

# Genome wide mRNA probing

Matthew Garrett

March 26, 2004

## 1 Background

mRNA localisation is a powerful mechanism for ensuring that protein production occurs only in the area of the cell where the product is desired. Rather than requiring large quantities of protein to be produced and then localised, a small quantity of RNA can be localised to the desired location and translation can occur in place. At the same time, translational control can then take place in situ rather than requiring translation followed by localisation and a consequent lag in protein availability.

mRNA localisation was first observed by Jeffery et al in Ascidian eggs. Much of the following work has also centred around eggs due to the larger cells requiring decreased resolution in order to observe non-uniform distribution. However, advances in technology have allowed observation of localisation throughout many different cell types, and it is now believed to take place in many (if not all) polarised cells.

mRNA structural regulation has been observed throughout eukaryotes, though is perhaps most studied in *Drosophila* development. Examples include bicoid mRNA which is localised along the anterior-posterior axis in order to generate a morphogenic gradient which in turn aids in the setting up of polarity and nanos, which in contrast is distributed throughout the cell but is translated only at the posterior pole with the control mechanism believed to be related to mRNA sequence.

Localisation is generally mediated by sequences in the 3' untranslated region of the mRNA being recognised by trans-acting proteins, though such sequences may be found elsewhere in the transcript. The bound protein is then responsible for the localisation of the mRNA. The mechanisms for this are somewhat outside the scope of this project, but are reviewed in Kloc et al 2002.

The sequences responsible for the regulation are currently poorly characterised. In general, it seems likely that some combination of structural and sequence specificity is required in order to function correctly. Localised transcripts generally contain multiple elements which are at least partially redundant, though full localisation efficiency depends on all being present. These sequences often interact or overlap with those responsible for translational control, which suggests that the two processes are linked. A reasonable hypothesis would be to suggest that some of the proteins involved in localisation remain bound to the transcript and block translation until the product is required.

Analysis of the localisation sequences generally suggests that both primary sequence and secondary structure are involved in their functioning. Directed mutation of the localisation sequence in the K10 transcript in *Drosophila* demonstrated that disruption of the secondary structure led to complete failure of localisation, whereas modification of certain parts of the primary sequence led to a slight reduction in efficiency. In order to allow computational analysis of localisation elements, it is therefore impractical to rely on applications such as BLAST as these will fail to take into account how similar sequences may result in different structures along with dissimilar sequences resulting in functionally equivalent structures. Instead, a mechanism that takes into account both primary sequence and secondary structure is required. RSEARCH is an algorithm that does just this, taking a primary sequence and accompanying structure (either experimentally determined or predicted based on the primary sequence) and then using that to generate a tree representing the search sequence. In contrast to BLAST, a single base difference in an otherwise homologous sequence may generate a significantly reduced score, while a compensatory change further along the sequence may raise it once more.

## 2 Proposed work

The existence of whole genome sequences and RNA structure prediction software makes it practical to search for areas of RNA structure and then seek homology between genes that follow similar localisation patterns, allowing for the identification of putative structural regions responsible for this localisation.

In contrast to nuclear DNA, mRNA is single stranded. As a result, rather than

being contained within a double helix, the bases on the mRNA strand are free to bind with each other. Given a sequence, the secondary structure of the RNA can be predicted using the algorithm described by Zuker and Stiegler. The predicted structure is determined by finding the most thermodynamically favourable set of binding interactions in the sequence, based on the assumption that the folding is a thermodynamically driven process rather than being mediated by a more complicated process. The amount of energy released will be primarily determined by the number of base pairs that are formed as the folding takes place, with more structure resulting in more binding and hence more energy release.

Katz and Burge describe a protocol for determining whether there is any bias towards RNA secondary structure over a genome. In summary, this involves comparing the minimum free energy involved in the folding of the mRNA sequences to the energy obtained by folding a randomised version of the same sequence. If a genome-wide bias existed, the folding energy in the actual sequence would be significantly more negative than that of the randomised sequence. The same mechanism can be used to look for more localised structure. Rather than attempt to identify stem-loop structures programatically, we can simply compare whether a region of RNA is significantly more thermodynamically favourable than a randomised version of the same.

This method will obtain a set of structural elements. However, this in itself is not a value judgement. In many cases, the presence of structure may be due to sheer chance. For useful conclusions to be drawn, it is necessary to be able to obtain a set of sequences that exhibit some amount of homology. The RSEARCH algorithm described above allows for calculating homology based on both the primary sequence and secondary structure of the search string. Searching for homology between identified structural sequences should then allow for grouping of families of elements.

Once smaller groups of elements have been found, attempts will be made to correlate these to gene function or behaviour. A first step may be to find sequences that have been experimentally noted. UTRdb contains a wide set of regulatory elements found in UTR regions, and this will be helpful in assigning preliminary functions to discovered families. Perhaps more important will be investigation of

families with several structural elements in common.

Once identified, there is obvious scope for experimental testing of the putative regulatory elements.

### **3 Progress**

At this point, code has been written to accomplish the first stage of this work. A Perl script takes each input sequence in turn and uses a 50 base sliding window to look at subsequences. Each 50 base sequence is then folded using the Vienna RNA folding prediction package which implements the Zuker algorithm. This returns a value for the minimum free folding energy, a measure of how much energy is released by the folding of the predicted structure. This value is then recorded and the sequence randomised. The randomised sequence is then folded in the same way. This randomisation and measurement process is repeated several times in order to generate a distribution, and the initial value is then tested against this distribution to determine whether the folding energy of the original sequence is significantly different to that expected for a sequence with that base composition. Subsequences which display significantly more structure than expected are recorded and written out to a FASTA file along with a measurement of the significance of their deviation.

In testing, this method has successfully located the RNA sequence located 681 bases into the 3' UTR of K10 that is responsible for its localisation. Running it over the entire 3L chromosome reveals several thousand items of structure. The majority of this is likely to be noise, so the next step of determining homology is required in order to filter this down to manageable proportions. A skeleton for homology checking has been written, and work is currently in progress to determine the best set of parameters for RSEARCH in order to perform this homology search.