Universiteit Leiden
Opleiding Informatica

Identificaton of Transposable Element Insertion

Into the *enod40* RNA

Name: Diederik Cames van Batenburg
Studentnr: 1045202

Date: 20/10/2016

1st supervisor: Alexander Gultyaev
2nd supervisor: Katy Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Identificaton of Transposable Element Insertion into the *Enod40* RNA

Diederik Cames van Batenburg

Abstract

The *enod40* gene in plants of the *Fabales* order is a root organ regulation gene that contains functional noncoding RNA. Whereas normal DNA is coding for proteins, noncoding RNA is folded into functional molecules. The program Mfold generates a predicted structure from a DNA sequence and gives insight into possible properties of noncoding RNA. Using the BLAST search algorithm on databases of sequenced genomes and Mfold, it is possible to identify noncoding RNA homologues by considering both sequence similarity and structure similarity. Transposable elements are short DNA sequences that can spread out in the genome and occur many times over. They have distinct structural properties. This thesis aims to use bioinformatic methods to identify an evolutionary process of transposable element insertion at the fourth domain of the *enod40* noncoding RNA contributing to its evolution. Results suggest that this transposable element insertion has occured at three points in time, but there is no conclusive evidence to define the moment in evolution precisely.

Acknowledgements

Thanks most of all to my supervisor,
mister Gultyaev,
who gave me an incredible amount of patience and support.

My mother for helping me push on.

My brother for sound advice and a sharp reading.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter we introduce the concepts that are key to the problem of this thesis, the initial observation that started it and the research question.

1.1 Noncoding RNA

Originally, the flow of genetic information within an organism was considered a single unidirectional path from DNA to regulatory proteins [Cri70]: The DNA code is transcribed into messenger RNA which is then translated into proteins. In this model, the "central dogma of molecular biology", the role of RNA is restricted to an intermediary translation step between DNA and proteins as seen in Figure 1.1a. However, during the following 46 years it became apparent that some RNA has functionality of its own.
This "noncoding" RNA, is transcribed from DNA and folds into a functional structure spontaneously (Figure 1.1b). It consists of a single string of nucleotides with four different types of bases: Adenine(A), Urasil(U), Cytocine(C) and Guanine(G). Just like in DNA, the molecular content and structure of each type of base give rise to a certain affinity to lie across from certain other types, following Watson-Crick basepairing. As a result, the RNA molecule is more likely to fold into shapes that contain many of the stable pairs A-U, C-G or U-G than into shapes with different pairs. The configuration of paired nucleotide bases is also called the secondary structure while its underlying sequence is the primary structure. Nucleotides in stems, long sequences of stable pairs, enhance eachothers stability because they will be oriented in a stacked manner [SWBP01]. Other patterns in secondary structure are illustrated in Figure 1.2. A common shape for secondary structures is a combination of stems and loops. In this way, it is possible for DNA to code for functional RNA molecules, based directly on the nucleotide sequence.

Naturally, as certain genes show to be dependent on noncoding RNA for their functionality, further re-

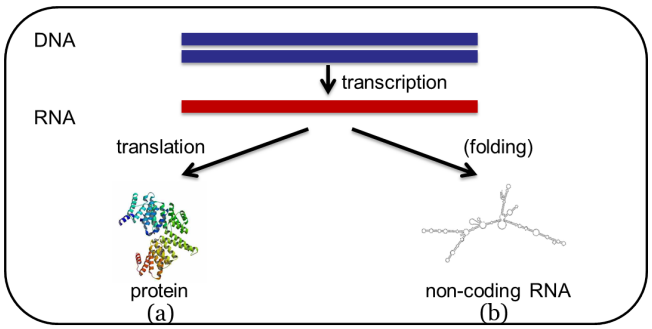

Figure 1.1: The different pathways of genetic information expression. The original model (A) assumes all functional DNA is transcribed into mRNA which is translated into functional proteins. Currently it is known that pathway (B) is also possible in which RNA will be functional itself after spontaneous folding.

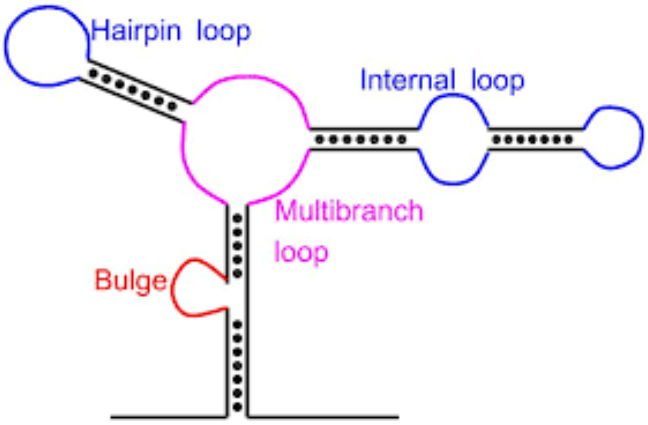

Figure 1.2: Names of different patterns in secondary structure of noncoding RNA.
Any sequence of paired nucleotides of a length greater than three is called a stem (black in this illustration). Unpaired sequences that are not part of loops or bulges are called free strands (at the base at either side of the structure).
From J.Zhang, M.Lin, R.Chen, W.Wang and J.Lang. Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *Journal of Chemical Physics*, 128, 2007.

search is useful for finding and understanding genes. Not only is the further research of noncoding RNA useful in understanding gene functionality, but also its evolutionary mechanisms. It has been suggested that noncoding RNA played a fundamental role in evolution of complexity since it is relatively prevalent in more complex eukaryotes. Moreover, it provides an additional pathway for regulation that works parallel to regulatory proteins [Mat04].

1.2 enod40

The gene that is studied in this thesis is called "Early nodulation gene 40" (*enod40*). This plant-wide conserved[1] gene, first found in the soyabean, was subsequently encountered in most legumes and many non-leguminous plants. The gene shows a overall maximum sequence similarity level of 30% in distantly related species but is about 70% in closely related species (reviewed in [Rut03]). It has a regulatory function in a symbiosis interaction with bacteria of the *Rhizobiaceae* family. After detecting the bacteria, the plant forms

[1] A conserved gene shows relatively low variation, which implies protection by natural selection.

specialized nodule organs at its roots providing shelter and carbon based nutrients for the bacteria. In exchange, the bacteria aid the plant by performing nitrogen fixation from dinitrogen into ammonia. This plays a role in the balancing of nitrogenbased molecules called the nitrogen cycle as shown in Figure1.3.

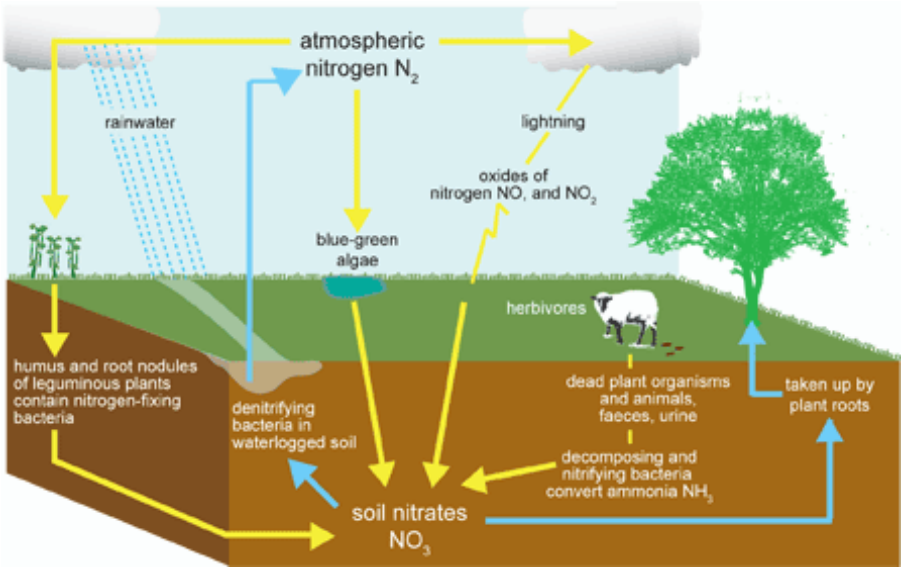

Figure 1.3: The nitrogen cycle. Nitrogen fixation plays an important role in the balance between gaseous and soil nitrates that serve as plant fertilizer. Nitrogen fixation in root nodules is a way to internalize the mutually beneficial relation between plants and nitrogen-fixing bacteria. From: Nitrous Oxide Focus Group [Gro04].

The gene is associated with regulating initiation of the nodules' growth proces [GRG+03]. Because a small subset of non-leguminous plants produce root nodules but most of them contain *enod40*, it is expected that *enod40* also has other developmental functions(reviewed in [Rut03]).

It has been shown that the gene codes for short peptides as well as noncoding RNA and that both are functional [SJC+01, RSM+02]. Studying this gene is well worth the effort since many species of the legume family are highly cultivated for food production. Moreover, the function of the gene in nitrogen fixation might be of importance for agricultural applications as a fertilising proces.

Species may contain several homologues[2] of the *enod40* gene. The *enod40* gene contains multiple noncoding RNA "domains" and two "regions" that have a conserved sequence as can be seen in Figure 1.4. Region I contains a small ORF(sORF I) that translates into a short peptide of 10–15 amino acids and shows high variance [PZF+09]. More conserved elements are the secondary structure of the noncoding RNA domains 2 and 3 and, most of all, the sequence of region II that lies in between. This part of the sequence is therefore called the conserved core consensus RNA structure. Certain motifs are identified that are especially conserved in domains 2 and 3 and consist of specific bases at key positions as well as their basepairings as shown in Figure 1.4. Domain 2 has a variable length and is more conserved near the stem of the structure. The domain 4 is not necessarily present within *enod40* genes and can vary more than domain 2 and 3 between species. It has thus far been identified only in the legume family and only in a small subset of leguminous plants. Remark-

[2] A gene homologue is a sufficiently similar gene so that it is considered the same gene.

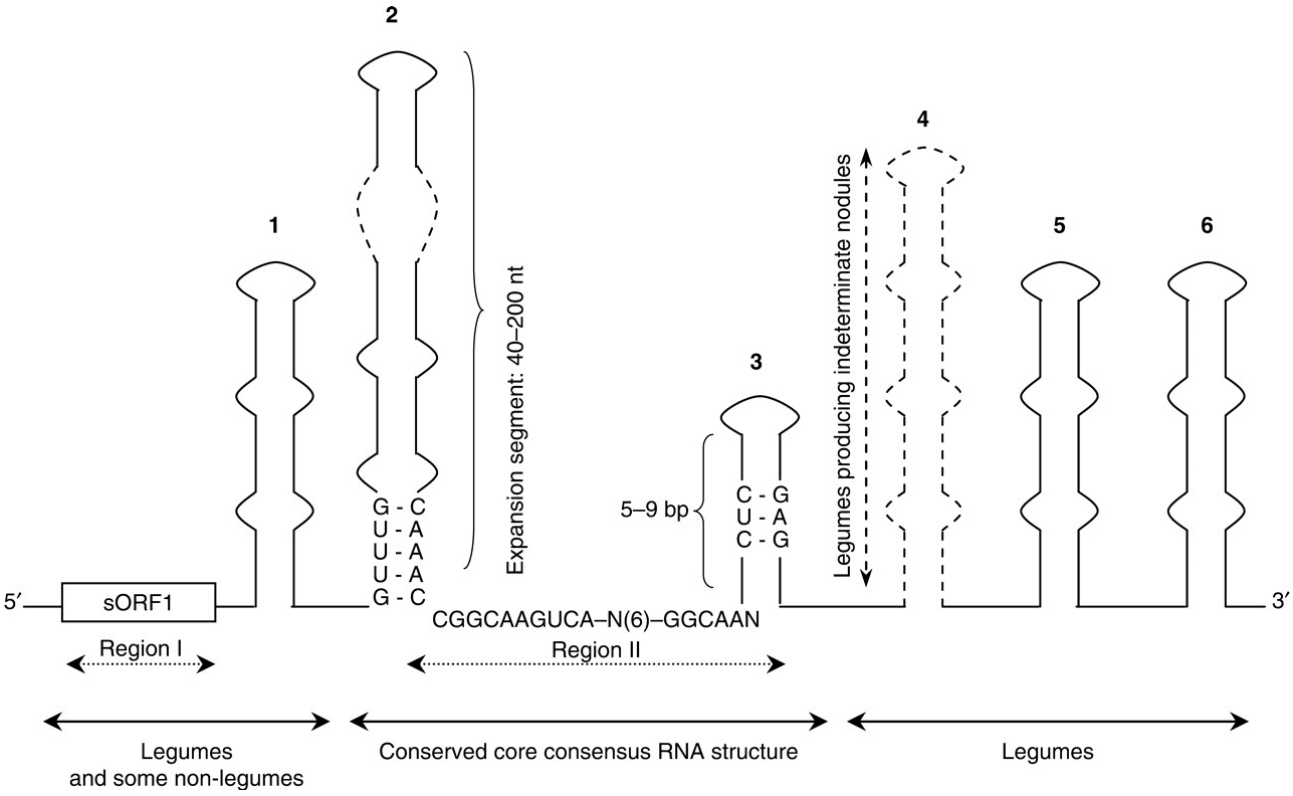

Figure 1.4: Schematic illustration of regions and domains in enod40.
Domains designate conserved secondary structures. Nucleotides written inside domains show conserved motifs. Regions are conserved in sequence, especially region II. The position of sORF I is also shown.

ably, these same species are also the ones that form indeterminate nodules while others form determinate nodules. The biggest difference is that indeterminate nodules grow in long shapes from one apical meristem [GRG+03]. Since legumes perform nitrogen fixation more effectively than other plants [SBF13], domain 4 is also correlated with an increased nitrogen fixation performance. According to unpublished data [Gultyaev, unpublished], this domain in some leguminous plants may originate from transposable elements that have similar stem-loop structures.

1.3 Transposable elements

Transposable elements, also known as transposons, are sequences of DNA within a genome that are capable of moving into another part of the genome. Much variation is found in the abundance of transposable elements in plant DNA from species to species and could be as high as 85% (for maize [SWF+09]). In contrast, the legume *Lotus Japonicus* genome, which is more relevant to this thesis, has a $\sim$ 30% transposable element composition [HZJ+06]. The subclass of autonomous transposable elements is able to move itself using the transposase protein that is encoded within the sequence itself. On the other hand, non-autonomous transposable elements depend on the protein from another transposable element to be able to move. Retrotransposons

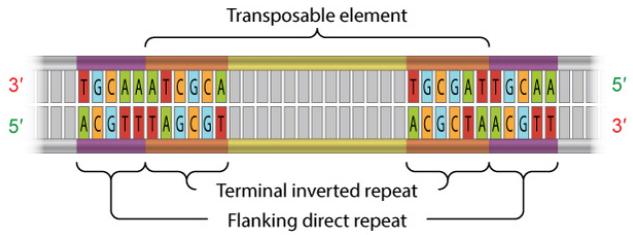

Figure 1.5: The structure of a DNA transposon. Non-autonomous DNA transposons are flanked at both ends by TIRs and TSDs flanking those. The inverted repeats are complements of each other (the repeat at one end is a mirror image of, and composed of complementary nucleotides to, the repeat at the opposing end). The target site duplications are pure duplicates. The middle area in this figure is not representative of the typical length of this sequence.

are transcribed into RNA and back into DNA on insertion. DNA transposons are cut and pasted over shorter distances at specific sequences and stay DNA during the process [Kaz04].

The family of miniature inverted-repeat transposable elements (MITEs) is similar to non-autonomous DNA transposons. They are $\sim$ 100–500 base pairs long and do not code for proteins. As illustrated in Figure 1.5, their sequence is flanked by characteristic target site duplications(TSDs) and conserved terminal inverted repeats(TIRs). However, they show higher copy numbers and uniformity (in structure) giving reason to classify them differently [FZW02]. If the MITE would also have a stable secondary structure as noncoding RNA, the TIRs would represent the stem of the structure.

The view of the role of transposable elements has changed multiple times since the intial discovery in 1951. At first it was expected to perform developmental or evolutionary change in regulation of nearby gene expression. Around 1980, the noncoding parts of DNA were considered "junk DNA" without functionality. This gave rise to the notion that transposable elements are "selfish" parasites in the form of DNA, although a symbiotic relation could occur by chance. Even so, it took a decade after the first complete sequences of the human genome in 2000, before transposable elements were given attention as functional elements. Currently, transposable elements are widely recognized contributors to gene regulation.

Moreover, study of evolutionary emergence of conserved genes in primates showed that transposable elements provided about 50% of the source material. Since this equals the relative amount of transposable elements, this suggests they are just as likely to evolve functionality as regular DNA. A large part of noncoding RNAs was also estimated to be orginated from and regulated by transposable elements [RDRP16]. This thesis is about identifying a transposable element insertion into a noncoding part of the *enod40* gene.

1.4 Research question

As *enod40* RNA structures of legumes *Cicer arietinum* and *Lupinus angustifolius* were studied by A.Gultyaev it appeared that sequences similar to one of the *enod40* domains were found many times over in its corresponding genomes suggesting the domains originate from transposable elements [Gultyaev, unpublished]. A similar sequence was also found manyfold in the *Medicago truncatula* species, another legume. After comparative analysis of highly stable folding structures of the C.arietinum *enod40* domain and its homologues, this sequence was found to be homologous to a known MITE according to the Plant Repeat Databases at Michigan State University [OB04]. This thesis aims to find a hereditary link between transposable elements and functional elements in the *enod40* RNA. The goal is to identify an event or multiple events in evolution of transposable element insertion into *enod40*, after which such element(s) became a functional part of the gene. This would be possible by studying the phylogeny[3] of the species at which we can find an insertion compared to those that do not show it. For example a shared insertion would make it likely the event occured before the species have diverged. Identification of such insertions require annotation of multiple copies of (previously unknown) *enod40* genes in available genomic sequences and genome-wide search for similar sequences and structures of transposable elements.

1.5 Related work

The most directly related work is those studies that try to identify conserved properties of secondary structures in *enod40*. The article of Girard et al [GRG$^+$03] focussed on finding its conserved domains using both a bioinformatic approach as well as a structure probing laboratory experiment. It is especially useful as a source of annotated *enod40* genes in the legume family. Somewhat more related is the article by Gultyaev et al [GR07] in their effort to find *enod40* homologues based on the conserved secondary structures. The work by Santi [SBF13] and Mus [MCG$^+$16] study the differences between leguminous and non-leguminous symbiosis with nitrogen-fixating bacteria and how they can be overcome. Comparison shows that there is some common ground in signalling pathways for both nodule-forming and other kinds of symbiosis with nitrogen-fixing bacteria. In agricultural context, the extensive fertilisation with nitrogen based molecules that can be directly absorbed by plants leads to longterm unbalances in the soil. Because the majority of plants grown for food are non-leguminous, an effort is made to transfer the more efficient root nodule symbiosis to these plant species in order to reduce fertilisation. The work of Lisch [Lis13] explores the role of transposable elements in plant evolution apart from noncoding RNA insertions.

[3] A phylogeny is a tree diagram of the hypothesized evolutionary relationships of a group of organisms.

1.6 Thesis Overview

The thesis is structured as follows: This chapter contains the introduction which also includes related work; chapter 2 includes the definitions, used tools and the workplan and choices that were made; chapter 3 discusses the results; chapter 4 concludes on a evolutionary level and discusses the impact of some choices and results on the research.

Chapter 2

Methods

In this chapter we describe what methods were used to perform the research.

2.1 NCBI BioProject database

With the advent of a new generation of cheap and fast DNA sequencing machines, called next generation sequencing, the amount of sequenced genome has grown considerably. Instead of doing our own laboratory experiments, we make use of the resourceful BioProject database of the NCBI organisation. This project provides an open access to many different databases including genome sequences generated by other scientists for their own research purposes. It uses accession numbers to uniquely identify individual submissions [TDB+13], and within them areas can be designated by specifying the nucleotide position numbers of the range. Apart from directly searching entries by accession number, there are several search algorithms available. These algorithms allow users to seach within specified groups of entries to compare them to a specific nucelotide sequence given by the user.

2.2 WGS

Whole genome sequencing (WGS) using shotgun methods is a sequencing technique that generates large amounts of randomly positioned short fragments of a sequence. In order to obtain a single read from the material, a computationally intensive algorithm reconstructs the sequence based on overlap between the fragments. The goal is to eventually sequence the whole genome of a species reliably. This method is especially attractive since the current sequencing machines are capable of generating short $\sim$ 25–500 basepair

sequences very fast. One of the side-effects of this method is that some parts might not (yet) be sufficiently covered at which point the WGS read contains gaps [TDB+13].

2.3 Alignment algorithms

In order to find similar sequences in genomes, the Basic Local Alignment Search Tool(BLAST) algorithm is used. It requires a query sequence and searches a specified range of genomes for similar or partly similar sequences. Performing a search returns a list of aligned sequences above a specified quality threshold, allowing deletions, insertions and mismatches with a certain penalty. Along with the results, it provides the Expect-value(E-value) metric for matching quality. It represents the expected value of the number of sequences that would have (at least) an equally well-aligned match, in a random database of the same size.
The BLAST is a heuristic algorithm for fast results and aligns locally, allowing partial alignments. It uses small words, which are subsequences of the query sequence, to find areas of partial similarity. In the following steps, BLAST keeps the sequences that can be expanded to include more of the query sequence without losing too much quality. High quality alignments are no longer expanded if a certain range of expansion only results in lower similarity.
Another useful alignment algorithm is the Needleman-Wunsch global alignment. It provides the best possible alignment between two sequences, always using the full query sequence which is useful for comparing two highly similar sequences. Using this algorithm for alignment, the differences, rather than the similarities, between the two sequences are more easily analyzed [TDB+13].

2.4 Mfold

Mfold is a secondary structure prediction tool for DNA or RNA molecules [Zuk03]. As said before, RNA folding is partly determined by base pairings that provide stability. This is quantified as the thermodynamic measure of free energy change: The most likely spontaneous folding of the RNA strand would decrease its free energy the most. It is estimated that, for short sequences, about 70% of secondary structures are predicted correctly based on only thermodynamics. Just like alignment algorithms, Mfold uses dynamic programming to calculate possible foldings efficiently by building on earlier conclusions to subproblems recursively [MMT10]. It generates many possible foldings and calculates a free energy level for each prediction and ranks them by free energy change.

2.5 Outline

In order to answer the research question, it is necessary to identify as many cases as possible of transposable element insertion at domain 4 in *enod40* genes, forming domain 4. Ultimately, we aim to deduce the moment in evolution and impact of the event, based on the interrelation within the phylogenetic tree of researched species. As said in chapter 2.1, the already present genomic information at the NCBI databases is used as primary data source, as well as annotations done in previous work (chapter 2.1 and unpublished work by A. Gultyaev). However, previous publications might not be complete in respect to current knowledge since the NCBI databases are continuously updated with newly sequenced material so there is reason to revisit the previously studied species. The first step is to list all known *enod40* containing species, and expand this list by looking for additional previously unknown *enod40* homologues (using the method described in 2.6). Since domain 4 is not always present in *enod40*, we filter the list of confirmed *enod40* genes to select those that contain domain 4. The domain 4 of these candidates is then tested for being an inserted transposable element by performing BLAST. Using domain 4 of each candidate as query, a BLAST search is performed in the whole genome of that candidates species for sequences similar to its domain 4 sequence. If many significant matches are returned that are not part of known *enod40* homologues, they are very likely a transposable element. To confirm their identity, the hits that give the best MITE-like secondary structure are used as BLAST query in databases of known plant MITEs.
Alternatively, domain 4 sequences of one species are used as BLAST queries against other *enod40* containing species. The same action would even be useful when the other species' *enod40* genes do not contain domain 4. This could find traces of yet unsequenced *enod40* genes which might contain domain 4. Otherwise, it could indicate cases where, after insertion at domain 4, the tranposable element lost its tendency to increase in number, while it did not in other descendants.

2.6 Finding enod40 homologues

As described in chapter 1.2, a large part of the gene shows no conserved sequences. At the same time, presence of its conserved secondary structures is necessary for its functionality. In order to identify a candidate as an *enod40* homologue, it requires the presence of region II, domain 2, domain 3 and the sORFI with a reasonably similar translation. However, any of the other domains also weigh in for the decision, and might make up for the lacking of other elements. The final judgment is a matter of collecting sufficient evidence, where highly conserved elements weigh in stronger than lesser conserved elements.
Since finding conserved secondary sequences is not automated as well as BLAST searching similar sequences, we focus on region II first. Taking the sequence of any region II of a known *enod40* gene as BLAST query would suffice since it is strongly conserved. We chose L32806 *Medicago sativa*, Ms-1 from the work by Girard

et al. [GRG+03], because it is well annotated and familiar. Since region II is very strongly conserved, it makes little difference at this stage what species is chosen for taking its region II as query. The resulting set of significant hits are filtered for exact duplicate submissions by aligning hits with eachother that show suspiciously similar statistics.

For each candidate, we perform a Needleman-Wunsch alignment against a known and closely related *enod40* gene, with a query of the aligned sequence, expanded in both directions to encompass the expected dom1–dom4 range. This allows for a full annotation of the gene, including the expected positions of structural non-coding RNA structures using Region II as anchorpoint. The sequences that roughly align with the respective domains are used as input for Mfold predictions to confirm their presence and location. The decision to call a structure homologue is again based on several properties: its relative position to other elements, length, shape, and amount of bulges and loops compared to known homologues, as well as conserved motifs. Although there is generally little conserved sequence in the domains, we also take sequence similarity into account. If the most likely predicted secondary structure is inconclusive, we take into account predictions of higher free energy level if the difference is not too large. Other Mfold settings can be investigated as well. Raising the suboptimality shows more of the less likely predictions, while changing the temperature gives different results.

Finally, we try to find sORFI using the surrounding sequence if its expected position as input in the ORF finder by Stothard et al. [Stooo] and compare the translated peptide to already known variants [PZF+09]. All positions of found elements are annotated on the corresponding sequence.

2.7 Scope of search space

Because of time limitations and in order to completely cover some part of of the plant species we chose to limit our search to species of the *Fabales* order. This is because the species of the initial observation were all member of the legumes(*Fabaceae*) family only, which is a subset of *Fabales*. Moreover, we expect to find *enod40* genes in most plants, while the domain 4 which we try to find has thus far been found only in legumes. Within this search space we were limited by the available sequence information on the NCBI databases.

Chapter 3

Results

Here we describe the results of the thesis.

3.1 enod40 homologues

Results for searching new *enod40* homologues are presented in Table 3.1. Ultimately, 18 different species from the *Fabales* order were considered. A total of 17 species contained at least one *enod40* homologue. In the case of *Vicia faba*, which returned an incomplete sequence, the only candidate gave inconclusive results. In some *enod40* RNAs the typical extended stem-loop structure of domain 4 could not be unambiguously identified. In these cases the regions downstream of the domain 3, roughly corresponding to the putative domain 4, were treated as "inconclusive domain 4" sequences. None of the newly found *enod40* homologues showed a clear presence of domain 4. Because of this, the 8 species from the article by Girard et al. [GRG$^+$03] were also included to search for additional copies. For the *Glycine max* species, the plant database of Ensembl, EnsemblPlants, was also used as a source for sequenced genome [KAA$^+$16].

3.2 Domain 4 similarity

Table 3.2 depicts the attempts to find new instances like the initial observation, by finding many sequences that are similar to domain 4. Since there were no new identifications of domain 4 - containing *enod40* genes, the query is restricted to domain 4 homologues that were already known. As seen in the original observation, there were many significant hits of similar sequences to its own domain 4 in the *Lupinus angustifolius* (334 with AOCW01181709.1) and *C.arietinum* (1034 with XM_004509513). Some cross species BLAST searching was done as well, where sequences similar to domain 4 of species A are searched in the genome of species B.

The *M.sativa* domain 4 of accession L32806.1 showed 173 significant hits of similar sequences in the *M.truncatula* genome. Alternatively, the second *M.truncatula* homologue returned 62 hits and using *M.truncatula* domain 4 itself as query returned only 24 hits. Since there is always a possibility that some of the judgments on domain 4 presence were wrong, we also considered the inconclusive regions corresponding to the putative domain 4. A newly found homologue of *C.arietinum* returned 32 similar sequences using its (inconclusive) domain 4 as BLAST query on the *C.arietinum* genome. However, this number is much lower compared to the number of hits from the known domain 4 containing homologue.

3.2.1 Medicago truncatula hits

Although there is no wgs available for *M.sativa*, its domain 4 of its *enod40* RNA returned 173 hits of significant sequence similarity in the *M.truncatula* genome (Table 3.2). The sequence similarity of the most significant hit is shown in Figure 3.1. The hits were researched while prioritizing those of high sequence similarity. Secondary structures were predicted with mfold(version 2.3 energies) using a 10 $^{\circ}$ C temperature setting. Exploring the structural properties of the hits showed a variety of structural properties of which many were MITE-like. All predicted structures showed at least partial stemloops of varying sizes. The difference in MITE-like properties is defined by the amounts of internal loops, bulges and multibranch loops that break up the stem. The ideal MITE structure would be a thermodynamically stable extended stem-loop structure with terminal palindromic sequences (Figure 1.5). Out of 173 putative MITE-like sequences similar to *M.sativa enod40* domain 4, ten promising sequences were selected for further consideration on the basis of their similarities to the query and stabilities of stem-loop structures. Choosing each of these promising sequences as a query on the same genome, seven additional sequences were discovered that were at least equally MITE-like. Of course, known domains 4 in the identified *enod40* homologues were excluded from the candidates. The same treatment was planned for inconclusive domains 4, but these sequences never occurred as a candidate. In Figure 3.2, we describe three of the most interesting cases and the following results gained by using their sequence as BLAST query.

3.2.2 MITE confirmation by databases

Taking the 17 best structures we performed BLAST searches in databases of known MITEs. The Plant Repeat Databases at Michigan State University returned only one hit (for APNO01002116.1) with an E-value of 0.85 for a *M.sativa* retrotransposon polyprotein gene interrupted by LTR retroelement MCIRE (complete sequence). This corresponds to a rather low sequence similarity. The database contains only 32 known MITEs of *Medicago* repeats and they are all from the *sativa* species. Another database, the P-MITE database [CHZ^{+}14],

Species	wgs available	accession	duplicates	enod40 homolog	sORF1	Domain 1	Domain 2	Region 2	Domain 3	Domain 4	enod40 annotation
Arachis duranensis	x	JQIN01000766		x	1236883-1236923		1237003-1237156	1237157-1237179	1237180-1237191		this article
		JQIN01000494		x	625055-625093	625155-625160	625160-625265	625266-625288			this article
Arachis ipaensis	x	JQIO01000187		x	8067216-8067253	8067265-8067328	8067335-8067448	8067449-8067470	8067471-8067488		this article
		JQIO01000138		?	1743167-1743205		?	1743375-1743399			this article
Cajanus cajan	x	AFSP01000939	AGCT01004525	x	4368-4333	4318-4272	4267-4142	4141-4120			this article
		AGCT01060158	AFSP01019676	x	31421-31386	31385-31332	31324-31183	31184-31163	31162-31140		this article
Cicer arietinum	x	NC_021166	AHII02010231.1 ANPC01013081.1 XM_004509513	x	16999054-16999092	16999103-16999157	16999161-16999289	16999290-16999311	16999312-16999329	16999382-16999510	Gultyaev (unpublished)
		AHII02012933.1	ANPC01002572.1	x	10060-10098	10111-10178	10190 -10339	10340-10362	10363-10384	?	this article
Glycine max	x	X69154 (Gm1)	ACUP02000722 BBNX02005343	x	47-82	91-137	148-272	273-295	296-312		Girard et al
		BBNX02000150	ACUP02000049	x	18038-18074	18093-18127	18142-18258	18259-18280	18281-18299		this article
		ACUP02005273	BBNX02038331	x	310717-310752	310765-310810	310815-310932	310933-310954	310955-310978		this article
		BBNX02070215	ACUP02010110		54982-54947	54936-54890	54867-54745				this article
		BBNX02101159	ACUP02002876		6850-6815	6793-6763	6738-6625				this article
		BBNX02001833	ACUP02000325		52059-52024	52004-51970					this article
Glycine soja	x	AZNC01025921		x	22830-22865	22886-22918	22932-23057	23058-23079	23080-23096		this article
		AZNC01055602		x	27086-27051		26975-26854	26855-26833	26834-26818		this article
		AZNC01029573		x	13061-13096	13109-13152	13157-13303	13304-13319	26819-13352		this article
		AZNC01058527		x	26959-26925	26911-26867	26862-26723	26743-26723	26722-26698		this article
Lotus japonicus	x	AJ271787 (Lj1)	BABK02038321.1	x	?	374-416	422-553	554-575	575-594		Girard et al
		AJ271788 (Lj2)	BABK02041305.1	x	?	375-434	441-616	617-638	639-656		Girard et al
Lupinus luteus	-	AF352375 (Ll)		x	?	109-137	153-288	288-309	310-327		Girard et al
Lupinus angustifolius	x	AOCW01106645.1		x	2387-2422	2427-2480	2485-2620	2621-2641	2642-2657		this article
		AOCW01112350.1		x	1687-1652	1623-1599	1581-1471	1470-1450	1449-1432	?	this article
		AOCW01181709.1		x		13054-13112	13118-13259	13260-13281	13282-13298	13299-13420	Gultyaev (unpublished)
		AOCW01036497.1		x	7038-7003	6983-6955	6939-6800	6799-6777	6776-6758	?	this article
Medicago truncatula	x	APNO01004230		?	7816-7867			8082-8104	?	?	this article
		X80264 (*Medicago truncatula*)	APNO01002340.1	x	96-134	145-199	211-339	340-361	362-380	405-468	Girard et al
Medicago sativa	-	L32806 (Ms1)		x	65-103	114-168	180-308	309-330	331-349	373-438	Girard et al
		X80263 (Ms2)		x	96-134	94-148	160-288	289-310	311-329	353-417	Girard et al
Phaseolus vulgaris	x	X86441 (Pv)	ANNZ01004416.1 LPQZ01041286.1			91-140	145-267	268-289	290-306		Girard et al
		LPQZ01024362.1	ANNZ01007225.1		5816-5781	5778-5722	5718-5580	5579-5558	5557-5534		this article
Pisum sativum	-	X81064 (Ps)		x	93-168	112-186	200-326	327-348	349-367	405-496	Girard et al
Trifolium repens	-	AJ000268 (Tr)		x	152-226	170-226	238-372	373-395	396-419	436-498	Girard et al
Vicia faba	x	CSVX01021651		?	?	?	?	16-37	39-54		this article
Vigna angularis	x	JZJH01053362	AUGG01022070	x	12071-12036	12013-11981	11967 -11845	11844-11824	11823-11800	?	this article
		JZJH01057672	AUGG01028463	x	1051-1086	1106-1132	1146-1287	1288-1309	1310-1333		this article
Vigna radiata	x	AF061818 (Vr)	LJIH01000289.1 JJMO01000319.1	x	?	21-55	68-190	191-212	213-229		Girard et al
		LJIH01000347.1	JJMO01001837.1	x	252438-252473	252485-252528	252533-252674	252675-25296	2597-26020		this article
Vicia sativa	-	X83683 (Vs)		x	?	110-193	207-334	335-356	357-375	425-517	Girard et al

Table 3.1:
Annotations of *enod40* candidate homologues within the *Fabales* order.
Presence of each element is indicated with the symbol 'x' or a specified range, absence with empty entries and inconclusive cases with the symbol '?'. The column *enod40* homologue depicts the conclusion whether the candidate is indeed a homologue. The last column indicates the source of the annotation. After accession numbers of some homologues, parenthesed codes are the names used in Girard et al. [GRG+03].

Target wgs	Query species	Query	Domain 4	Rough range domain 4	# significant hits
Lupinus angustifolius	Lupinus angustifolius	AOCW01106645.1		2640 - 2740	3
		AOCW01112350.1	?	1343 - 1448	3
(initial observation)		AOCW01181709.1	x	2054 - 2158	334
		AOCW01036497.1	?	6663 - 6749	3
Cicer arietinum	Cicer arietinum	AHII02012933.1	?	10385 - 10486	32
(initial observation)	Cicer arietinum	XM_004509513	x	483-611	1034
	Lupinus angustifolius	AOCW01181709.1	x	2054-2158	0
Vigna angularis	Vigna angularis	JZJH01053362	?	11803-11737	2
Medicago truncatula	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	173
(initial observation)	Cicer arietinum	XM_004509513	x	483-611	72
	Medicago sativa	X80263 (Ms2)	x	353-417	62
	Medicago truncatula	X80264 (Mt)	x	405 - 468	24
	Lupinus angustifolius	AOCW01181709.1	x	2054-2158	0
Vicia faba	Vicia sativa	X83683 (Vs)	x	425 - 517	1
Lotus japonicus	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0
Glycine max	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0
Arachis duranensis	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0
Arachis ipaensis	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0
Cajanus cajan	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0
Vigna angularis	Medicago sativa	L32806.1 (Ms1)	x	374 - 438	0

Table 3.2: Results for BLAST searching similar sequences to domain 4 in wgs targets, cross-species and intra-species. Significantly resembling hits are counted with a threshold of $E < 1e^{-3}$ and $\geq$ 50% cover. Presence of domain 4 is indicated with the symbol 'x', absence with empty entries and inconclusive cases with the symbol '?'. Parenthesed codes are the names used in Girard et al [GRG+03]. In this table, hits are counted without omitting known domain 4 homologues.

did give many hits for most of the structures as can be seen in Table 3.3. All hits were of partial sequences of known *M.truncatula* MITEs of the superfamily hAT. More specifically, hits belonged to different families DTA_Met2, 3, 9, 10 and 14 with a high proportion of familynumber 2, 9 and 10. The most hits were found for the APNO01002116.1 *M.truncatula* sequence.

Many of the MITE hits show comparable levels of MITE-like properties but also a much larger structure: MITEs of $\sim$ 80–210 nucleotides versus queries containing only 63 nucleotide queries. Parts of query sequences that showed similarity are folded differently when positioned in the larger structures. One particularly interesting hit was a fully identical MITE sequence from MITE family DTA_met2. Expanding the query to the MITE size still resulted in completely equal sequences, as can be seen in Figure 3.3. However, the orginal structure is very different from the one inside the larger MITE (Figure 3.4).

Looking at the secundary structure of the found MITE (Figure 3.4.C), it appears to be surprisingly uncharacteristic to MITEs. Most importantly, it lacks a clear palindromic stemloop at the base, corresponding to a terminal inverted repeat (Figure 1.5).

The sequence in the *C.arietinum* genome that was found to be similar to a *Medicago* MITE in the initial observation was compared with BLAST on the newly found *M.truncatula* MITE collection in Table 3.3. This sequence (ANPC01009927 10287-10414) showed no significant sequence similarity to any of these sequences. The initial observation also reported a collection of sequences in *M.truncatula* that were similar to this same sequence ANPC01009927 (10287-10414). These sequences were also compared to the newly found *M.truncatula* MITE collection that is represented in Table 3.3. The 100 most significant hits of both collections were all of distinct sequences.

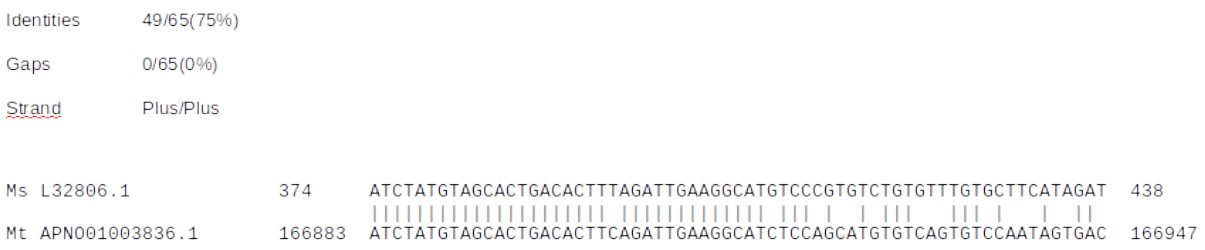

Figure 3.1: Alignment of *M.sativa* domain 4, highly similar to M.truncatula sequence outside of *enod40*.
The highly similar left side appeared as a BLAST hit using the complete *M.sativa* domain 4 sequence as query. This hit was expanded to cover the entire query with a Needleman-Wunsch global alignment.

Accession	range	P-MITE hits ($E \leq 1e^{-3}$)
APNO01000612.1	476177-476241	17
APNO01000622.1	680406-680470	18
APNO01001242.1	90350-90286	28
APNO01001506.1	17935-17871	9
APNO01001679.1	358027-358091	48
APNO01002116.1	140835-140744	63
APNO01002328.1	2385496-2385432	-
APNO01002484.1	256526-256462	3
APNO01002498.1	26283-26347	2
APNO01003836.1	166883-166947	30
APNO01001658.1	133185-133249	14
APNO01001651.1	227514-227578	11
APNO01003105.1	86823-86887	5
APNO01003892.1	43509-43445	7
APNO01002311.1	234880-234943	14
APNO01001109.1	44629-44565	11
APNO01002764.1	5710-5774	6

Table 3.3: The amount of significant hits found for comparing 17 of sequences in *M.truncatula* that are similar to the *M.sativa* domain 4 to known MITEs in the P-MITE database.
Using the *M.truncatula* domain 4 itself as query on the *M.truncatula* genome, four of these hits were also present with similar ranges: APNO01001242.1, APNO01002116.1, APNO01002328.1 and APNO01003836.1.

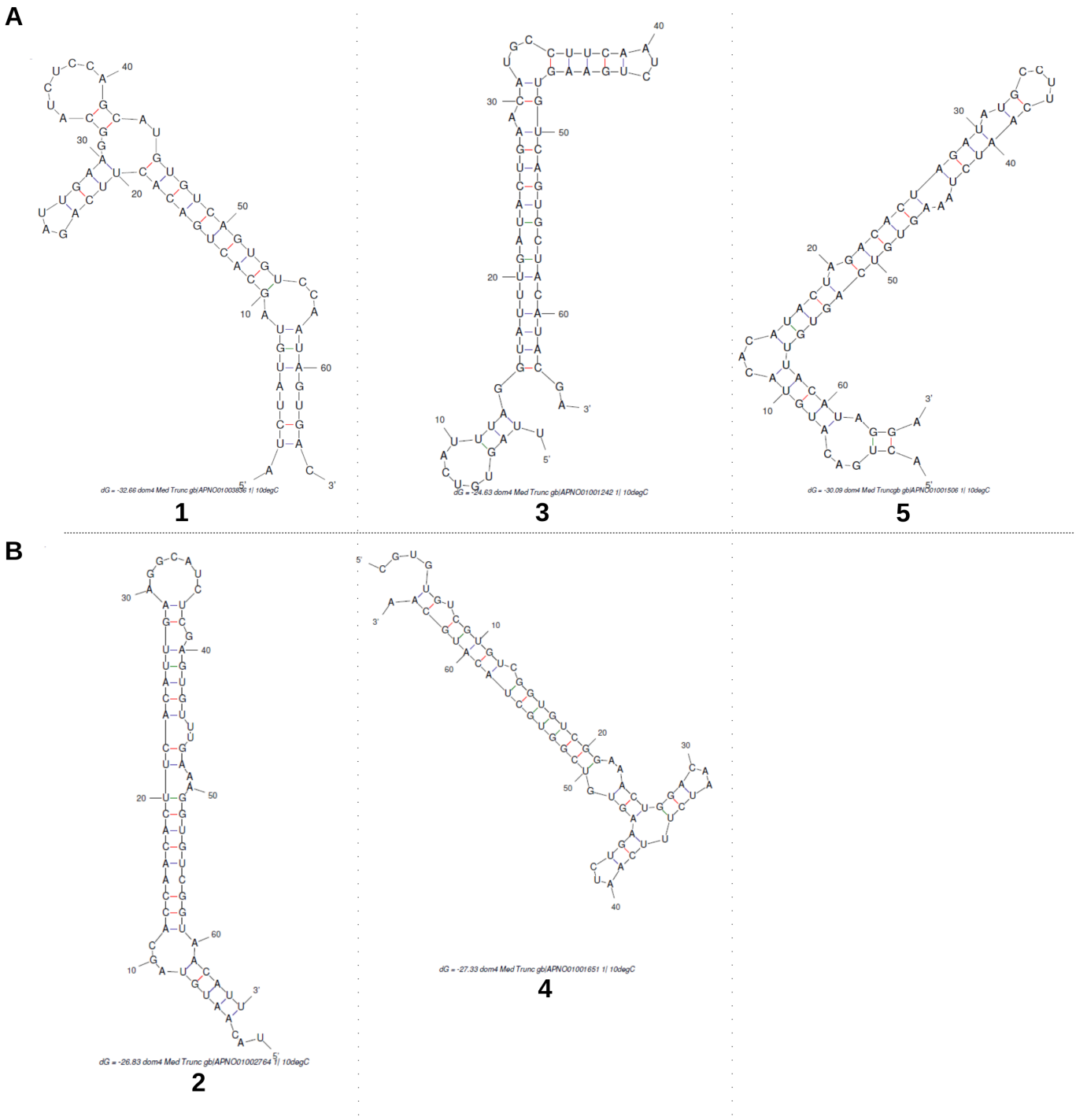

Figure 3.2: Predicted secondary structures of *M.truncatula* BLAST hits that most closely resemble a typical MITE. Structures in section **A** are from the initial BLAST search using *M.sativa* domain 4 as query. Structures in section **B** are hits from BLAST searches using the sequence directly above in **A** as query. Using structure 5 as query returned no noteworthy results.
Section **A**: **1** From the same BLAST hit as in Figure 3.1. This y-shaped structure shows a medium internal loop, a medium and very small hairpin loop and a small multibranch loop. Its base-pairing is strong, except for two bad pairings at the basestem. **3** This structure is preceded by a smaller hairpin with medium hairpin loop. The main structure has two small bulges (one of which a U-bulge) and a very small hairpin loop. Basepairing is strong in all stems. **5** This single stemloop contains four small bulges, one small internal loop and a small hairpin loop. Basepairing is strong in all stems.
Section **B**: **2** This single stemloop contains three small bulges (one of which a UU-bulge), a small internal loop and a medium hairpin loop. Basepairing is strong except for a single pair. **4** This T-shaped structure contains two small bulges, a small internal loop, a small multibranch loop and two very small hairpin loops. Basepairing is strong in all stems.

A

```
Query:      L32806.1 Medicago sativa domain 4, range 374-438
Subject:    APNO01001679.1 Medicago truncatula, range 358027-358091

Identities:     49/69(71%)          Gaps:  5/69(7%)          Strand: Plus/Plus

Query  374     ATCTATGTAGCACTGACACTTTAGATTGAAGGCATGTCCC-GTGTCTGTGTTTGTGCTTC  432
               ||||||||||| ||||||| ||||||||||||||||   ||||| |||  |   | |
Sbjct  358027  CACTATGTAGCACCGACACTTCAGATTGAAGGCATGTCATTGTGTC-GTGAGT---CGTG  358082

Query  433     ATAGAT     438
               | || |
Sbjct  358083  ACAGGTGTC  358091
```

B

```
Query:      APNO01001679.1 Medicago truncatula, range 358027-358091
Subject:    SQ203116789 Medicago truncatula (P-MITE number)

Identities:     172/172(100%)       Gaps:       0/172(0%)          Strand:      Plus/Minus

Query  358029  CTATGTAGCACCGACACTTCAGATTGAAGGCATGTCATTGTGTCGTGAGTCGTGACAGGT  358088
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1       CTATGTAGCACCGACACTTCAGATTGAAGGCATGTCATTGTGTCGTGAGTCGTGACAGGT  60

Query  358089  GTCTGACACGAAACATGTGGTTACCTTTAAATATTTCATTTTCTCAAATTATTGTTGGTG  358148
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61      GTCTGACACGAAACATGTGGTTACCTTTAAATATTTCATTTTCTCAAATTATTGTTGGTG  120

Query  358149  TCGGTGTCATGTTCGGTGTCTGTGTCGGTAAAAGTGTTTCGTAGGTAATAGC  358200
               |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121     TCGGTGTCATGTTCGGTGTCTGTGTCGGTAAAAGTGTTTCGTAGGTAATAGC  172
```

Figure 3.3: Alignment of (**A**) *M.sativa* domain 4 to similar sequence in *M.truncatula* and (**B**) of this (expanded) sequence to a known *M.truncatula* MITE.
The APNO01001679.1 sequence in *M.truncatula* is highly similar to domain 4 of Ms, and 100% equal to the known SQ203116789 MITE in *M.truncatula* from MITE family DTA_met2. Both alignments are done with a Needleman-Wunsch global alignment and input sequences were expanded to cover the entire query.

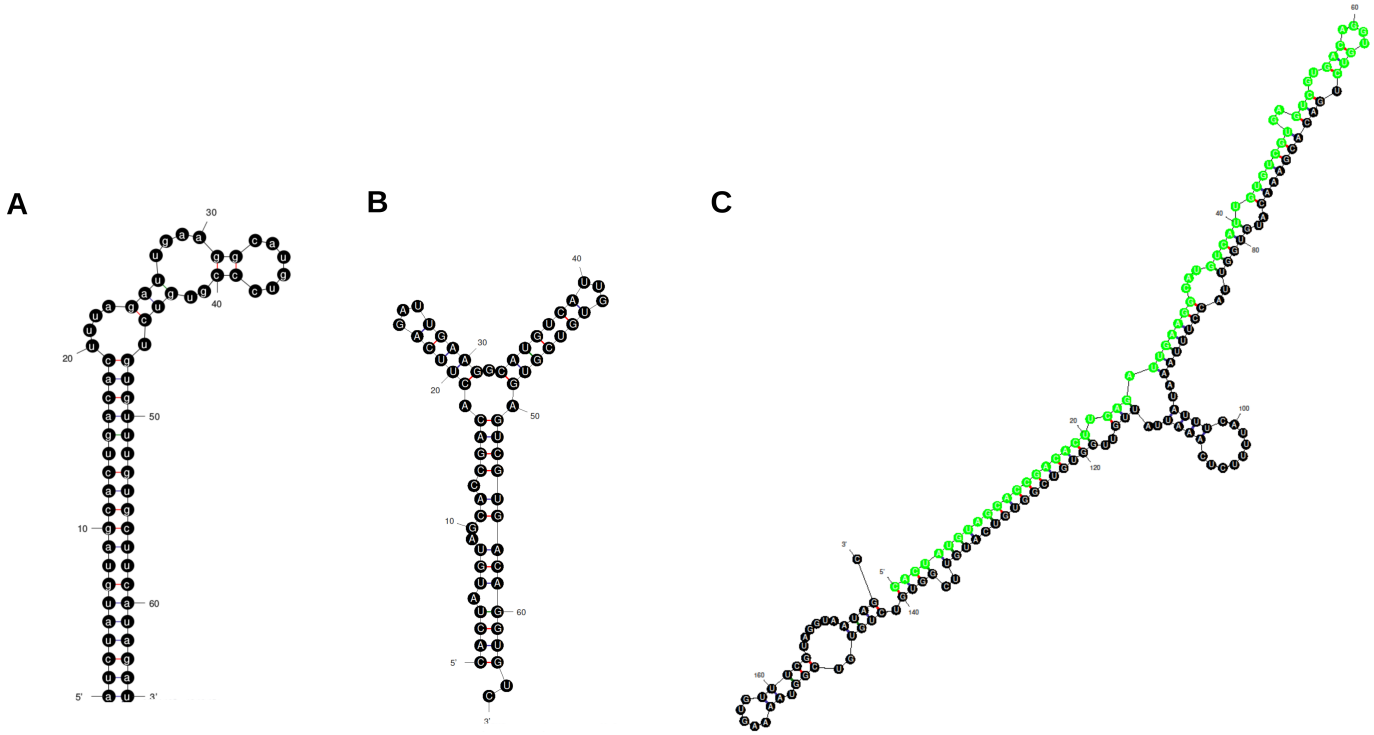

Figure 3.4: Predicted secondary structure of (**A**) domain 4 of *M.sativa*, (**B**) a *M.truncatula* sequence that is similar to **A**, APNO01001679 range 358027-358091 and (**C**) the known MITE that partially has an identical sequence, SQ203116789 from MITE family DTA_met2.
The secondary structure of the original domain 4 of Ms is a single stemloop **A** instead of the T-shape in **B**. **A** contains several U-internal loops which is typical of domain 4, while **B** does not. Green nucleotides in **C** indicate the position of the sequence **B** within this structure. It is part of one side of a stem and no longer pairs up within itself for most of the sequence.

Chapter 4

Discussion

All of 18 studied species in the *Fabales* order showed at least one homologue of the *enod40* gene, if enough genome sequencing data was available. This is support for the importance of the gene since it is so strongly conserved. Even though the gene is this common, presence of domain 4 is found in only 8 out of 36 conclusive *enod40* homologues, suggesting it is a rare domain. It also limited the ways we could search for instances where many domain 4 resembling sequences are found.

We have found two new instances of a domain 4 that shows many similar copies throughout a *Fabales* genome, one in the *M.truncatula* genome and a second one in the *C.arietinum* genome. The case of *M.sativa* domain 4 throughout the *M.truncatula* genome was extensively analyzed. Many of the hits were signficiantly similar to known MITEs of *M.truncatula*. Most interestingly, one of the hits appeared to be exactly a known MITE. It showed no sequence similarity to the *C.arietinum* MITEs from the initial observation. The MITE has a much greater length than the original hit. Expanding the query to match the MITE size still returned a perfect alignment.

However, the original structure in domain 4 is not comparable to its component within the MITE, where it is situated as the first half of a palindrome. Moreover, we were unable to find an example where the structures were comparable. Most often, the corresponding sequence was present in one of the strands of a double stranded component.

The evolutionary explanation of the difference in length is purely speculative. Possibly, only a small part of a MITE was inserted as domain 4 in *M.sativa*. Alternatively, a considerable amount of evolutionary change could have occured after the original insertion. This would explain the situation as a large deletion of an initial inserted MITE sequence in the *M.sativa enod40* RNA.

Moreover, the predicted structure of the associated MITE as in 3.2.C is not corresponding to an expected MITE structure. There is no clear representative for the typical terminal inverted repeat in the form of a strongly palindromic stem at the base of the structure. This is due to an additional stemloop structure connected to

the tail end of the structure. This could be explained by the fact that the found MITEs as represented in the P-MITE database are annotated as partial sequences of MITEs suggesting this sequence is incomplete. However, taking the MITE sequence from the database as query for a BLAST search on the *M.truncatula* genome, hits of partial sequence cover are found. The parts that are not aligned are predominantly at the end, corresponding roughly to location of the additional stemloop at the tail of 3.2.C. Therefore, an alternative explanation is that the MITE sequence as presented by P-MITE is incorrect in the sense that it is *longer* than the actual MITE. Omitting this last structure does provide us a satisfactory palindromic stem corresponding to a typical terminal inverted repeat.
Selecting the structures of the domain 4 resembling sequences in *M.truncatula* on their MITE-like properties did not contribute as much as expected. Many of the known MITEs that were associated with them were up to three times larger. Furthermore, the original structures similar to the conserved stem-loop of *enod40* domain 4 could not be traced in these MITEs. Since they were not maintained within the MITES, these original structures turned out to be less relevant. It could be argued that expanding the range of the queries of domain 4 resembling hits would have resulted in a better selection of MITE-like structures. However, the second purpose of structure selection was finding structural resemblance between the original domain 4 structure with the associated MITE structures. In this context, it does not make sense to increase the range of domain 4 resembling hits over its actual range.

We now know the following keypoints:

1. Based on the initial observations.

   (a) One of the many *C.arietinum enod40* domain 4 similarity hits in *C.arietinum* was similar to a known *M.truncatula* MITE.

   (b) One of the many *C.arietinum enod40* domain 4 similarity hits in *M.truncatula* was similar to two different known *M.truncatula* MITEs.

   (c) One *Lupinus angustifolius enod40* domain 4 showed 334 similarity hits in the *L.angustifolius* genome, but none were similar to known MITEs. *C.arietinum* and *M.truncatula* show no similar sequences to this domain 4.

2. Based on new results.

   (a) *M.sativa enod40* domain 4 has a similar sequence to many sequences in *M.truncatula* that are similar to known MITEs.

   (b) The *M.sativa* domain 4 and MITE resembling hits in *M.truncatula*(2a) are not significantly similar to the MITE resembling the *C.arietinum* hit (1a) or the *C.arietinum* domain 4 and MITE resembling hits in *M.truncatula*(1b).

The data show that in three species of the order *Fabales*, a MITE has been inserted in *enod40*. Following this insertion, the *enod40* RNA has then evolved from this genetic material a functional structural domain 4.
Taking into account the phylogeny of the species, we hypothesize a collection of events have occured, as illustrated in Figure 4.1. We expect that since *M.sativa* domain 4 was similar to *M.truncatula*s MITEs, the *M.sativa* genome would contain the same or very similar MITEs. Sadly, there is no whole genome sequence available for the *M.sativa* to test this. Querying the *M.truncatula* domain 4 itself resulted in a much lower amount of hits(24) compared to querying *M.sativa* domain 4, but three common hits showed strong MITE similarity. Possibly, a progenitor MITE, inserted into *enod40* and, while taking on the form of domain 4, evolved further away from its original sequence in *M.truncatula* than in *M.sativa*.
Not only the *C.arietinum* MITEs are different from the ones found in *M.truncatula*. The *C.arietinum* domain 4 resembling sequences in *M.truncatula* differ from *M.sativa* domain 4 resembling sequences in *M.truncatula* as well. This strongly suggests that the two cases have no hereditary link. One hypothesis is that the two insertions of MITEs into *enod40* have happened independently after which they have acquired structural and functional properties of the contemporary domain 4. They both happen after the branching of *C.arietinum*: one in its progenitor, another in the progenitor of *M.truncatula* and *M.sativa*, affecting both. Even though the 334 similar sequences of domain 4 in *L.angustifolius* could not be confirmed as a MITE, the high number of hits is sufficient to assume an insertion of a MITE at domain 4 has occured at some time in the past. We could also find no relation between the MITE collections of *L.angustifolius* versus those of *C.arietinum* and *M.truncatula*, suggesting again a seperate event.
Remarkably, insertions in the same region have occured in three different progenitors of contemporary species. In each case, the inserted sequence resulted in a conserved stemloop structure. Since the functionality of *enod40* domain 4 has emerged out of a MITE insertion, it appears that transposable elements contribute actively to the evolution of non-coding RNAs.

In order to make a fully informed and certain conclusion about the difference between MITE insertion in *M.sativa* and *M.truncatula*, more sequence data of the *M.sativa* genome is required. Similarly, to confirm the MITE identity of the domain 4 resembling sequences in *L.angustifolius*, there is need of a more extended knowledge of existing MITEs in current databases. Luckily, since the databases are continuosly being reinforced with new sequences and annotations, this is only a matter of time. Alternatively, a seperate study of this collection of sequences could be performed to further test for MITE-like properties or behaviour. If enough confirmation is achieved, an effort could be made to submit an entry and enrich the database actively.
Further study is required to determine the relation between the MITE insertion at *enod40* domain 4 of *L.angustifolius* to those of *M.truncatula* and *C.arietinum*. At this point, we know that domain 4 of *L.angustifolius* shows no similarity in the other two genomes. Furthermore, its domain 4 similarity hits were not recognized as MITEs by databases, in contrast to those of *M.truncatula* and *C.arietinum*. This suggests that the MITEs are not related, but a more satisfactory evidence would be achieved if the *L.angustifolius* MITE collection is

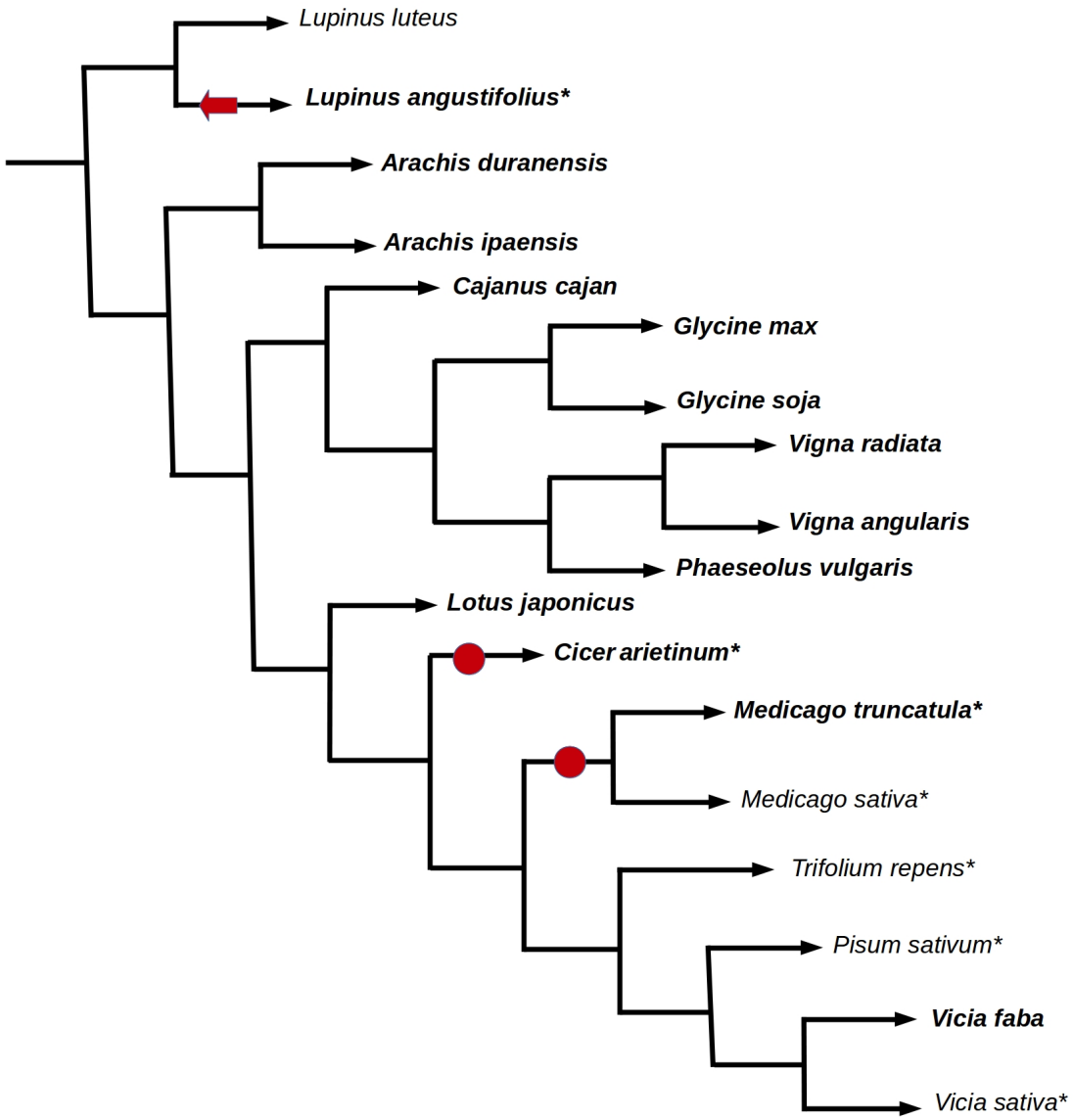

Figure 4.1: Hypothesized order of evolution and MITE insertion.
Red dots indicate MITE insertion. The red arrow signifies the insertion occured anywhere before this point. * denotates the confirmed presence of at least one *enod40* gene that contains domain 4. Bold names indicate that wgs is available for this species at the NCBI database. Figure based on Figures 1-3 of Lavin et al. [LHW05].

cross-aligned with BLAST to the other two collections to exclude similarity.

In order to understand transposable element insertion as evolutionary mechanism fully, a study of its transition from its originial form to a functional non-coding RNA in the host gene would be required. However, all we can do is reconstruct a narrative based on the current genomes and their evolutionary relations. For this, we need more examples of transposable element insertions in closely related species that can be extensively compared. Further research could be spend into identifying any other kinds of transposable elements besides MITEs that can contribute to non-coding RNA evolution. It would also be useful to search for occurrences of transposable element insertion in other non-coding RNA regions with different properties of the resulting non-coding RNA. This could provide more insight in the properties required for a transposable element in order to be able to succesfully insert and evolve into certain non-coding RNAs.

Bibliography

[CHZ+14] Jiongjiong Chen, Qun Hu, Yu Zhang, Chen Lu, and Hanhui Kuang. P-mite: a database for plant miniature inverted-repeat transposable elements. *Nucleic acids research*, 42(D1):D1176–D1181, 2014.

[Cri70] F. Crick. Central Dogma of Molecular Biology. *Nature*, 227:561–563, August 1970.

[FZW02] Cédric Feschotte, Xiaoyu Zhang, and Susan R Wessler. Miniature inverted-repeat transposable elements (mites) and their relationship with established dna transposons. *Mobile DNA II*, pages 1147–1158, 2002.

[GR07] A. P. Gultyaev and A. Roussis. Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants. *Nucleic Acids Res.*, 35(9):3144–3152, 2007.

[GRG+03] G. Girard, A. Roussis, A. P. Gultyaev, C. W. Pleij, and H. P. Spaink. Structural motifs in the RNA encoded by the early nodulation gene enod40 of soybean. *Nucleic Acids Res.*, 31(17):5003–5015, Sep 2003.

[Gro04] Nitrous Oxide Focus Group. Nitrous oxide focus group overview, 2004. [Online; accessed 2016-10-18].

[HZJ+06] Dawn Holligan, Xiaoyu Zhang, Ning Jiang, Ellen J. Pritham, and Susan R. Wessler. The transposable element landscape of the model legume lotus japonicus. *Genetics*, 174(4):2215–2228, 2006.

[KAA+16] Paul Julian Kersey, James E. Allen, Irina Armean, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, Paul Davis, Lee J. Falin, Christoph Grabmueller, Jay Humphrey, Arnaud Kerhornou, Julia Khobova, Naveen K. Aranganathan, Nicholas Langridge, Ernesto Lowy, Mark D. McDowall, Uma Maheswari, Michael Nuhn, Chuang Kee Ong, Bert Overduin, Michael Paulini, Helder Pedro, Emily Perry, Giulietta Spudich, Electra Tapanari, Brandon Walts, Gareth Williams, Marcela TelloRuiz, Joshua Stein, Sharon Wei, Doreen Ware, Daniel M. Bolser, Kevin L.

Howe, Eugene Kulesha, Daniel Lawson, Gareth Maslen, and Daniel M. Staines. Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, 44(D1):D574–D580, 2016.

[Kaz04] H. H. Kazazian. Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–1632, Mar 2004.

[LHW05] Matt Lavin, Patrick S Herendeen, and Martin F Wojciechowski. Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic biology*, 54(4):575–594, 2005.

[Lis13] Damon Lisch. How important are transposons for plant evolution? *Nature Reviews Genetics*, 14(1):49–61, 2013.

[Mat04] J. S. Mattick. RNA regulation: a new genetics? *Nat. Rev. Genet.*, 5(4):316–323, Apr 2004.

[MCG+16] Florence Mus, Matthew B Crook, Kevin Garcia, Amaya Garcia Costas, Barney A Geddes, Evangelia-Diamanto Kouri, Ponraj Paramasivan, Min-Hyung Ryu, Giles ED Oldroyd, Philip S Poole, et al. Symbiotic nitrogen fixation and challenges to extending it to non-legumes. *Applied and environmental microbiology*, pages AEM–01055, 2016.

[MMT10] D. H. Mathews, W. N. Moss, and D. H. Turner. Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol*, 2(12):a003665, Dec 2010.

[OB04] S. Ouyang and C. R. Buell. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.*, 32(Database issue):D360–363, Jan 2004.

[PZF+09] J. Podkowinski, A. Zmienko, B. Florek, P. Wojciechowski, A. Rybarczyk, J. Wrzesinski, J. Ciesiolka, J. Blazewicz, A. Kondorosi, M. Crespi, and A. Legocki. Translational and structural analysis of the shortest legume ENOD40 gene in Lupinus luteus. *Acta Biochim. Pol.*, 56(1):89–102, 2009.

[RDRP16] N. A. Rayan, R. C. Del Rosario, and S. Prabhakar. Massive contribution of transposable elements to mammalian regulatory sequences. *Semin. Cell Dev. Biol.*, May 2016.

[RSM+02] H. Rohrig, J. Schmidt, E. Miklashevichs, J. Schell, and M. John. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc. Natl. Acad. Sci. U.S.A.*, 99(4):1915–1920, Feb 2002.

[Rut03] T. Ruttink. Enod40 affects phytohormone cross-talk. *PhD Thesis*, 2003. ISBN:9058089797.

[SBF13] Carole Santi, Didier Bogusz, and Claudine Franche. Biological nitrogen fixation in non-legume plants. *Annals of botany*, 111(5):743–767, 2013.

[SJC+01] C. Sousa, C. Johansson, C. Charon, H. Manyani, C. Sautter, A. Kondorosi, and M. 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. *Mol. Cell. Biol.*, 21(1):354–366, Jan 2001.

[Stooo] Paul Stothard. The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *Biotechniques*, 28(6):1102–1104, 2000.

[SWBP01] Bruce A. Shapiro, Jin Chu Wu, David Bengali, and Mark J. Potts. The massively parallel genetic algorithm for rna folding: Mimd implementation and population variation. *Bioinformatics*, 17(2):137–148, 2001.

[SWF+09] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, et al. The b73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115, 2009.

[TDB+13] T Tatusova, M DiCuccio, A Badretdin, V Chetvernin, S Ciufo, and W Li. The ncbi handbook. 2013.

[Zuko3] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, Jul 2003.