J. theor. Biol. (1987) **127**, 63–78

Minimal Energy Foldings of Eukaryotic mRNAs Form a Separate Leader Domain

D. A. M. KONINGS,[1]† L. P. VAN DUIJN,[2] H. O. VOORMA[2] AND P. HOGEWEG[1]

[1]*Bioinformatics Group and* [2]*Department of Molecular and Cellular Biology, University of Utrecht, Padualaan 8, Utrecht, The Netherlands*

(*Received* 1 *July* 1986, *and in revised form* 3 *January* 1987)

We have investigated the minimal energy foldings of 38 mature mRNAs, including the globin family, the insulins, the growth hormones and interleukin-2, and have compared these foldings with those of fully and partly randomised sequences. The mRNAs differ from the random sequences in that they form a separate leader hairpin of 40–60 nucleotides, with the initiation codon typically located downstream of this hairpin, followed by a main fold in which a region flanking the initiation codon is basepaired with the trailer: resulting in a close proximity of the 5′ and 3′ end of the mRNA. The formation of this conformation depends not only—or primarily—on the structure of the leader, but on both the leader and trailer sequence and their interaction with the coding sequence. Thus if, as the frequent occurrence of this pattern suggests, the secondary structure of the leader regions plays a role in the initiation of translation, possibly accounting for the specificity of initiation and the different translational efficiencies of various mRNAs, we expect that these features may be influenced both by leader and trailer mutants.

1. Introduction

Messenger RNAs undergo the processes of transcription, splicing, transport and translation, in each of which several regulatory steps can operate. As a consequence it is almost a matter of course that mRNAs have evolved by exploiting all of the levels of information available, i.e, primary, secondary and tertiary structure.

The primary structure is important for coding as well as for consensus sequences in the regulation of processing and translation. In addition, many authors suggest that their experimental results may be best explained if secondary structure has a role in (pre-)mRNA 3′ cleavage and polyadenylation (Birchmeier *et al.*, 1983), splicing, and translation (e.g. Reed & Maniatis, 1986; Gheysen *et al.*, 1982; Lomedico & Andrew, 1982; Kuhne *et al.*, 1983; Vournakis & Vary, 1984; Spena *et al.*, 1985).

In the translation initiation the secondary and the tertiary conformation of the mRNA molecule may influence the efficiency of translation as well as the translation fidelity, i.e. the use of the correct initiation codon. Translational fidelity can be explained to a large extent simply by evaluating primary structures, the position of the initiation codon to the 5′ end of the mRNA and the consensus sequence surrounding the initiation codon, as proposed in the modified scanning model

† Present address: European Molecular Biology Laboratorium, Postfach 10.2209, 6900 Heidelberg, FRG.

0022-5193/87/130063+16 $03.00/0

(Kozak, 1981, 1984*a,b,c*). However, the information in the primary structure is not sufficient to predict the internal initiation of some mRNAs. This behaviour includes the bypassing of non-functional AUGs in relatively long leaders and the presence of a second initiation codon on one mRNA transcript (e.g. adeno E1b mRNA, Bos *et al.* 1981; CPMV, Van Wezenbeek *et al.*, 1983; and sendai P/C mRNA, Gupta & Kingsbury, 1985). Also the primary structure fails to account for the available data on translational efficiency. In contrast, the efficiency of translation is often related to the free energy of the secondary structure of the leader part (Isertant & Fiers, 1980; Pinck *et al.*, 1981; Gehrke *et al.*, 1983; Pelletier *et al.*, 1985), although no linear relation could be demonstrated (Johansen *et al.*, 1984; Godefroy *et al.*, 1985; Pelletier *et al.*, 1985). Only in the case of the α- and β-globin (Vournakis *et al.*, 1984) and of the four alfalfa mRNAs (Godefroy *et al.*, 1985), are secondary structure data concerning the leader part of the molecule available in addition to translational efficiency data. These data do not support a straightforward relation between translational efficiency and the local free energy of the leader structure. Nevertheless, the influence of the secondary structure in the translational initiation process can not be ruled out, because, amongst others, this influence depends not only on the (local) features of the leader but on those of the entire mRNA.

In order to understand the constraints which act on the secondary structure of mRNAs, the sequences of mRNA molecules can be investigated by minimal energy foldings. Previously we showed for four homologous E1a pre-mRNAs of adenovirus that a separate local hairpin at the 5′ end, with the AUG codon in an accessible position, and a backfold of the 3′ end of the pre-mRNA to the region near the AUG codon are important features (Konings *et al.*, 1987).

In the present paper we undertake a survey of minimal energy foldings of mature eukaryotic mRNAs (Table 1) to find out whether there is any consistent pattern in the secondary structures. These mRNAs, which cover a set of 23 different eukaryotic mRNAs, were arbitrarily selected from those of less than 1000 nucleotides. They include for example the family of the globins, the insulins, the growth hormones, and interleukin-2. We investigated in particular the leader region of the generated foldings. A class of related leader conformations is found in the minimal energy foldings of the studied mRNAs. We discuss these classes in view of the determination and efficiency of the translational initiation process.

2. Materials and Methods

2.1. INTRODUCTION

Secondary structure predictions of single RNA sequences are not very reliable. Therefore much work on secondary structure modelling looks for consensus secondary structures in a set of aligned sequences. This can be a suitable method for functional RNAs with a large degree of homology and functional similarity. It is not suitable for studying mRNAs because of the absence of homology and obvious functional equivalence. Moreover, even in case of homologous sequences similar secondary structures can be formed by non-homologous stretches (Konings *et al.*,

TABLE 1

List of the mRNA types of which minimal energy folding of the entire mRNA sequence were studied. The numbers assigned to each mRNA are also used in Fig. 1 to identify the different mRNA types. The bibliographic data of each mRNA sequence are abbreviated: first author, year, journal title, volume and first page

1. casein-γ	—rat	Hobbs (1982) NAR 10, 8079.
		Yu-Lee (1983) JBC 258, 10794.
2. cholecystokinin	—rat	Deschenes (1985) JBC 260, 1283.
3. E1a	—Ad5	Van Ormondt (1980) Gene 11, 299.
		Bos (1981) Cell 27, 121.
	—Ad7	Dijkema (1982) Gene 18, 143.
	—Ad12	Perricaudet (1980) Nature 288, 174.
		Sugisaki (1980) Cell 20, 777.
	—SA7P	Dekker (1984) JGV 65, 1699.
4. fatty acid binding protein	—rat	Gordon (1983) JBC 258, 3356.
5. gastrin	—human	Wiborg (1984) PNAS 81, 1067.
	—pig	Yoo (1982) PNAS 79, 1049.
6. globin-α	—human-1	Michelson (1980) Cell 22, 371.
	—human 2	Michelson (1980) Cell 22, 371.
	—rabbit	Baralle (1977) Nature 267, 279.
		Heindell (1978) Cell 15, 43.
7. globin-β	—chicken	Dolan (1983) JBC 258, 3983.
		Roninson (1982) Cell 28, 515.
	—human	Lawn (1980) Cell 21, 647.
	—rabbit	Baralle (1977) Cell 10, 549.
8. globin-γ	—human	Slighton (1980) Cell 21, 630.
9. glycoprotein-a	—human	Fiddes (1981) JMAG 1, 3.
10. growth hormone	—bovine	Woychik (1982) NAR 10, 7197.
	—human	Denoto (1981) NAR 9, 3719.
11. growth releasing hormone	—human	Mayo (1985) PNAS 82, 63.
12. histone-2a	—human	D'Andrea (1981) NAR 9, 3119.
13. insulin	—chicken	Perler (1980) Cell 20, 555.
	—human	Bell (1980) Nature 284, 26.
	—rat-1	Cordell (1979) Cell 18, 533.
	—rat-2	Cordell (1979) Cell 18, 533.
14. interleukin-2	—human	Taniguchi (1983) Nature 302, 305.
		Devos (1983) NAR 11, 4307.
15. apolipoprotein-II	—chicken	Van het Schip (1983) NAR 11, 2529.
16. lysozyme	—chicken	Jung (1980) PNAS 77, 5759.
		Grez (1981) Cell 25, 743.
17. metallothionein-I	—mouse	Glanville (1981) Nature 292, 267.
	—rat	Andersen (1983) DNA-2, 15.
18. metallothionein-II	—human	Karin (1982) Nature 299, 797.
		Karin (1982) NAR 10, 3165.
19. parathyroid hormone	—bovine	Kronenberg (1979) PNAS 76, 4983.
		Weaver (1982) MCE 28, 417.
	—human	Hendy (1981) PNAS 78, 7365.
		Vasicek (1983) PNAS 80, 2127.
20. somatostatin-14	—rat	Funckes (1983) JBC 258, 8781.
		Montminy (1984) PNAS 81, 3338.
21. steroid hormone receptor	—rat	Parker (1983) JBC 258, 12.
		Hurst (1983) EMBO 2, 770.
22. thyrotropin-α	—mouse	Chin (1983) PNAS 78, 5329.
23. vasopressin-II	—bovine	Land (1982) Natuer 295, 299.
	—rat	Schmale (1983) EMBO 2, 763.

1987). Thus, only independent secondary structure calculations can be used to study potential similarities in the secondary structures of mRNAs. Despite the limited reliability of such calculations, they can be used to detect common secondary structure features in sets of related sequences because:

(1) If a set of sequences possesses common secondary structure features these are very likely to show up in an appreciable fraction of the calculated secondary structures (as is shown by studies on small RNAs, the secondary structure of which can be investigated by a number of different methods).

(2) If a set of RNAs possess common features in their calculated secondary structures which are not observed in calculated secondary structures of sequences not belonging to the set (e.g. random sequences) these features are likely to reflect biologically relevant properties of the particular RNAs.

An additional check on the reliability of features in secondary structure found by such an approach is to see whether they are insensitive to small changes in the secondary structure calculations (e.g. changes in energy value, changes in algorithm): we know from small RNAs that conserved features are relative robust to such changes whereas features found in random sequences or rare mutants are very sensitive to such changes (Konings & Mattaj, 1987).

2.2. MATERIAL

Our principal aim in this survey was to fold complete mRNA sequences. The selection of the sequences was firstly based on their relatively short length, i.e. shorter than 1000 nucleotides (Table 1), because of computer limitations. Secondly, our analysis required full length mRNA sequences. Therefore we preferred those sequences which had been verified on the genome level and excluded those mRNAs for which the presence of a non-full-length cDNA copy was expected. Some of the latter class indeed showed incomplete patterns in the generated secondary structure, in particular in the 5′ and 3′ regions. However, we cannot rule out the possibility that some of the included cDNA sequences are incomplete. One has also to be aware that sequence errors could in certain cases influence the minimal energy folding in a crucial way (compare Hogeweg & Konings, 1985). Thirdly, we included sets of homologous mRNAs in order to get an insight into specific patterns for particular mRNA species, and to evaluate differences between homologous mRNAs. In total we analysed 38 mRNAs representing 23 different mRNA species (see Table 1).

For comparison with the patterns in the secondary structures revealed by the mRNA sequences we analysed in addition 40 random or partly randomised mRNA sequences. The random base compositions of these can be divided into the following classes:

(a) A specific base content which is uniform over the whole sequence. The different sequences of this class were matched to distinct base content patterns of the mRNAs.

(b) A dinucleotide preference, as tabulated by Nussinov (1981), between two out of three sequential positions (by analogy with the three coding positions).

(c) A random differentiation into three parts: the leader, coding and trailer regions of a specific mRNA sequence, e.g. the rabbit α-globin. The "coding" part in these were either generated with a preference over the three coding positions or without a preference.

(d) A randomisation of either the leader or trailer part of a particular mRNA sequence (e.g. the α-globin sequence of the rabbit and the 1Ea of the adenovirus), joined to the remaining mRNA functional parts. The base preference of the mutated parts was chosen to be either mRNA specific or to be of equal probability (see below 4.2).

Most random sequences were about 550 nucleotides in length.

2.3. METHOD

Minimal energy foldings were generated by the ENFOLD algorithm. ENFOLD is a generalisation of the maximum-match algorithm of Nussinov & Jacobson (1980), in which the non-monotonicity occurring during minimal energy folding is resolved via a backtracking mechanism: destabilising loops are entered provisionally if no stabilising bonds are available and are maintained if they lead to stabilising helices; otherwise they are ignored in later steps of the algorithm; destabilising bonds on unstable structures are not included (Hogeweg & Hesper, 1984). The free energy values for base-pairs tabulated by Zucker & Sankoff (1984) were used. We noted that by using slightly different energy values the conclusions drawn in this paper remain valid.

The algorithm generates the minimal energy folding of any selected subsequence as well as that of the entire sequence. To look for the presence of competitive secondary structures we analysed for each mRNA several sub-structures of the entire sequence. These included, amongst others, the sequential foldings (starting at the 5′ end and stepwise adding 50 nucleotides) of the mRNAs. Substructures may be important since mRNAs interact with proteins and probably undergo local unwinding of secondary structure features during processing.

The minimal energy foldings were represented according to the conventions proposed by Hogeweg & Hesper (1984). In order to minimise unwarranted suggestions about spatial relations, and in order to facilitate the comparison of foldings of sequences and/or subsequences, the linear structure of the RNA molecule is preserved in the representation. Base pairings are shown by horizontal connecting lines. Hairpin loops appear as "tops", while loops and bulges are represented as intermediate "plateaus". Helices show up as "hillsides" and branching regions are indicated by "valleys" (see Fig. 1).

4. Results

4.1. PATTERNS IN SECONDARY STRUCTURE OF mRNAs

The mRNAs show a great deal of variation in minimal energy folding (Fig. 1). Nevertheless some patterns emerge. A close proximity of the 5′ and 3′ termini in

most mRNAs is caused by a folding-back of the 3′ part to a region near to the 5′ end of the mRNA. Most often this backfold is formed by a large main stem on top of which one or only a few (2–3) large subfolds are located. The number of these substructures is partly related to the length of the mRNA concerned.

It is interesting to note that in some mRNAs the main stem does not include the leader. Thus these mRNAs form a separate leader hairpin, including 40–60 nucleotides (see Table 1 and Figs 1, 2). The main stem of the backfold in nearly all of these mRNAs is formed by at least 15 or more base pairs. This group of mRNAs includes 45% of all studied mRNAs, covering 57% of the different mRNA species.

The initiation codon is typically located downstream of the leader hairpin loop, most often in the branching region between this hairpin and the main fold of the mRNA. In the case of relatively long leaders, 80–100 nucleotides, the initiation AUG sits in the stem of the main fold of the mRNA, whether or not in a single-stranded region. In the separate 5′ terminal hairpin of the rabbit α-globin the initiation codon is present in the hairpin loop. This leader sequence, 39 nucleotides, is shorter than the average length.

The rest of the mRNAs do not show a uniform conformation of their leader region.

In order to evaluate the relevance of the 5′ terminal hairpin and the backfold of the mRNA, we analysed also minimal energy foldings of 40 random sequences. Only by comparison with non-biological sequences can we get an idea of what is to be expected in minimal energy foldings of mRNA sequences. The differentiated base contents of these random sequences reflect content patterns of the studied mRNAs (described in "materials and methods").

Figure 3 shows two "typical" minimal energy secondary structures of the 40 analysed random sequences. Comparison of the generated structures of the random sequences with the ones of the mRNAs shows that:

(1) The mere presence of a main stem is not exclusive to mRNAs.

(2) The separation of a 5′ terminal hairpin of about 50 nucleotides in close association with a folding-back of the 3′ terminus is restricted to mRNAs, a similar structural combination is generated in only 2 of the 40 analysed random sequences (i.e. in 5% of cases compared with 45% for mRNAs).

(3) The potential to form a hairpin within a stretch of about 50 nucleotides, however, is not a unique feature of leader sequences. Many randomly selected stretches of mRNA and random sequences do form a hairpin-like structure by energy minimisation. The energy content of hairpin-like structures formed by some of the arbitrary stretches ranges up to the average content (about $-11{\cdot}0$ kcal) of mRNA-leader sequences. Moreover there exists no correlation between the energy of the leader hairpin and its separation in the minimal energy folding.

(4) In the sequential subfoldings of the mRNAs the leader hairpin shows up in an appreciable number of the substructures, in contrast to a more instable conformation of a 5′ structure in the ones of random sequences. Furthermore, the subfoldings of the random sequences reveal a more variable pattern of local interactions throughout the whole sequence.

(5) There exists a clear relationship between the minimal energy and the base content for both the random sequences and the mRNA sequences. However, the

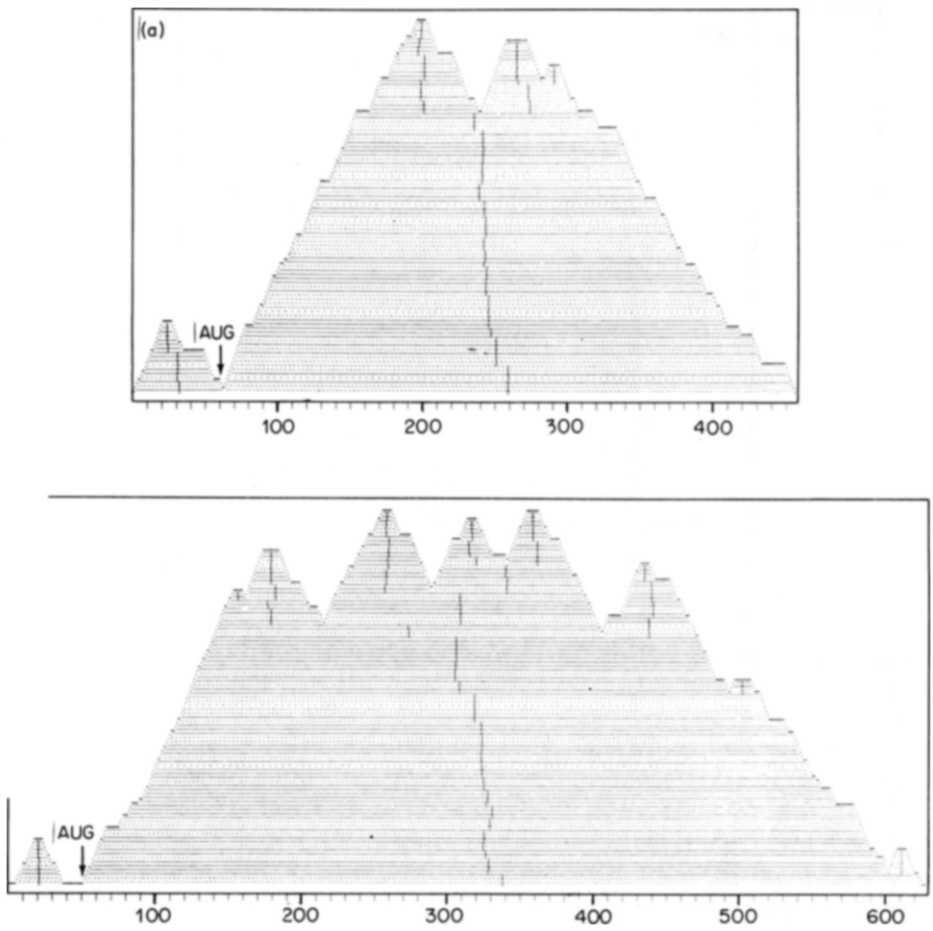
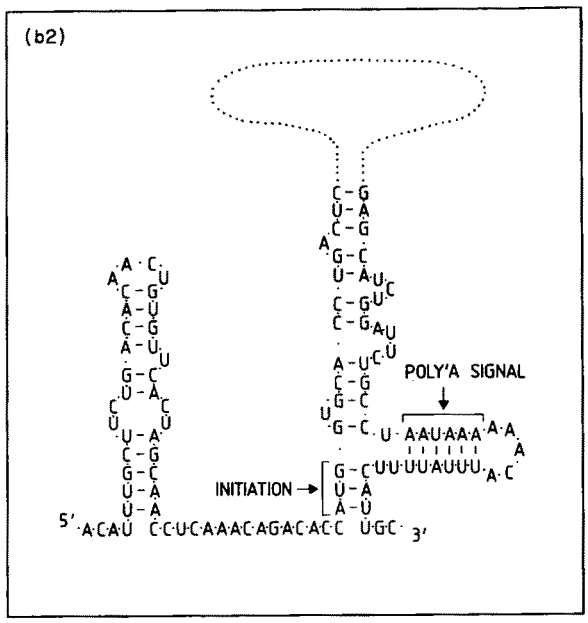


FIG. 1. Minimal energy foldings of four different mRNA types are shown as examples. Three of these (a, b, c) reveal a free leader domain, with either the initiation codon in the branching region (a, b) or in the 5′ flank of the main fold of the remaining mRNA structure (c). The last mRNA shown (d) shows a large backfold of the 3′ end to the 5′ end of the mRNA, without a separate leader hairpin. (a) 9S mRNA of adenovirus 12, calculated free energy: −136·5 kcal. (b) β-globine mRNA of human, calculated free energy: −265·5 kcal. (c) somatostatin-14 mRNA of rat, calculated free energy: −253·6 kcal. (d) steroid hormone receptor mRNA of rat, calculated free energy: −149·6 kcal. The mRNA folding of human β-globin (see b1) is also partly shown in a conventional way (b2) in order to make our representation clear and to emphasise the structural coexistence of the leader domain and the interaction of the 5′ and 3′ ends of the mRNA. The minimal energy foldings were generated by the ENFOLD-algorithm of Hogeweg & Hesper (1984), using the energy values as tabulated by Zuker & Sankoff (1984) (see "Material and methods"). In order to facilitate the comparison of foldings of sequences and/or subsequences, the linear structure of the RNA molecule is preserved in the representation. The nucleotides of the sequence are shown as symbols on the curve of the plot. Numbers below the bottom horizontal line indicate the position of the nucleotides of the mRNA with respect to the 5′ cap site. Vertical lines within the pattern show points midway between the nucleotides paired by the horizontal lines. Each base pair is shown by a horizontal line (——: G-C, - - -: A-U and - · - · -: G-U bonds). Hairpin loops appear as flat tops, interior loops and bulges as intermediate plateaus, helices as sloping lines and branching regions are shown as valleys (Hogeweg & Hesper, 1984). The arrows indicate the location of the initiation codons.



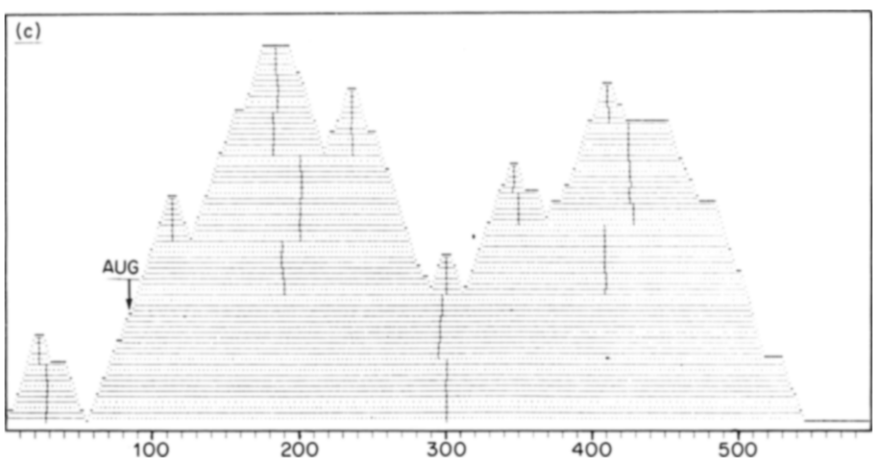


FIG. 1. (Continued).

mRNAs are slightly more stable, than their random counterparts. A similarly small difference holds true for the frequencies of base pairings. Thus, given the global base content of a sequence, the mRNAs seem to be only slightly stabilised by site-specific interactions.

Since base content strongly influences overall stability, the selection of base content may be important for regulating secondary structure (e.g. survival and

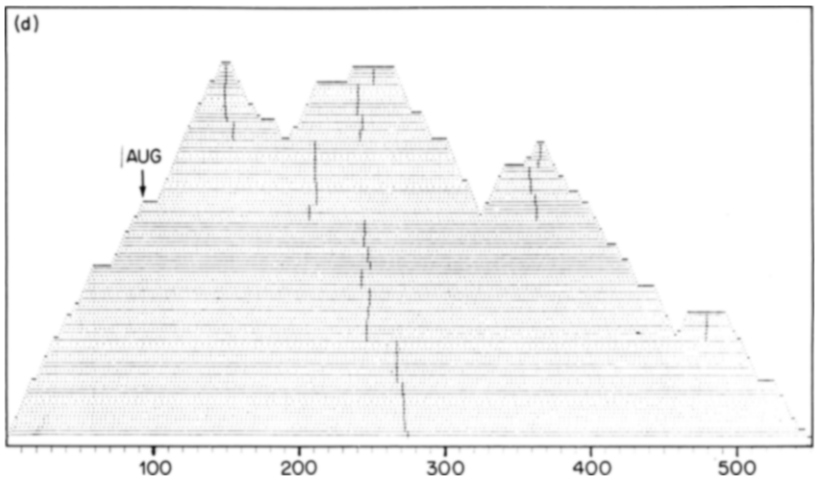


FIG. 1. (continued).

domain formation) of mRNAs, and may be accomplished without disrupting the coding capacity, if tuned by the third position of the reading frame (Konings & Hogeweg, in prep.). This comparison of the conformations and the energy patterns of the sequences studied shows that the overall differences between mRNA and random sequences are subtle. In contrast the frequent occurrence of the 5′ terminal hairpin is restricted to mRNAs.

Apart from the 45% of mRNAs which form the separate leader hairpin described by energy minimisation, another 30% (see Fig. 2) can form the same pattern, but with an energy content which is slightly higher than that of their minimal energy foldings. For the second group of mRNAs this alternative conformation is reasonable to consider because, in most cases, the overall conformation is identical to that of the minimal energy folding, or else the leader hairpin is exceptionally stable (around −25·0 kcal), and therefore may be favoured in a sequential folding process.

In 7 out of the 9 sets of homologous mRNAs, at least one of the members does form the leader hairpin spontaneously (see Table 1 and Fig. 2). A comparison of possible foldings for the different members of each set shows that the present consensus secondary structure for the leader is the one which fits into the described pattern. The AUG codon takes up an analogous position in these mRNA-specific structures. For example, in all insulins the AUG can be located just downstream of the leader hairpin, and in the growth hormone and β-globin mRNAs it can be in the 5′ flank of the following main helix. In the β-globin mRNA the minimal energy folding of the separated leader has a single stranded region between the hairpin and the AUG-codon (Fig. 1).

Within the set of adenovirus E1a mRNAs the Ad5 mRNA forms a slightly different conformation. In contrast to the other three members, no separate leader hairpin

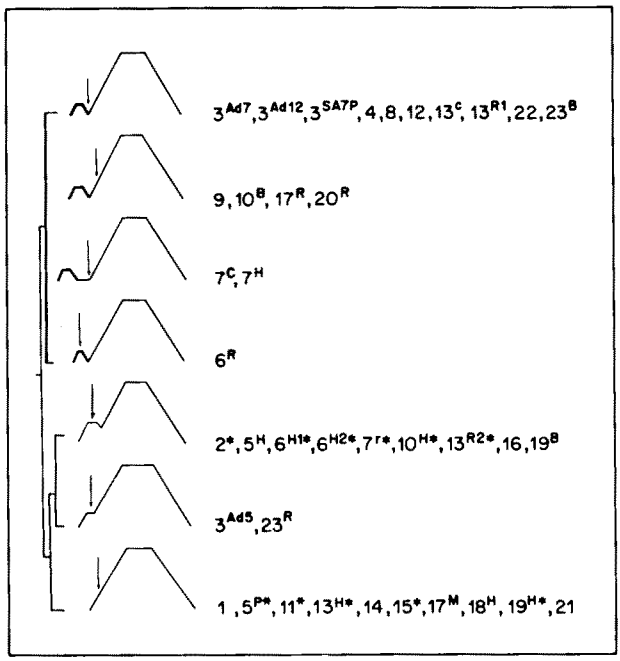


FIG. 2. Scheme of the different leader conformations generated by minimal energy folding of the entire mRNA studied (see Table 1). The representation of the leader conformations is a simplified version of the one shown in Fig. 1 for four of the mRNA foldings as described by Hogeweg & Hesper (1984). The location of the initiation codon is indicated by an arrow. To identify the mRNAs within this classification we have used the numbers as assigned in Table 1. Letters following the numbers refer to the origin of the mRNA when RNA from more than one species was analysed. The top cluster, with the mRNAs which reveal a more or less separate leader hairpin of 40–60 nucleotides in the generated minimal energy folding, includes 45% of all mRNAs studied, covering 57% of the different mRNA types. Those mRNAs in the lower cluster which can form the separate hairpin in a reasonable alternative secondary structure (see section 4.1) are indicated by an asterisk.

is present in the minimal energy folding of the Ad5 mRNA(s). However, the AUG codon in this mRNA is located in an accessible bulge downstream of a local leader hairpin loop. It is interesting that for this particular mRNA an extra initiation codon exists downstream (Spindler & Berk, 1984). This second AUG, like many other functional ones (see above), is located in an accessible position: a branching region flanking a main stem of the mRNA.

4.2. RANDOMISED mRNA SEQUENCES AND SELECTIONAL CONSTRAINTS

To pinpoint the selectional constraints on the secondary structures of the mRNAs—in particular on the observed leader conformation—we studied some randomised sequences of the adenovirus-12 9S-E1a mRNA and the rabbit α-globin. These sequences were generated by forming hybrids of random and biologically functional parts of the mRNAs: a base content-related random leader was combined with the remaining biological part of the mRNA, and the biological leader and

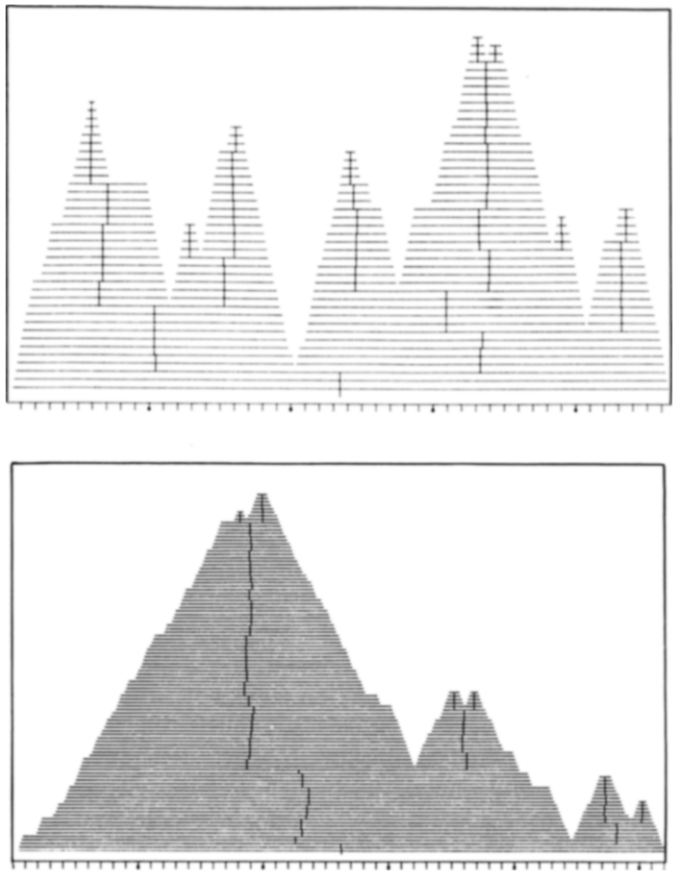


FIG. 3. Two minimal energy foldings of the 40 analysed random sequences (see 2.2. Material). The secondary structures shown are typical examples of the unspecific conformations formed by the random sequences. Unlike the mRNA foldings, these conformations show no significant common structure in their 5′ leading regions. The foldings are represented as described in figure 1. However, in this figure the nucleotides are not represented as symbols on the curve, neither are the three base-pairings (G-C, A-U and, G-U) represented as different horizontal lines.

coding part was combined with either a base content-related random trailer or with a random trailer with equal base probability.

The results of this study are shown in Table 2. Four of the random-leader hybrids show a loss of the typical leader hairpin (one of these is shown in Fig. 4), while the remaining two leader-hybrids reveal a separate 5′ terminal hairpin. However, in one of them the AUG codon is located in a position not observed elsewhere, namely in the 3′ flank of this hairpin. Three of the four random-trailer hybrids loose their 5′ terminal hairpin, whereas the fourth hybrid (see Fig. 4) keeps it by forming an alternative interaction between the randomised trailer and the 5′ part of the coding region, thereby maintaining the rest of the structure.

The analysis of these hybrids demonstrates that a selection for base-pair-specific interactions has to occur in order to obtain the different classes of leader hairpins.

TABLE 2

The conformation of the "leader" regions of the 6 random-leader and 4 random-trailer hybrids of the rabbit α-globin and the adenovirus-12 95-E1a mRNA sequences as generated by minimal energy folding of the entire sequence. These 5′ terminal structures are described in more detail in section 4.2

	5′ Terminal separate hairpin structure	No 5′ terminal separate hairpin structure
Leader hybrids		
9S mRNA adenovirus	1	2
α-globin rabbit	1	2
Trailer hybrids		
9S mRNA adenovirus	0	2
α-globin rabbit	1	1

The potential to form a hairpin by a limited 5′ stretch is always required, while the energy content of this hairpin does not seem to be crucial (see section 4.1). The hybrids show that this potential alone is not sufficient: a tuning of the downstream mRNA sequence is also essential. It seems that this tuning can be accomplished by a combination of a more or less site-specific backfold downstream of the leader structure and an avoidance of stretches which are strongly complementary to the leader sequence. This tuning can largely be done by the trailer. In addition specific interactions around the splice sites limit the stretches which should avoid such interactions with the leader.

5. Discussion and Conclusions

This paper describes an approach to uncover functional patterns in the secondary structure of mRNAs by comparing the minimal energy foldings of a large set of mRNAs, including single mRNAs and subsets of homologous mRNAs. Our analysis reveals a pattern in the secondary structure associated with the initiation codon, namely a hairpin structure at the 5′ end which is separated from a main fold of the remaining mRNA structure. The size of this hairpin is highly conserved (40–60 nucleotides), much more so than the location of the AUG codon. Typically the initiation codon occurs at the branching region just downstream of this leader hairpin, but in mRNA species with a long leader the AUG is located on the 5′ flank of the main fold while in mRNAs with a short leader the AUG may be found in the hairpin loop. In many cellular mRNAs potential hairpin structures which cover part of the leader have been pointed out (e.g. McReynolds *et al.*, 1978; Lomedico *et al.*, 1979; Cooke *et al.*, 1980; Kaempfer *et al.*, 1981). However, these data were thought not to be a general feature (McReynolds *et al.*, 1978; Kozak, 1980). The

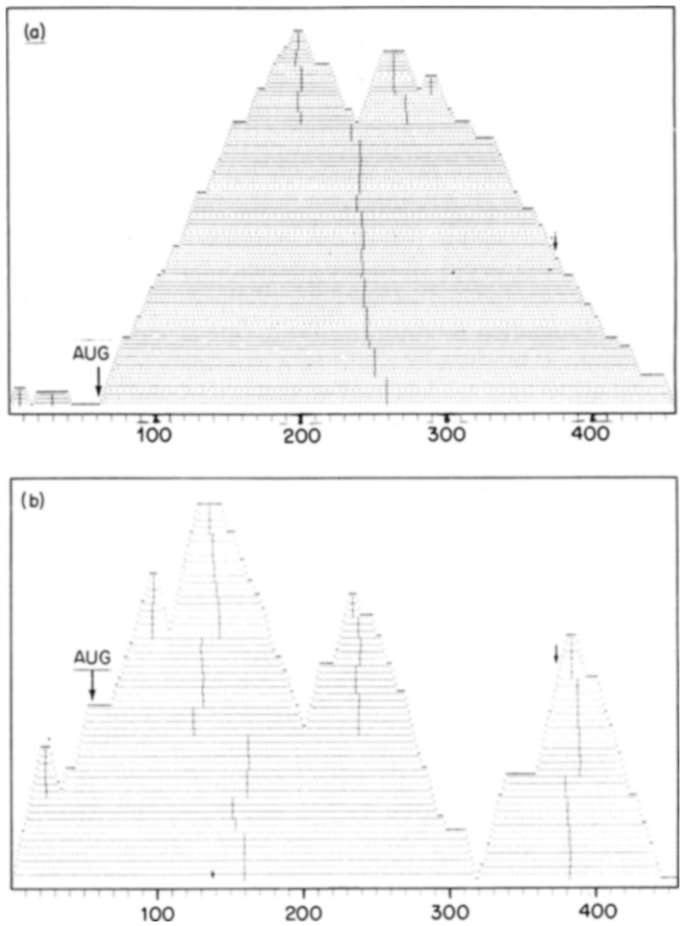


FIG. 4. Minimal energy folding of two of the adenovirus-12 9S-E1a hybrid sequences analysed (see section 3 and 4.2 for the generation of these hybrids). The leader hybrid shows a loss of the separate, single, leader hairpin, but keeps the backfold of the 3′ end to the region downstream of the randomised leader sequence (a). The trailer hybrid shows a loss of both the 5′ terminal hairpin and the backfold of the 3′ end (b). (a) leader hybrid of the 9S-E1a mRNA of adenovirus-12, calculated free energy: −127·6 kcal. (c) trailer hybrid of the 9S-E1a mRNA of adenovirus-12, calculated free energy: −147·9 kcal. The sites of the initiation codon, the termination codon are indicated by an arrow (large and small, resp.).

main fold of the mRNA following the leader hairpin results in a close proximity of the 5′ and 3′ ends in the secondary structure. This pattern is consistent with the observation that certain proteins are associated with both the cap and the poly-A tail of mRNAs (Kumar, pers. com.).

Translational initiation is a likely selectional force for the maintenance of a 5′ leader structure and the accessibility of the initiation codon. The folding pattern we find could play a role in different steps of the translational initiation process. The conserved 5′ hairpin structure seems to be present as an accessible domain which could act as a recognition signal for the interaction with initiation factors,

or with the 40S/80S ribosomal complex (compare Legon *et al.*, 1979; Kozak, 1979, 1980; Morgan & Shatkin, 1980). By analogy, similar substructures (primary and secondary) may function in the flagging of existing internal initiation areas (Perez-Bercoff, 1982). In addition, the characteristic position of the initiation codon in this hairpin could have implications for the determination of the start codon. The location of the AUG just downstream of the 5′ leader hairpin suggests an alternative to the scanning model, i.e. a "jump" from the cap to the initiation AUG. In this case an unwinding of the leader structure would not be required. Consequently, a different conformation of the AUG could affect the translational efficiency, for example the functional AUG could be located in the flank of the downstream helix in a buried position (compare Lomedico & Andrew, 1982; Spena *et al.*, 1985). An extended model of this "jumping" hypothesis could be a candidate to clarify the relationship between extraordinarily long leaders and high translational efficiency (Johansen *et al.*, 1984; Godefroy *et al.*, 1985; Kitamura *et al.*, 1981; Pelletier *et al.*, 1985). Thus it seems that a knowledge of both primary and secondary structure features of mRNA molecules are needed to understand translational phenomena.

Experimental tests of these hypotheses are currently in progress. The first results show that translational fidelity of a couple of AUG codons in the 5′ part of a mRNA transcript, which are indistinguishable in their primary contexts, is related to the local conformations of the different start codons and not to their distances to the 5′ terminal (van Duijn *et al.*, 1987, in prep.).

Since many biological processes illustrate that nature does not require an "all or nothing" mechanism to obtain certain effects, e.g. variable cap-dependence for translational initiation (Van Steeg *et al.*, 1981; Gehrke *et al.*, 1983; Ray *et al.*, 1983), some mRNAs may benefit from the leader hairpin while others manage to be translated efficiently enough without such a structure. The latter type may have compensating features (compare Both *et al.*, 1976; Muthukrishnan *et al.*, 1976; Goelet *et al.*, 1982; Yamaguchi *et al.*, 1982; Gehrke *et al.*, 1983). This could be the case for those mRNAs with a relatively short leader which is free of any secondary structure interaction, e.g. many viruses and interleukin-2. Furthermore, as suggested by Bahramian (1980) for the prokaryotic system a strong consensus primary structure (Shine-Dalgarno sequence) can compensate other features such as a specific secondary structure and vice versa.

We conclude that the secondary structure is one of the selectional constraints which act upon mRNAs, and affects e.g. translational initiation. This constraint can largely be met by the non-coding parts, including both the leader and trailer and the third codon position of the mRNA, and therefore does not interfere with the coding capacities (Konings *et al.*, 1987; Konings & Hogeweg, in prep.).

The model of the mRNA secondary structure presented here predicts the following counterintuitive, but still experimentally observed, properties of mRNAs.

(1) Translation efficiency is not only a function of the leader structure: the leader and the trailer regions should be tuned to the coding region as well as to each other in order to fulfill secondary structure constraints. As a consequence, mutations in either of these regions may affect the local conformation of the other in the secondary structure of the entire mRNA, and thus translational efficiency. There is some

experimental evidence from trailer mutants in support of this view (Spena *et al.*, 1985).

(2) Within a set of homologous mRNAs the leader and the 3′ part of the trailer are conserved to a degree similar to the coding part because of their close interaction in the secondary structure (compare Miyata & Yasunaga, 1981; Gunning *et al.*, 1984). In contrast, although non-homologous mRNAs are subject to the same secondary structure constraints and/or translation apparatus they do not show similarities in either of these domains.

We would like to thank Dr L. Philipson and Dr C. Zwieb for helpful discussions and suggestions concerning the paper, and G. Tebb for suggesting language corrections.

REFERENCES

BAHRAMIAN, M. B. (1980). *J. theor. Biol.* **84,** 103.
BIRCHMEIER, C., FOLK, W. & BIRNSTIEL, M. L. (1983). *Cell* **35,** 433.
BOS, J. L., POLDER, L. J., BERNARDS, R., SCHRIER, P. I., VAN DER ELSEN, P. J., VAN DER EB, A. J. & VAN ORMONDT, H. (1981). *Cell* **27,** 121.
BOTH, G. W., FURUICHI, Y., MUTHUKRISHNAN, S. & SHATKIN, A. J. (1976). *J. mol. Biol.* **104,** 637.
COOKE, N. E., COIT, D., WEINER, R. I., BAXTER, J. D. & MARTIAL, J. A. (1980). *J. biol. Chem.* **255,** 6502.
GEHRKE, L., AURON, P. E., QUIGLEY, G. J., RICH, A. & SONENBERG, N. (1983). *Biochemistry* **22,** 5157.
GHEYSEN, D., ISERENTANT, C., DEROM, C. & FIERS, W. (1982). *Gene* **17,** 55.
GODEFROY-COLBURN, T., THIVENT, C. & PINCK, L. (1985*a*). *Eur. J. Biochem.* **147,** 541.
GODEFROY-COLBURN, T., RAVELONANDRO, M. & PINCK, L. (1985*b*). *Eur. J. Biochem.* **147,** 549.
GOELET, P., LOMONOSSOFF, G., BUTLER, P., AKAM, M., GAIT, M. & KARN, J. (1982). *Proc. natn. Acad. Sci. U.S.A.* **79,** 5818.
GUNNING, P., MOHUN, T., NG, S-Y., PONTE, P. & KEDES, L. (1984). *J. mol. Evol.* **20,** 202.
GUPTA, K. C. & KINGSBURY, D. W. (1985). *Biochem. biophys. Res. Commun.* **131,** 91.
HOGEWEG, P. & HESPER, B. (1984). *Nucleic Acids Res.* **12,** 67.
HOGEWEG, P. & KONINGS, D. A. M. (1985). *J. mol. Evol.* **21,** 323.
ISERTANT, D. & FIERS, W. (1980). *Gene* **9,** 1.
JOHANSEN, H., SCHUMPERLI, D. & ROSENBERG, M. (1984). *Proc. natn. Acad. Sci. U.S.A.* **81,** 7698.
KAEMPFER, R., EMMELO, V. J. & FIERS, W. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 1542.
KITAMURA, N., SEMLER, B. L., ROTHBERG, P. G., LARSEN, G. R., ADLER, C. J., DORNER, A. J., EMINI, E. A., HANECAK, R., LEE, J. J., VAN DER WERF, S., ANDERSON, C. W. & WIMMEN, E. (1981). *Nature* **291,** 547.
KONINGS, D. A. M. & MATTAJ, I. W. (1987). *Exp. Cell Res.* (in press.)
KONINGS, D. A. M., HOGEWEG, P. & HESPER, B. (1987). *Mol. Biol. Evol.* (in press).
KOZAK, M. (1979). *Nature* **280,** 82.
KOZAK, M. (1980). *Cell* **19,** 79.
KOZAK, M. (1981). *Nucleic Acids Res.* **9,** 5233.
KOZAK, M. (1984*a*). *Nucleic Acids Res.* **12,** 857.
KOZAK, M. (1984*b*). *Nucleic Acids Res.* **12,** 3873.
KOZAK, M. (1984*c*). *Nature* **308,** 241.
KUHNE, T., WIERINGA, B., REISER, J. & WEISMAN, C. (1983). *EMBO J.* **2,** 727.
LEGON, S. (1979). *J. mol. Biol.* **134,** 219.
LOMEDICO, P. T., ROSENTHAL, N., EFSTRATIADIS, A., GILBERT, W., KOLODNER, R. & TIZARD, R. (1979). *Cell* **18,** 545.
LOMEDICO, P. T. & MCANDREW, S. J. (1982). *Nature* **299,** 221.
MCREYNOLDS, L., O'MALLEY, B. W., NISBET, A. D., FOTHERGILL, J. E., GIVOL, D., FIELDS, S., ROBERTSON, M. & BROWNLEE, G. G. (1978). *Nature* **273,** 723.
MIYATA, T. & YASUNAGA, T. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 450.
MORGAN, M. A. & SHATKIN, A. J. (1980). *Biochem.* **19,** 5960.
MUTHUKRISHNAN, S., FURUICHI, Y., BOTH, G. W. & SHATKIN, A. J. (1976). *Progr. Nucl. Acids Res., Mol. Biol.* **19,** 473.

NUSSINOV, R. & JACOBSON, A. B. (1980). *Proc. natn. Acad. Sci. U.S.A.* **77,** 6309.
NUSSINOV, R. (1981). *J. mol. Biol.* **149,** 125.
PELLETIER, J. & SONENBERG, N. (1985). *Cell* **40,** 515.
PEREZ-BERCOFF, R. (1982). In: *Protein Biosynthesis of Eukaryotes.* (Perez-Bercoff, R. ed.). New York: Plenum Press.
PINCK, M., FRITSCH, C., RAVELONANDRO, M., THIVENT, C. & PINCK, L. (1981). *Nucleic Acids Res.* **9,** 1087.
RAY, B. K., BRENDLER, T. G., ADYA, S., DANIELS-MCQUEEN, S., KELVIN MILLER, J., HERSHEY, J. W. B., GRIFO, J. A., MERRICK, W. C. & THACH, R. T. (1983). *Proc. natn. Acad. Sci. U.S.A.* **80,** 663.
REED, R. & MANIATIS, T. (1986). *Cell* **46,** 681.
SPENA, A., KRAUSE, E. & DOBBERSTEIN, B. (1985). *EMBO J.* **4,** 2153.
SPINDLER, K. R. & BERK, A. J. (1984). *J. Virol.* **51,** 884.
VAN STEEG, H., VAN GRINSVEN, M. VAN MANSFELD, F., VOORMA, H. O. & BENNE, R. (1981). *FEBS Lett.* **129,** 62.
VAN WEZENBEEK, P., VERVER, J., HARMESEN, J., VOS, P. & VAN KAMMEN, A. (1983). *EMBO J.* **2,** 941.
VOURNAKIS, J. N. & VARY, C. P. H. (1984). *Devel. Genetics* **4,** 313.
YAMAGUCHI, K., HIDAKA, S. & MIURA, K-L. (1982). *Proc. natn. Acad. Sci. U.S.A.* **79,** 1012.
ZUCKER, M. & SANKOFF, D. (1984). *Bull. Math. Biol.* **46,** 591.