EuPathDomains: The Divergent Domain Database for Eukaryotic Pathogens

Amel Ghouila, Nicolas Terrapon, Olivier Gascuel, Fatma Z. Guerfali, Dhafer Laouini, Eric Maréchal, Laurent Brehelin

EuPathDomains: the Divergent Domain Database for Eukaryotic Pathogens

Amel Ghouila (1, 2, 3), Nicolas Terrapon (1), Olivier Gascuel (1), Fatma Z. Guerfali(4), Dhafer Laouini (4), Eric Maréchal (5) and Laurent Bréhélin (1)

(1) Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Univ. Montpellier 2 161 rue Ada, 34095 Montpellier, France

amel.ghouila@lirmm.fr, terrapon@lirmm.fr, brehelin@lirmm.fr, gascuel@lirmm.fr

(2) Research Unit on Molecular Investigation of Genetic Orphan Diseases, Institut Pasteur de Tunis, Tunisia

(3) Computer Science Department, Faculty of Sciences of Tunis, Tunisia

(4) Laboratoire d'Immunopathologie, Vaccinologie et Génétique Moléculaire. Laboratoire International Associé (LIA), CNRS. Institut Pasteur de Tunis, Tunisia

fzguerfali@yahoo.fr, Dhafer_l@yahoo.ca

(5) Laboratoire de Physiologie Cellulaire Végétale - UMR CNRS CEA UJF INRA 5168, iRTSV, CEA Grenoble, 17 rue des Martys, 38054 Grenoble cedex 9, France

eric.marechal@cea.fr

Corresponding author:

Laurent Bréhélin - Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS, Univ. Montpellier 2, 161 rue Ada, 34095 Montpellier Cedex 5, France. Fax-Number: +33. 467 41 85 00 Email: brehelin@lirmm.fr

EuPathDomains: The Divergent Domain Database for Eukaryotic Pathogens

Journal: Infections, Genetics and Evolution

ABSTRACT

Eukaryotic pathogens (e.g. *Plasmodium*, *Leishmania*, *Trypanosomes*, etc.) are a major source of morbidity and mortality worldwide. In Africa, one of the most impacted continents, they cause millions of deaths and constitute an immense economic burden. While the genome sequence of several of these organisms is now available, the biological functions of more than half of their proteins are still unknown. This is a serious issue for bringing to the foreground the expected new therapeutic targets. In this context, the identification of protein domains is a key step to improve the functional annotation of the proteins.

However, several domains are missed in eukaryotic pathogens because of the high phylogenetic distance of these organisms from the classical eukaryote models. We recently proposed a method, Co-Occurrence Domain Detection (CODD), that improves the sensitivity of Pfam domain detection by exploiting the tendency of domains to appear preferentially with a few other favorite domains in a protein.

In this paper, we present EuPathDomains (http://www.atgc-montpellier.fr/EuPathDomains/), an extended database of protein domains belonging to ten major eukaryotic human pathogens. EuPathDomains gathers known and new domains detected by CODD, along with the associated confidence measurements and the GO annotations that can be deduced from the new domains. This database significantly extends the Pfam domain coverage of all selected genomes, by proposing new occurrences of domains as well as new domain families that have never been reported before. For example, with a false discovery rate lower than 20%, EuPathDomains increases the number of detected domains by 13% in *Toxoplasma gondii* genome and up to 28% in *Cryptospordium parvum*, and the total number of domain families by 10% in *Plasmodium falciparum* and up to 16% in *Cryptospordium parvum* genome. The database can be queried by protein names, domain identifiers, Pfam or Interpro identifiers, or organisms, and should become a valuable resource to decipher the protein functions of eukaryotic pathogens.

Keywords: Pathogens, protein domains, functional annotations, Pfam, HMMs, database.

Introduction

Pathogenic eukaryotes belong to several distinct phylogenetic lineages causing a wide variety of host infections. They have evolved the ability to infect a range of hosts, including humans, animals and plants [1]. Despite the availability of the complete genome of many of these organisms, their exact infection mechanisms and invasion strategies, as well as the function of more than 50% of their proteins, are still under investigation. This is mainly due to their high phylogenetic distance from the classical model organisms, as well as to their evolutionary mechanisms that are quite different from those envisioned in classical models of evolution due to the co-evolution of host and pathogens and the development of both systems in an ongoing process [2]. In fact, parasites have evolved a variety of mechanisms to evade host immune recognition and elimination. Such evolution is a direct consequence of the fact that they live at the expense of, and are dependent on, host tissues for long periods during their life cycles [3]. This contributed to the development of new gene functions and changes in the parasite's genome repertoire, occurring through gene acquisition and deletion [4].

Among the relevant annotations that can be attributed to a protein, domains occupy a key position. Protein domains are sequential and structural motifs that are found independently in different proteins and in different combinations and, as such, seem to be functional subunits of proteins above the raw amino acid sequence level [5]. Protein domain composition provides strong clues for the protein function. Indeed, two thirds of mono-domain proteins having the same domain also have the same functions. Likewise, 35% of multidomain proteins having one common domain present similar functions, while this rate increases to 80% when they share two common domains [6]. Several approaches and databases have been developed to define and identify domains. One of the most widely used domain scheme is the Pfam database [7], which is a collection of protein domain and families. Each family in Pfam is represented by a multiple sequence alignment and a profile Hidden Markov Model (profile HMM) [8]. The Pfam 23.0 release offers a large collection of 10 340 protein domains.

When analyzing a new sequence, each Pfam HMM is used to compute a score measuring the similarity between the sequence and the domain. If the score is above a given threshold provided by Pfam (each domain has a different score threshold), then the presence of the domain can be asserted in the protein. However, applied to divergent proteins, this strategy may miss several domains. This is the case for all eukaryotic pathogens, where no Pfam domains are detected in half or even more of their proteins. While the Pfam domain coverage of the classical model organisms is 73% on average for all organisms (Table 1), this proportion varies for eukaryotic pathogens from 38%, (for *Trypanosoma brucei*) to 53% (for *Plasmodium falciparum*). Although this situation may be explained by the existence of genes that are unique to these organisms or specific to a parasitic life style, it is further exacerbated by the high evolutionary distance between these organisms and the other organisms used as model Eukaryotes—such as yeast, worms, arabidopsis, drosophila and human [9]—, which makes homology detection particularly difficult.

We recently proposed a method named CODD that increases the sensitivity of Pfam domain detection in divergent proteins while controlling the false discovery rate (FDR) of the predictions [10]. Our method makes use of the well known tendency of protein domains to appear preferentially with a few other favorite domains in proteins [11]. This enables us to certify the presence of domains below the recommended thresholds, on the basis of the presence of another domain in the same protein. Moreover, a shuffling procedure provides an estimate of the FDR associated with each prediction.

In this paper, we present EuPathDomains, an extended database of known and new protein domains identified by CODD on several major human pathogens (three *Leishmania* species, three *Plasmodium* species, *Toxoplasma gondii*, *Cryptospordium parvum*, *Giardia lamblia* and *Trypanosoma brucei*) selected from EupathDB database [12]. In average, EuPathDomains increases the number of domains in a proportion of ~10% in each genome (with an FDR<20%). Several new domains types were discovered, leading to an enrichment of 10% to 16% of the total number of domain families in each genome. Some of these domains are localized in proteins that had no known domain. In addition, these newly predicted domains induce new GO annotations for about 15% of the proteins within each of the studied organisms. EuPathDomains database provides all these predictions along with the already known domains and annotations in a friendly interface that allows easy browsing and querying.

In the following, we briefly describe the studied pathogens. We explain the principle of CODD and the shuffling procedure used to estimate the confidence (FDR) associated with the newly discovered domains. We next present the results achieved on each species, and describe the *EuPathDomains* database interface. Finally, as a case study, we emphasize the results obtained with the three selected *Leishmania* species.

Data and Methods

Pathogens included in EuPathDomains

One organism affecting humans, representative of each phylum of the EupathDB[1] database was selected in this study [12]. These include *L. major* that causes leishmaniasis, three apicomplexans —*P. falciparum* that cause malaria, *T. gondii* and *C. parvum* that cause respectively toxoplasmosis and a diarrheal illness called cryptosporidiosis —, *T. brucei* that causes sleeping sickness, and *G. lamblia* that causes giardiasis. To allow comparative studies of close species, we also included two additional *Leishmania—L. infantum and L. braziliensis*—and two *Plasmodium—P. vivax* and *P. yoelli.*

***Leishmania* species.** Infection by the insect-transmitted *Leishmania* parasites represents a serious global health problem for which there is no vaccine and few effective but toxic drugs [13]. Intra-macrophagic protozoan leishmanies are single celled parasites that affect vertebrates including dogs and humans. At least 20 *Leishmania* species infect humans, and the spectrum of diseases that they cause can be categorized broadly into three types: visceral leishmaniasis, cutaneous leishmaniasis and mucocutaneous leishmaniasis. Three different *Leishmania* species included in this study, i.e.; *L. major*, *L. infantum* and *L. braziliensis* cause respectively cutaneous, visceral and mucocutanous leishmaniasis. These different *Leishmania* species contain each approximately 8 000 protein coding genes, more than half of them do not have either GO annotation terms or known protein domains.

***Plasmodium* species.** *Plasmodium* is a genus of parasitic protists. Infection by these organisms is known as malaria, which is one of the most devastating infectious diseases causing more than 1 million deaths each year worldwide [14]. Currently, over 200 species of this genus are recognized and new species continue to be described. Over all these species, at least 5 infect humans. Various types of malaria are caused by the different species. *P. falciparum* causes the most severe disease and is responsible for nearly all malaria-related deaths. Other species cause debilitating diseases that are less severe but highly persistent. The parasite always has two hosts in its life cycle: a mosquito vector and a vertebrate host. In this work, we studied three different *plasmodium* species: *P. falciparum*, *P. vivax* and *P. yoelli*. *P. falciparum* and *P. vivax* have approximately 5 500 genes, while *P. yoelii* genome contains about 7 720 genes. Nearly 50% proteins of each these 3 species do not have any known Pfam domain.

Toxoplasma gondii is *a* single-celled intracellular parasite that causes a disease known as toxoplasmosis. Infections by *T. gondii* are highly prevalent in humans and animals and the parasite is widely found throughout the world. The definitive host of *T. gondii* is the cat, but the parasite can be carried by many warm-blooded animals and cause severe and life-threatening disease in developing fetuses and immune-compromised individuals. The majority of infected persons are asymptomatic due to their protective immune status [15]. *T. gondi* has about 8 102 protein-coding genes, about 57% of them do not have any known Pfam domain.

Cryptosporidium parvum is a protozoan parasite that causes cryptosporidiosis, a disease affecting the mammalian intestinal tract and mainly characterized by a diarrheal illness. It is transmitted to the host via the fecal-oral route. Among healthy people, an acute self-limiting infection occurs, but cryptosporidiosis could have lethal effects on immuno-compromised individuals (HIV infected, immuno-suppressed etc.). Cryptosporidiosis can be found worldwide, and in developing countries 8-19% of diarrheal diseases are attributed to *cryptosporidium* [16]. Its genomic sequence shows a small, compact genome of 9.1Mb with a simple organization containing only about 3 805 protein coding genes [17]. A total of 1 123 different Pfam domains are known on these proteins covering 51% of the whole genome.

Giardia lamblia. Giardiasis is a parasitic disease caused by *Giardia* species, a flagellated protozoan parasite that occupies the small intestine of numerous hosts including humans. Like *Cryptosporidium*, infection can occur due to ingestion of contaminated water or food and by the fecal-oral route. Depending on its life cycle stage (external dormant form versus ingested active form), two different parasite forms exist. Through protein-surface modification (with antigenic variation), *Giardia* can efficiently evade the host's immune defense [18, 19]. Like other diplomonads, *Giardia* is characterized by the presence of two nuclei. Its genome was only recently published and revealed a sequence of roughly 12Mb and about 5 000 protein-coding genes [20].

Trypanosoma brucei are obligate parasitic protists of mammals to which they are transmitted by an insect vector commonly called tsetse fly. They cause African trypanosomiasis, commonly known as sleeping sickness. Its adaptation to different hosts, a common strategy among parasites, occurs through complex changes during its life cycle, and different parasite forms are known. A particular feature of kinetoplastids, in comparison with other eukaryotes, is a mitochondrial genome-containing kinetoplast associated with the basal body of the flagellum. Its genome is composed of 11 chromosomes and contains nearly 9800 protein coding genes. About 62% of these genes do not have any known Pfam domain.

CODD: Co-occurrence based domain detection

CODD is a computational approach that increases the sensitivity of HMMs for domain detection, while controlling the false discovery rate (FDR) associated with the predictions [10]. CODD utilizes the tendency (observed through the three kingdoms of life) of many domain families to occur preferentially with a few other favorite families of proteins [21]. Thus, the presence of one domain in a protein may be a strong clue for the presence of another one. CODD uses this tendency to warrant the presence of Pfam domains, on the basis of the presence of another domain in the same protein. The “validating domain” used for this purpose may be another Pfam domain or any other type of domain from the InterPro database [22]. CODD has been assessed both in *Yeast* and *P. falciparum* genomes [10]. It enabled the discovery of several hundreds new Pfam domain occurrences in both organisms, and contributed to the enrichment of the Gene Ontology (GO) annotation of their proteins.

The CODD principle is summarized in Figure 1. It involves five different steps:

1. The first step involves the identification of the Conditionally Dependent Pairs of domains (CDP), i.e. the domain pairs that tend to appear preferentially together in the same protein. The CDP list is computed from the whole set of domain pairs observed in Uniprot proteins of all organisms.

2. The second step involves the identification of potential domains in the proteins of the target organism. This is done by searching the protein sequences against Pfam HMM profiles with a loose score threshold.

3. The third step is the certification procedure that uses the CDP list to certify the presence of potential domains thanks to the presence of other domains in the same protein.

4. A shuffling procedure is then used to estimate the confidence (False Discovery Rate or FDR) of the newly certified domains.

5. Finally, GO annotations are deduced from the newly discovered domains themselves and also from their combinations with others.

We further detail below how this general method has been implemented and used to build the EuPathDomains database.

Selecting the CDPs

The list of Conditionally Dependent Pairs (CDP) of domains was built from the whole set of domain pairs observed in UniProt proteins. 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. For each pair, the number of proteins where both domains are present and where at least one domain is present is computed and used to measure the conditional dependence with a Fisher's exact test. All pairs with a *p-value* < 1% are considered as conditionally dependent and are added into the CDP list.

Selecting the potential domains

The sets of potential domains were inferred from the results of Pfam HMM searches using HMMER software [23]. 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 was 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.

Certification process

Each potential domain identified in the previous step is then queried for certification by another domain: if both domains are in the CDP list, the potential domain is certified. Three kinds of certifications are considered. The first and most accurate one involves the certification of the potential domain by a known Pfam domain of the protein. A complementary solution is to certify the potential domain with an InterPro non-Pfam known domain. This allows us to increase the number of certifications. However, due to the heterogeneity of the InterPro database, the certifications achieved in this way may be of lower quality than those achieved with Pfam domains. The first two solutions certify domains solely in proteins where at least one domain is already known. To overcome this limitation, a third solution is to certify the potential domain by another potential domain of the protein. With 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.

Estimating the number of false certifications

The certification procedure allows certifying several new domains among all the potential domains identified in the proteins of the organism. One issue is then to estimate the proportion of false positives among these new domains. To this end, one estimates the probability of certifying a potential domain under the null hypothesis H0 that it has been randomly predicted. This is done through computer simulations, by shuffling the potential domains of all proteins. The certification procedure is applied to the shuffled domains, and the number of certified random domains is computed. The entire procedure is repeated several times (typically 1000 times) to get a reliable estimate of the expected number of domains this procedure would certify under H0, and this number is used to compute an estimate of the False Discovery Rate (FDR), i.e. the proportion of false positives in the new domains.

Gene Ontology annotations

The GO currently serves as the dominant approach for machine-legible functional annotation. GO is a controlled and structured vocabulary describing three aspects of gene product function:

molecular function, biological process and cellular location. Attempts have been made to link the Pfam and InterPro domains with the terms of the GO. The pfam2go mapping associates a specific GO term with a Pfam domain if all proteins in which this domain is known share the term [24]. Thanks to this mapping, when a new domain is certified in a protein, all terms associated with this domain (if any) can be transferred to the protein.

Protein function usually results from combination of domains rather than from a single one [24]. Thus, we generated a second mapping that links whole domain combinations to GO annotations. This second mapping is built on the same principle as the pfam2go mapping. All observed domain combinations of the Swiss-Prot proteins are enumerated. Then, for each combination present in more than 10 proteins, if all proteins in which the combination is present share a specific GO term, this term is assigned to the combination. As with the pfam2go mapping, this second mapping can be used to transfer new annotations to the protein in which a new domain is certified, if this domain forms an annotated combination with other known or certified domains.

Building the database

CODD was applied to all selected and above described organisms, with the three types of certification (known Pfam, known non-Pfam, and potential Pfam). The set of known InterPro domains (Pfam and non Pfam) were downloaded from each organism's specific database. For each organism, two E-value thresholds (typically 1 and 50) were used to identify the potential domains. This led to the construction of two different sets of potential domains of increasing size. CODD was run on each set, and the FDR associated with the certifications achieved was computed. High E-value thresholds usually allow more certifications, but with also higher FDRs. All certifications with an FDR below 20% were transferred to the EuPathDomains database. For each newly certified domain, the GO annotations that can be deduced from this domain, either solely or in combination with other known or newly certified domains, were also transferred to the database.

Results and discussion

The EuPathDomains database

EuPathDomains (for "Eukaryotic Pathogens Domains database") is freely available at http://www.atgc-montpellier.fr/EuPathDomains/. A friendly interface allows easy browsing and querying. The database can be queried in various ways, by protein names, domain IDs, Pfam or Interpro identifiers, with the possibility to limit the search on an organism or a taxon (*Plasmodiums, Apicomplexans*, ...), and with a given FDR threshold. The result includes the list of proteins where the domain is already known and those where it has been predicted with the associated FDRs.

Figure 2 illustrates a typical output obtained when querying for the *L. major* protein LmjF36.4590. This figure shows the known domains for this protein as well as the predicted domains (with the domains that allowed their certification) and the related GO terms transferred to the protein. Some terms, such as "zinc ion binding" and "intracellular" are already known in the protein, but other annotations, such as "regulation of transcription, DNA dependent", are new.

The database contains several hundred new Pfam domains in each species. Some of the newly predicted domains have never previously been seen in these species. Thus, these predictions enabled the set of known domain families in all studied organisms to be expanded. With a FDR below 20%, the database expands the set of domains between 13% in *G. lamblia* to up to 28% in *C. parvum*. It also expands the set of known domain families in each organism, between 10% in *P. falciparum* to up to 16% in *C. parvum*. Moreover, the newly discovered domains also provide new functional annotations for the proteins of the different species. The number of additional GO terms ranges from 302 in *G. lamblia* to 684 in *T. brucei*. Table 2 details the number of new domains found in each species for the three types of certification. The numbers of new GO annotations that can be deduced from these domains are detailed in Table 3.

For example, in *L. major*, EuPathDomains contains a total of 719 new domains with a FDR below 20% (Table 2). This is an increase of 14% compared to the 5 216 already known Pfam domains in this organism (only one occurrence of each known/new domain per protein is considered here; Pfam release 23.0). Among these, 607 involve a new InterPro domain family in the protein. The known Pfam domains allow for the certification of 385 out of the 719 new domains, the known non-Pfam domains 417, and the potential domains themselves 316 (several new domains are certified by 2 or the 3 types of certifications). As we can see, the potential domains allow the certification of fewer domains than the known domains for a given FDR. This is not surprising, as these domains are potentially false. Hence, very low E-value thresholds are required to achieve low FDRs, which induce the selection of small amounts of potential domains. Moreover, 184 new domain types were discovered—i.e. which had never been previously detected in *L. major* proteins—, an increase of 11% in the total number of domain types known in this parasite. With an FDR<20%, the newly discovered domains led to the identification of 466 new GO terms (Table 3), representing 6% of the already known GO terms in *L. major* proteins. This allows us to propose GO annotations for 53 proteins without any annotation. Similar results were also achieved on the two other *Leishmania* species (*L. infantum and L. braziliensis;* see Supplementary Table S1 and S2, respectively).

Description of the newly discovered domains

We sought to characterize the specificities of the newly discovered domains, compared to the already known domains in each species. One of the first characteristics involves the length of the new domains. Usually, domains vary in length from about 25 to 500 amino acids. A comparison of the known and new Pfam domains shows that, on average, the newly certified domains are shorter than the already known domains (see Figure 3). This difference between known and new domains is not surprising, as short domains carry less information than longer ones, and are more difficult to detect, especially in divergent proteins. Thus, our certification procedure primarily enables recovery of these domains.

We next investigated the GO annotations provided by the newly predicted domains via Pfam2go mapping. GO annotations provided by combination of several domains are generally more specific than those deduced from the presence of a single one. For previously annotated proteins, this enables the discovery of more accurate functions confirming the already known annotations. For example, the term « RNA polymerase II transcription factor activity », which catalyzes DNA transcription to synthesize precursors of mRNA and most snRNA and microRNA, is assigned to different Leishmania proteins by combining both predicted and already known domains.

For hypothetical or very poorly annotated proteins, two major cases are observed based on the level of annotation of the newly detected domains. On the one hand, the new domains are ascribed to no or poorly informative GO terms. In this case, no precise functional annotation can be deduced from the present work, but the structural categorization of the proteins is refined, providing clues for future functional inferences. On the other hand, newly detected domains can be sufficiently informative and have accurate GO annotations, which are transferred to the hypothetical or poorly annotated protein.

Next, we used the GOstat tool [25] to compare the GO annotations associated with the known and the newly certified domains according to the three axis of the Gene Ontology: molecular function, biological processes and cellular component. For each species, we uploaded in GOstat the set of newly discovered domain families, the set of known domain families, and the Pfam2go mapping. Then, we asked GOstat for the GO annotations that are over-represented in the new domain families in comparison with the known domain families. Table 4 lists the results achieved in each species for domains certified with an FDR<10%. With the exception of *G. lamblia,* several terms appear to be overrepresented in the different organisms. For example, several overrepresented terms are related to DNA- or RNA-binding in *Plasmodium* species. This finding might be useful to decipher the currently debated mode of regulation of transcription in these organisms [26, 27].

Leishmania as a case study

We emphasized our analysis on the genus *Leishmania* with three species included in EuPathDomains database. At first glance, we notice that many terms related to DNA and RNA stability and transcription are found in the new described domains (Table 4) of these species, in comparison with the known ones. This should help to identify novel key elements involved in the complex mechanisms driving polycistronic mRNA transcription in *Leishmania*. Indeed, and although extensively studied, this process remains one of the major focuses of *Leishmania* biology research [28, 29].

First analyses of the predictions achieved in these species show that a great number of newly discovered domains are strikingly shared between them (see Table 5), which increases our confidence in the correctness of the approach. Indeed, the majority of newly discovered domains in *L. major* (87% with FDR<10%) were also newly discovered in *L. Infantum,* while 18 new domains were already known in *L. infantum* species.

When looking to new domain families that were previously unknown in the three Leishmania species, 45 domains with an FDR<10% (~40%) are common between them. RQC (PF09382) is one of these new domains that are common to the three species. This DNA-binding domain is specific to the RecQ family of helicases that bind and unwind G4-DNA. A recently described potential role [30, 31, 32] of this G4-DNA in gene regulation and genome stability is its participation in gene transcription by maintaining the DNA in an open conformation [33]. Moreover, RQC has a helix-turn-helix structure (HTH) that is commonly found in gene regulatory proteins, and known to bind DNA in a sequence-specific manner [34, 35]. RQC could thus represent a relevant feature of Leishmania gene regulatory processes. Related to another regulatory level is the “Response reg” domain (PF00072), which is associated in other organisms with a “two-component signal transduction system” (TCS) involved in signal transduction events [36, 37]. The TCS pathway was originally described in bacteria, and more recently in eukaryotes, as involved in sensing the environment for changes in stress or growth conditions [38, 39]. As TCS is involved in drug resistance, osmoregulation and motility among others, one can think that the associated Response_reg new domain could contribute to pathogen physiological status adaptation to a wide variety of stimuli.

The new 45 conserved protein domains have annotations related to different pathways. We classified them into nine different functional groups based on the pfam2go mapping but also on literature mining (Figure 4). The two most represented families are related to transcription and DNA binding functions. Besides the existence of all these common domains, a small proportion of the new families (16% in *L. major* for example) appear to be species specific, even if we cannot exclude the possible existence of false positives. As some of these domains may be related to species-specific mechanisms (e.g. tropism), they are worth further investigations.

Common new predicted domains between different pathogens

Within EuPathDomains database, eight domain families, which were previously unknown in the different studied pathogens, are identified in all these organisms. The presence of these new domains addresses the question of their specific occurrence in pathogens and on their potential relation with parasitic life style, including common infectious or surviving strategies that have already been described for pathogens with various clinical outcomes [40].

In the eight common domain families, three (Utp12, Utp21 and PWP2) are associated with proteins involved in the small-subunit (SSU) processome for the processing of the small-ribosomal-subunit rRNA [41]. It has been suggested that, through its association with specific ribosomal proteins, the SSU processome was probably involved in both pre-rRNA processing and ribosome assembly [42].

A fourth newly predicted domain family is the polyadenylate-binding protein (PABP)-interacting motif PAM2 that has been identified in several eukaryotic proteins, including ataxin-2 [43]. This domain interacts with a particular family of RNA helicases and participates in the recognition of the 3’end of mRNAs, with an essential contribution to eukaryotic translation initiation and mRNA stabilization/degradation [44]. In addition, a domain PF08147-DBP10CT corresponding to hypothetical RNA helicase has also been identified in all studied organisms using our approach. This could reflect the common need for rapid synthesis of molecules favoring survival of any organism and its resistance to environmental stress.

The last three new domains found in all organisms correspond to proteins involved in distinct functions. MutS_II is a domain found in the MutS family of proteins involved in mismatch repair [45]. Another domain with an unknown function, referenced as PF08953, has been associated in *Leishmania* with the LmjF23.1165 gene, coding for a putative Coronin, a WD-repeat containing protein. It is worth noting that the previously described Utp12 and PWP2 domains are members of this WD-repeat family. Hence, PF08953 domain might be involved in protein interactions in different biological processes such as signal transduction, transcriptional regulation, cytoskeleton remodeling, and regulation of vesicle trafficking [46]. Finally, PF09279, a domain called EFhandlike is newly described in the studied pathogens except Plasmodium species where it was already known. This domain is found in the phosphoinositide-specific phospholipase C family. The phospholipase activity has been described as participating in the degradation membranes containing pathogens by decreasing the number of phospholipids causing the weakening of membrane integrity [47].

We thus speculate that some of these new domains might be associated to some aspects of the parasitic life style. It is therefore now important to explore their occurrence in other organisms. Overall, all these new identified protein domains highlight the importance of domain discovery in assigning and fine tuning biological functions for poorly, and also for well annotated eukaryotic organisms, including important pathogens.

Conclusion

Domains are the building blocks of proteins and one of the key features to help decipher protein function. Indeed, identification of protein domain content is a crucial step in understanding their role. HMMs have proved to be a powerful and accurate tool for this purpose, and are the basis of the Pfam database, one of the most widely used resources for protein domains. However, these models may miss numerous domains in divergent proteins. This is especially true in eukaryotic pathogens where no Pfam domains are detected in more than half of the proteins of these organisms. We recently proposed a method named CODD that uses domain co-occurrence to help detect divergent domains. In this paper, this method is applied on ten human eukaryotic parasites. All predictions, along with their confidence values and GO annotations, have been integrated in a dedicated database named EuPathDomains.

On the whole, the EuPathDomains database significantly improves the domain coverage in all genomes, by localizing new occurrences of domains that are already known and, more interestingly, by detecting new domain families that have never previously been reported. For example, with an FDR<20%, EuPathDomains increases the number of domain occurrence between 13% to up to 28% in each organism, and the total number of domain families between 10% to up to 16%. Moreover, several of these new domains also provide new GO annotations. Thus, EuPathDomains appears to be a valuable new resource for domain annotation in eukaryotic pathogens, and should help deciphering the biology of these complex and crucial organisms.

References

[1] K. Haldar, S. Kamoun, N. L. Hiller, S. Bhattacharje and C. Ooij. Common infection strategies of pathogenic eukaryotes. Nature Reviews Microbiology 4 (2006), 922-931.

[1] K. Haldar, S. Kamoun, N. L. Hiller, S. Bhattacharje and C. Ooij. Common infection strategies of pathogenic eukaryotes. Nature Reviews Microbiology 4 (2006), 922-931.

[2] Julius P. Kreier. Infection, resistance and immunity. 2002, Second edition.

[3] H. Ochman and N. A. Moran. Genes Lost and Genes Found: Evolution of Bacterial Pathogenesis and Symbiosis. Science Vol. 292. no. 5519 (2001), pp. 1096 – 1099.

[4] Immunology of infectious diseases Par Stefan H. E. Kaufmann,Alan Sher,Rafi Ahmed (Mansfield, 2002) book chapter Immune evasion by parsites. J.M. Mansfield and M. Olivier.

[5] J. Richardson: The anatomy and taxonomy of protein structure. Adv Protein Chem., 34 (1981), 167-339.

[6] H. Hegyi and M. Gerstein. Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins. Genome Research 11(2001): 1632–1640.

[7] R. Finn. The Pfam protein families database. Nucleic Acids Research, 36 (2008). D281-D288.

[8] R. Durbin. Biological sequence analysis: probabilistic models of proteins and nucleic acids. (1998) Cambridge University Press, Cambridge, UK.

[9] P. Ward1, L. Equinet, J. Packer and C. Doerig. Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. BMC Genomics, 5 (2004):79.

[10] N. Terrapon, O. Gascuel, É. Maréchal and L. Bréhélin. Detection of new protein domains using co-occurrence: application to Plasmodium falciparum. Bioinformatics 2009 25(23):3077-3083.

[11] G. Apic. Domain combinations in archeal, eubacterial and eukayotic proteomes. J.Mol. Biol., 310 (2001).311-325.

[12] C. Aurrecoechea et al., EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Research, 38 (2010), Database issue D415-D419.

[13] C. S. Peacock et al, Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nature Genetics 39 (2007), 839 - 847

[14] Kooij TWA, Carlton JM, Bidwell SL, Hall N, Ramesar J, et al, A Plasmodium Whole-Genome Synteny Map: Indels and Synteny Breakpoints as Foci for Species-Specific Genes. PLoS Pathog vol 1(2005) , No 4, p, 349- 361.

[15] J. BLADER and J. P. SAEIJ, communication between Toxoplasma gondii and its host: impact on parasite growth, development, immune evasion, and virulence. APMIS 117 (2009), p458–476.

[16] Gatei W, Wamae CN, Mbae C, et al. "Cryptosporidiosis: prevalence, genotype analysis, and symptoms associated with infections in children in Kenya". Am. J. Trop. Med. Hyg. 75 (2006), nu (1): 78–82.

[17] Abrahamsen MS, Templeton TJ, et al. (2004). "Complete genome sequence of the apicomplexan, Cryptosporidium parvum." Science 304 (5669): 441–5.

[18] Svärd SG, Meng TC, Hetsko ML, McCaffery JM, Gillin FD (December 1998). "Differentiation-associated surface antigen variation in the ancient eukaryote Giardia lamblia". Molecular Microbiology 30 (5): 979–89.

[19] Tovar J, León-Avila G, Sánchez LB, et al. (2003). "Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation". Nature 426 (6963): 172–6.

[20] Morrison HG, McArthur AG, Gillin FD, et al. (September 2007). "Genomic minimalism in

the early diverging intestinal parasite Giardia lamblia". Science 317 (5846): 1921–6.

[21] F. Beaussart, J. Weiner and E. Bornberg-Bauer. Automated Improvement of Domain ANnotations using context analysis of domain arrangements (AIDAN). Bioinformatics (2007) 23(14): 1834-1836

[22] N. Mulder et al, New developments in the InterPro database. Nucleic Acids Res. 2007 Jan; 35(Database issue):D224-8.

[23] S. R. Eddy. Profile hidden Markov models. Bioinformatics. 14(1998):755-63.

[24] K. Forslund and E. L. Sonnhammer. Predicting protein function from domain content. Bioinformatics; 24 (2008):1681-1687.

[25] T. Beißbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of gene; 20 (2004): p 1464-1465.

[26] K. Essien, C. J. Stoeckert. Conservation and divergence of known apicomplexan transcriptional regulons. BMC Genomics 11 (1) 2010, art. no. 147

[27] E. Bischoff, C. Vaquero. In silico and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of Plasmodium falciparum. BMC Genomics 11 (1) 2010, art no. 34

[28] Haile S, Papadopoulou B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. Curr Opin Microbiol. 10(6):569-77 (2007).

[29] Clayton C, Shapira M. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. Mol Biochem Parasitol. 156(2):93-101 (2007).

[30] L.A. Hanakahi, H. Sun, N. J. High affinity interactions of nucleolin with G-G-paired rDNA. J. Biol Chem, 274 (1999) :15908-15912.

[31] G. N. Parkinson, M. P. Lee, S. Neidle. Crystal structure of parallel quadruplexes from human telomeric DNA. Nature, 417 (2002): p876-880.

[32] N. Maizels. Dynamic roles for G4 DNA in the biology of eukaryotic cells Nat Struct and Mol Biol, 13 (2006): p1055-1059

[33] Z. Du, Y. Zhao, N. Li. Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. Genome Res. 18 (2008): 233-241.

[34] Struhl K. Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. Trends Biochem Sci. 1989;14(4):137-40.

[35] K. Mathee and G. Narasimhan. Detection of DNA-Binding Helix-Turn-Helix Motifs in Proteins Using the Pattern Dictionary Method. Methods in Enzymology. 2003. 370, 250-264.

[36] G. Krauss. Biochemistry of Signal transduction and regulation. Second Edition. Wiley-VCH, 2001. Chapter 12. Other Receptor Classes. 377-384.

[37] G. M. Pao, M. H. Saier. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J Mol Evol. 1995. 40(2):136-54.

[38] J. M Skerker, M. S Prasol, B. S Perchuk, Ee G Biondi, and M. T Laub. Two-Component Signal Transduction Pathways Regulating Growth and Cell Cycle Progression in a Bacterium: A System-Level Analysis. PLoS Biol. 2005, 3(10): e334.

[39] D.E. Whitworth, P.J Cock. Evolution of prokaryotic two-component systems: insights from comparative genomics. Amino Acids. 2009; 37(3):459-66.

[40] Hybiske K and Stephens RS. Exit strategies of intracellular pathogens. Nat Rev Microbiol, 2008. 6: 99-110.

[41] Dragon F et al., A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature. 417, 967-970.

[42] Bernstein KA et al., The small-subunit processome is a ribosome assembly intermediate. Eukaryot Cell. 2004 Dec;3(6):1619-26.

[43] M. Albrecht, T. Lengauer. Survey on the PABC recognition motif PAM2. Biochem Biophys Res Commun. 2004; 316(1):129-38.

[44] Nonhoff et al., 2007. Ataxin-2 Interacts with the DEAD/H-Box RNA Helicase DDX6 and Interferes with P-Bodies and Stress Granules. Mol Cell Biol 2007; 18 :1385–1396.

[45] M. H. Lamers, H. H.Winterwerp, T. K. Sixma. The alternating ATPase domains of MutS control DNA mismatch repair. EMBO J. 2003; 22(3):746-56.

[46] P. J. Coronin. Diversity of WD-repeat proteins. Subcell Biochem 2008; 48:116-23 AND Smith TF. Subcell Biochem. 2008;48:20-30.

[47] Cummings, B. S., McHowat, J. & Schnellmann, R. G. Phospholipase A2s in cell injury and death. J. Pharmacol. Exp. Ther. 294, 793–799 (2000).

Tables

Table 1: Number of distinct Pfam domains and protein coverage in several Eukaryotes. Column « Pfam domains » lists the number of different known Pfam domains on a given genome. Column « Pfam coverage » gives the percentage of proteins with at least one known Pfam domain.

Organism	Proteome size	Number of Pfam domains	Pfam Coverage
Classical eukaryotes			
A. gambiae	12347	2991	74%
A. thaliana	34517	3125	74%
C. elegans	22637	2953	65%
D. melanogaster	16224	3129	72%
D. rerio	31884	3384	84%
H. sapiens	40252	3914	68%
S. cerevisae	5862	2369	76%
Eukaryotic pathogens included in EuPathDomains database			
P. falciparum	5460	1429	53%
P. vivax	5432	1415	50%
P. yoelii	7724	1313	42%
L. major	8406	1607	49%
L. infantum	8216	1607	49%
L. brazelinsis	8310	1587	48%
T. gondi	8102	1689	43%
G. lamblia	4889	845	49%
C. parvum	3805	1123	51%
T. brucei	9895	1554	38.00%

Table 2: Newly discovered domains in different organisms. “Certif. type” indicates the type of certification: by known Pfam domains (“Pfam”), by known InterPro non-Pfam domains (“Interp.”), or by potential domains (“Pot.”); Column “All” lists the results achieved when the 3 types of certifications were combined. “Certif. dom.” denotes the number of newly certified domains, “New Interp.” indicates the number of certifications allowing us to identify a new InterPro Entry for the protein, and “New Dom. Types” indicates the number of domain types that were not previously detected in any protein of the genome. In parentheses are reported the proportions of new domains or new domain types in comparison with the number of already known domains/types.

FDR		**≤ 10%**				**≤ 20%**			
Organism	**Certif. type**	**Pfam**	**Interp.**	**Pot.**	**All**	**Pfam**	**Interp.**	**Pot.**	**All**
L. major	Certif. Dom.	231	260	165	423 (8%)	385	417	316	719 (14%)
	New Interp.	187	200	143	341	326	333	279	607
	New Dom. Types	76	80	48	118 (7%)	104	106	98	184 (11%)
P. falciparum	Certif. Dom.	250	145	50	320(9%)	348	363	130	555 (15%)
	New Interp.	192	106	40	241	284	288	114	451
	New Dom. Types	64	36	14	80 (6%)	97	92	36	144 (10%)
T. gondii	Certif. Dom.	340	169	126	466 (8%)	436	552	255	776 (13%)
	New Interp.	284	138	107	392	379	475	228	671
	New Dom. Types	90	49	39	119 (7%)	118	148	67	198 (12%)
C. parvum	Certif. Dom.	292	246	123	417 (21%)	344	379	169	554 (28%)
	New Interp.	251	187	103	342	302	313	144	473
	New Dom. Types	109	78	42	135 (12%)	131	127	60	182 (16%)
G. lamblia	Certif. Dom.	126	63	18	144 (5%)	303	279	74	386 (13%)
	New Interp.	108	54	16	126	279	146	67	357
	New Dom. Types	44	23	13	52 (6%)	75	42	25	90 (10%)
T. brucei	Certif. Dom.	348	285 *	120	515 (9%)	513	616	270	908 (16%)
	New Interp.	285	201 *	109	418	442	501	243	768
	New Dom. Types	73	48 *	29	99 (6%)	111	118	60	176 (11%)

* these results are achieved with an FDR= 12%; no certifications with FDR<10% were achieved in *T. brucei* with the Interpro non-Pfam domains

Table 3: New GO annotations on different species proteins. “Single Domains” is the number of new GO annotations supplied by a single domain certified by our approach; “Combined domains” is the number of supplementary GO annotations (different from the previous column) that can be deduced from combinations of the newly certified domain and another known or new domain; “Unannot. Prot.” is the number of proteins without any annotation for which an annotation is now proposed.

Organism	**FDR**	**Single Domains**	**Combin. With Certified Dom.**	**Unannot. Prot.**
L. major	≤ 10%	158	108	28
	≤ 20%	302	164	53
P. falciparum	≤ 10%	117	77	20
	≤ 20%	250	111	39
T. gondii	≤ 10%	226	129	38
	≤ 20%	412	165	71
C. parvum	≤ 10%	176	114	44
	≤ 20%	258	141	56
G. lamblia	≤ 10%	51	37	14
	≤ 20%	203	99	59
T. brucei	≤ 10%	246*	175	37*
	≤ 20%	402	282	79

*: results achieved with an FDR= 12%; no certifications with FDR<10% were achieved in *T. brucei* with the Interpro non-Pfam domains

Table 4: GO terms that appear as overrepresented in the newly discovered domains for each organism compared to already known ones (black boxes). *L. m* for *L. major*; *P. f* for *P. falciparum*; *T. g* for *T. gondii*; *G. l* for *G. lamblia*; *C. p* for *C. parvum* and *T. b* for *T. brucei.*

GO term	*L. m*	*P. f*	*T. g*	*G. l*	*C. p*	*T. b*
Molecular function						
Signal transducer activity	■		■		■	
Protein_kinase activity	■				■	■
Transmembrane receptor protein kinase activity	■					
Zinc ion binding		■				
Purine ribonucleotide binding		■				■
ATP binding		■				
Structure specific DNA binding			■		■	
Enzyme inhibitor activity			■		■	
GTP binding						■
Biological process						
Post-translational protein modification	■					■
Cell communication	■					
Cell surface receptor linked signaling pathway	■		■			
Signal transduction	■					
DNA replication, synthesis of RNA primer	■		■			
Glucose catabolic process						■
Cellular component						
Nucleus	■		■			
Receptor complex	■		■			
Intracellular membrane-bounded organelle		■				
Clathrin adaptor complex		■				
Phosphopyruvate hydratase complex						■

Table 5: Proportion of known and new Pfam domains in *L. major* proteins with a known *L. infantum* homolog that are also present in the known/new Pfam domains of their *L. infantum* homolog. For example, 5279 known Pfam domains are in a *L. major* protein with a known *L. infantum* homolog. Among these, 5114 are also found in the *L. infantum* homolog. Similarly, 448 new domains with a FDR<10% in *L. major* are in proteins with a known homolog in *L. infantum*. Among these, 18 are already known in the *L. infantum* homolog, and 390 are also detected in the *L. infantum* homologue with an FDR<10%.

L. Infantum				
L. major		Known domains	New dom. FDR≤10%	New dom. FDR≤20%
	Known domains	5114/5279 (97%)	31/5279 (1%)	36/5279 (1%)
	New dom. FDR≤10%	18/448 (4%)	390/448 (87%)	392/448 (88%)
	New dom. FDR≤20%	19/776 (2%)	535/776 (69%)	604/776 (78%)

Legend of figures

Figure 1: The CODD certification process. The CDP list was computed from all domain pairs found in the UniProt proteins. The protein sequences were matched against the set of 10 340 Pfam profile HMMs with loose E-value thresholds to allow detection of more domains. Domains overlapping with known Pfam domains were discarded. In the remaining overlaps, domains with the best E-value were kept. The CDP list was used for the certification process of the set of potential domains. An FDR estimation process was then carried out to assess the level of confidence of the certified domains.

Figure 2: Known and newly predicted domains of *L. major* gene LmjF36.4590. One InterPro domain is already known (SSF57903). Localization as well as known GO annotation terms are given. Three new Pfam domains PF02178, PF00628 and PF00643 are predicted with E-values of 0.93 and 0.6 and 0.27 respectively. They have been certified by several validating domains with various FDRs. For example, the second domain have been certified by the known InterPro SSF57903 with an FDR equal to 20%, and the potentials Pfam PF02178 and PF00643 with FDRs below 10%. New GO terms are transferred to this protein from the predicted Pfam PF00628 itself: “zinc ion binding” and “protein binding”. The “zinc ion binding” is already known for that protein. Combined with other predicted domains, it brings new annotations such as “regulation of transcription, DNA related”.

Figure 3: Comparison of frequency of known (blue) and newly certified (red) *L. Major* domains according to their lengths.

Figure 4: GO functions of common new domain families predicted in *L.major*, *L. infantum* and *L. braziliensis.*