Journal of Biology

Minireview

Motifs from the deep

Tony W Hwang, Vlad Codrea and Andrew D Ellington

Address: Department of Chemistry and Biochemistry, Institute for Cell and Molecular Biology, University of Texas, Austin, TX78712, USA.

Correspondence: Tony W Hwang. Email: tony.hwang@mail.utexas.edu

Abstract

Because of the increasing recognition of the importance of noncoding RNAs in gene regulation, there is considerable interest in identifying RNA motifs in genomic data. In a recent report in *BMC Genomics*, Breaker and colleagues describe a new algorithm for identifying functional noncoding RNAs in metagenomic sequences of marine organisms, a strategy that may be particularly effective for discovering new and unique riboswitches.

Noncoding RNAs (ncRNAs) are increasingly recognized as mediators of disease [1] and as fundamental regulators of metabolic pathways in prokaryotes [2] and eukaryotes [3]. An unexpectedly large number of ncRNAs have been found to have key roles in essential cellular functions, including chromosome maintenance and DNA replication, RNA processing and translation, and protein translocation and stability [2,4]. The largest class of regulatory RNAs comprises microRNAs (miRNAs) of less than 30 nucleotides that bind to mRNAs and promote degradation or repress translation [4]. Less numerous than miRNAs, but widespread among bacteria, are riboswitches: structured RNAs located primarily in the 3' or 5' untranslated regions (UTRs) of bacterial mRNAs that bind metabolites and change conformation to regulate gene expression. Riboswitches are characterized by conserved motifs that include an 'aptamer domain' that recognizes the metabolite ligands and an 'expression platform' that can alter the conformation and function of regulatory elements involved in transcription or translation.

In recent years, experimental and bioinformatic strategies have been developed to discover ncRNA candidates in organisms ranging from *Escherichia coli* to humans. The laboratory of Ronald Breaker has now developed a novel method that extends this search to marine metagenomic data. The current work started with the genome of '*Candidatus* Pelagibacter ubique', which comprises as much as 20% of all marine metagenomic sequence reads [5], making it possibly the most abundant organism in the world, with estimates of approximately 10^{28} individual cells. '*C.* P. ubique' has the smallest genome yet found in a free-living organism, consisting of only 1.3 megabases, 1,354 genes, and very little noncoding DNA [6].

Computational methods have previously been successful at identifying structured RNAs, and the authors [5] developed a so-called comparative genomics pipeline for identifying regions in the '*C.* P. ubique' genome most likely to contain functional ncRNAs (Figure 1). Their strategy improves on a similar method developed by the same lab in 2007 [7] to search bacterial genomes. The 2007 work used complex criteria to define UTRs and identify structured sequences within them, but the current method treats all intergenic regions (IGRs), whether transcribed or not, as potentially harboring ncRNAs. The average IGR length in '*C.* P. ubique' is a meager 3 nucleotides, so the authors narrowed their search to a short list of IGRs that are longer than 100 nucleotides, known to contain the vast majority of previously identified functional RNAs. For comparison, the average IGR in *Saccharomyces cerevisiae* is 515 nucleotides and in humans is 12,000 nucleotides, so it is more difficult to construct a manageable list of potentially functional IGRs in these organisms using the size criterion alone.

Once candidate ncRNAs had been identified, a series of homology searches were used to further filter the structural hypotheses [5]. The first homology search used '*C.* P. ubique' IGRs as the reference sequences and looked for homologous IGRs in an ocean metagenomic database maintained by the CAMERA community (see references in [5] for details of programs and databases used). The second homology search looked for similarities between the IGRs and proteins in the NCBI nucleotide/protein sequence and CAMERA databases. IGRs similar to proteins were excluded. The authors were not limited by the novelty and incomplete curation of metagenomic sequences, and were able to predict (and avoid) unannotated protein coding regions using tools such as the MetaGene program and the Conserved Domain Database. CMFinder, a covariance analysis program that looks solely for RNA covariance and does not penalize sequences that show codon preservation, was used to align the '*C.* P. ubique' IGRs with their homologs and to predict a common secondary structure.

The program RAVENNA performed the third homology search, but in this case the consensus structures of IGRs, not their individual primary sequences, were used as

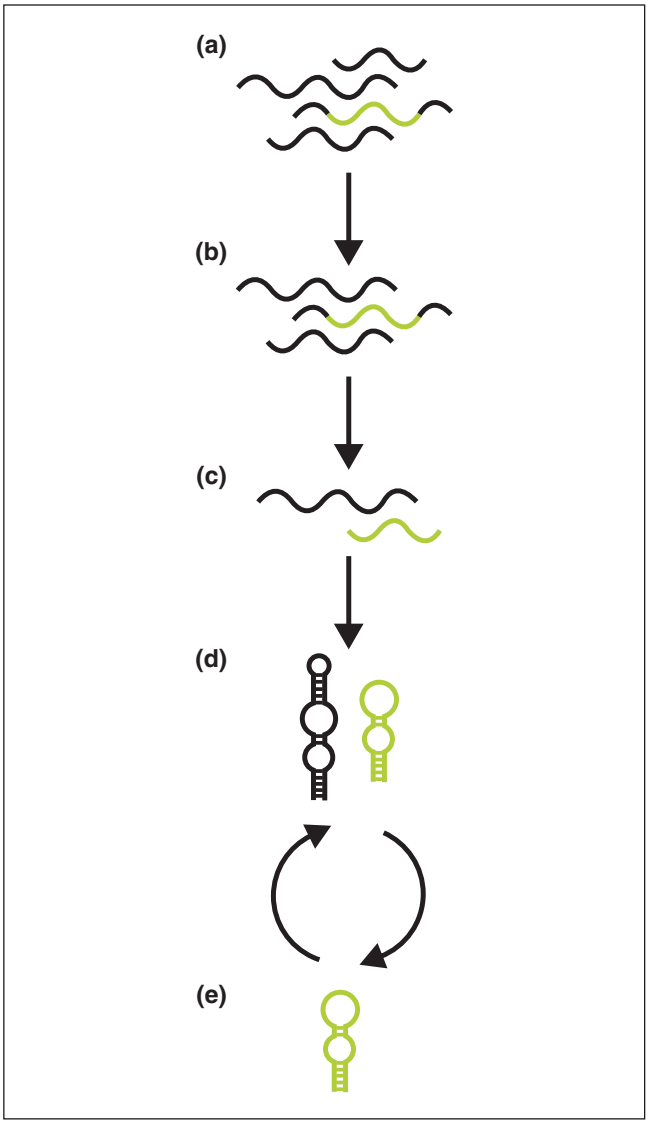

Figure 1

Flowchart of the computational methods used by Breaker *et al.* [5] in the identification of candidate ncRNA motifs. The steps in the process were as follows: **(a)** Identify IGRs by size, %GC content; **(b)** eliminate ncRNA motifs of known structure, such as tRNAs, rRNAs and annotated riboswitches; **(c)** find conserved IGR sequences in other genomes using BLAST analysis of the CAMERA database, and exclude protein-coding regions; **(d)** align IGRs and predict conserved secondary structures using CMFinder; and **(e)** search for homologs using conserved secondary structure criteria. Green indicates a candidate ncRNA motif.

references [5]. An ingenious aspect of this approach is that it is partially iterative: matches to the consensus secondary structures were used to refine those same secondary structures so that they could be used again to search for additional matches. This cycle can be performed any number of times, until a unique (with the exception of pseudoknots) and refined endpoint is achieved. This approach is reminiscent of its analog counterpart, the *in vitro* selection of functional RNA structures. In contrast, a purely statistical modeling of secondary ncRNA structure might not have been sufficient to identify homologs in other organisms. NcRNAs with the same function may fold differently as a result of having become adapted to the needs of a specific organism. That said, an alternative solution would have been to run the first homology search against the expansive set of genomic and metagenomic databases as opposed to just the CAMERA database. The resulting matches would be more comprehensive and could have possibly reduced the number of structure-based searches needed to arrive at a unique consensus structure.

By using consensus structures for comparison, the authors [5] could extend their studies from '*C.* P. ubique' to the enormous number of metagenomic sequences gathered from various environments, ranging from the ocean to acid mine drainage to mammalian intestines. Finally, the authors [5] looked at which genes appeared directly downstream of the putative functional ncRNAs in order to predict the pathways in which the ncRNAs might have a role.

A wide array of known ncRNAs were identified in the metagenomic sequences. These known ncRNAs include rRNAs, tRNAs, riboswitches, and the RNA components of RNase P and Signal Recognition Particle (SRP) [5]. In addition, eight novel structured RNA motifs were found. Four of these were unique to the metagenomic data, whereas the other four motifs include three possible *cis*-regulatory elements and a new *S*-adenosylmethionine-V (SAM-V) riboswitch class. One of the *cis*-regulatory elements, present upstream of the *rpsB* gene, had previously been characterized [8], but the others seemed to be novel, indicating that despite extensive genomic sequencing, many novel RNA motifs and functions may still remain to be found. The authors [5] also identified many IGRs that did not exhibit RNA structure but contained relatively short, conserved segments; these sequences may be protein recognition sites on the prokaryotic genophore.

The small '*C.* P. ubique' genome has a striking global AT bias (71% AT), hypothesized to be an adaptation to nutrient-poor environments such as the open ocean because of the fact that A and T are energetically cheaper to synthesize [9]. '*C.* P. ubique' probably cannot spend much energy creating regulatory proteins, but rather relies extensively on the interaction between metabolites and nucleic acids. Further evidence of the resourceful nature of these organisms is the fact that the motifs of RNase P RNA, SRP RNA, and two riboswitches were more than one standard deviation smaller in '*C.* P. ubique' than in other α-proteobacteria [5]. Interestingly, precisely because most of the identified sequences are AT-rich, the metagenomic structures may prove to be especially useful for the identification of

mechanistically important G and C residues in these structured RNAs. In prokaryotes, a high GC content is strongly correlated with structured RNAs and has been hypothesized to increase their stability, whereas there is no such correlation for genomes as a whole or for protein coding regions [10]. This analysis suggests that the disproportionately represented G-C base-pairs in the newly revealed pseudoknot of the SAM-V riboswitch [5] are probably particularly important for its structure and function.

Indeed, it can be argued that choosing metagenomic sequence information to scour for new riboswitches may have been particularly inspired. Small or streamlined genomes tend to be particularly AT-rich [9]. Such genomes may also have great need for small regulatory elements, and riboswitches are, in general, smaller than corresponding protein-based transcription or translation factors. Thus, examining the 'lifestyles of the small and AT-rich' may not only enable the quick identification of new and unique riboswitches, but also their functional sequences and structures.

References

1. O'Rourke JR, Swanson MS: **Mechanisms of RNA-mediated disease.** *J Biol Chem* 2009, **284:**7419-7423.
2. Waters LS, Storz G: **Regulatory RNAs in bacteria.** *Cell* 2009, **136:**615-628.
3. Hawkins PG, Morris KV: **RNA and transcriptional modulation of gene expression.** *Cell Cycle* 2008, **7:**602-607.
4. Hüttenhofer A, Vogel J: **Experimental approaches to identify non-coding RNAs.** *Nucleic Acids Res* 2006, **34:**635-646.
5. Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR: **Identification of candidate structured RNAs in the marine organism '*Candidatus* Pelagibacter ubique'.** *BMC Genomics* 2009, **10:**268.
6. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ: **Genome streamlining in a cosmopolitan oceanic bacterium.** *Science* 2005, **309:**1242-1245.
7. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR: **Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.** *Nucleic Acids Res* 2007, **35:**4809-4819.
8. Aseev LV, Levandovskaya AA, Tchufistova LS, Scaptsova NV, Boni IV: **A new regulatory circuit in ribosomal protein operons: S2-mediated control of the rpsB-tsf expression in vivo.** *RNA* 2008, **14:**1882-1894.
9. Rocha EP, Danchin A: **Base composition bias might result from competition for metabolic resources.** *Trends Genet* 2002, **18:**291-294.
10. Hurst LD, Merchant AR: **High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes.** *Proc Biol Sci* 2001, **268:**493-497.

Published: 2 September 2009
doi:10.1186/jbiol174