

Computational genomics of noncoding RNA genes

Sean R. Eddy

Howard Hughes Medical Institute & Department of Genetics,
Washington University School of Medicine
Saint Louis, Missouri 63110 USA
eddy@genetics.wustl.edu

Biologists should not deceive themselves with the thought that some new class of biological molecules, of comparable importance to proteins, remains to be discovered. This seems highly unlikely.

- Francis Crick (1958)

The number of known noncoding RNA genes is expanding rapidly. Computational analysis of genome sequences, which has been revolutionary for protein gene analysis, should also be able to address questions of the number and diversity of noncoding RNA genes. However, noncoding RNAs present computational genomics with a new set of challenges.

We often hear that we live in the post-genomic world, where “all genes” have been systematically tabulated in databases, but the idea of a complete enumeration is just a convenient fable. Even with genome sequences in hand, our ability to identify genes is largely limited to relatively large, evolutionarily conserved, moderately to highly expressed protein coding genes. We know there are exceptions that fly below our radar - tiny genes, rapidly evolving genes, genes expressed in only a few cells at special times - but we have hoped, with some justification, that they aren't too important or numerous. Perhaps nowhere else are the tools and assumptions of genefinding and genome sequence analysis more fundamentally challenged than in the rapidly developing field of noncoding RNA genes.

Noncoding RNA (ncRNA) genes make transcripts that function directly as RNA, rather than encoding proteins. Transfer RNA and ribosomal RNA are textbook examples. Other structural and regulatory ncRNAs are known, but their number and importance have seemed marginal. Of course, absence of evidence is not evidence of absence. Gene discovery methods are biased. Most assume the Central Dogma, and look for genes that make messenger RNAs and have open reading frames. What if we looked specifically for noncoding RNA genes? As described in the following minireviews, several lines of recent research suggest that there are many noncoding RNA genes that have evaded genetic, biochemical, and molecular detection until now.

The power of complete genomes and computational sequence analysis has revolutionized molecular genetics. The workhorses of sequence comparison, BLAST and FASTA, are as well-known as PCR. We are accustomed to browsing - and sometimes even believing - gene predictions made by genefinding programs. Will computational genome analysis tools prove as useful for ncRNAs as they have proved for protein gene analysis? What are the prospects for tabulating and annotating ncRNA genes in genome sequences?

What we're looking for

Noncoding RNA genes come in more than one flavor [[Eddy, 2001](#), Erdmann, 2000; see other minireviews in this issue]. This makes it difficult to imagine a single ncRNA genefinding approach to find them all.

The best known ncRNAs have complex three-dimensional RNA structures and play roles as catalytic or structural parts of RNA-protein machines; examples include transfer RNA, ribosomal RNA, and spliceosomal RNAs. Many other ncRNAs, especially many of the recently discovered ones, act in a relatively unsophisticated manner by base pairing to a target RNA, and either regulate gene expression directly (for instance, by sterically occluding a ribosome binding site), or provide RNA targeting specificity for a protein-based regulatory or modification mechanism; examples include the micro RNAs (miRNAs) [Ambros, 2001], *E. coli* translational regulatory RNAs [Wassarman et al., 1999], and small nucleolar RNAs [Eliceiri, 1999]. Other noncoding RNAs seem particularly prevalent in dosage compensation, such as Xist RNA in vertebrates or roX RNAs in *Drosophila*, and in imprinted regions of chromosomes, such as IPW (Imprinted in Prader-Willi) and H19 [Kelley and Kuroda, 2000]. The “cis-antisense” ncRNAs are transcribed from the opposite strand of protein-coding genes, overlapping one or more coding exons in an antisense arrangement; an example with a genetic phenotype is the human SCA8 ncRNA gene, which is mutated in one form of spinal cerebellar ataxia [Nemes et al., 2000].

Isolation of a new RNA species with no significant ORF is not sufficient evidence of a new ncRNA gene. Many different cellular processes throw off nongenic noncoding RNA species, including RNA processing intermediates, transcription from retroposed repetitive elements, and low-level background genomic transcription. There should be evidence of a function, either by computational means (e.g. sequence or structure conservation) or experimental means (e.g. genetic phenotype). There should also be evidence that the RNA does not code for a small protein. Here the best evidence is almost certainly comparative genome sequence analysis. Conserved coding regions generally show a very different pattern of mutation (e.g. synonymous codon changes) compared to noncoding RNAs, and this pattern can be obvious even for short ORFs. Several stories in the literature in which ncRNAs have been confused with genes for small proteins, and vice versa, have been rectified by use of comparative sequence analysis [[Eddy, 2001](#)].

Genefinding

Several promising approaches give us a tenuous clawhold on the problem of de novo ncRNA gene prediction. None of these approaches is yet as reliable as protein-coding genefinders.

One approach to computational ncRNA genefinding is to predict RNA transcript initiation, termination, and processing, and find all predicted transcripts that do not have open reading frames. However, accurate prediction of even simple transcription units

remains an open computational problem. Noncoding RNA genes present an even harder problem; eukaryotic ncRNAs are transcribed by different polymerases - rRNAs by pol I, small structural RNAs like tRNAs and 5S RNA by pol III, and most other ncRNAs by pol II - and some are not independently transcribed at all, such as vertebrate small nucleolar RNAs, which are processed out of the introns of host transcripts. Nonetheless, the approach is certainly feasible in microbes. Two successful screens for ncRNAs in *E. coli* used promoter and terminator identification combined with comparative genome analysis to identify conserved noncoding regions [[Argaman et al., 2001](#), [Wassarman et al., 2001](#)]. Aside from promoter/terminator prediction, there is also statistical signal in splice sites that can be used to predict transcription units; one can probably identify a subset of spliced noncoding RNAs by successfully detecting small, closely spaced, clustered introns [[Lim & Burge, 2001](#)].

Another approach is to examine sequence content statistics such as base composition [[Carter et al., 2001](#)]. In most organisms, though, ncRNAs do not show strong sequence composition biases - certainly nowhere near as strong as the codon bias statistics that protein genefinders exploit. The success of such approaches will probably be very organism dependent. In hyperthermophiles, for instance, highly structured ncRNA genes are driven to high GC content presumably for reasons of RNA thermostability, and in otherwise AT-rich genomes this produces a strong composition bias [[Galtier & Lobry, 1997](#)].

Comparative genome analysis provides what may be the most powerful computational ncRNA genefinding approach currently described. In a pairwise alignment of two structural ncRNAs that are similar enough in sequence to be reliably aligned, but dissimilar enough to show compensatory base changes that conserve the secondary structure, a statistical test can detect that the pattern of mutations observed is nonrandom and consistent with RNA structure conservation. This can even be done without knowing the structure a priori, if the approach is combined with a statistical model of RNA folding [[Rivas et al., 2001](#)]. A weakness of this method is that it detects any conserved RNA secondary structure, including cis-regulatory mRNA structures in addition to independent ncRNA genes; conversely, it also fails to detect ncRNA genes that have little conserved secondary structure.

Four papers recently described different screens for ncRNA genes in the same organism, *E. coli*, which enables some comparison of the performance of these approaches, although experimental work to test many of the candidate genes is still in progress [[Argaman et al., 2001](#); [Carter et al., 2001](#); [Rivas et al., 2001](#); [Wassarman et al., 2001](#)] (Table 1). Taken together, the three screens that show experimental data have confirmed expression of about 31 different new ncRNAs. No screen identified more than about two-thirds of these as candidates. The comparative structure analysis approach of Rivas et al. had the best sensitivity for detecting these RNAs (~22/31), and also found more confirmed ncRNAs that the other screens missed (6). On the other hand, it also had fairly poor specificity, with only 11/49 tested candidates showing expression on Northern blots (though this may be an underestimate, since only RNA from exponentially growing cells was tested, whereas the Argaman and Wassarman papers showed several RNAs to be

expressed only in other conditions, particularly stationary phase). Interestingly, it appears unlikely that any of these screens has fully saturated the *E. coli* genome for new ncRNAs.

Similarity searching

Searching databases for homologues is a fast way to get a clue of what a gene might be doing. Unfortunately, BLAST and FASTA are not as powerful as one would like for ncRNA similarity searches. Protein sequence comparison is much more sensitive and specific; nucleic acids have a smaller and less informative alphabet. Significant cross-phylum similarities, which we take for granted with protein sequences, are seen only for the most slowly evolving ncRNAs like ribosomal RNA. Many ncRNAs conserve a base-paired secondary structure, though. For these RNAs, much more discriminative power would be gained by scoring both conserved sequence and RNA structure in a database search. This raises interesting algorithmic issues.

Figure 1 illustrates the power of taking the secondary structure into account when scoring RNA sequence alignments. Even a simple ad hoc scoring system can be useful. All of the power in distinguishing homologous from nonhomologous alignments comes from the scoring system, though, so it is desirable to assign alignment scores in a statistically rational manner. The score matrices for primary sequence alignments, such as the BLOSUM62 matrix, assign high scores to identities, moderate positive scores to conservative amino acid substitutions, and negative scores to dissimilar residues. The mathematical theory for statistically estimating and optimizing these scores is well understood [Altschul, 1991].

Thus a central issue in developing an alignment scoring method for RNA is, how do we make a scoring system that sensibly combines contributions from conserved secondary structure and conserved primary sequence? For example, how should we score an alignment of two identical Watson-Crick pairs, versus an alignment of two different (compensatory) Watson-Crick pairs? How much weight should these base-pair scores get, relative to the primary sequence alignment scores for single stranded residues? A satisfactory statistically grounded solution to this problem has been elusive. The RNA structure/sequence alignment literature is instead almost exclusively devoted to applying a great variety of clever algorithms to the problem of optimal RNA structure/sequence alignment, while using ad hoc scoring systems.

The theory necessary for extending sequence alignment scoring approaches to RNA secondary structure alignment scoring came from an unexpected direction. Over the last ten years, sequence alignment scoring has become understood in greater mathematical depth using hidden Markov model (HMM) formalisms borrowed, interestingly, from the fields of computational linguistics and speech recognition [Durbin et al., 1998]. A higher-order cousin to HMMs in computational linguistics, “stochastic context-free grammars” (SCFGs), can deal not just with primary sequence but also with nested long-distance pairwise correlations in sequences - which is exactly what is needed for modeling base pairing in RNA secondary structure. SCFGs provide a statistical framework for scoring secondary structure and primary sequence alignment simultaneously, allowing us to

estimate alignment score parameters from trusted RNA multiple alignments in much the same way that the BLOSUM matrices are constructed from the BLOCKS alignment database.

SCFG methods for RNA sequence/structure analysis have been known for some time [[Durbin et al., 1998](#)], but few practical SCFG programs are available. Almost the only example is the tRNA gene prediction program used by most genome annotation groups, tRNAscan-SE [[Lowe & Eddy, 1997](#)]. SCFGs are computationally complex. They require much more time and memory than primary sequence alignment algorithms. Until recently, the number of known ncRNAs requiring similarity search analysis has not been large enough to justify the development of generalized SCFG-based, BLAST-like search tools. The field has instead made do with carefully developed patterns or special-purpose programs that only search for homologues of one RNA or RNA family of interest [[Dandekar & Hentze, 1995](#)]. With computers getting faster, and the number of new ncRNA gene sequences growing rapidly, it is time to deploy practical SCFG-based database search programs.

The modern RNA world

Interest in the function and structure of RNA has been spurred by a notion of a primordial "RNA World" [[Gesteland et al., 1999](#)]. Now, though, it appears that many RNA genes are phylogenetically recent innovations that are well adapted to their modern roles in posttranscriptional regulation, RNA processing, and RNA modification. Because base pairing allows a small RNA to target another nucleic acid with great specificity, complementary RNAs may evolve more easily than specific RNA-binding protein domains; 15 nucleotides of RNA can do the job of a 100 amino acid protein domain. Gene regulation can be effected by simple mechanisms like occluding an important mRNA site. RNA might therefore be an ideal material for making small nucleic-acid-binding regulatory molecules. This observation dates back to Jacob and Monod, who suggested in their classic 1961 paper on operons and messenger RNA that regulatory genes were likely to make small RNAs [[Jacob & Monod, 1961](#)]. If this notion is true, we can expect that cells with complex posttranscriptional regulation will have many small RNA genes that have yet to be discovered. A major goal of current computational (and experimental) screens is to get a handle on the numerology of the modern ncRNA genes. Are there just a few? Or a great many?

The current situation in RNA is reminiscent of the early days of protein sequence analysis. Not too long ago, the protein sequence database was published on paper, and algorithms for rigorous sequence comparison were well known to the cognoscenti but were too impractical and expensive to run on the computers of the time. Then the sequence database expanded rapidly, and fast, practical, heuristic tools like BLAST and FASTA appeared forthwith. If we are indeed at the forefront of a significant expansion of known ncRNA gene sequences, it is time for RNA computational biologists to step up and apply our known body of theory to the development of practical analysis programs and well-organized databases.

Selected Reading

Altschul, S.F. (1991). *J. Mol. Biol.* 219, 555-565.

Ambros, V. (2001). *Cell* 107, 862-864.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. & Altuvia, S. (2001). *Curr. Biol.* 11, 941-950.

Carter, R. J., Dubchak, I. & Holbrook, S. R. (2001). *Nucl. Acids Res.* 29, 3928-3938.

Dandekar, T. & Hentze, M. W. (1995). *Trends Genet.* 11, 45-50.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK.

Eddy, S. R. (2001). *Nat. Rev. Genet.* 2, 919-929.

Eliceiri, G.L. (1999). *Cell Mol. Life Sci.* 56, 22-31.

Erdmann, V.A., Barciszewska, M.Z., Szymanski, M., Hochberg, A., de Groot, N. & Barciszewski, J. (2001). *Nucleic Acids Res.* 29,189-193.

Galtier, N. & Lobry, J.R. (1997). *J. Mol. Evol.* 44, 632-636.

Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds (1999). *The RNA World*, Second Edition. Cold Spring Harbor Laboratory Press, New York.

Jacob, F. & Monod, J. (1961). *J. Mol. Biol.* 3, 318-356.

Kelley, R.L. & Kuroda, M.I. (2000). *Cell* 103, 9-12.

Lim, L. P. & Burge, C. B. (2001). *Proc. Natl. Acad. Sci. USA* 98, 11193-11198.

Lowe, T. M. & Eddy, S. R. (1997). *Nucl. Acids Res.* 25, 955-964.

Nemes, J.P., Benzow, K.A., Mosely, M.L., Ranum, L.P. & Koob, M.D. (2000). *Hum. Mol. Genet.* 9, 1543-1551.

Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001). *Curr. Biol.* 11, 1369-1373.

Wassarman, K.M., Zhang, A. & Storz, G. (1999). *Trends Microbiol.* 7, 37-45.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001). *Genes Dev.* 15, 1637-1651.

| Strategy | Reference | candidates | tested | expressed | #/31 | uniquely found/31 |
|--|------------------|------------|--------|-----------|------|-------------------|
| Promoter/terminator/seq conservation | Argaman et al. | 24 | 23 | 14 | 14 | 2 |
| Seq conservation/microarray | Wassarman et al. | 60 | 60 | 17 | 18 | 2 |
| Sequence composition/structure stability | Carter et al. | 370 | - | - | 13 | - |
| Comparative secondary structure | Rivas et al. | 275 | 49 | 11 | 22 | 6 |

Table 1. Comparison of four screens for ncRNAs in *E. coli*, showing the number of predicted ncRNA genes, the number tested for expression by Northern blot, and the number found to be expressed. The total number of different expressed ncRNA genes identified by these screens was 31. The final two columns indicate how many RNAs out of this total were identified by each screen, and the number of unique RNAs found by each method. Since many of the RNAs have not been mapped yet, the last two columns are estimates based on approximate genome locations and observed transcript sizes.

