Seqrutinator: Non-Functional Homologue Sequence Scrutiny for the Generation of large Datatsets for Protein Superfamily Analysis.

Agustín Amalfitano[1&], Nicolás Stocchi[2&], Hugo Marcelo Atencio[3], Fernando Villarreal[2], Arjen ten Have[2].
1 Laboratorio de Procesamiento de Imágenes, ICyTE-CONICET-UNMdP, Mar del Plata, Argentina; 2 Computational Biology and Comparative Genomics, IIB-CONICET-UNMdP, Mar del Plata, Argentina; 3 Banco de Germoplasma de Papa Andina, EEA-Balcarce INTA, Argentina. & These authors made equal contributions.
Corresponding author: tenhave.arjen@gmail.com.

Abstract

Background: In recent years protein protein bioinformatics has resulted in much improved algorithms for multiple sequence alignment (MSA) and phylogeny. Few attention has been paid to sequence selection whereas particularly recently published complete proteomes often have many sequences that are partial or derive from pseudogenes. Not only do these sequences add noise to the MSA, phylogeny and other downstream computational analyses, they actually also instigate many errors in the processing of the MSAs and therewith all other downstream analyses, including the phylogeny.
Objective: Here we aim to provide and test an objective, automated and flexible pipeline for the scrutiny of sequence sets from large, complex protein superfamilies, resulting in the classification of sequences into either functional or non-functional. The pipeline should identify few Swissprot sequences and result in an improved MSA.
Results: Seqrutinator is a pipeline that consists of five modules written in Python that identify and remove sequences that are likely Non-Functional Homologues. Here we tested the pipeline in its default order using default settings using three complex, related plant superfamilies (BAHD, CYP and UGT) that act in specialized metabolism, using the complete proteome of 16 plant species as input and using Swissprot as a control. Only 3.4% of Swissprot sequences were identified as NFH. Most NFH sequences are partial but, interestingly, their removal results in much shorter MSAs with many more reliable columns. Few but significant sequences that instigate gaps are found. The three superfamilies furthermore show similar scrutiny patterns.
Conclusion and perspective: Seqrutinator forms a consistent pipeline for sequence scrutiny that does result in sequence sets that generate high fidelity MSAs. Additional tests will be performed in order to further demonstrate Seqrutinators performance.

Introduction

Protein superfamilies, here defined as protein families with subfamilies that have different functional characteristics, are the subject of many computational studies (e.g. (1–8)) and form the target of many computational platforms (e.g. (9–12)). Structure-function analysis aims not only to identify which residues and/or subsequences are involved in functional diversification, it also tries to explain the functional differences and can identify hitherto nondescript subfamilies (e.g. (2, 3)). A large set of methods (e.g. (13–15)) is available and novel methods are published regularly (for review see (16)).

The basis for many protein bioinformatics tools is formed by phylogenies and their underlying Multiple Sequence Alignments (MSAs) and these come with a large set of algorithms and programs. Reliable methods for superfamily phylogeny use maximum likelihood (PHYML (17), RAXML (18); FASTTREE (19)) or Bayesian inference (MrBayes (20)). Also the construction of MSAs has improved significantly in recent years (21–25) but, likely since more complex protein families are now being analysed, is also still one of the major areas of research in bioinformatics (e.g. (26)).

Only very recently attention has been paid to sequence scrutiny, the selection of sequences that is to be included in the analysis (27, 28). This is a laborious task with many facets and which difficulty increases with the complexity of the protein family. One of the major facets is that of the identification of non-functional homologues (NFH). False signals that can derive from sequences that do not correspond to a functional homologue are often considered as mere noise that in all likelihood does not have a significant effect on the outcome of analyses. We argue that incorrect sequences also hinder the correct processing of the MSA, thereby generating indirect noise and erroneous signal.

There are two major sources of sequences of Non-Functional Homologues (NFHs): Pseudogenes and incorrect sequences. Pseudogenes show similarity levels from low (i.e. non detectable) to high (i.e. indistinguishable from functional homologues). Since they are no longer under functional constraint they can accumulate not only simple mutations but also obtain inserts and or deletions of subsequences, especially when the original gene contained introns. Incorrect sequences can, on their turn, result from sequencing errors or incorrect gene models. Particularly recently published complete proteomes often contain many incorrect gene models. The inclusion of many complete proteomes into a dataset results in an accumulation of various errors into that dataset. It can result in MSAs that are prohibitively long where inserts provide information that can derail proper MSA. Sequence scrutiny is therefore often required although, to the best of our knowledge, no objective, fully automated protocols exist. This is likely caused by the fact that defining inclusion thresholds is troublesome and will almost by definition result in both false positives and false negatives. In addition no real benchmarks on functional homologues exist and any such benchmark set will be biased towards too restrictive.

We developed a method for objective sequence scrutiny directed at the removal of sequences from NFH. This we did performing the sequence mining of three of the most complex single domain superfamilies in plants, to wit cytochrome P450 (CYP), UDP-Glycosyl Transferase (UGT), and BAHD acyltransferases (BAHD). Interestingly, many enzymes of these superfamilies function in specialized metabolism by which these families show a high evolution rate. This, in part, explains the high diversity and complexity of these superfamilies. The case studies

were used to improve the procedure with the final objective of constructing an automated pipeline. Here we present the pipeline, separate scripts and methods. We show numerical data generated by the three case studies directed at the validation of the method.

Design of the pipeline and its modules.

Objective

We define the objective of this work as: To provide and test an objective, automated and flexible pipeline for the scrutiny of sequence sets from large, complex protein superfamilies, resulting in the classification of sequences into either functional or non-functional. Input sequence sets represent all identified probable homologues of a protein superfamily as obtained by any sensitive data-mining methodology. Output sets consist of the functional homologuesequence set and the NFH sequence set. The functional homologue set should be aligned without much uncertainty. The NFH set should be subjected to a recovery analysis in order to prevent inadvertent false positive NFH incidences.

Definition of NFH

To scrutinize protein sequence datasets for non-functional homologues (NFHs), we first needed to define NFH in terms of sequence characters. We describe two major classes of NFH sequences that can be further subdivided. First, NFH-sequences can be the result of incorrect gene modelling or sequencing errors. Second, sequences can correspond with pseudogenes, where we define a pseudogene as a gene that no longer encodes its supposed function. As such, the scrutiny we develop is (super)family dependent. Note that he definition of pseudogene does not judge whether the gene might have another, possibly unknown or related function.

Incorrect gene modelling and sequencing errors can result in a number of issues. First, a number of sequences will lack either N- or C-terminal subsequences as a result of a missed start and/or stop codon This also applies to sequences that are incomplete as a result of incomplete sequencing/assembly. Second, missed start and stop codons can also lead to additional subsequences at either the N- or C-terminal. Third, not all introns are identified from eukaryotic sequences. Coding subsequences that are incorrectly identified as intron form a fourth problem. Both these intron issues can lead to intron-sized gaps or a switch in reading frame and the untimely stop of the coding sequence. A fifth problem, the incorrect calling of the exact intron splice donor and acceptor sites that lead to marginal errors are not tackled by our methods.

Pseudogenes are no longer under constraint and will as such accumulate mutations. This can result in two scenarios. It will result in an increased sequence and evolutionary distance, as can often be seen in phylogenetic trees. It may also result in the loss or gain of start- and stop-codons as well as splice donor and acceptor sites. As such, many pseudogenes will have issues similar to those found in NFH-sequences that result from incorrect gene modelling or sequencing errors.

Based on the above problem description we intend to identify NFH-sequences by the following characters:

1 Length: a number of NFH-sequences will be relatively short.

2 The presence of large contiguous gap-rich columns in MSAs, instigated by one or more sequences with intron derived subsequences incorrectly called as exon.

3 The presence of large gap-rich subsequences in MSAs, at otherwise amino acid rich columns, instigated by the absence of exon derived subsequences incorrectly called as intron.
4 Low similarity, either global or local.
The similarity issue can also correspond to non-homologous hits obtained in the initial sequence mining. This is a third, additional problem that depends on the sensitivity of the initial data mining. The modules and algorithms will be designed to find NFH-sequences based on these four characters.

Table 1: Seqrutinator modules and their actions.

Problem		SSR	NHHR	GIR	CGSR	PR
Sequence error	**ATG/Stop**	X		X	X	X
	Splice site	X		X	X	X
Pseudo-gene	**ATG/Stop**	X		X	X	X
	Splice site	X		X	X	X
	Similarity		X			X
Non-Homologous Hit		X	X			X

The Modules and the Default Pipeline

The fully automated pipeline is to be implemented in a larger procedure that starts with a user guided sequence collection and ends with user guided recovery analyses, as shown in figure 1. Sequence collection from multiple sequence sets is fully automated and requires a HMMER profile as well as the to be screened sequence sets as input. All sequence from a complete proteome with hits against the HMMER profile are collected in a single fasta and aligned by MAFFT. The reference sequence is selected by the user and added to each MSA using MAFFT-add. In all cases we used a sequence that corresponds with a resolved structure since these often correspond to the mature protein.

The MSAs are truncated manually, based on the reference sequence and the expert user's knowledge. This step is not automated since in a number of cases, truncation points will not correspond with the most N- or C-terminal residue of the reference sequence. Even when the reference sequence corresponds with a protein for which a structure has been resolved, differences in domain architectures can differ and we prefer user control at this important initial step. The truncated MSAs will as such already lack most N- and C-terminal extensions.

Although we distinguish four characters, it will be clear that they can be related and that they will not contribute equally to the final problem of generating a poor MSA. We developed five modules and a fully automated pipeline called seqrutinator. Seqrutinator is flexible as such that the user can not only select the modules and their order, but also change settings that will affect the stringency of the automated scrutiny based filter. Below we describe the default procedure (see Figure 1) and its reasoning. Table 1 shows which defective sequences will be detected by which modules. Details are in the complete description of the algorithms in supplemental file 2.

The first step is the Short Sequence Remover (SSR) module directed at sort sequences (character 1). As such, a reference sequence should be included. Sequences of proteins for which a structure has been resolved are preferred since these typically concern the final active protein, possibly following post-translational modifications such as propeptide cleavage. By default, sequences that have a length of 65% or less of the reference sequence are removed. .

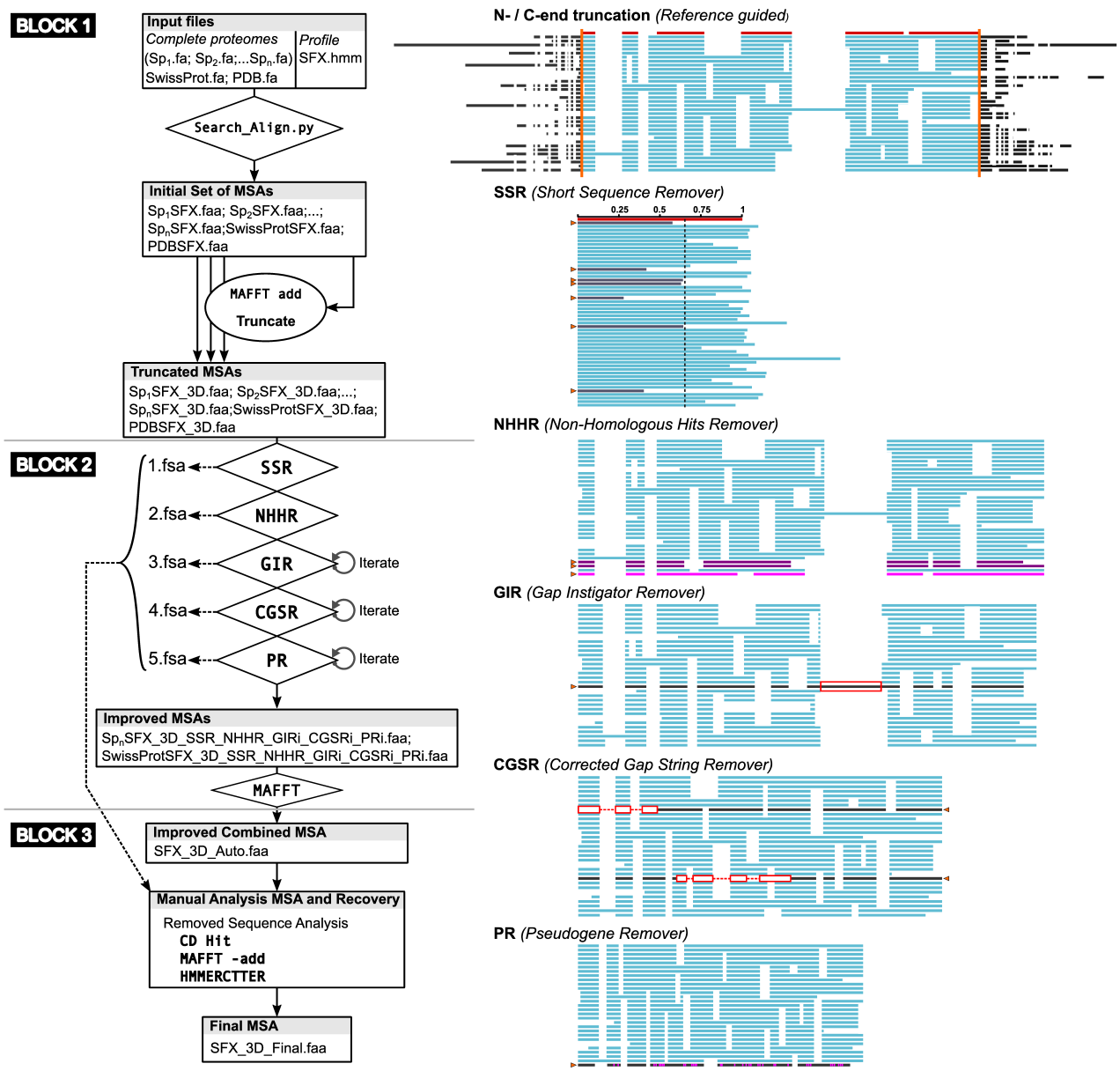


Figure 1: Schematic of the Default Seqrutinator Pipeline. *The complete workflow for protein superfamily sequence mining (left) consists of three blocks. The first block concerns the preparation of the input for the second block that contains the seqrutinator pipeline, which performs the fully automated sequence scrutiny. Each of its modules, here shown in its default order, and its iterations end with re-alignment of the remaining sequences. The third block concerns analyses for the recovery of inadvertently removed functional homologue sequences. 1.fsa, 2.fsa, 3.fsa, 4.fsa and 5.fsa are archives in which the removed sequences of the subsequent pipeline modules are temporarily stored for recovery analyses. The schematic MSAs on the right show how the truncation of block 1, and how the modules of the automated pipeline in block 2, function.*

The second step is directed at the removal of non-homologous hits (character 4). The Non-Homologous Hit Remover (NHHR) module identifies low sequence similarity outliers, hence it is an outlier remover. It constructs a HMMER profile and screens all sequences. Sequences are removed according to the three-sigma rule of thumb applied to HMMER scores, meaning that sequences that score below the mean minus three standard deviations are removed since they are not likely part of the superfamily. Hence, NHHR is an outlier remover. In a normal data set 99,7% of the sequences are significant and only few sequences should be removed.

The third step in the default pipeline removes sequences that instigate large gaps in MSAs (character 2). By default the Gap Instigator Remover (GIR) removes sequences that induce contiguous regions of over 30 gap columns, where gap column is defined as a column that has 90% or more gaps. This default region size is set at which is based on the minimal intron size observed among most eukaryotes. Sequences with additional N- and/or C-terminal sub-sequences will also be removed. Sequences are removed on a one-by-one basis (worst first) and GIR is iterated, where iterations start with realignment.

The fourth step uses the Continuous Gap Sequence Remover (CGSR) module. This removes sequences that show one or more instances of large continuous gaps in the MSA (character 3). Again default setting is at 30 columns whereas a gap only counts as a gap if less than 50% of the column has a gap, this in order to allow subfamily specific sub-sequences. CGSR is also iterated but removes sequences in batch. Note that not all sequences that lack a subsequence due to incorrect gene modelling will be detected. Residues that enclose the absent subsequence may align somewhere in the corresponding area, thereby splitting the region in two or more gap regions. In future versions of seqrutinator this problem will be tackled.

The last step of the default pipeline is taken by the Pseudogene Remover (PR) module (character 4). It is identical to the NHHR module except that it is iterated. The names of the NHHR and PR modules are as such based on the intent of the modules. Although there is no clear-cut threshold that can discriminate between a non-homologous hit and a pseudogene, the latter is expected to show more similarity and be less disturbing. By default PR uses the same setting as NHHR but has in practice two differences. First, NHHR is not iterated since the idea is that it should detect non-homologous hits which can be considered as real outliers that should all be detected in a single iteration. PR is iterated since pseudogene identification is more delicate and might require iteration. Second, the placement of the modules is such that NHHR works with an MSA that probably has many flaws whereas PR works with a much improved MSA. An improved MSA corresponds with a less flat HMMER score score distribution. As a result PR is more sensitive than NHHR.

Although all steps used in the pipeline are by default directed at the removal of only the most unlikely NFH-sequences, we envisage certain correct sequences will be inadvertently removed. As such we devise analyses that can be performed in order to make corrections against overzealous sequence removal. As for the scrutiny, we first listed the kind of erroneous sequence removal that could occur. Since the objective of the scrutiny is to obtain a dataset to be used for reliable classification, the sequences must be alignable without much uncertainties.

Design of the Performance Analysis

In order to develop automated sequence scrutiny we generated the algorithms while testing their performance using superfamily datasets. Although the basic problems are straightforward, the implementation of any quantitative scrutiny is hampered by the fact that it will be difficult to use reliable cut-off thresholds and settings. This is particularly true since no benchmark sets exist for sequence scrutiny. UniProtKB/Swiss-Prot (Swissprot) is, to the best of our knowledge, the best dataset for benchmarking the pipeline. It contains a large number of sequences that come with biochemical and / or transcript evidence. However, it does also contain sequences that do not have any actual evidence for functionality. Hence, we can expect that some sequences will actually be NFH-sequences.

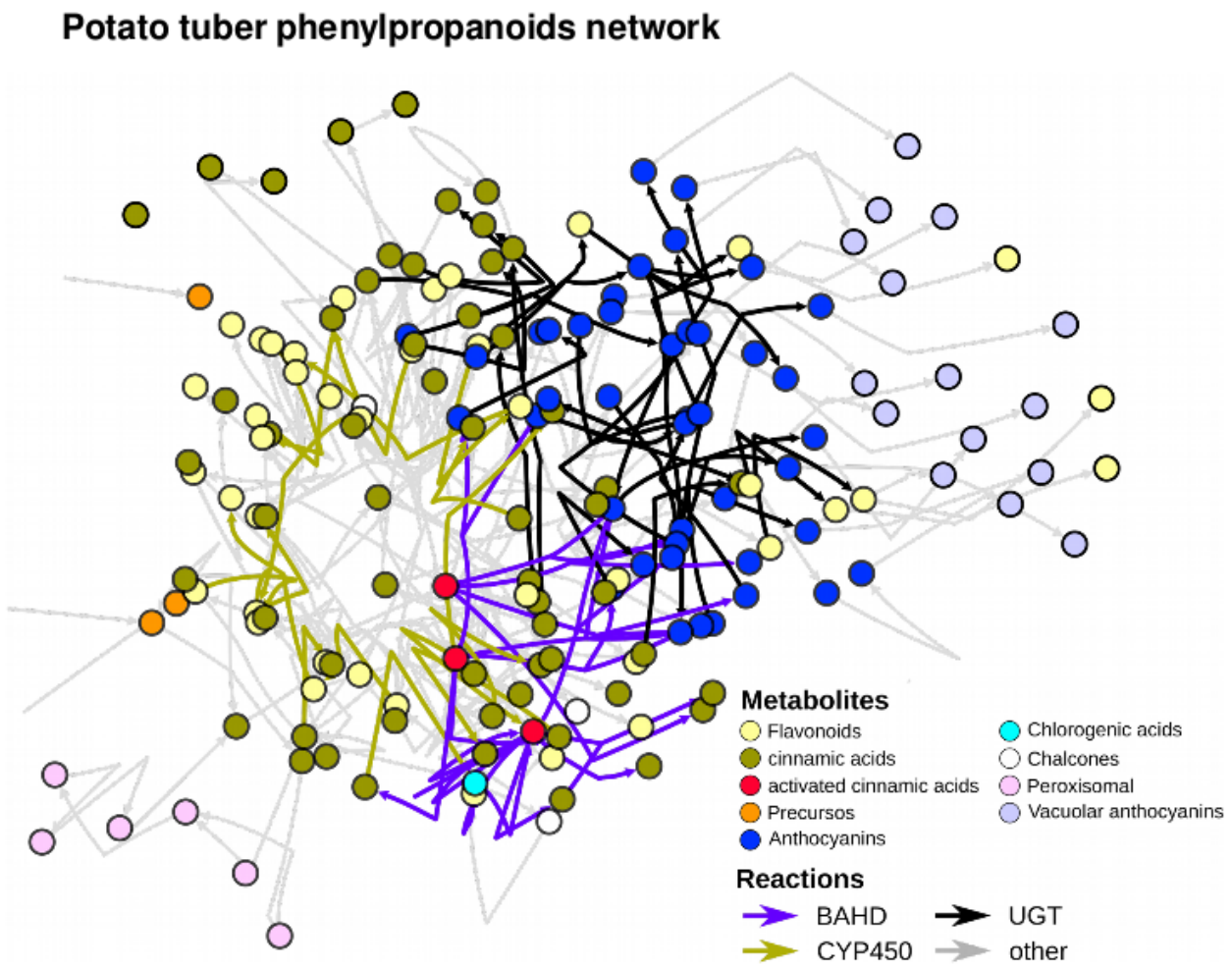


Figure 2: Cartoon Depicting the Phenylpropanoid/Flavonoid Biosynthesis Network in Potato. *Three major superfamilies cover approximately 25% of the enzymatic reactions. The superfamily subnetworks show a partial overlap. The BAHD and UGT superfamilies are shifted towards end-products that are conjugated and transported to the vacuole, thereby changing the effective sink.*

We tested the pipeline by simultaneously performing sequence mining for three functionally related and complex superfamilies in sixteen complete plant proteomes. We choose to select superfamilies with many homologues since large amounts (>100) of sequences are required for a statistical approach. We selected superfamilies involved in specialized metabolism since this provides a number of challenges. First, we argue that enzymes from specialized metabolism are under less functional constraint than enzymes from primary metabolism (29). As such we can expect high sequence diversity. Then, specialized metabolism generates a plethora of chemically related compounds via parallel, diverging and converging pathways (29). We envisage that promiscuous enzyme activity is often related to specialized metabolism and superfamily enzymes. Functional constraint will often act on (part of) the superfamily rather than on specific paralogues. The complexity of the superfamilies we will use is furthermore so high that no reliable function annotation is at hand for most of the homologues. Note that enzyme promiscuity can result in incorrect sequence annotation. A last issue we foresee is that certain subfamilies are taxon specific which will easily result in the inadvertent removal of functional homologue sequences.

Cytochrome P450 (CYP), UDP-Glysosyl Transferase (UGT) and BAHD acyltransferases constitute 4.17%, 7.81% and 11.98% respectively of the known flavonoid metabolism in potato, which contains 384 proteins that are classified in 59 Pfam domains. (See figure 2). The BAHD superfamily derives its name from its major enzymes: benzylalcohol O-acetyl transferase; anthocyanin O-hydroxycinnamoyl transferase; N-hydroxycinnamoyl anthranilate benzoyl transferase; and deacetylvindoline 4-O-acetyltransferase. As such, these three superfamilies form part of our major biological research interest which is to model flavonoid metabolism in potato. As such we need to functionally assign sequence of these three complex superfamilies. This can performed by for instance Panther (30), which assigns at subfamily level, rather than Pfam (9), which assigns at the superfamily level. Recently we developed HMMERCTTER (31) which outperformed Panther classifying three protein superfamilies. HMMERCTTER is a software for the clustering and classification of protein superfamily sequences. For this we need reliable sequence sets, i.e. sets that contain only few NFH-sequences but are nevertheless close to complete. In bioinformatics terms we need a sequence scrutiny with high precision and recall.

The automated sequence scrutiny of these three superfamilies in sixteen higher plants allows for a meta analysis that will shed light on the performance of the automated procedure. This is based on the hypothesis that complete proteomes, whether of low or high quality, will show similar sequence scrutiny performances when different superfamilies are analysed. Each of the applied algorithms should identify no or only few NFH-sequences in a high quality sequence set such as the complete proteome from the model organism *Arabidopsis thaliana* (TAIR10). On the other hand, complete proteomes that have either been published only recently or cannot count on a large research community, are more likely to contain many NFH-sequences. Although rather large differences consist in the number of paralogs in different plants, we expect a certain level of convergence. For *A. thaliana* we included TAIR v6, besides the latest and supposedly superb set of TAIR v10. The actual superfamily analyses will be published elsewhere, here we report the numerical data from the sequence scrutiny to show performance of the methods. We use Swissprot as a control since it forms the closest thing to a real benchmark dataset. We use a set containing all plant sequences as well as a set containing all plant sequences with either evidence at the protein or the transcript level.

Finally we considered how to detect inadvertently removed functional homologue sequences. To do so we must understand the method and the biases of the initial datasets. A complete proteome from a single organism will have a different bias than the Swissprot sequence set or, for instance, all sequences from a single Pfam seed alugnment. Hence, we do not propose a standard procedure but rather will present the analyses we performed. The results section contains a collection of methods that we used to analyse removed sequences as is also indicated in figure 1.

A first minor concern is that of sequences that have been removed by the PR module because of a poorly weighted HMMER profile. Although HMMER profiles are weighted, they cannot account for large differences in clade distance. In a superfamily with various equidistant subfamilies and a single, more distant subfamily, sequences of the distant subfamily will show low scores in a HMMER screening and might have been removed by the PR. This problem is excacerbated by our approach of scrutinizing complete proteomes, which come with divergent

MSAs and HMMER profiles with low precision and recall. There inadvertently removed sequences can be identified by cluster analyses of all the removed sequences (i.e. from the various initial sequence sets).

Another major concern is that of taxon or clade specific sequences. Complex superfamilies show a high rate of evolution by which taxon specific sequences should be expected. Any type of mutation can result in a novel functional subfamily and as such functional sequences can have been removed by GIR, CGSR or PR. Cluster analysis is likely to fail when the sequence is taxon specific, hence additional sequence mining may be required in order to enlarge the sequence space that is to be analysed.

Results and Discussion

Comparison of the pipeline on 16 complete proteomes shows it is consistent

We subjected the 19 sequence sets to HMMER searches with a HMMER profile for BAHD, CYP and UGT. This resulted in a total of three times 19 sequence sets representing the crude BAHDomes, CYPomes and UGTomes of the 16 plant species (Figure 3A) and the Swissprot plant sequence sets. All crude sequence sets were then prepared for and subjected to the seqrutinator pipeline using default settings. The numbers of homologues after each module were recorded and are in Supplemental File 1. Figure 3B shows similar patterns or the scrutiny and sequence removal on the three superfamilies. Hence, the scrutiny performance can be considered as consistent.

The most effective step is SSR, which indicates that many sequences in particularly the complete proteome of *Pinus taeda* are partial. The complete proteome of *A. thaliana* appears the only with few partials (4, 9 and 2 for BAHD, CYP and UGT, respectively). The CGSR module, which is a more elaborate algorithm but mostly removes partial sequences as well, also identified a significant amount of putative NFH sequences. Interestingly, Figure 3B shows the two complete proteomes with the most SSR hits for all three superfamilies, *P. taeda and Solanum tuberosum* (potato) also gave the most CGSR hits for all three superfamilies. The small complete proteome of lycophyte *Selaginella moellendorffii* also has relatively many SSR and CGSR hits (See supplemental file 1). GIR was the third most efficient module and removed a significant amount of sequences.

Figure 1C shows how many sequences are classified as functional homologues, relative to the initial number of sequences, for each species and each superfamily. Clearly, the complete proteome of *A. thaliana* appears as the best, with 85, 87 and 86% of the sequences classified as functional. On the other hand the recently published complete proteome of *P. taeda* appears as very poor with a mere 18, 16 and 20% of of the sequences classified as functional. The largest differences scrutiny numbers are found for *Z. mays* (76, 58 and 56%) and *S. lycopersicum* (61, 71 and 55%). However, this may be the result of differences in the initial number of homologues. For example, *Solanum lycopersicum* (tomato) has a total of 356 CYP homolgues of which 267 are considered functional whereas closely related *S. tuberosum* has 586 homologues of which 253 are considered functional. Also, *Z. mays* has relatively few BAHD homologues (120 as compared to 142 in *O. sativa*) of which 91 and 94 are considered as functional. Hence, the final numbers of accepted homologues show no clear indications that sequence scrutiny was inconsistent.

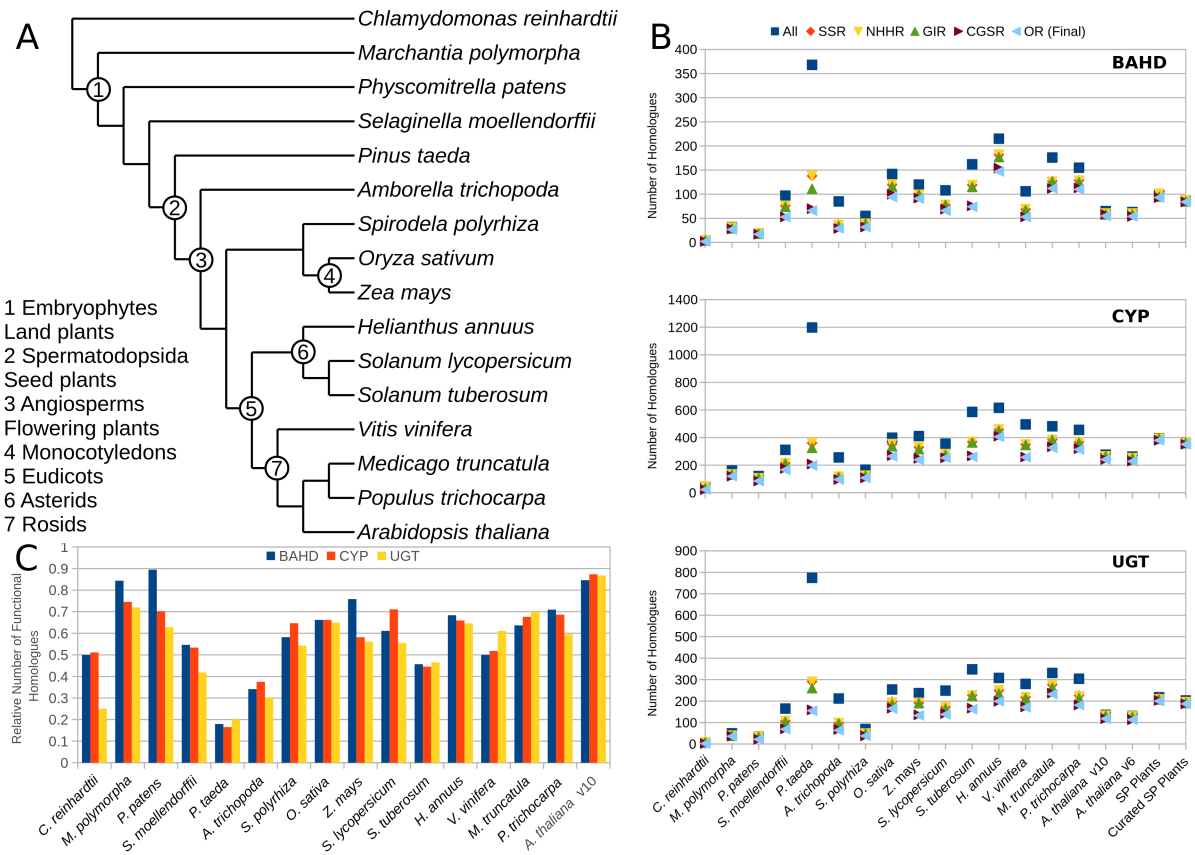


Figure 3: Seqrutinator Performance on 19 BAHDomes, CYPomes and UGTomes. A Approximate Taxonomy of *Selected Species.* ***B*** *Numbers of BAHD, CYP and UGT homologues per species (indicated below bottom plot) found and recovered after each step of the seqrutinator pipeline.* ***C*** *Seqrutinator performance for BAHDomes, CYPomes and UGTomes. Shown is the proportion of the numbers of finally accepted sequences and the numbers of initial sequences per species.*

Most importantly, the Swissprot sequence set appears to have only few sequences that were detected as NFH (7 and 3.4 % for the complete and curated Swissprot dataset, respectively (see supplemental File 1)). On the one hand the low number of removed sequences suggests the scrutiny was not overzealous. On the other hand, analysis of all removed sequences must be performed in order to show if the scrutiny was too stringent. Since the numbers of removed sequences are low this is feasible and will be performed for the second version of this manuscript.

A comparison of the final number of accepted sequences does not adequately reflect the efficiency of the sequence scrutiny since there are both technical and inherent differences in the number of functional homologues that could be detected in each species. This includes the final coverage of the genome, whole and partial genome duplications as well as lineage specific evolution of specialised metabolism. Clearly most seed plants, except *A. trichopoda* and *S. polyrhiza,* have more homologues than the other plants, The number of homologues among most seed plants varies with a factor 2.77, 2.05 and 1.97 for BAHD, CYP and UGT respectively. Closely related *S. lycopersicum* and *S. tuberosum* have 66 and 74; 253 and 267; and 138 and 162 homologues of BAHD, CYP and UGT respectively.

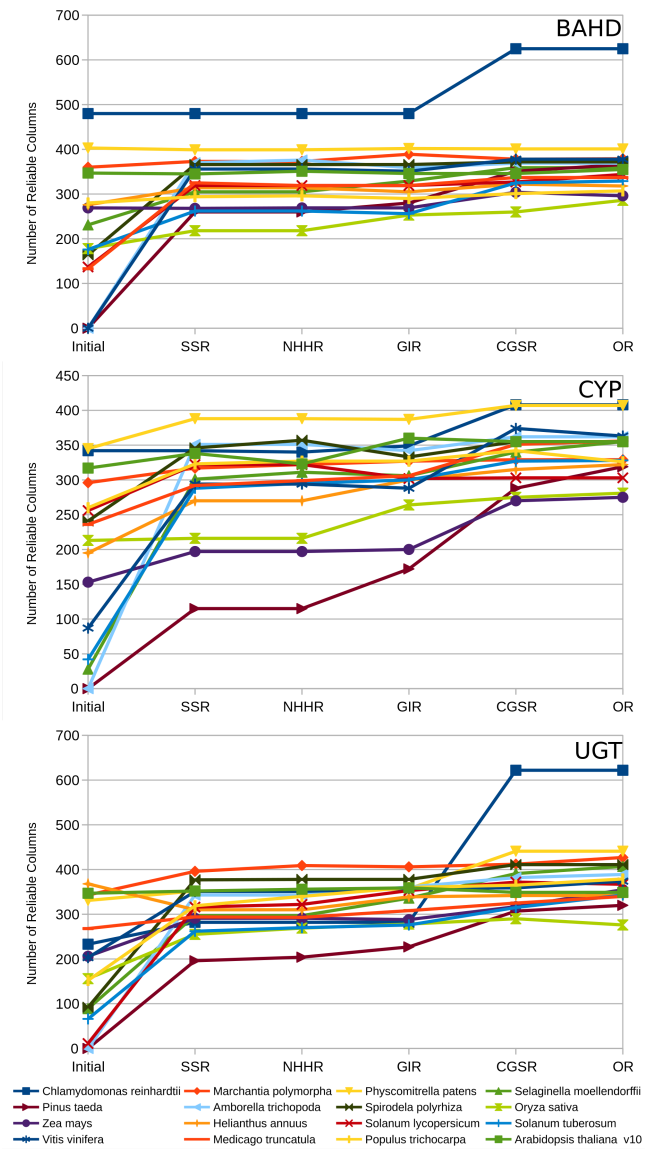
Another way to analyse the performance is by focussing on the amount of sequences that are removed from the various sequence sets in each step. Seed plants have mostly intermediate to high number of accepted sequences whereas non-seed plants have only few. Hence, here we focus on seed plants. Table 2

shows how many sequences were removed from the sequence sets in each step. This confirms that the *P. taeda* complete proteome is still a rather rough version and that *A. thaliana* but also *S. polyrhiza* (duckweed) have rather solid complete proteomes. The data suggest that certain but not all sequence consortia have performed for instance analyses dedicated at partial sequences.

Table 2: Number of Sequences Removed from Seed Plants. *Shown are the the numbers of removed sequence per module in the default pipeline. Red backgrounds indicate the number is above the average (SSR, GIR and CGSR only).*

	SSR			NHHR			GIR			CGSR			PR		
	BAHD	CYP	UGT	BAHD	CYP	UGT	BAHD	CYP	UGT	BAHD	CYP	UGT	BAHD	CYP	UGT
P. taeda	230	837	483	0	0	1	27	36	30	41	117	159	4	10	4
A. trichopoda	48	139	114	1	0	0	2	3	3	5	18	68	0	0	4
S. polyrhiza	16	42	20	0	1	1	0	2	0	7	14	38	0	0	0
O. sativa	22	46	57	0	0	5	4	16	6	16	68	167	6	5	2
Z. mays	20	81	45	1	0	0	0	14	2	7	67	136	1	10	3
S. lycopersicum	30	78	76	0	0	2	0	3	6	9	22	142	3	0	4
S. tuberosum	43	217	120	0	1	2	4	4	2	39	97	167	2	6	5
H. annuus	33	157	58	0	0	0	5	9	15	23	40	199	7	4	0
V. vinifera	37	146	62	0	0	0	2	4	5	14	87	173	0	2	2
M. truncatula	49	96	54	1	3	2	0	1	12	14	49	237	0	7	5
P. trichocarpa	26	88	78	1	3	4	3	3	8	12	42	182	3	7	1
A. thaliana v10	4	9	2	1	3	2	2	1	1	0	22	118	3	0	0
Average	46,50	161,33	97,42	0,42	0,92	1,58	4,08	8,00	7,50	15,58	53,58	148,83	2,42	4,25	2,50

Removal of NFH sequences Results in Improved MSA Quality



The ultimate goal of Seqrutinator is to obtain a sequence set that can be aligned with high fidelity. There are several measures that can used to calculate the quality of an MSA. TCS (32) is one of the most accurate measures but besides that it is computationally expensive, it does not provide highly discriminative scores, which makes it a difficult method for benchmarking. Hence, we sought alternative methods. A simple method is to look at that the length of the MSA, as compared to the length of the mature protein. Now although this is as quantitative measure, MSA length does not accurately describe its quality since, for instance, a single large insert leads to a large MSA but does not necessarily result in a poor MSA.

Figure 4: Number of Reliable Columns of MSAs After Each Step of the Seqrutinator Pipeline.

A more eloquent method is to determine the amount of columns of trimmed MSAs. Since MSAs of complex superfamilies by definition have regions that are either not too reliable or specific to certain subfamilies (hence not truly homologous), such MSAs are trimmed. Trimmers such as BMGE (33) or trimAl (34) remove columns with either high amounts of gaps or high entropy. As such, the length of an MSA following trimming conceptually reflects the amount of reliable columns and can as such be used as a quality measure. Figure 4 shows the result for the scrutinies we performed on all plant species we included. The number of reliable columns increases in almost all steps for all species and the three superfamilies. Besides that the MSAs clearly improve typically with a factor 2 to 4, there appears to be a convergence of MSA length. Note that the plots include the MSAs of the non-seed plants, that end with only few sequences and as such a set with only low complexity. As such the MSAs of these species are typically larger. Another interesting detail is that, among a few others, the initial *P. taeda* MSAs appears as completely unreliable. In general the largest increases in MSA length occur in the SSR step and in the CGSR step. This despite the fact that GIR removes sequences that induce large gap areas.

Thus, Seqrutinator results in sequence sets that show significant improvement in quality, as demonstrated by the increase of the number of reliable columns of MSAs. Note this does not necessarily reflect that the removed sequences are NFH sequences.

Materials and Methods

Initial Datamining

The initial sequence mining was performedby hmmsearch from HMMER (35) using Pfam profiles PF02458 for BAHD (36), PF00067 for CYP (37); and PF00201 for UGT (38) using HMMERs inclusion threshold as inclusion cut-off. Searches were performed in batch using an in house script that and the 16 plant species sequence sets obtained from Phytozome v12.1.6 (39), TAIR v6 that was obtained from TAIR (40), and the Swissprot datasets that were from Uniprot (41, 42).

Multiple Sequence Alignment and Trimming

All MSAs, inlcuding MSAs performed by Seqrutinator are performed using MAFFT-GINSI (43). Trimming was performed using BMGE using standard gap settings, BLOSUM62 and an entropy cut-off of 0.8.

Seqrutinator

A full description of Seqrutinator can be found in supplemental file 1.

Acknowledgements

AA and NS are CONICET fellows, AtH and VF are CONICET Career investigators. Funding was from ANPCyT, project PICT2017-1310.

Cited References and Information Resources

1. Villarreal F, Stocchi N, ten Have A. Functional Classification and Characterization of the Fungal Glycoside Hydrolase 28 Protein Family. J Fungi. 2022;8:217. doi:10.3390/jof8030217.

2. Bondino HG, Valle EM, ten Have A. Evolution and functional diversification of the small heat shock protein/a-crystallin family in higher plants. Planta. 2012;235:1299–313. doi:10.1007/s00425-011-1575-9.

3. Bustamante JP, Radusky L, Boechi L, Estrin DA, ten Have A, Martí MA. Evolutionary and Functional Relationships in the Truncated Hemoglobin Family. PLoS Comput Biol. 2016;12:e1004701. doi:10.1371/journal.pcbi.1004701.
4. Valiñas MA, Have A ten, Andreu AB. Identification of the functions of 4-coumarate-CoA ligase/ acyl-CoA synthetase paralogs in potato. bioRxiv. 2021;:2021.07.06.451337. doi:10.1101/2021.07.06.451337.
5. Revuelta M V., van Kan JAL, Kay J, ten Have A. Extensive Expansion of A1 Family Aspartic Proteinases in Fungi Revealed by Evolutionary Analyses of 107 Complete Eukaryotic Proteomes. Genome Biol Evol. 2014;6:1480–94. doi:10.1093/gbe/evu110.
6. Kumar K, Mhetre A, Ratnaparkhi GS, Kamat SS. A Superfamily-wide Activity Atlas of Serine Hydrolases in Drosophila melanogaster. Biochemistry. 2021;60:1312–24. doi:10.1021/ACS.BIOCHEM.1C00171.
7. Spence MA, Mortimer MD, Buckle AM, Minh BQ, Jackson CJ. A Comprehensive Phylogenetic Analysis of the Serpin Superfamily. Mol Biol Evol. 2021;38:2915–29. doi:10.1093/MOLBEV/MSAB081.
8. Lin LM, Guo HY, Song X, Zhang DD, Long YH, Xing Z Bin. Adaptive Evolution of Chalcone Isomerase Superfamily in Fagaceae. Biochem Genet. 2021;59:491–505. doi:10.1007/S10528-020-10012-Z.
9. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: The protein families database. Nucleic Acids Research. 2014;42:D222-30. doi:10.1093/nar/gkt1223.
10. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Research. 2003;31:371–3. doi:10.1093/nar/gkg128.
11. Orts F, ten Have A. Structure-function analysis of Sedolisins: Evolution of tripeptidyl peptidase and endopeptidase subfamilies in fungi. BMC Bioinformatics. 2018;19:464. doi:10.1186/s12859-018-2404-y.
12. 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;313:903–19. doi:10.1006/jmbi.2001.5080.
13. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res. 2013;41 Web Server issue:W8-14. doi:10.1093/nar/gkt427.
14. Mazin P V, Gelfand MS, Mironov AA, Rakhmaninova AB, Rubinov AR, Russell RB, et al. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. Algorithms Mol Biol. 2010;5:29. doi:10.1186/1748-7188-5-29.
15. Wilkins A, Erdin S, Lua R, Lichtarge O. Evolutionary trace for prediction and redesign of protein functional sites. Methods Mol Biol. 2012;819:29–42. doi:10.1007/978-1-61779-465-0_3.
16. Chagoyen M, García-Martín JA, Pazos F. Practical analysis of specificity-determining residues in protein families. Brief Bioinform. 2016;17:255–61. doi:10.1093/BIB/BBV045.
17. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010;59:307–21. doi:10.1093/sysbio/syq010.
18. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Oxford University Press; 2014. doi:10.1093/bioinformatics/btu033.

19. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490. doi:10.1371/journal.pone.0009490.
20. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4. doi:10.1093/bioinformatics/btg180.
21. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J Mol Biol. 2000;302:205–17. doi:10.1006/jmbi.2000.4042.
22. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2018;20:1160–6. doi:10.1093/bib/bbx108.
23. Löytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 2012;28:1684–91. doi:10.1093/BIOINFORMATICS/BTS198.
24. Szalkowski AM. Fast and robust multiple sequence alignment with phylogeny-aware gap placement. BMC Bioinformatics. 2012;13:1–11. doi:10.1186/1471-2105-13-129/TABLES/3.
25. Deorowicz S, Debudaj-Grabysz A, Gudys A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Sci Rep. 2016;6:1–13. doi:10.1038/srep33964.
26. Shen C, Zaharias P, Warnow T. MAGUS+eHMMs: improved multiple sequence alignment accuracy for fragmentary sequences. Bioinformatics. 2022;38:918–24. doi:10.1093/BIOINFORMATICS/BTAB788.
27. Tumescheit C, Firth AE, Brown K. CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. PeerJ. 2022;10:e12983. doi:10.7717/PEERJ.12983.
28. Chiner-Oms A, González-Candelas F. EvalMSA: A Program to Evaluate Multiple Sequence Alignments and Detect Outliers. Evol Bioinform Online. 2016;12:277–84. doi:10.4137/EBO.S40583.
29. Rieseberg TP, Dadras A, Fürst-Jansen JMR, Dhabalia Ashok A, Darienko T, de Vries S, et al. Crossroads in the evolution of plant specialized metabolism. Semin Cell Dev Biol. 2022. doi:10.1016/J.SEMCDB.2022.03.004.
30. 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:D377–86. doi:10.1093/nar/gks1118.
31. Pagnuco IA, Revuelta MV, Bondino HG, Brun M, Ten Have A. HMMER cut-off threshold tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS One. 2018;13.
32. Chang J-M, Di?Tommaso P, Lefort V, Gascuel O, Notredame C. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction: Figure 1. Nucleic Acids Res. 2015;43:W3–6. doi:10.1093/nar/gkv310.
33. 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.
34. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3. doi:10.1093/bioinformatics/btp348.
35. EDDY SR. A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE. In: Genome Informatics 2009. PUBLISHED BY IMPERIAL

COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.; 2009. p. 205–11. doi:10.1142/9781848165632_0019.
36. Pfam: Family: Transferase (PF02458). https://pfam.xfam.org/family/PF02458. Accessed 22 Mar 2022.
37. Nelson DR. Cytochrome P450 diversity in the tree of life. Biochim Biophys Acta - Proteins Proteomics. 2018;1866:141–54. doi:10.1016/j.bbapap.2017.05.003.
38. Pfam: Family: UDPGT (PF00201). https://pfam.xfam.org/family/PF00201.21. Accessed 22 Mar 2022.
39. Phytozome. https://phytozome-next.jgi.doe.gov/. Accessed 22 Mar 2022.
40. TAIR - Home Page. https://www.arabidopsis.org/. Accessed 22 Mar 2022.
41. Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69. doi:10.1093/nar/gkw1099.
42. UniProtKB. UniProtKB/Swissprot. https://www.uniprot.org/uniprot/?query=reviewed:yes. Accessed 22 Mar 2022.
43. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80. doi:10.1093/molbev/mst010.