Phylogenetic and Functional Classification of the Plant Cytochrome P450 superfamily by using SEQrutinator and HMMERCTTER inception.

Nicolás Stocchi[1], Fernando Villarreal[1], Agustín Amalfitano[2], Marcelo Atencio[3], Carlos Ray[4] and Arjen ten Have[1].
1 IIB-CONICET-UNMdP; 2 ICyTE-CONICET-UNMdP, Mar del Plata, Argentina; 3 EEA-Balcarce INTA, Argentina

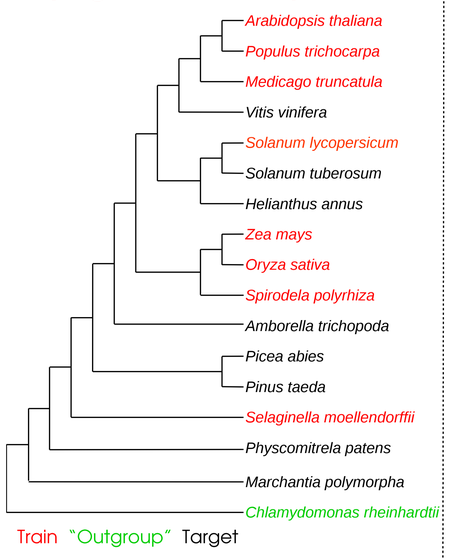
INTRODUCTION

- Cytochrome P450 (CYP) is a paradigm of protein superfamily analysis
- The existing classification of CYP into unrelated clans, families and subfamilies is based on identity which inherently results in conflicts
- Phylogenomics platforms as Panther and CDD do not classify CYP subfamilies, likely due to:
  - -The high complexity of CYP's bona fide sequence space
  - -The high amount of mala fide sequences
- Here we present a first phylogenetic classification of plant CYP space based on:
  - -Objective sequence scrutiny using SEQrutinator (Abstract 57)
  - -Reliable superfamily clustering using HMMERCTTER (Pagnuco et al)
  - -Evolutionary inception

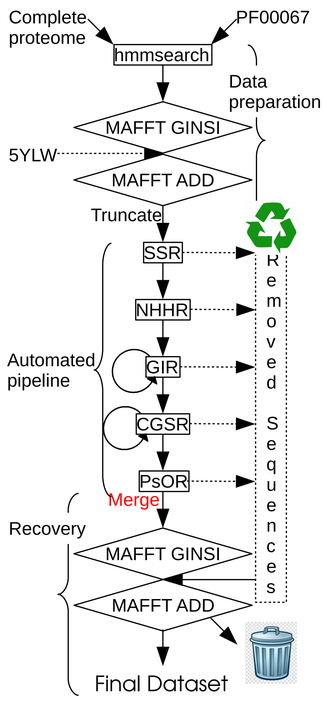
CONCLUSIONs

- Complex superfamilies cannot be phylogenetically clustered directly but require inception.
  - > Iterated, hierarchical clustering is to be used to improve the quality of phylogeny.
  - > A taxonomically well adjusted dataset can be used as training for effective clustering.
  - > Additional sequences are included by HMMERCTTER classification and subfamily phylogeny.
  - > The final tree can then constructed by hierarchically combining subtrees
- SEQrutinator consistently cleans complex sequence datasets yielding high quality MSAs and trees.
  - > Plant CYPs fall into 4 clusters that are 100% P&R-SD* and can be clustered to 182 clusters.
  - > Some clusters correspond to known enzymatic activities but most functions are unknown.
  - > A hifi phylogeny shows relationships between plant CYP clans and independent families.
  - > Many previously assigned subfamily codes are incorrect.
- Sunflower appears to have many additional subfamilies.
- Most of the diversification appears to have occurred in vascular plants.

Phylogenetic tree of proteomes



SEQrutinator pipeline



Sequence homologues were obtained by screening 8 (16) complete plant proteomes using Pfam's hmmer profile for CYP (PF00067) and subjected to SEQrutinator

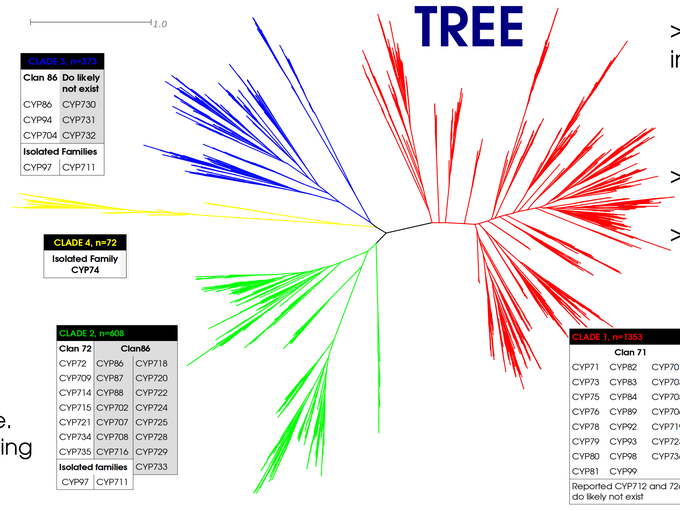
Preparation: MSA using MAFFT GINSI, MAFFT –add of 5ylw.fsa. C and N terminal subsequences lacking secondary structure elements were removed

Step 1: **S**hort **S**equence **R**emover. Length < 65% of reference sequence 5ylw.fsa
Step 2: **N**on-**H**omologous **H**it **R**emover. hmmbuild; hmmsearch: Mean -3SD Not iterated
Step 3: **G**ap **I**nstigator **R**emover. Sequence that induces longest gap > 30 column. Iterated
Step 4: **C**ontinuous **G**ap **S**tring **R**emover. Sequence that has longest (gap-columns excluded) gap > 30. Iterated
Step 5: **Ps**eudogene **O**utlier **R**emover. hmmbuild; hmmsearch: Mean -3SD. Iterated

Following this automated sequence scrutiny accepted sequences were combined and the rejected sequences were subjected to a recovery analysis in which a single, distant subfamily was identified and included in the accepted dataset

DATA

TREE

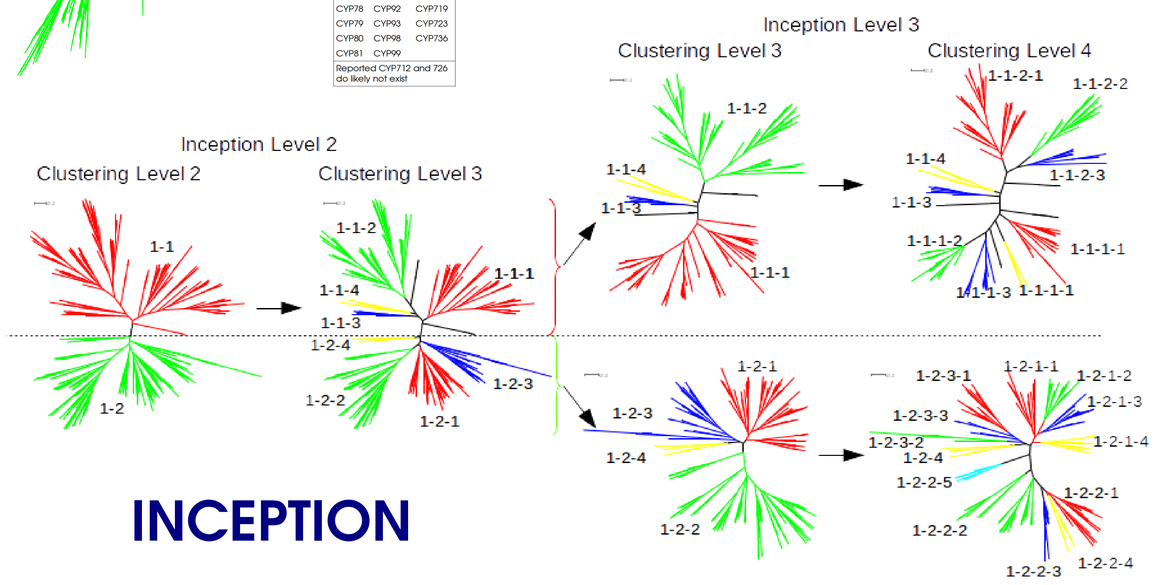


> Evolutionary inception is studying a subfamily without taking into account higher evolutionary hierarchic levels.
  Evolution of different subfamilies is independent.
  → improved signal noise ratio.
  → Improved tree topology.
> Here we show the principle for major clade 1.
  > 50% of the sequences and established plant subfamilies.
>Objective: To identify clusters of known functional subfamilies.
  -Subfamily MSAs and trees when tree topology was uncertain.
  -HMMERCTTER clustering with 100% P&R-SD* regarding of full sequence set.

NINE EIGHT SPECIES PHYLOGENY AND FIRST CLUSTERING

> The initial dataset included sequences from algae *Chlamydomonas rheinhardtii.*
> Its sequences in separate cladeswere removed in order to obtain a less complex tree.
> The obtained phylogeny with 2400 sequences was subjected to HMMERCTTER clustering
  → Only four clusters that are 100% P&R-SD*
> Tables show where annotated sequences (Swissprot or Dr. Nelson's dataset) classify
> Independent but functional family 74 was removed by SEQrutinator but recuperated.
> In order to obtain a more informative functional classification, each of the major clades was separated and subjected to iterative HMMERCTTER clustering. This process is demonstrated for clade 1.

INCEPTION



HMMERCTTER CLASSIFICATION

		182 Cluster Classification		4 Cluster Classification	
Species	**Homologs**	**Orphans**	**Coverage**	**Orphans**	**Coverage**
Solanum tuberosum	264	18	93	0	100
Helianthus annus	406	212	48	0	100
Vitis vinifera	265	29	89	1	100
Amborella trichopoda	95	43	55	2	98
Pinus taeda	209	102	51	4	98
Physcomitela patens	178	169	5	2	99
Marchantia polymorpha	134	112	16	8	94
Chlamydomonas rheinhardtii	39	31	21	23	41

CLASSIFICATION OF TARGET SEQUENCES

> A set 1590 CYP sequences from 8 species was classified using the 4- and 182-cluster transitions.
> Angiosperms were covered completely when using the sensitive 4-cluster partition trainingset.
> Particularly sunflower (*Helianthus annus)* has many sequences that are not covered when the 182-cluster was used.
> This suggests additional subfamilies exist in the four major clades.
> The gymnosperm and lower but multicellular plant species showed high classification coverage.
> The algae *Chlamydomonas rheinhardtii* shows a rather poor coverage, likely since it is monocellular.

ITERATED/HIERARCHIC CLUSTERING OF CLADE 1 PLANT/CLAN 71 CYPS: AN EXAMPLE OF ANALYZING A COMPLEX PHYLOGENY BY EVOLUTIONARY INCEPTION

> Subtree topologies do often not coincide with that of the initial major tree (not shown).
  The major tree has various subfamilies which different functional constraints.
  → Overlapping signals, AKA noise in the MSA underlying the tree.
> This shows the validity of applying the evolutionary inception principle.
> Functional clustering according enzyme activities corresponds well with the inception phylogeny.
> Hierarchic clustering was terminated when a cluster most likely corresponds to a single function.
  -Taxonomic distribution with only one "multiple taxon clade".
> Many of such "minimal" clusters appear to have no known enzyme activities.
> A number of clusters consist of species or family specific sequences suggesting concerted evolution albeit not mediated by hotspot recombination.

***100% Precision and Recall in Self Detection of HMMERCTTER Clusters**
Classification quality is a function of specificity and sensitivity. The basis of HMMEitaRCTTER's high accuracy is that clustering of training sequences result in clusters that detect their members with a HMMER score higher than that of any non-member of the training set. This we refer to as 100% Precision and Recall in Self Detection (100% P&R-SD). The same principle is maintained during classification of target sequences.

Citation: Inti Anabela Pagnuco, María Victoria Revuelta, Hernán Gabriel Bondino, Marcel Brun and Arjen ten Have. HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLOS-One. https://doi.org/10.1371/journal.pone.019375

CONICET — Universidad Nacional de Mar del Plata — I I B
INTA
BBC Bioinformatics and BioComputation UNMdP