Title
SEQUENTIAL FOLDING OF A MESSENGER RNA MOLECULE

Author
Tinoco, I.

Publication Date
1981-06-01

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

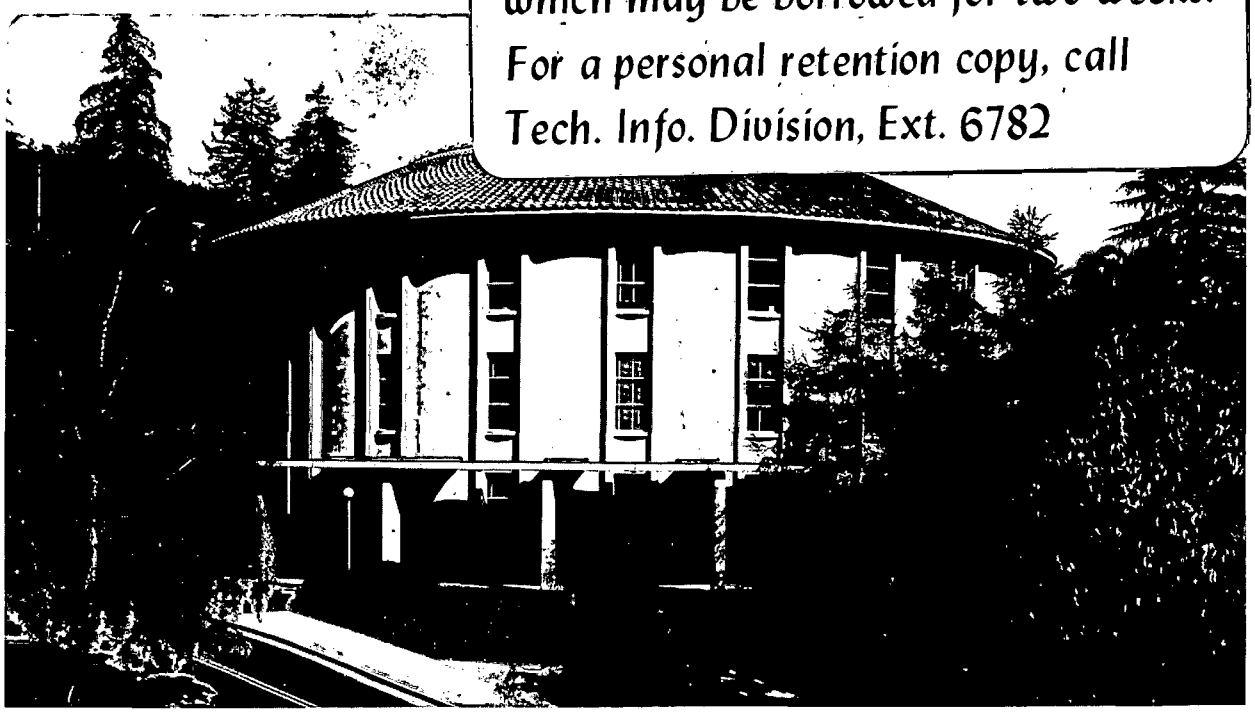
CHEMICAL BIODYNAMICS DIVISION

Submitted to Molecular Biology

SEQUENTIAL FOLDING OF A MESSENGER RNA MOLECULE

Ruth Nussinov and Ignacio Tinoco, Jr.

June 1981



Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

Sequential Folding of a Messenger RNA Molecule

Ruth Nussinov and Ignacio Tinoco, Jr.

Lawrence Berkeley Laboratory
Department of Chemistry and
Laboratory of Chemical Biodynamics
University of California, Berkeley
Berkeley, California 94720

This work was supported in part by National Institutes of Health Grant GM 10840 and by the Assistant Secretary for Environment, Office of Health and Environmental Research, Biomedical and Environmental Research Division of the U.S. Department of Energy under Contract No. W-7405-ENG-48.

Running title: RNA Folding

Abstract

The existence of a new, efficient algorithm for secondary structure prediction enables us to study the folding pattern of an mRNA chain. Our results indicate that successively longer RNA sequences with the same 5' end fold sequentially, usually keeping the stable, close range hairpin loops and rearranging the long range stems. This path will shorten the time the messenger RNA molecule needs in order to attain its preferred structure. It can also align splicing sites in a favorable orientation before the whole molecule is synthesized.

Our studies were carried out on the late SV40 precursor and processed mRNA.

Introduction

Most calculations of structure on long messenger RNA molecules have attempted to predict the secondary structures which have the lowest free energies. The present study investigates not just the final, molecule-specific lowest energy secondary structure, but rather the general, dynamic process of folding. Does the mRNA start the search for its preferred structure only when it is fully synthesized, or does it start folding, sequentially, as it is made, constantly keeping at least part of the already folded structure? The advantages of the latter process are clear. It reduces the time required to attain the final folded form.

NMR studies of the relatively short $tRNA^{Phe}$ molecule (Boyle, Robillard and Kim, 1980) imply that the tRNA folds sequentially during its synthesis. Despite the difference in the order of magnitude, the steadily growing long mRNA molecules do not constantly reshuffle their secondary structure either. Rather, for the most part the closely knotted short range interactions do not change; usually only the long range bonds are broken and closed again to form a different structure.

Figure 1 shows schematically our model for the secondary structure folding pattern of the growing messenger chain. The fact that the final lowest energy structure contains most of the already formed close range interactions indicates the higher stability of close range over long range interactions.

These studies were carried out on the late SV40 precursor mRNA. They were made possible by the recently developed, extremely rapid algorithm for search of lowest energy secondary structures (Nussinov and Jacobson, 1980).

Methods

In order to examine the secondary structure folding pattern of a growing mRNA chain, we have looked at successively longer RNA sequences with the same 5' end (see also Boyle et al., 1980). Starting with 20 nucleotides, the number of nucleotides was incremented each time by 20 (1-20, 1-40, 1-60, ...) until the maximum of 950 nucleotides was reached. For each RNA section the total energy of the minimal energy structure was computed and the structures themselves were given at intervals of 100 nucleotides.

Owing to the efficiency of the algorithm we had to run it only once to find out not only the best structure for the full length of the section specified (e.g., from nucleotide number 1 to number 950), but also for all subsections contained within it (e.g., 1-100, 1-200, ..., 1-950, since in this study we were interested in subsections with the same 5' end). Because of present disk space limitation we could not run the algorithm on more than 950 nucleotides. This run took 10 hours CPU (14 hours total elapsed time) on the VAX 11/VMS.

A description of the algorithm principle is given by Nussinov et al., 1978 and by Griggs, 1977; its application to nucleic acid sequences is described by Nussinov and Jacobson, 1980.

The lowest-energy algorithm contains an N^2 matrix (N being the length of the nucleic acid). At the end of a single computer run, one half of the matrix contains the lowest energy values of the best secondary structure of the whole section length and every part of it. For our 1-950 nucleotide example, we have also the lowest energies for 1-800, 200-600, 380-450, etc. From the other half-matrix one can easily read the actual secondary structures of all of these sections.

Clearly, this algorithm is more efficient than other existing secondary structure codes (Pipas and McMahon, 1975; Fitch, 1972; Studnicka et al., 1978). Not only is it about 100 times faster than another code, but a single computer run contains all the needed information about the sequence and all its subsections.

In calculating the free energy contributions of the different loop structures, we have followed the method which was developed and outlined by Tinoco et al., 1971; Tinoco et al., 1973; Delisi and Crothers, 1971; Gralla and Crothers, 1973a; Gralla and Crothers, 1973b; and Borer et al., 1974. We used the free energy values summarized by Salser, 1977. These values do not take into account the stabilizing contributions made by extra stacking interactions at the ends of helical stems (as shown by studies of dangling ends, Martin et al., 1971; Romaniuk et al., 1978). Other drawbacks implied by our dependence on these values are that (a) we do not evaluate whether a particular structure is sterically possible, nor do we (b) consider the non-Watson-Crick base pairing interactions present in the tertiary folding of tRNAs as revealed by X-ray crystallographic studies (Kim et al., 1978; Jack et al., 1976; Holbrook et al., 1978). In addition, (c) the stabilizing effect conferred by the binding of positively charged ions such as Mg^{2+} and spermidine^{3+} have not been taken into account in our calculations.

The differences in energies of successive structures are given in Figure 2. The detailed structures of recurring hairpin loops which contribute to the minima in Figure 2 are drawn in Figure 3. Figure 4 presents, schematically, several stages in the folding of the growing RNA chain.

We have chosen to carry out these studies on the SV40 late precursor and processed mRNAs. The nucleotide sequence of this virus is known (Fiers et al., 1978; Reddy et al., 1978) and the locations for initiation of several of its mRNAs (Lai et al., 1978; Villarreal et al., 1979; Ghosh et al., 1978; Bina-Stein et al., 1979) and some of its splicing sites are also known. Throughout this paper the numbering system of Reddy et al. (1978) is used.

Results

(a) The final structure attained in this study.

Flattening the 1-950 nucleotide structure onto a plane, the molecule has an elongated shape which is held together by long range base pairs. Out of this basic trunk grow, at intervals, short range hairpin loop structures (loops A though R, Figure 4). Most of these hairpin loops have been formed already in previous structures and are shown in detail in Figure 3. The 5' end is base paired in the CS_1 stem.

This structure represents only part of the messenger RNA. The whole late SV40 mRNA extends till nucleotide 2592, where the poly A is attached. It is therefore likely that the long range base pairs will rearrange and thus the gross structure will change. This is seen in Figure 4 (nucleotides 200-1150), which continues the mRNA and shows a branched structure. A different partial structure in Figure 4 (nucleotides 1-500 and 1330-1780, without the intron) is again elongated.

Further modification of the program or the availability of additional computer memory will allow studying the full messenger length.

(b) The general sequential folding process.

At the end of the single computer run on the specified nucleotide section (1-950), all the lowest energy structures for all subsections, as well as their corresponding energies, are presented in the N^2 matrix. We have chosen to print some of the latter for those subsections which have the same 5' end and the 3' end varies at increments of 20 nucleotides, i.e., for the subsections 1-20, 1-40, 1-60, ... 1-940, 1-950.

Clearly, as the messenger molecule grows and folds it yields progressively lower energies. The differences in the total energies of these successive structures (i.e., $\Delta G_{1-40} - \Delta G_{1-20}$, $\Delta G_{1-60} - \Delta G_{1-40}$, etc.) are shown in Figure 2. The minima in this graph are usually the result of formation of close range loops (denoted A to R). The detailed structures of these are presented in Figure 3. Most often the loop heads (i.e., the number of unpaired bases at the head of the loop) of these loop structures have 5 bases. The average number of base pairs in these close range loops which contribute to the minima is 8.4. If the number of unpaired bases in the stem is more than one, an internal

loop is preferred over a bulge.

The close range loops are thus relatively stable with an average energy of -9.2 kcal/mol. This value can be compared with the energies of the close range yeast $tRNA^{Phe}$ stems: The D stem is -2.2, the anti-codon stem is -4.5 and the TψC stem is -3.9. The only relatively long range stem in the tRNA molecule is the acceptor stem with an energy of -11.9. A long range stem which keeps recurring in our successive structures, the CS stem at the base of the A, B and C stems (see Figures 3 and 4) has a calculated energy of -18.9 and involves bases separated by 135 or 156 nucleotides, depending on its location. The sequence 167-174 pairs, in all our structures, with either 4-11 or 25-32, which are exact repetitions of each other. See also Figure 3.

Figure 4 presents the gradual folding of the molecule. Successive lowest energy structures for subsections which are at 100 nucleotide intervals (i.e., for 1-200, 1-300, ..., 1-800, 1-950) were also printed at the end of the single computer run. These are drawn in Figure 4. Following the process of folding of the partial structures, it is evident that, in general, the short range loops are kept. The recurring small loops are marked by darkened stems in Figure 4; one sees that loops A, B, C and CS are formed and kept in all structures. As the mRNA molecule is synthesized, the already folded, stable, relatively small loops usually do not open. It is the reshuffling of the long range stems that changes the gross structure of the growing chain. If a close range loop does open and rearrange, another variant small loop is formed instead, using part of the same nucleotides (e.g., loops F and F'). Other small loop

structures or their variants are formed in the successively larger structures and are mostly kept through the (presently) final structure.

Since at present a maximum of 950 nucleotides could be studied (see Methods) and since we have seen that the close range structure does not change with chain elongation, we have applied the algorithm to sections with a different 5' end. Figure 4 presents the results obtained for the section 200-1150. Loop structures A-D are missing since loop D ends at nucleotide 204. From loop E on, the close range loop structures are quite similar to those of previous sections. However, the long range gross structure is different, as expected.

Figure 4 presents the structure obtained for the late 16S mRNA. It contains the nucleotides of the exon, which remains after splicing, plus 50 nucleotides on each side of the intron; it is the structure for nucleotides 1-1780 with the segment 501-1330 removed. Even for this large segment, the A, B and C loop structures are found with the CS stem at their base. Next, the same close range structure through loop I of nucleotides 1-600 is found. Loop J is not found since part of its nucleotides are missing.

(c) Results pertaining to sites of special biological interest.

The capped late leaders start at nucleotides 110, 182 and the major leader is at 243 (Lai et al., 1978; Villarreal et al., 1979; Ghosh et al., 1978; Bina-Stein et al., 1979). These are marked in Figure 4 . The first leader starts 6 nucleotides before loop B and the second starts 6 nucleotides after the CS stem and the third is at the E stem (Figure 3). Thus, all start next to stable secondary structures.

Figure 5 presents the energy of part of the intron as compared to

that of part of the exon. Both start at a common 5' end. Whereas one continues into the intron (1-950), the other (1-500, 1330-1780) has most of the intron removed (the 16S splicing occurs at positions 444 and 1380) and continues to the exon on the 3' (acceptor) side of the splice. The consecutively longer segments of the intron have consistently lower energy, indicating a more stable structure than the exon.

Figure 4b schematically shows the lowest energy secondary structure for the whole 16S intron, including the splice junctions. The intron has a compact secondary structure with the donor and acceptor sites of the splice relatively close to each other in spite of the 935 nucleotides separating them in the primary structure. That this section of the structure is extremely stable can be seen from the detailed structure of Figure 6. Clearly, the tertiary structure could bring the splicing sites closer to each other and align them properly for splicing (Nussinov, 1980).

Basically, the intron has the same structure as before (e.g., compare with nucleotides 1-950). The donor, 5' side of the splice is at the head of a small loop which is competitive to the I-loop.

The late 19S splices occur at positions 292 (donor) and 475 (acceptor). The first is between loops F and G and the second just before loop J.

Discussion

RNA structure is important in understanding many of the processes that take place in the cell. Several experimental tools exist today for determining the structures of large RNA molecules. These include (a) enzyme cleavage, (b) electron microscopy which indicates compact regions in the molecule as well as stable long range pairings, (c) cross-linking of two regions of the RNA molecule, (d) oligonucleotide binding studies, and (e) chemical modification. It may be that the best approach is the combination of theoretical computations with experimental results. The two methods complement each other. Good agreement between our algorithmic approach and experimental studies was obtained in the structure of the potato spindle tuber viroid (Gross et al., 1978) and with electron microscopic studies of the MS2 RNA bacteriophage (Jacobson and Nussinov, unpublished results). However, as was already pointed out (Nussinov and Jacobson, 1980), the tRNA cloverlead form was not obtained for all tRNAs tested.

The theoretically predicted structures (Figs. 3,4) may be not entirely correct. The original simple algorithm was rigorously proved (Nussinov et al., 1978). However, during its conversion to a biologically applicable tool (Nussinov and Jacobson, 1980) some subtle errors may have been introduced. These are continuously searched for in repeated applications and are weeded out.

The only parameters introduced into the algorithm are the free energies of stacked base pairs and of single stranded regions. Since these were determined using short model compounds, while here the chain length is several hundreds of nucleotides, inaccuracies may result. Also, the stabilizing effect of dangling ends at the ends of helices (by ~1-2 kcal per helix) (Martin et al., 1971; Romaniuk et al., 1978)

was not included in our calculations.

The free energies have not been measured for all possible structural topologies. For example, the free energy of the single stranded centers in multibranched structures such as are present in tRNAs are unknown, making the assignment of these values arbitrary. In our calculation, the unpaired bases in tRNA-type centers are treated as bulges, if stacking can occur between any two base-paired branches. If no stacked base pairs between branches exist, single stranded centers are treated as internal loops.

Tertiary interactions were not considered while determining the most stable structure. The reasons for this omission are several. (a) Tertiary structures, whether knotted or pseudo-knotted, have a very different topology than the secondary, planar structures. The algorithm we use finds only planar solutions. To our knowledge, no algorithms searching for the complete three dimensional structure exist to date. (b) Even had the algorthmic tools existed, we would not have known the free energies associated with the different types of tertiary bond formations, as these have not been determined. However, as it was shown for tRNAs that first the secondary structure is formed, followed by tertiary structure formation (Boyle et al., 1980), we do not consider this deficiency critical in predicting the structures.

It should also be noted that our algorithm treats the RNA molecule as if it was in an <u>in vitro</u> test tube, in physiological conditions. However, <u>in vivo</u>, the mRNA chain is not free; it interacts with proteins to form ribonucleoprotein particles (RNPs). This association may cause a structural change or may stabilize an existing structure.

Our knowledge of the folding process of RNAs is still very scant. Except for the free energy values for short RNA stretches (e.g., Tinoco et al., 1973), we still do not know the rules governing RNA folding. Some insight may be gained from studies on the folding of the bacterial 16S rRNA (Woese et al., 1980). Their secondary structure model was derived on the basis of comparative sequence analysis of over 100 species of eubacteria, chemical modification studies and nuclease susceptibility data. Based on repeating structures in these phylogenetically close sequences, they propose that in several cases the hairpin that is formed is not the most stable one that can form in a given location. It thus appears that either the energy rules are incomplete or else there are other, additional factors which influence RNA folding, such as protein-nucleic acid interactions and folding during transcription.

Not withstanding all the potential difficulties outlined above, we are convinced that the qualitative results concerning the sequential folding of an mRNA molecule are valid. Experimental studies need to be done, but it is reasonable to believe that sequential folding is a general feature of all mRNA structure, while details of the secondary structure will vary.

Recently we have become aware of work by Zuker and Stiegler (1981) modeled on the same principles of Nussinov et al. (1978) and Nussinov and Jacobson (1980). The structures produced by both programs are now being compared. The Zuker and Steigler code does not, however, have the extremely useful feature of allowing studies of all sequential substructures by one single run of the program.

Conclusions

As an RNA molecule is synthesized on its DNA template, the growing chain starts folding sequentially from the 5' end. Since sequential folding limits the number of folding pathways, it is both simpler and faster than constant, complete structural rearrangment.

In general, short range hairpin loops are more stable than helical stems which are very far apart on the primary structure of the RNA chain. Indeed, since the first tRNA sequence (Holley et al., 1965) it was clear that hairpins are important constituents of secondary structures of any RNA molecule. Tinoco et al. (1971), Delisi and Crothers (1971), Tinoco et al. (1973), Gralla and Crothers (1973a) and others (see Methods) have found that the most stable hairpins have 5-7 unpaired bases at their heads.

Determination of a three dimensional structure of a tRNA molecule (Kim et al., 1974; Jack et al., 1976; Holbrook et al., 1978) has shown that different hairpins can be stacked and thus increase the RNA stability. As shown above, the hairpins found by our secondary structure algorithm for the mRNA are more stable than those found in the tRNA.

Boyle et al. (1980), who have studied the $tRNA^{Phe}$ folding process found that, whereas correct secondary structures are formed, very few tertiary hydrogen bonds are established in the "folding intermediates". It is possible that, since the long range stems keep rearranging, no tertiary hydrogen bonds are formed in "folding intermediates" of longer RNA chains too.

The fact that both a short RNA molecule as well as a long one fold sequentially suggests that this might be a general folding pathway. The hairpins are close to each other and, as the growing RNA chain is

released from the RNA polymerase, they are likely to form first. Their high stability would then prevent any major rearrangement. A rapidly formed stable structure might aid in preventing degradation as well as bring together in a proper alignment the sequences which are to be spliced and ligated (Nussinov, 1980). Changes in the folding pathway might bring about formation of alternative, competing structures and thus splicing at other locations (e.g., 16S vs. 19S late SV40 mRNAs).

The preferred formation of short range hairpins over long range interactions would enable unfolding and translation of part of the RNA messenger without major structural rearrangment of the whole molecule. The folding process of mRNA in the cytoplasm after ribosome-translation may be sequential too.

If indeed the sequential folding pathway is a general folding pattern, secondary structure prediction may be greatly facilitated. In any secondary structure prediction algorithm the major problem at the moment is that one cannot investigate very long sequences. In some codes (e.g., Pipas and McMahon, 1975) that is due to both computer time and memory with a limit of ~300 nucleotides. With the present algorithm (Nussinov and Jacobson, 1980) the limiting factor is the memory.

Using the sequential folding pathway, the computer can mimic nature. It lends credibility to running secondary structure prediction programs on successive sections of an RNA molecule, as is often done now with long RNA chains.

Acknowledgments

We wish to thank A. B. Jacobson, with whom the program was developed, for her interest and discussions, and D. Canaani and G. Pieczenik for discussions and suggestions. We also thank K. Wiley and S. Wong from the Laboratory of Chemical Biodynamics computer center. This work was supported in part by National Institutes of Health Grant GM 10840 and by the Assistant Secretary for Environment, Office of Health and Environmental Research, Biomedical and Environmental Research Division of the U.S. Department of Energy under Contract No. W-7405-ENG-48.

References

Bina, M., Stein, A., Thoren, M., Salzman, N. and Thompson, J. A. (1979) Proc. Natl. Acad. Sci. USA 76, 731-735.

Borer, P. N., Dengler, B., Tinoco, I., Jr. and Uhlenbeck, O. C. (1974) J. Mol. Biol. 86, 843-853.

Boyle, J. Robillard, G. T. and Kim, S.-H. (1980) J. Mol. Biol. 139, 601-625.

Delisi, C. and Crothers, D. M. (1971) Proc. Natl. Acad. Sci. USA 68, 2682-2685.

Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) Nature 273, 113-120.

Fitch, W. M. (1972) J. Mol. Evol. 1, 185-207.

Ghosh, P. K., Reddy, V. B., Swinscoe, J., Lebowitz, P. and Weissman, S. M. (1978) J. Mol. Biol. 126, 813-846.

Gralla, J. and Crothers, D. M. (1973a) J. Mol. Biol. 73, 497-511.

Gralla, J. and Crothers, D. M. (1973b) J. Mol. Biol. 78, 301-319.

Griggs, J. (1977) Ph.D. Thesis, M.I.T.

Gross, H. J., Domedey, H., Lossow, C., Jank, P., Raba, M., Alberty, H. and Sanger, H. L. (1978) Nature (London) 273, 203-208.

Holbrook, S., Sussman, J., Warrant, R. W. and Kim, S.-H. (1978) J. Mol. Biol. 123, 631-660.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. and Zamir, A. (1965) Science 147, 1462-1465.

Jack, A., Ladner, J. and Klug, A. (1976) J. Mol. Biol. 108, 619-649.

Kim, S.-H., Sussman, J. L., Suddath, F. L., Quigley, G., McPherson, S., Wang, A., Seeman, N. C. and Rich, A. (1974) Proc. Natl. Acad. Sci. USA 71, 4970-4974.

Lai, C.-J., Dhar, R. and Khoury, G. (1978) Cell 14, 971-982.

Martin, F. H., Uhlenbeck, O. C. and Doty, P. (1971) J. Mol. Biol. 57, 201-215.

Nussinov, R. (1980) J. Theor. Biol. 83, 647-662.

Nussinov, R. and Jacobson, A. (1980) Proc. Natl. Acad. Sci. USA 77, 6309-6313.

Nussinov, R., Pieczenik, G., Griggs, J. and Kleitman, D. (1978) Soc. Ind. Appl. Math (C) J. Appl. Math 35, 68-82.

Pipas, J. M. and McMahon, J. E. (1975) Proc. Natl. Acad. Sci. USA 72, 2017-2021.

Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. and Weissman, S. M. (1978) Science 200, 494-502.

Romaniuk, P. J., Hughes, D. W., Gregoire, R. J., Neilson, T. and Bell, R. A. (1978) J. Am. Chem. Soc. 100, 3971-3972.

Salser, W. (1977) Cold Spring Harbor Symp. on Quant. Biol. 42, 985-1002.

Studnicka, G. M., Rahn, G. M., Cummings, I. W. and Salser, W. A. (1978) Nucl. Acids Res. 5, 3365-3387.

Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. and Gralla, J. (1973) Nature New Biol. 246, 40-41.

Tinoco, I., Jr., Uhlenbeck, O. C. and Levine, M. D. (1971) Nature 230, 362-367.

Villarreal, L. P., White, R. T. and Berg, P. (1979) J. Virol. 29, 209-219.

Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. L., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J. J. and Noller, H. F. (1980) Nucl. Acids Res. 8, 2275-2294.

Zuker, M. and Stiegler, P. (1981) Nucl. Acids Res. 9, 133-148.

Figure Captions

Figure 1: A schematic drawing of the sequential folding pathway of a growing RNA molecule. The 5' end is the same in all (a-d) drawings. In (a) hairpin A is formed and in (b) hairpin B is added to it. (c) depicts the formation of hairpins C and D and a long range L_1 stem. (d) shows that while A, B, C and D are kept, there is a long range structural reshuffling to replace L_1 by hairpin E and produce another long range stem, L_2.

Figure 2: This graph presents the differences in the total free energies of successive structures in a successively longer RNA sequence. Starting with the section 1-20 of the SV40 late sequence (Reddy et al., 1978) the energies are printed by the computer at 20 nucleotide intervals, i.e., for 1-20, 1-40, 1-60, etc. Here the difference between them, i.e., $\Delta G_{1-40} - \Delta G_{1-20}$, $\Delta G_{1-60} - \Delta G_{1-40}$, ... $\Delta G_{1-940} - \Delta G_{1-920}$ are presented as function of the section location.

A-R are hairpins contributing to the graph minima. Their detailed structures are given in Figure 3. CS and LR are longer range stems with CS occurring repeatedly (see Figures 3 and 4).

Figure 3: The detailed structures of the stems contributing to the minima in Figure 2. A-R are close range loop structures. CS_1 and CS_2 (the latter is not drawn since it is composed of the sequence 25-32, which is an exact repetition of 4-11, base paired with 167-174) are variant structures denoted CS in Figures 2 and 4. LR is a long range stem.

Figure 4: A series of schematic drawings depicting the sequential folding pathway of the SV40 late precursor mRNA.

Structures are drawn at successive 100 nucleotide intervals after starting with the first 200 nucleotides, i.e., 1-200, 1-300, 1-400, etc. Also shown are nucleotides 200-1150, and the sections containing mainly the 16S mRNA exon (1-500, 1330-1780). Figure 4b shows the intron (nucleotides 440-1390). The stems of the recurring hairpin loops are darkened. The A, B, C,...R loops and the CS stem (see legend to Figure 3) are those shown in detail in Figure 3 and marked also in Figure 2. The primed loops, e.g., F', means that this loop structure is a variant of F and uses part of the same nucleotides used in the F loop. A loop marked by a letter with a subscripted number, e.g., G_1, means that this is a loop close, in location, to G, but does not contribute to the minima in Figure 2 and not shown in Figure 3. The various leader starts are marked in the 1-300 nucleotide structure, and the 16S splicing sites are shown in the intron and exon.

Figure 5: This graph compares the energies of the sequential folding of nucleotides in the intron and exon. Both have the same 1-500 section. Whereas one (1-950 in Figure 4) continues into the intron (500-950), the other (1-500, 1330-1780 in Figure 4) jumps over the intron and continues into the exon (1330-1780). This graph shows that consistently the intron has a lower energy than the exon. A section of the secondary structure of the intron is shown in Figure 6.

Figure 6: A detailed secondary structure of the splice region of Figure 4b. The splicing sites are marked by arrows.

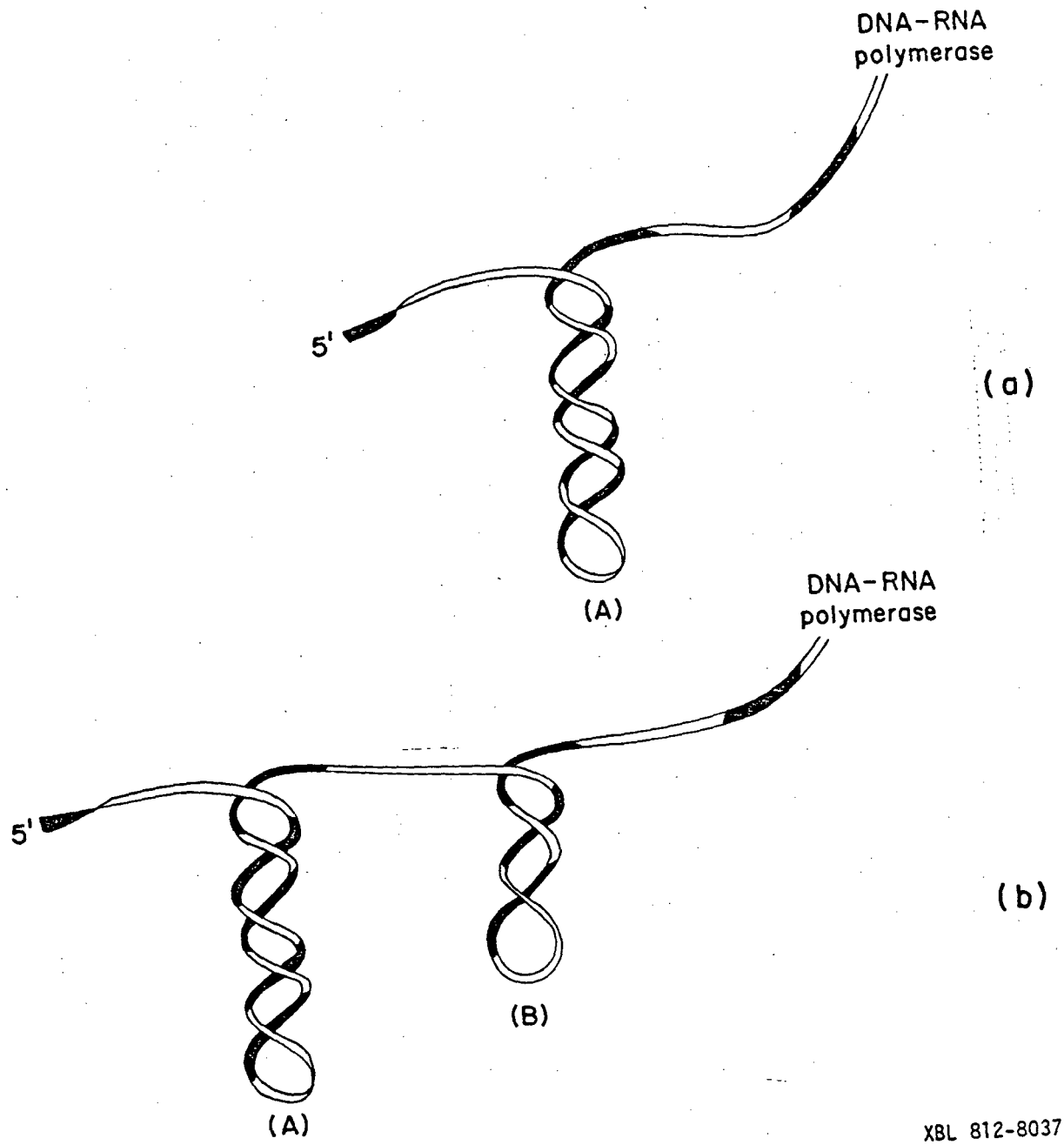


Figure 1

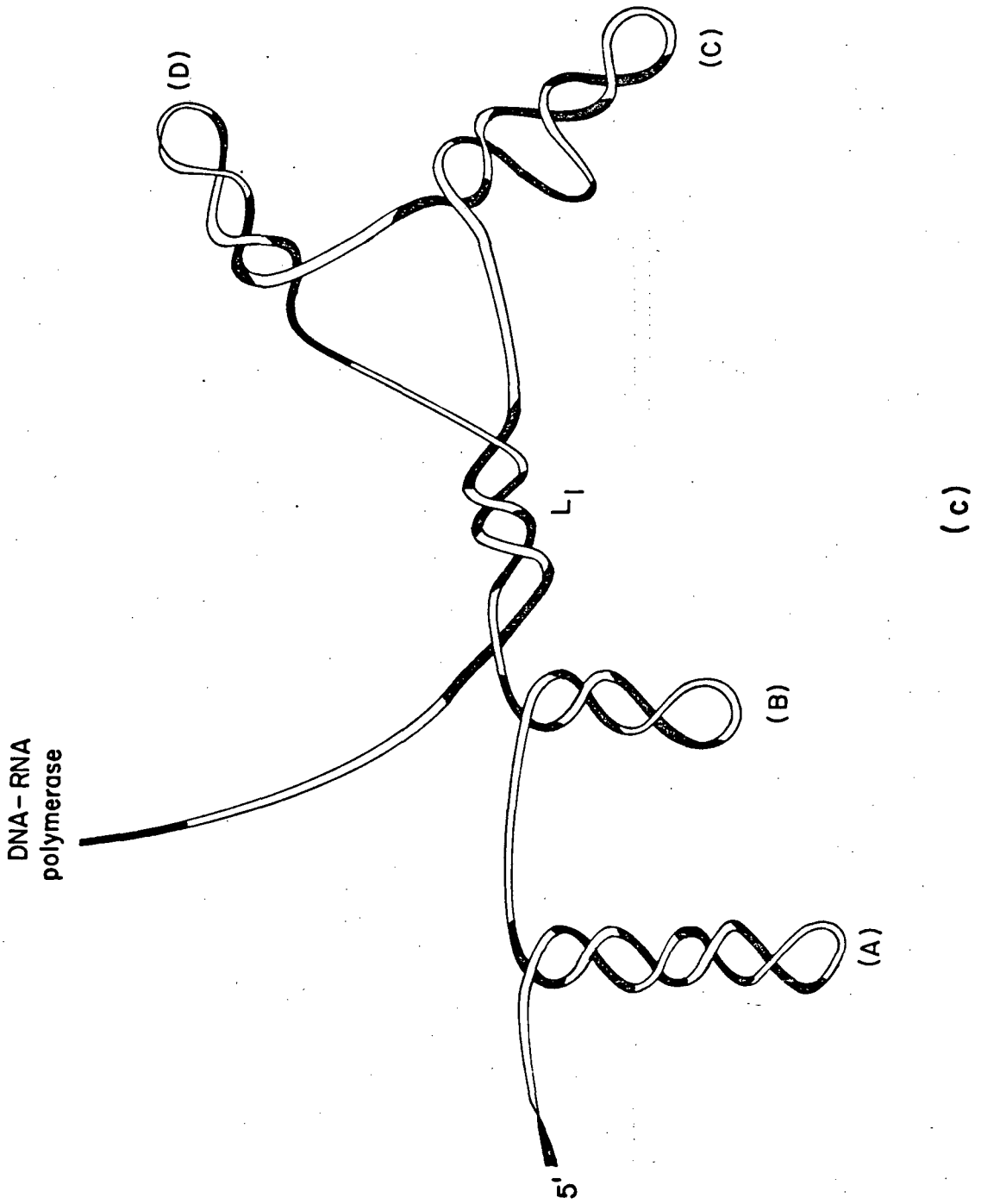
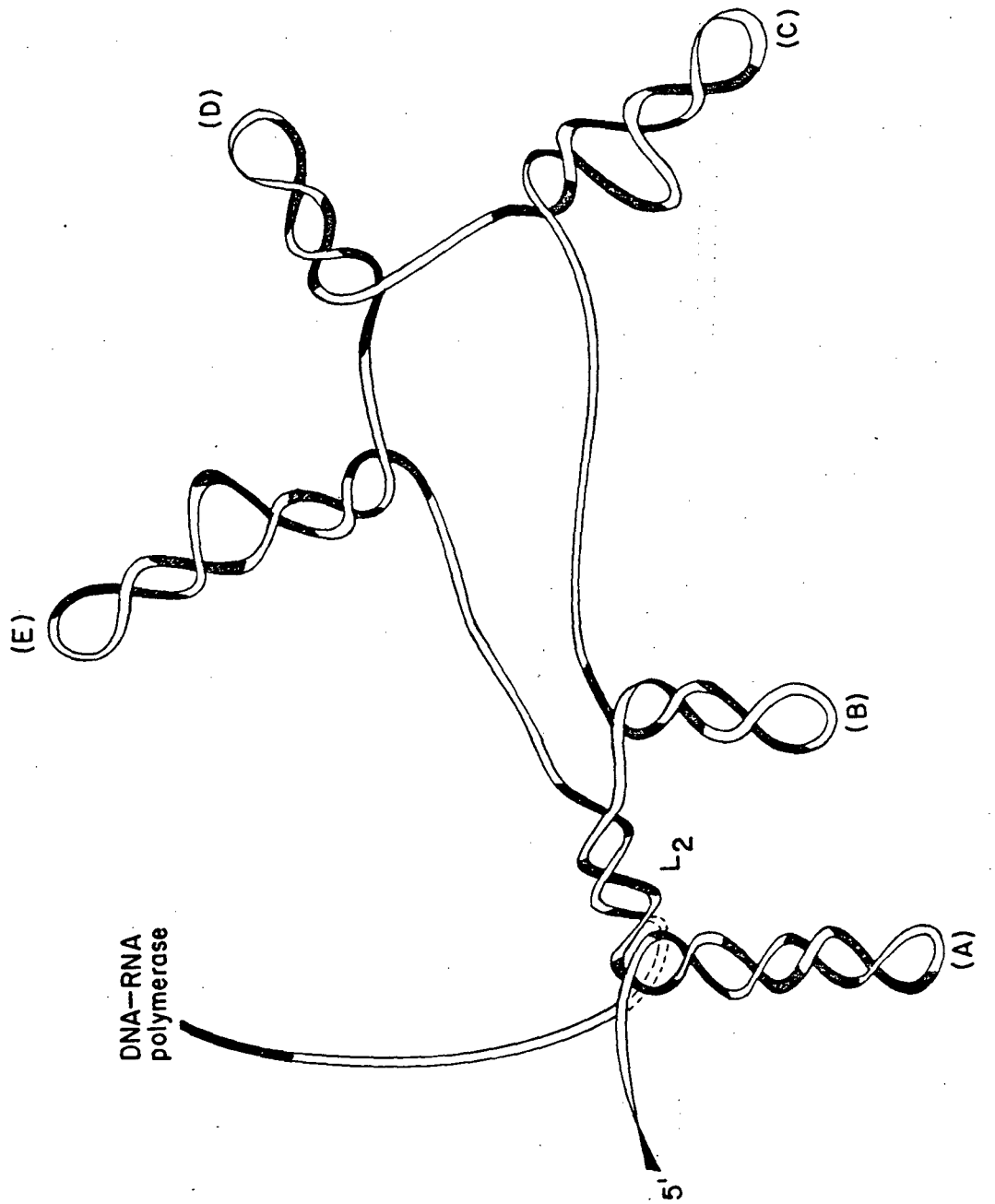


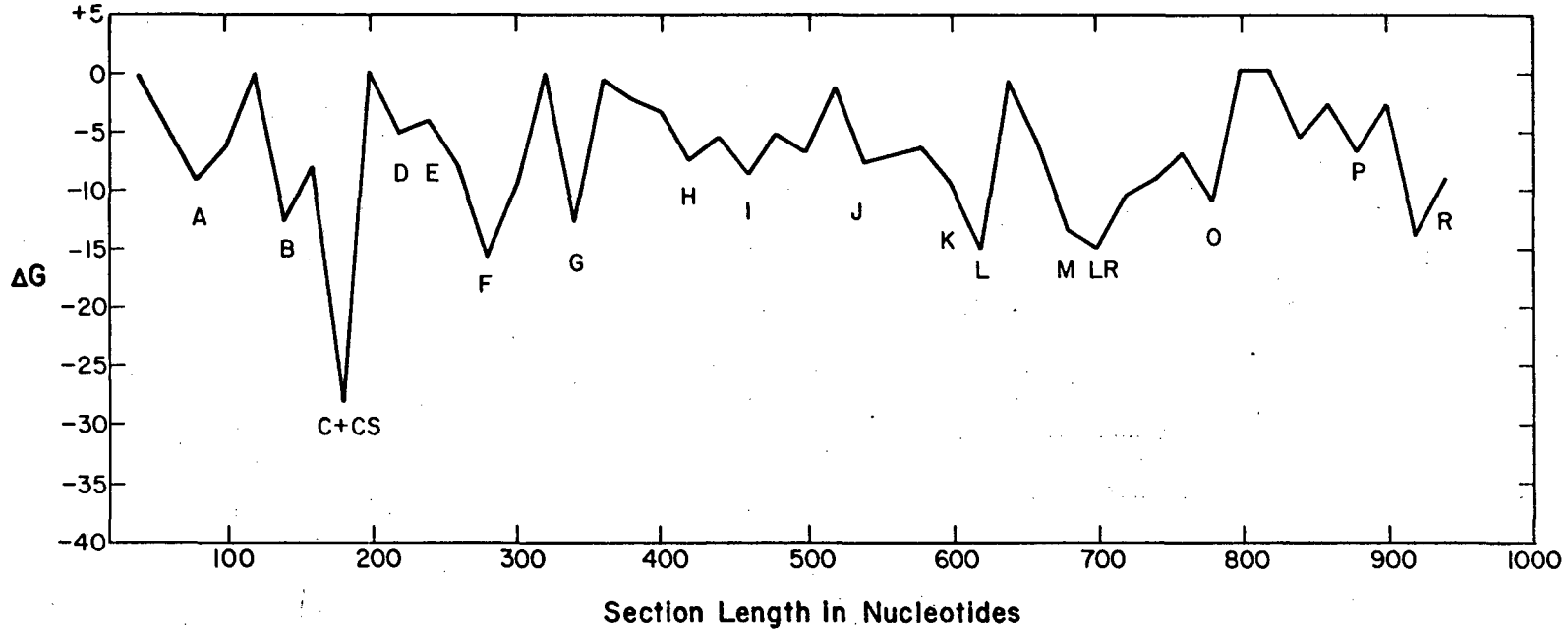
Figure 1



XBL 812-8040

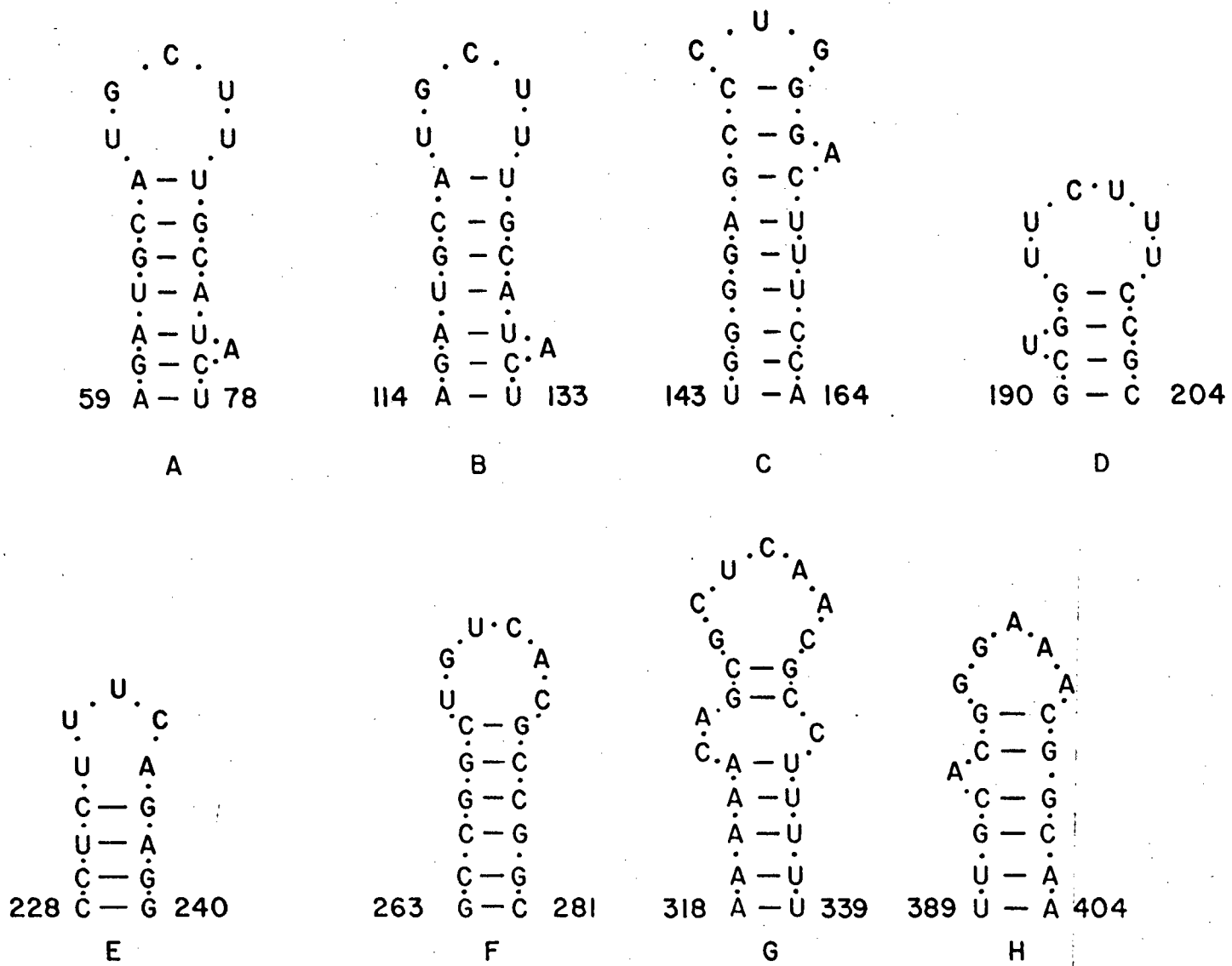
(d)

Figure 1



XBL 812-8044

Figure 2



XBL 812-8042

Figure 3

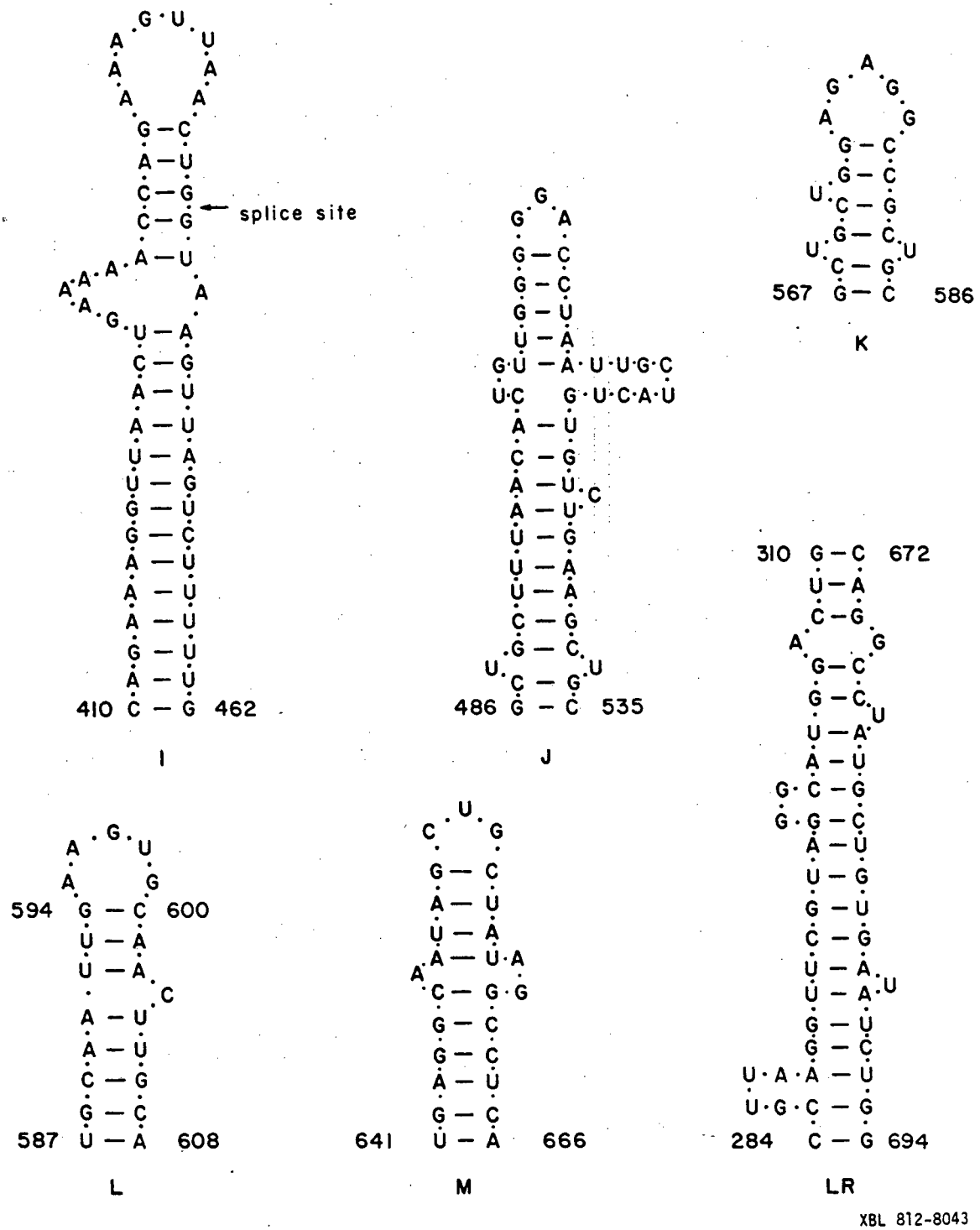

splice site
410 C — G 462
I
486 G — C 535
J
567 G — C 586
K
310 G — C 672
LR
284 C — G 694
594 G — C 600
L
587 U — A 608
641 U — A 666
M
XBL 812-8043

Figure 3

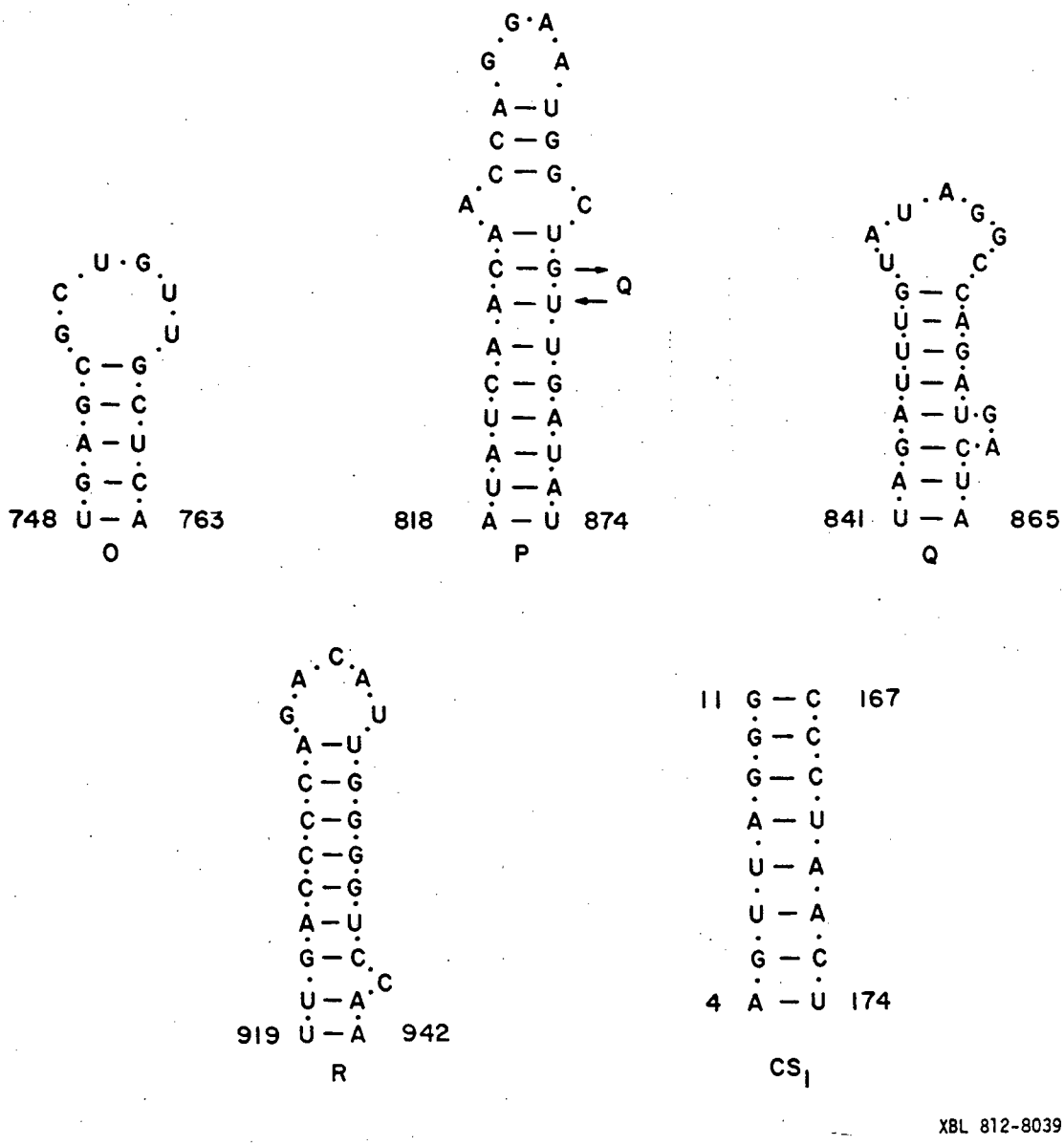

748 U — A 763
O
818 A — U 874
P
841 U — A 865
Q
919 U — A 942
R
11 G — C 167
4 A — U 174
CS1
XBL 812-8039

Figure 3

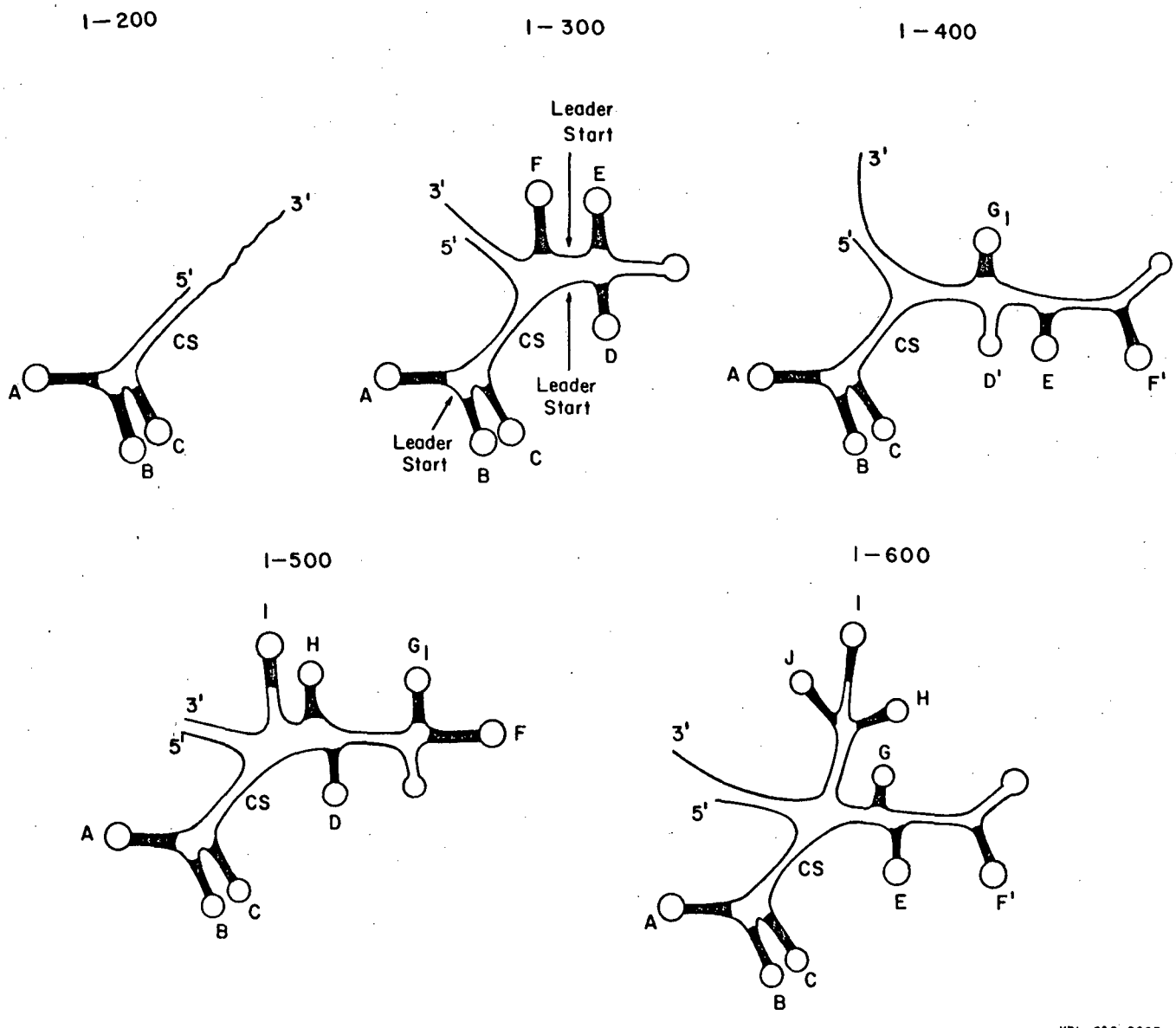
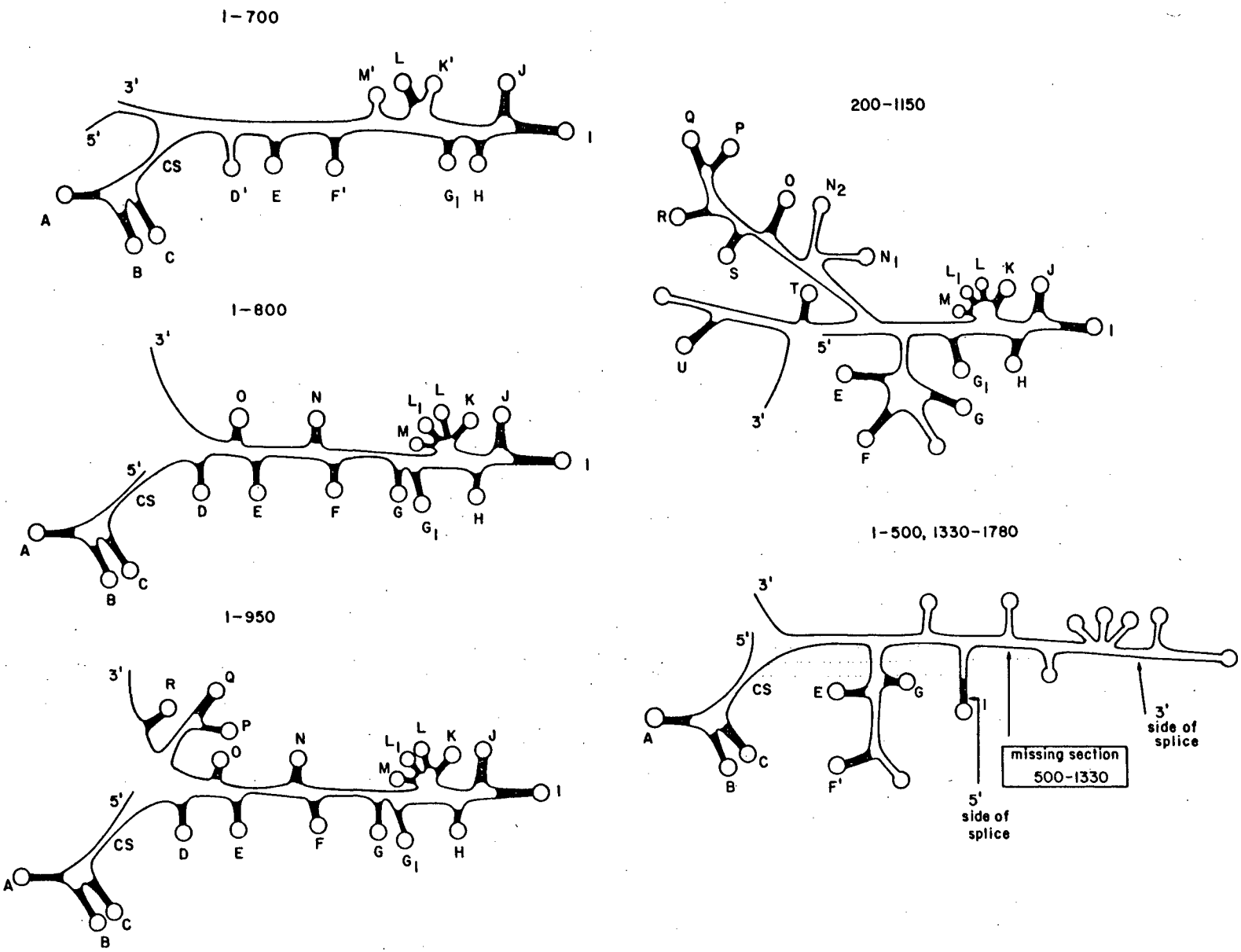

I–200
I–300
I–400
I–500
I–600
Leader Start
Leader Start
Leader Start
3'
5'
CS
A
B
C
D
E
F
G
H
I
J
D'
E
F'
XBL 812-8035

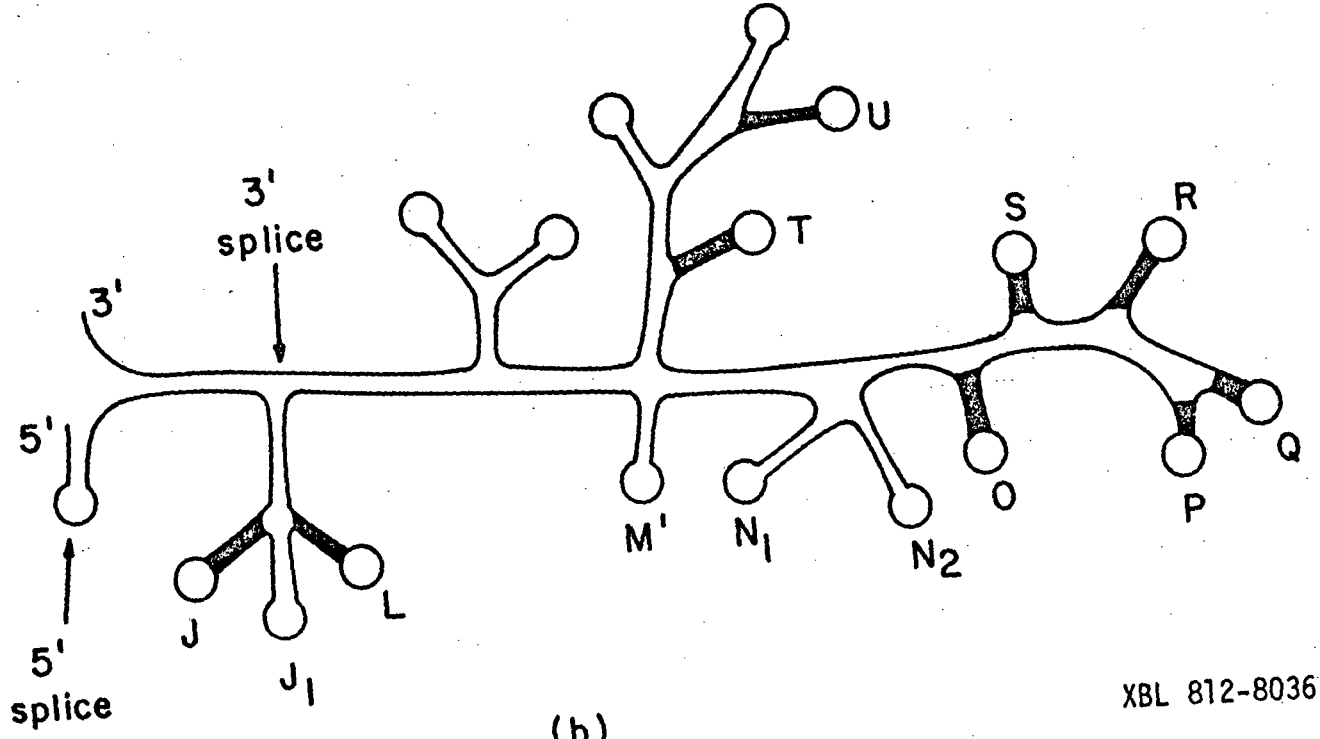
Figure 4



XBL 812-8045

Figure 4

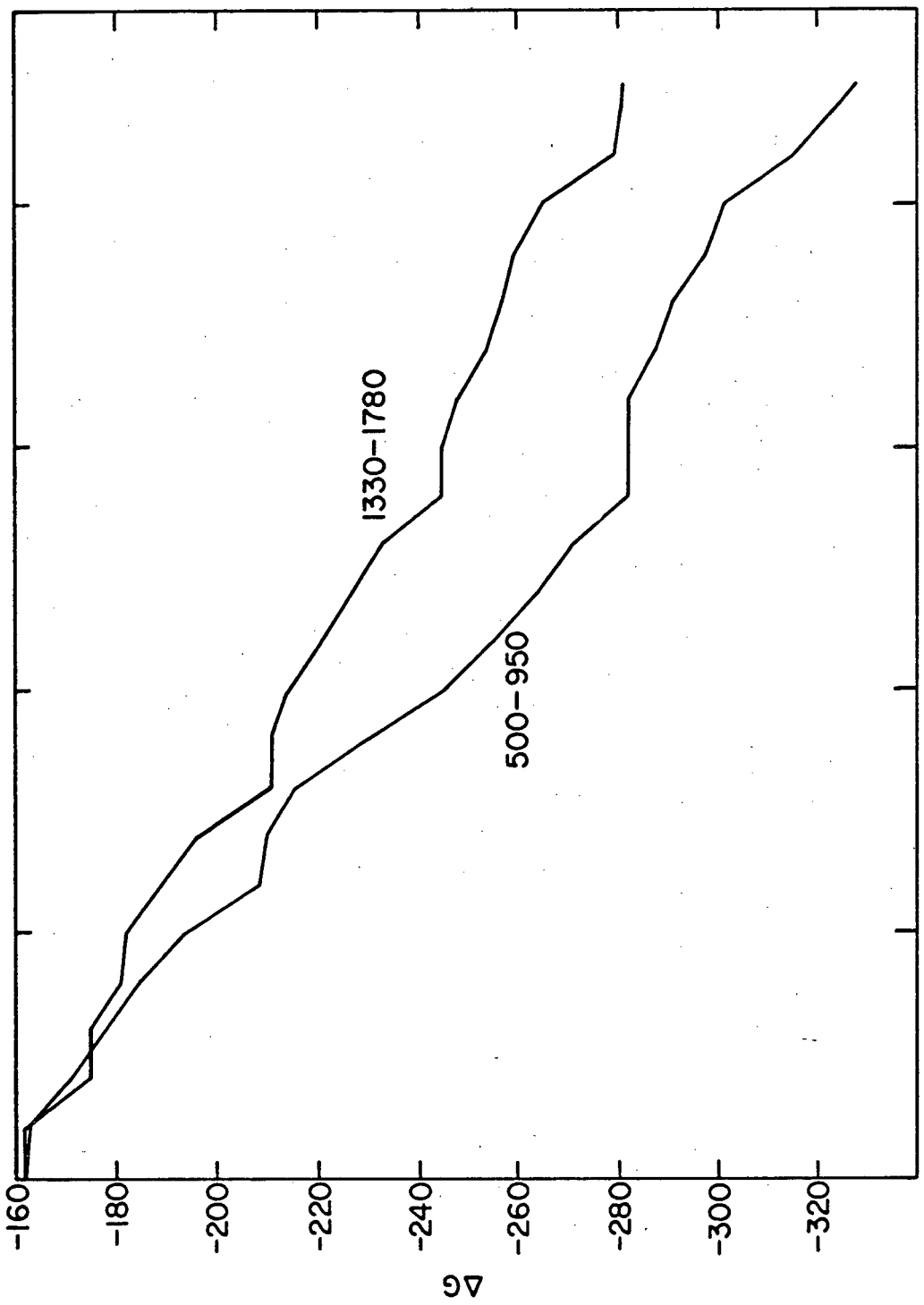
440-1390 intron



(b)

XBL 812-8036

Figure 4b



XBL 812-8038

Figure 5

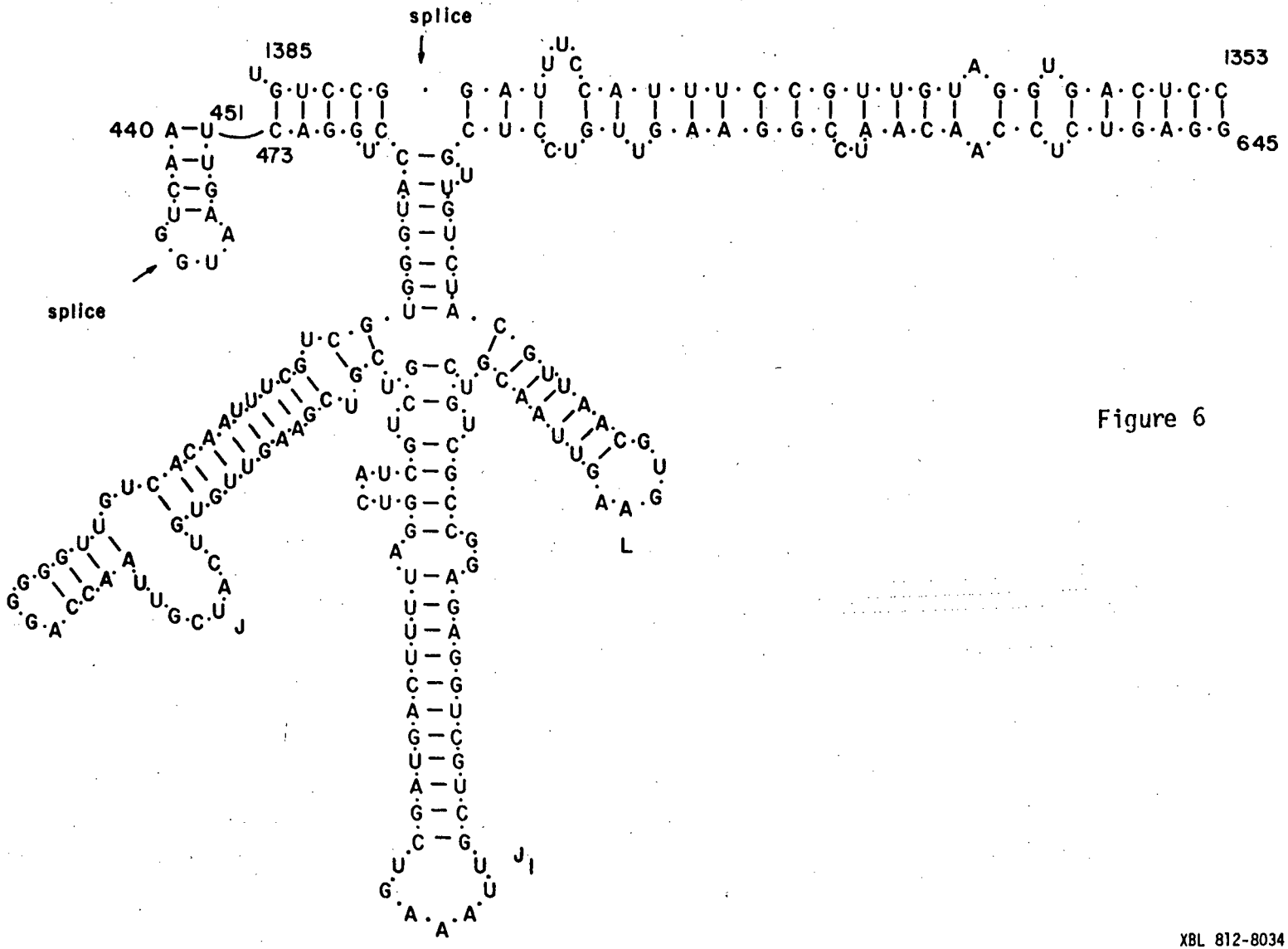


Figure 6

XBL 812-8034

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720