RESEARCH ARTICLE

HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold

Inti Anabela Pagnuco[1], María Victoria Revuelta[2¤], Hernán Gabriel Bondino[2], Marcel Brun[1], Arjen ten Have[2]*

1 Laboratorio de Procesamiento Digital de Imágenes, Instituto de Investigaciones Científicas y Tecnológicas en Electrónica (ICyTE), Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina, **2** Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

¤ Current address: Department of Medicine, Hematology and Oncology Division, Weill Cornell Medicine, New York, New York, United States of America
* atenhave@mdp.edu.ar

Citation: Pagnuco IA, Revuelta MV, Bondino HG, Brun M, ten Have A (2018) HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS ONE 13(3): e0193757. https://doi.org/10.1371/journal.pone.0193757

Editor: Andy T. Y. Lau, Shantou University Medical College, CHINA

Received: June 16, 2017

Accepted: February 4, 2018

Published: March 26, 2018

Data Availability Statement: All relevant data and source codes are available from the Github repository at the following URL: https://github.com/BBCMdP/HMMERCTTER.

Funding: IAP was a Consejo Nacional de Investigaciones Científicas y Técnicas doctoral fellow. MVR was a Consejo Nacional de Investigaciones Científicas y Técnicas doctoral fellow. HB was a Consejo Nacional de Investigaciones Científicas y Técnicas postdoctoral

Abstract

Background

Protein superfamilies can be divided into subfamilies of proteins with different functional characteristics. Their sequences can be classified hierarchically, which is part of sequence function assignation. Typically, there are no clear subfamily hallmarks that would allow pattern-based function assignation by which this task is mostly achieved based on the similarity principle. This is hampered by the lack of a score cut-off that is both sensitive and specific.

Results

HMMER Cut-off Threshold Tool (HMMERCTTER) adds a reliable cut-off threshold to the popular HMMER. Using a high quality superfamily phylogeny, it clusters a set of training sequences such that the cluster-specific HMMER profiles show cluster or subfamily member detection with 100% precision and recall (P&R), thereby generating a specific threshold as inclusion cut-off. Profiles and thresholds are then used as classifiers to screen a target dataset. Iterative inclusion of novel sequences to groups and the corresponding HMMER profiles results in high sensitivity while specificity is maintained by imposing 100% P&R self detection. In three presented case studies of protein superfamilies, classification of large datasets with 100% precision was achieved with over 95% recall. Limits and caveats are presented and explained.

Conclusions

HMMERCTTER is a promising protein superfamily sequence classifier provided high quality training datasets are used. It provides a decision support system that aids in the difficult task of sequence function assignation in the twilight zone of sequence similarity. All relevant data

fellow. AtH is a Consejo Nacional de Investigaciones Científicas y Técnicas career researcher. The project was funded by PIP11420100100286 from the Consejo Nacional de Investigaciones Científicas y Técnicas (AtH) and PICT2013-2296 from the Fondo para la Investigación Científica y Tecnológica (AtH and MB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and source codes are available from the Github repository at the following URL: https://github.com/BBCMdP/HMMERCTTER.

Introduction

Protein sequence function annotation is one of the major tasks of computational genomics. The most widely applied tools are based on the similarity principle: the higher the similarity between sequences, the higher the probability these have the same function. For instance BLAST [1] is a search machine, routinely used by biologists in order to identify the function of their query sequences. Although the functional protein sequence space conforms only a minor part of the polypeptide sequence space, similarity based sequence annotation is hampered by many problems.

The requirement of a set of reliably annotated sequences is fundamental and despite the steady development of UniProt [2], incorrectly annotated sequences form a major obstacle in sequence function annotation. A second problem is that of the high sequence variation that apparently is allowed for large numbers of protein families. The resulting sensitivity problem becomes more apparent when protein superfamilies are considered. The evolution of proteins has for a large part been instigated by gene duplications and the resulting process of functional redundancy and diversification [3]. Paralogs can obtain novel functions due to relaxed functional constraints, often while maintaining their original function. All together this results in intricate superfamilies where function annotation by similarity scoring is hampered by problems of sensitivity and specificity combined with imperfect annotation of reference sequences. This problem increases when taking into account the fact that, in the post genome era, biologists want to obtain annotations at the subfamily level, rather than the superfamily level. In other words, queries are performed to identify orthologs rather than homologs. Pattern based search strategies, such as provided by, for instance, Prosite [4] can be applied to increase specificity, as for instance when combined with BLAST [5].

HMMER [6] is another tool for sequence function annotation. Rather than comparing a query with a reference sequence database, it uses a database of mathematical profiles that describe multiple sequence alignments of protein families. Besides that the higher information usage generates a higher sensitivity, profile databases are more easily curated and improved curation results in significantly improved annotation. Several web servers that use HMMER to search their profile databases exist, all based on different objectives and principles. Pfam [7] is a collection of domain profiles whereas Superfamily [8] describes proteins based on the Structural Classification of Proteins [9]. Although both platforms show moderate levels of hierarchical organization, the objective of these major function annotation tools is to annotate at a superfamily rather than a subfamily level.

A more recent trend is that of phylogenomics [10] and hierarchical sub-clustering of superfamilies. Phylogenomics is based on the correct idea that phylogeny allows for a better clustering than similarity since it is based on evolutionary models. Since phylogeny is a form of hierarchical clustering, it also allows for the identification of subfamilies. A number of algorithms and phylogenomics-based sequence annotation platforms have been developed in the last two decades. RIO [11] is dedicated to the identification of orthologs and paralogs using bootstrapping to calculate a confidence value for orthology. SCI-PHY [12] and GEMMA [13] use agglomerative clustering. SCI-PHY implements a subfamily encoding cost minimization to define sub-clusters whereas GEMMA's sub-clustering is based on E-value. The coding cost minimization algorithm from SCI-PHY has been applied to Pfam and the identified subfamilies were analyzed for function shifts by means of the identification of Specificity Determining Positions (SDPs), resulting in the FunShift database [14]. GEMMA was used to subcluster the CATH-Gene3D resource resulting in FunFHMMER [15]. Here an automated cut-off was provided by a functional coherence index, also based on SDPs. CDD [16] is a domain database that includes an automated hierarchical classification of subfamilies, based on Bayesian analysis of what basically constitute SDPs [17]. Sifter [18,19] uses an empirical Bayesian approach to combine function evidence from for the Gene Ontology Annotation (GOA) [20] database with a phylogenetic tree. Panther [21] is a database with curated protein families that, besides GOA data and phylogenetic trees, includes manually curated metabolic pathways. Hence, basically phylogenomics platforms apply SDPs, either computationally predicted or empirically identified, to obtain more specific partitions.

Since the ultimate goal of the above mentioned methods is to provide HMMER databases that cover functional protein space, they depend on heuristics. Partitions and classifications might therefore contain errors when compared with the real phylogeny. Furthermore, most of the methods are fully automated, which can result in clusters that do not correspond with functional clusters as determined by experts, the identified clusters being either too small or too large. More important however, is the fact that current sequence function annotation methods lack a cut-off that results both in high precision and in high recall. Certain platforms, such as Pfam, use curated trusted thresholds providing high precision. Others, such as Panther and CDD, are redundant and show either the hit with best E value or all significant hits. Combining high specificity with high sensitivity is arguably problematic.

Interestingly, the high information usage of HMMER has principally been deployed to increase sensitivity whereas in principle high information content can also be applied to increase specificity. Basically, HMMER aligns a sequence to an MSA and computes a score of residue-profile correspondence. The variation among the sequences, which take part in the underlying MSA, affects the score of a query-profile alignment. Highly conserved sites have high information content and will give either high rewards or high penalties. On the other hand, highly variable sites have low information content and will hardly contribute to the total score. Thus, a HMMER profile made from a variable superfamily-MSA will be less specific than a HMMER profile made from a conserved subfamily-MSA. Thus, representing large, complex superfamilies by the various subfamilies' HMMER profiles will result in higher specificity, while presumably maintaining high sensitivity. Although this idea has been used by phylogenomics resources described above, it has not been used explicitly for the determination of a reliable cut-off. We developed a semi-automated, user-supervised procedure and pipeline that splits a superfamily into component subfamilies with the primary objective to cluster and classify its sequences with high precision and recall (P&R). Using a high quality phylogeny, HMMER Cut-off Threshold Tool or HMMERCTTER automatically identifies monophyletic sequence clusters that show self detection with 100% P&R in a HMMER screening. It does so by generating HMMER profiles from cluster-specific MSAs that by hmmsearch identify all the cluster's sequences with a score higher than that of any other sequence provided by the training set. Note that although all clusters are 100% P&R, not all training sequences are necessarily clustered. The cluster-specific score threshold provides a specific inclusion cut-off. Subsequently, these clusters can be accepted or rejected by the user assisted with information presented in hmmsearch score plots. In the classification phase, target sequences are classified using searches with the cluster-specific HMMER profiles and the established cut-off threshold as classifiers. Both profiles and corresponding cut-offs are iteratively updated upon the inclusion of novel sequences during an automated and a subsequent user-controlled classification, while imposing 100% P&R self detection. Note that not necessarily all homologs become classified.

The pipeline, which connects various existing softwares, is briefly described and demonstrated by detailed case studies of the alpha-crystallin domain (ACD) protein, the polygalacturonase (PG) and the phospholipase C (PLC) superfamilies. In the near future, HMMERCTTER will be extended towards the analysis of complete proteomes.

Results

Design of method and pipeline

Fig 1 outlines the HMMERCTTER training procedure and target sequence analysis, which are described in brief below and in detail in S1 Appendix. The training sequences are clustered using a user provided phylogeny. All possible monophyletic clusters are determined, sorted by size and tested as follows. The cluster's sequences are aligned and used to generate a HMMER profile that is subsequently used to screen the cluster's sequences as well as all training sequences. Obtained HMMER scores are compared and 100% P&R self detection is obtained when the lowest scoring cluster sequence has a higher score than the highest scoring non-cluster sequence. 100% self detection P&R clusters are provisionally accepted whereas non 100% P&R clusters are automatically rejected.

An interface showing score plots of cluster and training sequences as well as a tree with the provisional clustering is presented as shown in S1A Fig, at which point the user can reject or accept the cluster. Upon rejection, the program proceeds with the next cluster on the size-ordered list. Upon acceptance of a cluster, all its nested and overlapping clusters are removed from the list, and the program proceeds with the next cluster in the sorted list until no more clusters are encountered. This yields a number of clusters that show 100% P&R self detection in HMMER profiling as well as, possibly, a number of unclustered orphan sequences.

The HMMER profiles and corresponding cut-off scores form the initial classifiers that are used for screening the target dataset. In order to clarify whether we refer to the clustering or the classification phase of the pipeline, a cluster results from the clustering phase whereas a group is the result of the classification phase. Sequences with scores equal or above the cluster threshold are automatically accepted and added to the cluster, forming a group. We refer to these sequences as prior positives since they were not yet included in the cluster or group when tested. Sequences are realigned to construct a new HMMER profile with a new cut-off score in order to obtain higher sensitivity in subsequent HMMER profiling. As such, groups remain 100% P&R provided classification overlap is prevented. When a target sequence becomes classified by more than one group, all involved groups are excluded from subsequent iterations. Conflicting training sequences are removed from all but the original group whereas conflicting target sequences are removed from all groups and target dataset.

This automated step of classification terminates upon data convergence, when no novel sequences with a score above the threshold are identified. Hitherto, all accepted sequences were accepted based on a prior inclusion HMMER cut-off threshold, i.e. by a HMMER profile that did not include the to be accepted sequence(s). However, certain sequences might only be accepted once their information has been included into the profile, i.e. according to a posterior inclusion HMMER cut-off threshold. Hence, in the subsequent classification step, sequences with a score below the threshold are considered. Candidates are included in the group and tested with a novel HMMER profile that includes the candidate. An interactive interface (S1B Fig) allows the user to guide this process while 100% P&R self detection remains imposed and classification conflicts remain prohibited as described for the automated phase. The process is

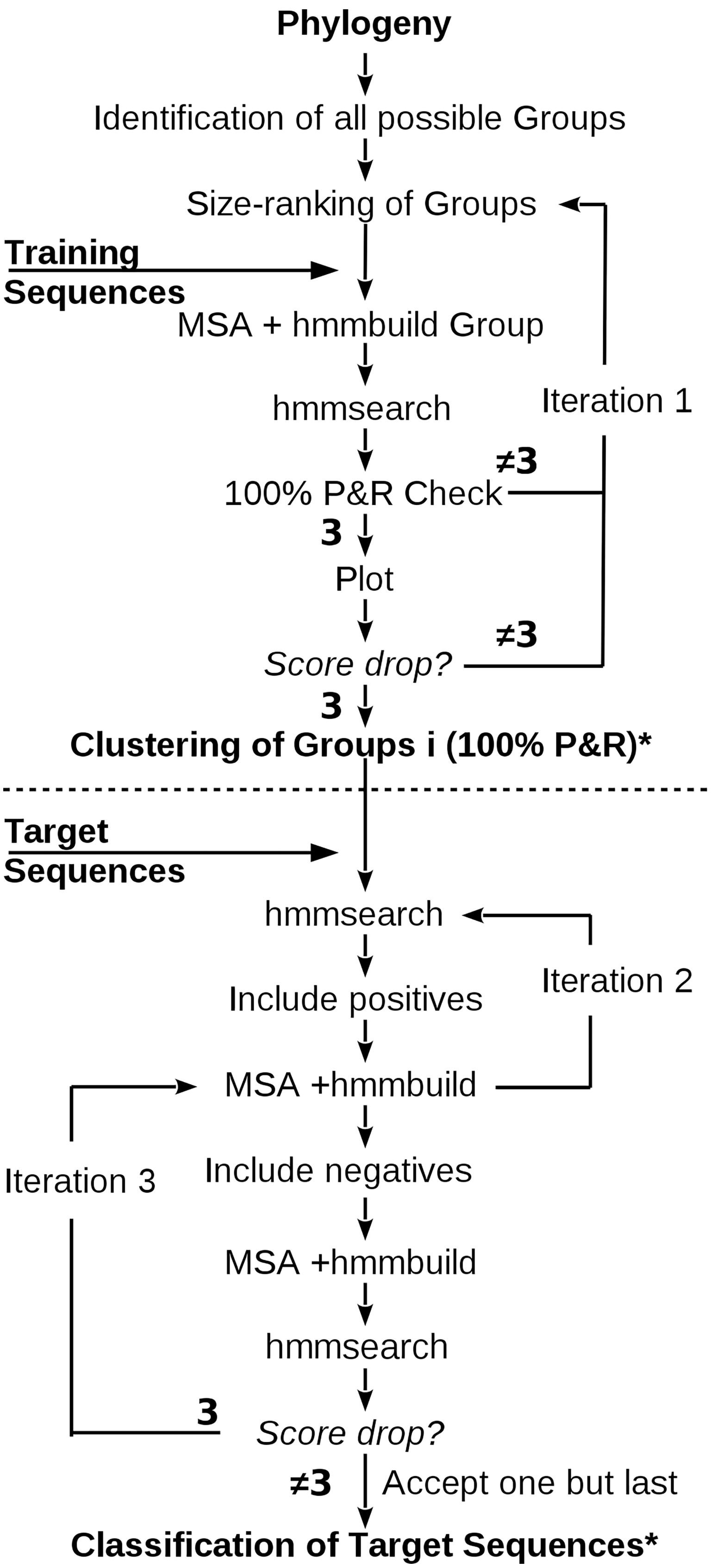

Phylogeny
Identification of all possible Groups
Size-ranking of Groups
Training Sequences
MSA + hmmbuild Group
hmmsearch
Iteration 1
≠3
100% P&R Check
3
Plot
≠3
Score drop?
3
Clustering of Groups i (100% P&R)*
Target Sequences
hmmsearch
Iteration 2
Include positives
MSA +hmmbuild
Iteration 3
Include negatives
MSA +hmmbuild
hmmsearch
3
Score drop?
≠3
Accept one but last
Classification of Target Sequences*

Fig 1. Flowchart of HMMERCTTER pipeline. Training and target phase are separated by the dotted line. Monophyletic clusters of the training set are tested for 100% P&R self detection, in descending size order. Iteration 1 is performed when a group is not accepted (either automatically or by user intervention) and the procedure is repeated with a smaller monophyletic group until no more groups are available for analysis. Accepted groups with corresponding HMMER profile and specific cut-off, defined by the 100% P&R self detection rule, are used later to classify target sequences. Automated iteration cycle 2 is performed upon inclusion of sequences with prior 100% P&R. Upon convergence and user acceptance, supervised iteration 3 includes seemingly negatives upon a test for posterior 100% P&R, i.e. upon construction of a novel profile. Note that iteration 2 is nested inside iteration 3, albeit user controlled. * indicates that final clustering and classification do not necessarily show 100% coverage of the corresponding sequence space.

https://doi.org/10.1371/journal.pone.0193757.g001

terminated by the user, resulting in updated groups and a file that indicates which sequences generated conflicts.

Algorithm performance

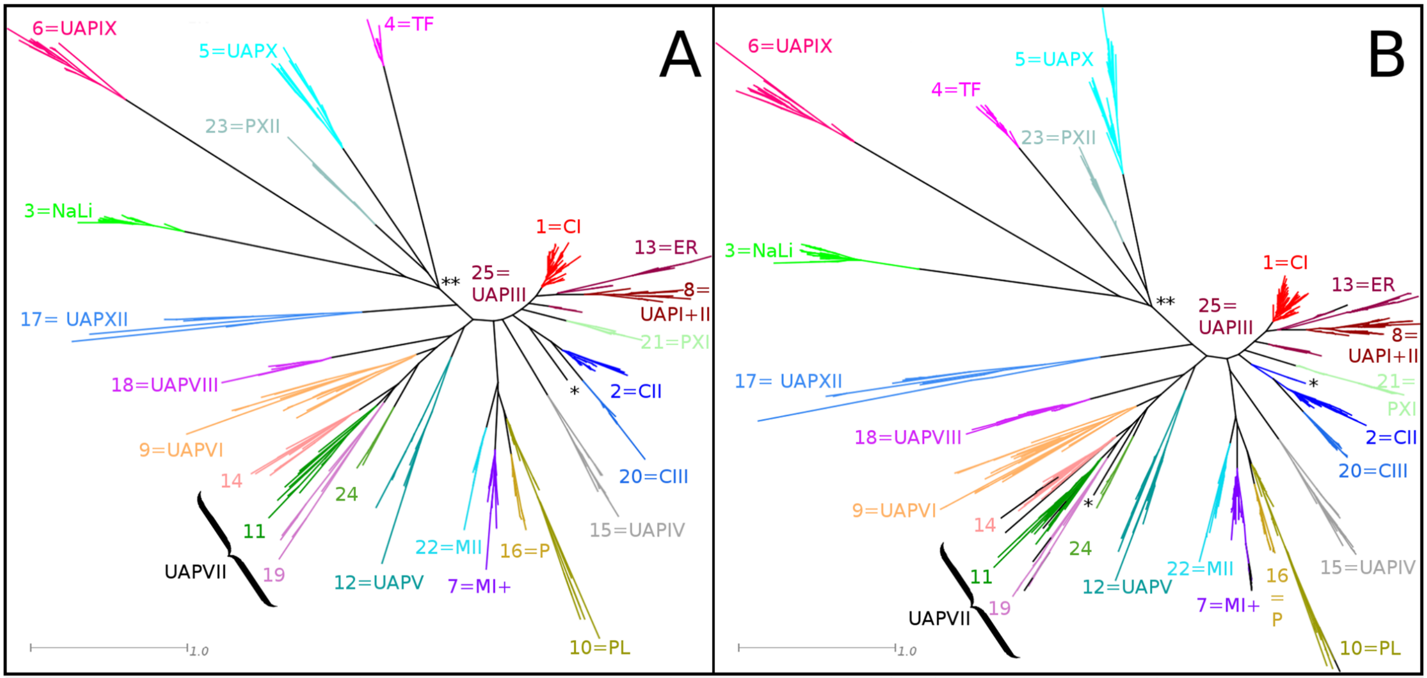
We set out to test the pipeline using three protein superfamilies. The major objective was to identify putative problems and limits of HMMERCTTER and to survey the general applicability of the procedure. In all cases classification was performed with optimal coverage of sequence space, as primary criterion. Manual override during classification (i.e. rejecting group updates) was applied only when the drop in the HMMER score decreased by at least an order of magnitude. Performance was measured as classification recall of the corresponding clade in the final reference tree (TP/(TP + FN)). Overall recall was determined as the micro average of recall as determined by

$$\frac{\sum_{\mathrm{i}=1}^{l} \mathrm{TP}_i}{\sum_{\mathrm{i}=1}^{l} (\mathrm{TP}_i + \mathrm{FN}_i)}$$

The first case consists of a published dataset that therefore restricts the sequence space to that of the published phylogeny, resulting in a fixed reference dataset with known classification. The other cases consist of novel datasets in which the sets of target sequences are formed by all homologous sequences identified from large reference proteomes dataset as described in materials and methods. In these cases no clear reference clustering exists and classification recall is expressed as sequence space classification recall interval which is the interval between the lowest and highest possible coverage of each group, restricted by monophyly or paraphyly.

The plant ACD protein superfamily: A complex case with paraphyletic groups and repeats

Alpha crystallin domain (ACD) proteins form a large superfamily that include various subfamilies of the well described small heat shock proteins (sHSPs) as well as a number of poorly or not described subfamilies [22]. Recently we identified 824 ACD proteins in 17 plant proteomes, using cluster-specific HMMER profiles manually made using a training set consisting of all ACD sequences identified in seven complete plant proteomes [22]. This suggested the existence of 24 major and five minor subfamilies alongside two orphan sequences. Approximately half of the subfamilies are sHSPs, which functional classification is largely based on the cellular component of function. The remaining clusters include a family of transcriptional regulators, a family of salt stress induced proteins and 11 subfamilies of Uncharacterized ACD Proteins (UAP) [22]. We took the datasets as previously used [22] but removed sequences of the five minor subfamilies and a single orphan, all distant sequences that fall inside the major sHSP-C1 cluster and prevent detection of the sHSP-C1 cluster since the algorithm imposes monophyly.



	R1		1	2	3	4	5	6	7	8		9	10	11	12	13		14	15	16	17	18	19	20	21	22	23	24	25		26
		R2	1	2	3	4	5	6	7	8	9	10	11		12	13	14	15		16	17	18	19	20	21		22	23		24	25
Tr	375	374	79	30	25	21	20	18	17	16	15	14	13	11	11	11	10	10	9	8	7	7	7	7	7	6	6	6	5	5	3
Ta	414	414	76	**21**	30	25	28	25	23	27	14	14	15	10	12	10	10	11	12	9	7	9	5	10	8	2	9	8	3	3	5
TP	387	403	76	**21**	30	25	28	25	18	27	14	13	14	0	12	**8**	8	11	8	9	7	9	3/5	10	8	1	9	8	2	3	5
FN	28	11	0	0	0	0	0	0	5*	0	0	1	1	11	0	**2**	2	0	4	0	0	0	2/0	0	0	1	0	0	1	0	0
FP	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
%	93	97	100	100	100	100	100	100	78	100	100	93	93	0	100	80	80	100	67	100	100	100	60/100	100	100	50	100	100	67	100	100

Fig 2. Optimized HMMERCTTER clustering and classification of plant ACD protein superfamily. (A) Training tree with clustering; clusters numbered according to HMMERCTTER and codes applied by Bondino et al., [22] (MI+ combines mitochondrial I (MI) and mitochondrial-like sHSPs (ML)). (B) Final and complete reference tree with classification made using clustering shown in A. Both A and B concern the second run as detailed in the text. Colors according to HMMERCTTER output, leaves in black could not be clustered or classified. Scale bars indicate 1 amino acid substitution per site. Note that UAPVII is represented by four clusters. UAPI and II were originally identified in the final 17 proteome dataset and correspond to a single clade in the training tree. * indicates a single orphan training sequence that becomes classified in a paraphyletic clade. ** points to a local difference in tree topology that does not affect clustering and classification. The table shows the numerical results of HMMERCTTER classification. R1: 1st run with specific columns in light-gray shade, R2: Optimized 2nd run with specific columns in dark-gray shade. R1-G11 is R2-G14; R1-G25 is R2-G24; and R2-G9 consists of R1-G15 and R1-G22. For details see context and S2 Fig. Tr = Train; Ta = Target; TP = True Positives; FP = False Positives; FN = False Negatives; % = classification recall (TP/(TP + FN)). Boldface numbers include paraphyletic sequences. * Concerns distant sequences as shown in S3 Fig. Note that both the total number of train and target sequences include one orphan sequence each and that the single false positive is a training sequence.

https://doi.org/10.1371/journal.pone.0193757.g002

We obtained a sequence clustering that is nearly identical to the described functional classification (Fig 2A). The major discrepancy consists of UAPVII that was not 100% P&R and further divided into groups 11, 14, 19 and 24. In order to determine how the several groups behave in HMMERCTTER classification we compared its classification with the final reference phylogeny of the complete sequence set, under the assumption that the phylogeny is correct.

The classification of a first run showed 93% overall classification recall (R1 in Table of Fig 1). However, group 11 had a classification recall of 0, meaning that not a single novel sequence was detected. We compared trees and analyzed hmmsearch output and encountered two dataset complications. First, the analysis is based on the assumption that both phylogenies are correct and as such comparable. This assumption appeared incorrect since one training sequence (VV00193000) clusters differently in both trees (S2A and S2B Fig) suggesting incorrect placement in the training tree, and, as a result, incorrect HMMERCTTER clustering and poor classification. This sequence was transferred from training dataset to target dataset. Furthermore, at least one target sequence was found to contain three partial ACDs, which resulted in an elevated total score in at least two groups, which generated another classification conflict. This sequence was removed from the target dataset. We repeated the analysis resulting in an optimized clustering and classification (Fig 2). Clusters R1_C15 and R1_C22 combined with sequence VV00193000 to form a larger cluster, R2_C9, with 100% P&R self detection in both clustering and classification. R2_C14 corresponds with cluster R1_C11 and shows 80% classification recall. Total coverage was 97%. In general false negatives are distant sequences as exemplified by the five false negatives indicated in S3 Fig. This concerns five sequences from *Sorghum bicolor* of which four appear to derive from the same locus.

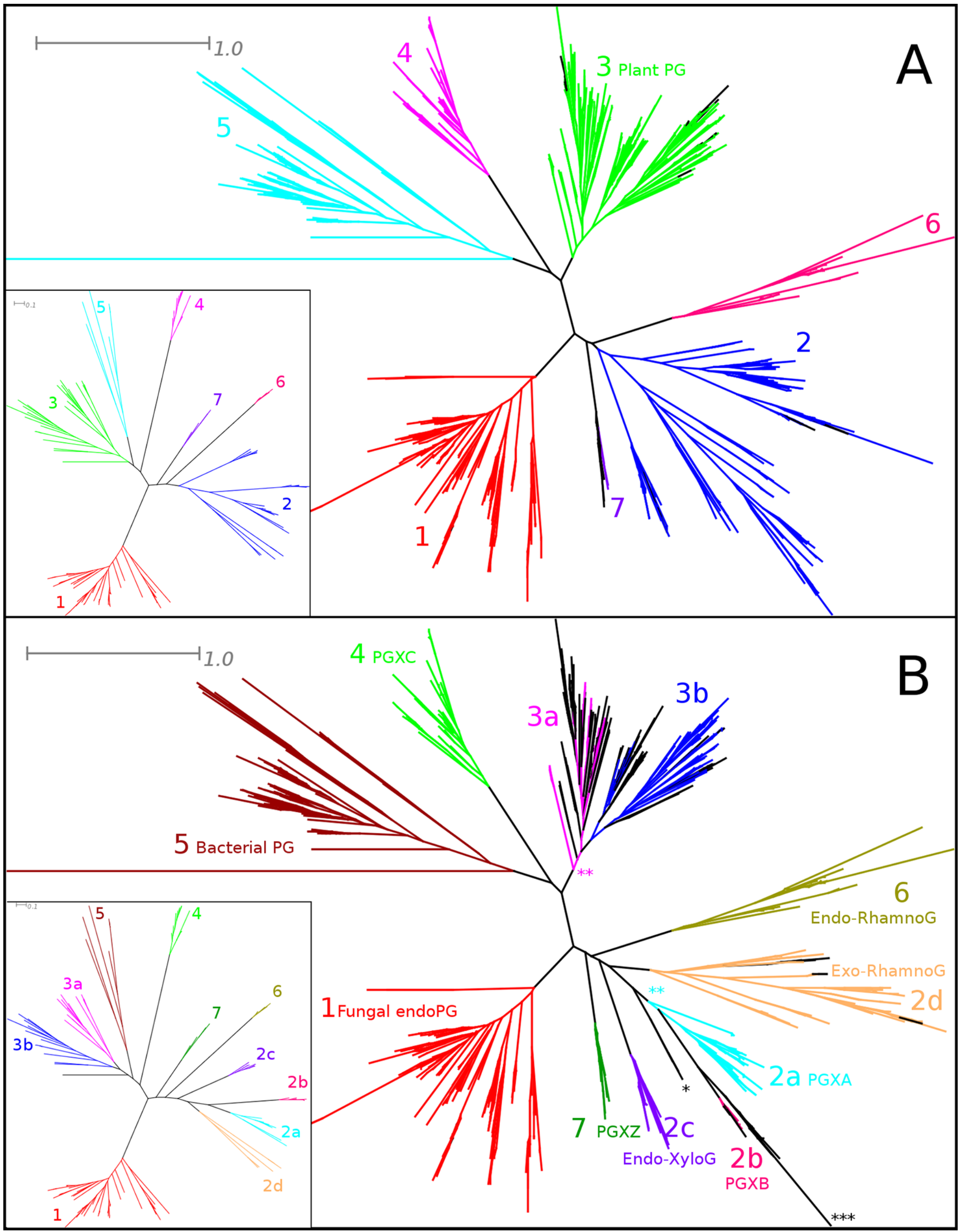
The PG superfamily: A case showing hierarchical clustering and compositional bias

Pectin is an important structural heteropolysaccharide component of plant cell walls formed of linear chains of α-(1–4)-linked D-galacturonic acid. Rhamnose and xylose can intervene in the main chain and sugar hydroxyl groups can be substituted by methyl groups and a variety of small sugar polymers, resulting in a complex mixture of polycarbohydrates (For review see [23]). Plants and many of their pathogens, therefore require a number of enzymes that can degrade these polycarbohydrates, among which those of the superfamily of galacturonases (G) and polygalacturonases (PGs, SCOP identifier 51137). Many isoforms have been described and are biochemically classified according to their mode of action (exoPGs, endoPGs) and substrate specificity (PGs, rhamnoGs and xyloGs) [24]. Then, the PG superfamily on its turn is part of the larger superfamily of pectin lyase like proteins (SCOP Identifier 51126). A number of 140 PG encoding sequences are documented in the UniProtKB/Swiss-Prot [2] database and were used to reconstruct a Maximum Likelihood phylogeny, together forming the training dataset. The target dataset consisted of 1255 PG homolog sequences identified from EBI's reference proteomes dataset amended with several complete proteomes from phytophagous organisms.

The training set could be clustered with 100% coverage in many ways (S4 Fig) and each clustering was used for classification of the target sequence dataset. Highest coverage (97%) was obtained using C7 (Fig 3A) (seven clusters named C7_C1 to C7-C7 and final groups C7-G1 to C7-G7), which is used as reference classification. However, the C7 clustering does not correspond perfectly with the functional clustering. C7-C2 contains two classes of exoPGs, the class of endo-xyloGs and the class of exo-rhamnoGs. C11, with C7-C2 and C7-C3 further divided into four and two subclusters respectively, does correspond with functional and hierarchical classification, albeit that exoPGs are represented by four polyphyletic clades. The C7 and C11 classifications are further discussed in detail (Fig 3).

As expected, partitions with few clusters (e.g. C2, C3, C4 see S4A and S4B and S4C Fig), show lower classification recall (See analysis output in Github repository). For instance, the combined plant (C3) and bacterial (C5) PG clusters do not detect any novel sequence using the classifiers determined by partitions C2, C3 and C4. Interestingly, partitions with more clusters such as C11 (Fig 3B) and C13 (See analysis output in Github repository), also show inferior performance.

The C7 classification performance was corrected for a small number of partial sequences from clades 1, 2 and 3 with scores slightly below the final thresholds. It also has an ambiguous reference classification. Eleven monophyletic and undetected sequences, properly detected in the C11 classification, are part of a larger monophyletic clade (S5A Fig). Classification recall interval is 22 to 100% but given the correct classification by C11, 22% should be considered as most meaningful. Strikingly, the C11-G7 score plot (S5B Fig) shows a distinguished group with a very sharp HMMER score drop following the cut-off (from 509.3 to 269.3). The reason why other partitions yield poor classification must therefore lie in conflicting sequence



C7	Total	1	2	3	4	5	6	7**
Tr	140	51	32	27	13	7	6	4
Ta	1255	316	316	279	136	140	50	4-18
TP	1214	313	307	264	136	140	50	4
FN	41	3	9	15	0	0	0	0-14
%	97	99	97	95	100	100	100	22-100
%	98	100*	99*	99*	100	100	100	22-100

C11	Total	1	2a	2b**	2c	2d	2O	3a	3b	4	5	6	7
Tr	140	51	10	8	8	6	1	14	12	13	7	6	4
Ta	1255	316	90	30-48	58	117	3	178	102	136	140	50	18
TP	1068	313	89	19	58	110	-	120	15	136	140	50	18
FN	170-188	3	1	11-29	0	7	3	58	87	0	0	0	0
%	85	99	99	40-63	100	94	-	67	15	100	100	100	100

Fig 3. HMMERCTTER analysis of the polygalacturonase superfamily. (A) C7 Clustering (inset) and classification with seven clusters. (B) C11 Clustering (inset) and classification with eleven clusters. Colors according to HMMERCTTER output, leaves in black could not be clustered or classified. Scale bars indicate 0.1 and 1 amino acid substitution per site as indicated. Functional classification is indicated in (A) for Plant PGs and (B) for the other classes where PGX stands for exo-polygalacturonase with A, B and C classes defined by Swissprot and Z corresponding to Zygomycete sequences. * Indicates a group of three orphan sequences that were not classified and could not be included in any reference classification. ** indicates instances of paraphyletic groups. *** indicates a subclade of unclassified sequences that complicate the reference classification of group 2b. Performance of the C11-2b cluster is therefore presented as an interval. The table shows the numerical results of HMMERCTTER classification. Tr = Train; Ta = Target; TP = True Positives; FN = False Negatives; % = Classification recall (TP/(TP + FN)). * Not included are 3, 7 and 11 partial sequences that appeared as false negatives in the C7 classification. ** The number of false negatives depends on the reference classification. Coverage is expressed as interval. 2O concerns the unclassifiable clade with three orphan sequences.

https://doi.org/10.1371/journal.pone.0193757.g003

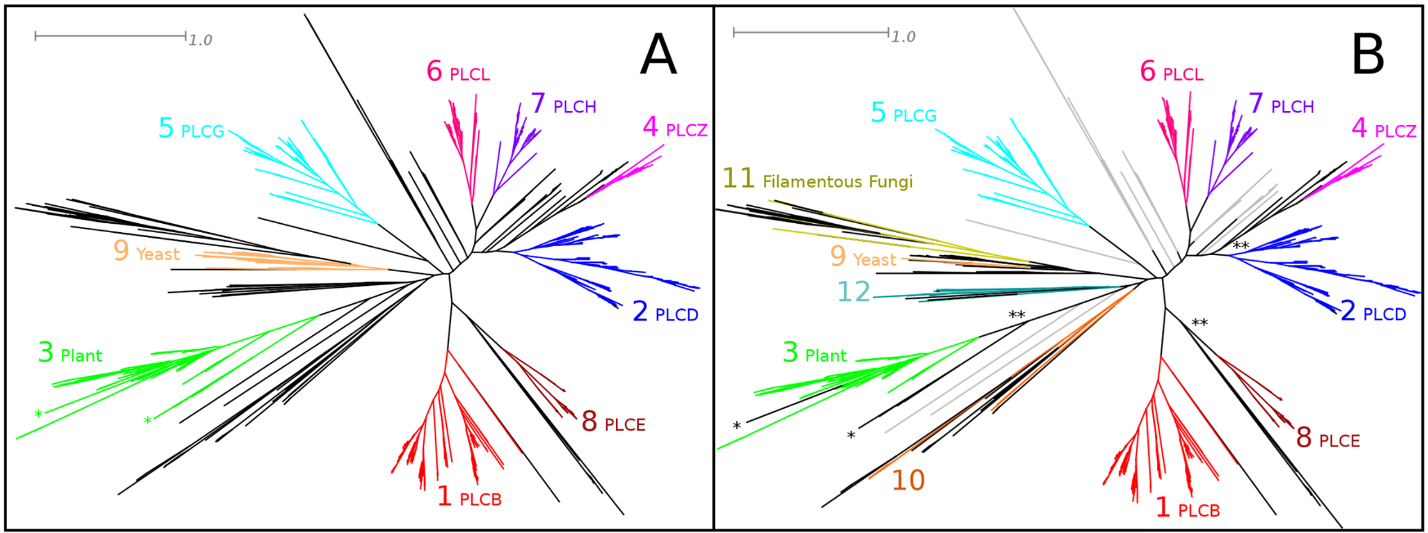
identifications. Indeed, the EFRo007836 sequence, part of the clade corresponding with C11_G7 (S5A Fig), is detected by both groups 2 and 7, and is therefore reported as conflicting sequence, arresting further analysis of either group.

The phospholipase C superfamily: A small, biased training set to classify a large target set

Phospholipase C (PLC) forms a class of enzymes that hydrolyze phospholipids [25]. They are involved in cell physiology and signal transduction and there are several reasons for functional diversification, as exemplified by the fact that PLCs can have a number of different regulatory domains. Six isotypes, B, D, E, G, H and Z, are discriminated in mammals that also contain PLC-like proteins (PLC-L), which lack the second catalytic His residue [25]. Fungi and plants also have PLCs: In tomato six isoforms have been reported [26] whereas in *Saccharomyces cerevisiae* only one homolog has been identified [27].

A total of 70 complete sequences was identified in UniProtKB/Swiss-Prot forming the training set. The target dataset consisted of 1047 sequences from EBI's Complete Reference Proteomes dataset. The training was guided by the functional classification. Nine clusters (B, D, E, G, H, Z, L, Plant and Yeast) were assigned based on the phylogenetic clustering and UniProtKB/Swiss-Prot annotation codes that reflect the diversity. The sequence of the PLC from *Dictyostelium* (lacking any additional specification in the UniProtKB/Swiss-Prot code) was not classified and regarded as orphan sequence. Fig 4 shows the results of the classification.

The classification of 939 sequences is nearly perfect. Strictly, only five single false negatives were identified when analyzing the classification on the reference tree (Fig 4). The performance of clusters 4 (PLCZ), 7 (Yeast) and 8 (PLCE) is ambiguous since larger clades can be considered, resulting in 49 false negatives and a classification recall of 95%. However, since the dataset consisted of 1047 PLC sequences, 60 additional sequences were not classified, which corresponds with 90% overall classification recall. This is explained by the fact that the training dataset was biased: A number of clades with no training representatives is found in the final tree (See Fig 4A). We added a total of nine target sequences to the training set in order to represent three of the major unrepresented clades and repeated the analysis (Fig 4B). Surprisingly, total classification recall was slightly lower at 94%. The major reason for this is the very poor performance of the yeast cluster (10%) and novel cluster 12 (0%), which contains sequences from filamentous fungi. Together they form a monophyletic clade that represents fungal PLCs, of which particularly the PLCs from filamentous fungi show high sequence divergence. Also the clade of group 10 is highly divergent and shows poor classification recall (29%). The plant PLCs showed a slightly lower amount of sequences correctly identified.



C9	Total	Orphans	1/PLCB	2/PLCD	3/Plant	4/PLCZ	5/PLCG	6/PLCL	7/PLCH	8/PLCE	9/Yeast
Tr	70	1	16	15	9	7	7	5	4	3	3
Ta	943-988	-	248	160	79	27-43	132	100	124	54-60	19-42
TP	939	-	247	158	79	27	132	100	124	54	18
FN	4-49	-	1	2	0	0-16	0	0	0	0-6	1-24
%	95-99.5	-	99.6	100	100	63-100	100	100	100	90-100	43-95

C12	Total	Orphans	1/PLCB	2/PLCD	3/Plant	4/PLCZ	5/PLCG	6/PLCL	7/PLCH	8/PLCE	9/Yeast	10	11/FF	12
Tr	79	1	16	15	9	7	7	5	4	3	2	4	3	3
Ta	992-1018	-	248	160	75-79	27-43	132	100	124	54-60	20	21	20	11
TP	931	-	247	158	74	27	132	100	124	54	2	6*	0	7
FN	44-87-117	30	1	2	1-5	0-16	0	0	0	0-6	18	15	20	4
%	89-91-94	-	99.6	100	94-100	63-100	100	100	100	90-100	10	29	0	64

Fig 4. HMMERCTTER analysis of the phospholipase C superfamily. (A) Initial C9 classification of nine described PLC subfamilies. (B) C12 Classification upon inclusion of additional sequences to the training set. Colors according to HMMERCTTER output of C9 clustering, leaves in black or gray could not be classified. Black leaves were included in determination of classification recall or classification recall interval, ** in B indicates classes where the reference classification is ambiguous. The 30 unclassifiable sequences are represented in gray. The * in A and B point to differences in the classification of Plant PLCs. The table shows the numerical results of the C9 and C12 classifications. Tr = Train; Ta = Target; TP = True Positives; FN = False Negatives; % = Classification recall (TP/(TP + FN)). * Includes the training orphan. Classification recall intervals are presented when the reference classification is ambiguous and are accompanied by intervals in the number of target sequences and false negatives. The 30 unclassifiable sequences indicated in gray in B were regarded as additional false negatives resulting in the lowest classification recall of the overall classification recall interval.

https://doi.org/10.1371/journal.pone.0193757.g004

Comparative analysis of classification

HMMERCTTER performance was compared with PANTHER[21], a major phylogenomics platform. Panther classification was performed with the complete datasets. i.e. both training and target sequences, since PANTHER uses its own HMMER profile database as training set. PANTHER reports to which subfamily or family HMMER profile a sequence scores best.

In the PG case, all except five sequences correspond with four major families (PTHR31736, PTHR31339, PTHR31375 and PTHR31884) with a large number of subfamilies. We report the analyses at the family as well as subfamily level. For the family analysis we collected all subfamily hits that correspond to the same family. Results are summarized in S6 Fig.

Interestingly, PTHR31736 contains all exoPGs as well as the RhamnoPGs and XyloPGs, even although these do not form a monophyletic cluster. PTHR31339 corresponds to bacterial PGs, PTHR31375 to plant PGs and PYHR31884 to fungal endoPGs. Only four false positives, and as such also false negatives, were identified. In addition five sequences were selected in rather different PANTHER families. Hence, 99% classification recall was obtained when subfamilies were grouped to their respective families, a result slightly better than what was obtained by HMMERCTTER in the C7 classification (98% see Fig 3). The latter shows a better correspondence with functional classification. The PANTHER classification into subfamilies shows a lower performance (compare S6 Fig with Fig 3). Both the bacterial and the plant families show a high quality classification with 94 and 89% classification recall, respectively, and 100% precision. However, sequences were assigned to 15 and 35 subfamilies for bacterial and plant families, respectively. This seems dis-proportionally high given the amount of known functional subfamilies. Classification of the two fungal families was however poor. The endoPG classification (S6 Fig) only shows a single subfamily, 31884:SF9 with 100% P&R self detection, all other subfamilies show many errors. ExoPG 31736:SF7 (S6 Fig) shows perfect correspondence with PGXC, 31736:SF9 corresponds to the endo-xylogalacturonases and 31736:SF5 corresponds with part of exo-rhamnogalaturonases. 31736:SF8 contains most but not all sequences from PGXA and PGXB, whereas subfamily 10 and the not further classified sequences contain a mixture of sequences. Hence, although a number of functional subfamilies are well classified, others show poor classification.

In the ACD case PANTHER identified six families (PTHR11527; PTHR15348; PTHR 33879; PTHR33981; PTHR34661; PTHR43670) of which three show 100% P&R self detection. Panther however, failed to correctly identify the sHSP as a family (See S7 Fig). The family containing all sHSP sequences also contained UAPXII and a number of other non-HSP sequences (See analysis output in Github repository). At the subfamily level, many sHSP classes were identified correctly but the C1 HSP consisted of four non polyphyletic subfamilies and a single false negative. Given the complexity of the classification pattern when plotted on the phylogeny, it is not possible to determine classification recall objectively. However, comparison of the classification patterns (Fig 2 and S7 Fig) clearly shows HMMERTCTTER yields a much better classification.

In the PLC case, many subfamilies of a single family (PTHR10336) were identified, which hampers analysis. We first analyzed all subfamilies that contain at least one training sequence plus the two subfamilies that contain more sequences than the smallest training sequence containing subfamily. Then we combined the PTHR subfamilies according to the initial functional classification of B, D, E, G, H, L, Z, Plant and Yeast PLC. Performance is shown in Table 1. The PLC-B, PLC-D, PLC-G and the PLC-Plant subfamilies appear represented by more than one PANTHER subfamily and clearly, combining these does improve classification recall. However, classification recall is still significantly below the classification recall obtained by HMMERCTTER (Fig 4). In addition a total number of 47 false positives were identified.

Discussion

We present and test the new protein superfamily sequence clustering and classification tool HMMERCTTER. HMMERCTTER uses a high fidelity phylogeny to identify clusters that show 100% P&R self detection. The resulting cluster- or subfamily-specific HMMER profiles with corresponding cut-off thresholds form the initial classifiers. Subsequent classification of target sequences is adaptive: In an iterative HMMER screen positively identified sequences are added to a subfamily and the HMMER profile with corresponding cut-off threshold is updated. This results in a high sensitivity while high specificity is safeguarded by imposing 100% P&R self detection to groups, combined with an iteration arrest when conflicting sequence identification occurs. Here we discuss method and pipeline based on issues identified in three case studies.

The high observed classification recall is an overestimate due to an inherent lack of reference. We emphasize that 100% P&R refers to group self detection, whether it concerns

Table 1. Panther classification of PLC case.

Functional Subfamily	PTHR Subfamily	Target	+	R+	F+	F-	%
B3	PTHR10336:SF10	176	103	103	0	73	59
B4	PTHR10336:SF106	326	108	108	0	218	33
B2	PTHR10336:SF12	63	44	42	**2**	21	67
PLCB	10-12-106	264	255	253	**2**	11	96
D4	PTHR10336:SF31	176	44	44	0	132	25
D4	PTHR10336:SF80	175	37	37	0	138	21
D1	PTHR10336:SF33	175	80	80	0	95	46
PLCD	31-33-80	175	161	161	0	14	92
At_PLCD9	PTHR10336:SF101	84	38	38	0	46	45
At_PLCD3	PTHR10336:SF105	204	18	12	**6**	192	6
At_PLCD2	PTHR10336:SF92	81	31	27	**4**	54	33
PLC Plant	92-101-105	175	161	161	0	14	92
Z	PTHR10336:SF29	49	49	49	0	0	100
G	PTHR10336:SF79	124	44	44	0	80	35
G	PTHR10336:SF91	139	72	72	0	67	52
PLCG	79–91	139	116	116	0	23	83
PLCL	PTHR10336:SF102	88	87	76	**11**	12	86
PLCH	PTHR10336	128	86	65	**21**	63	51
PLCE	PTHR10336:SF6	57	57	57	0	0	00
PLC Yeast	PTHR10336:SF36	49	37	24	**13**	13	65

Target corresponds with the size of the major clade containing (sub)family sequence, + corresponds with positives, R+ with real positives, F+ with false positives and F- with false negative. % Indicates classification recall (TP/(TP+FN).

https://doi.org/10.1371/journal.pone.0193757.t001

clustering or classification. This means that upon clustering, the partition can have unclustered orphan sequences. Furthermore, in theory HMMERCTTER yields highly specific classifications whereas certain homologs will not become classified. This particularly concerns remote homologs since including their sequences severely disrupts the group's MSA and therewith the HMMER scoring. In all three cases we found over 95% classification recall of the complete datasets with groups that show 100% P&R self detection according to the reference tree. It should be clear that the reference tree is not an ultimate benchmark dataset since *a priori* it is unknown which sequences should be considered as group member. HMMERCTTER classification performs iterative HMMER searches and includes sequences to groups while maintaining the groups with 100% P&R self detection. Sequence classification is arrested by conflicting sequence identification. In addition, it can be stopped when the user detects strong declines of the score drop that follow the lowest scoring group sequence. Clearly, sequence classification terminates in the twilight zone of detection, which is inherently subject to the lack of reference. Hence, although 95% might be an overestimate, the fact that conflicting sequence classification arrests the process shows HMMERCTTER is specific and that while it remains a high sensitivity, remote homologs will not be detected.

A high fidelity training set with no bias is fundamental for proper classification

The sequence incorrectly placed in the training tree of the ACD case (VV00193000, see Fig 2 and S2 Fig) had a severe impact on clustering and exemplifies the general rule that training data should be of high quality. Tree reliability is difficult to measure but in general trees with poor statistical support should be handled with care. The PLC case showed an additional training set issue: Although difficult or even impossible, bias should be avoided. HMMERCTTER is meant as a decision support system for the expert biologist, which presumably can provide a reliable training set. Still, although complete proteomes can be used as target, it is worthwhile to perform a preliminary sensitive data mining to obtain a set of target sequences restricted to possible homologs only, as we performed for the PG and PLC cases. Not only will HMMERCTTER run faster, it will also directly give an indication of performance, by which bias can be suspected, as was shown for the PLC case. Unfortunately, our attempt to correct for the bias in the PLC UniProtKB/Swiss-Prot dataset was not successful. This is at least in part due to biological complexity in the form of highly divergent subfamilies.

Orphan sequences in the training set should be avoided. They either represent incorrect sequences or result in bias since they are not included in the clustering. The ACD case had a single remaining orphan sequence. This sequence and a close homolog were classified into a group that forms a paraphyletic clade in the reference tree (Fig 2B). Paraphyletic groups were also identified in the PG classification (Fig 3B). Classification of paraphyletic sequences is possible since classification, rather than clustering, is based on HMMER profiling, basically an eloquent distance score. It should however be clear that an optimal clustering or classification corresponds with both tree topology and (known) functional classification, as was obtained for both the ACD and the PLC cases. Furthermore, the ACD case shows that sequences with repeats should be avoided. All similarity based search and classification tools inherently suffer from sequences with repeats.

Sensitivity of individual HMMER profiles and the clustering determine P&R of the overall classification

HMMERCTTER classifies sequences using controlled iterative HMMER searches. The sensitivity of the profiles not only determines the sensitivity but also the specificity of the classification. When classification is arrested upon conflicting sequence identification, the various HMMER profiles actually compete for the unclassified sequences. In general, a HMMER profile made from a variable subfamily-MSA will be more sensitive and less specific than a HMMER profile made from a conserved subfamily-MSA. This is demonstrated by the PG case in which the partitions with few clusters that, hence show high sequence variation, result in early classification conflicts of the plant and bacterial PG sequences.

On the other hand, further division of C7-C3 in C3a and C3b worsened classification, emphasizing that the final classification not only depends on the individual clusters but also on the exact partition. This is also demonstrated by the poor classification of C7-G7, as compared to C11-G7 (See Fig 3). Here the original C7-C7 and C11-C7 clusters are identical but the additional clusters of the two partitions differ, resulting in different sequence identification conflict scenarios. The main difference is that in C11 the C2 cluster is subdivided into four more specific subclusters that apparently no longer detect sequences that correspond to G7. Another part of the explanation for this classification error is the fact that PGs appear to have a moderately high compositional bias, which is known to negatively affect the accuracy of HMMER scores [28]. Similarly, convergent evolution that can be envisaged among the the four exoPG clades (2a, 2b, 6 and 7) might also negatively affect HMMER score accuracy [28] and therewith specificity. The fact that the C11 clustering is capable of correctly classifying the C7 sequences, suggests that the specificities and sensitivities of the profile combinations form an important factor in determining performance.

All together this demonstrates there is a balance between group size and variability and that it is difficult to predict how well a certain clustering will perform in classification. Two aspects that will define classification performance are compactness and separateness of a cluster. Both compact (e.g. sHSP-C1 of the ACD case Fig 2) and well separated clusters (e.g. C4 from the PG case Fig 3) will show good classification. The poor classification of ACD protein clusters 11, 14, 19 and 24 (Fig 2) as well as the fungal PLCs (Fig 4) can be explained by high sequence diversity, which equals low compactness of the cluster. Sequences at larges distances will not only obtain lower hmmsearch scores, but will also introduce high variation into the profile made by hmmbuild. A divergent profile corresponds with a low specificity, which is apparently still a major limit.

On the one hand the poor classification of distant sequences shows the limit of the HMMERCTTER method, on the other hand it points to dataset issues in the form of pseudogenes or sequences derived from incorrect gene models. To the best of our knowledge no function has been assigned to any of the members of the problematic and divergent UAPVII clade. The fungal PLCs are in large part orthologs and as such not pseudogenes. Thus, biological expertise remains required. Fortunately, distant sequences will, in general, only become accepted to a group during the interactive, user-controlled part of the classification phase. In the absence of an objective method for the reliable identification of problematic sequences, the interface of HMMERCTTER's interactive classification allows the user to use its expertise in order to make an educated decision. As such, HMMERCTTER is a decision support system.

The issue of dysfunctional sequences is problematic. Sensitive data mining, as for instance performed by the iterative JackHMMER [29], often results in heavily contaminated datasets, which results in severe problems while constructing an MSA. HMMERCTTER's 100% P&R self detection control, iteration arrest upon conflicting sequence identification and the fact that training sequences cannot be removed from the groups, prevents the inclusion of many problematic sequences and forms therefore an excellent method for sequence mining.

Prospects

We have developed HMMERCTTER that is capable of classification of protein superfamilies sequences with both high sensitivity and specificity. Performance is, depending on the protein family, similar or better than that of PANTHER, with the additional advantage of that the user is able to guide the clustering to its needs, whereas PANTHER seems to contain many subfamilies. The latter is likely related to the idea that subdivision of large superfamilies into small subfamilies should result in improved specificity. As such, the strict monophyletic approach of HMMERCTTER, combined with the possibility to guide clustering, using the knowledge of the expert biologist, seems a promising approach. It must be noted that as such HMMERCTTER is not a real phylogenomics tool since it depends on high quality phylogenies, which are not available at genomic scale.

HMMERCTTER provides an objective and computational approach rather than defining manually curated inclusion thresholds. The performance is high and limited mostly by aspects determined by the dataset such as training bias, errors in the training phylogeny, sequence repeats but also high sequence diversity. The 100% P&R self detection controlled iterative approach is arrested when conflicting sequence identifications are observed. Hence, performance in the twilight zone of sequence identification is determined by a balance between sensitivity and specificity. Current efforts toward future improvements include a profound mathematical modeling of the method dedicated at properties as correctness, convergence, classification recall, and measures of quality. It includes the determination of clustering quality, prediction of classification error rates, and the relationship between these two quantities. The development of this improved pipeline will be accompanied by a profound analysis of how thresholds should be defined for imbalanced classifications such as discussed by Zou and coworkers [30]

HMMERCTTTER is a phylogenomics tool since it uses phylogeny to classify sequences. However, since it is dedicated at sub-clustering single superfamilies, it is not comparable to databases such as FunFHMM, CDD or Panther. Applying maximum likelihood phylogeny rather than a heuristic sequence based clustering will, at least theoretically, improve sub-clustering but application to large, protein space spanning databases is not feasible. However, the 100% P&R self detection approach using HMMER appears to be applicable. The apparent error in the training phylogeny of the ACD shows that HMMER searches in a 100% P&R self detection setting typically correspond with phylogenetic clustering, which is in correspondence with our general experience with HMMER searches. Hence, the clustering phase could be replaced by a profile database provided that the profiles and corresponding sequence database correspond with the 100% P&R self detection rule. Current efforts are directed at constructing such a database that would be comparable to the aforementioned hierarchic HMMER profile databases.

Materials and methods

HMMERCTTER pipeline

The HMMERCTTER pipeline is written in MATLAB (The MathWorks Inc., Natick, MA, USA) and calls a number of PERL scripts that depend on Bioperl [31] and software packages. HMMER3 [6]: hmmbuild is used with default settings, hmmsearch with the option–noali. MSAs are constructed by MAFFTv7 [32]: with the settings–anysymbol–auto. Dendroscope 3 [33] is used for midpoint rooting and images representing clustering on the presented phylogeny on various user interfaces.

Datasets

The ACD training and target datasets were obtained from Bondino et al., [22] from which a single distant orphan sequence and the sequences from five distant subclusters were removed. PG training sequences were identified identified from UniProtKB/Swiss-Prot [2] by BLAST using endoPG sequence AAC64374.1 [34] as query. PLC training sequences were identified from Swissprot using human PLC-G sequence AAA60112.1 [35] as query. Target datasets for the PG and the PLC case were obtained using HMMER profiling. The PLC sequences were identified from EBI's Reference Proteomes, which consists of 122 eukaryotic and 25 prokaryotic complete proteomes (For details see http://www.ebi.ac.uk/reference_proteomes), whereas this was amended with a number of complete proteomes from phytophagous organisms for the PG case. Sequences identified by all the superfamilies training profiles were combined and filtered by CD hit [36] at 100% and scrutinized using Pfam [7]. In the PG case all sequences with a hit against the Glyco_hydro_28 domain were accepted, in the PLC case all sequences that hit both the PLC-X and PLC-Y domain and contain the first of two catalytic His residues were accepted. MSAs were constructed by MAFFTv7 [32] using the slow iterative global refinement (FFT-NS-i) mode for PGs and the multiple domain iteration (E-INS-i) mode for PLCs and subsequently corrected by Rascal [26]. PHYML3 [37] using the LG model was used for phylogenetic tree reconstruction following BMGE [38] trimming with BLOSUM62 matrix and an entropy cut-off of 0.9. Complete trees were constructed with all identified sequences. The training tree of the second PLC analysis is an excerpt of the PLC reference tree.

Supporting information

S1 Appendix. HMMERCTTER procedure.
(PDF)

S1 Fig. HMMERCTTER user interfaces. Shown are single examples obtained from the training (A) and the target phase (B). Blue asterisks are scores of sequences of the complete dataset (either training or combined), green circles are scores of sequences of the cluster or group in question. The red line shows the drop in scores among successive sequences, also indicated as score drop. The magenta line shows the current threshold and can be moved in order to determine the threshold in the interactive part of the classification. In the training the user must accept or reject the 100% P&R cluster. During classification the user can accept the cluster, return to the former or initial state or add seemingly negatives as indicated for posterior 100% P&R self detection testing.
(TIF)

S2 Fig. Incongruent clustering of ACD train tree severely affects classification. (A) Detail of poor classification of first run using clustering based on train-tree with sequence VV00193000 clustered differently than in the shown final tree. Sequences of GR1-11 including VV00193000 in red; GR1-15 in gray nested inside the clade containing GR1-22 light green lines. Target sequence ME22590va contains three ACDs also generating classification conflicts. (B) Detail of improved classification of second run obtained upon removal of VV00193000 and ME22590va from training and target set, respectively. GR2-14 shows 80% classification recall, formerly 0% with G1-11. GR2-9 shows 100% classification recall, which is an improvement of the 67 and 50 of constituents GR1-15 and GR1-22. Numbers also in Fig 1.
(TIF)

S3 Fig. Sequences at large distances impede classification recall. Detail of cluster 7/M+ demonstrating that particularly sequences at large distances (ML) are often not detected. The indicated sequences without shading are training sequences whereas the sequences in red shade are false negatives. The arrow points to a monocotyledon sub clade.
(TIF)

S4 Fig. C2, C3 and C4 clustering and classification of the polygalacturonase superfamily. (A1) C2 Clustering; (A2) C2 Classification; (B1) C3 Clustering; (B2) C3 Classification; (C1) C4 Clustering; (C2) C4 Classification. Colors according to HMMERCTTER output, leaves in black could not be clustered or classified. Scale bars indicate 1 amino acid substitution per site.
(TIF)

S5 Fig. Poor classification of group 7 by C7 clustering. (A) Classification of group 7 by C7 clustering. In purple sequences from C7-G7 group. Other leaves represent sequences identified by the C11 clustering. EFro007836 is identified by both C7-C2 and C7-C7, preventing classification. (B) Score plot of C11-G7. Blue asterisks are scores of sequences of the complete dataset, green circles are scores of sequences group C11-G7. The red line shows the drop in scores among successive sequences. The magenta line shows the current threshold.
(TIF)

S6 Fig. Classification of PG sequences using PSE-analysis and PANTHER. PANTHER classification was analyzed at combined family, as well as at individual subfamily level. "Panther 4 Families" shows a very good classification, false positives are indicated with colored labels, false negatives are indicated in black. The latter are indicated with their PANTHER identifier in the 31736 panel. The four families were analyzed at the subfamily level. Subfamily numbers are indicated in the same color as the corresponding edges,—indicates sequences correspond to a family but were not classified into a specific subfamily.
(TIF)

S7 Fig. Classification of ACD and PLC sequences using PANTHER. The top panels show the ACD family level and the sHSP subfamily level classifications. Each family is respresented by a different color. For the sHSP subfamily classification colors were chosen to represent separate clusters.
The lower panels show the classification of all PLC subfamilies containing a training sequence as well as two additional large subfamilies (PLC Top, random colors) and the classification upon combining subfamilies according to the functional classification according to SwissProt (For colors see Fig 4A). Data analysis is shown in Table 1.
(TIF)

Author Contributions

Conceptualization: Hernán Gabriel Bondino, Marcel Brun, Arjen ten Have.

Data curation: Arjen ten Have.

Formal analysis: Inti Anabela Pagnuco, María Victoria Revuelta, Arjen ten Have.

Funding acquisition: Marcel Brun, Arjen ten Have.

Investigation: María Victoria Revuelta, Arjen ten Have.

Methodology: Inti Anabela Pagnuco, Hernán Gabriel Bondino, Marcel Brun, Arjen ten Have.

Project administration: Arjen ten Have.

Resources: Arjen ten Have.

Software: Inti Anabela Pagnuco, Marcel Brun.

Supervision: Marcel Brun, Arjen ten Have.

Validation: María Victoria Revuelta, Hernán Gabriel Bondino.

Writing – original draft: Arjen ten Have.

Writing – review & editing: María Victoria Revuelta, Marcel Brun.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
2. Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Res. Oxford University Press; 2017; 45: D158–D169. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622
3. Eirín-López JM, Rebordinos L, Rooney AP, Rozas J. The Birth-and-Death Evolution of Multigene Families Revisited. Genome dynamics. 2012. pp. 170–196. https://doi.org/10.1159/000337119 PMID: 22759819
4. Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2013; 41: D344–D347. https://doi.org/10.1093/nar/gks1067 PMID: 23161676
5. Zhang Z. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 1998; 26: 3986–3990. https://doi.org/10.1093/nar/26.17.3986 PMID: 9705509
6. EDDY SR. A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE. Genome Informatics 2009. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.; 2009. pp. 205–211. https://doi.org/10.1142/9781848165632_0019 PMID: 20180275
7. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database [Internet]. Nucleic Acids Research. Oxford University Press; 2014. pp. D222–30. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

8. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001; 313: 903–919. https://doi.org/10.1006/jmbi.2001.5080 PMID: 11697912

9. Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2007; 36: D419–D425. https://doi.org/10.1093/nar/gkm9933 PMID: 18000004

10. Eisen JA. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Res. 1998; 8: 163–167. https://doi.org/10.1101/gr.8.3.163 PMID: 9521918

11. Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics. 2002; 3: 14. https://doi.org/10.1186/1471-2105-3-14 PMID: 12028595

12. Brown DP, Krishnamurthy N, Sjölander K. Automated Protein Subfamily Identification and Classification. PLoS Comput Biol. 2007; 3: e160. https://doi.org/10.1371/journal.pcbi.0030160 PMID: 17708678

13. Lee DA, Rentzsch R, Orengo C. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. Oxford University Press; 2010; 38: 720–737. https://doi.org/10.1093/nar/gkp1049 PMID: 19923231

14. Abhiman S, Sonnhammer ELL. FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res. 2004; 33: D197–D200. https://doi.org/10.1093/nar/gki067 PMID: 15608176

15. Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics. Oxford University Press; 2015; 31: 3460–3467. https://doi.org/10.1093/bioinformatics/btv398 PMID: 26139634

16. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017; 45: D200–D203. https://doi.org/10.1093/nar/gkw1129 PMID: 27899674

17. Neuwald AF, Lanczycki CJ, Marchler-Bauer A. Automated hierarchical classification of protein domain subfamilies based on functionally-divergent residue signatures. BMC Bioinformatics. 2012; 13: 144. https://doi.org/10.1186/1471-2105-13-144 PMID: 22726767

18. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE. Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res. 2011; 21: 1969–1980. https://doi.org/10.1101/gr.104687.109 PMID: 21784873

19. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE, Chervitz S. Protein Molecular Function Prediction by Bayesian Phylogenomics. PLoS Comput Biol. Morgan Kaufman Publishers; 2005; 1: e45. https://doi.org/10.1371/journal.pcbi.0010045 PMID: 16217548

20. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic Acids Res. Oxford University Press; 2009; 37: D396–D403. https://doi.org/10.1093/nar/gkn803 PMID: 18957448

21. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013; 41: D377–D386. https://doi.org/10.1093/nar/gks1118 PMID: 23193289

22. Bondino HG, Valle EM, ten Have A. Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. Planta. 2012; 235: 1299–1313. https://doi.org/10.1007/s00425-011-1575-9 PMID: 22210597

23. Willats WGT, Mccartney L, Mackie W, Knox JP. Pectin: Cell biology and prospects for functional analysis [Internet]. Plant Molecular Biology. Dordrecht: Springer Netherlands; 2001. pp. 9–27. https://doi.org/10.1023/A:1010662911148 PMID: 11554482

24. ten Have A, Tenberge KB, Benen JAE, Tudzynski P, Visser J, van Kan JAL. The contribution of cell wall degrading enzymes to pathogenesis of fungal plant pathogens BT—The Mycota XI, Agricultural Applications. Agricultural Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002. pp. 341–358. https://doi.org/10.1007/978-3-662-03059-2_17

25. Kadamur G, Ross EM. Mammalian phospholipase C. Annu Rev Physiol. 2013; 75: 127–154. https://doi.org/10.1146/annurev-physiol-030212-183750 PMID: 23140367

26. Vossen JH, Abd-El-Haliem A, Fradin EF, Van Den Berg GCM, Ekengren SK, Meijer HJG, et al. Identification of tomato phosphatidylinositol-specific phospholipase-C (PI-PLC) family members and the role of PLC4 and PLC6 in HR and disease resistance. Plant J. Blackwell Publishing Ltd; 2010; 62: 224–239. https://doi.org/10.1111/j.1365-313X.2010.04136.x PMID: 20088897

27. Andoh T, Yoko-O T, Matsui Y, Toh-E A. Molecular cloning of the plc1+ gene of Schizosaccharomyces pombe, which encodes a putative phosphoinositide-specific phospholipase C. Yeast. 1995; 11: 179–185. https://doi.org/10.1002/yea.320110209 PMID: 7732727

28. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. Oxford University Press; 2013; 41: e121. https://doi.org/10.1093/nar/gkt263 PMID: 23598997

29. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. BioMed Central; 2010; 11: 431. https://doi.org/10.1186/1471-2105-11-431 PMID: 20718988

30. Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the Best Classification Threshold in Imbalanced Classification. Big Data Res. 2016; 5: 2–8. https://doi.org/10.1016/j.bdr.2015.12.001

31. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. Cold Spring Harbor Laboratory Press; 2002; 12: 1611–1618. https://doi.org/10.1101/gr.361602 PMID: 12368254

32. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. Oxford University Press; 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

33. Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012; 61: 1061–1067. https://doi.org/10.1093/sysbio/sys062 PMID: 22780991

34. ten Have A, Mulder W, Visser J, Van Kan JAL. The Endopolygalacturonase Gene Bcpg1 Is Required for Full Virulence of Botrytis cinerea. Mol Plant-Microbe Interact. 1998; 11: 1009–1016. https://doi.org/10.1094/MPMI.1998.11.10.1009 PMID: 9768518

35. Ohta S, Matsui A, Nazawa Y, Kagawa Y. Complete cDNA encoding a putative phospholipase C from transformed human lymphocytes. FEBS Lett. 1988; 242: 31–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/2849563 PMID: 2849563

36. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics. Oxford University Press; 2010; 26: 680–682. https://doi.org/10.1093/bioinformatics/btq003 PMID: 20053844

37. Thompson JD, Thierry JC, Poch O. RASCAL: Rapid scanning and correction of multiple sequence alignments. Bioinformatics. Oxford University Press; 2003; 19: 1155–1161. https://doi.org/10.1093/bioinformatics/btg133 PMID: 12801878

38. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010; 10: 210. https://doi.org/10.1186/1471-2148-10-210 PMID: 20626897