Evolution of the Primary and Secondary Structures of the E1a mRNAs of the Adenovirus

D. A. M. Konings, P. Hogeweg, and B. Hesper
Bioinformatica

In this paper we investigate and compare (evolutionary) patterns in the primary and secondary structure of four homologous E1a mRNAs of the adenovirus. Our main results are as follows: (1) The similarity of the coding regions of the mRNA sequences reflects both similarity in function (i.e., oncogenicity) and evolutionary divergence. (2) The similarity of the leader and the trailer regions reflects host specificity (i.e., human or simian) and must therefore arise from convergence. (3) Minimal energy foldings of the mRNAs show similar secondary structures (in particular around the splice sites). The conservation of pre-mRNA secondary structure shows that mRNAs are subject to selection constraints in addition to those associated with proteins. (4) The conserved secondary (helical) structures consist of nonhomologous subsequences, i.e., shifts have occurred. The observed shifts near the splice sites seem to be the simplest way of dealing with the dual constraints.

Introduction

Informatic molecules in biotic systems are subject to many constraints. In this paper we investigate whether, in addition to coding, there are secondary structural constraints on (pre-)mRNAs. Selection pressures can be detected by comparing the similarities between secondary structures with those between the corresponding primary structures. Sequences that have nearly identical primary structures may nevertheless have very different secondary structures, and vice versa. If minimal energy foldings of RNAs are similar, one might expect that the secondary structure is important for the functioning of the RNA and is maintained by selection.

We examined these structural features on four homologous E1a mRNA sequences of the adenoviruses: Ad5, Ad7, Ad12 (human), and SA7P (simian [African green monkey, Cercopithecus aethiops]). The expression of the E1a transcription unit, together with that of E1b, is required for complete transformation and oncogenicity of the host cell (reviewed by Van der Eb and Bernards [1984]). We show that their evolutionary relationship, as seen in their primary structure, reflects function rather than host.

The minimal energy foldings of these pre-mRNAs show that they have certain secondary structures in common. These similarities suggest that there are multilevel selectional constraints on the secondary structure of the mRNA.

Material and Methods

We have chosen the E1a mRNA of four adenoviruses as the subject of our study primarily because their relatively small size (∼1,000 nucleotides) makes minimal

1 Key words: adenovirus E1a, mRNA minimal energy folding, mRNA evolutionary constraints, sequence alignment, adenovirus phylogeny, homology assessment.

Address for correspondence and reprints: Dr. P. Hogeweg, Bioinformatica, Padualaan 8, De Uitof, Utrecht, The Netherlands.

© 1987 by The University of Chicago. All rights reserved.
0737-4038/87/0403-0008$02.00

300
energy foldings of the entire mRNA feasible. In addition, conflicting "functional" patterns can be expected in the sequences because (1) three of them (Ad5, Ad7, and Ad12) are human adenoviruses and one (SA7P) is a simian (African green monkey [Cercopithecus aethiops]) adenovirus and (2) the degree of oncogenicity varies: Ad12 and SA7P are highly oncogenic; Ad7 is weakly oncogenic; and Ad5 is nononcogenic (oncogenicity is partly related to the E1a mRNA; Van der Eb and Bernards 1984).

The four adenovirus DNA sequences coding the E1a's have been determined: Ad5 (Perricaudet et al. 1979; Van Ormondt et al. 1980a; Bos et al. 1981; it is almost identical to Ad2), Ad7 (Dijkema et al. 1982), Ad12 (Perricaudet et al. 1980; Sugisaki et al. 1980), and SA7P (Dekker et al. 1984).

All four E1a pre-mRNAs can be processed to yield two main splice products—13S and 12S mRNA—early in infection. In addition, for the human adenoviruses, a 9S mRNA is known that mainly appears late in infection (for SA7P, no data are available concerning this latter splice product) (Spector et al. 1978; Chow et al. 1979; Ziff 1980; Wilson and Darnell 1981; D. Kimmelman, personal communication).

We have analyzed the pre-mRNA sequences from the 5' cap site down to the 3' poly-A attachment site, since it is known that the processing of the read-through mRNA by polyadenylation is usually the first event after transcription (Nevins and Darnell 1978). Thus, the sequence considered is the one relevant to splicing and, after splicing, to translation.

### Integrated Tree-Construction, Sequence-Alignment Procedure

Sankoff et al. (1972) were the first to use an estimated genealogical relationship among sequences to assist in aligning multiple sequences. This is appropriate, since an optimal alignment is biologically ill defined without reference to the tree of their relationship. A modified Sankoff et al. procedure was outlined earlier (Hogeweg and Hesper 1984b). That procedure, called TRIALS, modified and improved still further, is detailed here. Overall, it involves estimating the relationships of the sequences (tree construction), aligning the sequences in the light of the estimated tree, and checking to see whether the alignment suggests a different tree. If so, reiterate; if not, the process has converged. The method has five steps:

**Step 1.**—Calculate similarities between all pairs of sequences by the Needleman and Wunsch algorithm (1970), modified à la Fitch and Smith (1983), to (1) permit gap penalties of the form \( p = g + nr \) (where \( p \) = the total penalty; \( g \) = a penalty for their being a gap; \( r \) = the penalty per residue in the gap; and \( n \) = the number of residues spanned by the gap) and (2) permit penalizing gaps at the end of the sequence.

**Step 2.**—Formulate a starting tree from these similarities. We use the unweighted pair-group method of analysis of Sneth and Sokal (1973).

**Step 3.**—Realign the sequences according to the structure of the tree as follows: Each ancestral node has two immediate descendants, called sister sequences in analogy to sister taxa in taxonomy. Sister sequences are successively aligned and nodal sequences are computed so that an alignment of the set of sequences is obtained. The sister sequences are aligned by the Needleman and Wunsch (maximal match) algorithm with gap penalties of the form described in step 1. Among alignments with the same maximum similarity score, that alignment is used in which the gaps are arbitrarily located as far to the right (3') as possible. Gaps are treated as a fifth nucleotide when they already occur in the sequences to be aligned (which may be the case when two nodal sequences are aligned). However, newly formed gaps are fully penalized even
if they are introduced opposite gaps already present in the other sequences (otherwise, a profusion of gaps may arise in the later steps of the algorithm). A match was scored at full value irrespective of the ambiguity at that position in the two sequences; that is, in accordance with the idea of minimum mutation along the branches of the tree, no fractional matching was used. Nodal sequences must be generated by an algorithm equivalent to the first pass of Fitch’s parsimony procedure (1971), in which ambiguities are resolved as soon as a solution compatible with the parsimony paradigm is available.

**Step 4.**—Represent the tree with branch lengths corresponding to the number of nucleotide substitutions between nodes.

**Step 5.**—Reiterate the procedure, calculating a new set of similarities on the basis of sequence alignment, if necessary.

This alignment algorithm suffers from the same shortcoming as does any agglomerative clustering procedure: (arbitrary) alignment choices made earlier (i.e., near the tips [leaves] of the tree) are not changed in view of branches added later (Hogeweg and Hesper 1984b).

**Parameter Choices**

The gap penalty used in this study set \( r \) equal to zero so that the penalty was independent of length. The value of \( g \) was varied. End gaps were not penalized. Reiteration (step 5) was not performed.

These choices are now the default values in our algorithm. We believe that they are best for sequences having 50%–80% identity, such as the mRNAs studied here.

**Subsequences**

The mRNAs were divided into eight regions (see fig. 1) that were distinguished on the basis of coding capacity and degree of similarity. These regions were analyzed separately for the following reasons: (1) to prevent an outstanding local dissimilarity (e.g., a large gap for which it is hard to choose an appropriate weighting) from determining the initial tree and unduly influencing the final alignment, (2) to test whether the similarity differed over the various mRNA regions, and (3) if they should differ, to relate this to the conflicting functional patterns, the coding capacity, etc.

The final alignment was used to interpret the secondary structure. It is therefore essential that the alignment be derived by a uniform procedure and be independent of secondary-structure considerations.

**Secondary-structure Generation**

The generation of secondary structures by minimal energy folding is hampered by the absence of reliable free-energy estimates. The thermodynamic energy values reported in the literature vary considerably (e.g., Salser [1977] vs. D. H. Turner [personal communication]; Freier et al. 1985). Moreover, energy estimates from the correct foldings of tRNAs (Papanicolaou et al. 1984) yield yet other values. Therefore, many researchers resort to consensus folding without consideration of energy minimization.

Because no consensus in secondary structure can a priori be expected in mRNAs (unlike the case for small nuclear RNAs), we use minimal energy folding, but the occurrence of similar (consensus) secondary structures among the different sequences was used to reject alternative foldings.

The base content of the pre-mRNA sequences under consideration differs from that of most mRNAs in that the G and U bases are overrepresented, whereas most
Table 1

<table>
<thead>
<tr>
<th>EXTERIOR</th>
<th>INTERIOR CLOSING PAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOSING</td>
<td>GU</td>
</tr>
<tr>
<td>GU ......</td>
<td>-0.3</td>
</tr>
<tr>
<td>UG ......</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

NOTE.—Data are the stacking energies of the interior G-U closing pairs that were used for the generation of the presented foldings. The other base-pair energies used (both the remaining stacking energies and the bulge-, hairpin-, and interior loop-destabilizing energies) are those reported by Salser (1977). Data are in kcal/mol.

mRNA sequences show an overrepresentation of G and C bases vis-à-vis A and U bases, and of C vis-à-vis G bases. Therefore, potential G-U bonds play a more important role and the generated foldings are sensitive to the G-U energy values, which are quite varied in the literature. Only when the Salser energy values (1977) were used, with the increase of G-U stacking energy shown in table 1, was a consensus secondary structure generated in the molecules. Other energy estimates—the default values of Salser (1977), Papanicolaou (1984), and others—were tested, and each generated a different folding and showed no similarity among the pre-mRNA molecules. The minimal energy folding of the E1a pre-mRNA of human adenovirus 40 (A. V. Loon, personal communication) conforms perfectly to the pattern described in this paper (data not shown).

Minimal energy foldings were generated by ENFOLD, a generalization of the maximum-match algorithm of Nussinov and Jacobson (1980) in which the nonmonotonicity occurring during minimal energy foldings is resolved via a backtracking mechanism: destabilizing loops are entered provisionally if no stabilizing bonds are available and are maintained if they lead to stabilizing helices; otherwise, they are ignored in later steps of the algorithm (Hogeweg and Hesper 1984a).

The algorithm generates not only the minimal energy foldings of the entire sequence but also those of any selected subsequence. We generated the foldings of the mRNA in steps of 50 nucleotides because the messengers are formed sequentially during the transcription process and because folding commences as soon as parts of the molecules are formed. The molecule may well be caught in a local energy minimum instead of switching to the global one, since the latter can only be attained by disruption of previously formed helices (the sequential folding may perhaps also be maintained by interaction with other molecules, e.g., interaction with snRNPs during transcription). Therefore sequential folding can be considered in addition to the global folding when one is searching for similar secondary structures based on energy minimization (see fig. 4, Ad7).

The minimal energy foldings were represented according to the conventions proposed by Hogeweg and Hesper (1984a). To minimize unwarranted suggestions regarding spatial relations and to facilitate the comparison of foldings of sequences and/or subsequences, we preserved the linear structure of the RNA molecule in the representation. An important aspect of this representation is that this simple visualization generates new characterizations of locations in the secondary structure. For example, the “highest top” represents the location of the largest hairpin-like structure, however complicated its internal structure may be.
FIG. 1.—Fig. 1A: The consensus tree and the associated alignment of the four E1A sequences of the Ad5, Ad7, Ad12, and SA7P as generated by our computer program, TRIALS. The Ad sequences are from humans, and the SA sequence is from the African green monkey (Cercopithecus aethiops). The alignment
Results

Alignment

The entire mRNA sequences did not reveal clearly distinguishable pairwise similarities (the identities between pairs of the Ad7, Ad12, and SA7P sequences were all ~58% and the identities between these three sequences and the Ad5 sequence were all ~51%). Thus, while Ad5 is more distinct, no decision could be made on that basis regarding the phylogeny of the other sequences. Therefore we investigated the coding and noncoding sequences separately.

Conservation of the reading frame requires insertions and deletions to be in multiples of three bases. Hence, the generated alignments and the associated tree(s) for the coding sequences were evaluated in terms of whether the gaps introduced were in multiples of three nucleotides (triplets). Since this is not a requirement of the algorithm, the introduction only of triplets is a significant result. The tree calculated from combining the two (coding) exons (regions 2–5 and 7 in fig. 1) was the one that produced an alignment of nucleotide sequences consistent with the preservation of the reading frame. This same tree is formed by most of the coding regions considered separately, especially by the regions of average similarity (regions 2, 4, and 7 in fig. 1). The same gap penalty ($g = 3$) preserves the reading frame over nearly all coding regions, irrespective of the degree of similarity, when this particular tree is used; when other trees are used, gaps are introduced that do not preserve the reading frame.

We preferred the alignments (in terms of number and relative position of the gaps) of the noncoding regions (leader, intron, and trailer: regions 1, 6, and 8 in fig. 1) that were produced with a lower gap penalty—we.e., in which $g = 2$ and each produces a different tree. The tree generated by the coding sequences appeared to be the best common base for the alignment of the three noncoding regions. From this we conclude that this (consensus) tree most likely represents the phylogeny of the viruses.

The consensus tree and the associated alignment of the four E1a sequences as produced by the TRIALS algorithm are shown in figure 1. The sequences of Ad12 and SA7P are clearly the most similar. Since these are sequences of the highly oncogenic adenoviruses, the evolutionary relationship of these viruses reflects the degree of oncogenicity rather than the viral host (human vs. ape). Note, however, that the rooting of the tree is based on similarity and may not necessarily reflect time of divergence.

The above-cited alignment was produced on the basis of the regions indicated in figure 1 (regions 1–8), as well as on the basis of the entire mRNA. Because it aligned the functional pyrimidine stretches downstream of the splice acceptor, we much preferred the alignment of the noncoding intron region (region 6) generated on the basis of the (consensus) tree of the coding sequences to that generated on the basis of a tree

was produced based on (1) the regions indicated in the fig. (regions 1–8), which were distinguished on the basis of coding capacity and of degree of similarity and (2) the entire mRNAs as described in Alignment. The gap penalty used in the alignment was higher for the coding sequences ($g = 3$) than for the noncoding sequences ($g = 2$). Positions that are identical in at least three of the four sequences are in boldface and boxed. Numbers above the sequences refer to the 5' border of the regions (1–8) that were analyzed separately. The arrows indicate the splice junctions: the 12S-donor (12S-a, 12S-b, and 12S-c, respectively, for the site in SA7P, Ad5, and Ad7 and Ad12), 13S-donor, and the splice acceptor (SA) in the sequences. 5'D and 3'D are, respectively, the 5' and 3' element of the duplicate in the Ad5 sequence; yy and zz refer to the two subparts that can be realigned to form the alternative in fig. 1B (see Alignment). Fig. 1B: The alternative (preferable) alignment of part of region 7, including the 5' element of the duplicate (5'D), which is shown as one major gap in multiples of triplets (see Alignment).
calculated from the similarity in this region. The leader and trailer regions are also satisfactorily aligned on this consensus tree. Although each noncoding region generated different similarity trees when analyzed separately, both the leader and trailer regions reveal similarities reflecting host specificity: SA7P is clearly separate from the others. The (noncoding) intron region, however, does not seem to reflect this host relationship; thus, the intron seems to be tuned less to a host-specific transcription/processing system.

Since a good alignment of coding and noncoding sequences can be derived from the tree on the basis of coding sequence whereas the trees generated by the noncoding sequences produce nontriplet alignments of the coding sequences and a profusion of gaps in the noncoding sequences, we conclude that the tree based on the coding sequences reflects evolutionary divergence common to all stretches and that (other) local similarity structures are the result of convergence.

The alignment obtained is essentially the same as the one that we presented recently for the deduced amino acid sequences (see Dekker et al. 1984), but there are some small differences in the location of gaps (particularly in region 4). The alignment improves the one of Van Ormondt et al. (1980b) for the same three human adeno-viruses, particularly in that the gaps are in triplets over the entire coding region (see fig. 1, region 4).

The alignment has the following interesting features:

1. The 12S splice donors of Ad5 and of SA7P are located on sites that are not homologous to that of Ad7 and Ad12. However, the alignment shows that the Ad5 and SA7P sequences possess, at the site homologous to the 12S splice donor of Ad7 and Ad12, "remains" of a consensus donor; but both these remains are less complementary to the U1-interaction site involved in the splicing process than is the actual splice donor. Note that the transposition of splice sites (particularly the 12S donor site) does not reflect the evolutionary divergence, whereas parallels in length and hypervariability of the sequences do (cf. the results of Perricaudet et al. [1980]). This result might be connected with the suggestion that the 12S mRNA protein product may not be crucial: no specific function for it has been detected (Ricciardi et al. 1981; Montell et al. 1982).

2. The large gaps, as well as the regions of hypervariability as shown in the alignment, are associated with the location of the splice sites (i.e., the 12S and 13S splice donor and the splice acceptor; see also the main splice area [discussed below and shown in fig. 2]).

3. The Ad5 sequence shows a duplication in the region downstream from the splice acceptor (region 7), the 5' element (5'D in fig. 1) of which forms the stretch aligned with the large gaps in the other three sequences: the single (unique) element of the other mRNAs aligns with a slightly better fit to the 3' element (3'D in fig. 1) of the Ad5 duplicate (i.e., in all four cases, the 3' terminal coding region). This alignment corresponds with the alignment of the deduced amino acid sequences (Dekker et al. 1984). Two small subparts of Ad7, Ad12, and SA7P (yy and zz in fig. 1) have been aligned to intermediate parts of the 5' element of the duplicate in the Ad5 sequence. This fit, (fig. 1A), is only slightly better (in terms of number of gaps and substitutions) than the preferable in-reading frame alignment flanking the 5'element of the duplicate of Ad5. This alternative is shown in figure 1B.

4. The region 3' flanking the start codon (region 2) is so heterogeneous in its nucleotide and amino acid composition that the region aligns partly out of the reading frame.
FIG. 2.—Detail of the region closed off by the main stem, including the "early" used splice area. A, Ad7; B, Ad12; C, SA7P; and D, Ad5. By including the gap stretches as single-stranded parts, the primary and the secondary structure can be represented in direct relation. Similarity of the primary sequence is higher than average (■), average (—), or nonexistent (— —). Broken lines (— —) indicate gaps in the alignment (see fig. 1).
Fig. 3.—Minimal energy foldings of the four homologous E1a pre-mRNAs. The foldings are represented as described by Hogeweg and Hesper (1984a). Each base pair is shown by a horizontal line whose length spans the distance of that representation. Hairpin loops appear as flat tops, interior loops and bulges as intermediate plateaux, helices as sloping hillsides, and branching regions as valleys. The sequences and their calculated free energies are as follows: fig. 3A: Ad7, $-384.1$ kcal (the subsequences 1–250 and 251–999 were folded separately; see Material and Methods); fig. 3B: Ad12, $-369.7$ kcal; fig. 3C: SA7P, $-454.7$ kcal; and fig. 3D: Ad5, $-496.3$ kcal. L = Leader hairpin; ED = “early” donor hairpin; A = acceptor hairpin; ↓ = location of initiation AUG and of splice sites (13S and 12S = “early” splice donors; 9S = “late” splice donor, and SA = splice acceptor, [see Material and Methods]). Vertical lines within the pattern show the point midway between the nucleotides paired by the horizontal line.
Secondary Structure of the mRNA

The Ad5, Ad12, and SA7P secondary structures shown in figure 3 are the foldings corresponding to the overall energy minimization. The free energies are $-496.3$ kcal for Ad5, $-369.7$ kcal for Ad12, and $-454.7$ kcal for SA7P. The minimal energy folding of the entire Ad7 sequence was dissimilar to the other foldings in that, apart from the leader, it revealed only one main substructure (hill); this was not subdivided in the same way as it was in the other sequences. On studying the sequential folding (discussed in the following section), we found that this divergence did not occur until the last step. Prior to this step, the pattern corresponds to the pattern of the other viruses. The secondary structure of Ad7, shown in figure 3, was obtained by folding the subsequences 1-250 and 251-999 separately. The total energy of these foldings is only a little less negative than the total energy of the minimal energy folding, i.e., $-384.1$ kcal versus $-398.4$ kcal.

The overall structures show some striking similarities, some occurring among all four structures and others between pairs of them. All four foldings show the presence of a large main substructure (the largest fold), the accessible position of the initiating AUG codon at the base of a local hairpin, and the configuration associated with the main (early-used) splice area (including the 12S and 13S donor and their common acceptor site).

The mRNA of Ad5 is $\sim 130$ nucleotides longer than the mRNA of the other adenoviruses. According to our alignment (fig. 1), this is the result not of the presence

![Diagram of sequential folding pattern of Ad12 in steps of 50 nucleotides](image)

FIG. 4.—Sequential folding pattern of Ad12 in steps of 50 nucleotides. —— = Range of local hairpin in which represents the hairpin loop. Blank stretches are either single stranded or base paired to remote sites (long-range interactions), —— = Long-range interactions in the entire pre-mRNA. Long-range interactions are less constant than local interactions in the sequential folding. The splice sites are indicated (12S and 13S = “early” splice donors; 9S = “late” splice donor; and SA = splice acceptor). Numbers above the horizontal line indicate the nucleotides of the pre-mRNA sequence with respect to the 5’ cap site.
of additional sequence at the 3' terminus but rather of an insertion just downstream from the splice acceptor. The folding shows that, after forming the common main helix, Ad5 has 100 nucleotides left; these form a separate fold and do not disturb the previously formed correspondence.

*Sequential folding.*—The sequential folding of Ad12 is shown in figure 4 as an example. It shows that local helices are maintained and that the main stem is formed stepwise after the splice area has been included. Since the structures are formed gradually and very little reshuffling occurs, we conclude that this sequential folding represents an energetically feasible path for the in vivo process. It is interesting to see that the "stable" structures surrounding the splice sites (the major A and ED hairpins [see below]) are much larger—i.e., include more nucleotides—than the average close-range interactions described by Nussinov and Tinoco (1981).

*The structure around the splice sites.*—The large folds (see fig. 3 and, for detail, fig. 2) are strikingly similar: they contain two major hairpins (ED and A hairpins shown as the two highest hills) in which the 12S and 13S splice donors and the splice acceptor are located. This splice area forms a separate substructure; it is closed off by part of the main stem (helix). The splice acceptor is located in the second hairpin (A hairpin) of the large substructure (fold) common to all sequences (see fig. 3). The local structure is much the same in all cases (see fig. 2). This invariance in the secondary structure contrasts strongly with the relatively large variation, around the splice acceptor (fig. 1), in the primary structure. The pyrimidine-rich part flanking the splice acceptor site is the most conserved part of this region (it is the consensus sequence of the nuclear mRNA acceptor). In the secondary structure this part forms the top of the A hairpin. The 12S and 13S donors are separated by the first major hairpin (ED) loop. Both donors are located in an accessible position, either in a single-stranded area or in a branching area (valley) at the base of the fold concerned. Note that only Ad7 and Ad12 share a homologous 12S splice donor location, as can be seen in the alignment (fig. 1). The 12S donor of SA7P is located even more upstream than that of Ad5. We predicted this location on the basis of the primary structure (Dekker et al. 1984), which has now been experimentally verified by D. Kimmelman (personal communication). Its location in the minimal energy folding fits the patterns observed here, notwithstanding its 5' deviation.

Minimal energy folding generates similar secondary structures in all four messengers. We therefore conclude that there must be selection for conservation of the secondary structure of the pre-mRNAs and, therefore, that this structure should perform some function consistent with the variation observed. The primary purpose of the unspliced messenger is to be spliced. The primary purpose of the spliced messenger is to be translated. The observed secondary structure of the leader sequence is clearly selected for its role in the translation process (see McReynolds et al. 1978; Lomedico et al. 1979; Konings et al. 1987); the other features may play a role in the splicing process.

**Comparison of Primary- and Secondary-Structure Similarity**

Minimal energy secondary structures can differ greatly as a result of slight differences in the sequence (see, e.g., Hogeweg and Konings 1985). The overall identity (50%-60% pairwise) of the primary structure is certainly not sufficient to cause the similarities that appear in the secondary structure. That the degree of sequence identity and the similarity of minimal energy foldings do not vary in parallel is shown (fig. 5) by the dissimilar folding patterns of region 5, which is 85% identical in sequence. In contrast, the folding patterns of the overall mRNAs are quite similar (see fig. 3), although the sequences are much more variable.
Throughout the mRNA, nonhomologous stretches are used to form corresponding secondary structures. Figure 2 shows the interdependence of the similarity of primary and secondary structure for the main splice area, which includes the homologous region compared above. By including the gap regions as single-stranded parts, the primary and the secondary structures can be represented in direct relation (see fig. 2). The figure shows that the similarity of the secondary structure is conserved by shifts in the stretches that are used to form helices. Thus, in the course of evolution, mutations acceptable to the protein could be accommodated by small or larger shifts of base-pair stretches to conserve secondary structure. This seems to be an elegant solution.
for coping with the dual (multiple) constraints on the (one) sequence. The similarities between the configurations concerned (fig. 2), including the hypothesized shifts, are in agreement with the phyletic tree deduced from the sequences. The shifts and the tree emphasize the smallest divergence of Ad12 and SA7P (compare the range of base pairings of homologous stretches).

Discussion

We have examined patterns observable in the primary (coding as well as non-coding) and minimal energy secondary structure of four functionally identical mRNA nucleotide sequences and, in particular, the interdependence of these patterns.

On the basis of the similarity between the obtained minimal energy foldings and the protein structure, we concluded that there must be multiple constraints on mRNA sequences. No quantitative criterion is available for measuring secondary-structure similarity. However, the significance of the similarity can be further substantiated by experimenting with more or less scrambled sequences. Hybrid sequences (i.e., sequences into which the gaps and inserts from one sequence are inserted into another) and mutated sequences (in which the gaps and inserts are retained but randomly selected nucleotides are changed) all generated minimal energy foldings in which the pattern that was common to the original set of sequences was lost (mutation frequencies as low as that of the most conserved region [i.e., 15%] were tested).

Although the conservation of these structures suggest that they are under selection, no definite conclusion can be drawn relating particular secondary structures either to the different processes in which the molecule is involved or to their intramolecular interactions with associated molecules such as proteins and small-nuclear-RNP complexes. Nonetheless, the correspondence between known functional sites and the inferred structures suggests processes—such as translation, splicing, and survival (stability)—in which the structures could be involved. Furthermore, these patterns were obtained by considering the molecules in isolation.

Very heterologous regions of the nucleotide (and amino acid) sequences are associated with splice junctions. This feature, observed by others, has been related to the observation that exon junctions are generally located on the surface of the protein structure so that any variation flanking these junctions causes minimal disruption of the fundamental architecture of the molecule (Craik et al. 1982). This minimal disruption, however, is only possible if these variable flanking regions also consist of surface-exposable amino acids. In our set of sequences, the variable coding region 3' flanking the splice acceptor is used to stabilize the common configuration in the main splice region (see fig. 2). In the Ad5 mRNA, this part includes the 5' duplicate unit that we mentioned above (see Results and fig. 1). If we assume that this secondary pattern is important for the pre-mRNA, then the relevant (heterologous) stretches in the various mRNAs were primarily selected for that. A slight tuning of the corresponding amino acids has produced a hydrophilic region that causes only minor disturbance of the protein in that this region will be located on the surface.

On the other hand, the heterologous region enclosed by the 12S donors of SA7P and Ad12 (fig. 1) does not seem to play a crucial role in the secondary structure. If we relate this observation to the fact that the main (oncogenic) divergence of the adeno E1a mRNA is coupled with the first exon (5' flanking the 13S donor site), we conclude that this heterologous region could be a good candidate for a surface-exposed functional change (Craik et al. 1982, 1983). This conclusion is supported by the closest similarity of the amino acid sequence of the highly oncogenic Ad12 and SA7P (also see Dekker et al. 1984).
Both of the corresponding heterologous amino acid blocks are indeed hydrophilic and contain turn amino acids, so they are likely to be located on the surface. This feature applies over the whole acceptor-associated part that is common to all sequences and is significant for the 12S donor-associated part and for the 5' duplicate unit of Ad5. The duplicate of Ad5 has diverged so that the 5' unit is tuned to allow (1) the formation of a configuration of the pre-mRNA that is more or less the same as that of the others and (2) the surface exposure of the unit in the protein. The 3' unit, on the other hand, has evolved to be more hydrophobic than the corresponding parts of the other viruses; apparently this change was necessary to stabilize the protein.

The pattern that emerges from comparing the similarities of the primary and secondary structures of the four homologous mRNAs indicates that selectional constraints are dealt with, in so far as possible, by separate parts of the sequence. Notwithstanding the inherent globality of secondary structure (minimal energy folding), secondary structure seems to be conserved by specific stretches (which show little similarity between the sequences) that use slightly shifted parts of (highly conserved) stretches to form similar configurations in the various molecules. Since secondary-structure conservation seems to be stringent in the vicinity of splice sites, variable stretches are found near splice junctions.

Acknowledgments

We would like to thank H. Van Ormondt for helpful discussions and suggestions concerning this manuscript and S. M. McNab for suggesting language corrections.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received May 12, 1986; revision received November 14, 1986.