Genome Biology

RESEARCH HIGHLIGHT

The amazing world of bacterial structured RNAs

Eric Westhof*

See related research article by Weinberg *et al.*: http://genomebiology.com/2010/11/3/R31

Abstract

The discovery of several new structured non-coding RNAs in bacterial and archaeal genomes and metagenomes raises burning questions about their biological and biochemical functions.

Introduction

The compact genomes of bacteria contain 10 to 15% non-coding DNA sequences, which are transcribed into non-coding RNAs. Several classes of non-coding RNAs are small, less than 80 to 150 nucleotides, and act as post-transcriptional regulators by targeting mRNAs. Another large class of non-coding RNAs act *in cis* by binding structured elements in the 5' untranslated regions of mRNAs. Perhaps the best known are called riboswitches; upon binding a metabolite, the fold of the transcript is modified and this influences either the termination of transcription or the initiation of translation [1].

Some longer non-coding RNAs have also been detected in recent years. For example, RNAIII present in several Gram-positive bacteria is 500 nucleotides long and contains structured regions framing an open-reading frame [2]. However, two recent papers from Ron Breaker's group increase the number of large non-coding RNAs astonishingly [3,4]. Several new smaller non-coding RNAs are also identified. Strikingly, most of the new non-coding RNAs are structurally very complex. The complexity of some of the larger ones seems similar to that of the large ribozymes, such as the self-splicing group I and group II introns. These observations show, once again, how little we know about the microbial world: a great proportion of these new non-coding RNAs were identified in metagenomes or in environmental DNA sequences.

The search for non-coding RNAs

The search for non-coding RNAs in genomes is far from trivial [5]. Even for homologous and functionally well characterized RNA molecules, such as the ubitquitous RNaseP or the telomerase RNA, the search cannot be reliably automated because of the large and unpredictable variation in the length of the RNA transcript, with new insertions appearing in an otherwise globally similar secondary structure. On the other hand, the *de novo* search for the presence of non-coding RNAs within intergenic regions is plagued by false positives because of the poor discriminative power. Various computer tools have been produced for searching for potential non-coding RNAs in genomes by exploiting the thermodynamic stabilities of the helices formed [6,7]. The tools are generally dedicated to searching for either *cis*-acting RNAs (such as riboswitches) or *trans*-acting RNAs (such as the RNAs binding by full or partial complementarity to another RNA, either non-coding or coding).

Computer tools have been around for some time for searching RNAs on the basis of a known element of secondary structure. It has also been established several years ago that secondary structure alone is not enough for predicting non-coding RNA [8]. The computational pipeline followed by Weinberg and coworkers [3,4] exploits the power of comparative sequence analysis and involves sophisticated automatic techniques combined with manual intervention. The central tool used by Weinberg and coworkers [3,4] is CMfinder, which can derive RNA motifs and secondary structures from a set of unaligned RNA sequences [7]. However, in order to appreciate what these programs attempt to do, it is worth recalling how complex the structures of non-coding RNAs can be.

Structural complexity

What is meant with structural complexity? The first level of folding of the transcribed RNA is the fold-back hairpin capped by a loop. Such a simple single hairpin can have profound biological effects. In bacteria, insertion of selenocysteine (a version of cysteine containing selenium rather than sulfur) occurs because the stop codon to be read as a selenocysteine codon is followed by a small hairpin. Series of hairpins can form, which, upon binding a ligand (another RNA, a protein or a metabolite), will lead to a more complex fold or to cleavage of the RNA. Structural complexity starts to appear when hairpins

*Correspondence: E.Westhof@ibmc-cnrs.unistra.fr
Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue René Descartes, F-67084 Strasbourg, France

branch off from a hairpin, forming a three-way or multi-way junction. Naturally, further branching off of hairpins can occur within an already branched off hairpin. Because hairpins in three-dimensional space form RNA helices, which are bulky, the available space they can occupy is restricted, leading to co-axial or parallel stacking of some of them and, consequently, intricate three-dimensional architectures.

Such RNA architectures are maintained by a multitude of intramolecular contacts, with a resulting network of interactions dominated by non-Watson-Crick pairs. It has been observed that the non-Watson-Crick pairs organize themselves in RNA modules that are crucial for maintaining the three-dimensional structure. In RNA modules, various types of non-Watson-Crick pairs form a set that occurs in a conserved sequential order because of strong constraints due to chemical linkages and base-base stacking. Among those modules, a prominent one is the G-bulged module (Figure 1; also called the sarcin/ricin or loop E module because it occurs in the sarcin/ricin hairpin of the 23S rRNA and in the loop E of the eukaryotic 5S rRNA). In the example shown in Figure 1a, an internal loop of the secondary structure forms a set of non-Watson-Crick pairs typical of G-bulged modules with stacking of the bases and a compact helicoidal fold. RNA modules also organize multiple junctions of helices. In Figure 1b, the single strands joining the helices interact with each other, forming a G-bulged module and a three-way junction with a clear orientation of the helices. In addition, most RNA modules are adapted for binding to other elements or regions, contributing further to the overall architecture. For example, G-bulged modules contribute to RNA function either by RNA-RNA interactions or by RNA-protein contacts. In such instances, the set of non-Watson-Crick base pairs is maintained and the module binds as a whole to either RNA or protein [9].

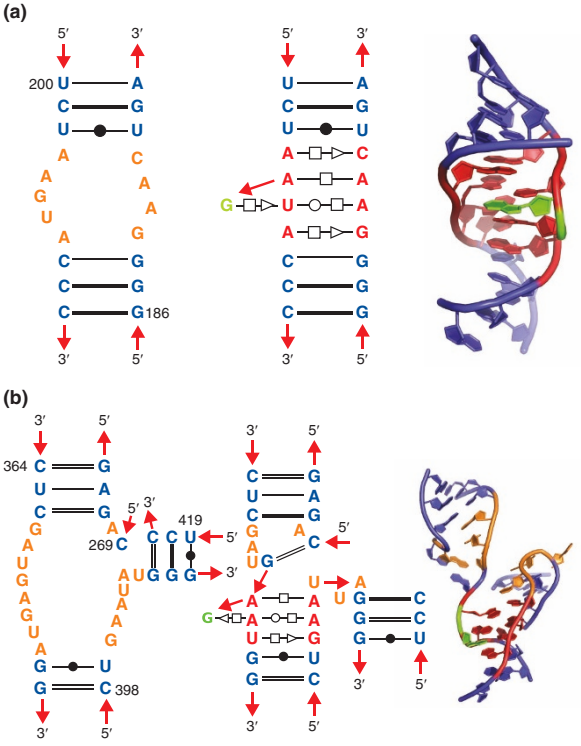

Figure 1. RNA secondary structures. Double lines between nucleotides indicate a strong Watson-Crick interaction between C and G; single lines indicate a weaker interaction between A and U. Nucleotides are colored as follows: blue, involved in Watson-Crick pairs; yellow, unpaired; red, involved in non-Watson-Crick pairs; green, the bulging G. The non-Watson-Crick pairs are named after the edges forming the H-bonded pairs and are indicated by: circle, Watson-Crick edge; square, Hoogsteen edge; triangle, Sugar edge. These symbols are blank when the two nucleotides approach in the *trans* orientation and dark when they approach in the *cis* orientation. Each panel shows the sequence with only Watson-Crick pairing on the left, the secondary structure with non-Watson-Crick pairing in the middle and the resulting three-dimensional structure on the right. **(a)** A G-bulged or loop E module completes a hairpin structure by forming non-Watson-Crick pairs within an internal loop. The sequential order of the usually observed set of non-Watson-Crick pairs is maintained, thereby defining a module. The structure of the G-bulged module shown is from helix H11 of the 23S rRNA of *Escherichia coli* (Protein DataBank (PDB) code 2AW4) [10]. **(b)** A G-bulged module organizes a three-way junction, leading to a rough co-axiality between two helical stems. The structure of the G-bulged module shown is the one at the junction of helices H16-H21-H22 from the 23S rRNA of *Escherichia coli* (PDB code 2AW4) [10]. Drawings courtesy of Jose Almeida Cruz.

Can we detect and assess structural complexity?

Such non-Watson-Crick pairs and the modules they form are an integral part of the tertiary structure; consequently, they are not predicted by the usual secondary structure programs that consider only Watson-Crick pairs. Correct secondary structure predictions should leave the bases that are potentially involved in non-Watson-Crick interactions as unpaired and single-stranded. Incorrect secondary structure predictions tend to predict that the bases that, in the native fold, would be forming non-Watson-Crick pairs are, instead, involved in secondary structure helices; this mis-prediction prevents the correct identification of structural elements key for the tertiary structure. Consequently, secondary structure predictions that allow for the possibility that single-stranded regions can form a known and recurrent RNA module have a higher probability of being functionally correct. Furthermore, given that such RNA modules are key elements of the tertiary structure, their presence indicates a potentially highly structured molecule.

Some striking cases are present in some secondary structures proposed for the newly reported RNAs. For example, the GOLLD (stands for Giant, Ornate, Lake- and Lactobacillales-Derived) RNA [3] contains two G-bulged modules, one internal loop within a hairpin,

and a second loop that forms a complex junction comprising four helices. In a very unusual example, the two strands forming the G-bulged modules exchange in the sequences (69% of the observed sequences start with 5'-AAA...AGUA-3' and 18% 5'-AGUA...AAA-3'; the remaining 5% adopt a simpler purine-rich module). Another RNA, dct-1, has a cluster of four G-bulged modules positioned around a three-way junction [3]. Interestingly, dct-1 is observed only in *Dictyoglomus thermophilum*, an extreme thermophile.

RNAs in metagenomes

As discussed by Weinberg and colleagues [3,4], several of the new RNAs could not have been discovered in the genomes of cultured bacteria known so far because such genomes do not contain the reported RNAs (except for some of the most recently sequenced genomes). Thus, the large collection of new RNAs are most probably just the tip of the iceberg, and an incredible number of still-to-be-discovered non-coding RNAs may be present in environmental sequences. The naming of the RNAs will continue to reflect the harvest of the sequences (for example, whalefall-1, Ocean-5, Soil-1 or Rhodopirellula-1).

The two recent papers [3,4] are extremely rich in information content, with large and complete supplementary material. They present many more RNAs, some of which are new riboswitches, with several containing various structural elements, such as interactions between loops or between a loop and a single-stranded region. Here, we highlight one particular aspect of the work. In the future, much more biochemical work, tedious and time-consuming, will be necessary to characterize the functions of the non-coding RNAs, to see whether they interact with a metabolite, another RNA or a protein and participate in regulatory networks, to identify those RNAs with catalytic power and to assess how widespread they are and why they were so elusive up to now.

Published: 15 March 2010

References

1. Barrick JE, Breaker RR: **The distributions, mechanisms, and structures of metabolite-binding riboswitches.** *Genome Biol* 2007, **8:**R239.
2. Benito Y, Kolb FA, Romby P, Lina G, Etienne J, Vandenesch F: **Probing the structure of RNAIII, the *Staphylococcus aureus* agr regulatory RNA, and identification of the RNA domain involved in repression of protein A expression.** *RNA* 2000, **6:**668-679.
3. Weinberg Z, Perreault J, Meyer MM, Breaker RR: **Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis.** *Nature* 2009, **462:**656-659.
4. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino KA, Moy RH, Breaker RM: **Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea and their metagenomes.** *Genome Biol* 2010, **11:**R31.
5. Menzel P, Gorodkin J, Stadler PF: **The tedious task of finding homologous noncoding RNA genes.** *RNA* 2009, **15:**2075-2082.
6. Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL: **De novo prediction of structured RNAs from genomic sequences.** *Trends Biotechnol* 2009, **28:**9-19.
7. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder - -a covariance model based RNA motif finding algorithm.** *Bioinformatics* 2006, **22:**445-452.
8. Rivas E, Eddy SR: **Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs.** *Bioinformatics* 2000, **16:**583-605.
9. Leontis NB, Stombaugh J, Westhof E: **Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules.** *Biochimie* 2002, **84:**961-973.
10. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JH: **Structures of the bacterial ribosome at 3.5 A resolution.** *Science* 2005, **310:**827-834.

doi:10.1186/gb-2010-11-3-108
Cite this article as: Westhof E: The amazing world of bacterial structured RNAs. *Genome Biology* 2010, **11:**108.