
Universiteit
Leiden
The Netherlands

Structural motifs in the RNA encoded by the early nodulation gene enod40 of soybean

Girard, G.A.O.; Roussis, A.; Gultyaev, A.P.; Pleij, C.W.A.; Spaink, H.P.

Citation
Girard, G. A. O., Roussis, A., Gultyaev, A. P., Pleij, C. W. A., & Spaink, H. P. (2003). Structural motifs in the RNA encoded by the early nodulation gene enod40 of soybean. *Nucleic Acids Research*, *31*(17), 5003-5015. doi:10.1093/nar/gkg721

Version:	Publisher's Version
License:	Licensed under Article 25fa Copyright Act/Law (Amendment Taverne)
Downloaded from:	https://hdl.handle.net/1887/3630081

Note: To cite this publication please use the final published version (if applicable).

Nucleic Acids Research, 2003, Vol. 31, No. 17 ***5003–5015***
DOI: 10.1093/nar/gkg721

Structural motifs in the RNA encoded by the early nodulation gene *enod40* of soybean

Geneviève Girard[1], Andreas Roussis[1], Alexander P. Gultyaev[1,2], Cornelis W. A. Pleij[2] and Herman P. Spaink[1,*]

[1]Institute of Biology, Leiden University, Wassenaarseweg 64, 2333 AL Leiden, The Netherlands and [2]Institute of Chemistry, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands

Received May 23, 2003; Revised June 30, 2003; Accepted July 17, 2003

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

ABSTRACT

The plant gene *enod40* is highly conserved among legumes and also present in various non-legume species. It is presumed to play a central regulatory role in the *Rhizobium*–legume interaction, being expressed well before the initiation of cortical cell divisions resulting in nodule formation. Two small peptides encoded by *enod40* mRNA as well as its secondary structure have been shown to be key elements in the signalling processes underlying nodule organogenesis. Here results concerning the secondary structure of mRNA of *enod40* in soybean are presented. This study combined a theoretical approach, involving structure prediction and comparison, as well as structure probing. Our study indicates five conserved domains in *enod40* mRNA among numerous leguminous species. Structure comparison suggests that some domains are also conserved in non-leguminous species and that an additional domain exists that was found only in leguminous species developing indeterminate nodules. Enzymatic and chemical probing data support the structure for three of the domains, and partially for the remaining two. The rest of the molecule appears to be less structured. Some of the domains include motifs, such as U-containing internal loops and bulges, which seem to be conserved. Therefore, they might be involved in the regulatory role of *enod40* RNA.

INTRODUCTION

The highly specific plant–prokaryote symbiosis between bacterial genera of the *Rhizobiaceae* family (*Rhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Bradyrhizobium* and *Azorhizobium*), also called rhizobia, and leguminous plants (1) induces on the plant root new organs, called nodules. The nodule essentially contains compartmentalised bacteria surrounded by a plant membrane and provides an ideal environment for the microsymbionts to perform the reduction of dinitrogen into ammonia, utilising in exchange plant-derived carbon sources.

The initiation and development of the nodule structure as a symbiotic organ is controlled at different levels. Highly species-specific signalling involves (iso)flavonoids from the plant side and a whole repertoire of β-1–4-linked-*N*-acetyl glucosamine compounds carrying different structural modifications secreted by the bacteria (known also as Nod factors, NFs), and these are the key molecules establishing the earliest stages of the interaction (2). NF perception in epidermal root hairs by a single or double component receptor (3,4) is associated with downstream Ca^{+} signalling (5,6) and possibly small trimeric GTPases along with phospholipase C and phosphoinosides (7,8). Phytohormone-mediated secondary signalling involving auxin, cytokinin and ethylene (9–11) is also suggested to be of major importance in nodule organogenesis and positioning.

A number of plant genes have been identified so far, which are co-ordinately induced or expressed in an enhanced manner in the nodules, and their products (nodulins) are regulated both temporally and spatially (2,12). Several results indicate that one of them, *enod40*, has a regulatory function in nodule initiation. *enod40* is highly conserved among leguminous plants and it has also been identified in the non-legumes tobacco, rice and maize (13,14). In the legume roots, it is expressed in the early stages of the interaction, specifically in the pericycle adjacent to the protoxylem pole and well before the initiation of cortical cell divisions that lead to nodule formation (15,16). Furthermore, it is induced by NFs and chitin pentamers (17,18) and the phytohormone cytokinin (19). Overexpression of *enod40* in *Medicago truncatula* results in a considerable increase of cortical cell divisions when plants are subjected to nitrogen-limiting conditions (20). Interestingly, such plants infected by *Sinorhizobium meliloti* nodulate faster and exhibit increased sensitivity to the Nod signals upon treatment with purified Nod factors or inoculation with *S.meliloti* mutants (21). In mature nodules, *enod40* is expressed in the pericycle of the vascular bundles (16,22) but also in several non-symbiotic tissues (11,16,22–24). Studies on rice *enod40* suggest that legume and non-legume *enod40* genes may share some common functions in differentiation and/or functions of vascular bundles, where rice *enod40* is mostly accumulated (14).

*To whom correspondence should be addressed. Tel: +31 71 527 5076; Fax: +31 71 527 5088; Email: spaink@rulbim.leidenuniv.nl
Present address:
Andreas Roussis, Center for Human and Clinical Genetics, Leiden University Medical Center, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands

The lack of a common, long open reading frame (ORF) in the otherwise conserved *enod40* genes has always been an intriguing aspect. Since the first isolation of the soybean *enod40* gene (16,22), several studies have focused on its expression pattern during nodule development in different legumes and in response to different stimuli, but also on the functional analysis of the two peptides encoded by two short ORFs (sORFI and sORFII). The polycistronic nature of *enod40* mRNA was recently demonstrated biochemically, with *in vitro* translation in wheat germ extracts resulting in the *de novo* synthesis of two peptides of 12 and 24 amino acid residues, subsequently immunoprecipitated by antipeptide antibodies (25). Additionally, these peptides were shown to specifically bind sucrose synthase (SuSy), thus implicating them in the regulation of SuSy activity and/or its intracellular targeting (25). Sousa *et al.* (26) showed that in alfalfa, the biological activity of *enod40* is directly related to the proper translation of sORFI and sORFII. The work of Crespi *et al.* (27) and Sousa *et al.* (26) suggests a possible regulatory role for RNA structure, indicated by an estimated low free energy of folding in some regions of *enod40*. Indeed, some local secondary structure elements in *enod40* were predicted by comparative analysis (26,28).

In order to contribute to our understanding of the molecular mechanism(s) of *enod40* action, we set out to study in detail the secondary structure of the soybean *enod40-1* RNA by means of chemical and enzymatic probing. The data from this analysis have been combined with computer-assisted studies, thus allowing for an ameliorated RNA structure prediction. Our studies reveal a number of particular characteristics of the soybean *enod40* RNA, such as, for example, the high amount of uridines in bulges and loops. Also for the first time, experimental data are provided for a strongly supported secondary structure element (hairpin 1) in the inter-sORF region.

MATERIALS AND METHODS

Prediction of RNA secondary structure

RNA secondary structures were predicted by folding simulations using a genetic algorithm (29), implemented in the package STAR (http://wwwbio.leidenuniv.nl/~batenburg/STAR.html). The thermodynamic parameters used were from version 2.3 of Turner *et al.* (http://www.bioinfo.rpi.edu/~zukerm/rna/energy/). Simulations were performed at various lowered (10–25°C) temperatures in order to mimic the conditions of the natural environment for the plant RNAs. For the initial search of conserved hairpins, the option for simulation of the folding during transcription was used. In order to refine the phylogenetically conserved structure, the conserved structures were 'forced' into the folding simulations of full-length molecules, so that at the next round only predictions containing these structures could be produced. Possible alternative foldings in the interiors of forced domains were also verified using predictions of optimal and suboptimal structures by the mfold package (30,31). Probing data were taken into account by forcing some regions of RNA to be single stranded in the simulations.

The following *enod40* sequences were used for the predictions: *Glycine max* (accession no. X69154), *Phaseolus vulgaris* (X86441), *Vigna radiata* (AF061818), *Sesbania rostrata* (Y12714), *Lotus japonicus* (AJ271787, AJ271788), *Medicago truncatula* (X80264), *Medicago sativa* (L32806, X80263), *Trifolium repens* (AJ000268), *Pisum sativum* (X81064), *Vicia sativa* (X83683), *Lupinus luteus* (AF352375), *Nicotiana tabacum* (X98716), *Oryza sativa* (AB024054), *Oryza branchyantha* (AB024055), *Hordeum vulgare* (AF542513) and *Lolium perenne* (AF538350). The numbering used on the figures starts at the 5′ end of the RNA molecules or follows the numbering of the corresponding database (*L.japonicus*).

RNA synthesis

DNA encoding the *G.max enod40-1* gene was amplified by PCR from a pBluescript-derived expression vector. This DNA was cloned under the control of the T7 promoter into a plasmid (pMP4800), and the expected sequence was verified. This plasmid was called pMP4000. For the enzymatic and chemical probing, *Gmenod40-1* was amplified by PCR from pMP4000 using the primers T7trans (5′ GGGCTAATACGACTCACTATAGGC 3′) containing the T7 promoter and *enod40*rev (5′ GAAAAGGACTCTGGAAACTTTTG 3′).

For technical reasons, the probing gels could not be read around the 3′ end. In order to study in more detail the structure of the 3′ end of *Gmenod40-1* RNA, probing of 3′-end-labelled *enod40* RNA was tried. PCR was done on pMP4000 using T7trans and Gm403′-pExtRNA (5′ GAAAGGACTCTGGAAACTTTTCTTTTTTTTTT 3′), so that the RNA would end with a short poly(A) tail, which would make the labelling of the 3′ end easier. However, structure probing on this 3′-end-labelled *enod40* RNA failed.

Transcription on the DNA fragments was performed using the RiboMax kit (Promega). The RNA was isolated by 2-fold phenol extraction and a single chloroform extraction followed by isopropanol precipitation.

Enzymatic structure probing

RNA (1 pmol) was renatured at room temperature for 5 min in a total volume of 6 µl containing 50 mM sodium cacodylate pH 7.5 and 10 mM $MgCl_2$. Reactions were performed in this mix (10 µl final volume) for 5 min at room temperature with 3 µg of bulk tRNA and 0.001, 0.0001 or 0.00001 U of RNase T1 (Sankyo), 0.01, 0.005 or 0.001 U of RNase V1 (Pharmacia Biotech), or 0.01, 0.001, or 0.0001 U of RNase T2 (Life Technologies). Subsequently, reactions were stopped by chilling on ice and adding 20 µl of buffer containing 0.4 M sodium acetate pH 5.2, 2 mM EDTA and 0.2 µg/µl carrier tRNA. The enzyme-cleaved RNA was subjected to phenol–chloroform extraction, followed by ethanol precipitation prior to primer extension analysis.

Chemical modification

RNA (1 pmol) was incubated at room temperature in 6 µl of 50 mM Na borate pH 8.0 and 10 mM $MgCl_2$. After addition of 3 µg of bulk tRNA, chemical modifications were performed in a total volume of 10 µl containing 12.6 µg (reaction time of 5 min) or 126 µg (reaction time of 5 or 10 min) of CMCT

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

[1-cyclohexyl-3-(2-morpholino-4-ethyl) carbodiimide methotosylate] (Lancer) or 3.36 mg (reaction time of 6, 12 or 18 min) of DEPC (diethyl pyrocarbonate) (Sigma). Reactions were stopped and RNA purified following the method used for enzyme-cleaved RNA.

Primer extension

Primer1 (5′ GAAAGGACTCTGGAAACTTT 3′, complementary to region 631–651 of *Gmenod40-1* RNA), primer 2 (5′ ACACAAACAAGCATGGAAAA 3′, complementary to region 523–543 of *Gmenod40-1* RNA), primer 3 (5′ TGAGCACTACATAGCCATAG 3′, complementary to region 313–333 of *Gmenod40-1* RNA) or primer 4 (5′ GTGAGGAGTGAGCACCCTCT 3′, complementary to region 109–130 of *Gmenod40-1* RNA) was 5′ end labelled with [γ-^{32}P]ATP using T4 polynucleotide kinase (Pharmacia).

Primer hybridisation was performed in a 4 µl reaction volume containing 0.2 pmol of enzyme-cleaved or chemically modified RNA, 1 pmol of labelled oligonucleotide, 50 mM Tris–HCl, 60 mM KCl and 10 mM dithiothreitol (DTT). Hybridisation was done by heating to 50°C and cooling down at 37°C for 15 min. Primer extension was carried out by adding 1 µl of 'reverse transcription mixture'. Reverse transcription mixture contains 2 U of enhanced AMV reverse transcriptase (Promega) in 50 mM Tris–HCl, 30 mM KCl, 10 mM DTT and 15 mM $MgCl_2$.

For the DNA sequencing reaction on pMP4000, the T7 sequencing kit (Pharmacia) was used in combination with the primers mentioned above, and samples were analysed according to standard protocols.

RESULTS

Prediction of phylogenetically conserved structures in *enod40* RNAs

The initial predictions of RNA secondary structures were produced by the iterative process of integrating folding simulations with sequence comparisons. At the first step, the structures of *enod40* RNAs from legume plants were predicted by the option of a genetic algorithm that simulated the folding during RNA transcription (see Materials and Methods) and was especially effective in searching for the most likely fast folding hairpins initiating structure formation. Using a combination of published (partial) alignments (14,19,27) with these predictions, the hairpins conserved in all sequences were identified. These hairpins were used in the input for the next predictions as 'forced' structures (see Materials and Methods) in order to locate conserved more distant pairings. The conserved long-range interactions were used to refine the alignment and as an input for the next round of predictions and so on until no conserved structures could be predicted.

Such an iterative process identified six conserved domains, one of them being folded only in part of the *enod40* RNAs containing an insertion of ~75–130 nucleotides, absent in *G.max* and some other sequences (Fig. 1E). The first stem–loop structure (Fig. 1A), located 7–13 nucleotides downstream of sORF1, is a hairpin of variable length, containing several internal loops. In all sequences, the 5′ half of the hairpin is purine-rich, whereas the 3′ part is pyrimidine-rich, resulting in possible 'flipping' of pairs. This structure, conserved in non-legume *enod40* RNAs, has also been predicted by Sousa *et al.* (26) and Hofacker *et al.* (28).

At the distance of 4–13 nucleotides downstream of hairpin 1, a much extended conserved stem–loop structure 2 was predicted (Fig. 1D). This domain, spanning 123–135 nucleotides (in *G.max*, positions 148–272), has a very peculiar feature: despite some structural variation in the interior loops, all structures contain multiple loops containing exclusively U residues. Such loops are represented by both bulges and symmetric or asymmetric interior loops. The top part of the structure demonstrates much variety, both in sequence and in possible alternative foldings. The existence of this structured domain 2 is also supported by the following observation: in *L.japonicus*, in spite of an insertion in the *enod40-2* sequence compared with *enod40-1*, the same terminal pairings are predicted. The 3′ terminus of domain 2 is located at the 5′ part of the so-called conserved region II (32).

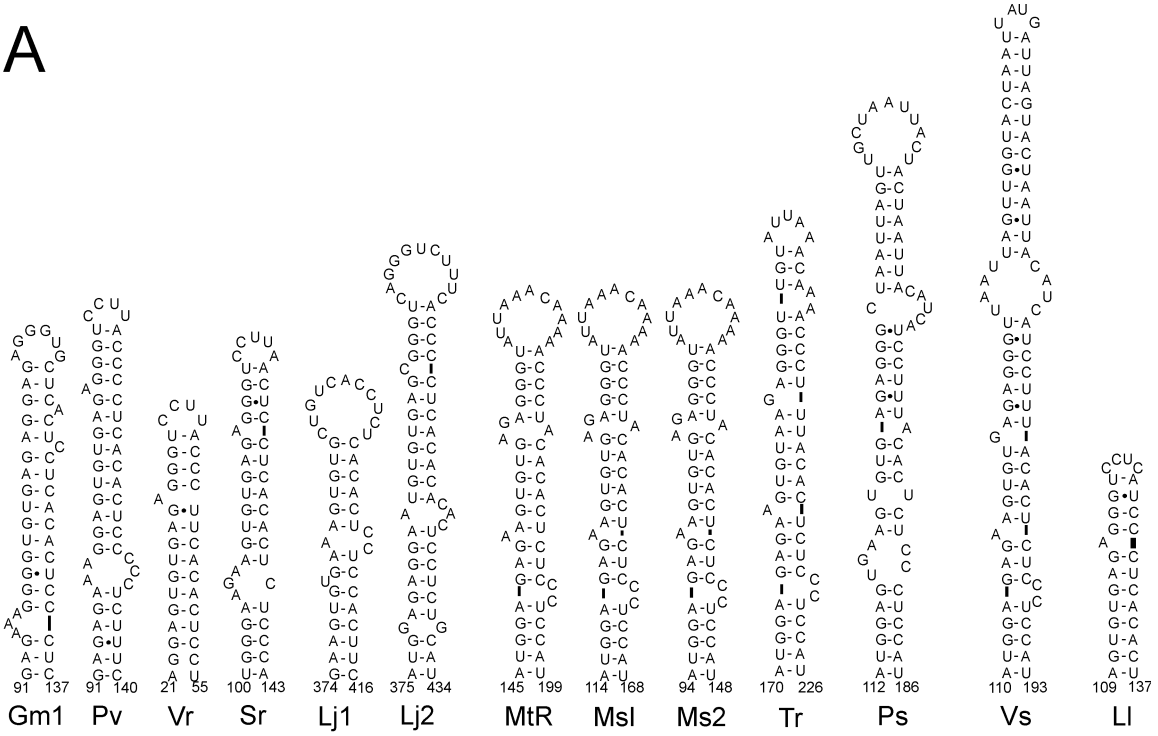
The 3′-proximal part of this region II is folded into a small hairpin 3 (Fig. 1B). It seems that the stem–loop structure 2 and hairpin 3 could also be traced as the peaks in the 'mountain plots' produced for a set of *enod40* sequences by the program RNAalifold for folding the aligned RNAs (28).

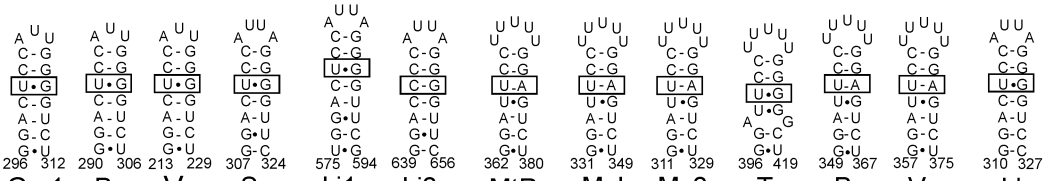
The *enod40* sequences downstream of the conserved region II are very variable in both sequence and length. In *M.truncatula*, *M.sativa* and *T.repens* sequences, an extended stem–loop structure can be predicted just downstream of the hairpin 3 (Fig. 1E). With a remarkable similarity to domain 2, this stem–loop structure (domain 4) is also characterised by a conserved pattern of multiple U-containing loops and bulges, while the top hairpin has a less conserved structure, with possible alternative foldings. Figure 1E shows the most conserved hairpin at the top, which corresponds to either the lowest free energy configuration or the second (suboptimal) folding of this domain, as predicted by the mfold package. Indirect support for the existence of such a stem–loop structure in these species could be derived from the fact that in some other *enod40* RNAs, including that of *G.max*, almost the whole structure is deleted, suggesting that such a large deletion (~5–130 nucleotides) could leave the remaining structure unperturbed. Thus this structure is absent in the *enod40* RNA of *G.max*.

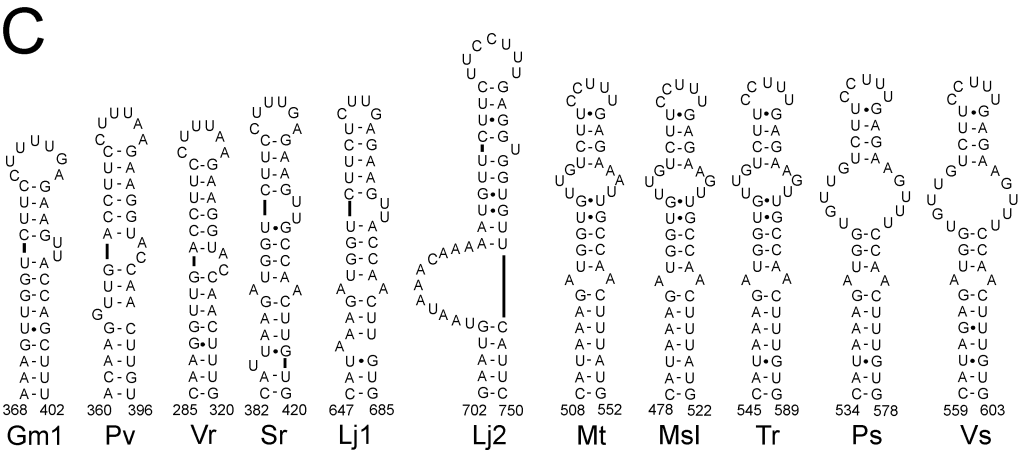
Further downstream, two conserved stem–loop structures can be predicted in all legume *enod40* sequences (structures 5 and 6, Fig. 1C and F, respectively). Despite some structural variation in the interiors of these hairpins due to substitutions and deletions, the terminal stems seem to be much conserved.

For all the domains, it is difficult to strictly define the nucleotide covariations. It is remarkable, however, that the predicted stem–loop conformations are conserved despite considerable sequence variability. In domains where stems are formed by conserved sequences, nucleotide substitutions and/or deletions cluster around internal loops and bulges, preserving an overall topology of alternating stems and loops, therefore also being indirect support for the proposed structural elements. In the small hairpin 3, formed by one of the most conserved *enod40* sequence regions, the substitutions and deletions occur mostly in the hairpin loop (Fig. 1B). One of the base pairs is also variable, being either U·G, U–A or C–G, which can be the result of single mutations preserving the pairing.

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023







Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

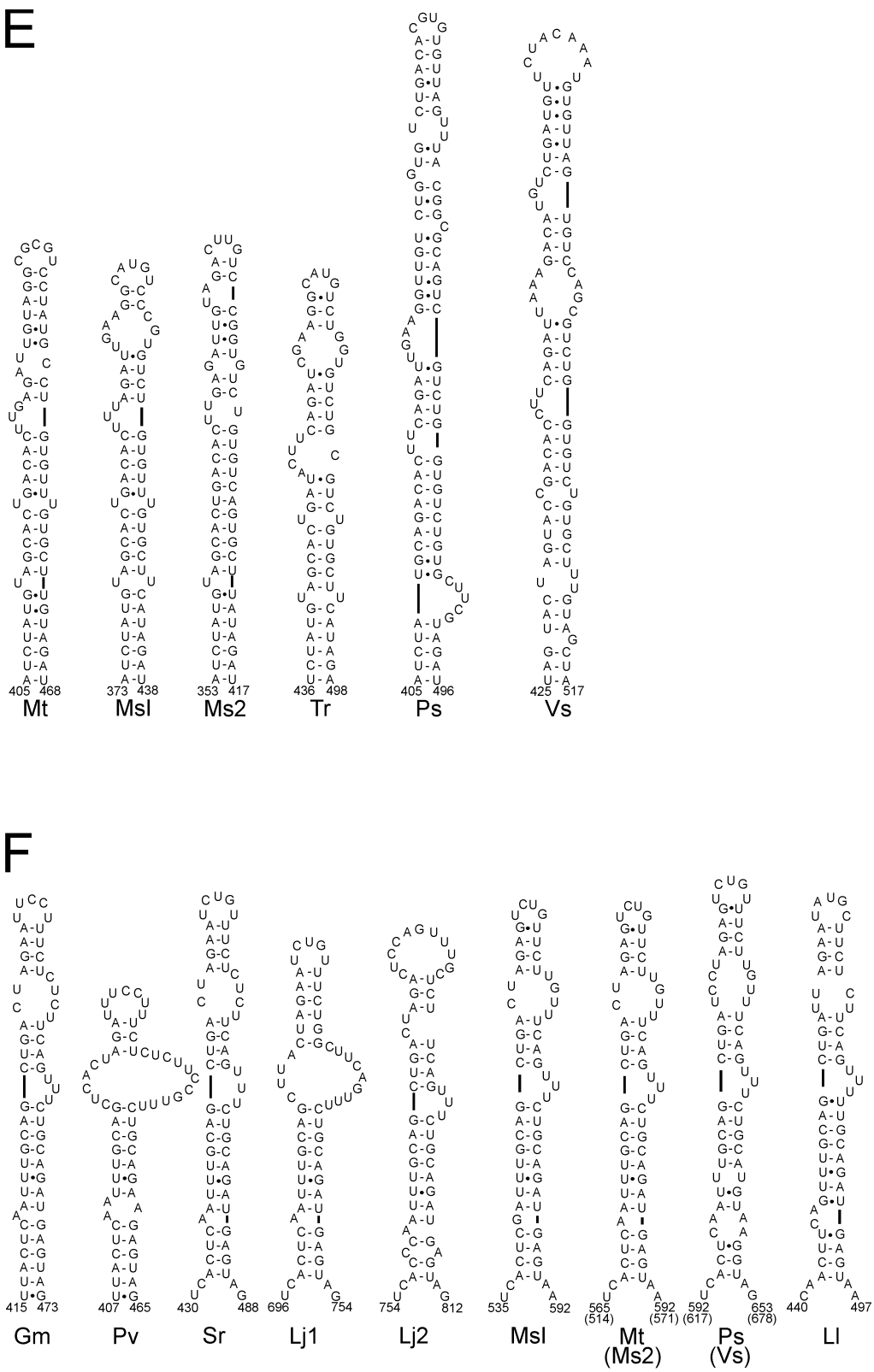


Figure 1. (Previous two pages and above) Phylogenetically predicted domains for *enod40* mRNA of several leguminous plant species abbreviated as follows: Gm1 (*Glycine max enod 40-1*), Pv (*Phaseolus vulgaris*), Vr (*Vigna radiata*), Sr (*Sesbania rostrata*), Lj1 (*Lotus japonicus enod40-1*), Lj2 (*L.japonicus enod40-2*), Mt (*Medicago truncatula*), MsI [*Medicago sativa* (cultivar Iroquois) *enod40*], Ms2 (*M.sativa enod40*), Tr (*Trifolium repens*), Ps (*Pisum sativum*), Vs (*Vicia sativa*) and Ll (*Lupinus luteus*). (**A**) Domain 1; (**B**) domain 3; (**C**) domain 5; (**D**) domain 2; (**E**) domain 4; (**F**) domain 6.

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

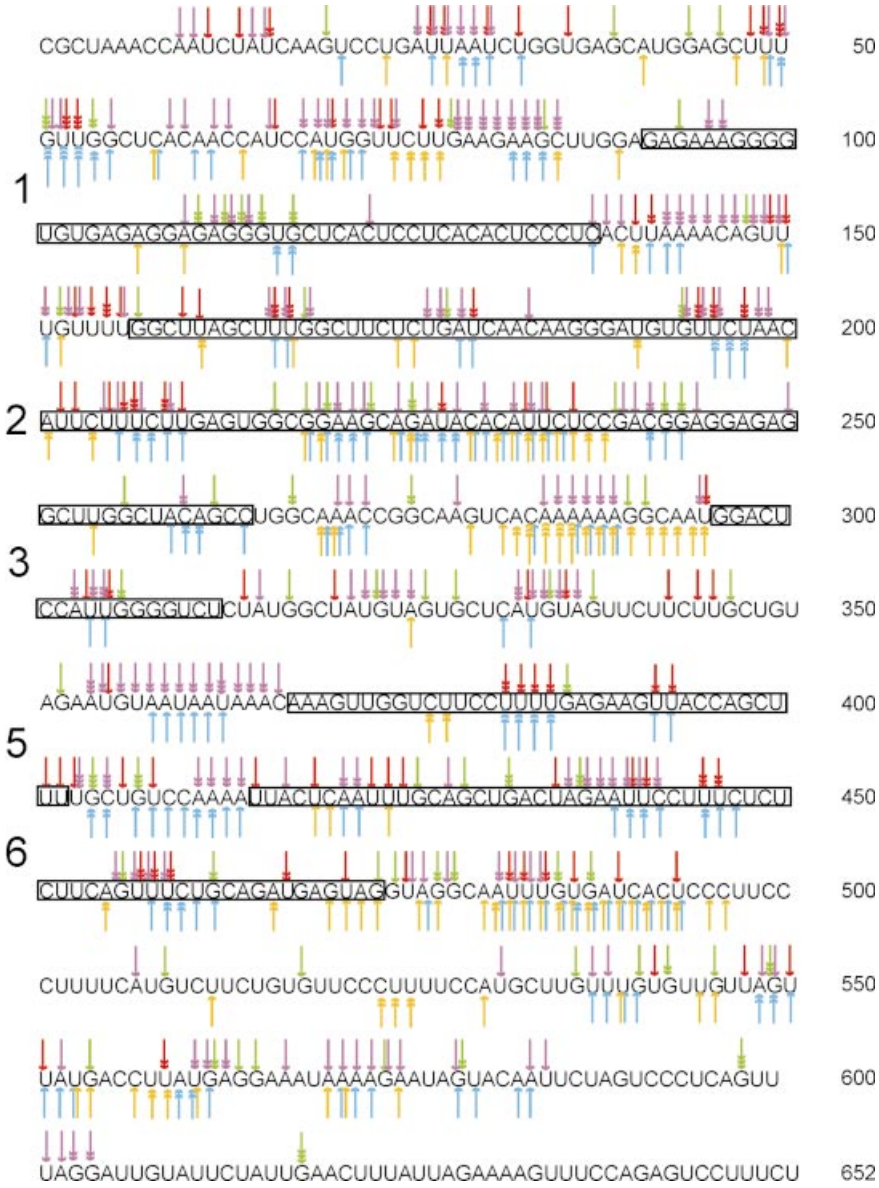


Figure 2. Results of the probing experiments represented on the linear sequence of *Gmenod40-1*. The reactivity of nucleotides to different probes is indicated by arrows, with one, two or three heads for weak, medium or strong reactivity, respectively. The nature of the probes is represented by the colour of the arrows: T1 in green, T2 in blue, V1 in yellow, CMCT in red and DEPC in purple. The black boxes indicate the different domains (predicted by phylogeny), and their corresponding numbers are on the left margin. The RNA sequence is numbered every 50 nucleotides on the right margin.

Structural probing of the *Gmenod40* RNA

The proposed secondary structure of *Gmenod40* RNA provided by the STAR program and by sequence comparison represents a rough indication of the folding of the RNA. However, experimental data were needed to support and refine our model. Therefore, structure probing was performed, using both enzymatic digestion and chemical modification in combination with primer extension.

First, DNA fragments containing the T7 promoter upstream of the gene *Gmenod40* were used as template to synthesise *Gmenod40* RNA. The RNA was treated by enzymatic digestion or chemical modification followed by primer extension using various radioactively labelled primers. Each region of the predicted structure was re-evaluated by the combination of the experimental data obtained with RNase T1, T2 and V1 digestions, and with CMCT and DEPC modifications. Most of these agents react in single-stranded regions. In principle, RNase V1 cuts in double-stranded regions, but results must be interpreted with care, since many double-stranded regions may react only weakly or not at all with RNase V1. The global results of probing are shown in Figure 2.

Domains 1, 3 and 5. The loops at the top of the three predicted hairpins of domains 1, 3 and 5 are strongly supported by probing data (Fig. 3A, C and D, respectively): the nucleotides present in these loops are highly reactive towards RNase T1

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

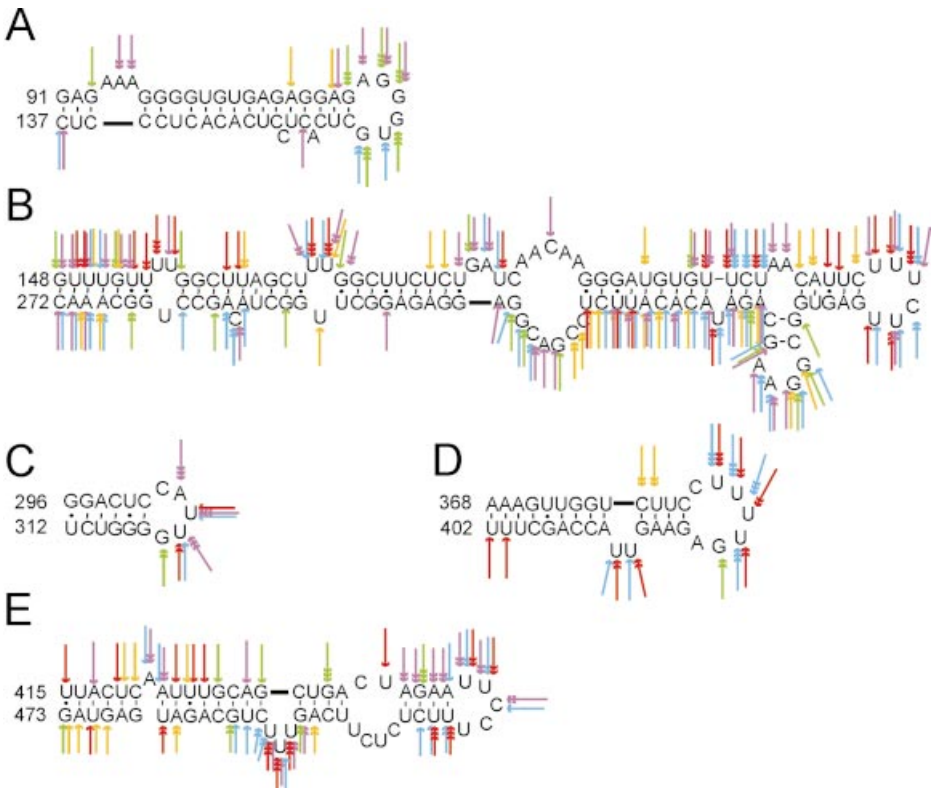


Figure 3. Results of the probing shown on the phylogenetically predicted domains. Concerning the probing indications, see legend to Figure 2. (**A**) Domain 1; (**B**) domain 2; (**C**) domain 3; (**D**) domain 5; (**E**) domain 6.

and T2, CMCT and DEPC. Furthermore, most of the bulges are also confirmed, e.g. DEPC strongly modifies A95 and A96 in hairpin 1, while CMCT strongly modifies U391 and U392 in hairpin 5. Some stem structures are also supported by the low reactivity of T1, T2, CMCT and DEPC in the corresponding regions or just by digestion by V1 (stem of hairpin 5). Only the very bottom of the stem in domain 1 is slightly more reactive than expected to CMCT, but this can be explained by the weakness of the G·U and A–U bonds present in that region. Taken together, the hairpins of domains 1, 3 and 5 are quite well supported by experimental data. The few discrepancies (some unexpected cuts) are of weak intensity.

Domains 2 and 6. The phylogenetic studies predict a long hairpin in domain 2 of *Gmenod40* RNA. The very top of the hairpin is not well defined by theoretical analysis which offers different alternatives: one consists of a branched structure with two small hairpins, supported by computer prediction, while the second one shows a straight hairpin (not shown), homologous to those from, for example, *P.vulgaris* or *V.radiata*. Although the latter seems more supported by phylogeny, the probing data (Fig. 3B) seem to confirm the branched structure. The probing validates very well the loop at the top of the structure, as in the hairpins of domains 1, 3 and 5. It also does so for some bulges, particularly for 166UU167, 177GA178, 198AA199 and 260C. The tetra-loop 218GGAA221 is also well confirmed by the probing, in addition to the fact that it is a stable GNRA tetra-loop. The bottom half of the predicted hairpin in domain 2 also looks reliable, with expected low reactivity to the probes (region 157–176 paired with 264–244) except in the bulges (166UU167, 260C). Only the very bottom is more reactive than expected. The median regions of domain 2 fit the probing less well. This is particularly the case for the region 181AACAA185, which is expected to be highly reactive to most of the probes and which is not. On the contrary, the region downstream of the stable tetra-loop reacts well with the probes, although it is predicted to be part of a stem. The probing results, however, are not very straightforward, since, in this region, V1 also reacts quite well, which would tend to confirm the stem structure. These mixed results can be interpreted as a secondary structure of low stability, not completely closed and probably containing non-canonical base pairs in the internal loops.

According to phylogeny, domain 6 also folds into a relatively long hairpin. Analysis of the probing (Fig. 3E) confirms the bottom stem of the structure, with only very weak cuts of the probes in most of the sequence, except on or close to the two bulges 421A and 457UUU459. Probing results are more contradictory to phylogeny in the top part of the structure. Indeed there is no clear pattern indicating an alternation of bulges and stem structures as expected. Also, although the probing clearly indicates two unpaired regions (436AGAAUUC442 and 445UU446) alternating with two non-reactive regions (443CU444 and 447CUCUCUCA455), we were unable to fold the region 436–455 in a structure that perfectly fits these observations. This situation is very comparable with the hairpin present in domain 2, involving

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

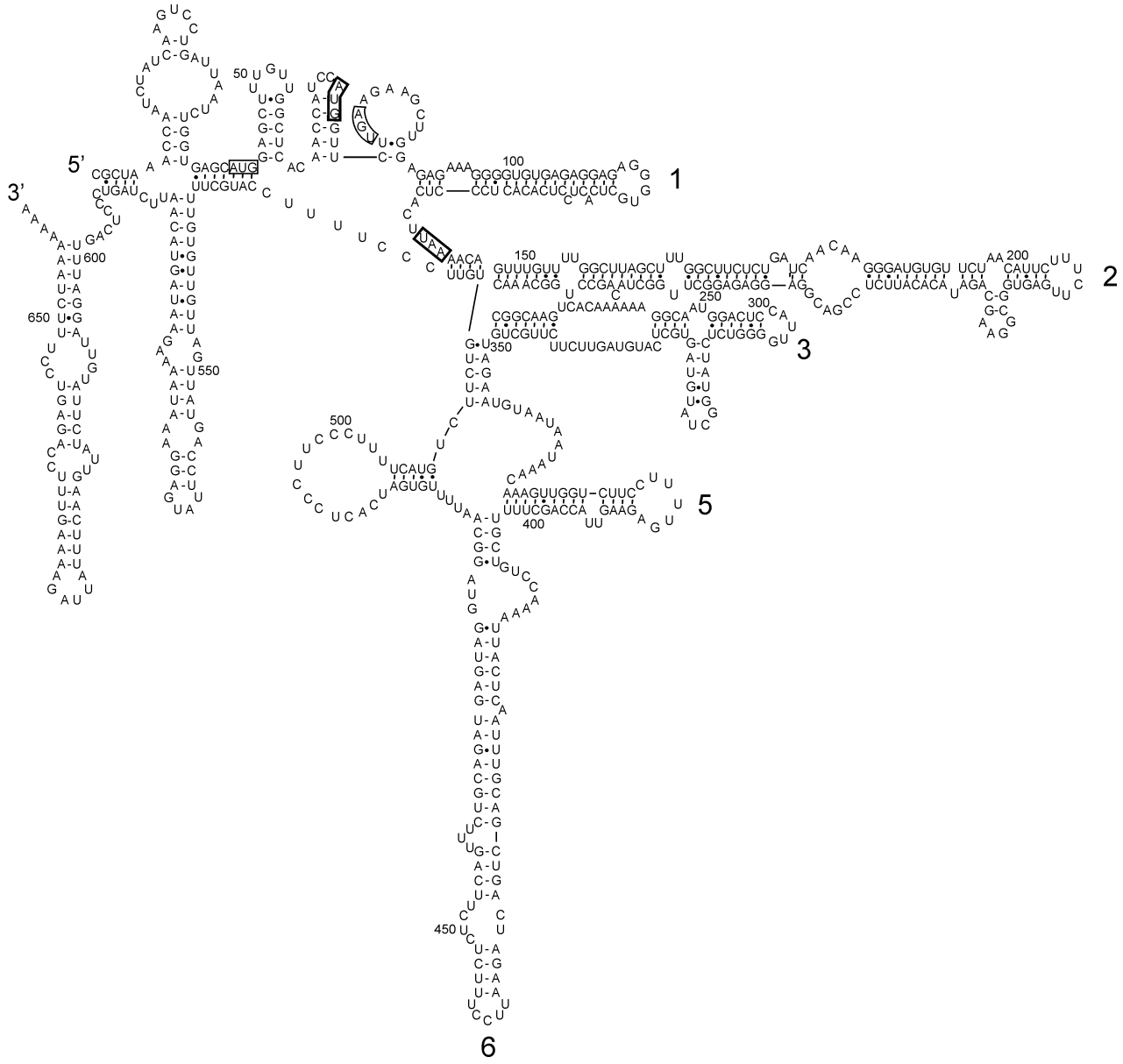


Figure 4. Proposed model for global folding of *Gmenod40-1*. The thick numbers from 1 to 6 indicate the corresponding different domains characterised by phylogenetic studies and probing. The RNA sequence is numbered every 50 nucleotides by the thin numbers. The boxed nucleotides indicate the start and stop codons for ORF A (thin boxes) and ORF B (thick boxes) discussed by Röhrig *et al.* (25).

a relatively strong stem bottom predicted by phylogeny and confirmed by probing, and a weaker top region predicted by phylogeny, partially contradicted by probing, but difficult to fold in a more optimal way.

Domain 4. Domain 4 is actually almost completely deleted in the *enod40* RNA of *G. max*. The probing does not support any particular structure in this region.

Global folding prediction

In order to produce a consistent model of *Gmenod40* RNA, phylogenetically conserved structures and probing data were combined as an input to more refined folding simulations, as described in Materials and Methods. Some regions were selected for their high reactivity to at least two different probes among T1, T2, CMCT and DEPC, which makes them therefore reliable candidates as single-stranded regions. In this way, the regions at positions 26–31, 78–84, 332–337 and 355–366 were forced as single-stranded in a new simulation. However, the folding simulations, taking into account the single-strandedness of these regions and the formation of conserved domains 1–5, still did not produce a single prediction of RNA secondary structure. Thus, with the described constraints, three runs of the genetic algorithm simulations yielded three different predictions with very close values of free energy (25°C): –232.9, –232.8 and –230.6 kcal/mol, suggesting the absence of a well-defined global structure. The structure with the lowest free energy (–232.9 kcal/mol) is shown in Figure 4.

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

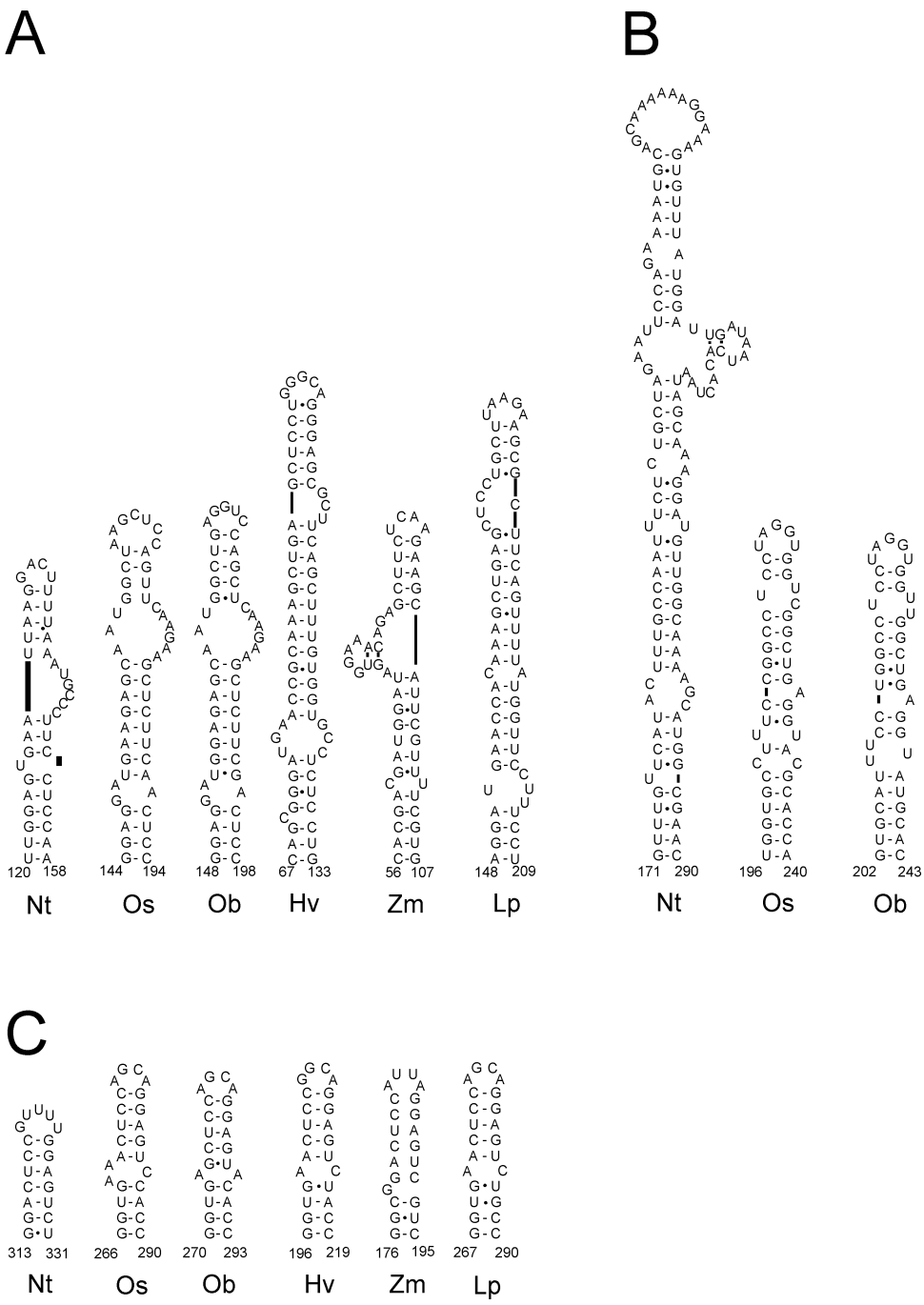


Figure 5. Phylogenetically predicted domains for *enod40* mRNA of several non-leguminous plant species abbreviated as follows: Nt (*Nicotiana tabacum*), Os (*Oryza sativa*), Ob (*Oryza branchyantha*), Hv (*Hordeum vulgare*) and Lp (*Lolium perenne*). (**A**) Domain 1; (**B**) domain 2; (**C**) domain 3.

DISCUSSION

The presented probing data, together with computer-assisted predictions and structural comparisons, indicate the presence of several structured regions in the soybean *enod40* RNA. Five domains of the *G.max enod40* RNA secondary structure are also conserved in other known leguminous *enod40* sequences (Fig. 1). The conservation of the domains points to their possible role in *enod40* RNA functioning. This is also consistent with the data on *Medicago enod40* RNA, which show that deletion of the structure located in the non-translated region, corresponding to domain 1 (Fig. 1A), impairs biological activity (26).

Domain 1, represented by a stem–loop structure, is strongly conserved in all leguminous plants (Fig. 1A) and also in non-legumes (Fig. 5A), and is particularly well supported by probing. This structure is very stable, and its presence has already been suggested on the basis of a comparison of folding free energies in the comparable region of related *enod40* RNAs with statistically expected values (26), and detected by

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

a program searching for conserved structures in related RNAs (28). In all sequences, the stem–loop structure is characterised by a purine-rich 5′ half, whereas the 3′ half is pyrimidine-rich, leading to possible alternative pairings in stem interiors.

Domain 2 is also represented by a stem–loop structure, but more extended in comparison with domain 1. Not well defined by predictions and phylogeny, the structure of the top of the hairpin can be better understood from the probing results. Remarkably, the interior (which is the part of domain 2 best supported by probing) exhibits an intriguing conserved feature, namely multiple internal loops and bulges consisting of U residues (Fig. 3B). Interestingly, only the locations of these loops are conserved (Fig. 1D), whereas their dimensions and symmetry vary due to deletions and/or insertions on both sides of the loops. Thus, the UU/U loop of the *G.max* structure (positions 166–167/254) has the same 2 × 1 symmetry at *S.rostrata*, *P.sativum* and *V.sativa* homologous positions, but 2 × 2 topology is found at equivalent position in *P.vulgaris* structure, 1 × 1 in *V.radiata*, 3 × 2 in *L.japonicus* and 4 × 3 in all three available *Medicago* sequences (Fig. 1D). Two U-containing loops, separated by a single G–C pair can be suggested in the homologous position of *L.luteus enod40-1* RNA structure, one of them being a UUU-bulge, while another is a symmetric UU/UU 2 × 2 loop (Fig. 1D). The same diversity is observed for other loops, still consisting of uridines only.

This conserved feature could be connected to the biological activity of *enod40* RNA. A similar pattern is observed in the structure folded in the region representing a rather large insertion in some legume sequences (domain 4), such as *Medicago*, *T.repens*, *P.sativum* and *V.sativa*. Again, a global topology of an extended stem–loop structure with duplexes separated by U-containing loops is preserved, while the dimensions of these loops are variable (Fig. 1E). Intriguingly, these sequences represent the species with indeterminate nodules. Indeterminate nodules are different in many ways from determinate nodules that are formed by species such as *G.max*, *P.vulgaris*, *V.radiata* and *L.japonicus*. The difference is most obviously noted in their morphology and ontological development: determinate nodules are round and lack a persistent meristem, whereas indeterminate nodules are long and continuously grow due to an apical meristem (1). If our hypothesis about the functional role of the *enod40* RNA U-containing loops is true, it is tempting to speculate that the presence of an additional stem–loop structure in the species with indeterminate nodules plays some role in defining the nodulation type.

Conserved patterns in internal loops, bulges and non-Watson–Crick base pairs are frequently recognition motifs in specific protein–RNA or RNA–RNA interactions, because, in contrast to regular RNA A-form helix, they can expose a variety of configurations of donor and acceptor groups, creating specific binding sites (33–35). As far as U-containing loops are concerned, one of the best studied examples of recognition motifs are U-rich bulges in HIV-1 and BIV TAR RNA hairpins, recognised by viral Tat proteins [reviewed by Hermann and Patel (35)]. The U·U base pairs in RNA helices are relatively stable (36) and can belong to different isostericity families (37). Crystallographic studies demonstrate that internal loops containing uracils can form consecutive non-canonical U·U base pairs with unique geometries of potential recognition sites (38,39).

Due to the diversity of the U-containing loops in domains 2 and 4 in *enod40* RNAs (Fig. 1D and E), it is difficult to derive a straightforward definition of a specific motif on the basis of available data. It is known that asymmetry and adjacent (non-canonical) base pairs are very important factors affecting both geometry and thermodynamics of internal loops (40,41). In addition to diversity of symmetric properties, described above, the *enod40* U-containing loops are also flanked by different base pairs. On the other hand, possible formation of U·U pairs within larger loops may constrain them, so that, for instance, 2 × 1 or 3 × 2 U-rich loops can be considered as a bulged U with adjacent U·U pair(s). Moreover, U·U pairs were shown to stabilise RNA internal 2 × 1 loops (42). Such a stabilising U·U pair in a UU/U loop is observed in the crystal structure of the large ribosomal subunit RNA from *Haloarcula marismortui* (40,43).

Another conserved structure in *enod40* RNA is the small hairpin 3, located at the 3′-proximal part of the conserved region II (Fig. 1B). Interestingly, this hairpin is also strongly conserved in non-legume sequences. Domains 5 and 6 seem to be conserved in all legume *enod40* RNAs, but we failed to recognise their equivalents in non-legume sequences.

The structure of the rest of the molecule is apparently less defined; what is seen are somewhat contradictory results from probing by various chemicals and enzymes. Sequence comparisons of available sequences also do not yield a consensus model for the secondary structure in these regions. This is consistent with our folding simulations predicting different structures outside the conserved domains 1–5 with very close free energies even when the simulations are constrained by single-strandedness of some sequences as evidenced by probing. Apparently, such structures can be characterised as poorly defined and are less likely to be functional (44,45).

Some of the described domains are conserved in non-legume *enod40* RNA molecules (Fig. 5). Domains 1 and 3 can be predicted in all known sequences. However, domain 2 is apparently less conserved. While in *N.tabacum* this domain folds very similarly to legume molecules, in both rice sequences a possible equivalent stem–loop structure is significantly shorter. Its existence in *H.vulgare*, *L.perenne* and *Z.mays* molecules is doubtful, because only relatively weak structures (not shown), if any, are possible, but they are not predicted by the folding algorithm and are not reliably supported by sequence comparisons (not shown).

To summarise, one part of the *enod40* RNA seems to be folded in a number of well-defined domains, while the rest of the RNA is poorly structured. Domains 1 and 3 are common to legumes and non-legumes, as well as domain 2 to a lesser extent. Domains 5 and 6 seem to exist only in leguminous plants, at least in examples known so far. Among leguminous plants from which *enod40* was sequenced, domain 4 appears well structured in plants known to produce indeterminate nodules, whereas it is completely deleted in plants with determinate nodulation, including *G.max*.

Well-defined structures of RNA are more likely to be functional. In order to understand further the regulating role of *enod40* RNA, the structured domains 1–6 are therefore the most promising to analyse, particularly domains 2 and 4 with their U-containing loops. For instance, these loops can serve as

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

specific recognition motifs for binding of proteins. It is interesting to note that deletion of some of the structured part of alfalfa *enod40* decreases its activity, while expression of two small ORFs is not affected (26), suggesting that the RNA structure is important for some interactions rather than for regulating translation. Furthermore, the expression of a 3′-proximal region of *Medicago enod40* RNA, lacking the peptide-coding sequence, was shown to elicit cortical cell division in transgenic plants at a frequency similar to that observed with complete *enod40* sequence (20). Thus the *enod40* RNA seems to have a rather unique combination of features typical of structural RNA and protein-encoding mRNA, because functional secondary structures in mRNAs are usually only involved in regulation of gene expression, while structural RNAs involved in other processes are mostly represented by non-coding molecules. Since there are several published bioassays for *enod40* function (20,26,27,46), it would be of interest to test in these assays *enod40* deletion mutants or recombinants based on domain swapping that are designed on the basis of our structural analysis.

ACKNOWLEDGEMENT

We thank Dr Nina Tsareva for technical expertise and helpful discussions.

REFERENCES

1. Hadri,A.E., Spaink,H.P., Bisseling,T. and Brewin,N.J. (1998) Diversity of root nodulation and rhizobial infection processes. In Spaink,H.P., Kondorosi,A. and Hooykaas,P.J.J. (eds), *The Rhizobiaceae*. Kluwer Academic Publishers, Dordrecht, pp. 347–360.
2. Bladergroen,M.R. and Spaink,H.P. (1998) Genes and signal molecules involved in the rhizobia–leguminoseae symbiosis. *Curr. Opin. Plant Biol.*, **1**, 353–359.
3. Ardourel,M., Demont,N., Debelle,F., Maillet,F., de Billy,F., Prome,J.C., Denarie,J. and Truchet,G. (1994) *Rhizobium meliloti* lipooligosaccharide nodulation factors: different structural requirements for bacterial entry into target root hair cells and induction of plant symbiotic developmental responses. *Plant Cell*, **6**, 1357–1374.
4. Schultze,M. and Kondorosi,A. (1998) Regulation of symbiotic root nodule development. *Annu. Rev. Genet.*, **32**, 33–57.
5. Felle,H.H., Kondorosi,E., Kondorosi,A. and Schultze,M. (1998) The role of ion fluxes in Nod factor signaling in *Medicago sativa*. *Plant J.*, **13**, 455–463.
6. Ehrhardt,D.W., Atkinson,E.M. and Long,S.R. (1992) Depolarization of alfalfa root hair membrane potential by *Rhizobium meliloti* Nod factors. *Science*, **256**, 998–1000.
7. Pingret,J.L., Journet,E.P. and Barker,D.G. (1998) *Rhizobium* nod factor signaling. Evidence for a G protein-mediated transduction mechanism. *Plant Cell*, **10**, 659–672.
8. Engstrom,E.M., Ehrhardt,D.W., Mitra,R.M. and Long,S.R. (2002) Pharmacological analysis of nod factor-induced calcium spiking in *Medicago truncatula*. Evidence for the requirement of type IIA calcium pumps and phosphoinositide signaling. *Plant Physiol.*, **128**, 1390–1401.
9. Penmetsa,R.V. and Cook,D.R. (1997) A legume ethylene-insensitive mutant hyperinfected by its rhizobial symbiont. *Science*, **275**, 527–530.
10. Mathesius,U., Schlaman,H.R., Spaink,H.P., Sautter,C., Rolfe,B.G. and Djordjevic,M.A. (1998) Auxin transport inhibition precedes root nodule formation in white clover roots and is regulated by flavonoids and derivatives of chitin oligosaccharides. *Plant J.*, 23–24.
11. Fang,Y. and Hirsch,A.M. (1998) Studying early nodulin gene *enod40* expression and induction by nodulation factor and cytokinin in transgenic alfalfa. *Plant Physiol.*, **116**, 53–68.
12. Crespi,M. and Galvez,S. (2000) Molecular mechanisms in root nodule development. *J. Plant Growth Regul.*, **19**, 155–166.
13. Matvienko,M., van de Sande,K., Pawlowski,K., Van Kammen,A., Bisseling,T. and Franssen,H. (1996) *Nicotiana tabacum* SR1 contains two *enod40* homologues. In Stacey,G., Mullin,B. and Gresshoff,P.M. (eds), *Biology of Plant Microbe Interactions*. International Society for Molecular Plant Interactions, St Paul, MN, pp. 387–391.
14. Kouchi,H., Takane,K., So,R.B., Ladha,J.K. and Reddy,P.M. (1999) Rice *enod40*: isolation and expression analysis in rice and transgenic soybean root nodules. *Plant J.*, **18**, 121–129.
15. Compaan,B., Yang,W.C., Bisseling,T. and Franssen,H. (2001) *enod40* expression in the pericycle precedes cortical cell division in *Rhizobium*–legume interaction and the highly conserved internal region of the gene does not encode a peptide. *Plant Soil*, **230**, 1–8.
16. Yang,W.C., Katinakis,P., Hendriks,P., Smolders,A., de Vries,F., Spee,J., Van Kammen,A., Bisseling,T. and Franssen,H. (1993) Characterization of *Gmenod40*, a gene showing novel patterns of cell-specific expression during soybean nodule development. *Plant J.*, **3**, 573–585.
17. Flemetakis,E., Kavroulakis,N., Quaedvlieg,N.E., Spaink,H.P., Dimou,M., Roussis,A. and Katinakis,P. (2000) *Lotus japonicus* contains two distinct *enod40* genes that are expressed in symbiotic, nonsymbiotic and embryonic tissues. *Mol. Plant–Microbe Interact.*, **13**, 987–994.
18. Minami,E., Kouchi,H., Cohn,J.R., Ogawa,T. and Stacey,G. (1996) Expression of the early nodulin, *enod40*, in soybean roots in response to various lipo-chitin signal molecules. *Plant J.*, **10**, 23–32.
19. vanRhijn,P., Fang,Y., Galili,S., Shaul,O., Atzmon,N., Wininger,S., Eshed,Y., Lum,M., Li,Y., To,V. *et al.* (1997) 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. *Proc. Natl Acad. Sci. USA*, **94**, 5467–5472.
20. Charon,C., Johansson,C., Kondorosi,E., Kondorosi,A. and Crespi,M. (1997) *enod40* induces dedifferentiation and division of root cortical cells in legumes. *Proc. Natl Acad. Sci. USA*, **94**, 8901–8906.
21. Charon,C., Sousa,C., Crespi,M. and Kondorosi,A. (1999) Alteration of *enod40* expression modifies *Medicago truncatula* root nodule development induced by *Sinorhizobium meliloti*. *Plant Cell*, **11**, 1953–1966.
22. Kouchi,H. and Hata,S. (1993) Isolation and characterization of novel nodulin cDNAs representing genes expressed at early stages of soybean nodule development. *Mol. Gen. Genet.*, **238**, 106–119.
23. Corich,V., Goormachtig,S., Lievens,S., Van Montagu,M. and Holsters,M. (1998) Patterns of *enod40* gene expression in stem-borne nodules of *Sesbania rostrata*. *Plant Mol. Biol.*, **37**, 67–76.
24. Asad,S., Fang,Y., Wycoff,K.L. and Hirsch,A.M. (1994) Isolation and characterization of cDNA and genomic clones of *Msenod40*: transcripts are detected in meristematic cells of alfalfa. *Protoplasma*, **183**, 10–23.
25. Rohrig,H., Schmidt,J., Miklashevichs,E., Schell,J. and John,M. (2002) Soybean *enod40* encodes two peptides that bind to sucrose synthase. *Proc. Natl Acad. Sci. USA*, **99**, 1915–1920.
26. Sousa,C., Johansson,C., Charon,C., Manyani,H., Sautter,C., Kondorosi,A. and Crespi,M. (2001) 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. *Mol. Cell. Biol.*, **21**, 354–366.
27. Crespi,M.D., Jurkevitch,E., Poiret,M., d'Aubenton-Carafa,Y., Petrovics,G., Kondorosi,E. and Kondorosi,A. (1994) *enod40*, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *EMBO J.*, **13**, 5099–5112.
28. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
29. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.
30. Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In Barciszewski,J. and Clark,B.F.C. (eds), *RNA Biochemistry and Biotechnology*. Kluwer Academic Publishers, Dordrecht, pp. 11–43.
31. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
32. Vijn,I., Yang,W.C., Pallisgard,N., Ostergaard,J.E., Van Kammen,A. and Bisseling,T. (1995) *Vsenod5*, *Vsenod12* and *Vsenod40* expression during *Rhizobium*-induced nodule formation on *Vicia sativa* roots. *Plant Mol. Biol.*, **28**, 1111–1119.
33. Hermann,T. and Westhof,E. (1999) Non-Watson–Crick base pairs in RNA–protein recognition. *Chem. Biol.*, **6**, R335–R343.
34. Hermann,T. and Patel,D.J. (1999) Stitching together RNA tertiary architectures. *J. Mol. Biol.*, **294**, 829–849.

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023

35. Hermann,T. and Patel,D.J. (2000) RNA bulges as architectural and recognition motifs. *Structure Fold. Design*, **8**, R47–R54.
36. SantaLucia,J.,Jr, Kierzek,R. and Turner,D.H. (1991) Stabilities of consecutive A·C,C·C,G·G,U·C and U·U mismatches in RNA internal loops: evidence for stable hydrogen-bonded U·U and C·C·+ pairs. *Biochemistry*, **30**, 8242–8251.
37. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
38. Baeyens,K.J., De Bondt,H.L. and Holbrook,S.R. (1995) Structure of an RNA double helix including uracil–uracil base pairs in an internal loop. *Nature Struct. Biol.*, **2**, 56–62.
39. Lietzke,S.E., Barnes,C.L., Berglund,J.A. and Kundrot,C.E. (1996) The structure of an RNA dodecamer shows how tandem U–U base pairs increase the range of stable RNA structures and the diversity of recognition sites. *Structure*, **4**, 917–930.
40. Schroeder,S.J., Burkard,M.E. and Turner,D.H. (1999) The energetics of small internal loops in RNA. *Biopolymers*, **52**, 157–167.
41. Schroeder,S.J. and Turner,D.H. (2000) Factors affecting the thermodynamic stability of small asymmetric internal loops in RNA. *Biochemistry*, **39**, 9257–9274.
42. Schroeder,S., Kim,J. and Turner,D.H. (1996) G·A and U·U mismatches can stabilize RNA internal loops of three nucleotides. *Biochemistry*, **35**, 16105–16109.
43. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
44. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
45. Le,S.Y., Zhang,K. and Maizel,J.V.,Jr (2002) RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res.*, **30**, 3574–3582.
46. Staehelin,C., Charon,C., Boller,T., Crespi,M. and Kondorosi,A. (2001) *Medicago truncatula* plants overexpressing the early nodulin gene *enod40* exhibit accelerated mycorrhizal colonization and enhanced formation of arbuscules. *Proc. Natl Acad. Sci. USA*, **98**, 15366–15371.

Downloaded from https://academic.oup.com/nar/article/31/17/5003/1192835 by Jacob Heeren user on 28 June 2023