Title page

Title: HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised Classification of Superfamily Protein Sequences with a reliable Cut-off Threshold

Short Title : HMMERCTTER Protein Superfamily Classification

Authors: Inti Anabela, Pagnuco[a]; María Victoria, Revuelta[b,c]; Hernán Gabriel, Bondino[b]; Marcel, Brun[a] and Arjen, ten Have[b,*¶]

Author affiliations:

a: Laboratorio de Procesamiento Digital de Imágenes, Instituto de Investigaciones Científicas y Tecnológicas en Electrónica (ICyTE), Facultad de Ingeniería, Universidad Nacional de Mar del Plata, J. B. Justo 4302, 7600 Mar del Plata, Argentina.

b: Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, CC 1245, 7600 Mar del Plata, Argentina.

c: Current address: Department of Medicine, Hematology and Oncology Division, Weill Cornell Medicine, New York, New York 10065, USA.

* To whom correspondence should be addressed: atenhave@mdp.edu.ar, +54 223 4818446

Author Contributions:

IAP: *Software;* assistance *Development;* assistance *Formal analysis; Writing-Review and Editing*

MVR: assistance *Development;* assistance *Formal analysis; Writing-Review and Editing*

HGB: *Development; Validation; Writing-Review and Editing*

MB: *Development;* supervision *Software; Writing-Review and Editing*

AtH: *Development; Formal analysis; Writing-Original Draft and Preparation*

Abstract

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.

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 100% precision and recall (P&R), thereby generating a specific threshold as inclusion cut-off. Profiles and threshold are then used as classifiers to screen a target dataset. Iterative inclusion of novel sequences to clusters and the corresponding HMMER profiles results in high sensitivity while specificity is maintained by imposing 100% P&R. In three presented case studies of protein superfamilies, classification of large datasets with 100% P&R was achieved with over 95% coverage. Limits and caveats are presented and explained.

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.

Author summary

The enormous amount of genome sequences made available in the last decade provide new challenges for scientists. An important step in genome sequence processing is function assignation of the encoded protein sequences, typically based on the similarity principle: The more similar sequences are, the more likely they encode the same function. However, evolution generated many protein superfamilies that consist of various subfamilies with different functional characteristics, such as substrate specificity, optimal activity conditions or the catalyzed reaction. The classification of superfamily sequences to their respective subfamilies can be performed based on similarity but since the different subfamilies also remain similar, it requires a reliable similarity score cut-off.

We present a tool that clusters training sequences and describes them in profiles that identify cluster members with higher similarity scores than non-cluster members, i.e. with 100% precision and recall. This defines a score cut-off threshold. Profiles and thresholds are then used to classify other sequences. Classified sequences are included in the profiles in order to improve sensitivity while maintaining specificity by imposing 100% precision and recall. Results on three case studies show that the tool can correctly classify complex superfamilies with over 95% coverage.

Introduction

Genome sequencing and the resulting sheer amount of protein coding sequences (PCS) provides many opportunities and challenges to biologists. BLAST [1] is a method for function assignation of PCSs that is also used for sequence mining directed at the bio-computational analysis of protein (super)families. PHI-BLAST [2] includes biochemical information in the form of a Prosite [3] formatted pattern, preventing the detection of many false positives at the cost of generating false negatives, whereas the laborious PSI-BLAST [4] iteratively increases the information of the original query by means of a position specific substitution matrix and is capable of detecting distant homologs.

HMMER [5] uses more information by comparing query sequences with multiple sequence alignments (MSAs), using probability profiles and a hidden Markov model. The combination of this sensitive tool with thoroughly annotated profile databases such as the domain database Pfam [6] or the protein database Superfamily [7], makes HMMER profiling an important tool in complete proteome function assignation. HMMER has been tested extensively, particularly regarding false positives and it is known to suffer from compositional bias, homology overextension but also homology overlap [8]. Hence, HMMER arguably suffers from poor precision and recall (P&R) when complex protein superfamilies are analyzed. Many applications for detection of remote homologs exist and methods to improve specificity have been applied successfully. The HHpred server [9] runs a PSI-BLAST with a single sequence, builds an MSA and HMMER profile that is subsequently compared to existing HMMER databases. HMMerThread [10] combines a relaxed sequence search with fold recognition, the latter in order to eliminate false positives. ConFunc [11] combines the sensitive PSI-BLAST with Gene Ontology annotations obtaining higher levels of precision and recall. However, to the best of our knowledge, the increased information usage of HMMER has principally been deployed to

increase sensitivity whereas in principle higher information content is also applicable to increasing 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. Sites with high information content, i.e. highly conserved sites, will give either high rewards or high penalties whereas sites with low information content, i.e. highly variable sites, 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. Based on this principle, 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 P&R. Here we report the first version, named HMMER Cut-off Threshold Tool or HMMERCTTER, that is directed at the classification of protein superfamily sequences, based on a training set that consists of a high quality phylogeny and corresponding protein sequences.

The training phase automatically identifies monophyletic sequence clusters that have 100% P&R in a HMMER screening, i.e. *hmmsearch* using profiles made from cluster-specific MSAs identify all the cluster's sequences with a score higher than that of any other sequence provided by the training set. This also defines a cluster-specific score threshold that prvides a reliable 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.

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 phospohlipase 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 S Appendix 1. 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 is obtained when the lowest scoring cluster sequence has a higher score than the highest scoring non-cluster sequence. 100% 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 S Fig 1A, 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 in HMMER profiling as well as, possibly, a number of unclustered orphan sequences.

The HMMER profiles and corresponding cut-off scores, which equal the lowest scores of the clusters' sequences, form the initial classifiers that are used for screening the target dataset. Sequences with scores equal or above the cluster threshold are automatically accepted and added to the cluster. We refer to these as prior positives since they were not yet included in the cluster 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, clusters remain 100% P&R provided classification overlap is prevented. When a target sequence becomes classified by more than one group, all 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.

The automated classification phase 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 interactive phase, sequences with a score below the threshold are considered. Candidates are included in the cluster and tested with a novel HMMER profile that includes the candidate. An interactive interface (S Fig 1B) allows the user to guide this process while 100% P&R remains imposed and classification conflicts remain prohibited as described for the automated phase. The process is 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, of which for one we used previously published datasets. The major objective was to 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 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 a factor 10.

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 [12]. 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 [12]. 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) [12]. We took the datasets as previously used [12] but removed the five minor subfamilies and a single orphan, distant sequences that fall inside the major sHSP-C1 cluster and prevent detection of the sHSP-C1 cluster since the algorithm imposes monophyly.

We obtained a sequence clustering that is nearly identical to the described functional classification (S Fig 2). 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% coverage (see Table 1, R1 with classes indicated numerically). However, group 11 had a coverage 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 (S

Fig 2A and B) 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 (see Supplemental Fig 2C and D, Table 1, R2). Groups R1_15 and R1_22 as well as sequence VV00193000 formed a larger cluster, R2_9, with 100% P&R in both clustering and classification. R2_14, corresponding with group R1_11 shows 80% coverage, total coverage was 97%. In general false negatives are distant sequences as exemplified by the five false negatives indicated in S Fig 2E. This concerns five sequences from *Sorghum bicolor* of which four appear to derive from the same locus.

Table 1: Numerical analysis of HMMERCTTER classification ACD case.

	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	6	6	6	5	5	3
Ta	414	413	76	19	30	25	28	25	23	27	14	14	15	11	12	10	10	11	12	9	7	9	5	10	8	2	9	8	3	3	5
TP	385	399	76	19	30	25	28	25	18	27	14	13	14	0	12	8	8	11	8	9	7	9	3	10	8	1	9	8	2	3	5
FN	28	13	0	0	0	0	0	0	5	0	0	1	1	11	0	2	2	0	4	0	0	0	2	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	50	100	100	67	100	100

R1: 1st run with specific columns in lightgray shade, R2: 2nd run with specific columns in darkgray shade. R1_11 is R2_14; R1_25 is R2-24; and R2_9 consists of R1_15 and R1_22 Tr = Train; Ta = Target; TP = True Positives; FP = False Positives; FN = False Negatives; Shown are the number of identified sequences. % = Coverage. For details see context and S Fig 2.

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 [13]). Plants and many of their pathogens, therefore require a number of enzymes that can degrade these polycarbohydrates, among which those of the superfamily of (poly)galacturonases (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, rhamnoPGs and xyloPGs) [14]. 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 [15] database and used to reconstruct a Maximum Likelihood phylogeny, together forming the training dataset. The target dataset consisted of 1260 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 (S Fig 3A, B, C, D and F) and each clustering was used for classification of the target sequence dataset. Highest coverage (96%) was obtained using C7 (with seven clusters named C7_c1 to C7-c7 and final groups C7-C1 to C7-C7 containing training and classified sequences), which is used as reference clustering. However, the C7 clustering does not correspond perfectly with the functional clustering. C7-c2 contains two classes of exoPGs, the class of endo-xyloPGs and the class of exo-rhamnoPGs. 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 clusters. The C7 and C11 classifications are further discussed (S Fig 3E and G) and their numerical analysis is shown in Table 2.

Table 2: Numerical analysis of HMMERCTTER classification PG case.

C7	Total R	Total C	1R	1C	2R	2C			3R	3C	4	5	6	7R	7C
Tr	140	140	51	51	32	32			27	27	13	7	6	4	4
Ta	1260	1247	316	313	315	302			279	268	136	160	50	4	17
TP	1213	1207	313	313	306	300			264	264	136	140	50	4	4
FN	47	20	3	0	9	2			15	4	0	20	0	0	14
%	96	97	99	100	97	99			95	99	100	88	100	100	22
C11	Total		1		2a	2b	2c	2d	3a	3b	4	5	6	7	
Tr	139		51		10	8	8	6	14	12	13	7	6	4	
Ta	1227		316		90	30	58	117	192	105	136	140	25	17	
TP	1042		313		89	19	58	110	120	16	136	140	23	17	
FN	146		3		4	11	0	7	58	61	0	0	0	0/2	
%	85		99		99	63	100	94	63	15	100	100	92	100	

Shown are the number of identified sequences by the C7 (raw (R) and corrected (C)) and C11 clustering analyses. Note that, except for the corrected C7 clustering, in the absence of a prior reference clustering, coverage is calculated using the smallest monophyletic clade containing all cluster members. In the C11 analysis clusters C7-c2 and C7-c3 were divided into 2a; 2b; 2c; 2d and 3a; 3b subclusters respectively. For details see context and S Fig 3.

As expected, clusterings with few clusters (e.g. C2, C3, C4 see S Fig 3A, B and C), show lower coverage (See Table 2). For instance, the combined plant (c3) and bacterial (c5) PG cluster does not detect any novel sequence in clusterings C2, C3 and C4 (not shown). Interestingly, clusterings with more clusters such as C11 (S Fig 3F and G) and C13 (not shown), also show inferior performance.

The C7 classification was corrected on two accounts. A small number of partial sequences from clades 1, 2 and 3 with scores slightly below the final thresholds were removed from the target dataset. Then, C7-C7R, only seems to show 100% P&R because by default we used the smallest monophyletic clustering for numerical analyses. Eleven undetected sequences, properly detected by the C11 classification, are part of a larger monophyletic clade (S Fig 3E). Hence, we corrected the coverage downwards for the C7 classification (group 7C in table 2). Strikingly the C11-C7 score plot (S Fig 3F) shows a distinguished cluster with a very sharp HMMER score drop following the cut-off (from 509.3 to 269.3). The reason why other clusterings yield poor classification must therefore

lie in conflicting sequence identifications. Indeed, the EFRo007836 sequence, part of clade C7_C7 (S Fig 3H), is detected detected by both clusters 2 and 7, and is therefore reported as conflicting sequence arresting futher analysis of either cluster.

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 [16]. 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 [16]. Fungi and plants also have PLCs: In tomato six isoforms have been reported [17] whereas in *Saccharomyces cerevisiae* only one homolog has been identified [18].

A total of 66 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. Two sequences were annotated as orphan sequences, the single PLC from *Dictyostelium* (lacking any additional specification in the UniProtKB/Swiss-Prot code) and the PLC_Z from chicken. Table 3 shows the statistics of the classification of the large dataset.

Table 3: Numerical analysis of HMMERCTTER classification PLC case

C9	Total	1/PLCB	2/PLCD	3/Plant	4/PLCZ	5/PLCG	6/PLCL	7/PLCH	8/PLCE	9/Yeast			
Tr	69	16	15	9	7	7	5	4	3	3			
Ta	941	248	158	79	27	132	100	124	54	19			
TP	939	247	158	79	27	132	100	124	54	18			
FN	2	1	0	0	0	0	0	0	0	1			
%	99.79	99.60	100	100	100	100	100	100	100	100			
C12	Total	1/PLCB	2/PLCD	3/Plant	4/PLCZ	5/PLCG	6/PLCL	7/PLCH	8/PLCE	9/Yeast	10	11	12
Tr		16	15	9	7	7	5	4	3	2	4	3	3
Ta	942	248	158	74	27	132	100	124	54	3	22	11	20
TP	925	247	158	74	27	132	100	124	54	2	7	7	0
FN	17	1	0	0	0	0	0	0	0	1	15	4	20
%	98.20	99.60	100	100	100	100	100	100	100	66.67	31.82	63.64	0.00

C9 indicates the initial clustering with the 9 described functional PLC classes. C12 corresponds with the clustering in which sequences were added to the training data in order to present the three major lacking clades. Shown are the number of identified sequences. For details see context and S Fig 4.

The classification of 940 sequences is nearly perfect, only two single false negatives were identified when analyzing the classification on the reference tree (S Fig 4). However, since the dataset consisted of 1047 PLC sequences, over a 100 sequences were not classified. Although part can be explained by poor classification such as for a number of PLCZ and PLCE sequences (See S Fig 4A), the major explanation is to be found in the fact that the training dataset was biased: A number of clades with no training representatives is found in the final tree. 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. As a result the plant PLCs showed a slightly lower amount of sequences correctly identified, although mathematically coverage remains 100% (See Table 3 and S Fig 4). Unfortunately only one of the novel clusters, C11, showed appreciable coverage (See Table 3). C10 only identified an additional 10 out of 22 sequences and cluster 12, containing fungal PLCs, did not identify any novel sequence and interfered with the clustering of the yeast PLCs.

Discussion

We present and test the new protein superfamily sequence classification tool HMMERCTTER. HMMERCTTER consists of two phases: a training phase depending on a hierarchic phylogenetic clustering and a target phase in which sequences and their information are added iteratively to their classifiers, providing high sensibility while specificity is safeguarded by imposing 100% P&R to the clusters and iteration arrest when conflicting sequence identification occurs. Here we discuss method and pipeline based on issues identified in three case studies.

The high observed coverage is an overestimate due to the lack of reference

In all three cases we found over 95% coverage of the complete datasets with clusters that show 100% P&R 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 cluster member. HMMERCTTER classification performs iterative HMMER searches and includes sequences to clusters while maintaining the clusters at 100% P&R. Sequence classification is arrested by conflicting sequence identification or can be stopped by the user when detecting strong declines of the score drop, which follows the lowest scoring cluster 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, at least suggests coverage is high and that HMMERCTTER is specific while is remains a high sensitivity.

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 S Fig 2) 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 very successful. To enrich the training set We selected target sequences randomly since random selection is key to obtaining unbiased datasets. However, EBI's reference proteome dataset does not cover the tree of life equally. In addition we selected very similar sequences from clades that contain rather divergent sequences. Selecting intermediately divergent sequences will improve both the training tree and the HMMER profiling based clustering and classification. All together this emphasizes both the importance and the problem of a high fidelity database such as UniProtKB/Swiss-Prot [15]. The quality of the sequences is excellent but it has a clear bias towards model organisms. It also shows that high sequence divergence inside a cluster does not favor HMMERCTTER classification.

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 in the corrected training set. This sequence and a close homolog were classified into a cluster that, according to the reference tree, is paraphyletic (S Fig1C and D). An additional paraphyletic group was identified upon

classification (not indicated). 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 clusterings with few, hence, variable clusters 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 clustering. This is also demonstrated by the poor classification of C7-c7, as compared to C11-c7 (See Table 2 and S Fig 3). Here the original c7 clusters are identical but the additional clusters 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 C11-C7. 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 [8]. 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 [8] 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 a major 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 S Fig 2) and well separated clusters (e.g. c4 from the PG case S Fig 3) will show good classification. C7-c5 (Plant PGs) from the PG case, is neither compact nor well separated and shows only 88% coverage. Interestingly, of the 20 false negatives, seven sequences were of bacterial origin. The poor classification of ACD protein clusters 11, 14, 19 and 24 (See S Fig 2) as well as PG clusters C11-c2b, C11-c2c and C11-c6 (See S Fig 3) might also be explained by high sequence diversity, which equals low compactness of the cluster. Sequences at larges distances will not only obtain lower *hmmsearch* scores in but will also introduce high variation into the profile made by *hmmbuild.* Divergent profile correspond with a low specificity.

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. Fortunately, distant sequences will, in general, only become accepted to a cluster 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. Obviously, the existence of problematic sequences is related to the fact that in general it is not possible to have a valid reference, as discussed above.

The issue of dysfunctional sequences is problematic. Sensitive data mining, as for instance performed by the iterative JackHMMER [19], often results in heavily contaminated datasets, which results in severe problem while constructing an MSA. HMMERCTTER's 100% P&R control, iteration arrest upon conflicting sequence identification and the fact that training sequences cannot be removed from the clusters, 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. This is achieved by 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 and compositional bias in general. The 100% P&R 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, coverage, and measures of quality. It includes the determination of clustering quality, prediction of classification error rates, and the relationship between these two quantities.

HMMERCTTTER consists of two phases of which the clustering phase can be replaced by a HMMER profile database, provided that the profiles and corresponding sequence database are 100% P&R. Current efforts are directed at constructing such a database. Based on the idea of HMMERCTTER we envisage that this database will consist of relatively many profiles, as compared to Pfam and SUPERFAMILY, in order to describe functional protein sequence space.

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 [20] and software packages. HMMER3 [5]: *hmmbuild* is used with default settings, *hmmsearch* with the option –noali. MSAs are constructed by MAFFTv7 [21]: with the settings –anysymbol –auto. Dendroscope 3 [22] 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., [12] 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 [15] by BLAST using endoPG sequence AAC64374.1 [23] as query. PLC training sequences were identified from Swissprot using human PLC-G sequence AAA60112.1 [24] 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 phytophagues organisms for the PG case. Sequences identified

by all the superfamilies training profiles were combined and filtered by CD hit [25] at 100% and scrutinized using Pfam [6]. 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 [21] 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 [27] using the LG model was used for phylogenetic tree reconstruction following BMGE [28] trimming with BLOSUM62 matrix and an entropy cut-off of 0.9. Complete trees were constructed with all sequences identified by the various clusterings analyzed.

Citations

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2
2. Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin E V, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 1998;26: 3986–3990. doi:gkb628 [pii]
3. 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. doi:10.1093/nar/gks1067
4. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs [Internet]. Nucleic Acids Research. 1997. pp. 3389–3402. doi:10.1093/nar/25.17.3389
5. Eddy SR. A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE. Genome Informatics. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.; 2009;23: 205–211. doi:10.1142/9781848165632_0019
6. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. Oxford University Press; 2014;42: D222-30. doi:10.1093/nar/gkt1223

7. 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. doi:10.1006/jmbi.2001.5080

8. 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. doi:10.1093/nar/gkt263

9. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33: W244-8. doi:10.1093/nar/gki408

10. Bradshaw CR, Surendranath V, Henschel R, Mueller MS, Habermann BH. HMMerThread: detecting remote, functional conserved domains in entire genomes by combining relaxed sequence-database searches with fold recognition. PLoS One. Public Library of Science; 2011;6: e17568. doi:10.1371/journal.pone.0017568

11. Wass MN, Sternberg MJE. ConFunc--functional annotation in the twilight zone. Bioinformatics. 2008;24: 798–806. doi:10.1093/bioinformatics/btn037

12. 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. doi:10.1007/s00425-011-1575-9

13. Willats WGT, McCartney L, Mackie W, Knox JP. Pectin: cell biology and prospects for functional analysis. Plant Cell Walls. Dordrecht: Springer Netherlands; 2001. pp. 9–27. doi:10.1007/978-94-010-0668-2_2

14. 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. Agricultural Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002. pp. 341–358. doi:10.1007/978-3-662-03059-2_17

15. UniProt: the universal protein knowledgebase. Nucleic Acids Res. Oxford University Press; 2017;45: D158–D169. doi:10.1093/nar/gkw1099

16. Kadamur G, Ross EM. Mammalian Phospholipase C. Annu Rev Physiol. 2013;75: 127–154. doi:10.1146/annurev-physiol-030212-183750

17. 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. 2010;62: 224–239. doi:10.1111/j.1365-313X.2010.04136.x

18. 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. doi:10.1002/yea.320110209

19. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and

iterative HMM search procedure. BMC Bioinformatics. BioMed Central; 2010;11: 431. doi:10.1186/1471-2105-11-431

20. 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: 1611–8. doi:10.1101/gr.361602

21. 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. doi:10.1093/molbev/mst010

22. Huson DH, Scornavacca C. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst Biol. 2012;61: 1061–1067. doi:10.1093/sysbio/sys062

23. Have A ten, 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. doi:10.1094/MPMI.1998.11.10.1009

24. 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

25. 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. doi:10.1093/bioinformatics/btq003

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

27. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52: 696–704. Available: http://www.ncbi.nlm.nih.gov/pubmed/14530136

28. 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. doi:10.1186/1471-2148-10-210

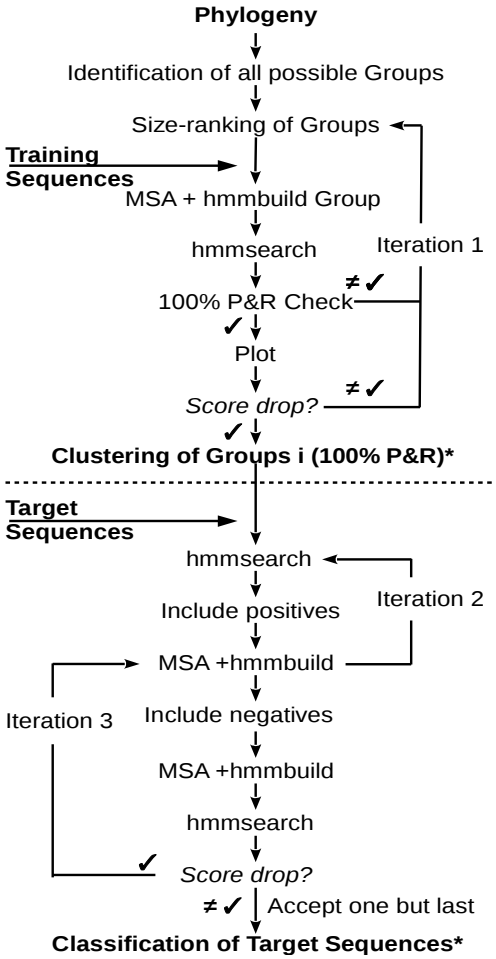

Fig 1: Flowchart of HMMERCTTER pipeline

Training and target phase are separated by the dotted line. Monophyletic clusters of the training-set are tested for 100% P&R, 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 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.*