MyCLADE: a multi-source domain annotation server for sequence functional exploration

Riccardo Vicedomini, Clémence Blachon, Francesco Oteri, Alessandra Carbone

W452–W458 *Nucleic Acids Research, 2021, Vol. 49, Web Server issue* *Published online 22 May 2021*
https://doi.org/10.1093/nar/gkab395

MyCLADE: a multi-source domain annotation server for sequence functional exploration

Riccardo Vicedomini[1,2], Clémence Blachon[1], Francesco Oteri[1] and Alessandra Carbone[1,*]

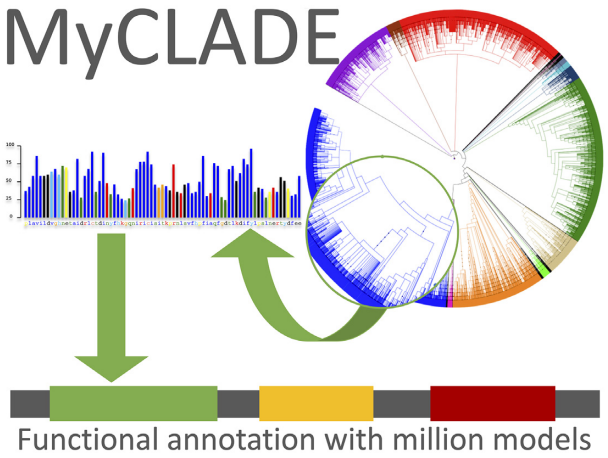
[1]Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), UMR 7238, Paris 75005, France and [2]Sorbonne Université, CNRS, Institut des Sciences du Calcul et des Données (ISCD), France

Received March 08, 2021; Revised April 27, 2021; Editorial Decision April 27, 2021; Accepted April 29, 2021

ABSTRACT

The ever-increasing number of genomic and metagenomic sequences accumulating in our databases requires accurate approaches to explore their content against specific domain targets. MyCLADE is a user-friendly webserver designed for targeted functional profiling of genomic and metagenomic sequences based on a database of a few million probabilistic models of Pfam domains. It uses the MetaCLADE multi-source domain annotation strategy, modelling domains based on multiple probabilistic profiles. MyCLADE takes a list of protein sequences and possibly a target set of domains/clans as input and, for each sequence, it provides a domain architecture built from the targeted domains or from all Pfam domains. It is linked to the Pfam and QuickGO databases in multiple ways for easy retrieval of domain and clan information. E-value, bit-score, domain-dependent probability scores and logos representing the match of the model with the sequence are provided to help the user to assess the quality of each annotation. Availability and implementation: MyCLADE is freely available at http://www.lcqb.upmc.fr/myclade.

GRAPHICAL ABSTRACT



INTRODUCTION

Current sequence databases contain hundreds of billions of nucleotides encoding genes whose functional profiling is a primary problem in genomics and metagenomics. This task is typically carried out by annotating protein domains, which are functional units, much shorter than proteins, with an average size of 100 aa (1–3). Despite their short length, they are accurate enough to inform us about the potential functional activity of a protein. Recently, a marked improvement in annotation has been achieved with the multi-source strategy, using multiple probabilistic models, compared to the mono-source strategy, using one single consensus model, employed by the two most commonly used annotation tools HMMER (4) and HHblits (5,6). CLADE (7,8) and MetaCLADE (9) rely on millions of probabilistic models for Pfam domains (10) that have proven to be more specific and functionally predictive than the widely used consensus models (11–13). They have been shown to improve domain architectures in complete genomes (8) and the catalog of functions in microbiomes (9,14). A fine degree of accuracy has been achieved in domain annotation for both prokaryotic and eukaryotic organisms.

*To whom correspondence should be addressed. Tel: +33 1 44 27 73 45; Fax: +33 1 44 27 42 52; Email: alessandra.carbone@lip6.fr

MyCLADE is an online server that allows the community to access domain annotations, based on a large dataset of probabilistic models, which enhance Pfam (10) and InterPro (15) annotations. It finds domains for proteins that are annotated for the first time, it enriches known architectures with new domains and might provide alternatives for previously annotated domains. To reconstruct the most appropriate domain architecture, MyCLADE integrates a new improved version of MetaCLADE and DAMA (16) in the same environment. It also provides hit scores, model logos and GO terms to evaluate the confidence in a domain annotation. Rebuilding such a computational environment on local machines can be a stumbling block for many users and MyCLADE offers a solution to this limitation. Since users are often interested in understanding a given function (17) or biochemical pathway (18,19) across multiple conditions or samples, in practice, only a limited number of domains (a few dozens) must be annotated during profiling. So, to answer this practical need, MyCLADE searches for a few selected domains (a specific Pfam clan or a user-provided subset) in large sets comprising thousands of sequences, or it considers all Pfam domains to annotate small sets of tens or hundreds of sequences. It can exploit a very large model library providing up to 350 models per domain or a reduced library of at most 50 models per domain. Its computational time and performance are evaluated on datasets of increasing sizes, against all domains, few domains and clans.

METHODS

Model library for Pfam domains

MyCLADE uses a probabilistic model library that includes Pfam sequence consensus models (SCM) (20–23) and at most 350 clade-centered models (CCMs) (7,9), with an average of 161 models, per domain. Both SCM and CCM are profile Hidden Markov Models (pHMM) generated using HMMER v3, http://hmmer.org.

To construct a CCM for a Pfam domain D^i, we consider the FULL set of homologous sequences S^i in Pfam (10) associated to D^i, and for some representative sequences s_j in S^i (see below), we construct a model by retrieving with HHblits (5) a set of sequences similar to s_j from the Uniprot Uniclust30 database. The probabilistic model generated in this way displays features that are characteristic of the sequence s_j and that might be very different for other sequences s_k in S^i. The more divergent the homologous domain sequences s_j and s_k are, the more models constructed from these sequences are expected to display different features. It is therefore important for a domain D^i to be represented by several models that can characterise its different pathways of evolution within different clades. These probabilistic models are the CCMs used in MyCLADE. The details of their construction are described in (9).

For a domain, representative sequences are selected in order to span the tree of life as much as possible as in (7) by considering different clades. The rationale is that evolutionary patterns can be found in species that are very far apart in the tree. For this, we considered the tree of life and fixed a list of clades. The selection of representative sequences was designed as follows: given a Pfam32 domain, all Pfam32 sequences for the domain were clustered with MMseqs2 (https://github.com/soedinglab/MMseqs2) with a default sequence identity threshold of 50%. From the list of these clusters (randomly ordered), we iteratively selected as the representative sequence of a cluster, the sequence with 1. the longest domain hit and 2. an associated species whose phylogenetic clade had not already been selected in a previous step of the iteration. Both conditions should be met and, if not, a cluster will not have a representative sequence. Once a representative sequence is selected, a CCM is generated from it and integrated into the model library. As soon as 350 CCMs are built for a domain, the procedure stops. If at the end of the analysis of the listed clusters the number of models is smaller than 350, all clusters without a representative sequence are reconsidered and sequences that satisfy only condition 1 are chosen. Note that this step allows to select sequences (possibly paralogs) sufficiently divergent from those already selected due to the initial clustering.

This is equivalent to building >2.5 million probabilistic models for the whole Pfam32 database (17 929 domains). A reduced model library was created, keeping only the first 50 models built from the aforementioned procedure.

By construction, CCMs span regions of the protein sequence space that are usually not well represented in a SCM. These regions might highlight motifs, structural or physico-chemical properties characteristic of divergent homologous sequences. Thus, if a domain is associated with many divergent homologs, the CCMs are expected to describe properties that might not be detected by a SCM. For this reason, CCMs allow to find divergent homologous sequences in species that might be phylogenetically distant. Note that the construction guarantees the spread of the species within the tree of life also for the reduced library.

The MyCLADE approach

MyCLADE integrates MetaCLADE and DAMA, and an interactive interface organises the information for an easy evaluation of the annotation confidence.

MyCLADE runs a new version of MetaCLADE (9) (MetaCLADE v2 at www.lcqb.upmc.fr/metaclade/) with a library of more than two million probabilistic domain models and an intelligent algorithmic strategy filtering the high number of hits produced by the models to retain only the most reliable ones. Each sequence is scanned with the model library in order to identify all domain hits. Each hit is defined by a bit-score, that is the HMMER score associated to the match, and by a mean-bit-score, that is the bit-score of the hit divided by its length. These two scores are used to evaluate the probability p of the hit to represent a true annotation (computed after a domain-dependent estimation). The output of this first step of MetaCLADE is a set of hits, each one defined by a domain family D, a probabilistic model M associated to D, a bit-score and a mean-bit-score. Since each domain can be represented by a large number of models, a large number of domain hits might be associated to each sequence. MyCLADE uses three criteria to filter them, based on the bit-score, the probability p of being a true positive, and the identity percentage of the hit (9). The output of this filtering step is the sequence annotation where hits of different domains might overlap for at most 30

aa and hits of the same domain are non-overlapping. Compared to the original version of MetaCLADE (9), in MetaCLADE v2: (i) the code has been optimized, (ii) the selection of representative sequences spanning the tree of life follows a new strategy, as described above, (iii) the search for similar sequences is realized with `hhblits` instead of `psiblast` for the construction of HMMs instead of PSSMs.

Domain co-occurrence is expected to enhance the level of confidence in a domain prediction (7,24,25). Intuitively, co-occurrence suggests functional cooperation, that is, two or more domains can interact to determine the protein function (26–28). Once domains are selected, the user can decide to call DAMA (29) (www.lcqb.upmc.fr/dama/), a tool that considers domain co-occurrence and domain overlapping, and that combines several domains into most probable architectures.

For each domain annotation, MyCLADE provides *E*-value, bit-score and domain dependent probability score (ddProb) to allow the user to evaluate the confidence. Significant amino acid residues and subtle sequence patterns can be visualised through model logos aligned against the annotated sequences.

MyCLADE input and parameters

MyCLADE can be run on three different library types defined by: (i) few (at most 10) user-provided domains, (ii) all Pfam32 domains, (iii) a Pfam32 clan. Each of the three sets of domains characterizes a library of models that is based on either 350 or 50 models per domain in the set. An option allows the user to decide on the number of models per domain.

The first and third library types require a list of up to 2000 sequences in FASTA format (possibly uploaded). The second library type requires a small dataset of up to 200 sequences. The list of input sequences is checked for format requirements. The help section in the online interface provides information on the expected format.

An option allows the user to filter out all hits with an *E*-value <1 which is greater than the chosen threshold (by default set at 1e−3). The reconstruction of the best domain architecture is possible by selecting the DAMA option, together with its three dedicated parameters: an *E*-value (set by default at 1e−10), a possible overlap size between domains (set by default at a maximum of 30 aa) and an allowed domain overlap percentage (set by default at a maximum of 50%). By changing the parameters, the user can explore potentially new annotations. When MyCLADE is run without DAMA, domain overlapping is allowed for at most 10aa as in MetaCLADE. The criteria used in MetaCLADE to accept an overlapping domain among different domain hits are described in detail in (9).

Logos can be generated for all domain hits in MyCLADE annotations. The user can decide to build logos either by marking the corresponding input option or after building the architectures.

The user can provide an e-mail address to obtain an identifier to access the data online after the job is completed.

Annotation files produced in previous run of MyCLADE can be provided as input and displayed graphically in the webserver.

MyCLADE output

MyCLADE outputs are organised in two main pages, 'Results' and 'Architecture'.

The 'Results' page describes MyCLADE annotations by listing the input sequences with their: sequence id, list of distinct annotated domains, best e-value obtained within domain hits, number of domain hits including domain repetitions (Figure 1A). By hovering over the domain name, a tooltip synthesizes information on the domain annotation. The number of sequences with either no hit or at least one hit among the total number of input sequences is given.

The 'Architecture' page is accessible from the 'Results' page by clicking on a sequence id. It displays an interactive graphical representation of the domain architecture of the sequence with the description of the annotation (Pfam family, initial and final position of the domain in the sequence corresponding to the HMMER envelope, the species from which the probabilistic model used to annotate was generated, the E-value, the bit-score and the domain-dependent probability scores of the hit), the associated GO-terms (with clan identifier and clan family) if available, and the logo of the match between the model and the sequence displaying significant amino acid residues and conserved patterns (Figure 1B−E). A presentation of the information on the annotated domain is also available through a tooltip, by hovering the mouse over the graphical representation of the domain. Three tables collect all details of domain annotation, the associated GO-terms (31) and the logos matching the hits. Multiple links to Pfam (10) and QuickGO (32) databases are available for an easy retrieval of general domain and clan information (Figure 1A,B,D). For each domain hit, start and end positions of the model hit against the sequence are reported in the table presenting the logos. Note that these positions are different than the envelope positions, usually corresponding to a larger interval.

Construction of the testing dataset

The Harmonizome database (http://amp.pharm.mssm.edu/Harmonizome/ dataset/InterPro+Predicted+Protein+Domain+Annotations) (33) contains a set of 18 002 genes organised in 11 015 domains annotated with Interpro (15). Pfam domains were randomly selected and the associated protein sequences were retrieved to create a FASTA file for testing: 200 sequences were recovered for a total of 64 domains. The set of sequences was subdivided into subsets to evaluate MyCLADE time complexity on very small (5, 10, 25 and 50) and larger (100 and 200) sets of sequences.

Domain (re)annotation of genomes

The *Staphylococcus aureus* genome comprises 2767 protein sequences, having average size of 282 amino acids (aa). MyCLADE annotation was compared with HMMER (`hmmscan`) annotation (34). Two overlapping domains, identified by the two methods, were considered the same if they belonged to the same Pfam clan. The genome was downloaded at NCBI (https://www.ncbi.nlm.nih.gov).

Figure 1. Visualization of MyCLADE interfaces. Examples of the 'Results' page (A) and of the 'Architecture' page (B,C,D,E). (**A**) The table summarizes MyCLADE annotation on four different sequences. By hovering over a domain name, a synthesis of the annotation for that domain hit is reported in a tooltip. (**B**) The architecture of the 511 aa long protein (the human AMY1A) is graphically represented with colored domains. Two of the domains are overlapping. A synthetic table lists the domains and some associated information. Links to Pfam are highlighted in green. (**C**) Zoom on the architecture in B; a pop-up box appears by passing the mouse on the domain name. (**D**) A second table lists the GO-terms for each domain, with additional information on the clan (clan identifier and clan family identifier) to which the domain belongs. A 'NA' value means that the domain is not yet linked to a clan or a GO-term. Multiple links to Pfam and QuickGo databases are highlighted in green. (**E**) For each domain, MyCLADE provides a logo matched to the sequence. Zoom on the logo associated to the PF02806 domain. Hovering over the sequence provides the position number in the sequence matching the position in the logo.

Logos

Logos have been generated with hmmlogo in HMMER (http://hmmer.org) and Logomaker (30). hmmlogo computes letter height of amino acids in a position of a HMM model and Logomaker produces the logo image. The alignement between the HMM model and the sequence is generated with hmmsearch in HMMER. Within a logo, the height of the stack of letters corresponds to the conservation at that position, and the height of each letter within a stack depends on the frequency of that letter at that position. Residues are colored according to the ClustalX coloring scheme grouping amino acids by their physico-chemical properties: glycine (G) in orange, proline (P) in yellow, small or hydrophobic (A, V, L, I, M, F, W) in purple, hydroxyl or amine amino acids (S, T, N, Q) in green, charged amino-

acids (D, E, R, K) in red and histidine or tyrosine (H,Y) in cyan. The symbol '*' shows a perfect match between the most frequent letter in the logo and the letter in the sequence, and the symbol '+' shows that the letter in the sequence and the most frequent letter in the logo share the same physico-chemical group.

RESULTS

The performance of MetaCLADE has been extensively evaluated across multiple genomic and metagenomic datasets (9). On a dedicated page of the MyCLADE server, five annotations of known and less known proteins are discussed: the Mediterranean Fever gene MEFV and its pyrin protein, the human amylase AMY1A, a UDP-*N*-acetyltunicamine-uracil synthase TunB-like protein (B5GL39), the hypothetical protein YP_499998 in *Staphylococcus aureus* and the RING finger protein in *Plasmodium falciparum* PFE0100w. These case studies illustrate how the user can explore domain annotations and can acquire confidence in them. We also discuss the motivation to search for specific domains in datasets of a few thousand sequences and give an example of the application of MyCLADE to a case study. Here, we (re)annotate the entire *Staphylococcus aureus* genome with MyCLADE for the discovery of new domains. Additionally, we assessed the time complexity of MyCLADE on different datasets based on the complete and restricted model libraries.

Evaluation of the execution time of MyCLADE

MyCLADE can annotate sequences with a limited number of targeted domains or with all Pfam32 domains. A runtime evaluation of MyCLADE was performed on these two use cases (Figure 2) for small (5, 10, 25 and 50 sequences) and large (100 and 200 sequences) protein datasets. On targeted domains, from 10 to a few hundred, MyCLADE annotates hundreds of sequences in less than a minute while on all Pfam32 domains it takes less than a hour. The estimates are realized with the complete model library. The three plots in Figure 2 (for 10 and 340 domains in Figure 2A and all domains in Figure 2B) show that by augmenting the number of sequences, for instance from 100 to 200, the ratio of the execution times decreases when the number of models increases: 3.16 on ∼2000 models for 10 domains, 2.5 on roughly 60 000 models for 340 domains and 1.2 on more than 2.5 million models, for all 17 929 domains of the Pfam32. This means that the bottleneck of the annotation pipeline resides on the alignment of the models against the sequences and not on the number of sequences. Therefore, the user should expect essentially the same time of execution for the annotation of few sequences (e.g. 1, 10, 20) with a given set of domains. To evaluate the dependence of the computation time on the number of available models, we considered the restricted library and observed that the computation time is greatly reduced: on 10 domains it is reduced by a factor of 6, on 340 domains by a factor of 3.65 and on 17 929 domains of a factor of 2.3 (Figure 2). The time reduction is obtained at the cost of an expected decrease of the number of annotated domains: on the 200 sequences of the test dataset, 637 domains are annotated with the complete library and 561 with the restricted one. MyCLADE

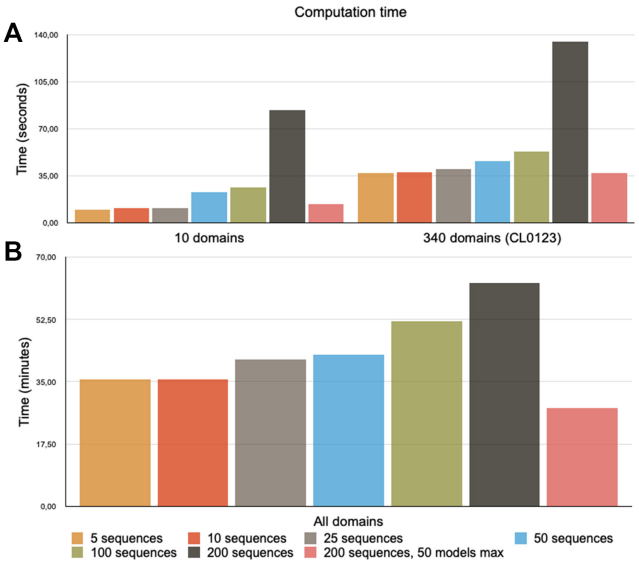


Figure 2. Performance analysis. MyCLADE time performance is evaluated on sets of sequences of various size with (**A**) either 10 domains, or 340 domains belonging to the largest Pfam clan CL0123 (in seconds) or (**B**) with all Pfam domains (in minutes). A dataset of 200 sequences, of length varying from 494 aa up to 671 aa, has been constructed from the Harmonizome database and three different subsets of 5, 10, 25, 50 and 100 sequences have been randomly extracted from it and tested. The plot reports the computing time needed for MyCLADE annotation without DAMA averaged over the three datasets for each experiment. The sets have been evaluated with the complete library of models (<350 models per domain) and with a reduced library (<50 models per domain).

performance has been evaluated without DAMA because DAMA computing time is negligeable (for a few hundred proteins, the architecture reconstruction takes less than a few seconds) as described in Table 2 of (16).

Annotation of the *Staphylococcus aureus*

The emergence of antibiotic-resistant strains of *S. aureus* is a worldwide problem in clinical medicine (35). MyCLADE, run with DAMA, annotates 4,184 domains versus the 4,374 domains annotated by HMMER (`hmmscan`). 1220 HMMER annotated domains overlap for >30 up to 104 aa while MyCLADE annotations allow for a limited overlap of 30 aa. In addition, MyCLADE annotates 809 new domains, 459 of which occupy regions which were left with no annotation by HMMER. Moreover, MyCLADE annotations based on CCM models (Figure 3 A) provide smaller E-values than the corresponding HMMER annotations based on HMMER v3 models increasing confidence in the predictions. Note that only 940 MyCLADE domain annotations were identified from HMMER v3 models (Figure 3B, C) and all others were best identified by CCMs generated mostly by bacterial sequences (2533) but also by sequences classified in Fungi, Viridiplantae and Metazoa clades (Figure 3C). When MyCLADE is run without DAMA, it identifies a total of 3849 domains, 682 of which are new domains.

SOFTWARE AVAILABILITY

For the analysis of large files, not allowed in MyCLADE, the user can locally install MetaCLADE (www.lcqb.upmc.

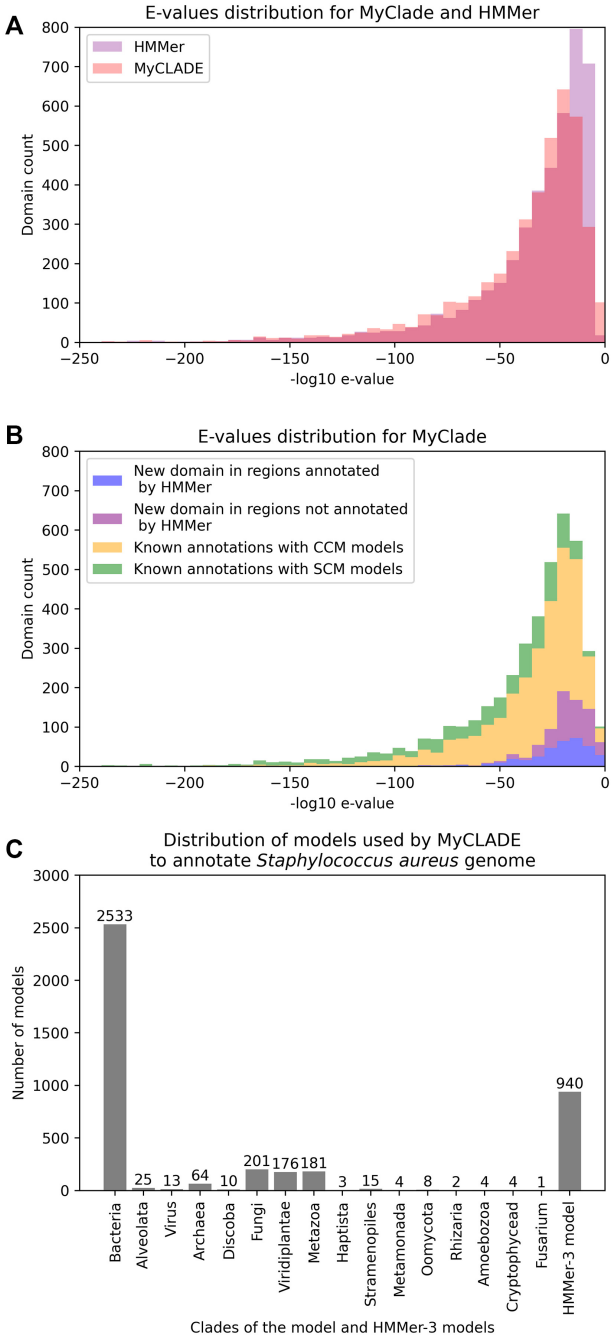


Figure 3. Analysis of the domain annotation of *Staphylococcus aureus* genome. (**A**) Distribution of best E-values for MyCLADE (pink) and HMMER (`hmmscan`) (purple) domain annotations of the *S.aureus* proteins. The two transparent shapes are superimposed and they show that MyCLADE E-values are shifted on the left towards smaller values highlighting an increased confidence in the annotation. (**B**) Domains are organised by MyCLADE best E-values and partitioned in four classes: domains annotated with CCMs and identified with SCMs by HMMER (yellow), domains annotated with SCMs (green), new domains identified in regions annotated by HMMER differently (blue), new domains identified in regions with no HMMER annotation (purple). The plot shows the cumulative distribution of the four classes. The shape of the distribution is the same as in (A). (**C**) Distribution of the species generating the CCMs which were selected by MyCLADE for the annotation of the *S.aureus* genome. They are organised by phylogenetic clades. The number of annotated domains identified with SCMs (generated by HMMer v3) is reported in the last column.

fr/metaclade/) along with the corresponding model library. DAMA can be retrieved at www.lcqb.upmc.fr/dama/. Note that DAMA uses Pfam knowledge for building its architectures and the user can check whether the domains in a MyCLADE architecture are known to co-occur or not at http://pfam.xfam.org/search (on the 'domain architecture search').

DISCUSSION

MyCLADE is a server that provides an online version of MetaCLADE v2. It gives the user access to a quick way to annotate protein sequences by exploring a wide range of probabilistic models that can achieve more accurate domain annotation than more classic approaches. Also, several options allow to filter out domain hits with E-value greater than the chosen threshold, the reconstruction of the best domain architecture is parametrizable, the user can provide an e-mail address to obtain an identifier and access the data online after job completion, annotation files produced in previous runs of MyCLADE can be provided as input and displayed graphically with the server.

With MyCLADE, the user can explore domain annotations to search for new domains and possibly to find hints for a functional annotation. He/she can search for the best domain hits (without DAMA) or for the best domain architecture based on already observed combinations of co-occurring domains (with DAMA). These two modes lead to two important observations on domain overlapping and on functional annotation. First, domain overlapping is an important problem in the reconstruction of domain architectures and it should be remarked that DAMA has been designed to solve a combinatorial optimization problem taking into consideration various parameters, including co-occurrence, to accept overlapping annotations. In order to explore potential annotations without DAMA, MyCLADE accepts overlaps up to 10 aa in length, the same threshold used in MetaCLADE and a more binding condition compared to the 30 aa accepted by DAMA. Second, protein function for a combination of domains is not easily deducible based on the functions of single domains. Yet, the use of known domain combinations and the access to GO-terms for single domains can help the user for a fast exploration of potential protein functions (13) with MyCLADE.

FUNDING

LabEx CALSIMLAB [ANR-11-LABX-0037-01 constituting a part of the "Investissements d'Avenir" program - reference : ANR-11-IDEX-0004-02 to R.V.]. Funding for open access charge: LabEx [ANR-18-CE13-0004].
Conflict of interest statement. None declared.

REFERENCES

1. Janin,J. and Wodak,S. (1983) Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Molec. Biol.*, **42**, 21–78.
2. Richardson,J. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.
3. Xu,D. and Nussinov,R. (1998) Favorable domain size in proteins. *Structure*, **3**, 11–17.

4. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
5. Remmert,M., Biegert,A., Hauser,A. and Soeding,J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
6. Soeding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
7. Bernardes,J., Zaverucha,G., Vaquero,C. and Carbone,A. (2016a) Improvement in protein domain identification is reached by breaking consensus, with the agreement of many profiles and domain co-occurrence. *PLoS Comput. Biol.*, **12**, e1005038.
8. Bernardes,J., Vaquero,C. and Carbone,A. (2017) Plasmobase: a comparative database of predicted domain architectures for *plasmodium* genomes. *Malaria J.*, **16**, 241.
9. Ugarte,A., Vicedomini,R., Bernardes,J. and Carbone,A. (2018) A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. *Microbiome*, **6**, 149.
10. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
11. Fortunato,A.E., Jaubert,M., Enomoto,G., Bouly,J-P., Raniello,R., Thaler,M., Malviya,S., Bernardes,J.S., Rappaport,F., Gentili,B. *et al.* (2016) Diatom phytochromes reveal the existence of far-red-light-based sensing in the ocean. *Plant Cell*, **28**, 616–628.
12. Briquet,S., Ourimi,A., Pionneau,C., Bernardes,J., Carbone,A., Chardonnet,S. and Vaquero,C. (2018) Identification of *Plasmodium falciparum* nuclear proteins by mass spectrometry and proposed protein annotation. *PLoS One*, **13**, e0205596.
13. Vicedomini,R., Bouly,J-P., Laine,E., Falciatore,A. and Carbone,A. (2021) Multiple probabilistic models extract features from protein sequence data and resolve functional diversity of very different protein families. bioRxiv doi: https://doi.org/10.1101/717249, 09 March 2021, preprint: not peer reviewed.
14. Amato,A., Dell'Aquila,G., Musacchia,F., Annunziata,R., Ugarte,A., Maillet,N., Carbone,A., D'Alcalà,M.R., Sanges,R., Iudicone,D. *et al.* (2017) Marine diatoms change their gene expression profile when exposed to microscale turbulence under nutrient replete conditions. *Sci. Rep.-UK*, **7**, 3826.
15. Mitchell,A., Chang,H.-Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
16. Bernardes,J.S., Vieira,F.R.J., Zaverucha,G. and Carbone,A. (2016c) A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics*, **32**, 345–353.
17. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S., Sharma,A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.
18. Tagliabue,A., Bowie,A.R., Boyd,P.W., Buck,K.N., Johnson,K.S. and Saito,M.A. (2017) The integral role of iron in ocean biogeochemistry. *Nature*, **543**, 51–59.
19. Vital,M., Karch,A. and Pieper,D.H. (2017) Colonic butyrate-producing communities in humans: an overview using omics data. *mSystems*, **2**, e00130-17.
20. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) In: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
22. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
23. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 4355–4358.
24. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) Cdart: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.
25. Vogel,C., Berzuini,C., Bashton,M., Gough,J. and Teichmann,S.A. (2004) Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.*, **336**, 809–823.
26. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mole. Biol.*, **310**, 311–325.
27. Marcotte,E.M., Pellegrini,M., Ng,H.-L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
28. Wuchty,S. and Almaas,E. (2005) Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.*, **5**, 24.
29. Bernardes,J., Vieira,F., Zaverucha,G. and Carbone,A. (2016b) A multi-objective optimisation approach accurately resolves protein domain architectures. *Bioinformatics*, **32**, 345–353.
30. Tareen,A. and Kinney,J. B. (2020) Logomaker: beautiful sequence logos in Python. *Bioinformatics*, **36**, 2272–2274.
31. Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
32. Binns,D., Dimmer,E., Huntley,R., Barrell,D., O'donovan,C. and Apweiler,R. (2009) QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**, 3045–3046.
33. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Maayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**, baw100.
34. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. and Finn,R.D. (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.
35. Fuchs,S., Mehlan,H., Bernhardt,J., Hennig,A., Michalik,S., Surmann,K., Pané-Farré,J., Giese,A., Weiss,S., Backert,L. *et al.* (2018) AureoWiki - the repository of the staphylococcus aureus research and annotation community. *Int. J. Med. Microbiol.*, **308**, 558–568.