Among relevant annotations that can be attached to a protein, classes of proteins (Murzin et al., 1995). Others are inferred by clustering conserved subsequences (Finn et al., 2008; Mulder et al., 2007). 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, which account for >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 Pfam library—see Supplementary 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 Pfalciparum, which is composed of >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 Pfalciparum proteome, in which six amino acids account for >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 Pfalciparum. Our method involves lowering the thresholds provided by Pfam for detecting the 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.

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, whereas this rate increases to 80% when they share two identical domains (Gerstein and Hughey, 2001). This is the basis of approaches used to predict protein functional annotations. Scott et al. (2004) use Bayesian networks of co-occurring motifs to predict subcellular localizations of proteins. McLoughlin 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 article, 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 shuffling procedure that allows estimation of its false discovery rate the sensitivity of Pfam domain detection.

(2002) present the CDART tool that allows users to find proteins 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 article, we propose to use domain co-occurrence to improve the sensitivity of Pfam domain detection.

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-Gilad 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-off’), 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 IntePro database (Mulder et al., 2007), a meta-database of different domain databases: PROSITE, PRINTS, ProDom, SMART, TIGRFam, PIRSF, SUPERFAMILY, Gene3D, PANTHER and Pfam (see Mulder et al. (2007) for references). IntePro incorporates this knowledge into a single resource by organizing entries in IntePro families pooling all representations of the same domain. The known IntePro 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, in order 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 hmmsearch 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 CDPs

The list of CDPs 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 \(\chi^2\). 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 (respectively \(A\) is absent but \(B\) is present), and the number \(z\) of proteins where...
We first assessed the potential of the method to improve the Pfam HMMs used with their score thresholds to determine the set of validating domains. All proteins with at least one domain are retained, and the number of new domains is even higher for lower substitution rates. 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 a newly discovered domain. For example, for the 0.1 substitution rate, 130 of the 274 new discovered domains are annotated, among which 12 (7%) have annotations already known in the corresponding 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 Pfalciparum proteins

The method was applied to Pfalciparum using the three different types of validating domains discussed in Section 2: (i) known Pfam domains; (ii) known InterPro domains (excluding Pfam domains); and (iii) potential domains themselves. Known Pfam and InterPro domains were obtained using the InterProScan software (release 1.9.0) that was applied to the Pfalciparum proteome (PlasmoDB release 5.5). For potential Pfam domains, different 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.
The Supplementary Figure 1 shows the number of certifications as a function of the E-value. Number of certifications (thick line), expected number of certifications under H0 (thin line) and FDR (dashed line) achieved when modulating the E-value thresholds (x-axix). The number of certifications and errors (thin and thick lines) are read on the left y-axis, whereas the FDR (dashed line) is read on the right one.

which illustrates the potential of the method to control the FDR by simply modulating this threshold.

Table 2 summarizes the results achieved for FDRs <10% and 20% with the three types of validating domains. For example, for FDR < 20%, 585 new domains are certified. This is an increase of ∼16% compared with the 3683 already known Pfam domains in Pfalciparum 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 Pfalciparum proteins—an increase of 11% of the total number of domain types known in Pfalciparum (see Supplementary Table 1). The Supplementary 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 Section 1, Pfalciparum 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 (Supplementary 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 (Supplementary Fig. 2), no bias toward 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 Section 1, 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 2235 Pfam domain combinations associated with at least one specific GO annotation: 2115 domain pairs, 119 domain triplets and 1 quartet. All associations between domain combinations and GO terms are available at http://www.lirmm.fr/~terrapon/codd/. Altogether, single domains and domain combinations improve the annotations of several Pfalciparum proteins. Table 3 summarizes the results. For example, with a FDR < 20%, the newly certified domains leads to the discovery of 273 + 114 = 387 new GO annotations, i.e. 6% of the 5791 already known GO annotations of this organism (in three ontologies). All results with FDR ≤ 40% are available in an online database at http://www.lirmm.fr/~terrapon/codd/. The newly predicted Pfam domains of Pfalciparum 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_0289 gene. When browsing this database, several points can be noted. First, annotations of already annotated genes can be enriched with new GO annotations. 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 PP00035 and PP04851. Next, a function is predicted...
Detection of new protein domains using co-occurrence

Fig. 2. Known and newly predicted domains of gene PF11_0189. Four domains are already known: three InterPro 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 an 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 mitochondrial and cell part.

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...
Table 3. New GO annotations of \emph{Pfalciparum} proteins

\begin{tabular}{cccccc}
\hline
\text{FDR (%)} & \text{Single domains} & \text{Combin. known dom.} & \text{Combin. with certified dom.} & \text{Total prot.} & \text{Unannot. prot.} \\
\hline
\leq 10 & 128 & 122 & 74 & 194 & 20 \\
\geq 20 & 273 & 122 & 114 & 267 & 39 \\
\hline
\end{tabular}

\begin{itemize}
\item \text{‘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 two 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.}
\end{itemize}

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 \textit{PFPL1315w, PFPL0975w \textit{or PFPL07_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 (Callebaut et al., 2005; Coulson et al., 2004). 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 \textit{Plasmodium} species

Finally, we applied the procedure to the proteomes of \textit{P.vivax} and \textit{P.yoelii}, two other sequenced \textit{Plasmodium} species infecting humans and rodents, respectively. The results can be browsed at \url{http://www.lirmm.fr/~terrapon/codd/}. Statistics on the number of newly certified domains and GO annotations are in Supplementary Tables 3 and 4 and in Supplementary Figure 3. The number of newly certified domains is slightly higher in \textit{P.falciparum} than in the other species. Importantly, a large part of the newly certified domains in \textit{Pfalciparum} proteins are also certified in \textit{P.vivax} and \textit{P.yoelii}, while another part corresponds to already known domains in these organisms (see Supplementary Tables 5 and 6). For example, among the newly certified domains (FDR $\leq 10\%$) in the \textit{Pfalciparum} proteins with a known homolog in \textit{P.vivax}, 14\% are already known in the \textit{P.vivax} homolog and 69\% are also certified in this homolog. Thus, 83\% of the new domains are also found in \textit{P.vivax} homologs. These results strongly support our method and our findings in the \textit{Pfalciparum} proteome, and can be seen as a third indicator that our new domains are still functional.

4 DISCUSSION AND CONCLUSIONS

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 article, 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 to 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 \textit{Pfalciparum} proteins present in the Swiss-Prot40 data that the authors analyzed. Swiss-Prot40 involved 491 \textit{Pfalciparum} proteins. In these proteins, our method allows the certification of 36 new domains with an FDR $\leq 20\%$, whereas Coin et al. (2003) propose nine new domains. Among these five domains, one is also certified by our method. Among the four remaining domains, a close inspection of the protein compositions reveals that two proteins contain repeats of a single domain. In other words, for two 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 the 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 \textit{Pfalciparum}. 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 \textit{P.vivax} and \textit{P.yoelii}. Moreover, experiments on yeast showed that the method could also benefit 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 and Teichmann, 2009), it is likely that taking some ordering information into account would improve the approach.

AUTHORS’ CONTRIBUTIONS

N.T., O.G and L.B. conceived and designed the experiments. N.T. implemented the approach. E.M. biological results. N.T. and L.B. drafted the manuscript, revised the manuscript. O.G. initiated the project.
Detection of new protein domains using co-occurrence

**Funding:** PlasmoExplore project of the French National Research Agency (ANR-06-CIS6-MDCA-14).

**Conflict of Interest:** none declared.

**REFERENCES**


