Non-protein-coding RNAs

Diego Mauricio Riaño Pachón

October 21, 2004

Contents

1 General Information

The perspectives of biological research broadened enormously in recent years. Due to several completed and still ongoing genome sequencing projects the "code of life" is now publicly available for an increasing number of species. Among them are the three plant species *Arabidopsis thaliana* (thale cress), *Medicago truncatula* (barrel medic), and *Oryza sativa* (rice). However, the main value of genomic sequence data relies on its annotation, i.e., identification of the genes encoded by the genome. This systematic annotation is mainly based on computer assisted gene prediction algorithms. One of the most important criteria for gene prediction is the presence of an 'open reading frame' (ORF) that must extend over at least 300 nt (MacIntosh *et al.*, 2001). In this way most of the smaller genes are being missed by default, unless experimental evidence supports their existence. Furthermore, a few years ago most of the genes were thought to code for

proteins. However, now it is a well accepted fact that beside these protein-coding genes, there is a vast amount of non-protein-coding genes (being transcribed into ncRNAs), which code for RNA as a final product rather than for a protein (Mattick and Gagen, 2001; Mattick, 2003). These ncRNAs are assumed to play important roles in processes such as transcriptional regulation, chromosome replication, RNA processing and modification, messenger RNA stability, translation, protein stability and protein translocation (Storz, 2002). Non-protein-coding RNAs have been found experimentally in bacteria, insects, mammals and plants (Eddy, 2001; Erdmann *et al.*, 2001*b*; MacIntosh *et al.*, 2001; Marker *et al.*, 2002; Storz, 2002). Three well known classes of non-protein-coding RNAs (rRNA, tRNA, and snRNA) are characterized by very clear features having allowed to develop specialized gene-finding programs to search for this kind of genes (Rivas and Eddy, 2001). However, definitive gene-finding programs to predict ncRNAs of other types do not exist.

Because of these limited available approaches, in the ongoing process of the annotation of the genome of the model plant *A. thaliana*, like in other genomes, the identification of ncRNAs is not promoted on a systematic scale, with the notable exception of (MacIntosh *et al.*, 2001; Hüttenhofer *et al.*, 2002). Thus it is very likely that an important amount of genes are overlooked from genome annotation of *A. thaliana* because either they lack significant open reading frames or encode RNA as their final product.

1.1 Identification of ncRNAs in the wet-lab

First ncRNAs were identified by chance. After the realization of their existence and importance, systematic approaches started to be employed (Storz, 2002; Hüttenhofer *et al.*, 2002; Sunkar and Zhu, 2004). Briefly, different approaches rely in the separation and enrichment of small RNAs by gel electrophoresis, followed by the elimination of well know RNAs, by BLAST searches or experimentally through hybridization (Lagos-Quintana *et al.*, 2001; Storz, 2002; Sunkar and Zhu, 2004). Size-fractionated RNA populations have been isolated from several species (e. g. *A. thaliana*, *Archaeoglobus fulgidus*, *C. elegans*, *Drosophila*). Northern blots have been used to confirm the expression of small transcripts, providing information about spatial and temporal expression patterns (Storz, 2002).

Some of the non coding RNAs (ncRNA) are transcribed by the polymerase II and therefore have mRNA-like structure (Erdmann *et al.*, 2000; Lee *et al.*, 2004), like polyadenylated tails, which means that they can be caught in EST libraries.

1.1.1 ncRNA flavors

Several types of ncRNAs have been identified experimentally. The most studied among them are snoRNA, miRNA and siRNA:

- snoRNAs. One class of ncRNAs, called small nucleolar RNA (snoRNA), has been found in the nucleolus. Their size ranges between 70 and 250nt. Some snoRNAs have a role in rRNA processing. Up to now snoRNAs can be divided into two subfamilies, "C/D box" and "H/ACA", that direct site-specific 2'-O-ribose methylation and pseudouridylation, respectively. Both snoRNA types form a complex with a protein methylase or pseudo-U synthetase, and the specificity to the target rRNA is provided by the snoRNA (Eddy, 2001; Hüttenhofer *et al.*, 2002). SnoRNAs also play a role in the modification of tRNA and snRNA.

- miRNAs. MicroRNAs usually act as transcriptional repressors in a single-stranded conformation (Eddy, 2001; Lagos-Quintana *et al.*, 2001) . Their sizes usually range from 21 to 23 nt (Hüttenhofer *et al.*, 2002).

- siRNAs. Small interfering RNAs act as a guide to target complementary RNA sequences for destruction. Their sizes usually range from 21 to 23 nt (Hüttenhofer *et al.*, 2002).

1.2 Computational identification of ncRNAs

Experimentally identified ncRNAs can be used to follow a similarity-based approach for the identification of more ncRNAs. However, this approach has the drawback that new structurally divergent ncRNAs cannot be identified. In ncRNAs no signatures like a start and stops codons have been uncovered. Additionally, a secondary structure-based approach could be used, based on the fact that the secondary structure that is adopted by

non-protein-coding RNAs (e. g. snoRNAs, tRNAs) is fundamental for its biological function. Accordingly, (Griffiths-Jones *et al.*, 2003) have created models for RNA families, which are deposited in the Rfam database. This resource allows to identify previously unrecognized members of existing RNA families, but a completely novel ncRNA that does not belong to one of the described families will fail to be recognized. Consequently, there is no definitive computational method to identify novel ncRNAs. Nevertheless, there are some guidelines that could help in the prediction of novel ncRNA genes. According to (Eddy, 2001) one of the best lines of evidence to distinguish between small peptide coding genes and ncRNA genes is comparative genome analysis. An ORF should be conserved in other related species. Therefore, the pattern of mutations in the related genes should favor synonymous and conservative aminoacid exchanges. These would not happen in an ncRNA gene. Instead, in an ncRNA gene, it might be possible to find an intramolecular secondary structure and comparative analysis should show compensatory base substitutions (Rivas and Eddy, 2001; McCutcheon and Eddy, 2003). Additionally, given that the secondary structure in ncRNAs is important, one can use this feature to distinguish ncRNAs from random sequences. It has been shown (Rivas and Eddy, 2000) that secondary structure is not useful to distinguish ncRNAs from random sequences. Although, recent studies had shown that at least in miRNA precursors (Bonnet *et al.*, 2004) and H/ACA snoRNAs (Edvardsson *et al.*, 2003) secondary structure can be successfully used to distinguish those kinds of genes from random sequences. Anyway, for the reliably prediction of ncRNAs the use of secondary structure prediction should be used in association with other kind of evidence, as the presence of RNA motifs and comparative genomics (Edvardsson *et al.*, 2003; Eddy, 2002). Consequently, it is necessary to conduct a detailed systematic study on the usefulness of secondary structure as a useful feature in the prediction of ncRNAs.

The main method to evaluate secondary structure in RNAs is by the Minimum Free Energy (MFE) of the folded structure (Le *et al.*, 1988; Le *et al.*, 1989; Chen *et al.*, 1990), but this approach has a serious drawback. Biologically active RNAs might not adopt the secondary structure with the Minimum Free Energy but a sub-optimal structure (Meyer and Miklos, 2004; Giegerich *et al.*, 2004) and the search of sub-optimal structures increases greatly the computational complexity of the problem. Therefore it is possible

that approaches relying on the computation of MFE have limited usefulness. Accordingly, some recent approaches have been proposed based on an additional level of abstraction of the RNA secondary structure. Those approaches represent RNA secondary structures as graphs, that can be studied within the framework of the graph theory (Diestel, 2000). A graph is a structure composed by a set of nodes and a set of edges between those nodes. This kind of representation could help to uncover features that underlay the secondary structure (Bermúdez *et al.*, 1999; Gan *et al.*, 2003; Giegerich *et al.*, 2004). Bermúdez et al. (Bermúdez *et al.*, 1999), worked on the graph representation of tRNAs. The set of nodes corresponding to nucleotides, and the set of edges representing covalent or hydrogen bonds between nucleotides, weighting edges and nodes based on quantum properties. Then the authors evaluated the structural similarity of different graphs representations which allowed them to find "a correlation between tRNAs that shared structural features with aminoacids belonging to similar biosynthetic pathways" (Bermúdez *et al.*, 1999). In another study (Gan *et al.*, 2003), the set of nodes corresponds to loops and bulges and the set of edges represents stems. Then the authors proceed to enumerate graph motifs. Gan et al., found that the number of natural motifs is smaller than the number of mathematically possible (random) motifs. Both approaches offered important insight into ncRNAs features, it is feasible to find correlations between structures and biological function as shown by (Bermúdez *et al.*, 1999), and the number of real graphs motifs is much smaller that than the theoretical number, which means that are serious constraints for those motifs, that can be use in gene finding approaches. Anyway, the utility of this approaches in the search for novel ncRNAs still remains to be tested.

1.2.1 ncRNAs and small peptide-coding genes databases

Several ncRNAs and small peptides have been found in different organisms. Their sequences have been compiled in different databases. Thereby the main focus is set to ncRNAs whereas small peptide coding genes do not receive special attention. Among those databases are for example:

- Plant snoRNA. This database compiles small nucleolar RNAs involved in cleavage of precursor ribosomal RNA and small nuclear spliceosomal RNAs (Brown *et al.*, 2003). `http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home`

- Noncoding RNAs in Plants. This database compiles lists of known or annotated ncRNAs in several plant species and list ESTs with characteristics of ncRNAs or small peptide-coding RNAs (MacIntosh *et al.*, 2001). `http://www.prl.msu.edu/PLANTncRNAs/`

- Noncoding RNA Database. This database contains information about ncRNAs which do not have long open reading frames and that act as riboregulators (Erdmann *et al.*, 2001*a*). `http://biobases.ibch.poznan.pl/ncRNA/`

1.3 Small peptides and ncRNAs in *A. thaliana*

One of the first approaches to address the systematic and computational prediction of ncRNAs in *A. thaliana* has been carried out by (MacIntosh *et al.*, 2001). Their work started with a pre-clustered collection of ESTs performed by a different group. With this approach 15 potential ncRNAs and 10 potential small peptide coding genes were predicted. However, further surveys in our group weakened the confidence in these predictions. The main point of criticism concerns the fact that the approach did not take into account the genome annotation of the recently (at that time, 2001) sequenced *A. thaliana* genome. Taking now a closer look at the predictions of MacIntosh et al., and comparing them against the *A. thaliana* genome annotation it is found that most of the predicted ncRNAs have high similarity with known proteins or at least contain characteristic protein patterns. Nevertheless, the proposition to base the computational search for ncRNAs on ESTs seems to be promising. Therefore, and as shown by (Riano-Pachón *et al.*, 2004) the method of MacIntosh et al., could be extended and improved. The new approach started with the clustering of the complete (at the moment available) EST collection of *A. thaliana* and then took into account the information embedded in the existing annotation of the *A. thaliana* genome.

References

Bermúdez,C., Daza,E. and Andrade,E. (1999) Characterization and comparison of *Escherichia coli* transfer RNAs by graph theory based on secondary structure. *J Theor Biol,* **197** (2), 193–205.

Bonnet,E., Wuyts,J., é,P. and Peer,Y.V.D. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics,* **Advance Online Publication**.

Brown,J.W.S., Echeverria,M. and Qu,L.H. (2003) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends Plant Sci,* **8** (1), 42–9.

Chen,J., Le,S., Shapiro,B., Currey,K. and Maizel,J. (1990) A computational procedure for assessing the significance of RNA secondary structure. *Comput Appl Biosci,* **6** (1), 7–18.

Diestel,R. (2000) *Graph Theory*, vol. 173, of *Graduate Texts in Mathematics*. Second edition,, Springer-Verlag.

Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat Rev Genet,* **2** (12), 919–29.

Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell,* **109** (2), 137–40.

Edvardsson,S., Gardner,P.P., Poole,A.M., Hendy,M.D., Penny,D. and Moulton,V. (2003) A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics,* **19** (7), 865–73.

Erdmann,V., Barciszewska,M., Hochberg,A., de Groot,N. and Barciszewski,J. (2001*a*) Regulatory RNAs. *Cell Mol Life Sci,* **58** (7), 960–77.

Erdmann,V., Barciszewska,M., Szymanski,M., Hochberg,A., de Groot,N. and Barciszewski,J. (2001*b*) The non-coding RNAs as riboregulators. *Nucleic Acids Res,* **29** (1), 189–93.

Erdmann,V., Szymanski,M., Hochberg,A., Groot,N. and Barciszewski,J. (2000) Non-coding, mRNA-like RNAs database Y2K. *Nucleic Acids Res,* **28** (1), 197–200.

Gan,H.H., Pasquali,S. and Schlick,T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res,* **31** (11), 2926–43.

Giegerich,R., Voss,B. and Rehmsmeier,M. (2004) Abstract shapes of RNA. *Nucleic Acids Res,* **32** (16), 4843–51.

Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res,* **31** (1), 439–41.

Hüttenhofer,A., Brosius,J. and Bachellerie,J.P. (2002) RNomics: identification and function of small, non-messenger RNAs. *Curr Opin Chem Biol,* **6** (6), 835–43.

Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science,* **294** (5543), 853–8.

Le,S., Chen,J., Currey,K. and Maizel,J. (1988) A program for predicting significant RNA secondary structures. *Comput Appl Biosci,* **4** (1), 153–9.

Le,S., Chen,J. and Maizel,J. (1989) Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. *Nucleic Acids Res,* **17** (15), 6143–52.

Lee,Y., Kim,M., Han,J., Yeom,K.H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J,* **Advance Online Publication**.

MacIntosh,G., Wilkerson,C. and Green,P. (2001) Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol,* **127** (3), 765–76.

Marker,C., Zemann,A., Terhörst,T., Kiefmann,M., Kastenmayer,J.P., Green,P., Bachellerie,J.P., rgen Brosius and ttenhofer,A. (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr Biol,* **12** (23), 2002–13.

Mattick,J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays,* **25** (10), 930–9.

Mattick,J.S. and Gagen,M. (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol,* **18** (9), 1611–30.

McCutcheon,J.P. and Eddy,S.R. (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res,* **31** (14), 4119–28.

Meyer,I.M. and Miklos,I. (2004) Co-transcriptional folding is encoded within RNA genes. *BMC Mol Biol,* **5** (1), 10.

Riano-Pachón,D.M., Dreyer,I. and Müller-Röber,B. Orphan transcripts in *Arabidopsis thaliana*: several hundred previously unrecognized genes. In preparation.

Rivas,E. and Eddy,S. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics,* **16** (7), 583–605.

Rivas,E. and Eddy,S. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics,* **2** (1), 8.

Storz,G. (2002) An expanding universe of noncoding RNAs. *Science,* **296** (5571), 1260–3.

Sunkar,R. and Zhu,J.K. (2004) Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *Plant Cell,* **16** (8), 2001–19.