

Biological Appraisal of the new Pfam domains detected in *Plasmodium falciparum* proteins based on co-occurrence with Pfam or other Interpro domains

We conducted a biologic appraisal of all *P. falciparum* sequences in which new Pfam domains have been detected. Our method being based on domain co-occurrence in sequences of well annotated genomes, the information gained by the detection of new Pfam in the *Plasmodium* proteins, assessed by the corresponding GO annotations, shows a strong level of biological consistency.

For previously annotated *Plasmodium* sequences, the new domains and the GO annotations they provided always confirm or enrich the initial functional inference of the genes. For instance, the annotation of PF10_0122 as a putative phosphoglucomutase, initially supported by the detection of the InterPro domain SSF53738, G3DSA:3.40.120.10 and the Pfam domain PF02878 (“PGM_PMM_I”), is now abundantly supported by the detection of the Pfam domains PF02879 (“PGM_PMM_II”), PF02880 (“PGM_PMM_III”) and PF00408 (“PGM_PMM_IV”). No contradiction has been noticed in the whole set of certified domains.

Some initial annotations are substantially improved. Thus, MAL13P1.83 had been annotated as a karyopherin in version 5.5 of PlasmoDB, based on sequence alignments, and contained only one domain with a GO annotation, *i.e.* the InterPro domain SSF48371, associated to the molecular function term GO:0005488, corresponding to “binding”. This GO term is poorly informative since it is defined as the “interaction with one or more specific sites on another molecule”. MAL13P1.83 also contains a Pfam domain without any GO annotation, PF08389, which was therefore initially non-informative. Interestingly, based on PF08389, we detected the co-occurring PF03810 domain, which has a rich GO annotation, *i.e.* GO:0008565, “protein transporter activity”, GO:0000059, “protein import into nucleus, docking”, GO:0006886, “intracellular protein transport”, GO:0005634, “nucleus”, GO:0005643, “nuclear pore” and GO:0005737, “cytoplasm”. Thus, the *P. falciparum* karyopherin, a transporter of the nuclear envelope, is now richly detailed at the level of domain annotation, which now allows the retrieval of this sequence when mining *Plasmodium* data based on protein domain descriptions. Likewise, MAL7P1.91, annotated as an “exported serine/threonine kinase” based on the occurrence of PS50011, SSF56112 and PF00069 that support a protein kinase activity, can now be ascribed a novel possible molecular activity, *i.e.* an involvement in protein ubiquitination, based on the detection of PF04564, or “U-box”, co-occurring with the PS5001 and PF00069 domains. This novel Pfam domain brings new GO annotations to the protein: GO:0000151 “ubiquitin ligase complex”, GO:0004842 “ubiquitin-protein ligase activity” and GO:0016567 “protein ubiquitination”.

In some cases, unsuspected new function can also be hypothesized for some previously annotated proteins. For example, MAL7P1.12 is known as an erythrocyte membrane-associated antigen (Kun et al., 1991¹). It initially contained three InterPro domains with no GO annotations, SSFS2540, SM00487 and G3DSA:3.40.50.300 but none Pfam domains. Based on the co-occurrence with SM00487, we detected the PF00035 or “dsrm” Pfam domain, which has been ascribed the GO term GO:0003725, for “binding to double stranded RNA”. Co-occurring with the dsrm domain, we further detected the PF04851 or “ResIII” Pfam domain, annotated with the following GO terms: GO:0003677, “DNA binding”, GO:0005524, “ATP binding”, and GO:0016787, “hydrolase activity”. The detection of these novel domains suggests that MAL7P1.12 might be involved in some of the very diverse, but essential cellular processes regulated by double stranded RNA (dsRNA), including dsRNA regulation/signalling and/or defense mechanism against pathogens, like viruses, implying ARN degradation, or else control of RNA levels in the parasite cell during its life cycle (Saunders and Barber, 2003²). The particular detection of a domain binding dsRNA (“dsrm”) and a domain possibly involved in a hydrolase activity (“ResIII”) suggests that MAL7P1.12 might cleave dsRNA into smaller fragments, possibly generating short interfering RNAs (siRNAs) and / or microRNAs (miRNAs). This would infer MAL7P1.12 part of the molecular function of the eukaryotic Dicer protein (Jaskiewicz and Filipowicz, 2008³), which generates siRNAs and miRNAs and helps load these fragments into the RNA-induced silencing complex (RISC). No Dicer or RISC proteins have been detected in *Plasmodium* genome so far, and a recent study by Xue et al. (2008)⁴ further supports that no miRNA is produced in this parasite. MAL7P1.12 might therefore have a close but distinct role, possibly cleaving some specific dsRNAs. In the current view of the regulation of the *Plasmodium* genome expression at the level of RNAs (including a possible tuning of the RNA decay) rather than at the transcriptional level, the detection of PF00035 and PF04851 Pfam domains in an erythrocyte membrane-associated antigen raises therefore puzzling questions regarding the function of this protein, which should be now examined carefully in the context of RNA control.

For *Plasmodium* sequences listed as hypothetical proteins in PlasmoDB, or very poorly annotated, two major cases can be observed based on the level of annotation of the newly detected domains.

On the one hand, the new domains might be ascribed no or poorly informative GO terms. In this case, no precise functional annotation can be deduced from the present work,

¹ Kun J, Hesselbach J, Schreiber M, Scherf A, Gysin J, Mattei D, Pereira da Silva L, Müller-Hill B. Cloning and expression of genomic DNA sequences coding for putative erythrocyte membrane-associated antigens of *Plasmodium falciparum*. *Res Immunol*. 1991 Mar-Apr;142(3):199-210.

² Saunders LR, Barber GN. The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J*. 2003 Jun;17(9):961-83.

³ Jaskiewicz L, Filipowicz W. Role of Dicer in posttranscriptional RNA silencing. *Curr Top Microbiol Immunol*. 2008;320:77-97.

⁴ Xue X, Zhang Q, Huang Y, Feng L, Pan W. No miRNA were found in *Plasmodium* and the ones identified in erythrocytes could not be correlated with infection. *Malar J*. 2008 Mar 10;7:47.

but the structural categorisation of the proteins is refined, providing clues for future functional inferences. For instance, the “WD40” or PF00400 domain is found in all eukaryotes, in proteins implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis. Based on the PF00400 description, the repeated WD40 motifs act as a site for protein-protein interaction, and proteins containing WD40 repeats are known to serve as platforms for the assembly of protein complexes or mediators of transient interplay among other proteins. The specificity of the proteins is determined by other domains occurring in the sequence, outside the WD40 repeats. The WD40 domain was initially reported in 63 *Plasmodium* proteins and it is now detected in 12 additional sequences, making the family of WD40-containing proteins, the third largest domain containing protein family in the *P. falciparum* genome. Combination of WD40 with other domains allows a classification of the corresponding proteins, and might ease the annotation of this family in future works: it is for example possible to define domain combinations in hypothetical proteins, like the LisH/WD40 combination in MAL13P1.54, PFE0540w, PFE0930w and PFE0930w that differs from the LisH/RanBPM combination in MAL13P1.182 and MAL13P1.308. Likewise, the “DEAD” or PF00270 domain, initially reported in 42 proteins, is now detected in 11 additional sequences. Occurrence of a DEAD domain indicates that the protein is likely involved in various aspects of RNA metabolism, including nuclear transcription, pre mRNA splicing, ribosome biogenesis, nucleocytoplasmic transport, translation, RNA decay and organellar gene expression. The precise function of the protein cannot therefore be speculated based on the detection of a single DEAD domain, and the combination with other domains will therefore be essential for future analyses.

On the other hand, newly detected domains can be informative enough to allow the proposition of an annotation for a hypothetical or poorly annotated protein. Three situations can be envisaged, depending on previously known domains of these proteins:

- Firstly, proteins can be initially devoid of any InterPro or Pfam domains. For instance, PF14_0380, which is recorded as a hypothetical protein in PlasmoDB 5.5, is possibly a subunit of the cohesion complex involved in chromatid segregation based on the new Pfam domains PF04824 “Rad21_Rec8” and PF04825 “Rad21_Rec8_N”. Likewise, PFF0910c, MAL13P1.78 and PFF1045w are possibly proteins kinases (PF06743, which has been ascribed the corresponding GO annotation GO:0004672, and PF08373 domain).

- Secondly, proteins can initially contain informative Pfam domains yet still annotated by “unknown function”. Several unannotated proteins possess informative Pfam domains which have not been taken into account yet. In this case we discover domains confirming previously known domains and hypothetical function can thus be inferred with a high degree of confidence. This is particularly the case for PFF1490w, where a PF02882 domain (“THF_DHG_CYG_C”) is detected by Pfam recommended threshold and allows to certify the PF00763 domain (“THF_DHG_CYG”) supporting a tetrahydrofolate dehydrogenase/cyclohydrolase, a new enzyme of the folate metabolism in *P. falciparum*, or in PF14_0052 where certification of the PF07683 (“CobW_C”) domain confirms PF02492 (“CobW”) and indicates that PF14_0052 might be implied in the synthesis of cobalamin

(vitaminB12), a molecule necessary to the parasite development in very small amounts, but that can display anti-malarial properties in larger amounts.

- Thirdly, proteins can initially contain Interpro domains providing little information. In this category, PF10_0040 is possibly a nuclease involved in DNA repair (based on detection of PF00867 and PF00752 domains); PF10_0152 a nucleotidyltransferase (PF01909 and PF03828 domains); PF11_0244 an ATP-dependent protease (PF02190 domain); PF11_0276 a lipase (PF00561 domain); PF11_0368 a transporter of the nuclear envelope (PF03810 domain), PF11_0375 a signal recognition particle related protein (PF08492 domain), PF11_0469 a DNA-dependent RNA polymerase (PF08221 and PF05645 domains); PF14_0031 a Ca²⁺-binding protein involved in vesicular trafficking (PF00637, PF00008 and PF07645 domains); and finally a very interesting domain certified in PF14_0479 is the PF05605 domain (“Di19”), since it is extremely specific to the plant kingdom, associated to the response of plants to drought, and likely acquired in *P. falciparum* following the ancestral endosymbiosis with an alga at the origin of the plant-like features in Apicomplexa.

All function that could be proposed cannot be listed here but we noticed that many of them revealed protein involved in chromatin interaction (such PFF1385w, PFL0975w or PF07_0106) and numerous transcription factor associated proteins which are of a particular interest for future investigation considering the apparent lack of such proteins in initial studies⁵. We did not list either all the hypothetical proteins which have been inferred a function after the 2008 reannotation workshop, since all new domains were consistent with the revised annotations. Based on this work, novel potential targets for antimalarial therapeutics might be envisaged, such as the putative tetrahydrofolate dehydrogenase/cyclohydrolase.

Interestingly, we did not add a single new sequence to the largest family of domain-containing proteins, *i.e.* proteins with the Rifin_STEVOR domain, and very little changes to Duffy binding- and PFEMP-containing protein families. These domains are all defined based on *Plasmodium* surface antigens, and consistently, our method did not allow any advance in their detection. It is noticeable that the families of proteins containing domains related to RNA binding, modification and/or processing (Helicase_C, RRM, DEAD) are amongst the largest in *Plasmodium* genome. It also appears that the domains involved in protein-protein interaction, *e.g.* WD40, together with TPR_1 (initially identified in 15 sequences, now in 28 proteins) or TPR_2 (initially identified in 8 sequences, now in 24 proteins), are also detected in large families of proteins. The present work allows therefore an in-depth analysis of these families, with an improved genomic coverage.

⁵ Coulson RMR, Hall N, Ouzonis A. Comparative Genomics of Transcriptional Control in the Human Parasite *Plasmodium falciparum*, Genome Research, 2004; 14:1548-1554

Callebaut I, Prat K, Meurice E, Mornon JP et Tomavo S. Prediction of the general transcription factors associated with RNA polymerase II in *Plasmodium falciparum*: conserved features and differences relative to other eucaryotes, BMC Genomics, 2005; 6:100