RNA: Searching For Pattern, Finding Only Particular Relations

Dennis Mishler, 2002
Bioinformatics Final project/paper

Genome-wide sequence analysis has been present for much of the last decade. It has steadily improved to an extent that accurate predictions as to the number of genes present in an organism, whose genome is known, can be made. However, this same progress can not be said of noncoding RNA (ncRNA). The progress made in finding ncRNA is that there now appears to be a lot more ncRNA than anyone would have guessed. Currently, there is no obvious description or understanding of what key elements are needed to make a ncRNA. There are some obvious markers, such as RNA Pol III promoters. But this does not encompass all possible ncRNA, nor does it necessarily indicate how large the ncRNA is (1,2). Current methods in trying to find general ncRNA are not succeeding, but methods that search for specific types of ncRNA (such as tRNA, rRNA, siRNA, stRNA) may prove more useful.

The largest effort being undertaken to find ncRNA is headed up by Sean R Eddy. His most recent work has focused on algorithms that attempt sequence alignment based upon secondary structure (3,4). However, this secondary structure is based solely upon known primary sequences. Thus, the algorithms can only search for ncRNA that is similar to ncRNA already known. Even then, it is not perfect as the programs rely heavily upon compensatory mutations in attempting to find potential ncRNA sequences. It is often as right as it is wrong, and even then the preliminary results are not conclusive. This type of searching lacks the ability to find new types of ncRNA or ncRNA that have a non-canonical sequence or structure. The reason it lacks the ability to efficiently find ncRNA is that ncRNA secondary structure is not based so much upon its own primary sequence, but rather is often based upon the protein or nucleic acids it interacts with. Additionally, RNA secondary structure is often formed between bases that are relatively distant from one another (in terms of primary sequence). These reasons make progress in finding ncRNA genes unlikely without already knowing many sequences of ncRNA, which means knowing the identity of many different types of ncRNA, and even then there are often exceptions to the general sequences.

The current successes in finding ncRNA based on sequence alignment have been ncRNA-type dependent. For example the signal recognition particle RNA genes were determined using a conserved helix 8 motif found within these RNA, but not necessarily in other RNA (5). Similarly, short introns analysis works ok, but only when searching for short introns and only when using short intron-specific features (6). Even when using covariance models for the nucleotides, these programs do not find all the members. This has been particularly true of short intron analysis. Thus, it may be ideal, to focus on specific types of ncRNA until a better all encompassing algorithm can be found. It may prove more practical to be more specific in searching for ncRNA since RNA structure can be so variable.

The apparent slowness in finding certain ncRNA is problematic. It would be far easier if one could use sequence alignment to locate potential ncRNA and then test for function rather than first having to identify a function that may use a ncRNA and then attempting to determine its sequence. In C. elegans it has been shown that stRNA (lin-4, let-7) plays a very important role in development (7). Similarly, throughout

eukaryotes, siRNA has been shown to be an incredible silencer of gene expression. Despite, all the interest in these two topics, finding stRNA or siRNA via genomic or bioinformatic techniques has shown little, if any, progress. There are two main reasons, the first being that there is not yet a large enough sample population to create a reliable algorithm. The second reason should perhaps be the real focus for it may yield data more quickly in quantities that would allow reliable algorithms to be created. This second reason focuses not on the ncRNA, but rather on the things they may interact with. And rather than looking at what it might "base pair" to, it focuses on what interactions the protein or nucleic acid could make. Thus, if one could model potential domains or conformations that could interact with the protein/nucleic acid in question, one might be able to come up with a library of possible ncRNA structures suited to this particular protein. From this library, one could then carry out RNA sequence alignment in hopes of satisfying the needed secondary structure or even tertiary structure.

Using something other than ncRNA to develop an algorithm that efficiently finds ncRNA eliminates the need for large numbers of already discovered ncRNA. This is a problem because you often wish to find ncRNA that has only a few related known sequences and thus there exists no efficient algorithm. The basis for developing a better algorithm may at first not be ncRNA dependent. When creating the algorithm look at known proteins or proteins suspected of interacting with a particular type of RNA and attempt to model potential interactions that a string of nucleic acids could make with this protein. In other words, the molecular dynamics of a protein interacting with a potential RNA. These potential interactions can be narrowed down dramatically using the protein's sequence and structure to only take into consideration conserved sequences or motifs within the protein. These interactions can also be narrowed down by just considering areas within the protein that may lend themselves well to specific recognition of ncRNA (these domains may not have conserved sequence, but are still recognizable through protein structure, which is much more well understood). Once this has been done, and the database compiled, a regression from structure to sequence could be done to highlight potential sequences that could give the potential structures of interest. Using potential protein interactions to create an algorithm for finding ncRNA uses that which is better known, protein interactions, as opposed to using the that which is being searched for, the unknown ncRNA.

Early trials of this method can use currently well known proteins and the ncRNA that they interact with. The best example may be tRNA, which has been well studied for quite some time. Looking at a protein family like tRNA sythetases and then classifying what interactions are possible with the domain(s) that interact with tRNA might yield a number of potential structures and from this a number of likely sequences could be derived. A comparison between these sequences and the actual tRNA sequences could test how effective this process is or whether it is even feasible. I think this is a much more pro-active approach to finding ncRNA than the sit-back-and-wait-for-more-sequences approach. If there are as many ncRNA as is believed, then they must be doing something, for otherwise it would be an incredible waste of energy and resources. The question is what do they interact with and how? Based upon current knowledge the possibilities may be limitless, but some of the possibilities can be quantified. And those proteins or nucleic acids that they interact with can only be interacted with in a certain number of ways. If these ways can be determined or at least narrowed down to a reasonable number, then perhaps certain "consensus structures" can be identified without being observed directly. From these

structures, certain sequences with higher propensities to fold as such could be found and from these sequences an algorithm (or possibly even current algorithms) could find as yet unknown ncRNA. The question is not the algorithm in use, but rather the data the algorithms use to align sequences.

References:

1. Eddy, SR. Non-coding RNA Genes and the Modern RNA World. Nature Reviews.etics 2, 919-928 (2001).

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

3. Eddy, SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3, (2002).

4. Eddy SR, Rivas E. Noncoding RNA gene detection using comparative sequence. BMC Bioinformatics 2, (2001).

5. Regalia M, Rosenblad MA, Samuelsson T. Prediction of signal recognition particle RNA genes. Nucleic Acids Res. 30(15), 3368-77 (2002).

6. Lim LP, Burge CB. A computational analysis of sequence features involved in recognition of short introns. PNAS 98(20), 11193-8 (2001).

7. Reinhart, BJ, et al. The 21-nucleotide let-7 RNA regulates developmental timing in C. elegans. Nature 403, 901-906 (2000).