Proposal of a novel bioinformatic technique that searches for similar RNA secondary structures irrespective of primary sequence

RNA secondary structure is important for many biological phenomenon. For example group I and II introns, the hammerhead ribozyme, mRNA splicing, mRNA localization and ribosome processing all depend on RNA secondary structure. Interestingly, it has recently been shown that the latter two examples, mRNA localization and ribosomal processing depend on unique secondary structures that are not dictated by primary sequence (Chartrand et al., 1999; Klein et al., 2000).

mRNA localization is essential for proper development in several different organisms. It has been shown that proper localization of some mRNA's in human, *drosophila* and yeast depends not on primary sequence, but rather secondary structure (Chartrand et al., 1999; Ferrandon et al., 1997; Macdonald and Kerr, 1998; Ross et al., 1997). Recently, a lot of research has focused on genomic wide searches for other mRNA's that are localized using known transport proteins and finding other mRNA's that associate with them as detected by microarrays. However this type of approach excludes mRNA's that may be localized via unknown transport proteins. One simple approach to overcome this limitation would be to search whole genome sequences, irrespective of primary sequences, for secondary structures involved in mRNA localization. Since the secondary structure is conserved and not the primary sequence it is currently impossible to use bioinformatic approaches to find candidate mRNA's that could be localized.

Recently studies have shown that ‘K turn' (also known as the box C/D motif), found in snoRNA’s, is also involved in mRNA splicing (Watkins et al., 2000). Moreover, 8 examples of this motif have been found in ribosomal RNA. This indicates that this motif is important for many different biological processes. However there is no consensus primary sequence for this motif, therefore more examples of this motif cannot easily be searched for using bioinformatic techniques.

To date there are a sundry of bioinformatic techniques to search, score, compare and profile nucleic acid and protein primary sequence (BLAST, FASTA), protein secondary (GORE) and even protein tertiary structure (RMS fitting, Refine method). However, there as of yet, does not exist a method for searching nucleic acid sequences to find common secondary structure motifs that are not dictated by primary sequence. In other words, one cannot search for similar secondary structures that may have very different primary sequences.

I propose a technique that searches for RNA secondary structure elements irrespective of primary sequence. This can be accomplished by taking a given set of mRNA’s (for example, the known ORF’s in yeast) and determining their secondary structure using Michael Zucker’s M-FOLD (Zuker and Stiegler, 1981). M-FOLD takes into account thermodynamic parameters to determine the optimal fold for a given sequence. In addition M-FOLD takes into account reactivity of certain nucleotides to modifications, phylogenetic data and long-range interactions. M-FOLD can be used to create a database of folded RNA’s. This database can then be queried using a Hidden Markov Model (HMM) of the secondary structure of interest. For example the over 15 known examples of the K-turn motif can be used to train and produce a HMM. However,

the HMM would not be trained to emit a primary sequence rather it would emit secondary structural elements (for example S for stem-loop, P for hair pin, L for loop and B for bulge. In actuality there can be many different states for each of the nucleotides. Limiting the number of states will have the cost of lower resolution search, but will have the advantage of speed.) The database of folded RNA's would also be encoded as structural elements, thereby allowing the HMM to pick out similar secondary structure motifs. One advantage of using a HMM driven search is its flexibility. It is conceivable that a hairpin is important to function, however the length of the hairpin is not important. The HMM, having seen many examples, would allow for the variance in lengths by creating alternate paths to other elements after reaching the termination of a particular element and attach an appropriate probability to the alternate paths based on their occurrence in the training set. Also HMM's weigh certain elements based on their occurrence, so fluctuations (stem loop here a bulge there) can be tolerated. A technique similar to this was used to pick out proteins that were distantly homologous and was 74% successful (Geetha et al., 1999).

There are several considerations for getting this technique up and running. First, the technique needs to be tested. This can be done using the K-turn motif and a small database of folded RNA's. A HMM can be produced by using the known examples of this motif. A few examples will be withheld from HMM training, but will occur in the database of folded RNA's. This will serve as a cross validation technique. The HMM should pick out the known examples of K-turns that were not used in it's training. Another consideration is the many suboptimal folds that M-FOLD can produce (Zuker, 1989). Suboptimal folds are folds that don't maximize thermodynamic and structural

considerations and are usually quite different than the ‘optimal’ structure. Some of the suboptimal folds may actual score better that optimal folds. To have a more thorough search suboptimal folds should be included in the database and weighted according to their deviance from optimal folding (a score M-FOLD produces). The advantage of including suboptimal fold is a more thorough database, but the disadvantage is an increasing in data bins by δN (where δ = the number of suboptimal folds included and N = the number of RNA sequences folded by M-FOLD).

One obvious disadvantage of using HMM’s is that they are only as effective as the set of examples used to train the HMM. Fortunately for the K-turn motif several examples are known, however only a few (six) examples of putative secondary structure elements are known for mRNA’s that are localized (although, this would still be sufficient for training). Another downside to this technique is that some of the most important RNA structural elements are in noncoding regions. This can be overcome by separating genomic sequences into two categories: ORF’s and intergenic regions and have both categories folded using M-FOLD. It would be interesting to see if some secondary structure elements are biologically relevant in both coding and noncoding regions. It will also be interesting to determine all the mRNA’s that associate with RNA recognition proteins. This can be done by using the existing sequences that are known to interact with an RNA Recognition Motif (RRM) and develop an HMM from the RNA sequences that interact with RRM's. This will produce an HMM that search for all mRNA’s with a secondary structure similar to ones that associate with RRM’s. This brings up a very important point, that although this technique will be able to search out

similar secondary structure elements, they will still need to be tested for biological relevance using biochemical techniques.

The wealth of bioinformatic tools has neglected RNA secondary structural searches. With the completion of several genome sequences, it is now more relevant than ever to develop a RNA secondary structure search tool. This can be done by creating databases of folded RNA's using M-fold and searching these databases with a HMM trained by a specific secondary structure. A technique of this type will be very powerful in finding additional RNA secondary structures that have a biological role. Such as secondary structures that mediate mRNA localization, ribosomal processing, mRNA splicing, mRNA export out of the nucleus and post transcriptional regulation. Not only would this type of technique find other RNA molecules involved in these processes, but it could also be used to determine how these elements have evolved to participate in their role.

References

Chartrand, P., Meng, X. H., Singer, R. H., and Long, R. M. (1999). Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. Curr Biol *9*, 333-6.

Ferrandon, D., Koch, I., Westhof, E., and Nusslein-Volhard, C. (1997). RNA-RNA interaction is required for the formation of specific bicoid mRNA 3' UTR-STAUFEN ribonucleoprotein particles. Embo J *16*, 1751-8.

Geetha, V., Di Francesco, V., Garnier, J., and Munson, P. J. (1999). Comparing protein sequence-based and predicted secondary structure-based methods for identification of remote homologs. Protein Eng *12*, 527-34.

Klein, D., Schmeing, M., Nissan, P., Moore, P., Steitz, T., Manuscript in preparation

Macdonald, P. M., and Kerr, K. (1998). Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA. Mol Cell Biol *18*, 3788-95.

Ross, A. F., Oleynikov, Y., Kislauskis, E. H., Taneja, K. L., and Singer, R. H. (1997). Characterization of a beta-actin mRNA zipcode-binding protein. Mol Cell Biol *17*, 2158-65.

Watkins, N. J., Segault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., and Luhrmann, R. (2000). A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP [In Process Citation]. Cell *103*, 457-66.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science *244*, 48-52.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res *9*, 133-48.