Detection of new protein domains using co-occurrence: application to Plasmodium falciparum

Nicolas Terrapon, Olivier Gascuel, Eric Maréchal, Laurent Brehelin

BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–7

Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*

Nicolas Terrapon [1,2], Olivier Gascuel [1], Éric Maréchal [2] and Laurent Bréhélin [1,*]

[1]Méthodes et algorithmes pour la Bioinformatique, LIRMM, Univ. Montpellier 2, CNRS, 161 rue Ada 34392 Montpellier Cedex 5 France
[2]CEA Grenoble iRTSV/LPCV, 17 rue des Martyrs, 38054 Grenoble cedex 9 France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Hidden Markov Models (HMMs) have proved to be a powerful tool for protein domain identification in newly sequenced organisms. However, numerous domains may be missed in highly divergent proteins. This is the case for *Plasmodium falciparum* proteins, the main causal agent of human malaria.

Results: We propose a method to improve the sensitivity of HMM domain detection by exploiting the tendency of the domains to appear preferentially with a few other favorite domains in a protein. When sequence information alone is not sufficient to warrant the presence of a particular domain, our method enables its detection on the basis of the presence of other Pfam or InterPro domains. Moreover, a shuffling procedure allows us to estimate the false discovery rate associated with the results. Applied to *P. falciparum*, our method identifies 585 new Pfam domains (versus the $3\,683$ already known domains in the Pfam database) with an estimated error rate below 20%. These new domains provide 387 new Gene Ontology annotations to the *P. falciparum* proteome. Analogous and congruent results are obtained when applying the method to related *Plasmodium* species (*P. vivax* and *P. yoelii*).

Availability: Supplementary Material and a database of the new domains and GO predictions achieved on *Plasmodium* proteins are available at http://www.lirmm.fr/∼terrapon/codd/

Contact: brehelin@lirmm.fr

1 INTRODUCTION

Among relevant annotations that can be attached to a protein, domains occupy a key position. Protein domains are sequential and structural motifs that are found independently in different proteins, in different combinations and, as such, seem to be functional sub-units of proteins above the raw amino-acid sequence level (Richardson, 1981). Several approaches have been developed to define and identify domains. Some are based on observed distinct structural classes of proteins (Murzin *et al.*, 1995). Others are inferred by clustering conserved subsequences (Mulder *et al.*, 2007; Finn *et al.*, 2008). One of the most widely used domain schemata is the Pfam database (Finn *et al.*, 2008). In this database, each domain family is defined with a set of distinct representative protein sequences, manually selected and aligned, and used to learn a *Hidden Markov Model* (HMM) (Durbin *et al.*, 1998) of the domain. The current release of the Pfam database (version 23.0 of July 2008) offers a large collection of $10\,340$ HMMs/domains, which account for over 73% of all proteins in the Uniprot database (UniProt Consortium, 2009). Some Pfam HMMs have been annotated by the InterPro consortium (Mulder *et al.*, 2007) in the *Gene Ontology* (GO) (Gene Ontology Consortium, 2000). According to the InterPro annotation policy, a domain is annotated with a given GO term if all proteins where this domain is known also share this GO term. This stringent rule allows, when a new domain is detected in a protein, transfer of its annotations to this protein.

When analyzing a new protein sequence, each Pfam HMM is used to compute a score that measures the similarity between the sequence and the domain. If the score is above a given threshold provided by Pfam (score thresholds differ depending on the HMMs), then the presence of the domain can be asserted in the protein. However, when applied to highly divergent proteins, this strategy may miss numerous domains. For example, with *Plasmodium falciparum*, the main causal agent of human malaria, no Pfam domains are detected in nearly 50% of its proteins, while many domain types seem to be missing from the *P. falciparum* library—see Supp. Table 1 for a comparison of the protein coverage and domain numbers between several eukaryotes. This can be partly explained by the highly atypical genome of *P. falciparum*, which is composed of above 80% A+T, and involves long low-complexity insertions of unknown function believed to form non-globular domains (Pizzi and Frontali, 2001). This strongly biases the amino-acid composition of *P. falciparum* proteome, in which six amino acids account for more than 50% of the protein composition (Bastien *et al.*, 2005). In this article, we propose a new method to increase the sensitivity of Pfam domain detection in divergent proteins like those of *P. falciparum*. Our method involves lowering the thresholds provided by Pfam for detecting domains. This enables more domain detections, but at the expense of numerous false positive predictions. The core of the method is a filter procedure based on domain co-occurrence properties which selects the potential domains that are most likely true.

*to whom correspondence should be addressed

Several authors have studied domain combinations in proteins (Apic *et al.*, 2001). They showed that numerous domains are frequently found together. As a result, the number of observed domain combinations in nature is clearly less than the number of possibilities. The domains tend to appear with a few other favorite domains. For example, when computing the number of distinct Pfam domain pairs observed in Uniprot proteins, only 20 000 out of the ∼12.5 million possibilities are observed (a ratio of 1.6‰). This property suggests functional cooperation. Indeed, two thirds of mono-domain proteins having the same domain also have the same functions. For multidomain proteins, 35% of proteins having one identical domain have similar functions, while this rate increases to 80% when they share two identical domains (Gerstein and Hegyi, 2001). This is the basis of approaches used to predict protein functional annotations. Scott *et al.* (2004) use Bayesian networks of co-occuring motifs to predict sub-cellular localizations of proteins. McLaughlin *et al.* (2007) characterize domain assembly (DASSEM) units, *i.e.* groups of domains that cooperate to achieve a given function. Forslund and Sonnhammer (2008) propose an approach to predict specific GO annotations from domain groups. Geer *et al.* (2002) present the CDART tool that allows users to find proteins having a domain composition similar to that of the query protein. In this paper, we propose to use domain co-occurrence to improve the sensitivity of Pfam domain detection.

In the following, we first describe our approach and present a shuffling procedure that allows estimation of its false discovery rate (FDR). Next, the method is assessed by searching for domains in artificial proteins obtained by simulating evolution events in the yeast proteins. Finally, our approach is applied to *P. falciparum*. In this organism, it detects 558 additional domains with an estimated FDR below 20%. Among these new domains, 159 domain types have never before been detected in *P. falciparum* proteins. Moreover, these domains provide 387 new Gene Ontology annotations. Analogous and congruent results are obtained when applying this method to related *Plasmodium* species (*P. vivax* and *P. yoelii*).

2 METHOD

In the following, the *domain composition* of a protein is the set of domains it contains. Thus, both the sequential order of the domains, as well as the number of times each domain occurs is ignored. This is done on the grounds of the assumption that the presence/absence of a domain is the prime information for deciphering the protein function (Cohen-Gihon *et al.*, 2007). For each protein, we distinguish two types of known domains: Pfam domains and InterPro non-Pfam domains. Known Pfam domains are all Pfam domains above the stringent score thresholds provided by Pfam (*"gathering cut-offs"*), or whose presence has been asserted by experts and can be found in dedicated databases of the query organism—for example, PlasmoDB for *P. falciparum* (Bahl *et al.*, 2003). The InterPro domains come from the InterPro database (Mulder *et al.*, 2007), a meta-database of different domain databases: PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, PANTHER and Pfam (see Mulder *et al.* (2007) for references). InterPro incorporates this knowledge into a single resource by organizing entries in *InterPro families* pooling all representations of the same domain. The known InterPro domains of a protein are inferred with InterProScan software, and can be found in the dedicated databases of the organisms.

We aim to enrich the known Pfam domains of a query organism (*P. falciparum*). The principle is to use domain co-occurrence properties to certify the presence of a new *potential* Pfam domain in a protein, thanks to the presence of another *validating* domain. To this end, all Uniprot proteins with known domain composition were used to extract domain pairs showing strong co-occurrence (as assessed by a statistical test). These domain pairs are stored in a reference list of Conditionally Dependent Pairs (*CDP*), which is then used as follows. Let us consider a protein of the query organism where one or more potential Pfam domains are detected by lowering the score thresholds of the Pfam HMMs. If one of these potential domains forms, along with another non-overlapping domain of the protein, a pair that belongs to the *CDP* list, then the presence of the potential domain is considered as *certified*. Hence, to apply this method we need a set of *validating* domains V_i and a set of *potential* domains P_i for each protein i of the query organism. We also have to infer the *CDP* list $\{(A, B), A \neq B\}$ from all proteins of known domain composition.

2.1 Set of potential domains

The sets of potential domains (P_i) are inferred from the results of Pfam HMM searches using *hmmer* software (Eddy, 1998). Given a set of proteins and an HMM, this tool computes a score that measures the similarity between each protein sequence and the domain modeled by the HMM. Additionally, this score can be used to compute an E-value estimate that represents the expected number of random sequences that would obtain a score above that achieved by the protein. Here, the set of potential domains of each protein is built by considering all HMM hits that differ from the already known Pfam domains and which have an E-value below a given permissive threshold (*e.g.* 10). This E-value threshold is chosen to be much less conservative than the thresholds recommended by Pfam for each HMM. The results are then processed for each protein to obtain a list of non-overlapping potential domains. The policy applied for this selection is to favor domains with the most significant (lowest) E-values. Overlaps with already known Pfam domains are forbidden. The same domain may appear several times in this non-overlapping domain list, but redundancies are removed to obtain the final set of potential domains P_i.

2.2 Set of validating domains

Different sets of validating domains (V_i) are considered here. The first solution is to use known Pfam domains of the protein. A complementary solution is to use InterPro non-Pfam domains already known in the protein. This allows us to increase the number of validating domains of each protein and thus the expected number of certifications. However, due to the heterogeneity of the InterPro database, the certifications achieved this way may be of lower quality than that achieved with Pfam domains. With these first two solutions, domains can be certified solely in proteins where at least one domain is already known. To overcome this limitation, a third solution is to consider the potential domains themselves as validating domains. In this solution, all pairs of potential domains of the protein are enumerated and, if the pair belongs to the CDP list, the two domains are certified. Of course this procedure is more prone to false positives than the two others, but we will see below how this can be controlled. Thus, the three types of validating domains are mutually exclusive and of decreasing quality *a priori*. Note that when certifying a potential domain, only the validating domains that do not overlap this domain are considered. In the experiments below, the three types of validating domains are used and tested independently.

2.3 Selecting the *CDP*s

The list of Conditionally Dependent Pairs is computed from the whole set of domain pairs observed in Uniprot proteins of all organisms but the query organism. Only pairs of *different* domains are considered in the CDP list, in order to avoid certifying one domain by itself. These pairs must reveal a conditional dependence between a Pfam domain and an InterPro (Pfam or non-Pfam) domain, that is, the presence of the InterPro domain has to be a strong clue of the presence of the Pfam domain. Testing the conditional

dependence of a domain pair involves measuring the association between two variables. This can be done with a correlation test like χ^2. Here we use a one-tailed Fisher's exact test to cope with small sample sizes. To this end, we compute, for each domain pair $(A, B), A \neq B$, the number x of proteins where both A and B are present, the numbers w (respectively y) of proteins where A is present but B is absent (resp. A is absent but B is present), and the number z of proteins where both A and B are absent. The probability of observing x or more proteins with the B domain in the set of $x+w$ proteins having the A domain, under the null hypothesis that domains are independent, is computed as a sum of hypergeometrical laws:

$$p\text{-}value_{(A,B)} = \sum_{t=0}^{\min(y,w)} P(x+t, y-t, w-t, z+t), \qquad (1)$$

with $P(\alpha, \beta, \gamma, \delta)$ being the probability of observing exactly α proteins with B among $\alpha+\gamma$ proteins, knowing that there are a total of $\alpha+\beta$ proteins with B among a total number of $\alpha+\beta+\gamma+\delta$ proteins. Hence, a *p-value* is computed with expression (1) for each domain pair. If this *p-value* is below a given threshold—1% in our experiments—the null hypothesis is rejected, the domains are considered as conditionally dependent, and the pair is added to the *CDP* list.

2.4 Estimating the number of false certifications

From the potential domains, the validating domains, and the *CDP* list, we can certify a certain number of new domains. One issue is then to estimate the proportion of false positives among these new domains. To this end, we estimate the probability of certifying a potential domain under the null hypothesis H_0 that it has been randomly predicted. This is done through computer simulations, by shuffling the potential domains of all proteins. This procedure randomly permutes the potential domains, while avoiding associating the same domain to a given protein more than once. This creates a situation where the potential domains are independent of the validating domains, while preserving the domain distribution and the number of validating and potential domains in each protein. The certification procedure is applied to the shuffled domains, and the number of random domains certified is computed. The entire procedure is resumed several times (typically 1000 times) to get a reliable estimate of the expected number of domains our procedure would certify under the hypothesis that all potential domains are random. This number is then used to compute an estimate of the False Discovery Rate (FDR) of the certification process with the formula

$$\widehat{FDR} = \frac{\text{expected number of certifications under } H_0}{\text{number of certifications on original data}}. \qquad (2)$$

This approach is similar to that proposed in Soriç (1989) and Benjamini and Hochberg (1995) to control the FDR associated with the multiple testing of several hypotheses. We shall see that by modulating the E-value threshold used to define the set of potential domains, we can control the FDR associated with our certifications through this procedure.

3 RESULTS

3.1 Method assessment

We first assessed the potential of the method to improve the Pfam sensitivity in divergent proteins. In this experiment, *Saccharamyces cerevisiae* was chosen for the quality of its annotations. The principle is as follows. Pfam HMMs were used with their score thresholds to determine a set of reference domains in yeast. All proteins with at least two Pfam domains were considered for the following steps. We simulated the evolution of these proteins using *seqgen* (Rambaut and Grassly, 1997) in order to change their global amino-acid composition and make it close to that of *P. falciparum*. *seqgen* was used with the *WAG* substitution rate matrix, but we replaced the default standard amino-acid composition by that of *P. falciparum* (PlasmoDB 5.5). Starting from any sequence and applying substitutions based on this modified matrix, we obtained a sequence with an amino-acid composition converging toward the *P. falciparum* distribution. Four different substitution rates were applied—0.1, 0.25, 0.5 and 0.75—yielding four sets of artificial protein sequences of increasing divergence from the original yeast proteins. For each of these sets, each HMM was used with the Pfam score thresholds to determine the set of validating domains. Some reference domains were no longer detected because of the sequence drift. The certification procedure was then applied with these validating domains to retrieve the lost domains.

Table 1. Results on yeast after sequence drift

Subst. Rate	Lost Dom.	Potent. Retr.	Retr. Dom.	FDR	New Dom.	Known GO
0.1	149	145	134	11.5%	274	97/130
0.25	346	301	265	9.2%	171	72/93
0.5	907	645	491	5.4%	60	20/31
0.75	1436	747	501	4%	12	7/12

"Subst. Rate": substitution rate applied to the sequences; "Lost Dom.": number of domains that are not found with Pfam thresholds in the drifted sequences; "Potent. Retr.": number of domains which might be retrieved (*i.e.* lost domains in proteins where at least one other domain is still detected by Pfam thresholds); "Retr. Dom.": number of lost domains retrieved by our approach; "New Dom.": number of certified domains that do not belong to the reference domain set; "Known GO": proportion of annotated new domains with annotations already known in the corresponding proteins.

Table 1 summarizes the results. For example, with a substitution rate of 0.5, 907 domains are lost; among these, 645 might be retrieved (*i.e.* belong to a protein where at least another domain is still detected using the Pfam thresholds), and 491 (76%) are actually retrieved. Moreover, 60 new domains (absent from the original reference set) are also certified (with a small 5.4% FDR). The number of new domains is even higher for lower substitution rates, and may appear surprisingly high in a well annotated organism like yeast. This questions the validity of these new domains. Addressing this issue is not easy. One solution is to refer to the GO annotations associated with domains. Indeed, when GO annotations associated with a newly discovered domain are already known in the corresponding protein, the domain is likely to be present. The last column of Table 1 reports the proportion of new domains with GO annotations that are in agreement with those of the corresponding proteins. For example, for the 0.1 substitution rate, 130 of the 274 new discovered domains are annotated, among which 97 (75%) have annotations already known in the associated proteins. This high proportion suggests that a large part of the new domains are not false positives, but rather true domains recovered by our approach.

3.2 Reannotation of *P. falciparum* proteins

The method was applied to *P. falciparum* using the three different types of validating domains discussed in the Method section: 1) known Pfam domains; 2) known InterPro domains (excluding Pfam domains); and 3) potential domains themselves. Known Pfam and InterPro domains were obtained using the InterProScan software (release 19.0) that was applied to the *P. falciparum* proteome (PlasmoDB release 5.5). For potential Pfam domains, different

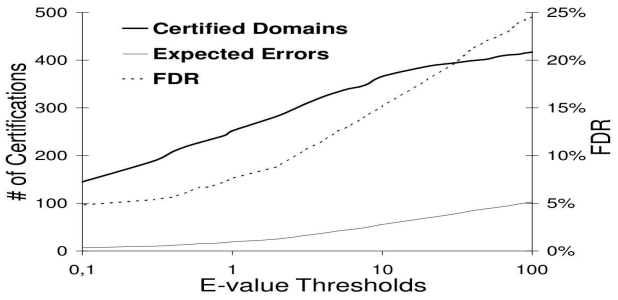

Fig. 1. Evolution of the number of certifications and FDR as a function of the E-value. Number of certifications (thick line), expected number of certifications under H_0 (thin line), and FDR (dashed line) achieved when modulating the E-value thresholds (x-axis). The number of certifications and errors (thin and thick lines) are read on the left y-axis, while the FDR (dashed line) is read on the right one.

Table 2. Results on *P. falciparum*

FDR	$\leq 10\%$				$\leq 20\%$			
Valid. dom.	Pfam	Interp.	Pot.	All	Pfam	Interp.	Pot.	All
Certif. Dom.	259	185	47	340	363	395	130	585
New Interp.	200	138	39	259	298	318	114	479
New Dom. Types	70	51	13	90	106	105	36	159

"Valid. dom." indicates the type of validating domains used for certifications: "Pfam", known Pfam domains from InterProScan; "Interp.", known InterPro (non-Pfam) domains from InterProScan; "Pot.", potential domains themselves. "All": results achieved when combining the 3 types; "Certif. dom.": number of new certified domains; "New Interp.": number of certification allowing us to identify a new InterPro Entry for the protein; "New Dom. Types": number of domain types that were not previously detected in any *P. falciparum* proteins.

E-value thresholds were used in order to obtain predictions with various FDRs.

Figure 1 reports the number of certified domains and FDR achieved with known Pfam domains when varying the E-value threshold used to select the potential domains. Both the number of certified domains and the FDR increase with the E-value threshold, which illustrates the potential of the method to control the FDR by simply modulating this threshold.

Table 2 summarizes the results achieved for FDRs below 10% and 20% with the three types of validating domains. For example, for $FDR \leq 20\%$, 585 new domains are certified. This is an increase of $\sim 16\%$ compared to the $3\,683$ already known Pfam domains in *P. falciparum* proteins (only one occurrence of each known/new domain per protein is considered here; Pfam release 23.0). Among these, 479 involve a new InterPro domain family in the protein. The known Pfam domains allow certification of 363 of the 585 new domains, the known InterPro domains 395, and the potential domains themselves 130 (several new domains are certified by 2 or 3 of the validating domain types). Moreover, 159 new domain types are discovered—*i.e.* which had never been previously detected in *P. falciparum* proteins—, an increase of 11% of the total number of domain types known in *P. falciparum* (see Supp. Table 1). The Supp. Figure 1 shows the number of certifications as a function of the FDR achieved by each type of validating domain. For a given FDR, the potential domains allow the certification of fewer domains than the two other types. This is not a surprise, as these validating domains are potentially false. Hence, very low E-value thresholds are required to achieve low FDRs, which induces the selection of small amounts of potential-validating domains.

We then addressed the difficult issue concerning conservation of the functionality of the new domains. We tried to answer this question by looking at two different indicators. First, as explained in the Introduction, *P. falciparum* proteins often involve long low complexity regions, and these regions are suspected to primarily affect the non-functional parts of sequences. However, a comparison of the proportions of low complexity regions in newly certified domains and in already known domains does not reveal significant differences (see Supp. Table 2). We next investigated the positions where the new domains are discovered. Indeed, Weiner *et al.* (2006) showed that domain divergence events (especially domain loss due to loss of functionality) occur primarily at the ends of the proteins. Here again, when comparing the distances separating known and new domains from protein ends (see Supp. Figure 2), no bias towards the ends can be observed in the new domains. Hence, no bias is found in these two indicators, which seems to indicate that the proportion of new domains that are non-functional is not higher than that of the already known domains.

We next investigated GO annotations that could be deduced from these newly identified domains. As described in the Introduction, some domains have been associated with specific GO terms by the InterPro consortium. The policy is to associate, with a given domain, annotations shared by all annotated proteins possessing this domain. Moreover, by extending this policy to domain combinations—as described in Forslund and Sonnhammer (2008)—, several additional GO terms can be deduced from the combination of two or more domains. To this end, we enumerated all Pfam domain combinations in the proteins of Swiss-Prot, and identified for each combination the GO terms shared by all annotated proteins where the combination is present (only combinations observed in at least 10 annotated proteins were considered). We found $2\,235$ Pfam domain combinations associated with at least one specific GO annotation: $2\,115$ domain pairs, 119 domain triplets and 1 quartet. All associations between domain combinations and GO terms are available in the Supplementary Material[1]. Altogether, single domains and domain combinations improve the annotations of several *P. falciparum* proteins. Table 3 summarizes the results. For example, with a $FDR \leq 20\%$, the newly certified domains leads to the discovery of $273 + 114 = 387$ new GO annotations, *i.e.* 6% of the $5\,791$ already known GO annotations of this organism (in the three ontologies).

All results with $FDR \leq 40\%$ are available in an online database at URL http://www.lirmm.fr/∼terrapon/codd/. The newly predicted Pfam domains of *P. falciparum* proteins are displayed along with their associated FDR and the Pfam/InterPro validating domains used for certification. Figure 2 presents an extract from the database for the `PF11_0189` gene.

When browsing this database, several points can be noted. First, annotations of already annotated genes can be enriched

[1] http://www.lirmm.fr/∼terrapon/codd/

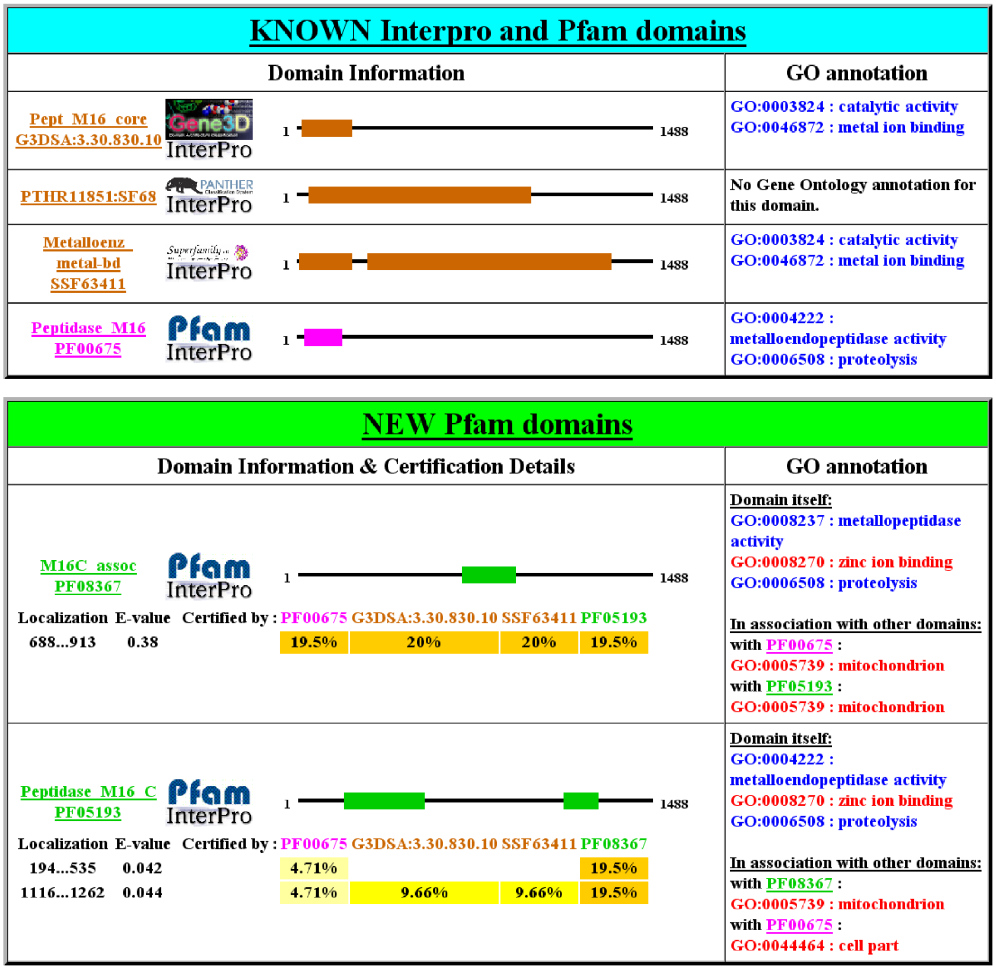

Fig. 2. Known and newly predicted domains of gene `PF11_0189`. Four domains are already known: three InterPro (Gene3D, Panther and Superfamily) domains (`G3DSA:3.30.830.10`, `PTHR11851:SF68`, and `SSF63411`) and one Pfam domain (`PF00675`). We see the localizations of these domains and the associated GO terms. Moreover, two new Pfam domains have been discovered: `PF08367` and `PF05193`. For example, `PF05193` has been discovered in two positions: the first one with an E-value of 0.042 and the second one with an E-value of 0.044. These E-values are too high to be safely considered according to the recommended Pfam threshold for this domain. However, they have been certified by several validating domains. For example, the domain in second position is certified by the known Pfam domain `PF00675` with a $FDR = 4.71\%$, the known InterPro domains `G3DSA:3.30.830.10` and `SSF63411` ($FDR = 9.66\%$), and the potential (and also newly certified) Pfam domain `PF05193` ($FDR = 19.5\%$). Note that the domain in first position is not certified by the two InterPro domains because it overlaps these domains. `PF05193` is associated with two GO annotations already known for the gene (`proteolysis` and `metalloendopeptidase activity`) and one new annotation (`zinc ion binding`). Moreover, since it is found together with domains `PF08367` and `PF00675`, it is also associated with GO terms `mitochondrion` and `cell part`

.

in numerous cases. For example, `MAL7P1.12`, annotated as an "erythrocyte membrane-associated antigen", is ascribed a novel possible molecular activity related to RNA control, based on detection of `PF00035` and `PF04851`. Next, a function is predicted for several genes listed as hypothetical in PlasmoDB 5.5. For example, `MAL13P1.78` is possibly a protein kinase (based on the detection of the `PF06743` domain), `PF10_0040` a nuclease involved in DNA repair (based on `PF00867` and `PF00752` domains), `PF10_0152` a nucleotidyltransferase (`PF01909` and `PF03828` domains), `PF11_0244` an ATP-dependent protease (`PF02190` domain), etc. In this list of new putative annotations, novel potential targets for antimalarial therapeutics might be envisaged, including, for example, the `PF14_0052` protein for which certification of the `PF07683` domain (confirming `PF02492`) indicates that it might be involved in the synthesis of cobalamin (vitamin B12), a molecule necessary for the parasite's development. Similarly, the putative tetrahydrofolate dehydrogenase/cyclohydrolase `PFF1490w` (confirmation of `PF-02882` by `PF00763`), is a new enzyme of the folate metabolism, which should now be experimentally investigated. Detection of the `PF05605` domain in `PF14_0479` is also very interesting, since it is extremely specific to the plant kingdom, associated with the

Table 3. New GO annotations of *P. falciparum* proteins

FDR	Single Domains	Combin. Known Dom.	Combin. with Certified Dom.	Total Prot.	Unannot. prot.
$\leq 10\%$	128	122	74	194	20
$\leq 20\%$	273	122	114	267	39

"Single Domains" is the number of new GO annotations brought by a single domain certified by our approach; "Combin. Known Dom." is the number of GO annotations that can be deduced from combinations of already known domains thanks to inferred associations between domain combinations and GO annotations; "Combin. with Certified Dom." is the number of supplementary GO annotations (different from the 2 previous columns) that can be deduced from combinations involving a newly certified domain. "Total Prot." is the total number of proteins involved, and "Unannot. Prot." is the number of proteins without any annotation and for which an annotation has been proposed.

response of plants to drought, and likely acquired in *P. falciparum* following ancestral endosymbiosis with an alga at the origin of the plant-like features in Apicomplexa (Kohler *et al.*, 1997).

Overall, we observe that families of proteins containing domains related to RNA binding, modification and/or processing (Helicase C, RRM, DEAD) are amongst the largest in the *P. falciparum* genome. It also appears that domains involved in protein-protein interactions, *e.g.* WD40, together with TPR 1 (initially identified in 16 sequences, now in 28 proteins) or TPR 2 (initially identified in 12 sequences, now in 27 proteins), are also detected in large protein families. Moreover, the newly certified domains reveal proteins involved in chromatin interaction (such `PFF1385w`, `PFL0975w` or `PF07_0106`) and numerous transcription factor associated proteins, which are of particular interest for future investigation considering the apparent lack of such proteins in initial studies (Coulson *et al.*, 2004; Callebaut *et al.*, 2005). The present work therefore allows an in-depth analysis of these families, with greater genomic coverage. An extended biological appraisal of these results, with several additional examples, is available in the Supplementary Material.

3.3 Application to other *Plasmodium* species

Finally, we applied the procedure to the proteomes of *P. vivax* and *P. yoelii*, two other sequenced *Plasmodium* species infecting humans and rodents, respectively. The results can be browsed at the same URL[2]. Statistics on the number of newly certified domains and GO annotations are in Supp. Tables 3 and 4 and in Supp. Figure 3. The number of newly certified domains is slightly higher in *P. falciparum* than in the other species. Importantly, a large part of the newly certified domains in *P. falciparum* proteins are also certified in *P. vivax* and *P. yoelii*, while another part corresponds to already known domains in these organisms (see Supp. Tables 5 and 6). For example, among the newly certified domains ($FDR \leq 10\%$) in the *P. falciparum* proteins with a known homologue in *P. vivax*, 14% are already known in the *P. vivax* homologue and 69% are also certified in this homologue. Thus, 83% of the new domains are also found in *P. vivax* homologues. These results strongly support our method and our findings in the *P. falciparum* proteome, and can be seen as a third indicator that our new domains are still functional.

[2] http://www.lirmm.fr/∼terrapon/codd/

4 DISCUSSION AND CONCLUSION

Enhancing domain detection is a complex task. Practically, domain models are designed to ensure the presence of a domain thanks to manually curated score thresholds. Beyond the thresholds, avoiding false positives is no longer guaranteed. In this paper, we propose a method to filter out false positives from hits with scores in the twilight zone below the thresholds. To the best of our knowledge, two previous works address related issues. Beaussart *et al.* (2007) designed a tool that helps identify possible annotation artifacts, notably missing domains. This is achieved by searching for clusters of proteins with similar domains, and aligning the proteins of each cluster on the basis of their domain arrangement. Then a missing domain can be detected in a protein by looking at the domain composition of all proteins in the same cluster which have high similarity with the query sequence. This can be an efficient strategy if a protein homologous to the query protein is already known and correctly annotated. Coin *et al.* (2003) propose an elegant approach to increase the sensitivity of HMM domain detection by incorporating context information. Rather than independently detecting each domain of a protein sequence, the authors propose a Markov model that allows global detection of the domain composition of the protein. With this model, the score achieved by a domain at a given position is a function of both the protein sequence and the other potential domains of the protein. A precise comparison of the results achieved with this approach is difficult, as the databases they used have been enriched. However, we can get a rough comparison by concentrating on *P. falciparum* proteins present in the Swiss-Prot40 data the authors analyzed. Swiss-Prot40 involved 491 *P. falciparum* proteins. In these proteins, our method allows the certification of 36 new domains with a $FDR \leq 20\%$, while Coin *et al.* (2003) propose 5 new domains. Among these 5 domains, 1 is also certified by our method. Among the 4 remaining domains, a close inspection of the protein compositions reveals that 2 proteins contain repeats of a single domain. In other words, for 2 cases, the domain has been used to improve its own detection, which is a certification mechanism we did not consider here.

Compared with previous works, our approach has several appealing features. First, it is a simple and intuitive approach which has low computing time and can potentially be used on any genome. Second, each prediction can be explained by exhibiting the validating domain(s) that enables discovery of the new domain. Third, we can benefit from all types of domain information already known in the InterPro database, as well as in any other domain database. Finally, and most importantly, an estimate of the confidence of the certifications can be computed with our shuffling procedure.

The approach proved to be promising with *P. falciparum*. With $FDR \leq 20\%$, it allows us to increase the total number of known domains by 16%, the number of different known domain types by 11%, and the number of known GO annotations by 6%. Analogous and congruent results are obtained on *P. vivax* and *P. yoelii*. Moreover, experiments on yeast showed that the method could also benefit better annotated organisms. Since domain co-occurrence is the strongest source of information, this work did not consider the adjacency and sequential order of the domains. However, as these features are also often conserved (Kummerfeld *et al.*, 2009), it is likely that taking some ordering information into account would improve the approach.

AUTHORS' CONTRIBUTIONS

NT, OG and LB conceived and designed the method and experiments. NT implemented the approach. EM analysed the biological results. NT and LB drafted the manuscript. EM and OG revised the manuscript. OG initiated the project.

ACKNOWLEDGEMENT

This research was supported by the PlasmoExplore project of the French National Research Agency (ANR-06-CIS6-MDCA-14).

REFERENCES

Apic, G., Gough, J., and Teichmann, S. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, **310**(2), 311–325.

Bahl, A. *et al.* (2003). Plasmodb: the plasmodium genome resource. a database integrating experimental and computational data. *Nucleic Acids Research*, **31**(1), 212–215.

Bastien, O., Roy, S., and Marechal, E. (2005). Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C R Biol*, **328**(5), 445–453.

Beaussart, F., 3rd Weiner, J., and Bornberg-Bauer, E. (2007). Automated improvement of domain annotations using context analysis of domain arrangements (aidan). *Bioinformatics*, **23**(14), 1834–1836.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **85**, 289–300.

Callebaut, I., Prat, K., Meurice, E., Mornon, J., and Tomavo, S. (2005). Prediction of the general transcription factors associated with rna polymerase ii in *Plasmodium falciparum*: conserved features and differences relative to other eucaryotes. *BMC Genomics*, **6**, 100.

Cohen-Gihon, I., Nussinov, R., and Sharan, R. (2007). Comprehensive analysis of co-occuring domain sets in yeast proteins. *BMC Genomics*, **11**(8), 161.

Coin, L., Bateman, A., and Durbin, R. (2003). Enhanced protein domain discovery by using language modeling techniques from speech recognition. *PNAS*, **100**(8), 4516–4520.

Coulson, R., Hall, N., and Ouzonis, A. (2004). Comparative genomics of transcriptional control in the human parasite *Plasmodium falciparum*. *Genome Research*, **14**(8), 1548–1554.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Eddy, S. (1998). Profile hidden markov models. *Bioinformatics*, **14**(9), 755–763.

Finn, R. *et al.* (2008). The pfam protein families database. *Nucleic Acids Research*, **36**(Database Issue), D281–D288.

Forslund, K. and Sonnhammer, E. (2008). Predicting protein function from domain content. *Bioinformatics*, **24**(15), 1681–1687.

Geer, L., Domrachev, M., Lipman, D., and Bryant, S. (2002). Cdart: protein homology by domain architecture. *Genome Research*, **12**(10), 1619–1623.

Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Gerstein, M. and Hegyi, H. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, **11**(10), 1632–1640.

Kohler, S. *et al.* (1997). A plastid of probable green algal origin in Apicomplexan parasites. *Science*, **275**, 1485–1489.

Kummerfeld SK., Teichmann SA. (2009). Protein domain organisation: adding order. *BMC Bioinformatics*, **10** (39).

McLaughlin, W., Chen, K., Hou, T., and Wang, W. (2007). On the detection of functionally coherent groups of protein domains with an extension to protein annotation. *BMC Bioinformatics*, **16**(8), 390.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., and *et al.* (2007). New developments in the interpro database. *Nucleic Acid Research*, **35**(Database Issue), D224–228.

Murzin, A., Brenner, S., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**(4), 536–540.

Pizzi, E. and Frontali, C. (2001). Low-complexity regions in plasmodium falciparum proteins. *Genome Research*, **11**(2), 218–229.

Rambaut, A. and Grassly, N. (1997). Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**(5), 235–238.

Richardson, J. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.

Scott, M., Thomas, D., and Hallett, M. (2004). Predicting subcellular localization via protein motif co-occurrence. *Genome Research*, **14**(10A), 1957–1966.

Soriç, B. (1989). Statistical 'discoveries' and effect size estimation. *Journal of Am. Statist. Ass.*, **84**, 608–610.

UniProt Consortium (2009). The universal protein resource (uniprot) 2009. *Nucleic Acids Research*, **37**(Database Issue), D169–D174.

Weiner 3rd, J., Beaussart, F., and Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS Journal*, **273**(9), 2037–2047.