Phylogenetic relationships and characterization of UDP-glycosyltransferases (UGTs) in dicotyledonous plants

Hugo Marcelo Atencio; Nicolas Stocchi; Agustín Amalfitano; Fernando Villarreal; Arjen ten Have

Bioinformatics and BioComputation
UN MdP

CONICET
Universidad Nacional de Mar del Plata
I I B

INTA

BACKGROUND

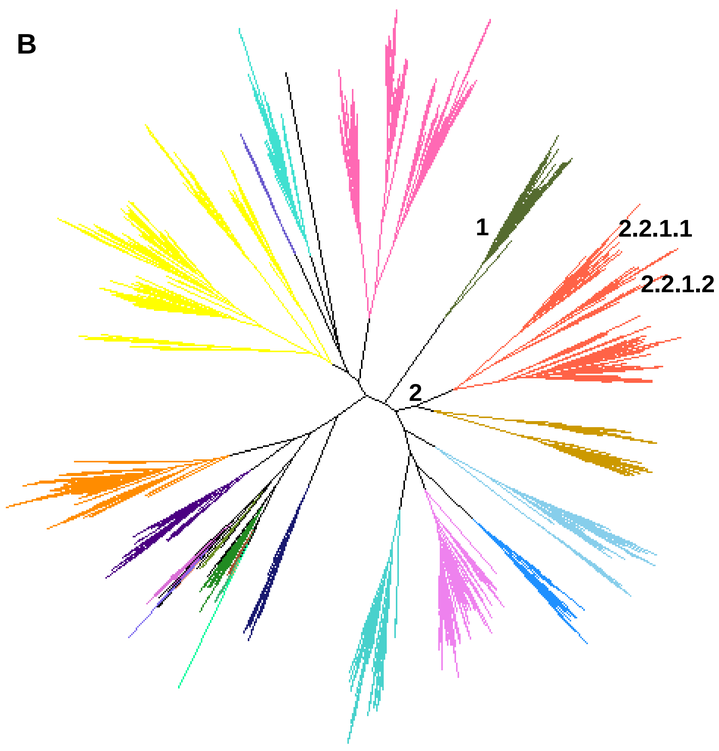
- The protein superfamily UDP-Glycosyl Transferase (UGT) is highly divergent in plants.
- UGTs transfer sugar moieties from a variety of UDP-activated sugars donor to a variety of acceptor molecules.
- Plant UGTs have a consensus motif of 44 amino acids denoted Plant Secondary Product Glycolsyltransferase (PSPG motif).
- UGTs are important in secondary metabolism and they have evolved towards a complex superfamily.
- Substrate promiscuity has likely resulted in erroneous function assignation and might explain why many UGT phylogenies do not correspond with functional classification.
- We are interested in plant UGTs involved in flavonoid and anthocyanin synthesis and need to identify the paralogs involved.
- In the initial sequence mining we identifiied many partial sequences as well as many pseudogenes in datasets.
- The presence of these “bad” sequences in datasets results in erroneous MSA and therewith erroneous phylogenies.

OBJECTIVES

> To reconstruct and characterize a high fidelity phylogeny for structure-function prediction of plant dicotyledoneous UGTs.
> To identify Specific Determining Positions (SDPs) of UGT, involved in the catalytic or structural function on anthocyanin metabolism.

CONCLUSIONS

>A high fidelity phylogenetic tree of UGTs was obtained and will contribute to a more precise functional annotation based on studies of structure-function prediction.
>The UGTs involved in anthocyanidin and anthocyanin glycosilation (3-O-UGT and 5-O-3UGT respectively, are NOT monophyletic despite the clear substrate similarity.
> Identification of SDPs in hampered by the fact that the 3-O-UGT subfamily has no sister clade, by which comparison has to be done indirectly.



A

Seqrutinator / module order	Sequences removed	Sequences accepted
1st SSR	1783	-
2nd NHHR	24	-
3rd GIR	113	-
4th CGSR	632	-
5th PsOR	42	-
Total (4950)	2693	2257

C

Cluster/ Subcluster	MIT	Network MICS	Network nodes
1 vs. 2	7.7	1.37	13
1 vs. 2.2.1.1	5.32	0.85	48

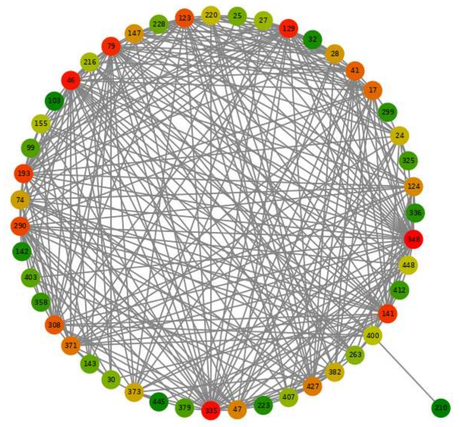
Subcluster	MIT	Network Score	Network nodes
2.2.1.1 vs. 2.2.1.2	4.28	0.8	16

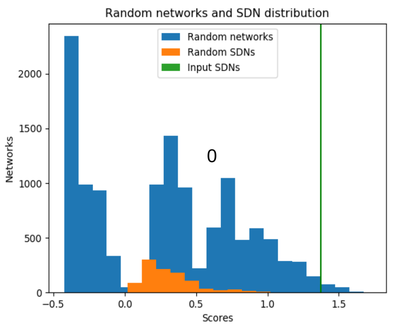
1)- UGT phylogenetic tree and clustering for identification of CDPs and SDPs.

A Of 4950 homologs, a total of 2257 UGT sequences passed SEQrutinator (See table for details) and were considered bona fide sequences used for Multiple Sequence Alignment phylogeny. **B** HMMERCTTER clustering using the obtained high quality phylogeny resulted in 24 clusters that show 100% Precision and Recall. The minor clade 1 corresponds with 3-O-UGT whereas 2.2.1.1 is 5-O-3UGT (UGTs that glycosylate at position 3 and 5 of anthocyanidins and anthocyanin, respectively). Note that these subfamilies evolved from different ancestors, which hampers the identification of SDPs using SDPFox, Mistic and our novel Specificity Networks Mutual Information Connectivity Score (MICS) **C**.

1 vs. 2.2.1.1

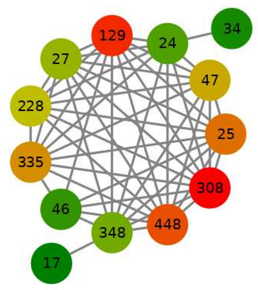
Nodo / position	Node / MICS	Number edges
348	2,07	32
335	1,99	29
46	1,90	27
129	1,87	25
79	1,87	26
141	1,79	25
193	1,69	22
290	1,68	22
123	1,67	21
308	1,57	17
41	1,50	18
17	1,50	19
371	1,48	17
427	1,45	18
47	1,42	17
124	1,40	15
147	1,35	15
28	1,31	14
74	1,23	15
373	1,20	14
382	1,17	13
24	1,16	13
220	1,12	12
448	1,10	12
400	1,10	12
155	1,07	12
216	0,94	10
27	0,94	10
407	0,93	11
263	0,93	10
30	0,72	7
25	0,63	7
228	0,62	7
143	0,47	6
403	0,44	6
99	0,25	5
379	0,23	5
325	-0,11	3
412	-0,14	3
358	-0,19	3
299	-0,41	2
336	-0,41	2
223	-0,42	2
32	-0,51	2
142	-0,53	2
445	-0,59	2
103	-0,61	2
210	-1,16	1

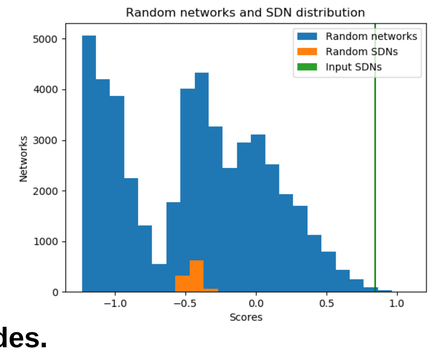




1 vs. 2

Nodo / position	Node / MICS	Number edges
308	1.90	10
129	1.85	10
448	1.78	10
25	1.74	9
335	1.72	9
47	1.68	9
228	1.62	8
27	1.61	9
348	1.59	9
24	1.56	9
46	1.28	6
34	-0.20	1
17	-0.31	1





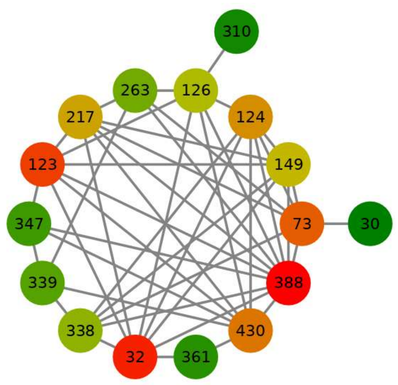
2)- Structure Function Prediction of Distantly related Clades.

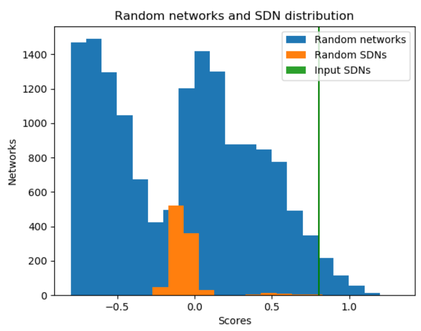
Comparing the 3-O-UGT (1) with 5-O-3UGT (2.2.1.1) is hampered by the lack monophyly. SDPs identified might be involved in the diversification of clades 1 and 2, rather than the diversification of 3-O-UGT (1) with 5-O-3UGT (2.2.1.1) only. The green line on the two plot “Random networks and SDN distribution” corresponds with the MICS of the SDP network. A total of 48 SDPs are identified when comparing 1 with 2.2.1.1., of which many might actually result from the diversification of 1 and 2. Indeed, the 13 SDPs identified by comparing 1 with 2, are identified when comparing 1 with 2.2.1.1.

Although the SDPs identified in the 1-2 comparison are not exclusive to the diversification we analyzed, they might involve mutations that have allowed for the diversification toward the 5-O-3UGT activity. A more careful and stepwise of the family is required to obtain more reliable predictions.

2.2.1.1 vs. 2.2.1.2

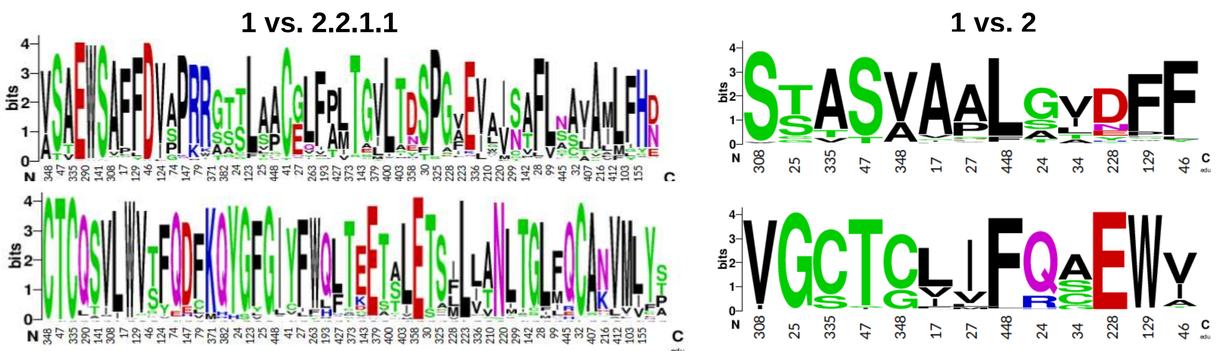
Nodo / position	Node / MICS	Number edges
388	1.59	11
32	1.30	8
123	1.26	7
73	1.17	7
430	1.17	8
124	1.16	7
217	1.15	7
149	1.13	7
126	1.12	7
338	1.12	6
263	0.84	5
339	0.59	4
347	0.55	4
361	0.08	2
310	-0.63	1
30	-0.72	1





4)- Contrast within the monophyletic clade.

Comparing 5-O-3UGT (2.2.1.1) wiht sister clade (2.2.2.2). Only 4 of the 13 predicted SDPs are shared with SDPs identified between clades 1 vs. 2.2.1.1, but do not match those SDPs detected by comparing 1 vs 2.



3)- Specificity Determinig Network (SDN): selection of sensitive and specific threshold.

Application of the MIT at 5.32 identified a fully connected network has the highest Score SDN= 0.85 and contains 48 SDPs/Nodes between 1 vs 2.2.1.1. The MIT determined for clades 1 vs 2 was 1.37 with the largest network MICS. The logo sequences for each of the comparisons are shown.

Data Mining and Methods

- UGT homologues with HMMER. 17 complete proteomes of dicotyledonous plants and the control-set of SwissProt. - 2257 UGT homologs sequences scrutinized objectively with SEQrutinator. - MSA: MAFFT GINS-i, Trimming: BMGE. - Phylogeny: PHYML; Clustering: HMMERCTTER (100% P&R), CDPs (Cluster determining Positions): SDPFox. - MI: mutual information (MISTIC: http://mistic.leloir.org.ar/index.php). - MICS: mutual information connectivity score. - **MICS= $\log_{10}$(C*[Σ((1,5*MI)-MIT)]/N-1**. MI= mutual information, MIT= mutual information threshold, N= number of nodes, C= connectivity. - Network Score: average of MICS (SDN score is compared to 1000 randomly selected, connected subnetworks inferred by the same MIT). - Logos: https://weblogo.berkeley.edu/logo.cgi.