Prof. dr Piotr Zielenkiewicz

Institute of Biochemistry and Biophysics PAS

Warsaw University, Faculty of Biology

Warszawa, 07.09.2015

Evaluation

of the **PhD** dissertation of Mr **Ge Qu** for the Faculty of Biology of Adam Mickiewicz University

The PhD dissertation of Mr. Ge Qu, MSc entitled "Identification of novel non-coding RNA genes in plants using conserved promoter regulatory elements" fits perfectly into research profile of the Laboratory of Computational Genomics of the Institute of Molecular Biology and Biotechnology of the Faculty of Biology, Adam Mickiewicz University in Poznań. The research interests of the laboratory head and supervisor of the presented thesis, Professor Wojciech Karłowski from many years concentrate on predictive sequence analysis of nucleic acids, the field in which he very successfully cooperates with experimental groups. The task which was put in front of Mr Ge Qu to be solved in the frame of his PhD work was to check if prediction of different non-coding plant RNAs is possible on the basis of the sequence analysis of promoter regions.

The existence of the non-coding RNAs was one of the major discoveries of the recent decades, allowing to truly understand processes underlying gene expression and regulation. On the other hand, the diversity of non-coding RNAs and difficulty to distinguish them from RNA molecules present due to degradation of larger molecules or natural variability in gene expression occurring between cells in isogenic populations (transcriptional noise) makes prediction of functional ncRNAs a difficult task.

Mr Qu decided to apply a novel methodology for the *in silico* prediction of ncRNA in plants, namely based on the fact that promoter regions contain conserved regulatory elements. Two combinations of evolutionarily conserved elements were already distinguished in plants.

For snoRNAs the promoter regions are enriched with Telo-box associated with SiteII element. Mr Qu constructed a computational pipeline to search Telo-SiteII signature in *Arabidopsis thaliana* genome. He assumed that a region would constitute a putative promoter if both Telo and SiteII could be found in a common region of less than 1 kb. Next, he considered sequences up to 0,5 kb away and downstream from the Telo-SiteII location to analyse gene annotations from TAIR, miRBase and PLncDB. The predicted ncRNAs were compared with known RNAs from RFAM. This procedure generated 3896 loci in Arabidopsis. This part of the calculations led to the discovery of 61 novel ncRNAs under control of Telo-SiteII including 26 new sno- and sca- RNAs. Thirteen of the new ncRNAs are of sca- type targeting snRNAs, which more than doubles the number of scaRNAs known in Arabidopsis. However, probably the most interesting result of this research is identification of sno-miRNA genes producing precursor leading to both mature sno- and miRNAs. In addition to in silico calculations, this result was also confirmed experimentally, as well as other results with novel ncRNA predictions.

Similarly to the presence of Telo-SiteII, also for snRNA gene promoters there is a combination of conserved regulatory elements, namely the upstream sequence element (USE) and TATA-box. For the Pol II-directed genes the distance between USE and TATA is 32-36 bp, for Pol III it is 23-26 bp. Both distances are highly conserved. The USE-TATA motif has also been used in a pipeline of computations meant for the prediction of ncRNAs. For this purpose the pairs of USE and TATA with distances restricted to values characteristic for both polymerases were searched in the genome. The position of TATA box was used to point to the transcription start site, the adjacent (0,5 kb) sequences of which were searched for annotations in the sum of the above mentioned databases (with addition of ASRG). Comparison to RFAM enabled identification of 197 non-redundant USE-TATA loci in Arabidopsis genome. Again, RT-PCR experiments allowed to confirm correctness of the *in silico* predictions. The multiple sequence alignment of the ncRNAs allowed Mr Qu to identify (in the Pol III type ncRNA subset) a new sequence motif present 22 nt downstream the TATA-box, with no similarity to any known RFAM motifs. It has therefore been named UTAM (USE-TATA Associated Motif) and probably is the most important result from this part of study. It has further been shown that this motif is present in other plants.

The PhD dissertation of Mr Qu is written in English and divided into Introduction, Materials and Methods, Results, Discussion and Perspectives sections. Although the division seems to be the standard one, the content of the sections is not necessarily what one would expect. For example, this conservative referee would like to be convinced from reading the introduction that the objectives are relevant. For this work, it would mean that the introduction contains critical review of the current methods of ncRNAs prediction leading to recognition of the need for new prediction methods. In the present introduction, however, there is only a standard review of ncRNA types and not even a word about previous methods for ncRNA sequence prediction. Similarly, in the Discussion section, I would expect comparison of the results obtained with the methods developed by the Author with the results obtained by other ncRNA prediction software on the same test set.

It would be beneficial if we could hear one or two examples of this kind during the public defense of Mr Qu thesis.

The critical comments above do not nullify generally high rating of the dissertation text, in which the results are discussed in detail in the context of recent literature.

I conclude that the work presented by Mr Ge Qu fulfills the customary and legal criteria required from PhD dissertations in Poland and urge the Council of the Faculty of Biology of Adam Mickiewicz University to grant the candidate access to further stages of the procedure.