Prof. dr hab. Jerzy Ciesiołka
Instytut Chemii Bioorganicznej PAN
ul. Noskowskiego 12/14
61-704 Poznań

Poznań, 2015-09-22

**Review of the PhD thesis of mgr Ge Qu**

presented in order to obtain a PhD degree from the Faculty of Biology, Adam Mickiewicz University in Poznań.

Thesis title: **Identification of novel non-coding RNA genes in plants using conserved promoter regulatory elements**

The research presented in the Thesis was supervised by prof. dr hab. Wojciech Karłowski and was conducted at the Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań.
The work aimed at searching for novel non-coding RNA genes in plants, mainly in Arabidopsis and rice, using the computational approach accompanied by experimental validation of the presumed transcriptional units. Two regulatory elements, Telo–SiteII and USE–TATA, were applied as functional, transcribed gene indicators in computational analysis. Selected, newly identified ncRNA genes were further validated experimentally at the RNA level.
In light of an enormous increase of interest in non-coding RNA molecules in recent years I found the topic of the Thesis highly on time and scientifically very important.

The Introduction section (17 pages) provides the description of the current stage of knowledge on non-coding RNAs, transposable elements, and selected *cis*-regulatory elements. I understand the Author's intention to characterize all the molecules and processes to which he refers in the discussion of his own results. This is very beneficial to the reader, but in some places the text is organized into very short chapters giving an impression of information fragmentation. In my opinion, the Author could have presented some data in tables and illustrate them with figures. It is worth noting, however, that the Introduction section covers a

high number of studies and is well documented in the References list (for a few citations incomplete details were given, yet they can easily be found in the Medline database).

The Materials and Methods section (8 pages) is complete and clearly presented, providing easy to follow and precise protocols. A wide spectrum of methods was used – from bioinformatics procedures: retrieving genome databases and RNA sequencing data with several currently available computer programs to wet-laboratory approaches: DNA and RNA isolation, gene identification by PCR, identification and characterization of RNA transcripts by RT-PCR, 5' and 3' RACE, and Northern blot analysis.

The Results section of the Thesis (45 pages) is divided into three major parts. In the first part, in order to identify novel ncRNA genes, two promoter element combinations, *Telo*-box associated with the Site II element (Telo–SiteII) and USE associated with the TATA-box (USE–TATA), were used to screen the whole genome of Arabidopsis. The initial screening procedure for promoter elements Telo–SiteII generated approximately 3900 non-redundant loci in the Arabidopsis intergenic regions. After thoughtful evaluation, the Author excluded most of these loci from further analysis. Finally, the collection of the remaining loci was restricted to those covered by cDNA/EST sequences, along with available RNA-Seq datasets. This led to the identification of 12 Telo–SiteII gene loci, corresponding to ten polycistronic and two monocistronic snoRNA genes. Many of the predicted snoRNA sequences had a high level of sequence conservation between the closely related species *Arabidopsis lyrata*, *Capsella rubella* and *Brassica rapa*. No similar sequences were found in *Medicago truncatula* or *Oryza sativa* plant genomes. Seven snoRNAs were predicted to direct the methylation of 18S or 25S rRNA residues, and 13 other RNAs were found to target the spliceosomal snRNAs U2, U4, U5 or U6. The expression of a majority of the predicted RNAs was supported by RNA-Seq data. Importantly, the expression of some newly identified snoRNAs was validated by the Author experimentally, using the RT-PCR analysis. Downstream of the Telo–SiteII element some other approx. 50 ncRNAs were located. However, they did not share any similarity with known RNA families. Out of these, 16 loci were supported by sequence databases on the RNA level. Computer analysis suggested that these RNAs were polyadenylated although they did not code any protein. Several of these novel ncRNAs were validated by the Author experimentally by RT-PCR. However, he did not analyze them any further.

The second part of the Thesis describes a novel class of polycistronic sno-miRNA genes that were revealed by an analysis of Telo–SiteII loci. The Author decided to focus on the analysis of the novel dicistronic genes of snoRNA–miR775 and snoRNA–miR779. For snoRNA–miR775, in addition to the validation of its expression by RT-PCR, transcription start and termination sites were determined by 5' and 3' RACE. The secondary structure of the sno-miR775 precursor sequence as well as the interactions of snoR775 with the complementary region of target 25S rRNA were also proposed. Moreover, the Author confirmed the expression of snoRNA as the predicted length transcript by Northern blot analysis in 3 different plant tissues. Undoubtedly, such a thorough analysis was fully justified by the novelty of the discovery of this dicistronic transcript. Its identification raises several questions concerning the functional significance of producing snoR775 and miR775 from the same precursor. The Author observed a similar genomic organization for the sno-miR779 gene. Characterization of expression of this gene was also performed, similarly as earlier for snoRNA–miR775. What was, however, the reason for not showing the transcription termination site in Figure 9? In view of the very interesting discovery of these dicistronic transcripts in Arabidopsis, I wonder whether further studies of such transcripts are planned, including, for example, characterization of the ncR100-miR158b gene or elucidating functional significance of such dicistronic arrangements? Some functional studies are mentioned in the Thesis but these experiments and the mutants which were used are not described in detail. Moreover, while identifying non-coding RNA molecules, one has to keep in mind that most of them are involved in the development, differentiation, stress response and regulation of cellular processes. Thus, their expression may vary considerably depending on the state of the cell or tissue as well as on environmental conditions. Could the Author comment on the issues raised above during the public defense of his dissertation?

In order to find out whether a similar sno-miRNA dicistronic arrangement also exists in the rice genome, the genomic regions flanking the annotated rice miRNA genes were scanned. This procedure enabled the identification of 20 new sno-miRNA candidates. Eight candidates were selected and tested by RT-PCR experiments. The results confirmed the expression of six precursors encoding the predicted snoRNA–miRNA genes. In some cases, the assay was ambiguous due to alternative splicing events or sequence overlaps.
Subsequently, the Author performed a comparative analysis of four sno-miRNA genes in 12 plant organisms. It turned out that although the mature snoRNA and miRNA components could be identified in other species, only in Arabidopsis they seemed to be a part of a

common transcript. However, a preliminary analysis of two chosen sno-miRNAs in rice suggested the presence of similar common transcript arrangements. Is there any explanation/suggestion why the snoRNA and miRNA components seemed to be a part of a common transcript only in Arabidopsis (and possibly in rice)?

In the third part of the Thesis the Author describes the use of an approach analogous to that applied earlier for the analysis of Telo–SiteII elements, but with the USE–TATA combination as a promoter indicator. Fortunately, the nucleotide composition of USE–TATA is highly conserved, and the distance between these elements clearly distinguishes between Pol II and Pol III-directed genes. After screening of the genome of Arabidopsis and filtering the results, 26 potential novel ncRNA candidates were found; 5 and 21 loci were of the Pol-II- and Pol-III-type, respectively. The expression of nine transcripts was confirmed by RT-PCR assay. Sequence alignment analysis of the predicted ncRNA genes revealed an additional conserved motif, named UTAM, 25 nucleotides in length, which could be considered as an internal promoter element. Subsequent scanning of the genome of *Arabidopsis thaliana* with the newly discovered motif resulted in six protein-coding genes, in which the motif was located in the intronic regions. Therefore, the Author hypothesizes that isoforms of these genes may exist.
A motif similar to the UTAM motif found in Arabidopsis was also identified in the rice genome but its consensus sequence was supposed not to be the same. The data concerning the consensus sequences of the UTAM motifs, described in the text and shown in Figures 17 and 19 is, however, confusing. I would appreciate if the Author could clarify this point. Finally, the Author noted that some of the identified ncRNA loci contained multiple copies of USE-TATA-UTAM motifs. This observation has a rather preliminary character and requires experimental validation of its functional importance.

The Discussion section (6 pages) gave the Author a possibility to present several additional remarks concerning the promoter-based methodology for the annotation of ncRNAs. However, most importantly, a few new hypotheses were formulated regarding the biological significance of the obtained data. The Author also suggested the directions of further work towards establishing the functions of the discovered ncRNAs. Moreover, in the next, last section of the Thesis titled Perspectives (2 pages) he proposed concrete experiments aimed to solve several questions that remain to be answered. In my opinion, in the last two sections the Author documented a high level of expertise in the non-coding RNA field.

To conclude this report, the results presented in the Thesis are novel and very interesting. They shed a new light on non-coding RNA genes present in plants, mainly in Arabidopsis. Some of the results have been described in a manuscript that is under review, with Mr Ge Qu as the first author. I have no doubt that the other obtained data will provide the basis for further studies and subsequent valuable publications.

Therefore, I recommend the Scientific Board of the Faculty of Biology, Adam Mickiewicz University in Poznan, proceed with further procedural steps to confer a PhD degree to Mr Ge Qu.