Universiteit Leiden
The Netherlands

Opleiding Informatica

Discovery of ENOD40 RNA in Poaceae

Daan Vijfvinkel

Supervisors:
Dr. A.P. Gultyaev & Dr. K.J. Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

10/07/2021

Abstract

ENOD40 is a plant gene that produces long functional noncoding RNA with small open reading frames. This RNA product has been documented as having highly conserved secondary structure elements, which are suggested to be more important to its functioning than its protein-coding potential. In the family of Poaceae ENOD40 has a unique property; One of the structure elements, known as domain 2, is significantly less conserved than in other families. Since the last research into ENOD40 in Poaceae, the databases containing genome data of various organisms have grown greatly. In this thesis, a set of new ENOD40 homologues found in the new genome data is presented, containing a multitude of duplicate homologues. Using this set of ENOD40 homologues, a set of secondary RNA structure consensus groups has been constructed for domain 2.

Contents

1 Introduction

1.1 Background

1.1.1 ENOD40

Early nodulin 40 or ENOD40 is a plant gene present in all Magnoliopsida (flowering plants) first identified in Glycine max (soybean) [YKH+93]. Early nodulin genes [SvdWZ+89] are genes involved in the early development of root nodules in legumes. Root nodulation happens in a symbiotic process between legumes and bacteria like rhizobia after which the bacteria fix nitrogen inside the root nodules. ENOD40 is one of the early nodulin genes which shows upregulation during this root nodulation process, possibly having a regulatory function [GRG+03] and also shows activation during symbiosis with phosphate-acquiring mycorrhizae fungi [VRFG+97]. It is not completely clear however how exactly ENOD40 contributes to both processes.
ENOD40 produces long functional noncoding RNA [Edd99], with only small open reading frames instead of longer peptide encodings. The shorter peptide products have been documented as being functional [SJC+01, RSM+02]. Maybe more interesting however, is that the ENOD40 gene RNA product contains strongly conserved secondary RNA structure elements (nomenclature will be used from [GRG+03]). These structure elements are presented in Figure 1.

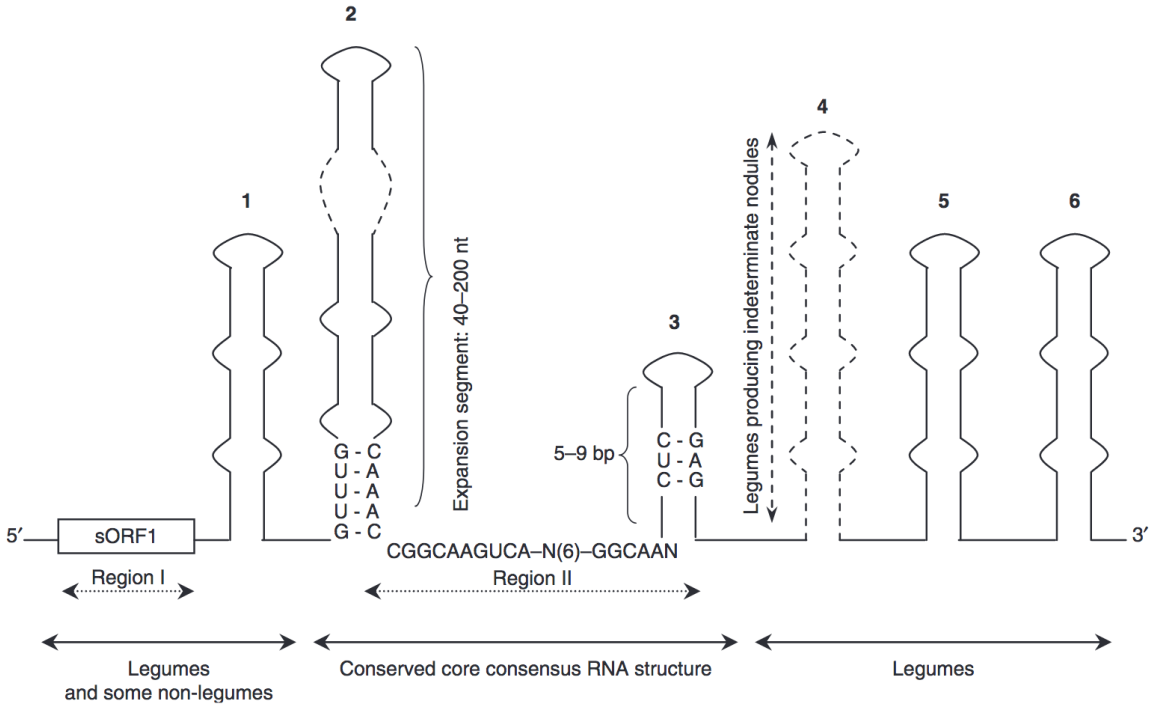

Figure 1: "Conserved structures and sequence motifs in enod40 RNAs. Domains 1–6 are shown schematically and not in scale, domain numbering isfrom (29). The locations of sORF1 and regions I and II (dashed double arrows) are also shown. Double arrows indicate the extent of conservation ofparticular structures. Some deviations from the shown consensus sequences are possible.", figure and description directly from [GR07].

While ascertaining ENOD40's function, it is of interest to consider these conserved structural elements, as they seem to be more conserved than the protein-coding potential. The structure elements are not all present in all species; only domain 2 and domain 3, as well as the sequence of region II intermediate of these domains are conserved across all orders [Rut03, GRG+03]. The short open reading frame sORF1 in region I is also well conserved. Other domains are more specific for legumes, with previous research suggesting that the presence of these domains relates to the root nodulation and nitrogen fixing symbiosis exclusive to legumes. This is further corroborated by the presence of domain 4 in species being matched by a difference in determinate or indeterminate nodule forming [GRG+03]. Denoted in Figure 1, domain 2 and domain 3 have well conserved motifs within the hairpin structure, while domain 2 is more variable in length compared to the other domains.

1.1.2 Poaceae

The family of Poaceae, trivially known as grasses, embodies one of the largest families of Magnoliopsida. They inhabit every continent, including Antarctica [KAC+18]. To humans, Poaceae is considered the most important plant family, consequently carrying heavy economical importance [Wat90]. The reason for this is because they are the main food source for humans; About half of our diet consists of Poaceae [Gna09]. Furthermore, they can be used for other products as well, ranging from bamboo baskets to modern biomass fuel.
As a result of its importance to humans, species in the Poaceae family have been researched extensively and have been well defined in the field of taxonomy. As an example, Oryza sativa (Asian rice) is one of the most annotated species in available in the NCBI database.
The family is commonly divided into two major clades, the BOP clade and PACMAD clade, which contain the majority of Poaceae species [Hod18]. A major difference between these two clades is that species of the BOP clade exclusively use C_3 photosynthesis, while the PACMAD clade contains species that independently have developed C_4 photosynthesis [II12].

1.1.3 ENOD40 in Poaceae

ENOD40 in Poaceae has been documented as having some interesting circumstances compared to other plant families in research considering ENOD40 in different orders.
Firstly, there is a difference in conservation level noted in a specific part of the structural domain 2 of the RNA product (Figure 1). Domain 2 is of variable length and is compared to expansion segments [GR07]. However, in other families there is the consistent and conserved motif GUUUG/CAAAC in the base of the stem-loop double stranded hairpin structure. This motif was only found in two homologues in Poaceae, while the motif was not present in all other species studied [GR07]. This means that domain 2 in Poaceae ENOD40 is very variable, in both stem-loop length and nucleotide sequence, as well as a less conserved structural base motif.
The second point of interest is that species in the Poaceae family have been found to contain more than one homologue of ENOD40 [CRA+03]. These duplicate homologues have been documented for Oryza sativa and Zea mays (maize). Interestingly, the homologues per species are distinctly different which calls into question the evolutionary aspect, since the species belong to the BOP clade and PACMAD clade, respectively. This is more the case considering that the only 'perfect' GUUUG/CAAAC domain 2 motif is found in only one of the two Oryza sativa homologues.

After these distinctive features were known, an effort was made to more precisely define domain 2 in Poaceae, which also added more species containing ENOD40 homologues [Kos12]. In this work, two nucleotide sequences were found in domain 2 that were more conserved than the GUUUG/CAAAC. There is a lesser conserved motif present in the stem UUCCGUGGU/GAGGCGUGCA, which nearly always has at least one mismatch and/or bulge. Furthermore, the loop at the end of the stem has a strongly conserved CC/GG pairing closing the loop.

1.2 Thesis overview

First, the tools that have been used during research for this thesis are introduced. Afterwards, the specific way these tools have been used is described.
The set of newly discovered ENOD40 homologues is joined with the set of previously known homologues. Some previously known homologues have had specifications updated. The full set of all known ENOD40 homologues in Poaceae is presented in a table. Utilizing the information in this full set, a series of secondary RNA structure predictions for domain 2 was done and then compiled in order to create a framework of domain 2 variant grouping.

2 Materials and Methods

2.1 Tools

To understand the method used to generate the results, it is important to understand the bioinformatic tools used. For the purpose of finding new homologues, a method was selected for searching through all Poaceae genome data currently available. As a means to make this search more efficient, a taxonomy inspection method was also used. In order to determine the structure for domain 2 of the different homologues, a method was used to predict secondary RNA structures. All tools used are publicly available on different web servers.

2.1.1 BLAST

The search for new ENOD40 homologues in Poaceae was conducted using the Basic Local Alignment Search Tool (BLAST) on the NCBI servers (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`) [AGM+90]. BLAST is an algorithm used to find similarity in biological sequences, such as DNA, RNA and proteins, based on a scoring system, positively scoring for sequence matches and negatively scoring for mismatches and gaps. The algorithm first tries to match short subsequences (default 28nt or 6aa) of the given input query sequence with subsequences of the queried sequences. If found, it subsequently tries to expand this match until the score quality of the match decreases. If the quality of the full match is deemed high enough, it is selected and added to the output.
With BLAST on the NCBI servers it is possible to query a given sequence on databases which contain all available genome data from the International Nucleotide Sequence Database Collaboration (INSDC) [KMNC12] using the Entrez cross-database search system [SEOK96]. If used like this, BLAST will produce a set of database accession hits, along with an Expect value (E-value) describing the chance that the found sequence from the database was found by chance. These hits can then further be manually inspected for validity.
The BLAST queries were performed on the INSDC databases Nucleotide Collection (NR/NT), Whole-genome Shotgun Contigs (WGS) and RefSeq Genome Database (refseq_genomes). The NR/NT database contains non-redundant nucleotide sequences from anywhere on an organisms genome. The WGS database contains incomplete sections of nucleotide sequences called 'contigs' and 'scaffolds'. Contigs are sequenced and afterwards reassembled by an algorithm, creating scaffolds consisting of contigs and gaps. Scaffolds or contigs can overlap, creating redundant sequences in the database. The refseq_genomes database contains non-redundant full reference genomes for a specific set of organisms, which are used amongst others for stable annotation and identification for genes.

2.1.2 Taxonomy browser

Not all Poaceae species have been combed through, because the Poaceae family contains at least 12.000 species [CB16]. A search through all of these would be computationally inefficient and would be aborted by NCBI BLAST for memory overconsumption. Instead, the NCBI taxonomy browser [Fed12] was used in order to select tribes and genera from the two major clades of the Poaceae family, the BOP clade and PACMAD clade. The selection was hand picked, based on how much genome data is available for each respective tribe or genera according to the taxonomy browser. This information is gathered from the Entrez cross-database search system [SEOK96]. An example of these presented statistics is seen in Figure 2.

Entrez records

Database name	Subtree links	Direct links
Nucleotide	2,281,135	319,531
Protein	438,703	61,811
Structure	249	70
Genome	1	1
Popset	1,178	1,028
Conserved Domains	12	5
GEO Datasets	20,398	14,961
PubMed Central	29,321	29,321
Gene	95,324	149
HomoloGene	9,787	9,787
SRA Experiments	86,658	65,625
GEO Profiles	670,939	670,939
Protein Clusters	15,559	-
Identical Protein Groups	204,577	44,070
Bio Project	6,323	4,916
Bio Sample	94,492	50,484
Bio Systems	1,265	751
Assembly	91	44
Probe	186,169	92,681
PubChem BioAssay	478	448
Taxonomy	9	1

Figure 2: Example of Entrez records of Oryza sativa in the NCBI taxonomy browser.

2.1.3 Secondary RNA structure prediction algorithm

Domain 2, the region of interest in Poaceae ENOD40 has, contrary to other families' ENOD40, a low conservation for the sequence of domain 2. In order to determine if the found domains 2 still have a secondary structure, and if so which range of nucleotides makes up this structure the Fold program on the RNAstructureWeb servers was used (`https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Fold/Fold.html`) [BRSM13]. Like similar secondary RNA structure prediction algorithms, it is a dynamic programming algorithm that computes the optimal thermodynamic free energy differential for a given input nucleotide sequence and outputs the structure that provides this optimal free energy.

2.2 Genome database search

Using BLAST the set of tribes and genera selected from the Poaceae family based on data from the taxonomy browser was queried upon to find new putative ENOD40 (see section 2.1.1). A selection was made from the previously found ENOD40 in Poaceae to be used as queries, this selection is depicted in Table 1. These accessions were truncated prior to querying to include only the range from the startcodon of sORF1 to the end of domain 3. This range contains the conserved signature sequences present in all Poaceae ENOD40.

Species	Accession
Zea mays(1)	CD990776
Zea mays(2)	DN209550
Oryza sativa(1)	AB024054
Oryza sativa(2)	AU101849
Sorghum bicolor	BE362667
Hordeum vulgare	AF542513
Triticum aestivum	BJ278615
Brachypodium dystachyon	DV479239
Avena sativa	CN815024
Lolium perenne	AF538350

Table 1: Selected species and accesions for query on database, selected from [GR07].

Taking into account the highly variable domain 2, the BLAST variant BLASTn was used, which is optimized for "somewhat similar sequences". All tribes and genera selected were individually blasted, using the organism taxonomy ID limitation option. From the BLAST results, hits with an E-value lower than 0.01 were selected for manual analysis to filter out the hits which did not contain ENOD40. This analysis determined the validity of the hits by searching for the most conserved signature sequence of region II (CGGCAAGUCA-N(6)-GGCAAN). If any doubts remained, the presence of sORF1 and the bottom of the hairpin of domain 3 (CUC/GAG) were also searched for further confirmation.

2.3 Domain 2 secondary structure predictions

Once a set of ENOD40 in Poaceae was found, the secondary RNA structure of their domain 2 were determined using the RNAstructureFold server (see section 2.1.3). In this program, all fold options were kept at default values, except for the temperature, which was set at 293.15 °K instead of the default 310.15 °K.
Regions with similarity to domain 2 of known ENOD40 were manually selected and given as input. Depending on the output the input sequences were refined, removing unnecessary nucleotides until only the hairpin secondary structure was isolated. The sequence which made up the hairpin was then added to a list for comparison of the differences of the various domains 2 of the different Poaceae ENOD40.

3 Results

During the timespan between the last efforts to document ENOD40 and now, the databases queried have grown greatly with newly annotated sequences. This increase logically also entails an increase in sequenced ENOD40 homologues. Although both the nonconserved regions and domain 2 are highly variable in ENOD40 in Poaceae, performing BLAST searches on the various databases of the INSDC has proven fruitful in producing accesions of putative ENOD40 homologues. Further

analysis confirmed the existence of the other conserved regions, supporting the validity of the sequences as ENOD40.

The region of interest in Poaceae ENOD40, the less conserved domain 2, has been examined for all currently known Poaceae ENOD40 in order to find a secondary RNA structure consensus. Instead of a single consensus, a set of structure consensuses has been defined. The different ENOD40 homologues have been grouped based on distinct sequence similarities.

3.1 Homologues

In this work, 103 new or updated ENOD40 homologues in Poaceae are introduced. All currently known ENOD40 in Poaceae are listed in Table 2, split based on the two major clades of the Poaceae family.
In the table, the nucleotide positions of the found ENOD40 within their respective accessions are noted from the start of sORF1 to the end of domain 3, as well as the position of the region of interest domain 2 within this range. While some sequences miss or have distorted variants of a signature ENOD40 conserved region, the other regions are present and the sequences are indubitably ENOD40.
If known, the chromosome on which the gene was found is also reported. The previously known sequences, mostly from the Expressed Sequence Tags (EST) database, and a set of new sequences from the WGS database have no chromosome known. All new sequences from the same species without chromosome notation are distinct; If duplicate homologues with similar sequences lacking chromosome notation exist in the same species, only one is included in the table.
Some homologue information has been updated; If a species was already known to have ENOD40 then the original accession and genome position are kept, except when a new accession is available which contains chromosome information. For some homologues, the domain 2 was refined in later research. In these cases, the latest domain 2 is noted, with annotation credit to both the original finder and refiner of the domain.

BOP clade					
Species	Accession	Location	Domain 2 (size)	Chromo	Annotation
RC = Reverse Complement. [a]Region II is missing/distorted. [b]sORF is missing/distorted. [c]End of domain 3 is missing/distorted. † = accession from EST database.					
Oryza sativa(1)	AB024054†	2256-2455	2361–2405 (45)	-	[KTS+99]
Oryza sativa(2)	AU101849	29-271	151–225 (75)	-	[Rut03]
Oryza branchyantha(1)	XM_015836663	1333-1527	1442-1484 (43)	4	This work
Oryza branchyantha(2)	XR_001549656	88-339	219-299 (81)	2	This work
Oryza glaberrima	ADWL01008900	34490-34684	34595-34639 (45)	-	[Kos12]
Oryza longistaminata	HS384133†	18-212	123-167 (45)	-	[Kos12]
Oryza rufipogon	CU405621	59-301	178-258 (81)	-	[Kos12]
Lolium perenne	AF538350	93-290	199-244 (54)	-	[Lar03],This work
Festuca arundinacea	DT701589†	(<1)-196	97-150 (54)	-	[GR07], [Kos12]
Festuca pratensis	GO893341†	17-209	115-169 (55)	-	[Kos12]
Hordeum vulgare	CABVVH010000002	516075054-516075254	516075155-516075209 (54)	2H	This work
Hordeum vulgare cultivar	JAFEGY010024544	497077756-497077956	497077857-497077911 (54)	2H	This work
Leymus chinensis	CN465797†	(<1)-61	ns	-	[GR07]
Triticum aestivum	BJ278615†	41-223	118-181 (64)	-	[Rut03],[CRA+03], [Kos12]
Triticum dicoccoides isolate(1)	NC_041382	610722571-610722764	612915131-612915178 (48)	2A	This work
Triticum dicoccoides isolate(2)	NC_041382	612915045-612915221	610722674-610722722 (48)	2A	This work
Triticum dicoccoides isolate(3)	NC_041382	561645219-561645395	561645305-561645353 (48)	2B	This work
Tricitum monococcum	BQ802914†	15-198	106-154 (49)	-	[Kos12]

Avena sativa	CN815024†	(<1)-86	-	-	[GR07]
Avena barbata	GR366484†	316-516 RC	356-409 RC (54)	-	[Kos12]
Brachypodium dystachyon	NC_016135	18709399-18709590	18709499-18709550 (51)	5	This work
Phyllostachys edulis(1)	FP097810	12-205	113-163 (51)	-	[Kos12]
Phyllostachys edulis(2)	WJQQ01000519	311542-311758 RC	311582-311646 RC (56)	-	This work
Phyllostachys edulis(3)	WJQQ01000319	716274-716481 RC	716315-716369 RC (66)	-	This work
Aegilops tauschii	NC_053036	461535595-461535788	461535662-461535709 (47)	2D	This work
Dendrocalamus latiflorus isolate(1)	JACBGG010000019	12984279-12984468 RC	12984320-12984370 RC (51)	19	This work
Dendrocalamus latiflorus isolate(2)	JACBGG010000020	37905807-37905996	37905848-37905898 (51)	20	This work
Dendrocalamus latiflorus isolate(3)	JACBGG010000029	26631657-26631848	26631758-26631809 (51)	29	This work
Dendrocalamus latiflorus isolate(4)	JACBGG010000030	10910428-10910619 RC	10910468-10910518 RC (51)	30	This work
Dendrocalamus latiflorus isolate(5)	JACBGG010000025	26523860-26524079	26523977-26524036 (59)	25	This work
Dendrocalamus latiflorus isolate(6)	JACBGG010000026	11449547-11449763 RC	11449591-11449649 RC (59)	26	This work
Dendrocalamus latiflorus isolate(7)	JACBGG010000031	15303932-15304150 RC	15303972-15304037 RC (66)	31	This work
Dendrocalamus latiflorus isolate(8)	JACBGG010000032	35219093-35219314	35219209-35219275 (66)	32	This work
Dendrocalamus latiflorus isolate(9)	JACBGG010000047	11034817-11034997 RC	11034770-11034903 RC (40)	47	This work
Dendrocalamus latiflorus isolate(10)	JACBGG010000048	28722533-28722713	28722627-28722667 (40)	48	This work

PACMAD clade

Species	Accession	Location	Domain 2 (size)	Chromo	Annotation
RC = Reverse Complement. [a]Region II is missing/distorted. [b]sORF is missing/distorted. [c]End of domain 3 is missing/distorted. † = accession from EST database.					
Zea mays(1)	CD990776†	58-252	152-207 (56)	-	[Rut03],[CRA+03]
Zea mays(2)	DN209550†	82-360	221-314 (94)	-	[Rut03],[CRA+03]
Sorghum bicolor(1)	NC_012875	50361806-50361985[c]	50361894-50361948 (54)	6	This work
Sorghum bicolor(2)	NC_012873	56600787-56601040	56600921-56601001 (80)	4	This work
Sorghum bicolor(3)	NC_012873	56609660-56609887[a]	56609790-56609860 (70)	4	This work
Saccharum hybrid cultivar(1)	QPEU01304195	4475-4663	4570-4624 (54)	-	This work
Saccharum hybrid cultivar(2)	QPEU01330982	5991-6219	6113-6180 (67)	-	This work
Saccharum officinarum	CA155599†	54-248	149-202 (54)	-	[GR07]
Saccharum spontaneum cultivar(1)	QVOL01000017	23135930-23136118 RC	23135970-23136023 RC (54)	5A	This work
Saccharum spontaneum cultivar(2)	QVOL01000017	31159077-31159263 RC	31159117-31159170 RC (54)	5A	This work
Saccharum spontaneum cultivar(3)	QVOL01000018	17197699-17197887 RC	17197739-17197792 RC (54)	5B	This work
Saccharum spontaneum cultivar(4)	QVOL01000019	14702748-14702936 RC	14702788-14702841 RC (54)	5C	This work
Saccharum spontaneum cultivar(5)	QVOL01000013	24226335-24226567 RC	24226375-24226441 RC (67)	4A	This work
Saccharum spontaneum cultivar(6)	QVOL01000013	24235895-24236127 RC	24236035-24236001 RC (67)	4A	This work
Saccharum spontaneum cultivar(7)	QVOL01000014	20359581-20359813	20359621-20359687 (67)	4B	This work
Saccharum spontaneum cultivar(8)	QVOL01000015	24002351-24002583 RC	24002391-24002457 RC (67)	4C	This work
Saccharum spontaneum cultivar(9)	QVOL01000016	24557884-24558105 RC	24557924-24557990 RC (67)	4D	This work
Setaria italica(1)	NC_028456	24494045-2449423	3407-3458 (51)	7	This work
Setaria italica(2)	NC_028450	31431744-31431984	69-145 (76)	1	This work
Setaria viridis(1)	XM_034747225	3372-3569	3471-3522 (51)	7	This work
Setaria viridis(2)	XR_004641991	261-490	386-461 (76)	1	This work
Setaria viridis cultivar(1)	CP050801	23339809-23340006	23339908-23339599 (51)	7	This work
Setaria viridis cultivar(2)	CP050795	31531944-31532173	31532058-31532134 (76)	1	This work
Panicum hallii(1)	NC_038048	39151637-39151828	39151737-39151788 (51)	7	This work
Panicum hallii(2)	NC_038042	49820833-498210506	49820935-49821011 (76)	1	This work
Panicum miliaceum(1)	PQIB02000013	29172885-29173076	29172985-29173036 (51)	15	This work
Panicum miliaceum(2)	PQIB02000015	24565177-24565368 RC	245652218-24565268 RC (51)	16	This work
Panicum miliaceum(3)	PQIB02000009	32610089-32610228[b] RC	32610130-32610180 RC (51)	6	This work
Panicum miliaceum(4)	PQIB02000012	33491665-33491897	33491785-33491855 (70)	12	This work
Panicum miliaceum cultivar(1)	PPDP02000014	9884748-9884939 RC	9884848-9884899 RC (51)	14	This work
Panicum miliaceum cultivar(2)	PPDP02000015	24286432-24286623 RC	24286532-24286583 RC (51)	15	This work
Panicum miliaceum cultivar(3)	PPDP02000017	8686451-8686584[b] RC	8686492-8686542 RC (51)	17	This work
Panicum miliaceum cultivar(4)	PPDP02000010	34933757-34933890[b] RC	34933798-34933803 RC (51)	10	This work
Panicum miliaceum cultivar(5)	PPDP02000012	33350933-33351165	33351053-33351123 (70)	12	This work
Panicum miliaceum cultivar(6)	PPDP02000006	10781689-10781880 RC	10781729-10781804 RC (76)	6	This work
Panicum virgatum(1)	JABWAI010000055	711150-711374	711276-711327 (60)	1N	This work
Panicum virgatum(2)	JABWAI010000023	4268914-4269149	4269037-4269107 (70)	1K	This work
Panicum virgatum(3)	XR_005677875	1496-1803	1687-1758 (71)	7N	This work
Panicum virgatum(4)	XM_039919048	3466-3635	-	7K	This work
Digitaria exilis(1)	LR994616	22355198-22355389	22355296-22355350 (54)	7A	This work
Digitaria exilis(2)	LR994617	19445047-19445237	19445145-19445198 (53)	7B	This work
Digitaria exilis(3)	LR994604	27431556-27431786	27431674-27431744 (70)	1A	This work
Digitaria exilis(4)	LR994605	25489084-25489317	25489205-25489275 (70)	1B	This work
Alloteropsis semialata(1)	QPGU01000595	58722927-58723171	58723057-58723132 (75)	1	This work
Alloteropsis semialata(2)	QPGU01000685	32132411-32132655	32132541-32132616 (75)	4	This work
Alloteropsis semialata(3)	QPGU01000082	51321439-51321630[a]	51321540-51321589 (49)	7	This work

Cenchrus americanus cultivar	LKME02052029	253386136-253386331	253386233-253386284 (51)	3	This work
Dichanthelium oligosanthes cultivar(1)	LWDX02073916	14913-15099	15011-15057 (45)	-	This work
Dichanthelium oligosanthes cultivar(2)	LWDX02058855	3993-4225	4110-4186 (76)	-	This work
Chrysopogon serrulatus isolate(1)	JADLZL010007362	80690-80878 RC	80730-80783 RC (54)	-	This work
Chrysopogon serrulatus isolate(2)	JADLZL010003248	336723-336960	336845-336921 (76)	-	This work
Echinochloa crus-galli(1)	OAMR01000183	1629808-1629999	1629908-1629959 (51)	-	This work
Echinochloa crus-galli(2)	OAMR01000237	278515-278705	278613-278664 (51)	-	This work
Echinochloa crus-galli(3)	OAMR01000510	208142-208373	208262-208331 (69)	-	This work
Eleusine indica(1)	QEPD01000755	75189-75382	75292-75337 (45)	-	This work
Eleusine indica(2)	QEPD01000043	400207-400443 RC	400257-400326 RC (70)	-	This work
Eleusine coracana cultivar(1)	LXGH01289804	66659-66852	66762-66807 (45)	-	This work
Eleusine coracana cultivar(2)	LXGH01420798	4251-4487	4368-4438 (70)	-	This work
Eragrostis curvula cultivar(1)	RWGY01000031	18072955-18073147 RC	18073004-18073041 RC (38)	-	This work
Eragrostis curvula cultivar(2)	RWGY01000011	11879222-11879448 RC	11879265-11879334 RC (70)	1	This work
Eragrostis nidensis cultivar(1)	JAAXCT010000013	770932-771123 RC	770975-771021 RC (47)	-	This work
Eragrostis nidensis cultivar(2)	JAAXCT010001162	141764-141988 RC	141807-141876 RC (70)	-	This work
Eragrostis tef cultivar(1)	LAPY01002384	52682-52881	52792-52837 (45)	-	This work
Eragrostis tef cultivar(2)	LAPY01001660	57115-57306 RC	57158-57204 RC (47)	-	This work
Eragrostis tef cultivar(3)	LAPY01002226	71554-71780 RC	71597-71666 RC (70)	-	This work
Hyparrhenia diplandra(1)	JADLZK010012887	110619-110808	110715-110769 (54)	-	This work
Hyparrhenia diplandra(2)	JADLZK010004634	236342-236580	236462-236541 (79)	-	This work
Hyparrhenia diplandra(3)	JADLZK010010082	96330-96519[a] RC	96381-96414 RC (34)	-	This work
Miscanthus lutarioriparius(1)	CAJGYO010000011	26393706-26393894 RC	26393754-26393791 RC (38)	11	This work
Miscanthus lutarioriparius(2)	CAJGYO010000012	27600832-27601020 RC	27600880-27600917 RC (38)	12	This work
Miscanthus lutarioriparius(3)	CAJGYO010000008	25656461-25656702	25656587-25656663 (76)	8	This work
Miscanthus sacchariflorus cultivar(1)	PUID01021797	(<12950)-13062[b]	12977-13015 (38)	-	This work
Miscanthus sacchariflorus cultivar(2)	PUID01053440	1744-1986 RC	1784-1863 RC (80)	-	This work
Themeda triandra isolate(1)	JAENPR010000383	193081-193269 RC	193125-193166 RC (42)	-	This work
Themeda triandra isolate(2)	JAENPR010000272	417638-417882	417763-417843 (80)	-	This work
Urochloa ruziziensis cultivar(1)	PVZT01000902	20638-20838[c] RC	20683-20733 RC (51)	-	This work
Urochloa ruziziensis cultivar(2)	PVZT01009743	7585-7820	7705-7781 (76)	-	This work
Urochloa ruziziensis isolate(1)	WEIB01000006	41235053-41235245[c]	41235150-41235201 (51)	8	This work
Urochloa ruziziensis isolate(2)	WEIB01000003	11822955-11823190	11823075-11823151 (76)	6	This work
Zoysia japonica(1)	BCLF01000004	1784170-1784361	1784273-1784318 (45)	-	This work
Zoysia japonica(2)	BCLF01000007	5398060-5398301 RC	5398100-5398179 RC (80)	-	This work
Zoysia matrella(1)	BCLG01003487	9068-9259 RC	9112-9156 RC (45)	-	This work
Zoysia matrella(2)	BCLG01005723	3278-3522 RC	3321-3400 RC (80)	-	This work
Zoysia pacifica(1)	BCLH01001441	49959-50150	50062-50107 (45)	-	This work
Zoysia pacifica(2)	BCLH01000647	31753-31994 RC	31753-31793 RC (80)	-	This work

Table 2: Complete table with all currently known documented ENOD40 in Poaceae.

For many species duplicate homologues have been found. It is not fully clear yet which homologues are paralogues or orthologues (or combinations of both). The amount of homologues found per species is not consistent; While for some species there has only a single homologue found, other species have more than 2 variants, with *Dendrocalamus latiflorus* having the most at 10. Moreover, the location of homologues shows an interesting pattern. For species with accessions of which chromosome notation is available, it stands out that homologues are mostly found on different chromosomes, such as with *Alloteropsis semialata*. More sparsely, some species have two homologues on the same chromosome like *Triticum dicoccoides*. Nearly all polyploid species for which ENOD40 was found have homologues on different chromosomes both in number and variant such as *Digitaria exilis*. Some polyploid species have only variants of the same chromosome with homologues or have only one homologue found. It is not sure if these exceptions do not abide to the pattern of the other polyploid species or that the homologues just have not been sequenced.

3.2 Structure consensus

An attempt was made for all ENOD40 homologues in Poaceae to form a family-wide consensus secondary RNA structure of domain 2 using a secondary RNA structure prediction algorithm.

Because the conservation of domain 2 in this families' ENOD40 is so low, this has proven difficult. Instead of one singular consensus, groups of different types of consensuses have been formed, based on common sequences found in the 3' bottom of the hairpin structure along with some unique cases which did not fit any grouping. A new structural motif for the bottom of the hairpin has also been found and separately documented.
This attempt of grouping the domains 2 based on the predicted RNA structures does not result in a coherent concept. These results are to be warily interpreted as they are a result of *in silico* predictions, and have no support of phylogenetic or experimental research.
The consensus groups created are identified by the variation in the 3' CAAAC part of the motif, since this has the most consistent variation appearances across multiple species, usually in the form of an insertion of 2 contiguous nucleotides. In Table 3 the groups for the species in the PACMAD clade are listed. In Table 4 the groups for the species in the BOP clade are listed. Both tables also show the sequences that define each group.

The recognizable semi-conserved GUUUG/CAAAC sequence is not (fully) part of the hairpin structure any more in more than half of the species due to substantial variation in domain 2. Because of this variation, the hairpin does not always form with the same motif at the bottom, while the GUUUG/CAAAC sequence is usually still present in some variation.
There is no standard for comparing RNA structures and consensuses which is usable in this context, therefore in the tables there is also a 'match' defined based on observation. A 'perfect' match means that the group defining sequence is fully intact and fully part of the bottom of the secondary structure. A 'slight difference' match means that the sequence slightly deviates from the group defining sequence, but is fully part of the bottom of the secondary structure. 'Partially in structure' and 'not in structure' are more self-evident, describing that the group defining sequence is fully intact, but not (fully) part of the secondary structure.

In earlier research, it was noted that domain 2 in ENOD40 of species in the order of Poales did not contain the GUUUG/CAAAC motif at the bottom of the hairpin structure [GR07]. The structure results further prove that this is the case for nearly all Poaceae; Only 3 homologues, *Oryza branchyantha 1*, *Oryza rufipogon* and *Oryza sativa 2*, have this motif still intact. All other species have some kind of variation on this motif. Later research has shown that there are more conserved motifs to be found in Poaceae ENOD40 [Kos12]. The predicted structures corroborate this.

Interestingly, the two more newly discovered motifs, UGGUGCCUU/GAGGCGUGCA in the stem and CC/GG closing the loop at the end of the stem, are found in nearly all homologue structure predictions. As expected, the stem motif occurs in variants, such as the frequent UGGUGCCUU/GAGGUACGCA. The loop closing motif however is really strongly conserved, but it occurs as GG/CC instead of CC/GG in group 4 and 5. In species where the bottom of the hairpin motif GUUUG/CAAAC is distorted, it seems to take priority in deciding how the secondary structure is folded. Another interesting observation comparing Table 3 with Table 2 is that the same groups in which the GG/CC is present, contains domains 2 in which the hairpin is longer (∼70nt instead of ∼50nt).

PACMAD domain 2 consensus groups		
Group & Sequence	Species	Match
Group 1 CAAAC	Cs1, Hd1, Sh1, Zmay1	Perfect
	Ei1, Eco1	Not in structure
Group 2 CAUAAAC	De3, De4	Perfect
	Ca, Pm1,3,4, Pmc2,4,5,6, Si2, Sv2, Svc2, Urc2, Uri2	Partially in structure
	Pv4, Hd3	Not in structure
Group 3 CACAAAC	Sb3, So, Ss6,7,8,9	Perfect
	Ecg1,2, As3	Partially in structure
Group 4 CAGAAAC	Ph1, Pmc1, Si1, Sv1, Svc1, Urc1, Uri1, Cs2	Perfect
	Zj2, Zmat2, Zp2	Slight difference
	Pm2, Pmc3, Pv1,2	Partially in structure
Group 5 CAGAGAC	As1,2, Do2, Hd2, Ml1, Ms2, Sb1, Sh2, Ss1,2,3,4,5, Zmay2	Perfect
	Tt1	Slight difference
	Ecg3, Eco2, Ecu1, Ei2, En2, Et3, De1,2	Partially in structure

Table 3: The different consensus groups devised for the secondary RNA structure of domain 2 of ENOD40 in the PACMAD clade of the Poaceae family. Species abbreviations are derived from Table 2. See Figure 3 for examples of structure predictions.

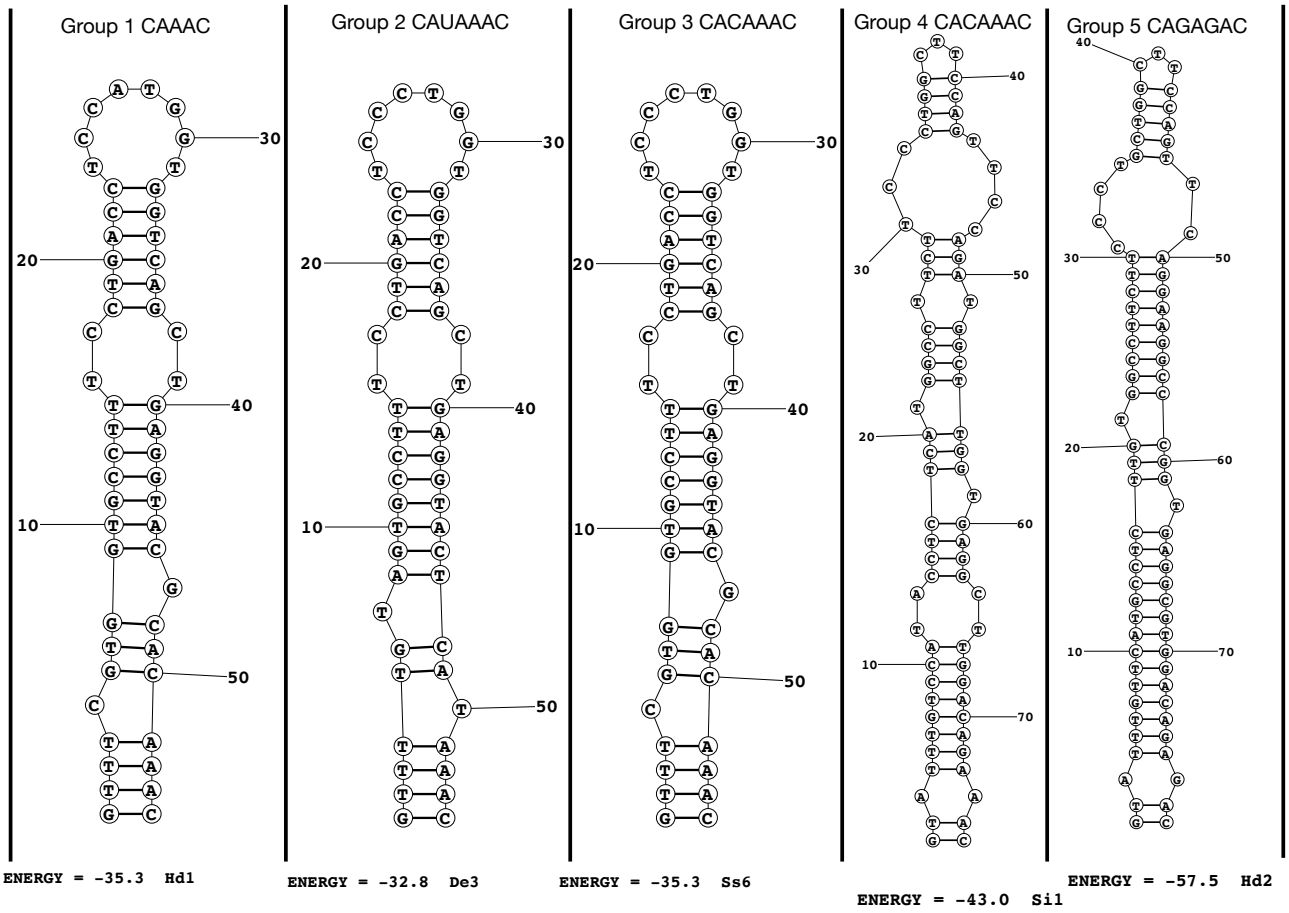

Figure 3: Structure prediction examples of PACMAD ENOD40 domain 2 groups. See Table 3.

BOP domain 2 consensus groups		
Group & Sequence	Species	Match
Group 1 CAAAC	Ob1, Or, Os2	Perfect
	Os1	Not in structure
Group 2 CACGAAC	Fp, Lp	Perfect
	Bd	Partially in structure
	Ab	Slightly different
Group 3 CAAGAAU	At, Ta, Td1,2,3, Tm	Partially in structure
	Hv, Hvc	Partially in structure & slightly different
Other	Ab, Fe1	

Table 4: The different consensus groups devised for the secondary RNA structure of domain 2 of ENOD40 in the BOP clade of the Poaceae family. Species abbreviations are derived from Table 2. See Figure 4 for examples of structure predictions.

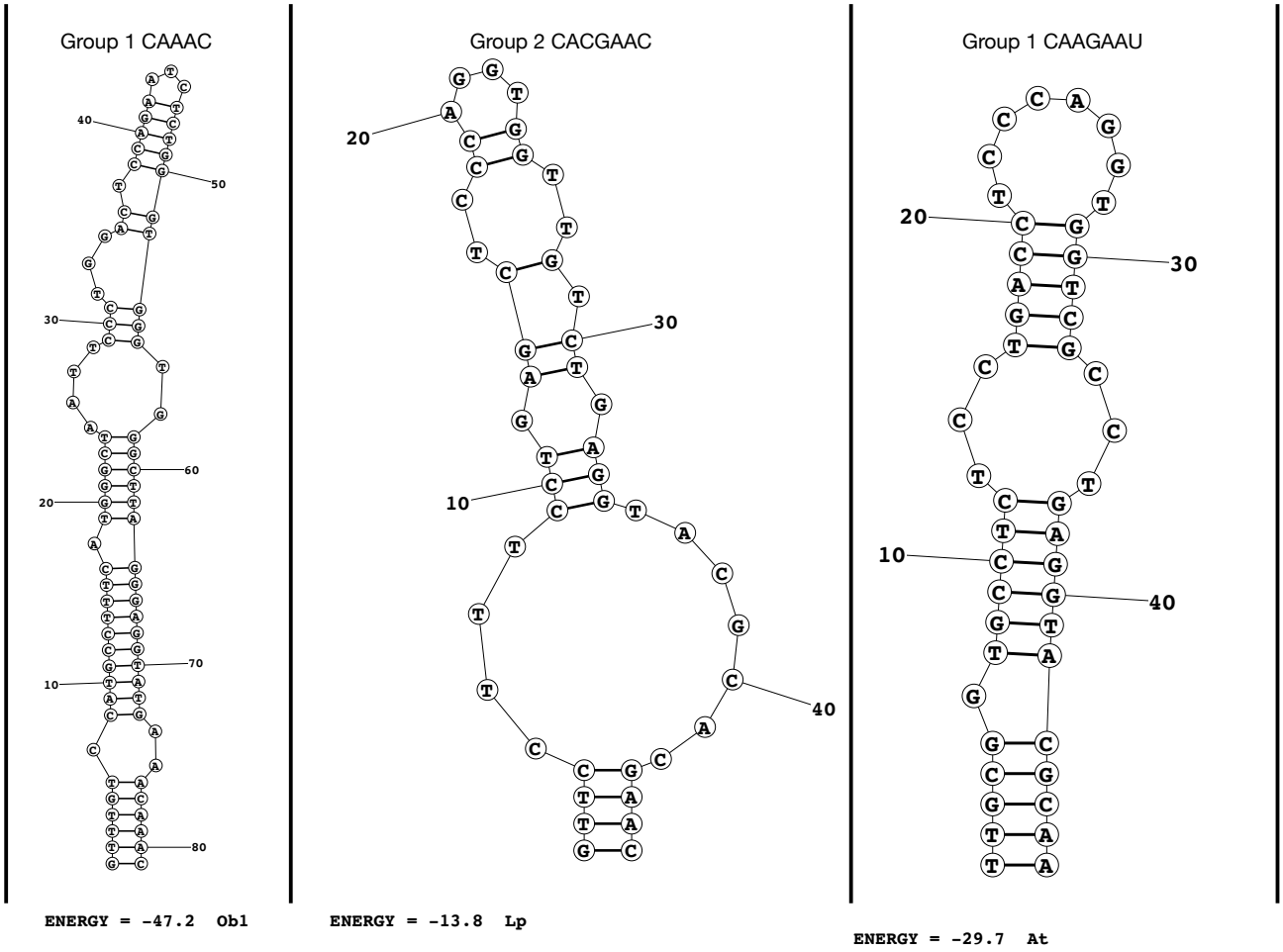

Figure 4: Structure prediction examples of BOP ENOD40 domain 2 groups. See Table 4.

There are two homologues from species which do not belong in any group, displayed in Table 5. The first, *Sorghum bicolor 2*, contains only 6 nucleotides of the normally highly conserved region II. Between this short leftover and sORF1 only patches resembling the stem motif UGGUGC-CUU/GAGGCGUGCA are present. While 2 potential hairpin folds with the same sequence have been predicted, they have little similarity to other secondary structures. The sequence which forms these predicted hairpin folds is nonetheless marked as domain 2. Secondly there are *Panicum virgatum* and *Festuca arundinacea*, for which no sensible secondary structure prediction could be found, in *Panicum virgatum* partially because of the short sequence between sORF1 and region II.

Unique cases	
Species	Reason
Sb2	Only half of region II present, arbitrary range for domain 2
Pv3, Fa	No sensible secondary structure predicted

Table 5: Unique cases for the secondary RNA structure of domain 2 of ENOD40 in the PACMAD clade of the Poaceae family. Species abbreviations are derived from Table 2.

Interestingly, a new consistent structural motif at the bottom of the hairpin has been observed. For certain species in which the GUUUG/CAAAC motif was distorted the RNA secondary structure prediction algorithm predicted that a shorter sequence would form the hairpin instead, with a GUG/CAC motif at the bottom of the hairpin. The species containing this motif and minor variations on this motif are listed in Table 6, regardless if there is one of the previously defined group identifying sequences present in the ENOD40 gene. Note that both the BOP clade and PACMAD clade find representation in this table. Looking at other domains 2, the GUG/CAC motif appears in a number of homologues and it seems that when the sequence further towards the bottom of the hairpin is distorted and not able to form stable Watson-Crick base pairings, it is the next stable motif.

GUG/CAC domain 2 consensus groups	
BOP clade	
Variant	Species
GUG/CAC	Ob2
GUG/UAC	Og, Ol
PACMAD clade	
Variant	Species
GUG/CAC	Tt2, Zj1, Zmat1, Zp1, Do1, Ecu2
UGUG/CACA	En1, Et1,2
GUG/UAC	Ml3,4, Ms1

Table 6: The GTG/CAC consensus groups devised for the secondary RNA structure of domain 2 of ENOD40 in the Poaceae family. Species abbreviations are derived from Table 2. See Figure 5 for examples of structure predictions.

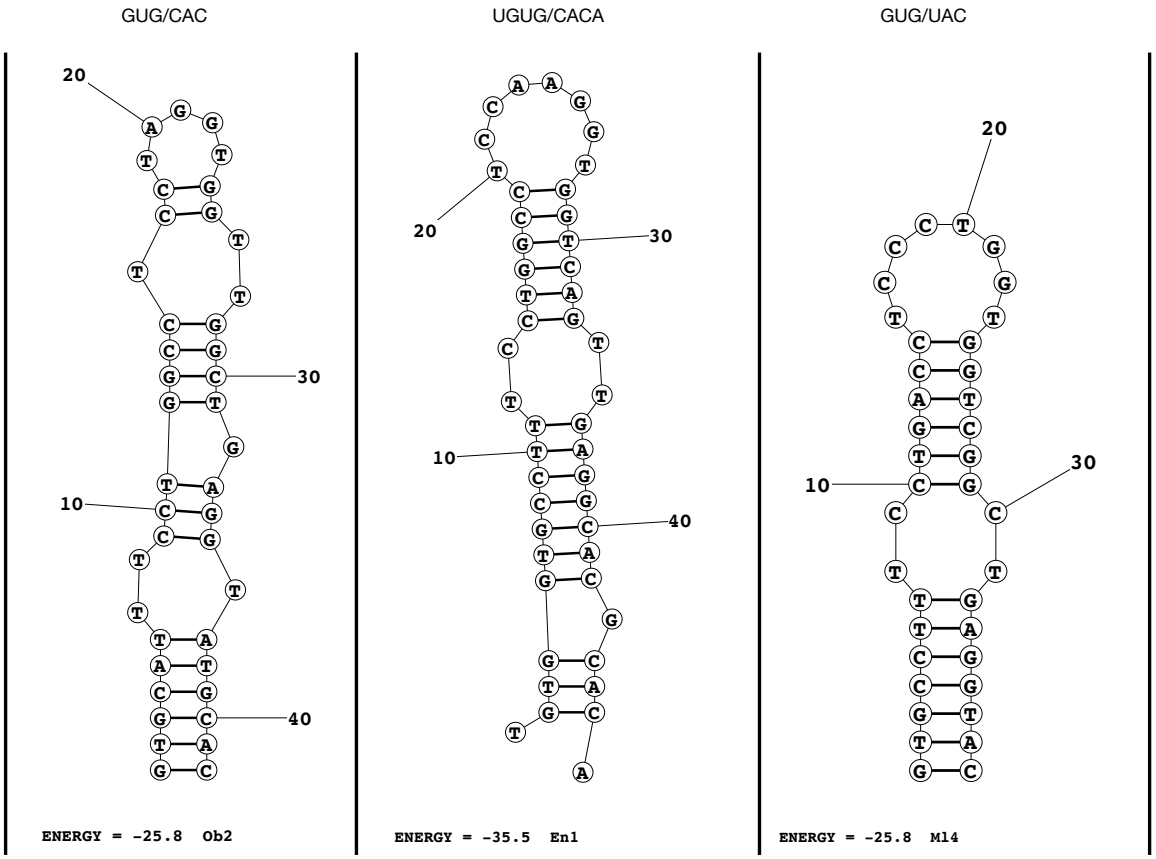

Figure 5: Structure prediction examples of GUG/CAC ENOD40 domain 2 groups. See Table 6.

4 Discussion

When interpreting the two sets of results presented, different levels of confidence should be bestowed to each set. The set of genome database accessions with putative ENOD40 genes is quite promising and stable, while the domain 2 structure prediction grouping of these accessions is less reliable and open for debate.

The set of all 121 known putative ENOD40 genes in Poaceae presented offers ample opportunity for further research, particularly for identifying in which order duplication and speciation happened across the clades and tribes. Earlier research has already provided a suggestion for this order for *Zea mays* [CRA+03], but this could be extended for many more species. The suggestion given was that duplication and mutation occurred before speciation based on homology levels between homologues within *Zea mays* and between these homologues and their counterparts in *Oryza sativa*. This does not explain the large number of homologues found for certain species in the set of newly found ENOD40 however, so potentially it is the case that duplication after speciation also plays a role.

The groups created for the domain 2 secondary RNA structure prediction attempt to provide scaffolding for further research, but are most likely not fully accurate or final. For the general groups presented in Table 3 and 4, there are doubts surrounding the *in silico* predictions they are based on and the factor of human interpretation of these predictions. These groups primarily show that the CAAAC motif found in other families is still present and relatively conserved to the rest of domain 2, having little variance. The GUUUG motif is less conserved with more erratic variance. While present, the CAAAC motif and its variants are not always (fully) part of the secondary hairpin structure anymore, caused by more intense variation upstream. The other two more recently identified motifs are present
The secondary RNA structure of domain 2 in Poaceae shows very little consensus across all species, but when grouped, comparisons of some patterns between the groups can be made. For instance, groups 4 & 5 in the PACMAD grouping have clearly larger domains 2 and a difference in the strongly conserved CC/GG loop ending. It might be that the change in size caused the other two phenomena, but this can only be confirmed or denied by experimental research. Moreover, phylogenetic work could be done with the groups as a basis. Considering the differences between groups and the fact that species have different homologues in different groups could provide further insight into the process of duplication and speciation.

As noted in Section 2.3, the process included isolating the hairpin structure, removing unnecessary nucleotides. During this process it was noticed that for some species a larger domain 2 could be possible by inclusion of a part of the beginning of region II. As seen in Figure 6, this does not change the rest of the structure, but adds a few nucleotide pairings at the bottom of the stem. In earlier research, this portion of region II has been included in some domains 2 [Kos12]. In this work however, it is assumed that region II is *not* part of the hairpin structure of domain II. If it can be confirmed that region II can actually be part of domain 2, this would mean that the grouping presented here is too strict and that some species might have larger domains 2 than identified here.
For one of the three unique structure cases (Table 5) there is a distinct difference between previous research and this work. The secondary RNA structure prediction algorithm used for this work did not produce a viable structure for the domain 2 of *Festuca arundinacea*, while previous research has [Kos12]. This could be caused by a difference in prediction algorithm used, since the structure predicted in the previous research seems to be more lenient with mismatches which cause large loops with significantly more nucleotides in one strand than the other.

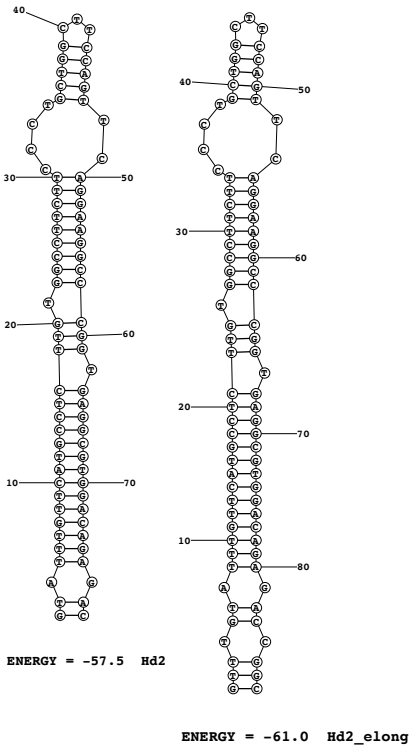

Figure 6: Example of *Hyparrhenia diplandra* domain 2 secondary RNA structure with and without region II inclusion. Structure exluding region II is left, structure including region II on the right.

5 Conclusion

In this thesis a set of new and updated Poaceae ENOD40 homologues has been introduced. As expected there is an assortment of duplicate homologues found, with some species having more than 2, which was the previous upper limit. For the homolgue set presented in this thesis, a combination of previously known Poaceae ENOD40 homologues and newly found Poaceae ENOD40 homologues, secondary RNA structure consensus groups have been formed for the highly variable domain 2.
In order to find the Poaceae ENOD40 homologues, BLAST was used to query species selected with assistance from the NCBI taxonomy browser with queries consisting of previously found Poaceae ENOD40. The secondary structure consensus groups were compiled based on predictions by a secondary RNA structure prediction algorithm.
The nature of the homologues found and the volatility of domain 2 both invite for a future evolutionary insight. For domain 2, it might be of interest to understand in which order which changes have happened, with the constructed consensus groups functioning as a rough indication. The homologues found consist largely of duplicate homologues, which leads to questions about when the duplication of the homologues took place during evolution and if there is a reason for the large disparity between the amount of homologues per species.

Acknowledgements Thanks to my supervisor Dr. A.P. Gultyaev for providing assistance and guidance. Thanks to all who have lived with me during the process for picking up the responsibilities I had no time for. Thanks to Niels van Belle for being an amazing partner.

References

[AGM+90] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[BRSM13] Stanislav Bellaousov, Jessica S Reuter, Matthew G Seetin, and David H Mathews. Rnastructure: web servers for rna secondary structure prediction and analysis. *Nucleic acids research*, 41(W1):W471–W474, 2013.

[CB16] Maarten JM Christenhusz and James W Byng. The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3):201–217, 2016.

[CRA+03] Bert Compaan, Tom Ruttink, Cathy Albrecht, Robert Meeley, Ton Bisseling, and Henk Franssen. Identification and characterization of a zea mays line carrying a transposon-tagged enod40. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1629(1-3):84–91, 2003.

[Edd99] Sean R Eddy. Noncoding rna genes. *Current opinion in genetics & development*, 9(6):695–699, 1999.

[Fed12] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.

[Gna09] Samuel S Gnanamanickam. Rice and its importance to human life. In *Biological control of rice diseases*, pages 1–11. Springer, 2009.

[GR07] Alexander P. Gultyaev and Andreas Roussis. Identification of conserved secondary structures and expansion segments in enod40 rnas reveals new enod40 homologues in plants. *Nucleic acids research*, 35(9):3144–3152, 2007.

[GRG+03] Geneviève Girard, Andreas Roussis, Alexander P Gultyaev, Cornelis WA Pleij, and Herman P Spaink. Structural motifs in the rna encoded by the early nodulation gene enod40 of soybean. *Nucleic acids research*, 31(17):5003–5015, 2003.

[Hod18] Trevor R Hodkinson. Evolution and taxonomy of the grasses (poaceae): a model family for the study of species-rich groups. *Annual Plant Reviews Online*, pages 255–294, 2018.

[II12] Grass Phylogeny Working Group II. New grass phylogeny resolves deep evolutionary relationships and discovers c4 origins. *New Phytologist*, 193(2):304–312, 2012.

[KAC+18] Justyna Koc, Piotr Androsiuk, Katarzyna Joanna Chwedorzewska, Marely Cuba-Díaz, Ryszard Górecki, and Irena Giełwanowska. Range-wide pattern of genetic variation in colobanthus quitensis. *Polar Biology*, 41(12):2467–2479, 2018.

[KMNC12] Ilene Karsch-Mizrachi, Yasukazu Nakamura, and Guy Cochrane. The international nucleotide sequence database collaboration. *Nucleic acids research*, 40(D1):D33–D37, 2012.

[Kos12] Céline Koster. Search for enod40 homologous genes in plants and further exploration within the poales family. *Honours College, University of Leiden*, 2012.

[KTS+99] Hiroshi Kouchi, Ken-ichi Takane, Rollando B So, Jagdish K Ladha, and Pallavolu M Reddy. Rice enod40: isolation and expression analysis in rice and transgenic soybean root nodules. *The Plant Journal*, 18(2):121–129, 1999.

[Lar03] Knud Larsen. Molecular cloning and characterization of a cdna encoding a ryegrass (lolium perenne) enod40 homologue. *Journal of plant physiology*, 160(6):675–687, 2003.

[RSM+02] Horst Röhrig, Jürgen Schmidt, Edvins Miklashevichs, Jeff Schell, and Michael John. Soybean enod40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences*, 99(4):1915–1920, 2002.

[Rut03] Tom Ruttink. *ENOD40 affects phytohormone cross-talk*. 2003.

[SEOK96] Gregory D Schuler, Jonathan A Epstein, Hitomi Ohkawa, and Jonathan A Kans. Entrez: Molecular biology database and retrieval system. *Methods in enzymology*, 266:141–162, 1996.

[SJC+01] Carolina Sousa, Christina Johansson, Celine Charon, Hamid Manyani, Christof Sautter, Adam Kondorosi, and Martin Crespi. Translational and structural requirements of the early nodulin gene enod40, a short-open reading frame-containing rna, for elicitation of a cell-specific growth response in the alfalfa root cortex. *Molecular and Cellular Biology*, 21(1):354–366, 2001.

[SvdWZ+89] Ben Scheres, Clemens van de Wiel, Andrei Zalensky, Ann Hirsch, Ab Van Kammen, and Ton Bisseling. Identification of rhizobium leguminosarum genes and signal compounds involved in the induction of early nodulin gene expression. In *Signal molecules in plants and plant-microbe interactions*, pages 367–377. Springer, 1989.

[VRFG+97] P Van Rhijn, Y Fang, S Galili, O Shaul, N Atzmon, S Wininger, Y Eshed, M Lum, Y Li, V To, et al. Expression of early nodulin genes in alfalfa mycorrhizae indicates that signal transduction pathways used in forming arbuscular mycorrhizae and rhizobium-induced nodules may be conserved. *Proceedings of the National Academy of Sciences*, 94(10):5467–5472, 1997.

[Wat90] L Watson. The grass family, poaceae. *Reproductive versatility in the grasses*, pages 1–31, 1990.

[YKH+93] Wei-Cai Yang, Panagiotis Katinakis, Peter Hendriks, Arie Smolders, Floris de Vries, Johan Spee, Ab van Kammen, Ton Bisseling, and Henk Franssen. Characterization of gmenod40, a gene showing novel patterns of cell-specific expression during soybean nodule development. *The Plant Journal*, 3(4):573–585, 1993.