Pair Stochastic Context-Free Grammar Model in Comparative Sequence Analysis to Identify New Noncoding RNAs

Li Li
Biochem 218
15 March 2007

Introduction

A growing line of evidence has pointed to the importance of noncoding RNAs (ncRNAs) in the regulation of gene expression at multiple levels in both prokaryotes and eukaryotes. NcRNAs are defined as all RNA transcripts that lack protein-coding capacity. Recent studies have suggested that the human genome contains ~21,561 protein-coding genes, while the predicted number of transcribed genes is far higher, ~69,185 (5). In addition to the abundance of ncRNAs in the human and other mammalian genomes, these molecules execute a diverse array of functions. Besides the housekeeping ncRNAs such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear RNA (snRNA), small nucleolar RNAs (snoRNAs), RNAse P RNAs, and telomerase RNA, numerous other ncRNAs participate in regulatory function (4,5). In mammalian genomes, for instance, introns have not only been implicated in nucleosome formation, alternative pre-mRNA-splicing, and scaffold/matrix-attachment, but have also been shown to encode microRNAs (miRNAs) and repetitive elements (5). NcRNAs are involved in genomic imprinting, dosage compensation, and translational modulation through the RNAi pathway or by acting as natural antisense transcripts, all of which participate in the development of an organism (5). Thus, the biology of ncRNAs is rich and complex.

Attempts to identify new ncRNAs, however, have been particularly challenging. As many of the cellular mechanisms dispose of nongenic noncoding RNA species, scientists have proposed two principle criteria for confirming ncRNAs. First, ncRNAs need to be shown to have function; and second, there should be evidence showing that they do not encode for a small peptide (3). However, using genetic screen to identify new ncRNAs has given low yields because ncRNAs are usually variable in size, lacking in ORF, and relatively immune to point mutations (5). Computational analysis seems to be a more promising approach by scanning genomes for ncRNAs. Yet, unlike in identifying new protein-coding genes, ncRNAs sequences do not give strong statistical signals (6). The crux of the problem lies in the fact that ncRNA sequences diverge across phyla, making sequence comparisons difficult (5).

Comparative Sequence Analysis

The Three-Model Comparison

One of the better methods developed so far is by Rivas *et al.* and uses comparative sequence analysis to detect novel structural RNA genes by incorporating both the sequence and the secondary structure information. The method extends from concepts of previous work by Badger & Olsen. In Badger & Olsen's work, BLASTN program is first used to locate regions of significant sequence similarity between two bacterial species (6). Then, a program analyzes these

ungapped, aligned regions for evidence of coding structure (6). For instance, synomynous substitution would receive a positive score while dissimilar amino acid substitution would get a negative score (6). From this basic framework, Rivas *et al*. had four extensions: 1) the use of fully probabilistic models; 2) the addition of a third model of pairwise alignments constrained by structural RNA evolution; 3) the allowance of gapped alignment; and 4) the allowance of partial pairwise alignment to represent structural RNA (6).

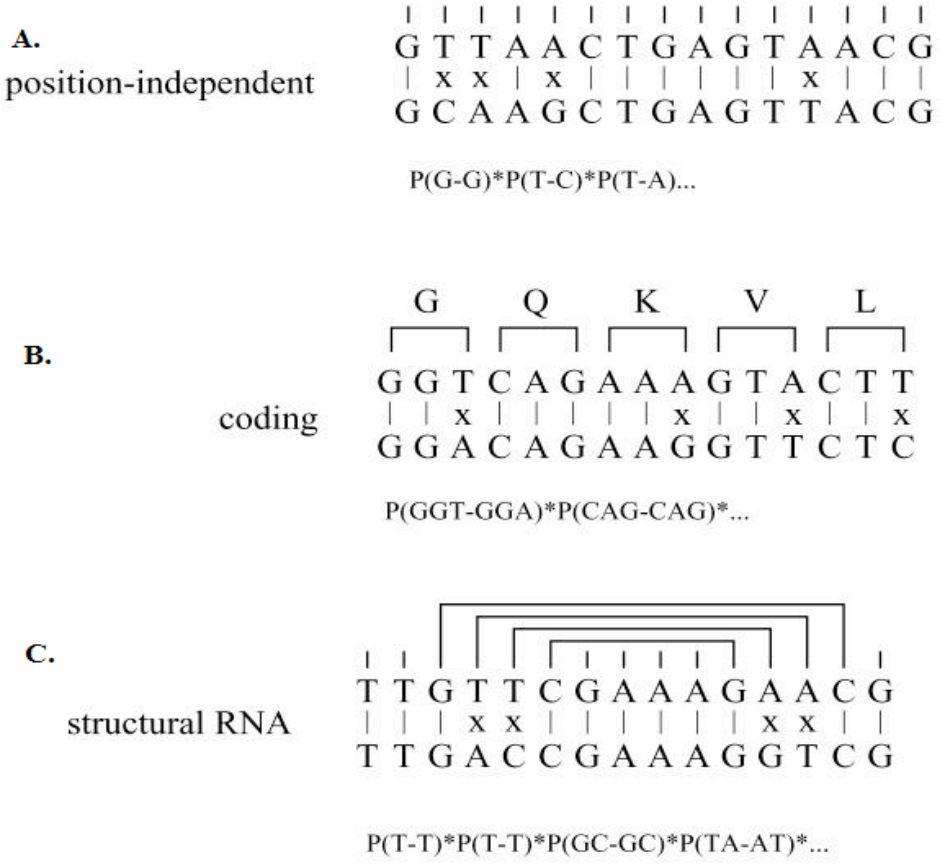

Figure 1. The Three Models. (A) The null hypothesis model in which mutations are position-independent. (B) The coding model where codons of homologous proteins are compared. (C) The RNA model that uses both sequence and its secondary structure for comparison. Note: Figure taken from (6).

The use of fully probabilistic models allows differentiation between coding, RNA, and null class classification of the genome by employing different evolutionary constraints. Essentially, pair hidden Markov models (pair HMMs) were proposed to be used for the null class control and the protein-coding class. The pair stochastic context-free grammar (pair-SCFG) was introduced to incorporate the secondary structures of the sequence. In the null hypothesis model, one assumes that mutations occur in a position-independent fashion (6). Thus, one examines each base pair separately, and calculate the alignment probability as the product of the probabilities of the individually aligned positions (see Fig 1A). This model is appropriate as a

control. In the coding model, one assumes that the aligned sequences encode homologous proteins (6). As one expects substitution of amino acid to be mostly synonymous, a table for the probabilities of correlated emission of two codons is constructed (6). Also, because the reading frame can be in six different orientations, the overall alignment probability is the sum of the alignment probabilities from the six frames and the assumption is that all six frames are equiprobable (see Fig 1B) (6). As this model is used in a rudimentary differentiation between coding sequences and junk DNA, the pair HMM work well in determining protein sequences.

The RNA model is more challenging, as one needs to differentiate structural RNAs from nonstructural sequences. Currently, the pair-SCFG (Fig 1C) is the widely used model to compare sequence structures during alignment. The model uses three states of substring end base-pairing and two types of emission probabilities (6). In general though, predicting RNA secondary structure involves two different approaches, one thermodynamic and the other comparative (1). In comparative analysis, one takes into account the covariation of homologous sequences to determine which base-pairings are preserved (1). Pair-SCFG is advantageous in discovering new ncRNAs because it is not restricted by the available RNA folding patterns, as is by other models (1). It is, however, very computationally labor-intensive (1). One possible way to refine the pair-SCFG is to overlay another program that looks at clustering. It is unlikely that a structured ncRNA to contain a single hairpin, as specificity would likely to require more spatial determinants such as multiple hairpins. Thus, the presence of hairpin clustering could be indicative a potential catalytic ncRNA or a multi-functional one. The clustering technique may be a rudimentary way of providing evidence that tertiary structures exist for these RNA sequences, even though to determining the actual tertiary structures may be too computationally complex. In addition, by creating a hairpin map, with the hairpin length information kept intact, one may align these hairpin maps to determine whether conserved tertiary folds might exist. This hierarchical clustering, from sequence to hairpins, from hairpins to "hairpin modules", may help lower false positives and offer a more complex picture of conserved tertiary structure of RNA (see Fig 2).

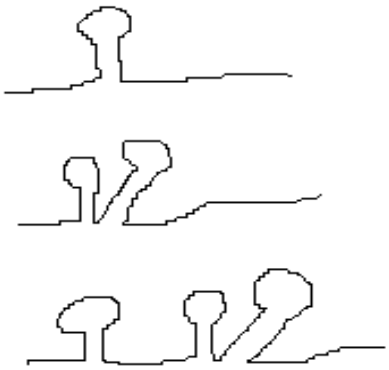

Figure 2. Hairpin clusters. The existence of hairpin clusters or repeating hairpin structure motifs may be indicative of higher

order conservation in structure.

In the thermodynamic approach, on the other hand, one takes into account base-pairing, base-stacking, and near-neighbor forces to calculate the minimum free energy (1). The existence of folded RNAs at their minimum free energy may not be the best assumption as RNAs in nature tend to be within 10% of their minimum free energies (1). MFOLD is the most commonly used program to fold RNA, with a prediction accuracy around 50% (1). However, as MFOLD does not work efficiently on large scale, it may not particularly useful in this case in searching for new ncRNAs.

Limitations of the Three-Model Comparison

One weakness that Rivas *et al.* pointed out with the three models is the introduction of the transition probability parameters in addition to the emission probability. The authors assigned arbitrary values to the transition parameters by first discriminating model-tested data and random sequence alignments (6). In the null hypothesis, the transition parameters should be similar for each paired letter of the sequence unless the sequence composition is biased towards a particular base composition. For instance, in hyperthermophiles, structured ncRNA genes have higher GC content presumably to allow RNAs more thermostability at high temperatures (3). The transition parameters for the coding model may be better determined if amino acids were assigned hydrophobicity values as hydrophobic amino acids tend to cluster together. This may be of some use for the null-hypothesis and coding models, but for the large part, the values may still be arbitrary. For the RNA model, it is difficult to determine whether the stem of a hairpin has a higher chance of growing than forming the loop. One could use a training set and determine the average the hairpin lengths and the probabilities of sequences in stem extensions and in loops. In all cases, the transition parameters will remain a difficult problem to resolve.

Another weakness in the approach Rivas *et al.* is discussed briefly by the authors. The models they use detect conserved RNA secondary structures, which would include the *cis*-regulatory mRNA structures as well. Thus, an algorithm may be needed to distinguish between the *cis*-regulatory and *trans*-regulatory structures. One approach that may be employed was developed by a group at Harvard. The group looked into *cis*-regulatory modules (CRMs) in DNA using a hierarchical mixture method under two assumptions: 1) eukaryote genes are not regulated by a single site, but by *cis*-regulatory modules that have multiple transcriptional factor (TF) binding sites, and 2) these TF binding sites also have specific motifs (8). This group's method analyzes the genome sequence at two levels. At one level, the method distinguishes a mixture of CRMs from the background sequence, and at the second level, it examines the motifs within each CRM. Because one scans the whole genome for ncRNA, it is possible that the structured sequence may be either a *cis*-regulatory DNA or mRNA structure. Although one may use the hierarchical mixture method to search for *cis*-regulatory sequences to differentiate from *trans*-regulatory sequences, one still has to differentiate between DNA and mRNA structure. One possibility is that the binding motifs for DNA is different than that of mRNA, though one can imagine a case where mRNA structure is similar to that of DNA, and the RNA may help sequester certain TFs from binding to the DNA sequence. However, one can use specific motifs to search for specific ncRNAs. A group in China, for example, used a probabilistic model and conserved primary and secondary motifs to search orphan C/D or H/ACA snoRNAs while the

usual approach is based on simple sequence complementarity to rRNAs or snRNAs (7). Other ncRNA-specific motifs can also be used to see whether hairpins overlap and whether they are in close proximity to those motifs. The other concern is that the regulatory regions of the mRNA may have multiple binding sites for proteins and that they may not exist as modules. In that case, a non-hierarchical search can be used to search for motifs. In both scenarios, an efficient way of lowering the number of false positives may be to use known *cis*-regulatory sequences to develop an algorithm to filter out sequences from the original model.

Finally, an obvious limitation of the comparative analysis using structural information is that ncRNAs with relatively little secondary structure will not be identified (3). These ncRNAs might be relatively conserved, but with little known function. In this case, the computational approach may not be the most appropriate; rather microarray experiments might better determine the existence of these nonstructural transcripts.

Testing the Models

To test the comparative three-model system, Rivas *et al.* generated potential structural RNAs and randomly shuffled the basepairs to test for specificity and adjust for false positives and false negatives (6). In principle, the RNA-generated data should give a positive score while the shuffled data would have the same composition as the RNA-generated data, but lack the correlative information, and thus should not return a positive score (6). The simulated data test given by Rivas *et al.* produced a frequency of 0.023 false positives and 0.081 false negatives at a threshold of 1.4 bits for the RNA posterior log-odds score. A 5-bit threshold was set for to lower the false positives. For known RNAs, however, the specificity degrades quickly for >90% identity in alignments (6). In either case, the performance of the models seemed promising as one can choose the percent identity from the BLASTN alignment that corresponds to the desired specificity.

What would be particularly interesting is to change the training set for a more specific model in ncRNA identification. In this study, the training set comprised of tRNAs and rRNAs, however, other training sets can be tested as well. In a recent study to identify miRNA, a control group and a training set were used that search for *distinctive properties* such as *structural features* such as hairpin length, hairpin-loop length, thermodynamic stability, base-pairing, bulge size, location, and distance of miRNA form the loop of its hairpin precursor; and *sequence features* such as nucleotide content and location, sequence complexity, repeat elements, and internal and inverted sequence repeats (2). Using miRNA as a training set can lead to a better understanding of the similarity in ncRNA structure. For instance, by comparing values of the training set used for a particular ncRNA class versus a global ncRNA search, one can gain more insight into variations in ncRNAs and possible adjustments to the parameters to obtain new results. Thus, using other training sets such as snRNAs or snoRNA, one can determine the dependency of the new ncRNAs search on the training set.

Conclusion

The emerging field of ncRNAs provides an exciting opportunity for new discovery, especially since the regulatory functions of ncRNAs are so diverse. Although the biology of

ncRNAs is fascinating, the process of identifying new ncRNAs is particularly challenging. Properties of ncRNAs such as size, composition, location that make them much harder to detect than proteins. Even computational methods in scanning the genome for ncRNA face many obstacles. In this evaluation of the three-model comparison to determine whether the aligned regions are protein, RNA, or neither, the logic, as well as the strengths and weaknesses of using Pair HMM and Pair-SCFG, were examined. Overall, pair-SCFG has much greater potential than other models in the field, which usually use known RNA folding patterns to search for ncRNAs. I introduced the notion of a hairpin map where clustering and repeating hairpin motifs can be indicative of conserved higher order structure. I also proposed expanding the training set, to determine the variations between ncRNA classes and to decide on the extent of bias established by the choice of the training set. With new findings derived from this computational search algorithm, a larger database of potential ncRNAs can be created and more insights into the ncRNA evolution can be gained.

References

1. Baird, SD, Turcotte, M, Korneluk, RG and Holcik, M. (2006) Searching for IRES. *RNA*. 12:1755-85.

2. Bentwich, I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett*. 579:5904-10.

3. Eddy, SR. (2002) Computational Genomics of Noncoding RNA Genes. *Cell*. 109:137-140.

4. Eddy, SR. (1999) Noncoding RNA genes. Current Opinion in *Genetics Dev*. 9:695-699.

5. Prasanth, KV and Spector, DL. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev*. 21:11-42.

6. Rivas, E and Eddy, SR. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*. 2:8.

7. Yang, J-H, Zhang, X-C, Huang Z-P *et al*. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res*. 34(18):5112-23.

8. Zhou, Q. and Wong, WH. (2004) CisModule: De novo discovery of cis-regulatory modules by heirarchical mixture modeling. *Proc. Natl. Acad. Sci. USA*. 101:12114-119.