multiGenome-snoScan: A comparative approach to snoRNA annotation

Christoph Rau

Introduction and Background

The mechanisms which govern the proper functioning of an organism or cell are diverse. However, the most crucial cellular mechanisms are responsible for the transcription and translation of DNA to RNA and proteins. Without these functions, a cell would quickly die. The most common molecules created from DNA are proteins, however, many RNA molecules, such as rRNA and tRNA are also encoded in the DNA.

One such RNA molecule is small nucleolar RNA (snoRNA). SnoRNAs methylate other RNA molecules by binding a guide sequence to a corresponding sequence on the target molecule, and recruiting a methylase to methylate the target molecule. SnoRNAs have been found in both prokaryotes and eukaryotes, and typically act on ribosomal RNA and transfer RNA. Some evidence exists that suggest that snoRNAs can also act on mRNAs, but this is not well documented (Omer et al, 2003). These methylation events act to alter the behavior of the targeted molecules, and may also serve as recognition sites for other molecules (Caboche et al, 1977). Oddly, evidence suggests that methylation does not play a major role in determining the efficacy of rRNA. When methylation was prevented, rRNA showed only a small decrease in efficiency (Caboche et al, 1977). Clearly, methylation, and by extension the snoRNAs that cause it to occur, are performing other functions than purely kinetic ones.

In finding snoRNAs, scientists are aided by the conserved structure of the molecule. One class of snoRNAs is CD-box snoRNAs. These snoRNAs are characterized by two box motifs, labeled C and D, and two related motifs, C' and D', arranged as seen in figure 1.

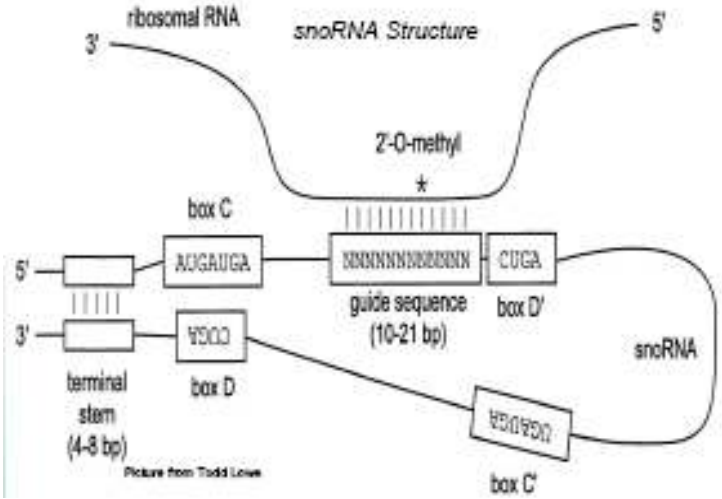

Fig. 1 Typical structure of a CD-box snoRNA. SnoRNAs are characterized by two motifs, labeled C and D, a terminal base-pairing stem, and guide sequence(s). (Lowe et al, 1999).

Regions which are not a part of the terminal stem, boxes or guide sequence are typically not conserved. Despite these non-conserved regions, the conservation of the other regions allows researchers to search for snoRNAs in a genome computationally.

SnoScan, a program designed by Todd Lowe, searches for snoRNAs by examining genomes for the above motif. It has been used to assist in the annotation of snoRNAs in *sacromyoces cervaece*, as well as various archaeal species, such as the *pyroccoccous* genus (Lowe et al, 1999. Omer et al, 2000). Unfortunately, every sequence that follows the CD-box motif is not necessarily a snoRNA. Therefore, the program requires a tedious examination by hand of each snoRNA for validity.

This paper describes an extension to snoScan, snoScan-multiGenome. The extension was designed to reduce the amount of human analysis necessary to identify snoRNAs by using computational biology maxims to increase the accuracy of the program. A tenet of computational biology is that if a genomic region has an important role, then it will be conserved across multiple genomes. This is especially the case in RNA-coding regions. SnoScan-multiGenome takes the results of multiple snoScans across multiple genomes, and finds the most conserved sequences, which are more likely to be snoRNAs than sequences that share no synteny with other genomes.

Design and Results

At its core, snoScan-multiGenome relies on three major programs to return relevant results. The first is snoScan (Lowe et al, 1999), the next is Blastz (Schwartz et al, 2003) and the third program is Multiz (Blanchette et al, 2004). Each of these programs has been described before, and snoScan-multiGenome acts as an efficient conversion and organizational tool for combining all three programs.

The program was trained against the *pyrobaculum* genus, and specifically *pyrobaculum aerophilum,* which had been hand-annotated by Todd Lowe. The false positive rate of the program was quite low (8% (2/25)), especially when compared to the original snoScan. On the other hand, many annotated snoRNA genes were not found by snoScan-multiGenome. It appears that there are many snoRNAs that are unique to an individual genome.

Two other interesting results appeared from the snoScan-multiGenome program. The first was the sharp decline in the number of potential snoRNAs identified by the program when the *pyrobaculum* genomes were compared against two other archaeal species, namely *Thermoproteus tenax* and *Caldivirga maquilingensis*. Only one putative snoRNA was identified in *tenax*, and no snoRNAs were identified in *Caldivirga maquilingensis* at all. This could indicate that snoRNAs are not especially highly conserved, which seems to match up with the high numbers of unique snoRNAs not detected by snoScan-multiGenome (snoScan itself detected many putative snoRNAs for each other genome).

The other interesting result involved the gene pattern surrounding the snoRNAs. Breaks were seen between the snoRNA's native genome and other genomes that had been blasted against it. A break is defined as a region where synteny is lost between two genomes. This synteny loss can be large (i.e. chromosomal rearrangement) or small (i.e.

a small viral insert or the like). Both types of breaks are seen in the *pyrobaculum* snoRNAs.

Conclusions, Analysis and Further Work

SnoRNAs are important molecules that regulate the methylation of other RNA molecules. SnoScan-multiGenome works to discover snoRNAs that are conserved in multiple genomes. These snoRNAs are more likely to be functional, since they appear in more than one genome. An interesting observation is that the number of snoRNAs that the program finds is significantly smaller than the number of snoRNAs that have been hand annotated and confirmed. This implies that many snoRNAs are unique to individual genomes, instead of common to multiple genomes. An interesting experiment would be to examine the unique snoRNAs for common features, such as ribosome target sites, genomic position, etc. It is possible that the changed snoRNAs target changed regions in the ribosomal RNA.

Another interesting observation involves the lack of snoRNA conservation between the *pyrobaculum* genomes and the other archaeal genomes tested. A single snoRNA was found in common between any of the *pyrobaculums* and *Thermoproteus tenax.* This may indicate that snoRNA evolution rates are higher than anticipated, or that the methylation sites have shifted quicker in either species. It is known that rRNA exhibits a slow mutation rate. However, since rRNA methylation is not crucial to the functioning of the ribosome, snoRNA mutation might proceed at a much faster rate (Caboche et al, 1977). This could also account for the high numbers of unique snoRNAs seen in the different *pyrobaculum* genomes.

Finally, interesting breaks were noticed in the *pyrobaculum* genomes near the multiGenome found snoRNAs. These breaks were either complete rearrangements (ex. snoRNA near PAE0163) or small inserts (ex. snoRNA near PAE3507). It is currently unclear as to exactly what is occurring at these sites. It is possible that these sites are naturally sites of recombination events, or it may be that these regions are used as recognition sites for viruses and transposons to insert viral genes into the genome.

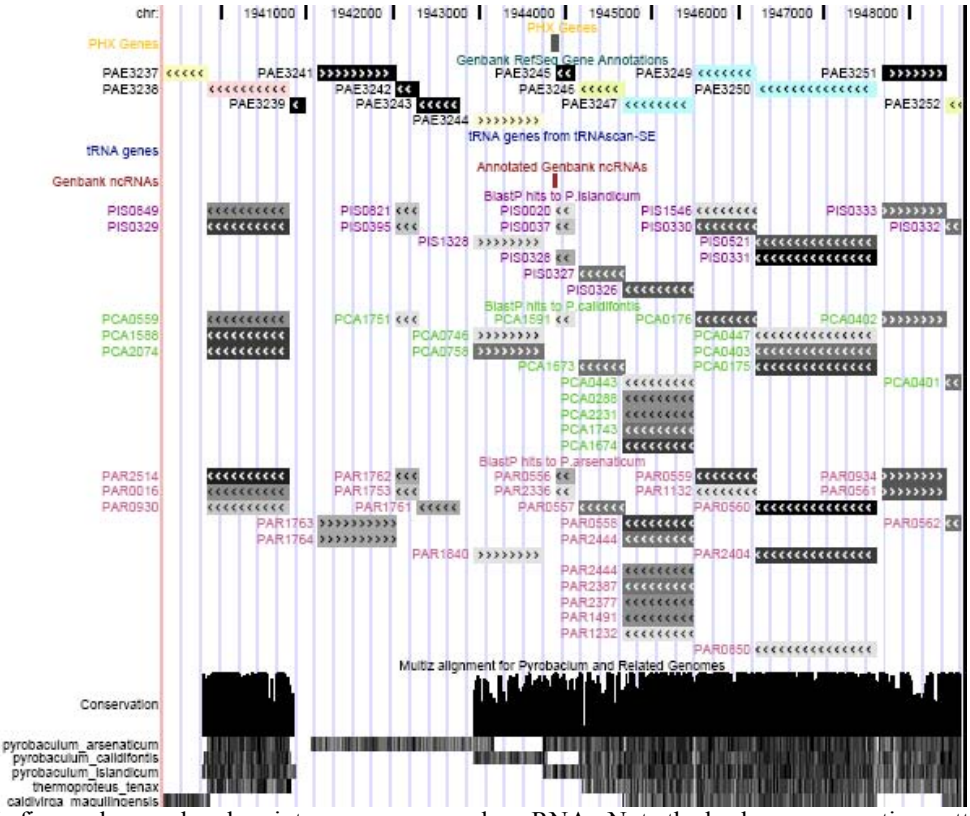

Fig. 2 This figure shows a break point near a conserved snoRNA. Note the broken conservation patter between the genomes at the bottom of the image, as well as the interruption in the gene progression in the blasted genomes.

This second hypothesis is supported by the fact that the number of breaks are much lower in the unique snoRNAs of the *pyrobaculums*. The fact that the multiGenome snoRNAs are conserved across multiple genomes may cause them to be preferentially targeted over unique snoRNAs, since a protein that recognizes a conserved snoRNA in one organism will recognize it in others. In the future, the connection between conservation and breaks should be explored, not only in other archaeal species, but also in other organisms.

Bibliography

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin K, Baertsch R, Rosenbloom J, Clawson H, Green E, Haussler D, Miller W. *Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner*. Genome Research 2004, 14:708-715.

Caboche M, Bachellerie JP *RNA Methylation and Control of Eukaryotic RNA Biosynthesis. Effects of cycloleucine, a specific inhibitor of methylation, onribosomal RNA maturation.* European Journal of Biochemistry. March 1977, 74(1):19-29.

Lowe TM, Eddy SR. *A Computational Screen for Methylation Guide snoRNAs in Yeast.* Science. Feb 1999, 283:1168-1171.

Omer AD, Lowe TM, Russel AG, Ebhardt H, Eddy SR, Dennis PP. *Homologs of small nucleolar RNAs in Archaea.* Science April 2000. 288:517-22

Omer AD, Ziesche S, Decatur W, Fournier M, Dennis, P. *RNA-modifying Machines in Archaea.* Molecular Microbiology. May 2003, 48:617

Schwartz S, Kent W, Smit A, Zhang Z, Baertsch R, Hardison R, Haussler D, Miller W. *Human-Mouse Alignments with BLASTZ.* Genome Research January 2003. 13:1:103-107.