Computational and Mathematical Methods in Medicine,
Vol. 7, No. 1, March 2006, 37–43

Taylor & Francis
Taylor & Francis Group

Mathematical and biological modelling of RNA secondary structure and its effects on gene expression

T. A. HUGHES†* and J. N. MCELWAINE‡

†Pathology and Tumour Biology Section, Level 4 Leeds Institute for Molecular Medicine, St James's University Hospital, Leeds University, Leeds LS9 7TF, UK
‡Department of Applied Maths and Theoretical Physics, Centre for Mathematical Sciences, Cambridge University, Wilberforce Road, Cambridge CB3 0WA, UK

(Received 26 May 2006; revised 12 June 2006; in final form 23 June 2006)

Secondary structures within the 5′ untranslated regions of messenger RNAs can have profound effects on the efficiency of translation of their messages and thereby on gene expression. Consequently they can act as important regulatory motifs in both physiological and pathological settings. Current approaches to predicting the secondary structure of these RNA sequences find the structure with the global-minimum free energy. However, since RNA folds progressively from the 5′ end when synthesised or released from the translational machinery, this may not be the most probable structure. We discuss secondary structure prediction based on local-minimisation of free energy with thermodynamic fluctuations as nucleotides are added to the 3′ end and show that these can result in different secondary structures. We also discuss approaches for studying the extent of the translational inhibition specified by structures within the 5′ untranslated region.

Keywords: UTR; RNA folding; Dynamic programming; Translational regulation

1. Introduction

Gene expression is tightly regulated at several separate stages to establish and maintain appropriate expression levels. Much research attention has focused on transcriptional regulation while, until recently, there has been less interest in the regulation of translation of specific gene products. However, in the last few years, misregulation of translation has been shown to contribute to the pathogenesis of numerous human diseases, such as cardiac hypertrophy, ataxia, some neurodegenerative diseases including Huntington's and Alzheimer's and cancer [1,2], therefore mechanisms controlling gene-specific translation are becoming more intensively studied. The major mechanism for initiation of translation is cap-dependent scanning. This requires binding of the molecular initiation machinery to the mRNA cap at the 5′ end of the molecule and subsequent scanning through the 5′ untranslated region (UTR) until the machinery arrives at and recognizes the initiation

*Corresponding author. Email: t.hughes@leeds.ac.uk. Tel.: + 44-113-3438624. Fax: + 44-113-3438702.

Computational and Mathematical Methods in Medicine
ISSN 1748-670X print/ISSN 1748-6718 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10273660600906416

codon of the reading frame. At this point, more factors are recruited and protein synthesis starts. 5′ UTR sequences can modulate translation of specific mRNAs by interfering with either the recruitment or scanning of translation factors [3]. For example, 5′ UTRs containing regions of secondary structure, i.e. regions where intra-molecular base pairing folds the molecule, can have a particularly potent inhibitory effect on translation [4]. Only a minority of genes express these structured 5′ UTRs, but this minority is notable since it contains many oncogenes, tumour suppressors and other genes associated with cell proliferation or disease [5]. In addition, regulation of the effects of RNA structure is significant since it acts as a point of gene misregulation in cancer [6].

Our understanding of the regulation of gene-specific translation, however, is limited by our inability to predict accurately the structures formed by 5′ UTR sequences. A number of computer algorithms have been developed to predict RNA secondary structures (examples, Vienna RNAfold [7] and mfold [8]). These use dynamic programming methods to identify the structure with the global-minimum free energy for any input sequence. The free energy (ΔG, a negative value) is calculated as the sum of the stabilising contributions of each base pair (negative values) and the destablising effects of un-paired bases forming loops or bulges (positive values) as determined experimentally [9]. Algorithms have been used with some success to predict structures of certain RNAs, such as ribozymes, but have proved less useful for prediction of 5′ UTR structure, although results can be improved when additional, experimentally-determined, structural information is also used to inform the algorithm [10].

RNA molecules are synthesised by sequential addition of nucleotides to the 3′ end of an extending linear chain by RNA polymerases. Therefore, rather than folding into a secondary structure as a complete molecule, folding actually occurs progressively from the 5′ end during on-going synthesis of 3′ sequences. Similarly, the passage of translational machinery along the complete molecule in a $5'-3'$ direction causes local and general structural unfolding with de novo folding occurring from the 5′ end as it emerges after scanning. A consequence of this is that initial base pairing is more likely between bases that are relatively close on the linear chain, since those more distant in a 3′ direction are not yet available. Of course, the structures formed may refold as more sequence is released from the polymerase or the scanning translational machinery, but only if the free energy of the new structure is more favourable and importantly, if thermodynamics allow that the activation energy required for unfolding of the existing structure is overcome. These considerations mean that the RNA may not actually fold into the structure with the global-minimum free energy. A key point about the current folding algorithms is that they attempt to find this global-minimum free energy and assume that all the possible configurations in the energy landscape are possible; hence the history of how the RNA structure arose does not influence the prediction and this aspect of the biology is not taken into account. In addition, modelling accurate structures is only half the battle, since understanding of how and to what extent these structures influence translation is required to gain insight into their regulatory effects. We have no clear mechanism of linking 5′ UTR structures to the extent of their translational inhibition other than by experimentation, aside from the generalisation that 5′ UTRs with lower minimum free energies may have greater inhibitory effects. In this article, we describe ways in which modelling 5′ UTR structures and their effects on translation can be imbued with a more biological approach with a view to gaining greater understanding of the importance of 5′ UTR structure in gene regulation and disease.

2. Results and discussion

2.1 mRNAs fold progressively from the 5′ end—implications for modelling

We have examined how the minimum free energies and RNA structures determined by one of the most-used folding algorithms vary with increasing length of RNA chain as if emerging from synthesis or translational scanning by analysing progressively longer sections of test mRNA using mfold v3.2 (default settings; http://www.bioinfo.rpi.edu/applications/mfold/old/rna/). We have used the human axin2 mRNA with 5′ UTRb (axin2 can express three alternative 5′ UTRs) as our test sequence since we have previously determined that its structure causes substantial translational inhibition [11]. We have extended the RNA in 10 nucleotide sections, starting with only 10 through to 100 nucleotides into the coding region. We show minimum free energies (figure 1A), free energies per nucleotide (figure 1B) and a summary of the predominant structure for critical lengths highlighting significant base pairings (figure 2).

Minimum free energy decreases with the length of RNA since more bases are available for pairing (figure 1A). Free energy per nucleotide initially decreases with increasing length of RNA to reach two low points (at 60 and 120 nucleotides) before increasing gradually towards

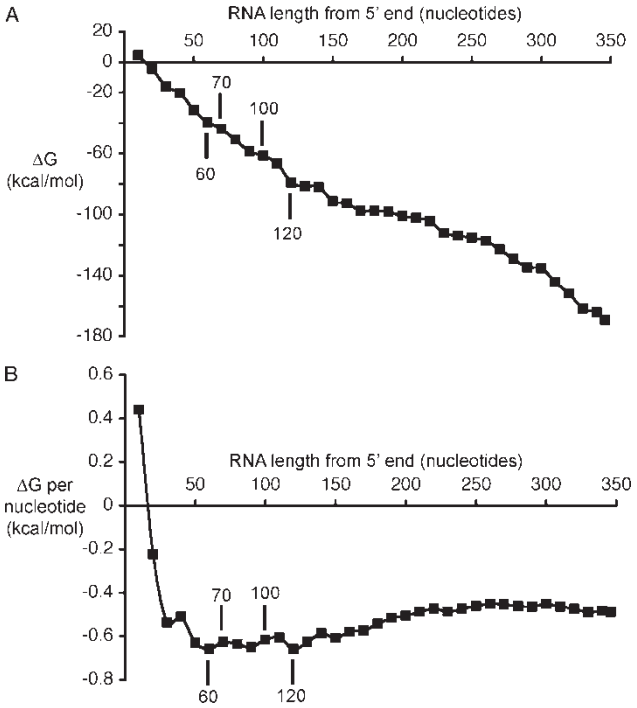

Figure 1. Modelling the effects of RNA chain extension on determination of minimum free energies. Increasing lengths of the human axin2 5′ UTRb mRNA sequence were analysed using mfold, starting with the most 5′ 10 nucleotides only, then the most 5′ 20 nucleotides and so on. The minimum free energy (ΔG) and minimum free energy per nucleotide of the resultant structures are shown (figure 1A and B, respectively). The data for RNAs of certain lengths are highlighted: 60 and 120 nucleotides, since these have the lowest free energy per nucleotide and form stable structures (figure 2B and C) and 70 and 100 nucleotides, since these have particularly small decreases in minimum free energy from their previous shorter chain lengths and their predicted structures are likely to misleading (see Results and Discussion).

A

Length of RNA (nucleotides)	Structures *(sections at least 80% base paired)*	Figures
60	stem-loop *(11-26 bind 47-30)*	Figure 2B
70	no similarity to 60 *(2-6 bind 63-67)*	
80	'60' stem-loop; 61-80 form separate structure	
90	'60' stem-loop; 61-90 form separate structure	
100	no similarity to 60, 70, 80 or 90 *(6-15 bind 88-98)*	
110	'60' stem-loop; 61-110 form separate structure	
120	Double stem-loop: '60' and new *(54-76 bind 110-82)*	Figure 2C
130-270	'120' double stem-loop; 130+ form separate structures	
280	'120' structure broken; '60' stem-loop maintained *(62-67 bind 279-274 and 91-112 bind 270-251)*	
290	'60' stem-loop maintained; remainder folds differently *(61-71 bind 285-274)*	
300	'120' double stem-loop returns; 122-300 forms separate structures	
310-350	'120' structure broken; '60' stem-loop maintained 300+ binds 60-120 in various combinations	

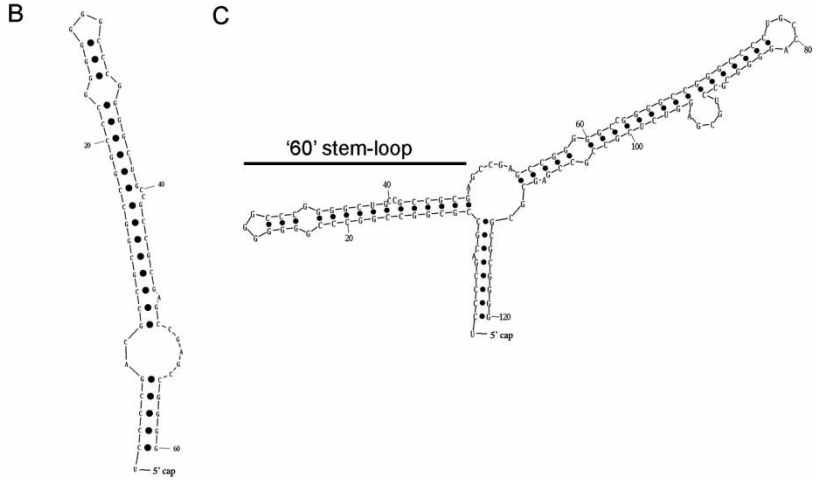

Figure 2. Modelling the effect of RNA chain extension on structural prediction. Increasing lengths of the human axin2 5′ UTRb mRNA sequence were analysed using mfold as for figure 1. The structures formed at each length were examined and compared. About 60 and 120 nucleotides form structures with particularly low free energies per nucleotide and these structures are maintained in the majority of subsequent foldings. Structures at each length are summarised (figure 2A) so as to describe the base pairing and to note the inclusion of the patterns formed by 60 or 120 nucleotides (shown in figure 2B and C, respectively).

a plateau value when the full 5′ UTR is analysed (the reading frame starts at nucleotide 244 therefore, the first 250 include the entire 5′ UTR) (figure 1B). In some cases, successive extensions to the RNA chain do not alter the 5′ structure (for example 130–270 nucleotides), while in others, the new structure for each longer molecule is a modification of, or entirely replaces, the previous structure (for example 70, 80, 100 and 110 nucleotides) (figure 2). We consider that detailed analyses of these data give insights into structures that may be formed and maintained during the extension of folding RNAs during synthesis or after translational scanning. For example, a stem-loop structure is predicted for the first 60 nucleotides (figure 2B). This structure has a low free energy per nucleotide suggesting that is particularly stable (figure 1B) and it is present in every predicted structure from 110 nucleotides onwards since nucleotides 3′ of position 60 pair with themselves rather than with the 5′ end (figure 2A). However, it is absent from the predictions for 70 or 100 nucleotides where the most 3′ nucleotides break the existing structure and base pair with the extreme 5′ end of the molecule. In each of these cases, there are particularly small decreases in minimum free energy from

the previous structures (figure 1A) and consequent increases in free energy per nucleotide (figure 1B). We suggest that the structures predicted for 70 and 100 nucleotides represent structures that, despite having slightly lower minimum free energies than the previous shorter structures, would not form since this would require unfolding of existing stable structures (in terms of free energy per nucleotide). Similarly, the double stem-loop structure formed from 120 nucleotides (figure 2C) is present in each folding from 120 through to 270 nucleotides (as well as being present when the full length mRNA is folded; not shown). However, it is not present in the patterns formed by 280, 290 and 310–350 nucleotides where, again, the new 3′ bases pair with critical elements of the "120" structure (figure 2A). This double stem-loop structure has the lowest free energy per nucleotide of the data set, suggesting stability (figure 1B). We suggest that this stable structure is unlikely to unfold readily in response to the presence of alternative base pairing partners over 160 nucleotides further along the molecule.

First, we conclude that analysis of multiple different lengths of an RNA molecule may be required to establish whether a structure has functional significance, or alternatively is transient at best, or artefactual at worst. Secondly, we conclude that algorithms for RNA structural prediction that ape the biology of an extending RNA chain are likely to be more accurate. One approach would be to couple the minimisation algorithm directly to the thermodynamics of the system as it is constructed in time and to simulate an ensemble of molecules. That is, the secondary structure of the RNA molecule would evolve randomly in time as a Markov chain, with transition probabilities to a state proportional to $e^{-\Delta G_i/T}$, where ΔG_i is the free energy of state i and T is the temperature. In this approach, states with unbound bases must be included as they represent the energy barriers that must be overcome to change states. This is repeated for a number of random steps corresponding to the time necessary for the next base to be added to the 3′ end and the process repeated. Running the simulation procedure thousands of times will then produce a distribution of final secondary structures. The necessary background for this is in place in the Vienna code, which allows computation of the partition function, however implementing a stochastic evolution algorithm that is efficient enough to provide predictions in reasonable time is not trivial. The key difficulty to address is choosing the appropriate intermediate states and defining their energy. The simulation would be greatly accelerated by choosing large effective time steps that allow direct transitions between different secondary structures. Instead the simpler approach we have taken, using mfold, could be developed. Rather than considering the full stochastic evolution of the secondary structure, a series of local optimisation procedures can be used. Given a chain of length N, the structure of the first (measuring from 5′) $N - n$ bases can be fixed before finding the optimal configuration of the remaining n nucleotides at the 3′ end. As each new nucleotide is added the procedure can be repeated. n can be chosen as a fixed number or chosen so as a fixed amount of energy is necessary. This can be regarded as an acceleration of the stochastic procedure where we move straight to the state with minimum energy that only requires moving over intermediate states with certain energy.

2.2 Modelling the effect of 5′ UTR structures on translation

Although longer 5′ UTRs tend to have lower minimum free energies (figure 1A), this does not necessarily seem to cause greater inhibition of translation, indeed it is the shortest of the three 5′ UTRs expressed from the human axin2 gene that inhibits translation most [11]. The free energies per nucleotide of specific structures within the 5′ UTR are likely to be more relevant since these may relate to their stabilities and therefore to their potentials to block the

recruitment of translation factors or scanning. An approach to assess this is to determine the free energy per nucleotide of a relatively small number of nucleotides within the whole folding pattern in a ”scanning window” moving from the 5′ to the 3′ end of the molecule. Regions of high local structure determined in this way (high local free energy per nucleotide) are likely to confer considerable inhibitory effects on translation. It should be noted that structures close to the 5′ cap or to the initiation codon, and those within the body of the 5′ UTR should be treated separately since these are thought to target factor recruitment and scanning respectively and therefore to have different effects [12,13].

A number of strategies allow accurate measurement of translational inhibition specified by particular 5′ UTRs within cells [11,14]. In these experiments, a reporter gene is expressed with either a control 5′ UTR, which lacks significant regulatory motifs, or with a test 5′ UTR. Reporter mRNA and protein expression are determined in both cases and the effect of the test 5′ UTR on translation accurately quantified by comparison to the control. These strategies could be developed to allow high-throughput determination of the influence of any biological or artificial 5′ UTR. This unlocks the attractive prospect of combining the extent of the biological influences of 5′ UTRs with predictions of their structures in order to investigate the link between structure and translational inhibition. This would be achieved by simultaneously predicting structures for hundreds of 5′ UTRs and precisely measuring their degrees of inhibition. These would then be used by a computer classification system to identify the structures that are responsible for the inhibition. The key difficulty here is to identify a low-dimension parameterisation of the structure so that the classification system is effective. This could be done by describing the structure in terms of the lengths of each base paired section and how they are connected together. To what extent a purely geometric classification like this can explain the inhibition is a very interesting question.

3. Conclusions

Accurate prediction of RNA structure within 5′ UTRs will give substantial insights into the functional regulation of many important genes in the context of normal physiology and pathology. We may have reached the limits of the current modelling techniques that are based on our understanding of RNA chemistry. Incorporation of either a more biological approach or additional biological data into this approach in the future is likely to enhance the accuracy of these predictions, and therefore their worth in the study of gene regulation.

Acknowledgements

T. A. H. is supported by the Breast Cancer Research Action Group, BCRAG (UK Charity No. 1075308). J. N. M. is supported by an Advanced Research Fellowship (GR/T02416/01) from the Engineering and Physical Sciences Research Council (EPSRC, UK).

References

[1] Pickering, B.M. and Willis, A.E., 2005, The implications of structured 5′ untranslated regions on translation and disease, *Seminars in Cell and Developmental Biology*, **16**, 39–47.
[2] Pandolfi, P.P., 2004, Aberrant mRNA translation in cancer pathogenesis: an old concept revisited comes finally of age, *Oncogene*, **23**, 3134–3137.

[3] Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F. and Liuni, S., 2001, Structural and functional features of eukaryotic mRNA untranslated regions, *Gene*, **276**, 73–81.
[4] Kozak, M., 2005, Regulation of translation via mRNA structure in prokaryotes and eukaryotes, *Gene*, **361**, 13–37.
[5] Kozak, M., 1991, An analysis of vertebrate mRNA sequences: intimations of translational control, *Journal of Cell Biology*, **115**, 887–903.
[6] Stoneley, M. and Willis, A.E., 2003, Aberrant regulation of translation initiation in tumorigenesis, *Current Molecular Medicine*, **3**, 597–603.
[7] Hofacker, I.L., 2003, Vienna RNA secondary structure server, *Nucleic Acids Research*, **31**, 3429–3431.
[8] Zuker, M., 2003, Mfold web server for nucleic acid folding and hydrization prediction, *Nucleic Acids Research*, **31**, 3406–3415.
[9] Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H., 1986, Improved free-energy parameters for predictions of RNA duplex stability. in *Proceedings of the National Academy of Sciences of the United States of America*, **83**, 9373–9377.
[10] Mitchell, S.A., Spriggs, K.A., Coldwell, M.J., Jackson, R.J. and Willis, A.E., 2003, The apaf-1 internal ribosome entry segment attains the correct structural conformation for function via interactions with PTB and unr, *Molecular Cell*, **11**, 757–771.
[11] Hughes, T.A. and Brady, H.J.M., 2005, Expression of axin2 is regulated by the alternative 5′ untranslated regions of its mRNA, *Journal of Biological Chemistry*, **280**, 8581–8588.
[12] Babendure, J.R., Babendure, J.L., Ding, J.H. and Tsien, R.Y., 2006, Control of mammalian translation by mRNA structure near caps, *RNA*, **12**, 851–861.
[13] Nakamoto, T., 2006, A unified view of the initiation of protein synthesis, *Biochemical and Biophysical Research Communications*, **341**, 675–678.
[14] Brenet, F., Dussault, N., Delfino, C., Boudouresque, F., Chinot, O., Martin, P.M. and Ouafik, L.H., 2006, Identification of secondary structure in the 5′-untranslated region of the human adrenomedullin mRNA with implications for the regulation of mRNA translation, *Oncogene*, in press.