Computational identification of three-way junctions in folded RNAs: a case study in *Arabidopsis*

Adaya Cohen[1#], Samuel Bocobza[2,3#], Isana Veksler[1], Idan Gabdank[1], Danny Barash[1], Asaph Aharoni[3], Michal Shapira[2] and Klara Kedem[1,*]

[1] Department of Computer Science, Ben-Gurion University, Beer-Sheva 84105, Israel
[2] Department of Life Sciences, Ben-Gurion University, Beer-Sheva 84105, Israel
[3] Department of Plant Sciences, Weizmann Institute of Science, Rehovot 76100, Israel

[#] AC and SB contributed equally to this work

* Corresponding author
Email: klara@cs.bgu.ac.il

Edited by H. Michael; received April 25, 2007; revised and accepted January 28, 2008; published March 12, 2008

Abstract

Three-way junctions in folded RNAs have been investigated both experimentally and computationally. The interest in their analysis stems from the fact that they have significantly been found to possess a functional role. In recent work, three-way junctions have been categorized into families depending on the relative lengths of the segments linking the three helices. Here, based on ideas originating from computational geometry, an algorithm is proposed for detecting three-way junctions in data sets of genes that are related to a metabolic pathway of interest. In its current implementation, the algorithm relies on a moving window that performs energy minimization folding predictions, and is demonstrated on a set of genes that are involved in purine metabolism in plants. The pattern matching algorithm can be extended to other organisms and other metabolic cycles of interest in which three-way junctions have been or will be discovered to play an important role. In the test case presented here with, the computational prediction of a three-way junction in *Arabidopsis* that was speculated to have an interesting functional role is verified experimentally.

Keywords: three-way junctions, folding prediction by energy minimization

Introduction

The RNA molecule, which was once considered to serve only as an intermediate between DNA and proteins, is attracting major attention in the recent decade. Studies focusing on non-coding and small RNAs, as well as understanding the structural details of assemblies in various ribosomes and ribozymes, have led to some interesting discoveries of RNA constructs that possess diverse functional roles. Some structures of folded RNAs were found to have favorable properties that can facilitate tertiary interactions with other biomolecules. One such example is the structure of 3-way junctions; these can form the catalytic core of the hammerhead ribozyme, as well as make up the adenine- and guanine-sensing domains in purine riboswitches. They consist of a multibranch loop linking three stems. More comprehensive information about their topology and characterization is available in [Lescoute and Westhof, 2006].

The focus of our current study is on devising an algorithm capable of filtering out 3-way junction patterns given a set of genes. In general, if an arbitrary data set is used, all types of 3-way junctions can be found. If, however, the given data set is restricted to particular genes, for example those involved in purine metabolism, a hammerhead ribozyme-

like sequence most likely will not appear. Moreover, only a few 3-way junction patterns will emerge, and they can be checked by a manual inspection of each one to determine whether their sequence is of particular interest. Thus, such a strategy can be used to try and find traces of adenine- and guanine-sensing domains of purine riboswitches. As a test case for demonstrating our proposed approach, we take the aforementioned task of trying to find a purine riboswitch in eukaryotes.

RNA genetic control elements known as riboswitches were discovered in bacteria [Mironov *et al.*, 2002; Winkler *et al.*, 2002]. These are highly structured domains within mRNAs that precisely sense and bind metabolites, resulting in structural alterations that serve as a basis for the control of gene expression. Riboswitches are typically composed of two functional domains [Winkler and Breaker, 2003; Nudler and Mironov, 2004]: an aptamer [Ellington and Szostak, 1990; Tuerk and Gold, 1990; Hermann and Patel, 2000] that selectively binds its target metabolites, and an expression platform [Winkler and Breaker, 2003] that responds to the metabolite binding and controls gene expression by allosteric means. The aptamer domain is well-conserved, whereas the expression platform can vary widely in both its sequence and secondary structure. Riboswitches regulate several metabolic pathways including the biosynthesis of vitamins and the metabolism of methionine, lysine and purine [Mandal and Breaker, 2004; Nudler and Mironov, 2004; Vitreschak *et al.*, 2004]. Riboswitches have been found experimentally in prokaryotes at the level of transcription termination [Mironov *et al.*, 2002] and translation initiation [Winkler *et al.*, 2002a]. The thiamine pyrophosphate (TPP) riboswitch was also discovered in eukaryotes [Kubodera *et al.*, 2003; Sudarsan *et al.*, 2003a; Sudarsan *et al.*, 2005]. It was identified using a clever and also a relatively simple bioinformatics approach from the algorithmic standpoint, described in [Sudarsan *et al.*, 2003b], which mostly relies on sequence conservation.

Many interesting RNAs have a common secondary structure without sharing a significant sequence similarity [Durbin *et al.*, 1998]. Thus, searching for RNA motifs by sequence alone may miss important findings. There are several tools to search for RNA motifs, based mainly on sequence similarities with a minor emphasis on structure, such as the SequenceSniffer used in [Sudarsan *et al.*, 2003b] or programs that incorporate information about numbers/lengths of stems/loops such as the RNA-Pattern used in [Rodionov *et al.*, 2003]. There are more extensive search methods with structure considerations such as the RNAMotif [Macke *et al.*, 2001], FastR (http://bioinf.ucsd.edu/~bchaas/fastr/) [Zhang *et al.*, 2005], RNAProfile [Pavesi *et al.*, 2004], RSEARCH [Klein and Eddy, 2003], and the STR^2 search [Bergig *et al.*, 2004] developed in our group.

In our case study in *Arabidopsis*, in addition to applying our proposed approach that has been much improved and extended in similarity scores and statistical estimates since the preliminary version was put forth in [Bergig *et al.*, 2004], we also examined an alternative search strategy. It is a combined search composed of several methods, for the purpose of comparing and considering other search methods that can be applied in tandem. As a preliminary work, we compared various methods and integrated some in a hierarchical manner. For example, methods such as FastR and RNAMotif can work on inputs that consist of a whole genome, whereas RNAProfile works best as a post processing step. Subsequently, our proposed approach of scanning genes for traces of 3-way junctions allowed us to filter out candidates based on pure structural considerations, and then check for sequence similarity. Our methods of comparison in the combined search strategy and validation in our proposed approach were performed on prokaryotic data, in which riboswitches have been found [Mandal *et al.*, 2003]. We then applied it on data we collected from fungi and plants, with the goal of finding new riboswitches in eukaryotes using these methodologies. We identified an interesting case study in *Arabidopsis* that was speculated to have a structural resemblance to the bacterial guanine riboswitch and we concluded the case study by checking its structure experimentally. It should be noted that when searching in distant relatives in eukaryotes based on verified box motifs in bacteria, the secondary structure of common motifs is expected to play a more prominent role relative to sequence. Therefore, methods that mainly consist of sequence information are probably not enough to discover novel riboswitches in higher organisms, whereas our searches were meant to overcome this limitation.

Materials and methods

Overview

In this work we focus on five established bioinformatics methods that were previously tested by searching for known small RNAs. Below, we summarize the conceptual basis for each one. We have attempted to combine the first four methods described below to operate in tandem. In parallel we extended and considerably improved the fifth method,

our Structure to String (STR^2) [Bergig *et al.*, 2004].

FastR [Zhang *et al.*, 2005; 2006] is a method that takes as input a query which is an RNA sequence with a known secondary structure, and a database which consists of target sequences. It computes and outputs all the structural homologues. The method is based on structural filters that eliminate a large portion of the database. Filters are based on the basic *(k-w)bp-unit* filter defined in [Zhang *et al.*, 2005] and meet four criteria: *generality*, *efficiency*, *sensitivity* and *specificity*. This algorithm works in three stages; first, it filters the whole database using the *(k-w)bp-unit* filter. Second, it aligns the filtered substrings to the query. Third, it excludes hits that are likely to be obtained by chance, using P-value as a screen. The search can be performed on a large scale database such as a whole genome.

RNAMotif [Macke *et al.*, 2001] is a method that takes as input a database of nucleotide sequences and a constraint description file that consists of four sections called *parameters*, *descriptor*, *sites* and *score*. A structural pattern defined in a *descriptor* uses a special pattern language. In addition, one can provide a *score* section that allows the user to rank imperfect matches to desired sequence/structural elements. This type of information gathering is motivated by the goal of developing a computer program that can describe an RNA structural element of any complexity and then search any nucleotide sequence database. The algorithm outputs candidates that meet the conditions of the *descriptor* and the scoring scheme. RNAMotif uses a two-stage algorithm to perform motif searches. The first stage is a compilation stage that analyzes the specified motif, the *descriptor*, and converts it into a search tree based on the helical nesting of the motif. The second stage is a depth first search of the tree that was created by the compilation stage. Each time a complete solution to the original descriptor is found, the candidate is passed to an optional *score* section for scoring and ranking. If the *score* section is absent then the candidate is automatically accepted. The search can be performed on a large scale database such as a whole genome.

RNAProfile [Pavesi *et al.*, 2004] is a method that takes as input a parameter, which denotes the number of distinct hairpins a motif has to contain, and a set of unaligned RNA sequences expected to share a common motif. It outputs the regions that are most conserved throughout the sequences, according to a similarity measure that takes into account both the sequence of the regions and the secondary structure they can form according to base-pairing and thermodynamic rules. The algorithm works in two phases; first, it extracts from each input sequence a set of candidate regions whose predicted optimal secondary structure contains the number of hairpins given as input. Second, the regions selected are compared with each other to find the groups of most similar ones, formed by a region taken from each sequence. This method is recommended as a post-processing step for the output generated by other motif finding programs.

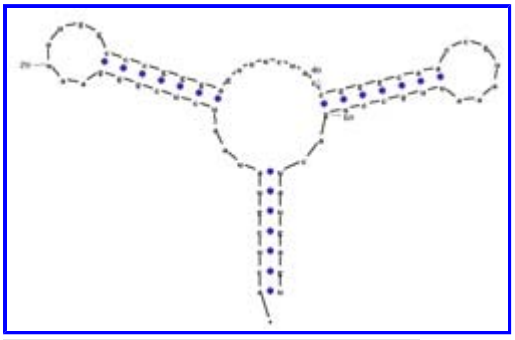
SeqSniff_Imitate is a program written by our group that conceptually imitates Breaker and coworkers' SequenceSniffer, as described in [Sudarsan *et al.*, 2003b; Barrick *et al.*, 2004]. The SequenceSniffer takes as input a pattern that consists of sequence information, variable gaps, and base pairing brackets representing Watson-Crick base pairs and a strong non-canonical base pair. It outputs candidate locations of the sequence that meet constraints imposed by the method. The algorithm uses simple matching, based on the constraints, as opposed to dynamic programming. Thus, the search can easily be performed on a large scale database such as a whole genome.

STR^2 [Bergig *et al.*, 2004] is a method, developed in our group, that takes as input a query sequence and a set of target sequences, and outputs candidates (sub sequences) such that the predicted structure of the query is similar to the predicted structure of the candidate. The idea here is to convert a secondary structure to shape representing characters, after which known sequence-based search algorithms with an efficient implementation can be used. The STR^2 transforms the problem of structure similarity to *inexact string matching*. The algorithm works in three phases; first, it folds the target sequence in windows of small lengths using mfold [Zuker, 2003], where suboptimal foldings [Zuker, 1989] are also taken into account. Second, it converts the structures obtained into a shape representing character string. Third, it perceives the strings as a problem of *inexact string matching*, and solves it by known search algorithms for this problem. The search, if performed on a large scale database such as a whole genome, will be computationally expensive. Thus, it is recommended to apply this method after filtering the genome with other methods or tools that perform an initial scan.

In addition, **RSEARCH** [Klein and Eddy, 2003] and the **CMFinder** (covariance model finder; http://bio.cs.washington.edu/yzizhen/CMfinder/) [Yao *et al.*, 2006] are more expensive search methods that utilize Hidden Markov Models (HMMs) and Stochastic Context Free Grammar (SCFG). They are well-established methods and the latter was used successfully in [Mandal *et al.*, 2004].

Method comparison

For the comparison, the database that is used in this work is that of the *Bacillus halodurans* genome (NCBI accession number BA000004). The tests described below were performed on the purine riboswitch (G-Box) known structure and members taken from RFAM [Griffiths-Jones *et al.*, 2003]. We used the consensus for the purine riboswitch available in RFAM [Griffiths-Jones *et al.*, 2003] as the query sequence that consists of 69 nt (See Fig. 1).



Click on the thumbnail to enlarge the picture

Figure 1: Consensus of purine riboswitch aptamer taken from RFAM [Griffiths-Jones *et al.*, 2003], the folding is predicted by mfold [Zuker, 2003] (same prediction of the Vienna package [Hofacker *et al.*, 1994] as available in RFAM). The consensus shares sequence and structure conservation among numerous bacterial types.

Datasets

Three types of datasets were used for examining the methods: (1) a **genomic** dataset, full genome sequence of *Bacillus halodurans* (4202353 nt); (2) a **noisy** dataset consisting of 59 upstream fragments of 500 nt each, taken from different genes, including their corresponding 5'-UTRs. It contains 26 "noise" genes of which most encode for ribosomal proteins, and 33 genes that are involved in purine metabolism (including the reported locations taken from RFAM); (3) a dataset of **purine metabolism genes** that consists of 33 5'-UTRs of 500 nt each, taken from upstream regions of purine metabolism genes (including the reported locations taken from RFAM).

Our goal was to examine the relative performance of the methods described above. For comparison we focused on two measures: **Sensitivity**, which is the fraction of the true matches that are actually predicted by the method, and **Positive Predictive Value** (PPV), which is the fraction of the sequences predicted as matches that are indeed true matches. We used the following definitions for the calculations: True Positives ($\boldsymbol{TP}$) are homologous findings (found in RFAM database) and are considered hits; False Positives ($\boldsymbol{FP}$) are hits that are non-homologous; False Negatives ($\boldsymbol{FN}$) are homologous, but not hits. We can now define Sensitivity = $TP/(TP + FN)$, and the Positive Predictive Value, $PPV = TP/(TP + FP)$. One often combines both these measures into Receiver Operating Characteristic curves (ROC curves), but because of the small portion of True Positives (TPs) the ROC curves are less likely to serve as good indicators in our case. The comparison results are shown in Tab. 1, where each method uses a different input as follows: **RNAMotif** was executed using a descriptor file based on the structure of the query; **SeqSniff_Imitate** and **RSEARCH** were executed using a query as described in [Mandal *et al.*, 2003]; **CMFinder** was executed with default parameters as they appear in the CMfinder webserver [Yao *et al.*, 2006] with the change of "number of stem-loops=2" and "minimum length of motif=15"; **FastR** used seeds taken from RFAM, each of length ~100 nt; **STR**^{2} was executed using the sequence of the query; and **RNAProfile** was used in iteration mode on dataset 3, candidates considered with fitness $\geqslant 0$. In examining Tab. 1, it should be noted that a low value for sensitivity indicates that many genuine riboswitches are missed, whereas a low value for PPV indicates that many false alarms are being inappropriately categorized as a riboswitch.

Table 1: Methods comparison on *Bacillus halodurans*.

Method	Dataset	Sensitivity	PPV	hits/TP/FP	False Negatives	Actual riboswitches
RNAMotif	1	0.4	0.285	7/ 2/ 5	3	5
RNAMotif	2	0.6	1.0	3/ 3/ 0	2	5
RNAMotif	3	0.6	1.0	3/ 3/ 0	2	5
FastR[3]	1	1.0	0.71	7/ 5/ 2	0	5
SeqSniff_Im	1	0.8	1.0	4/ 4/ 0	1	5

SeqSniff_Im	2	1.0	1.0	5/ 5/ 0	0	5
SeqSniff_Im	3	1.0	1.0	5/ 5/ 0	0	5
RSEARCH	2	1.0	1.0	5/ 5/ 0	0	5
RSEARCH	3	1.0	1.0	5/ 5/ 0	0	5
CMFinder	2	0.6	0.056	53/3/50	2	5
CMfinder	3	1.0	0.625	8/ 5/ 3	0	5
RNAProfile	3	0.8	0.19	21/ 4/ 17	1	5
RNAProfile	4*	0.6	1.0	3/ 3/ 0	2	5

* Dataset 4 is described in the following section

Both RNAMotif and SeqSniff_Imitate fail to find the complementary matches on the genomic dataset. However, when applying the methods on the noisy dataset and the dataset of purine metabolism genes, where complementary sequences were available, both succeeded to locate the missing match. CMfinder, on the other hand, is less successful in the noisy dataset. It finds correctly a motif that consists of one stem-loop only, which is only part of the riboswitch motif. Within its hits it reports 3 True Positives (TPs) and 50 False Positives (FPs). However, it should be noted that there can be numerous two stem-loops in a dataset containing arbitrary sequences and therefore the results reported for CMfinder in the noisy dataset do not reflect its high capabilities and potential to find novel riboswitches as is evident in [Mandal *et al.*, 2004]. Admittedly, the comparison is performed specifically for the purine riboswitch and not for other riboswitches, which is limited in scope, the reason being that **STR**^{2} can only address at present motifs that are well predicted by energy minimization methods. RSEARCH is successful in its performance in our comparison as reported in Tab. 1. However, neither RSEARCH nor any other of the methods being compared to **STR**^{2} succeeded to find an attractive purine riboswitch candidate in eukaryotes that is worthwhile investing on by following with a biochemical experiment, whereas our STR^2 found such an attractive potential candidate as reported in the Results section.

Integration of three methods

In our first search strategy, we integrated FastR, RNAMotif and SeqSniff_Imitate. The strategy is based on the assumption that the intersection of all the methods output (hits) would include the most robust candidates, by using both sequence and structure considerations. Following the union we could apply RNAProfile to construct a profile of the candidates.

The **merge (4)** dataset consisted of the output sequences (hits) of RNAMotif, FastR and SeqSniff_Imitate. This dataset mostly contained sequences of length 50-80 nt. The candidates were filtered by location: coding sequences (CDS) areas or sequences that were not related to purine metabolism genes were excluded. We presumed that sequences (candidates) that do not meet the structure of the query are most likely to be excluded from the best profiles, or at least have a negative fitness score. An input file for RNAProfile was created from these sequences, including the original query sequence and predicted structure, as an anchor. RNAProfile was used in an "all versus all" mode on dataset 4. The hits given by RNAProfile were then filtered by their fitness score. The final output, candidates with fitness = 0, showed a sensitivity of 0.6 and a PPV of 1.0 (see Tab. 1).

The integration was also tested on a different genome: the *Oceanobacillus iheyensis* HTE831 genome (NCBI accession number BA000028, 3630528 nt). Tests were conducted on the purine riboswitch known structures and members, taken from RFAM [Griffiths-Jones *et al.*, 2003]. RNAProfile was executed on the **merge** dataset, as described above. The output, candidates with fitness $\geqslant$ 0, showed a sensitivity of 1.0 and a PPV of 0.8.

In-line probing

In-line probing experiments were conducted as previously described [Mandal *et al.*, 2003]. The DNA template was obtained by PCR and the RNA fragment was transcribed *in vitro* using T7 RNA polymerase. The RNA was dephosphorylated and end labeled at the 5' end with T4 polynucleotide kinase. The in-line probing assay was performed by incubating the labeled RNA during 40 h at 25°C in a buffer containing 50 mM Tris-HCl (pH = 8.3 at 25° C), 20 mM $MgCl_2$, 100 mM KCl. Reactions were performed with increasing concentrations of guanine that ranged between 0 to 330 μM. The partially cleaved fragments were resolved on a denaturing PAGE (8% polyacrylamide - 8 M

urea gels) and the gels were subjected to a phosphorimager analysis.

Results

STR^2 search in prokaryotes

The pioneering approach of STR^2 in [Bergig *et al.*, 2004] has been considerably extended and adapted for our needs. In the following, we describe the improvements that were added to STR^2 for RNAs, and their validation on prokaryotes.

Automation of STR^2

The extended version was automated to allow the user to search a large data set. The automation enables the user to input a whole genome in a GenBank format and extract from it specific segments (UTRs) of required genes, based on the name and product of the gene. The automation works in several phases; first, the genome is filtered by extracting segments of interest from the genome to create the database. Second, a pre-processing scheme is used. From each sequence of the database, subsequences are created, using a sliding window of a specific length (usually the length of the query). Subsequently, the structures of the subsequences are determined using mfold [Zuker, 2003], are converted into shape representing strings, and a file with a string representation of the structure (str file) of each is produced. As an alternative to the shape representing strings, Shapiro's tree-graphs [Shapiro, 1988] can be used. Third, the main routine of the program called **Protrace** is applied on the str files database to search for similar structure. Fourth, post-processing of **Protrace** output is applied. The output is first subjected to statistical estimation, as described in the Statistical estimation Subsection below, and candidates with z-value below a given threshold are excluded. The preprocessed output is then displayed on an HTML file. Each phase is a stand-alone process. The search can be applied as a pipeline or as separate phases to enable flexibility and efficiency.

Refining the similarity scores

In the previous version of STR^2 [Bergig *et al.*, 2004], the similarity scores were decided according to the minimum RMS distance between the fragments associated with the characters, whereas in the current version we added knowledge based insight to the scores. The STR^2 method is based on analyzing secondary structure drawings. A nucleotide that is represented by a linear secondary structure is most likely to be in a double strand, while a nucleotide represented by a curved structure is most likely to be in a single strand. Thus, replacing one of the structures by the other, although sharing similar appearance, can be erroneous. In our current improved version this results with high penalty: the similarity score penalty is the highest possible (10.0).

Statistical estimates

For the STR^2, it was decided to use z-values as our estimate. In order to evaluate the constant values of the equation we used Pearson's **regress1** method as described in [Pearson, 1998]. In our case, as in FASTA, the matches are of different lengths and the random database could be biased. We used 50 sequences of 500 nt each and conducted the search with 120 random query sequences. We obtained scores of different match lengths, and applied the **regress1** method resulting with the coefficients for calculating the z-value of each length.

Validation of STR^2

We validated the performance of this method by using the "noisy" dataset and query as described in the Datasets section. We carried out a test on *Bacillus halodurans*, and considered only output with z-value $\geqslant 8$ as candidates. The sensitivity obtained was 0.6 and the PPV was 1.0. The results are shown in Tab. 2.

Table 2: STR^2 output on *Bacillus halodurans* dataset 2

Position (nts)	Source CDS	z-value	TP/FP
676492-676558	purE(BH0623)	9.2	TP
650328-650393	BH0608	9.17	TP
648459-648526	guaA(BH0607)	8.7	TP

A similar test was carried out with *Oceanobacillus iheyensis* HTE831 genome (NCBI accession number BA000028). Once again we considered the output with z-value $\geqslant 8$ as candidates. The sensitivity obtained was 0.75 and the PPV was 1.0. The results are shown in Tab. 3.

Table 3: STR^2 output on *Oceanobacillus iheyensis* dataset 3

Position (nts)	Source CDS	z-value	TP/FP
786783-786851	OB0739	9.82	TP
760490-760558	guaA	9.16	TP
1103959-1104025	OB1061	9.2	TP

In STR^2, only sequence information is required as an input. Consequently, using the automation makes the processing of a large database relatively efficient. The method is based on the fact that the structure of interest is stable enough and assumes that energy minimization methods [Zuker, 1989; 2003; Hofacker *et al.*, 2004] predict the structure fairly accurately. Once the correct structures are generated, STR^2 is proved to be highly sensitive. Furthermore, if the correct threshold is used for the z-value, it shows high PPV as well.

STR^2 search in eukaryotes

Before applying the different methods that were used to eukaryotes, several adaptations were made. As mentioned earlier, sequence conservation is most likely to be abandoned and thus SeqSniff_Imitate's query must be less strict. With RNAMotif, we needed to add a filtering scoring scheme on the sequence. Results from FastR were not publicly available for eukaryotes and thus were not included. As for the STR^2, the search remained the same, with no modifications.

A search in *Saccharomyces cerevisiae*

Using the merge strategy, the search was conducted on the complete genome of *Saccharomyces cerevisiae* (NCBI accession numbers are specified in the Appendix, Table A). The findings were as follows. SeqSniff_Imitate's output gave 11 candidates, and RNAMotif output gave 15 candidates. None of these candidates were related to a purine metabolism gene.

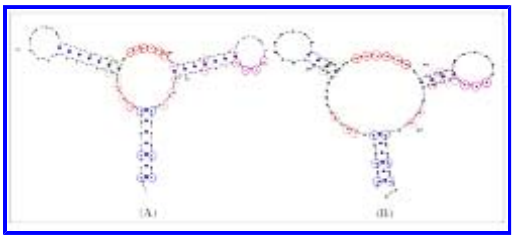
Using the STR^2, the search was performed on both the 5' and 3' UTR regions of the genes (details are specified in the Appendix, Table C). STR^2 gave 6 candidates as output. These candidates did not show strong structure similarity, and had almost no sequence similarity.

A search in the model plant *Arabidopsis thaliana*

Using the merge strategy, the search was conducted on the complete genome of *Arabidopsis thaliana* (NCBI accession numbers are specified in the Appendix, Table B). Similar to the yeast findings, an intersection of methods gave two candidates but they were not related to a purine metabolism gene.

Using STR^2 the database contained both 5' and 3' UTR regions of purine *de novo synthesis* gene list, taken from the *Arabidopsis* Functional Genomics Network (AFGN) website (http://www.uni-tuebingen.de/plantphys/AFGN/boldt.htm). The STR^2 output contained several 3-way junction structures that resemble the consensus guanine riboswitch aptamer. We examined the 3-way junction candidates in terms of sequence similarity and identified a potential for a

guanine riboswitch aptamer located downstream to the *Arabidopsis At2g05140* gene coding region (Fig. 2). The functionality of the STR[2] was tested by using "random" genes that served as a negative control. For this purpose we selected fifty genes that are not related to purine metabolism and conducted a similar search using these genes as our database. As output we obtained 5 candidates, however none of them showed strong structure similarity or any sequence homology to the consensus. Thus, our 3-way junction finding (Fig. 2) that share common nucleotides with the consensus in its core is a case study that was found interesting for further experimental verification.

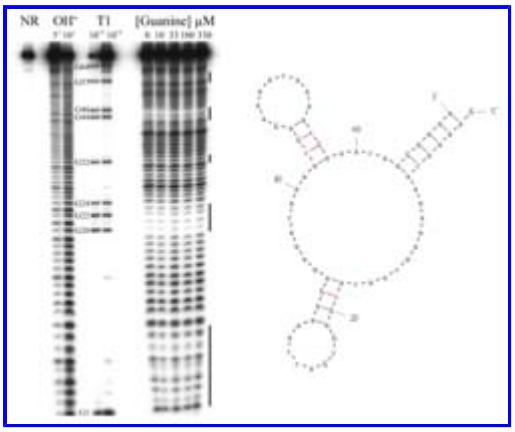


Click on the thumbnail to enlarge the picture

Figure 2: (A) The consensus purine riboswitch as available in RFAM. (B) Our 3-way junction prediction finding in the 3' region of the *Arabidopsis At2g05140* gene putatively encoding a phosphoribosylaminoimidazole carboxylase. The locations of sequence similarity between the consensus and our 3-way junction finding are marked with colored circles. All four nucleotides that participate in the hypoxanthine binding [Batey *et al.*, 2004; Holbrook, 2005; Lescoute and Westhof, 2005] are present in our 3-way junction finding: U22 (denoted by t8), U47 (denoted by t34), U51 (denoted by t38), and C74 (denoted by c60). Most of the common nucleotides appear in the core of the 3-way junction.

Experimental examination of the candidate

To verify that the output found downstream of the *Arabidopsis At2g05140* gene coding region indeed folded into a 3-way junction as predicted, we subjected the corresponding RNA fragment that was synthesized *in vitro* to in-line probing analysis [Sudarsan *et al.*, 2003; Winkler *et al.*, 2002]. Using this methodology, the end-labeled RNA fragment is cleaved more efficiently in single-stranded regions, whereas double-stranded stems are more resistant to the self-cleavage that occurs under mild basic conditions. Although the results of the in-line probing experiment verify that the region identified by STR[2] indeed folds into a 3-way junction as predicted, increased concentrations of guanine did not result in structural changes (Fig. 3).

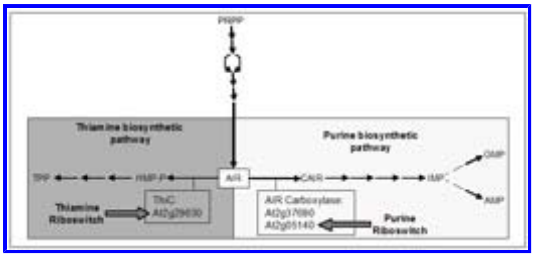


Click on the thumbnail to enlarge the picture

Figure 3: Structural verification of the 3-stem loop junction candidate by in-line probing. RNA was synthesized *in vitro*, end labeled and gel purified. The radiolabeled RNA was subjected to in-line probing by incubation in mild-basic conditions at 25°C for 40 h, in the presence of increasing guanine concentrations: 0, 10, 33, 100, and 330 µM (see Material and methods for details). The partially cleaved fragments were resolved on a denaturing gel (8 M urea - 8% polyacrylamide), which was further subjected to a phosphorimager analysis. Double stranded regions are marked by vertical lines. NR represents non reacted RNA. Lanes designated OH- contain RNA that was first denatured and then subjected to partial alkaline digestion for 5 and 10 minutes. Lanes designated T1 contain RNA subjected to partial digestion with RNAse T1 (G-specific cleavage) that was diluted to contain 10^{-2} and 10^{-3} units per reaction.

The predicted 3-way junction is situated in a putative 5'phosphoribosyl-aminoimidasol carboxylase (AIRC) pseudogene

The plant *de novo* pathway for the synthesis of IMP (the precursor for purine metabolism) involves ten enzymatic steps [Stasolla *et al.*, 2003] (Fig. 4). Although this pathway is very similar among all organisms, the enzymatic steps differ between prokaryotes and eukaryotes [Henikoff, 1987]. In the sixth step of this pathway, 5'-phosphoribosyl-5-aminoimidazole (AIR) is converted by AIR carboxylase (AIRC) into 5'-phosphoribosyl-4-carboxy-5-aminoimidazole carboxylate (CAIR; Fig. 4). In eubacteria, AIR carboxylase consists of two different enzymes encoded by *purE* and *purK* genes [Ebbole and Zalkin, 1987]. In *Bacillus*, the purine metabolic pathway is controlled by the purine riboswitch located in the 5'-UTR of the *xpt-pbuX* operon [Mandal and Breaker, 2004]. In yeast purE and purK homologous domains are fused into a single polypeptide (purKE) [Szankasi *et al.*, 1988].



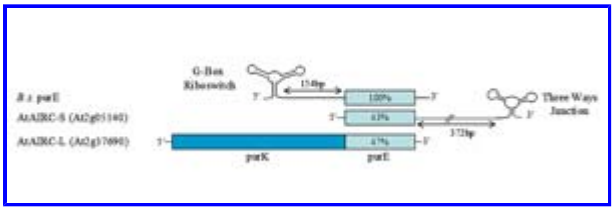
Click on the thumbnail to enlarge the picture

Figure 4: Thiamine and purine metabolism share AIR as a common precursor and both pathways encompass a putative riboswitch. PRPP: 5-phospho-ribosyl-pyrophosphate; AIR: 5'-phosphoribosyl-5-aminoimidazole; CAIR: 5'-phosphoribosyl-4-carboxy-5-aminoimidazole; IMP: 5'-inosine monophosphate; AMP: 5'-adenosine mono-phosphate; GMP: 5'-guanosine-monophosphate; HMP-P: 4-amino-2-methyl-5-phosphomethylpyrimidine; TPP: thiamine pyrophosphate; ThiC: thiamine C synthase.

In the plant *Vigna aconitifolia*, reference [Chapman *et al.*, 1994] identified a single gene coding for AIRC, which is able to complement a *purE* deficient *E. coli*. The authors suggested that the AIRC enzyme in plants is derived from a fusion of ancestral *purE* and *purK* enzymes. Moreover, the order of the K- and E-domains in the purKE fusion protein in *Vigna aconitifolia* is conserved in yeast but does not coincide with the order of the separate *purE* and *purK* cistrons in the *pur* operon of *B. subtilis* or with the dicistronic *purEK* operon of *E. coli*. It was therefore suggested [Chapman *et al.*, 1994; Chung *et al.*, 1996] that an inversion from *purE* and *purK* gene order (present in eubacteria) to *purK* and *purE* gene order (in eukaryotes) may have been a prerequisite to their fusion, imposed by structural constraints on the tertiary structure of the fusion protein.

In *Arabidopsis*, we have identified a putative 642 amino acid long AIRC protein (AtAIRC-L, *Arabidopsis thaliana* AIR carboxylase-long; At2g37690) that retains 67% overall identity to the *Vigna aconitifolia* AIRC protein. Publicly available microarray gene expression data (www.genevestigator.ethz.ch) revealed that it is expressed in multiple tissues, mainly in roots, radicles, juvenile leaves, seeds, carpels, shoot apex, cell suspension and callus. It is therefore conceivable that the AtAIRC-L is the orthologues protein of *Vigna aconitifolia* AIR carboxylase in *Arabidopsis* and both might resemble a fusion protein of the ancestral purK and purE enzymes.

The putative 3 way junction we have identified in the course of this study is situated in a second and much shorter *Arabidopsis* AIRC-like gene [*AtAIRC-Short* (S); *At2g05140*]. It encodes a 162 residues long polypeptide that is similar to the *purE* protein identified in *B. subtilis* (43% identity). The putative protein encoded by AtAIRC-S shows 82% identity with the C-terminus of the *AtAIRC-L* gene. The potential riboswitch in the 3' region of *AtAIRC-S* is located 372 bp downstream from its stop codon (Fig. 5). We could not detect expression of the *AtAIRC-S* gene in *Arabidopsis* (in RT-PCR experiments) and could neither identify any Expressed Sequence Tag (EST) sequence that corresponds to this gene. In contrast to the *AtAIRC-L* gene, there is no evidence for significant levels of expression of *AtAIRC-S* in specific tissues or under different conditions in publicly available microarray gene expression data (www.genevestigator.ethz.ch).



Click on the thumbnail to enlarge the picture

Figure 5: A potential riboswitch located downstream of the *Arabidopsis AtAIRC-S* gene stop codon. Amino acid sequence identities between the purE-domains are noted relatively to *B. subtilis purE* protein. The purine riboswitch in *B. subtilis purE* is located 154bp upstream from the start codon while the putative riboswitch in *Arabidopsis thaliana* AIR carboxylase-short (*AtAIRC-S*) is located 372bp downstream from the stop codon. *Bs*, *Bacillus subtilis*; *AtAIRC-L*, *Arabidopsis thaliana* AIR carboxylase-long.

Moreover, public database annotation of the *Arabidopsis* genome (www.arabidopsis.org) does not indicate a putative 3' UTR in the *AtAIRC-S* gene and it was therefore not possible to determine whether the 3-way junction that we have identified is part of the predicted transcript. In attempt to identify the 3' UTR region of *AtAIRC-S* we performed 3' Rapid Amplification of cDNA Ends (RACE)-PCR experiments using cDNA derived from ten days old seedlings grown under normal conditions. However, in line with our inability to monitor the corresponding transcript (see above) the 3' RACE assay did not generate any product using *AtAIRC-S* specific oligonucleotides. In contrast to the *AtAIRC-S* 3' region , we could amplify the 3' end of the *AtAIRC-L* gene. These results raise the possibility that *AtAIRC-S* might be

a pseudogene.

Discussion

Purine nucleotides participate in many essential biochemical processes in plants from nucleic acids synthesis to energy production [Zrenner *et al.*, 2006]. They also serve as precursors for the synthesis of primary metabolites (e. g. sucrose and phospholipids) and as cofactors in multiple reactions. The *de novo* biosynthetic pathway of purine nucleotides, AMP and GMP, from PRPP (5-phospho-ribosyl-pyrophosphate) has been established in microorganisms, animals and plants [Henderson and Paterson, 1973; Neuhard *et al.*, 1987; Zalkin and Nygaard, 1996; Zrenner *et al.*, 2006].

The putative 3-stem loop junction we have identified in the course of this study is situated in a gene putatively catalyzing the conversion of AIR to CAIR. AIR also serves as a substrate to the thiamine C synthase enzyme (ThiC) that is part of pathway leading to the formation of the B-class vitamin thiamine. Most interestingly, both bacterial as well as plant *ThiC* genes contain a thiamin riboswitch [Sudarsan *et al.*, 2003a]. The riboswitch element located in the 3' end of the *Arabidopsis ThiC* gene (*At2g29630*) has been shown to be active *in vitro* under increasing thiamine concentrations. It is conceivable that two genes encoding enzymes using the same substrate might be regulated by similar mechanisms. However, although the 3-way junction that we identified downstream of the *AtAIRC-S* gene shares structure and sequence similarities with the purine riboswitch consensus (Fig. 2), it does not respond to the addition of guanine by structural alteration, and therefore does not comply with the definition of a riboswitch. Thus the *Arabidopsis AtAIRC-S* gene might be a dormant descendant of the ancestral *purE* gene.

It is also worthwhile noting that since the riboswitch discovery in bacteria in 2002, only the TPP-riboswitch was found in eukaryotes [Kubodera *et al.*, 2003; Sudarsan *et al.*, 2003a; Sudarsan *et al.*, 2005] and only in very few organisms. The failure of a more sophisticated bioinformatics search by structure as presented in this article to discover more, alongside the attempts by others since 2002, suggests that a search dictated by the sensor domain conservation consideration alone has been exhausted and as proposed in [Nudler, 2006] the expression platform needs to be considered as well. Other types of riboswitch architectures may exist in eukaryotes.

Conclusions

We have proposed a new bioinformatics approach for finding 3-way junctions in a given set of genes. The approach was presented and a comparison was performed with other methods on a known limited set of data in bacteria. The method we have devised is currently limited to riboswitches such as the purine riboswitch for which the folding prediction by energy minimization is accurate. In the future, there is a potential to overcome this limitation by an improvement in the parameterization by biochemical experiments and other continual advances in RNA folding predictions. Subsequently, we have implemented our new approach on a test case involving genes that participate in purine metabolism in *Arabidopsis*.

Based on our preliminary studies in prokaryotes, we opted to use the following two procedures in eukaryotes. The first procedure is a hierarchical global search, starting from the **merge** dataset obtained by applying FastR, RNAMotif, SeqSniff_Imitate on complete genomes to be followed by RNAProfile. The second procedure is a focused search, starting from the **genes** dataset and applying our proposed approach together with other constraints imposed.

As a result of the second procedure, we computationally found an interesting candidate that we thought may have a chance to become a purine riboswitch in plants. The candidate was predicted to have a structure that is highly similar to the consensus available in RFAM and a considerable sequence similarity. To verify if this is indeed a true riboswitch or an interesting false positive, we have performed an in-line probing experiment. We found that although the in-line probing verifies the 3-way junction secondary structure prediction, the candidate does not respond to the binding of guanine and hence should be classified as a false positive one despite its structure and sequence similarity to the consensus. Nevertheless, our bioinformatics approach that was described for attempting to locate a purine riboswitch in fungi and plants and its structure prediction that was successfully verified in the biochemical experiment can be used to search for any aptamer of interest in various other metabolic cycles and other genomes in the future.

Acknowledgements

Michal Shapira is an incumbent of the Thaler Chair for Plant Genetics at BGU. Plant research in the Shapira laboratory is supported by the Israel Science Foundation and by the Faculty of Natural Sciences of BGU. Asaph Aharoni is an incumbent of the Adolfo and Evelyn Blum Career Development Chair at the Weizmann Institute. The work in the Aharoni laboratory was supported by the William Z. and Eda Bess Novick New Scientists Fund and the Henry S. and Anne Reich Family Foundation. The research was partially supported by a grant from the Israel USA binational science foundation BSF 2004291.

References

- Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J. K. and Breaker, R. R. (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA **101**, 6421-6426.
- Batey, R. T., Gilbert, S. D. and Montange, R. K. (2004). Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature **432**, 411-415.
- Bergig, O., Barash, D., Nudler, E. and Kedem, K. (2004). STR^2: A structure to string approach for locating G-box riboswitch shapes in pre-selected genes. In Silico Biology **4**, 593-604.
- Chapman, K. A., Delauney, A. J., Kim, J. H. and Verma, D. P. (1994). Structural organization of de novo purine biosynthesis enzymes in plants: 5-aminoimidazole ribonucleotide carboxylase and 5-aminoimidazole-4-N-succinocarboxamide ribonucleotide synthetase cDNAs from Vigna aconitifolia. Plant Mol. Biol. **24**, 389-395.
- Chung, S. O., Lee, J. H., Lee, S. Y. and Lee, D. S. (1996). Genomic organization of purK and purE in Brevibacterium ammoniagenes ATCC 6872: purE locus provides a clue for genomic evolution. FEMS Microbiol. Lett. **137**, 265-268.
- Durbin, R., Eddy, S., Krogh, A. G. and Mitchison, G. (1998). Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
- Ebbole, D. J. and Zalkin, H. (1987). Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for *de novo* purine nucleotide synthesis. J. Biol. Chem. **262**, 8274-8287.
- Ellington, A. D. and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature **346**, 818-822.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. R. (2003). Rfam: an RNA family database. Nucleic Acids Res. **31**, 439-441.
- Henderson, J. F. and Paterson, A. R. P. (1973). Nucleotide Metabolism: An Introduction. Academic Press, London.
- Henikoff, S. (1987). Multifunctional polypeptides for purine de novo synthesis. Bioessays **6**, 8-13.
- Hermann, T. and Patel, D. J. (2000). Adaptive recognition by nucleic acid aptamers. Science **287**, 820-825.
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125**, 167-188.
- Holbrook, S. R. (2005). RNA structure: the long and the short of it. Curr. Opin. Struct. Biol. **15**, 302-308.

- Klein, R. J. and Eddy, S. R. (2003). RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics **4**, 44.

- Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S., Gomi, K. and Hanamoto, H. (2003). Thiamine-regulated gene expression of *Aspergillus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. FEBS Lett. **555**, 516-520.

- Lescoute, A. and Westhof, E. (2005). Riboswitch structures: purine ligands replace tertiary contacts. Chem. Biol. **12**, 10-13.

- Lescoute, A. and Westhof, E. (2006). Topology of three-way junctions in folded RNAs. RNA **12**, 83-93.

- Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. **29**, 4724-4735.

- Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. and Breaker, R. R. (2003). Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. Cell **113**, 577-586.

- Mandal, M. and Breaker, R. R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. **5**, 451-463.

- Mandal, M., Lee, M., Barrick, J. E., Weinberg, Z., Emilsson, G. M., Ruzzo, W. L. and Breaker, R. R. (2004). A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science **306**, 275-279.

- Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A. and Nudler, E. (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell **111**, 747-756.

- Neuhard, J. and Nygaard, P. (1987). Biosynthesis and conversions of nucleotides: Purine and pyrimidines. *In:* Neidhardt, F. C. (ed). Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology. Vol 1. Amer Soc Microbiol, Washington, DC pp. 445-473

- Nudler, E. and Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. Trends Biochem. Sci. **29**, 11-17.

- Nudler, E. (2006). Flipping riboswitches. Cell **126**, 19-22.

- Pavesi, G., Mauri, G., Stefani, M. and Pesole, G. (2004). RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Res. **32**, 3258-3269.

- Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. **276**, 71-84.

- Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2003). Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. J. Biol. Chem. **278**, 41148-41159.

- Serganov, A., Yuan, Y.-R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R. and Patel, D. J. (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem. Biol. **11**, 1729-1741.

- Shapiro, B. A. (1988). An algorithm for comparing multiple RNA secondary structures. Comput. Appl. Biosci. **4**, 387-393.

- Stasolla, C., Katahira, R., Thorpe, T. A. and Ashihara, H. (2003). Purine and pyrimidine nucleotide metabolism in higher plants. J. Plant Physiol. **160**, 1271-1295.

- Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M. and Breaker, R. R. (2005). Thiamine

pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. Chem. Biol. **12**, 1325-1335.

- Sudarsan, N., Barrick, J. E. and Breaker, R. R. (2003a). Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA **9**, 644-647.

- Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. and Breaker, R. R. (2003b). An mRNA structure in bacteria that controls gene expression by binding lysine. Genes Dev. **17**, 2688-2697.

- Szankasi, P., Heyer, W. D., Schuchert, P. and Kohli, J. (1988). DNA sequence analysis of the ade6 gene of *Schizosaccharomyces pombe*. Wild-type and mutant alleles including the recombination host spot allele ade6-M26. J. Mol. Biol. **204**, 917-925.

- Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249**, 505-510.

- Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. and Gelfand, M. S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet. **20**, 44-50.

- Winkler, W. and Breaker, R. R. (2003). Genetic control by metabolite-binding riboswitches. Chembiochem. **4**, 1024-1032.

- Winkler, W., Nahvi, A. and Breaker, R. R. (2002a). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature **419**, 952-956.

- Winkler, W., Cohen-Chalamish, S. and Breaker, R. R. (2002b). An mRNA structure that controls gene expression by binding FMN. Proc. Natl. Acad. Sci. USA **99**, 15908-15913.

- Yao, Z., Weinberg, Z. and Ruzzo, W. L. (2006). CMfinder - a covariance model based RNA motif finding algorithm. Bioinformatics **22**, 445-452.

- Zalkin, H. and Nygaard, P. (1996). Biosynthesis of purine nucleotides. *In:* Neidhardt, J. *et al.* (eds). Escherichia coli and Salmonella typhimurium Cellular and Molecular Biology. American Society for Microbiology, Washington, DC pp. 561-579.

- Zhang, S., Haas, B., Eskin, E. and Bafna, V. (2005). Searching genomes for noncoding RNA using FastR. IEEE/ACM Trans. Comput. Biol. Bioinform. **2**, 366-379.

- Zhang, S., Borovok, I., Aharonowitz, Y., Sharan, R. and Bafna, V. (2006). A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. Bioinformatics **22**, e557-565.

- Zrenner, R., Stitt, M., Sonnewald, U. and Boldt, R. (2006). Pyrimidine and purine biosynthesis and degradation in plants. Annu. Rev. Plant. Biol. **57**,805-836.

- Zuker, M. (1989). On finding all suboptimal folding of an RNA molecule. Science **244**, 48-52.

- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**, 3406-3415.

Appendix

Table A: Accession numbers of *Saccharomyces cerevisiae*

Description	NCBI Accession Number	Version
Chromosome #1	NC_001133	NC_001133.5
Chromosome #2	NC_001134	NC_001134.7
Chromosome #3	NC_001135	NC_001135.3
Chromosome #4	NC_001136	NC_001136.6
Chromosome #5	NC_001137	NC_001137.2
Chromosome #6	NC_001138	NC_001138.4
Chromosome #7	NC_001139	NC_001139.5
Chromosome #8	NC_001140	NC_001140.4
Chromosome #9	NC_001141	NC_001141.1
Chromosome #10	NC_001142	NC_001142.5
Chromosome #11	NC_001143	NC_001143.5
Chromosome #12	NC_001144	NC_001144.3
Chromosome #13	NC_001145	NC_001145.2
Chromosome #14	NC_001146	NC_001146.3
Chromosome #15	NC_001147	NC_001147.4
Chromosome #16	NC_001148	NC_001148.3
Mitochondrion	NC_001224	NC_001224.1

Table B: Accession numbers of *Arabidopsis thaliana*

Description	NCBI Accession Number	Version
Chromosome #1	NC_003070	NC_003070.5
Chromosome #2	NC_003071	NC_003071.3
Chromosome #3	NC_003074	NC_003074.4
Chromosome #4	NC_003075	NC_003075.3
Chromosome #5	NC_003076	NC_003076.4
Mitochondrion	NC_000932	NC_000932.1
Chloroplast	NC_001284	NC_001284.2

Table C: Genes involved in the STR^2 search on *Saccharomyces cerevisiae*.

	Gene name	Source Chromosome #		Gene name	Source Chromosome #
1	PRS4	2	19	PRS3	8
2	HIS7	2	20	HIS5	9
3	HIS4	3	21	HIS6	9
4	APT2	4	22	SER33	9
5	ADE8	4	23	PRS1	11
6	MTH1	4	24	PNP1	12
7	SER3	5	25	ADE16	12
8	PRS2	5	26	ADE13	12
9	FCY21	5	27	MEU1	12
10	FCY22	5	28	ADE4	13
11	FCY2	5	29	ADE17	13
12	HIS1	5	30	APT1	13
13	HIS2	6	31	AAH1	14
14	SER2	7	32	ADE12	14
15	ADE3	7	33	SER1	15
16	ADE6	7	34	ADE2	15
17	TPN1	7	35	PRS5	15
18	ADE5,7	7	36	HIS3	15
			37	FCY1	16