Rnomics: Gene Prediction for Non-Coding RNAs

Kyle Friend
Bioinformatics
12/13/01

RNA plays a large role in numerous cellular processes, but this molecule is not the current focal point for genomics. Not only is RNA a substrate for translation, but multiple central cellular processes utilize RNA as an effector. RNA is largely responsible for, among others, splicing, translation, and signal peptide recognition. Moreover, within mammalian systems, a substantial portion (up to 95%) of the transcriptome is non-coding RNA (ncRNA) (1). This includes introns and functional RNA genes – ribosomal RNA for example. Given the substantial role RNA plays in cellular processes as well as the abundance of ncRNAs within the transcriptome, it is ironic that so much attention is paid to coding RNA (2).

Less is known about functional RNA when compared to protein. Few structures of catalytically active ribozymes have been solved or components of ribonucleoproteins (RNPs) whereas there is a wealth of protein structural information. In addition, lacking traditional start and stop codons, ncRNA genes are difficult to define. The above necessitate a novel approach to deal with this problem. To define RNA genes, it becomes necessary to explore non-proteincentric ideas. Comparative approaches are still utilized, but microarray technology serves as a direct determinant of RNA transcripts. In addition, techniques such as SAGE as well as structural considerations are employed.

Given the relative dearth of RNA sequence and structure information, it is rare to discover systems about which much is known – a useful exception is snoRNA. SnoRNA processes pre-ribosomal RNA, and has a defined set of structural characteristics (3). Using a computational technique and relying on a larger amount of sequence conservation within transcripts, novel snoRNAs were identified within yeast (4). However, most novel RNA genes are within poorly characterized categories, and at some point, more complicated genomes than that of yeast will be considered. Also, mammalian genomes contain a larger intron to exon ratio – a complication for this technique.

Comparing genomes is an extension of the above. Based on the principal of gene conservation, sequence homology is used as a gene predictor. Moreover, the pattern of conservation within ncRNA transcripts differs from those encoding protein. Degenerate base mutations within a codon are more prevalent; however, in ncRNA, mutations are more likely to occur at random, provided structural features and catalysis are maintained (5). Gene prediction success has been encountered in *E. coli* by comparison to a close relative, *Salmonella typhimurium* (repeated with *S. cerevisiae* compared with two other closely related family members) (6,7). Additionally, microarrays were designed to detect transcripts within the intergenic regions from *E. coli*, and extra ncRNA transcripts were identified (6).

Since screens will often identify large numbers of potential ncRNAs, some means must be used to pair down candidates. The *Arabidopsis* genome-sequencing project has constructed a library of expressed sequence tags, or ESTs. By simply searching the known sequences from these libraries, it is possible to ascertain what transcripts are produced and examine whether there are any without start and stop codons. Positive hits

are identified as potential ncRNA genes and then further analyzed to see if they encode non-translated RNAs. This is a useful strategy in *Arabidopsis* since a significant number of closely-related sequenced genomes is lacking for comparative analysis (8).

A genomics approach requires cDNA isolation from an organism of interest subtractive hybridization using cDNAs from a different part of that organism. An enriched pool is produced, limiting the number of transcripts to be identified. Consequently, identified genes tend to be tissue or treatment specific, and larger numbers of cDNA pools must be constructed to cover the genome. However, detection of messages with lower or tissue-specific expression is augmented; murine *G90* (9) as well as yeast ncRNAs (10) were identified in this fashion. An interesting twist on this approach uses known transcripts for the subtractive hybridization; cDNAs enriched in this manner should encode novel transcripts, a subset of which would be ncRNAs.

A concern with the aforementioned approaches is that they target poly-adenine containing transcripts whereas numerous ncRNAs do not contain poly-adenine tails. Serial analysis of gene expression, SAGE, should deal with this difficulty. SAGE can examine a total RNA library and is high-throughput, so whole genome screens are possible (11). As a caveat, multiple rounds are required to determine low-copy transcripts. In addition, restriction enzyme sites within transcripts are required – shorter transcripts are less detectable. In *C. elegans*, two short ncRNAs are involved in development (12,13), and SAGE is more likely to miss these transcripts. However, this technique was utilized successfully to identify several ncRNAs within the yeast transcriptome.

The minimum cut-off for most open-reading frame prediction programs is a gene coding for over 100 amino acids – shorter transcripts are missed. Using a total RNA library, the transcriptome can be resolved on a gel and size-selected. Selecting for smaller transcripts, approximately 200 small murine ncRNAs were identified (14). Potential degradation products contaminate the sequenced pool, and not all ncRNAs fall within the size-selection used.

Prediction based on cellular transcripts has complications; however, promoter and termination sites for RNA polymerase also define genetic loci. Where promoter and termination sequences are more highly conserved, gene identification is simplified, and this was successful in both *E. coli* and *S. cerevisiae* (15,16). Probes against intergenic regions of the yeast genome also identified several ncRNA genes using Northern analysis. Poorly conserved promoter and termination sites as well as an increased prevalence of intergenic regions complicate this approach, and serve as a sticking point in mammalian and plant genetics.

The previous predictive models have associated problems, but features of ncRNAs themselves may identify them as such. Most notably, ncRNAs maintain structural features, but these are practically poor gene prediction elements (17). Another approach examines transcript thermostability – functional transcripts should have higher thermostability, and several untested transcripts have been identified in this fashion (18).

Overall, ncRNA gene prediction is complicated, and no one technique is foolproof, so some combinatorial approach is required. Additionally, the above give little consideration to intronic RNA, the most prevalent cellular ncRNA (19). Some introns are functionally relevant, but discussion is beyond the scope here. Computational and

genomic approaches will be applied to this question as they already are to gene prediction.

1. Mattick J. S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO *reports*. **2:**986-991
2. Erdmann V. A. *et al*. (2001) Regulatory RNAs. Cellular and Molecular Life Sciences. **58:**960-977
3. Tycowski K. T. *et al*. (1996) A mammalian gene with introns instead of exons generating stable RNA products. Nature. **379:**464-466
4. Lowe T. M. and Eddy S. R. (1999) A Computational Screen for Methylation Guide snoRNAs in Yeast. Science. **283:**1168-1171
5. Rivas E. *et al*. (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Current Biology. **11:**1369-1373
6. Wassarman K. M. *et al*. (2001) Identification of novel small RNAs using comparative genomics and microarrays. Genes and Development. **15:**1637-1651
7. Cliften P. F. *et al*. (2001) Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. Genome Research. **11:**1175-1186
8. MacIntosh G. C. *et al*. (2001) Identification and Analysis of Arabidopsis Expressed Sequence Tags Characteristic of Non-Coding RNAs. Plant Physiology. **127:**765-776
9. Krause R. *et al*. (1999) Identification and characterization of *G90*, a novel mouse RNA that lacks an extensive open reading frame. Gene. **232:**35-42
10. Watanabe T. *et al*. (2001) Comprehensive isolation of meiosis-specific genes identifies novel proteins and unusual non-coding transcripts in *Schizosaccharomyces pombe*. Nucleic Acids Research. **29:**2327-2337
11. Velculescu V. E. *et al*. (1997) Characterization of the Yeast Transcriptome. Cell. **88:**243-251
12. Reinhart B. J. *et al*. (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature. **403:**901-906
13. Olsen P. H. and Ambros V. (1999) The *lin-14* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Developmental Biology. **216:**671-680
14. Huttenhofer A. *et al*. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO. **20:**2943-2953
15. Argaman L. *et al*. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Current Biology. **11:**941-950
16. Olivas W. M. *et al*. (1997) Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. Nucleic Acids Research. **25:**4619-4625
17. Rivas E. and Eddy S. R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics. **16:**583-605
18. Le Shu-Yun *et al*. (2001) Local Thermodynamic Stability Scores Are Well Represented by a Non-central Student's *t* Distribution. Journal of Theoretical Biology. **210:**411-423
19. Mattick J. S. and Gagen M. J. (2001) The Evolution of Controlled Multitasked Gene Networks: The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms. Molecular Biology of Evolution. **18:**1611-1630