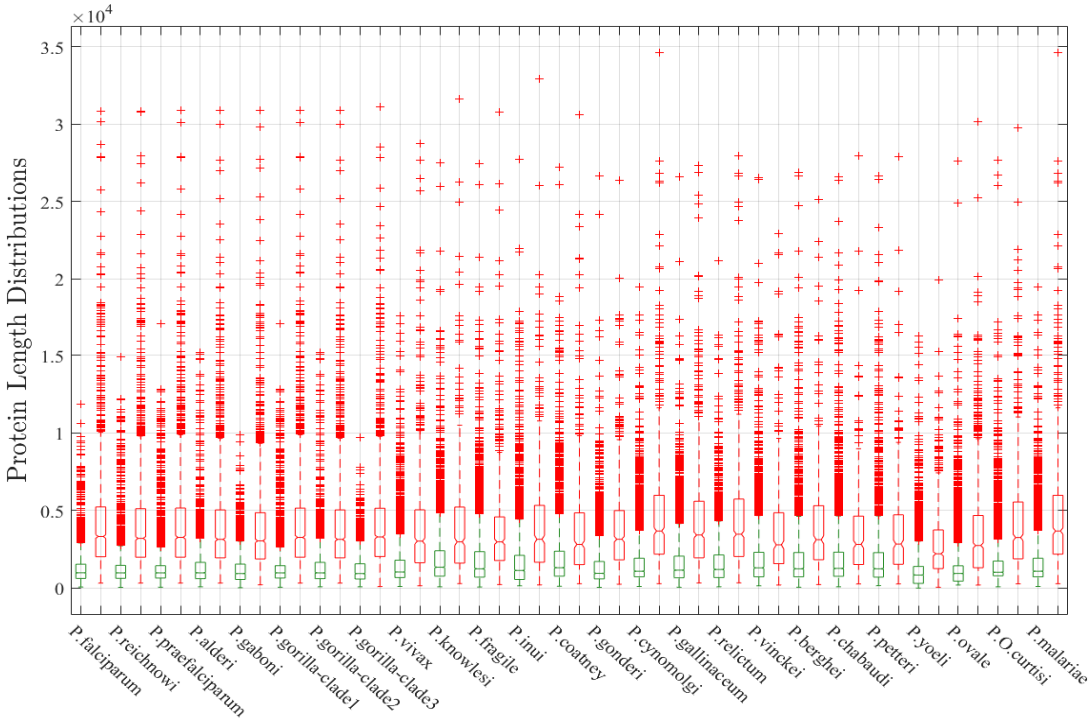


Figure 5 In this LCP and nLCP protein length distributions are represented. LCP arerepresented in red while nLCP are represented in green

LCPs are systematically longer than nLCPs. However, LCP are not as rich as P.falciparum in LC content in their entirety. Looking at Table 2 we can observe that LC content is not so sotrongly

correlated to protein length. So, length seems to be a standing alone characteristic suggesting that LCP seems to be longer compared to nLCP despite the LC content.

The pairwise comparisons result to be interesting because protein length is a rather important factor in protein evolution. Interestingly, it has been reported how newly evolved proteins are, on average, longer than older ones (Wang et al.,2004). This affirmation correlates well with other studies that asserts how recently evolved proteins seems to contain a greater number of Low complexity residues (Toll-Riera et al.,2012). Given that, and given the puzzling situation of LCPs found to be longer of nLCPs comprehended the previous results of CUB related to ENC plot and SPI, we speculate that LCPs being relatively newer could represent a signal of evolutionary and adapting path for these parasites.

Plamsodia	ρ
P.falciparum	0.7
P.reichnowi	0.71
P.praefalcipaum	0.71
P.alderi	0.67
P.gaboni	0.7
P. gorilla clade 1	0.65
P.gorilla clade 2	0.60
P.gorilla clade 3	0.69
P.vivax	0.35
P.knowlesi	0.26
P.fragile	0.32
P.inui	0.42
P.coatney	0.26
P.cynomolgi	0.22
P.gonderi	0.33
P.gallinaceum	0.38
P.relictum	0.22
P.vinckei	0.10
P.berghei	0.26
P.chabaudi	0.47
P.petteri	0.37
P.yoeli	0.45
P.ovale	0.43
P.o.curtisi	0.41
P.malariae	0.6

Discussion

The analysis of the evolutionary pathways of proteins is of fundamental importance for molecular evolution, comparative genomics and structural biology (Pal et al., 2016). So, in this study we focused on ENC extending the classic results obtainable from this tool through the SPI leaving other approaches (i.e. Neutrality and Pr2

plot (Sueoka,1988; Sueoka,1995; Sueoka, 1999) to further analysis. The ENC has been implemented in line with what previously done by Sun and colleagues (Sun et al., 2012) where the codon usage patterns for each protein calculated by the ENC are restricted between 61 and 20 codons correcting the typical fluctuations commonly found (Wright, 1990) due to the failure to divide the six-codon amino acid families into two and four-codon sub-branches (Sun et al., 2012) which can be selected in an extremely different way. To have a reference frame to see the equilibrium between selective pressure and mutational bias we divided the proteomes of each plasmodium in two branches, i.e. the LCPs (Low Complexity Proteins) or proteins with at least one low comlexity domain in their amino acid chain and the nLCPs, proteins that do not contain any of these stretches. Furthermore, interested in the relationship between LC and protein length reported in various other works, we decided to explore this feature. This last part of our work has revealed that there is no strong correlation between low complexity and protein length in all the plasmodes and this is a more marked characteristic of the plasmodes evolutionarily closer to P.falciparum. However, in an extremely interesting fashion we observed how LCPs are the longest proteins within each organism we considered, despite the low complexity content being extremely varied and profoundly different depending on the organism. Using the ENC we had a first glimpse of how evolution had weighed the two evolutionary forces .LCPs are expressed with a greater number of codons with respect to nLCPs, the first sign of a greater interference of the mutational pressure on this portion of the proteome. As previously suggested, two proteins with the same codon bias, however, based on the distance from the Wright curve, can represent two cases where evolution has operated in a profoundly different way in the choice of codons subverting the balance between mutational and selective pressure. The SPI helped us to solve this problem eliminating the masking introduced by the ENC. This analysis sees the LCPs, closer to the Wright curve. Considering also the correlation between SPI and protein length we see the longer the protein the lower the SPI. Some studies have reported that new proteins from an evolutionary point of view contain a greater content in Low Complexity (Toll-Riera et al., 2012; Ohno Epplen, 1983; Ohno Epplen 1984) and how these can favor the development of new domains and protein functions (Green wang, 2004). Another interesting observation is that these low complexity regions do not destroy the phenotype of proteins by being expressed between functional domains and not within them (Pizzi & Frontali, 2001;Frugier et al., 2009).Some studies show that the increase in protein length would increase the energy cost of biosynthesis (Wang et al. 2004) seeing the selective pressure favoring shorter proteins to increase the effectiveness of the synthesis (Akashi, 2003).Furthermore, the insertion of peptides can increase the stability of proteins and their functions (Wang et al 2004). More important is what has been expressed by Gemayel and colleagues (Gemayel et al. 2002) where the mutability of the low complexity traits would be compromised even by a few nucleotide mutations that would stabilize the Low Complexity trait by preventing or lowering the mutability of the same.So the purer the trait greater its ability of lengthening or shrinking. Other results show that the newest proteins from an evolutionary point of view would have more low complexity residues (Toll-Riera et al 2012; Ohno and Epplen 1983-1984). Our hypothesis is that the low complexity traits are inserted where there is less selectivity in the use of codon patterns allowing more frequent mutations. So, the lower the selection the more the LC insertion. This leads us to make some considerations: the accumulation of mutations would lead to the stabilization of the low complexity traits and the consequent formation of new protein domains. The insertion of the traits in proteins longer and less subject to selective pressure would serve to develop new protein functions. Overall, the Low Complexities would contribute to the adaptability and evolution of these organisms.

Conclusion

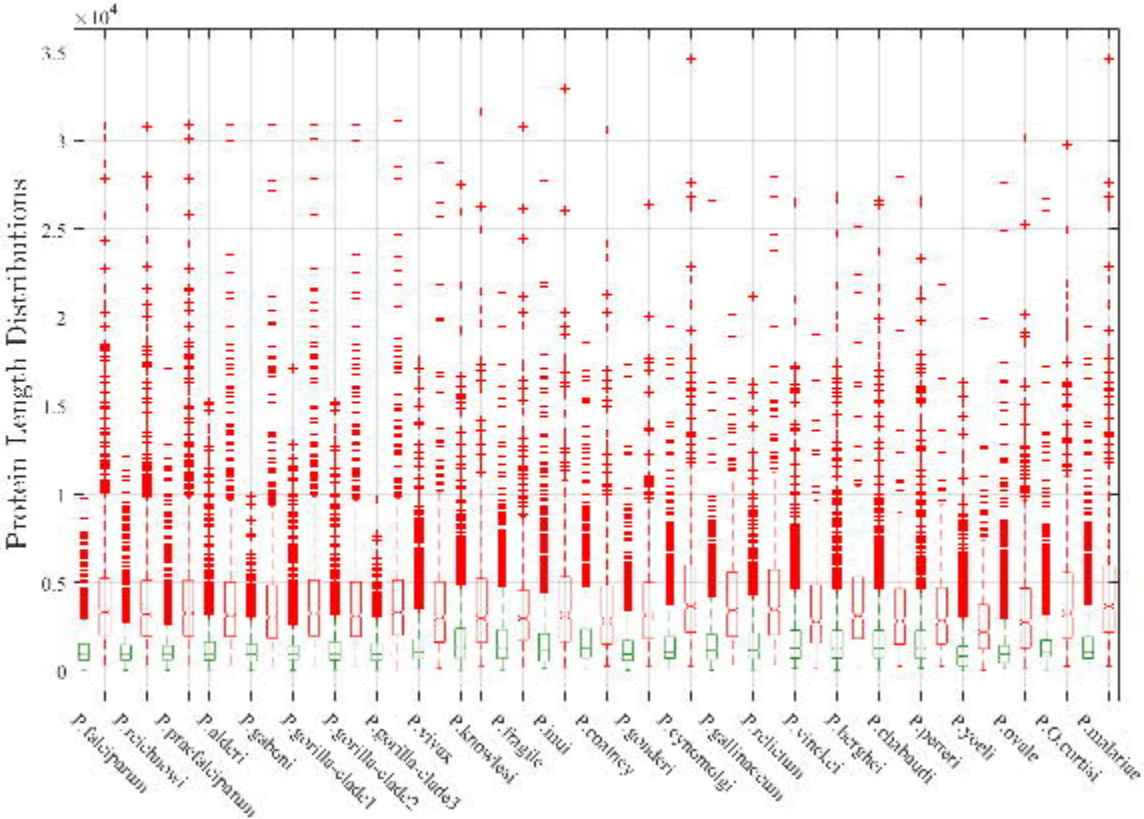
In this study, we performed a systematic analysis of the evolutionary pressures (natural selection and mutational bias) that shape synonymous codon usage patterns of the plasmodial genes focusing on ENC analysis and extending the classical results obtainable from this tool through the definition of the Selective Pressure Index. Firstly, we observed that LCPs portion of the proteome represents the longest portion of the proteome itself. Secondly, LCPs are nearer the expected curve of Wright than nLCPs. Importantly, we found that low complexity stretches tend to appear in longer protein. Another important feature deriving from our

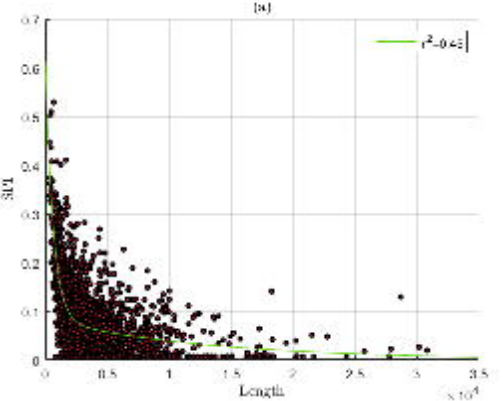
analysis is that the equilibrium between selective pressure and mutational bias shifts toward the latter for what concerns codon usage patterns suggesting a crucial role of Low Complexity for evolutionary adaptability and protein evolvability.

References

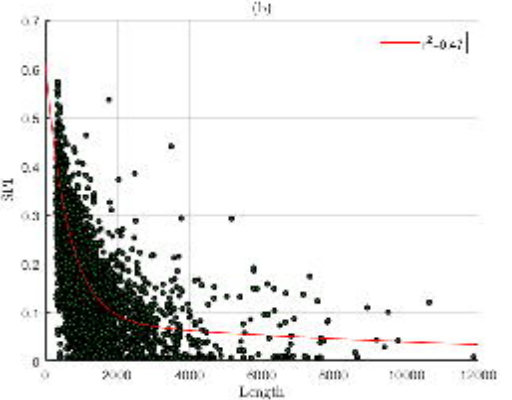
- Akashi H, Translational selection and yeast proteome evolution. Genetics 164: 1291-1303
- Battistuzzi FU, Schneider KA, Spencer MK, Fisher D, Chaudhry S, Escalante AA. Profiles of low complexity regions in Apicomplexa. BMC Evol Biol. 2016;16:47. Published 2016 Feb 29. doi:10.1186/s12862-016-0625-0
- Belalov IS, Lukashev AN (2013). Causes and implications of codon usage bias in RNA viruses. PLoS ONE 8:e56642 Brown LY, Brown SA. Alanine tracts: the expanding story of human illness and trinucleotide repeats, Trends Genet. , 2004, vol. 20 (pg. 51-58)
- Cobb DW, Florentin A, Fierro MA, Krakowiak M, Moore JM, Muralidharan V. The Exported Chaperone PfHsp70x Is Dispensable for the Plasmodium falciparum Intraerythrocytic Life Cycle. mSphere. 2017;2(5):e00363-17. Published 2017 Sep 27. doi:10.1128/mSphere.00363-17
- D. M., Koulov, A. V., Balch, W. E., Kelly, J. W. (2007). Functional amyloid – from bacteria to humans. Trends in Biochemical Sciences, 32(5), 217–224. doi:10.1016/j.tibs.2007.03.003 14
- Ellegren H. Microsatellites: simple sequences with complex evolution, Nat Rev Genet. , 2004, vol. 5 (pg. 435-445)
- Filisetti, D., Th´eobald-Dietrich, A., Mahmoudi, N., Rudinger-Thirion, J., Candolfi, E., Frugier, M. (2013). Aminoacylation of Plasmodium falciparum tRNA Asn and Insights in the Synthesis of Asparagine Repeats. Journal of Biological Chemistry, 288(51), 36361–36371. doi:10.1074/jbc.m113 .522896
- Forcelloni, S., Giansanti, A. Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. J Mol Evol (2019) doi:10.1007/s00239-019-09921-4 Fowler,
- Frugier, M., Bour, T., Ayach, M., Santos, M. A. S., Rudinger-Thirion, J., Th´eobald-Dietrich, A., Pizzi, E. (2009). Low Complexity Regions behave as tRNA sponges to help co-translational folding of plasmodial proteins. FEBS Letters, 584(2), 448–454. doi:10.1016/j.febslet.2009.11.004
- Fuglsang A (2006) Accounting for background nucleotide composition when measuring codon usage bias: brilliant idea, difficult in practice. Mol Biol Evol 23:1345–1347
- Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles, Nat Rev Genet. , 2005, vol. 6 (pg. 743-755)
- Green H, Wang N. Codon reiteration and the evolution of proteins. Proc Natl Acad Sci U S A. 1994;91(10):4298–4302. doi:10.1073/pnas.91.10.4298
- Halfmann R, Alberti S, Krishnan R, Lyle N, O'Donnell CW, et al. (2011) Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. Mol Cell 43: 72–84.
- John Federhen, Scott. (1993). Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. Computers Chemistry. 17. 149- 163. 10.1016/0097-8485(93)85006-X. Xiao H, Jeang KT. Glutamine-rich domains activate transcription in yeast Saccharomyces cerevisiae, J Biol Chem. , 1998, vol. 273 (pg. 22873-22876)
- Kashi, Y., King, D. (2006). Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics, 22(5), 253–259. doi:10.1016/ j.tig.2006.03.005
- Kumari, B., Kumar, R., Kumar, M. (2015). Low complexity and disordered regions of proteins have different structural and amino acid preferences. Molecular BioSystems, 11(2), 585–594. doi:10.1039/c4mb00425f

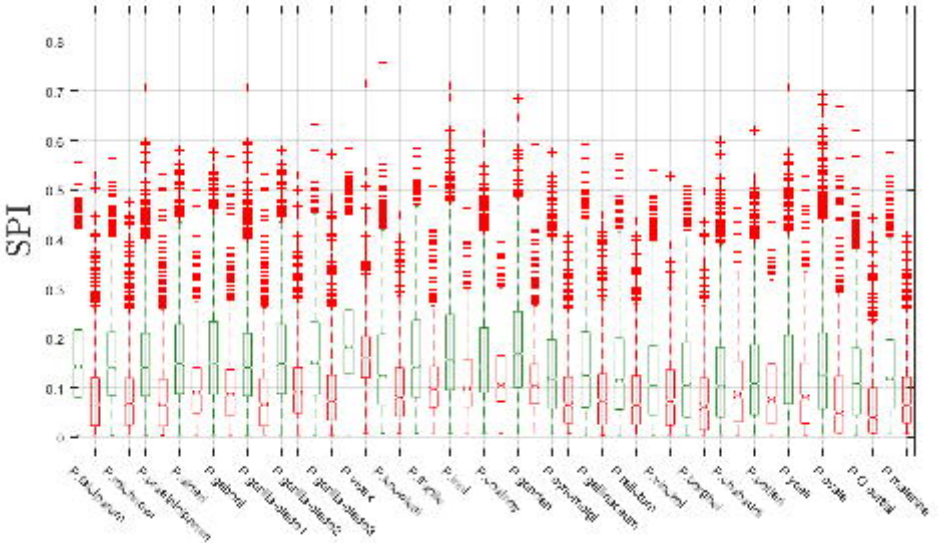
- Moriyama, E. N., and J. R. Powell. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. Nucleic Acids Res. 26:3188–3193.
- Muralidharan V, Goldberg DE. Asparagine repeats in Plasmodium falciparum proteins: good for nothing?. PLoS Pathog. 2013;9(8):e1003488. doi:10 .1371/journal.ppat.1003488
- Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19:1390–1394
- Ohno, S. (1984). Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings of the National Academy of Sciences, 81(8), 2421–2425. doi:10.1073/pnas.81 .8.2421
- Ohno, S., Epplen, J. T. (1983). The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. Proceedings of the National Academy of Sciences, 80(11), 3391–3395. doi:10.1073/pnas.80.11.3391
- Pal, C., B. Papp, and L. D. Hurst. 2001. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol. Biol. Evol. 18:2323–2326
- Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 2008;72(4):686–727. doi:10.1128/MMBR.00011-08 15
- Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657
- Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325
- Sueoka N (1999) Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. Gene 238:53–58
- Toll-Riera, M., Rado-Trilla, N., Martys, F., Alba, M. M. (2011). Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. Molecular Biology and Evolution, 29(3), 883–886. doi:10.1093/ molbev/msr263 Wootton

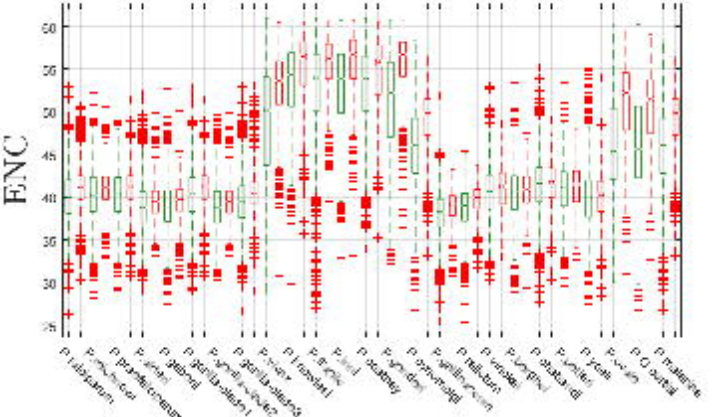

Protein Length Distributions
×10^4
3.5
3
2.5
2
1.5
1
0.5
0
P.falciparum
P.reichenowi
P.praefalciparum
P.alderi
P.gaboni
P.gorilla-clade1
P.gorilla-clade2
P.gorilla-clade3
P.vivax
P.knowlesi
P.fragile
P.inui
P.coatney
P.gonderi
P.cynomolgi
P.gallinaceum
P.reichenowi
P.vinckei
P.berghei
P.chabaudi
P.berrori
P.yoelii
P.ovale
P.O.curtisi
P.malariae

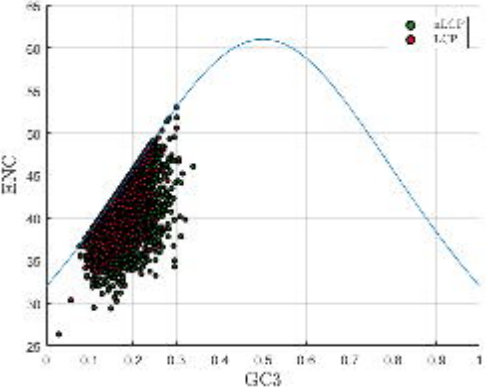


(a)
SPL
Length
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.5
1
1.5
2
2.5
3
3.5


(b)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SDI
0
2000
4000
6000
8000
10000
12000
Length


SPI
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


ENC
60
55
50
45
40
35
30
25
p_labiatum
p_malmquistii
p_parviflorum
p_pinetorum
p_pubescens
p_guttatum
p_montanum
p_montanum_var_rupicola
p_guttatum_var_yukonense
p_yukonense
p_fissipes
p_rosulatum
p_fuscum
p_stamineum
p_cypripedium
p_macranthum
p_spectabilis
p_passerinum
p_fallax
p_vivianae
p_calceolus
p_parviflorum
p_shastense
p_ventricosum
p_reginae
p_ram
p_c_occidentale
p_californicum


aLCP
LCP
ENC
GC3
65
60
55
50
45
40
35
30
25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1