Volume 10 Number 1 1982 Nucleic Acids Research

Pattern recognition in nucleic acid sequences. II. An efficient method for finding locally stable secondary structures

Minoru I.Kanehisa and Walter B.Goad

Theoretical Biology and Biophysics Group, University of California, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received 15 September 1981

ABSTRACT

We present a method for calculating all possible single hairpin loop secondary structures in a nucleic acid sequence by the order of N^2 operations where N is the total number of bases. Each structure may contain any number of bulges and internal loops. Most natural sequences are found to be indistinguishable from random sequences in the potential of forming secondary structures, which is defined by the frequency of possible secondary structures calculated by the method. There is a strong correlation between the higher G+C content and the higher structure forming potential. Interestingly, the removal of intervening sequences in mRNAs is almost always accompanied by an increase in the G+C content, which may suggest an involvement of structural stabilization in the mRNA maturation.

INTRODUCTION

As a working hypothesis it is generally assumed that base-paired segments form the primary motifs in functional three-dimensional RNA structure. Further, it is natural to think of such motifs as potential elements of biological recognition and function. These may in fact function in conditions in which contacts with other molecules or with other regions of the same molecule contribute to the thermodynamic stability of a particular structure. But to locate candidate motifs from sequence informantion we can start by selecting from the very many different possible foldings of a typical RNA molecule those with good inherent thermodynamic stability as judged by secondary structure alone.

This has been the strategy, implicit or explicit, of many authors. To briefly make it explicit here, we look to the one RNA structure known in three-dimensional detail, yeast phenylananine tRNA,[1] where four motifs or the four stems together realize the cloverleaf model of secondary structure. These are defined by and stabilized by the set of specific base pairing and stacking interactions of the double-helical regions. But in reckoning the inherent thermodynamic stability of each double-helical region, account also

needs to be taken of their ends. A consistent, and the customary, procedure is to attribute a free energy to the unpaired regions at either a hairpin end or in the interhelix regions that constitute various types of loops as if they were free of any specific interactions other than at the helix ends. Taken together these idealized helical and loop free energies constitute the measure of inherent stability of secondary structural motifs.

In the phenylalanine tRNA structure we also see the basic cloverleaf structure folded so as to bring many atoms and atomic groups belonging to different motifs into proximity, indicating a rich inventory of intramolecular interactions that are no part of the helical and loop free energies. These are the tertiary interactions; the hydrogen bonds among them have been tabulated by Rich,[1] but there are many other types of interactions as well. At present this must be considered a background which the inherent stability of secondary structural motifs operates as an important determinant of overall structure.

Consensus estimates of free energies of base pairing and stacking and of various kinds of loops, have been given by Tinoco et al.[2-4] These are based on measurements by a number of investigators using model oligomers as filled out by plausible interpolation and extrapolation. Salser[5] has interpolated a few more entries to provide a complete and compact formulation that we will use. The method we present, however, does not depend on this specific choice; it can easily accomodate an expanded description--for example the loop free energy closed by a GU pair might be specified.

A number of algorithms have been given for the combinatorial problem of calculating for an RNA of given sequence that topology of paired segments which minimizes the total free energy of secondary structural features. The most efficient of these, described recently by Nussinov and Jacobson[6] and by Zuker and Stiegler[7] requires of the order of N^3 operations for a molecule of N bases. Efficiency is achieved by attending at each point in the calculation to only those structures which remain candidates for extension into the final optimal structure, others being in effect dropped as soon as there is, however marginally, a better one. The strategy is similar in spirit to that of the Needleman-Wunsch-Sellers algorithm;[8] in the computer literature it would be called dynamic programming.

However, there are uncertainties in the free energies assigned to various secondary structural features, and since it is usually the case that the optimal structure is merely the best of many closely competing ones, the true optimal secondary structure is usally uncertain. Further, tertiary inter-

actions and/or interactions with other molecules may well stabilize a particular secondary structure over a more optimal one. Less efficient algorithms, including those of Pipas and McMahon[9] and Studnicka et al.,[10] offer a means of generating a range of competing structures for further consideration.

The alternative explored here is to identify all locally stable secondary structures. A locally stable secondary structure is a sufficiently stable connection of double helices in a localized region of the sequence, bounded by one hairpin loop and possibly containing bulges and internal loops. It is a potential structural motif in the sense mentioned above. We expect this repertoire of all locally stable secondary structures to include, among many others, those motifs present in the true overall structure. We present an algorithm that accomplishes this in the order of N^2 operations. One use of the repertoire is in ascertaining whether two sequences have a potential for displaying the same motif. The two sequences might, for example, be known to exercise a particular biological function, and the question is whether there is a secondary structural feature that might mediate it. This has been done for 5S and 16S rRNAs[11,12] using the repertoire of all possible double helical regions, which can contain a huge number of possibilities even if one limits the minimum length of a double helix. On the other hand, our repertoire provides a quite manageable number of possibilities for a molecule of up to 10^4 or so bases. In this paper we present another use of the method; the statistics of locally stable secondary structures in known sequences as compared with random sequences.

METHODS

A. Free Energy Calculation Rules

The calculation of free energy is based on the modified Tinoco rule[3] with the free energy data compiled by Salser.[5]

(A1) A helical (base paired) region consists of continuous GC, AU, and GU pairs. It must have at least two GC/AU pairs, and GU pairs cannot occur at either end of a helical region.

(A2) The stacking free energy is assigned to each doublet of base pairs in a helical region.

(A3) A hairpin loop must have at least three bases. Its free energy depends on the GC or AU pair closing the loop.

(A4) The free energy of an internal loop also depends on whether it is closed by two GC pairs, by one GC and one AU pairs, or by two AU pairs.

(A5) The free energy of a bulge loop is corrected by the stacking free energy of two base pairs across the bulge.

B. Connection of Helical Regions by Induction

The Needleman-Wunsch-Sellers algorithm[8] for homology search finds the alignment of two sequences for which an additive measure of their similarity is maximal. If we substitute for that measure the negative free energy of base pairing and have one sequence run in the 5' to 3' direction and the other run oppositely, the same procedure can be used to find an arrangement of the dublex molecule for which there is, overall, maximal stability of the base paired regions. Waterman and Smith[13] used this correspondence to develop an algorithm for RNA secondary structure calculation.

It is helpful to follow the computational process in terms of paths through a matrix as illustrated in Fig. 1. The sequence is written as the first row and, in inverted order, as the first column. A diagonal line seg-

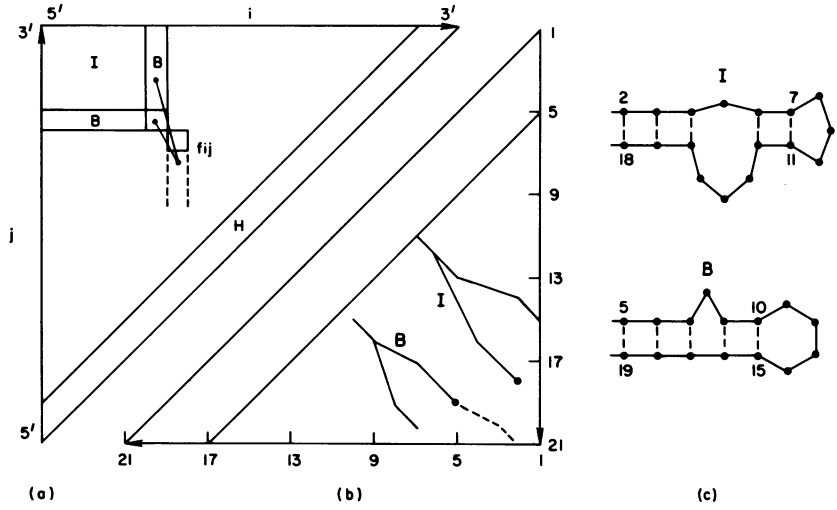


Figure 1. (a) In calculating the free energy f_{ij}, Waterman and Smith's method examines all possible connections of bulge loops (region B) and internal loops (region I) as well as helix extension. Instead, our method keeps track of the positions that give the minimum free energy loop connections, and they are updated at every iteration. Waterman and Smith's algorithm performs the iteration up to the minimum length of a hairpin loop (region H). (b) Our method performs the iteration in the opposite direction making hairpin loops first. A diagonal path corresponds to a helix extension, and others to bulge B or internal I loop formations depending on the increment of base numbers. A locally stable secondary structure is found by the traceback procedure. (c) Two locally stable secondary structures containing bulge B and internal I loops corresponding to the paths leading to the dots in (b).

ment connecting base pairs (i-1, j-1) and (i,j) indicates that they belong to a helical region, and any other segment indicates a loop of some kind. Advancing from left to right and top to bottom, at each possible base pair (i,j) the Waterman-Smith algorithm (Fig. 1(a)) computes the free energy f_{ij} of the optimal structure that can be formed by associating the first i bases and the last j bases of the sequence. This is done at each base pair (i,j) by induction, choosing the structure of lowest free energy from among those that can be built by extending the optimal structures already found for $i' < i$ and $j' < j$. At base pair (i,j) three kinds of connections are possible:

(B1) Extension of a helical region; connected to (i-1, j-1).

(B2) Formation of a bulge loop; connected to (i-1, ℓ) where ℓ ranges from 1 to j-2, or connected to (k, j-1) where k ranges from 1 to i-2 (region B in Fig. 1(a)).

(B3) Formation of an internal loop; connected to (k, ℓ) where k ranges from 1 to i-2 and ℓ ranges from 1 to j-2 (region I in Fig. 1(a)).

Note that the connection of helical regions by a loop (B2) or (B3) may be made with an inner base pair of a helical region as well as an end base pair. The best connection is that minimizing the free energy f_{ij}:

$$f_{ij} = \min \{ f_{i-1,j-1} + \alpha_{ij} , \min_{k>0} (f_{i-k-1,j-1} + \beta_k) , \min_{k>0} (f_{i-1,j-k-1} + \beta_k) , \min_{k,\ell>0} (f_{i-k-1,j-\ell-1} + \gamma_{k+\ell}) \} \quad (1)$$

Here $\alpha_{i,j}$ is the stacking free energy of base pairs (i-1, j-1) and (i,j), and β_k and γ_k are the free energies of a bulge and an internal loop, respectively, of length k. The free energy f_{ij} is defined only at base pairs, which can formally be managed by entering plus infinity (or a large positive value) when i and j do not base pair.

In Fig. 1(a) matrix elements lying along the principal antidiagonal represent the same base in both horizontal and vertical sequences. If we assume that a hairpin loop must contain at least three looped-out bases due to steric restrictions the last four antidiagonals define a sector labelled H in Fig. 1(a) in which base pairing is not permitted. When the induction by Eq. (1) is complete the optimal single hairpin loop structure is found by considering the possibility of each base pair closing a hairpin loop:

$$F = \min_{(i,j)} (f_{i,j} + \delta_{N-i-j}) \quad (2)$$

where δ_k is the free energy of a hairpin loop of length k and F is the optimal

overall free energy. Note that the algorithm performs of the order of (i-1) x (j-1) operations to calculate the free energy f_{ij}, thus, the entire procedure requires the order of N^3 operations.

We modify the Waterman-Smith algorithm in three principal respects: (i) Another level of induction is introduced on possible loop ends (see below). This reduces the computational order of the algorithm from N^3 to N^2. (ii) To find and analyze all single hairpin loop structures, we start the calculation at the hairpin antidiagonal and advance toward the origin (see Fig. 1(b)). (iii) We allow all loop free energies to depend on the specific base pairs closing the loop.

C. Inductive Calculation of Optimal Loop Ends

Just as the Waterman-Smith algorithm replaces a cumbersome enumeration of all possible foldings of helical regions by a simple induction, so in their algorithm defined by Eq. (1) we can replace the searches over all possible loop connections by an inductive construction of the single best loop connection for each type of loop. If the loop free energies β_k and γ_k are linear functions of length k, Eq. (1) can be reduced to:

$$f_{i,j} = \min (f_{i-1,j-1} + \alpha_{i,j} , f_{BI(i,j),j-1} + \beta_{i-BI(i,j)-1} , f_{i-1,BJ(i,j)} + \beta_{j-BJ(i,j)-1} , f_{I(i,j),J(i,j)} + \gamma_{i-I(i,j)+j-J(i,j)-2}) \quad (3)$$

Here the base pairs (BI, j-1), (i-1, BJ), and (I,J) are the best possible loop ends in the horizontal bulge region B, the vertical bulge region B, and the inner loop region I, respectively, in Fig. 1(a).

Consider bulge loops in which bases from the vertical copy of the sequence are looped out. Let BJ(i,j) be the row index of the base pair which closes a possible bulge loop extending from (i-1, BJ) to (i,j). To make it the optimal loop, we determine BJ(i,j) by the induction:

$$BJ(i,j+1) = \begin{cases} j-1 & \text{if } f_{i-1,j-1} < 0 \text{ and } f_{i-1,j-1} + \beta_1 \leq f_{i-1,BJ(i,j)} + \beta_{j-BJ(i,j)} \\ BJ(i,j) & \text{otherwise} \end{cases} \quad (4)$$

Namely, if (i-1, j-1) is a base pair the bulge connection from (i-1, BJ) to (i, j+1) and that from (i-1, j-1) to (i, j+1) are compared (see Fig. 1(a)) considering the free energy values $f_{i-1,BJ}$ and $f_{i-1,j-1}$ and the loop lengths. The initial values of BJ are zero, which represents there are no possible bulge forming base pairs. Then, providing β_k is proportional to k, of all possible vertical bulge loop connections which end with the base pair at (i,j)

the particular connection starting at (i-1, BJ) has the lowest free energy. This is so because whenever, as the induction proceeds, a new loop starting point is picked up by satisfying the first condition of Eq. (4), this condition holds for any subsequent extension of a loop, that is,

$$\text{if} \quad f_{i-1,j-1} + \beta_1 \leq f_{i-1,BJ(i,j)} + \beta_{j-BJ(i,j)}$$

$$\text{then} \quad f_{i-1,j-1} + \beta_{1+k} \leq f_{i-1,BJ(i,j)} + \beta_{j+k-BJ(i,j)} \quad \text{for} \quad k > 0$$

Similar provisions must be made for horizontal bulge loops and internal loops.

Figure 2 shows the actual loop free energies, experimental values and theoretical extrapolations, according to Salser.[5] Since the use of Eqs. (3) and (4) is to approximate the curves B and I's of Fig. 2 by the initial slopes, it gives satisfactory results for comparisons of loops of six bases or less, but it tends to favor shorter loops when comparing with a longer loop. It is possible to incorporate the nonlinear dependence into our algorithm by using Eq. (1) to a limited length say, up to eight bases, and making linear approx-

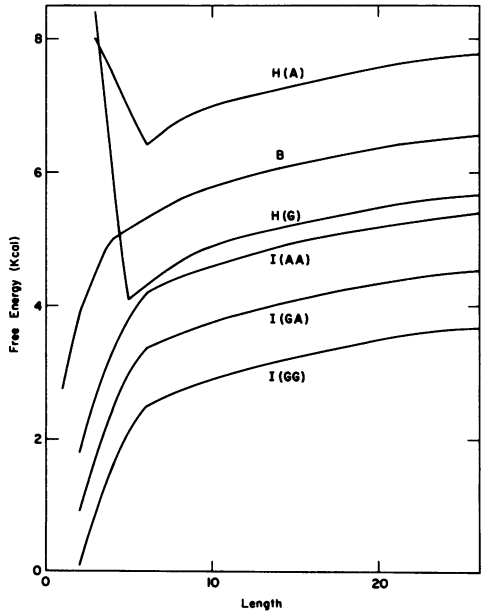


Figure 2. The loop free energies versus the loop length according to Salser: H(A), H(G), hairpin loops closed by an AU and a GC pair; B, bulge loop; I(AA), I(GA), I(GG), internal loops closed by two AU, one AU and one GC, and two GC pairs.

imations by using Eq. (3) beyond that.

In order to correctly incorporate the end base pair dependence of loop free energies, the possible base pair positions for loop ends must be remembered separately for each type. According to the free energy calculation rules (A4) and (A5), GC/CG and AU/UA pairs are distinguished for internal loops, and all GC, CG, AU, and UA pairs are distinguished for bulge loops. Then, Eq. (3) must be modified so that the comparison is made for twelve possible paths: helix extension, two types of internal loop formation, four types of bulge loop formation on each side, and formation of a new hairpin loop. The last possibility must be examined because the opening up of base pairs adjacent to a hairpin loop can give an energetically favorable structure. In addition to the free energy calculation rules (A1)-(A5) we assume that loops must be closed by helical regions; i.e., a single base pair cannot exist alone. Therefore, the possibilities of loop formation are examined only when there is another base pair next to the one closing the loop.

D. Locally Stable Secondary Structures

The algorithm locates all possible hairpin loops and extends helical regions by induction. As illustrated in Fig. 1(b) the paths representing optimal connections of helical regions frequently branch. A branch is terminated when the free energy f_{ij} becomes positive. However, owing to the small positive free energy associated with long loops, a branch may grow until it reaches an end of the sequence. Accordingly, we have introduced the maximum permitted length that can be set as a parameter of the calculation

When the induction is complete all locally optimal folded structures having exactly one hairpin loop have been constructed. Now each path is traced back from its lower right-hand end in Fig. 1(b), and when two or more branches lead to the same trunk only the one of lowest free energy is kept. The free energy is recalculated during the trace back procedure. A branch may well begin with a segment whose contribution to the free energy is positive as illustrated by the broken line in Fig. 1(b); such segments are removed. The structure corresponding to each remaining path represents what we have called a locally stable secondary structure.

E. Programming Notes

As noted above we have defined a parameter which limits the total number of bases in any structure. In addition, the maximum lengths of hairpin, internal, and bulge loops are also set and may be changed by the user. The program prints out all locally stable secondary structures with free energies

below a threshold set by the user. An example of output is shown below:

```
#86(34) -64.0
5502        5512        5522        5532
AGCGG CGCAUUAA GCGC GGCGGGUGUGGUGGUUACGCG
::::: :::  ::  :::: :::..::::.:.:::.:::: C
UCGCCCGCG  AUCCCGCGACCGUUCACAUCGCCAGUGCGA
5573        5563        5553        5543
```

This is the lowest free energy structure in DNA phage M13[14] found by the method (86th found structure), and it has the total free energy of -64 kcal and 34 base pairs formed. In this simple representation the hairpin loop is shown at the right end and it is not difficult to distinguish between a bulge loop and an internal loop.

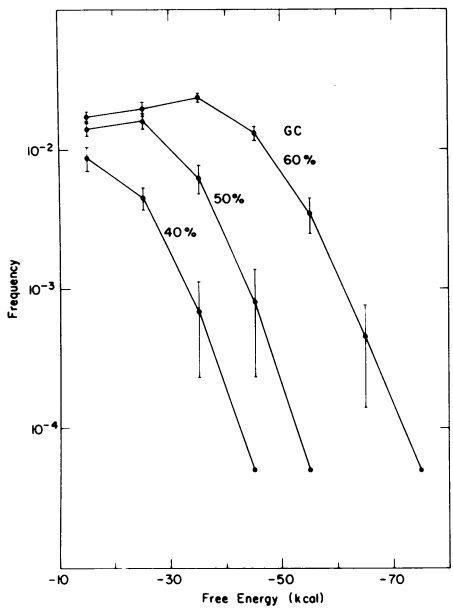
Our program is written in Fortran, and it may be obtained upon request. It requires the storage area of about $N^2/2$ to search the entire region of an N base sequence. The maximum length of N in one search is 500 because of the limitation of the maximum array size (131 K) in CDC Fortran IV. It takes less than 5 seconds of CPU time on CDC 7600 to search the entire region of 500 bases. For longer sequences the program repeats the procedure by shifting the search region. There is no limit of sequence length that can be analyzed if the search length is limited. It takes about 12 seconds on CDC 7600 to find all locally stable secondary structures of 100 bases or less in a sequence of 5,000 bases.

RESULTS AND DISCUSSION

For a particular nucleic acid sequence, we might look to a plot of the number of locally stable secondary structures as a function of their free energy, as an indication of the structure forming potential of the sequence. To evaluate this notion we have analyzed a sample of random sequences, each of length 4,000 bases. Bases were drawn independently with probabilities corresponding to the desired composition. By limiting the maximum length of a local secondary structure at 100 bases, the observed numbers of structures with free energies in the ranges of -10 to -20, -20 to -30, etc. were counted and normalized by the sequence length. In Fig. 3 the observed frequency for the average of 10 random sequences is plotted against free energy with standard deviations shown by the bars. The three lines represent the variation of the G+C content; the compositions in G and C, and in A and U are assumed to be equal. As is apparent in Fig. 3 the structure forming potential is higher for higher G+C content. If a GC pair and an AU pair had been treated equally as in the search for dyad symmetries, the case of 40% G+C (60% A+U) would have had a higher frequency distribution than that of 50% G+C. Although this

result may not be surprising considering the stability of a GC pair compared to an AU pair, it also suggests some risk of just counting the length of a double helical region in the base pairing matrix[2] where GC and AU pairs are treated alike.

Figure 4 shows the structure forming potential of some of the known sequences that we have analyzed. RNA phage MS2 (3,569 bases, 52.1% G+C),[15] E. coli ribosomal RNAs 23S (2,904 bases, 53.4% G+C)[16] and 16S (1,542 bases, 54.4% G+C)[17] have high potential, but they also have relatively high G+C contents. In comparison, single-stranded DNA phages φx174 (5,386 bases, 44.8% G+C),[18] G4 (5,577 bases, 45.7% G+C)[19] and M13 (6,407 bases, 40.8% G+C)[14] have relatively low G+C contents and low structure forming potential (of course, according to the free energy values of ribonucleotides). In fact, it appears that most



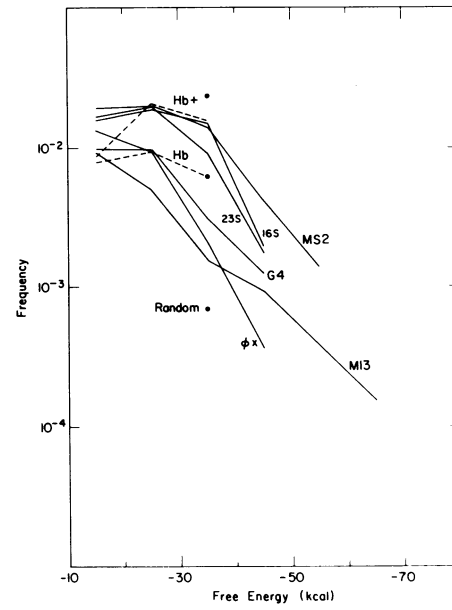


Figure 3. (Left) Frequency per base of locally stable secondary structures up to 100 bases long versus the free energy in random sequences. The average and standard deviation of 10 random sequences are shown for each of the three G+C contents. The maximum lengths of hairpin, internal, and bulge loops were limited to 20, 10, and 5, respectively, in this and next calculations.
Figure 4. (Right) Frequency per base of locally stable secondary structures up to 100 bases long in natural sequences. The solid circles represent the results of random sequences shown in Fig. 3. The broken lines represent the variation of frequency before (Hb) and after (Hb+) splicing of a rabbit globin mRNA.

real sequences are not distinguishable from random sequences of the same G+C content in their potential to form secondary structures. An exception is MS2;[8] the observed frequencies in the free energy ranges of -10 to -20, -20 to -30, -30 to -40, -40 to -50, and -50 to -60 were, respectively, 67, 70, 50, 15, and 5, while they were 50±4, 67±5, 36±7, 7±3, and 1±1 in 10 randomized sequences of the same composition. Thus, the lowest free energy structure in phage M13 around the origin of replication[14] is not unexpected in a random sequence of the same length and composition.

Our conclusion is that the secondary structure forming potential of most existing RNA molecules is indistinguishable from that of random sequences. A similar conlusion has been reached by Gralla and DeLisi.[20] It does not imply, however, that the distribution of hairpin structures within an existing molecule is in any sense random. The specific location of a hairpin structure may well be preserved during evolution. According to our method, yeast phenylalanine tRNA (76 bases, 54% G+C)[1] contains four relatively stable (less than -5 kcal) local secondary structures, and three of them are in fact used in the three loops of the cloverleaf. (The acceptor stem is found in the extension of a structure containing the anticodon loop.) By randomizing this sequence we counted the frequency and examined the location of locally stable secondary structures. We observed average of 3.5 ± 1.8 structures in 10 randomized sequences, but in no sequence were as many as three structures topologically realizable at the same time. Therefore, although the structure forming potential of tRNA may not be distinguishable from random sequences, the actual sequence must have preserved its specific positions of hairpin loop structures against severe topological restrictions. In seeking relationships between sequences, this points to the value of comparing their specific repertoires of locally stable secondary structures as mentioned in the Introduction.

We have found that the best index of structure forming potential is simply G+C content. This suggests that if secondary structure (or its absence) plays a role either in the removal of intervening sequences or in some aspect of gene regulation for which splicing is also a factor, then this may be reflected in differences in local G+C content as between intervening sequences and the rest of the message. Table 1 summarizes the variation of G+C contents before and after splicing of a number of eukaryotic messenger RNAs. It is interesting that the G+C content always increases, in one case more than 10%, by removal of intervening sequences from the initial transcript, except the two which already have high G+C contents. This indicates that intervening sequences are relatively A+T rich as compared with the rest of the message.

TABLE 1. Increase of G+C content after splicing of eukaryotic messenger and transfer RNAs

	Total bases	IVS† bases	G+C content(%) before	after	change	Ref.
Human α_2 globin gene mRNA	833	257(2)	66.7	64.2	-2.5	21
Human β globin gene mRNA	1607	980(2)	40.2	51.5	11.3	22
Human δ globin gene mRNA	1641	1017(2)	40.9	50.5	9.6	23
Human ε globin gene mRNA	1594	977(2)	42.1	51.7	9.6	24
Human $^{A}\gamma$ globin gene mRNA	1573	988(2)	46.9	50.9	4.0	25
Human $^{G}\gamma$ globin gene mRNA	1590	1008(2)	46.9	51.2	4.3	25
Mouse α globin gene mRNA	819	256(2)	56.8	56.7	-0.1	26
Mouse β globin gene mRNA	1389	762(2)	45.7	50.7	5.0	27
Rabbit β_1 globin mRNA	1289	699(2)	42.9	50.5	7.6	28
Rat preproinsulin gene I mRNA	563	119(1)	56.8	57.9	1.0	29
Rat preproinsulin gene II mRNA	1062	618(2)	54.7	58.8	4.1	29
Chicken ovalbumin gene fragment	1698	1232(3)	38.8	44.2	5.5	30
Human Ig κ chain v region	476	125(1)	48.9	52.4	3.5	31
Human Ig κ chain v region	452	104(1)	51.3	54.9	3.6	32
Mouse Ig γ^1 chain c region	1546	575(3)	51.9	54.0	2.1	33
Mouse Ig γ^{2b} chain c region	1542	535(3)	51.5	52.4	0.9	34
Mouse Ig λ_1 chain v region	441	93(1)	49.0	49.7	0.7	35
Silkworm fibroin gene fragment	1473	970(1)	38.6	43.5	4.9	36
Yeast actin gene	1433	309(1)	41.3	43.9	2.6	37
Yeast cytochrome oxidase gene	9979	8440(7)	25.2	31.1	5.9	38
Yeast tyrosine tRNA	89	14(1)	51.7	56.0	4.3	39
Yeast phenylalanine tRNA	92	19(1)	48.9	53.4	4.5	39

† Number of intervening sequences in parentheses.

Accordingly, the structure forming potential can also be significantly increased after splicing; the broken lines in Fig. 4 show this increase explicitly for rabbit β_1 globin messenger RNA.[28]

We have also examined the splicing in SV40, polyoma, BKV, and andenovirus, but these animal viral messenger RNAs show alomost constant G+C content; variation of less than 1% after splicing. In the secondary structure

model for splicing in yeast transfer RNAs,[39] whose G+C content increases are also shown in Table 1, the specific location of base-paired and loop regions are the recognition factors by the cutting enzyme. The tendency of forming double helices in G+C rich regions in turn suggests the tendency of forming open loops in A+T rich regions. Therefore, the clustering of G+C rich and A+T rich regions can exhibit a contrast of secondary structures, which may help structural recognition by other molecules.

ACKNOWLEDGMENTS

We thank Drs. Michael Waterman and George Bell for helpful discussions. This work was performed under the auspices of the U.S. Department of Energy.

REFERENCES

1. Rich, A. and RajBhandary, U.L. (1976) Annu. Rev. Biochem 45, 805-860.
2. Tinoco, I., Jr., Uhlenbeck, O.C., and Levine, M.D. (1971) Nature 230, 362-367.
3. 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.
4. Borer, P.N., Dengler, B. and Tinoco, I., Jr. (1974) J. Mol. Biol. 86 843-853.
5. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 62, 985-1002.
6. Nussinov, R. and Jacobson, A.B. (1980) Proc. Nat. Acad. Sci. USA 77, 6309-6313.
7. Zuker, M. and Stiegler, P. (1981) Nucl. Acids Res. 9, 133-148.
8. Goad, W.B. and Kanehisa, M.I., the preceding paper.
9. Pipas, J.M. and McMahon, J.E. (1975) Proc. Nat. Acad. Sci. USA 72, 2017-2021.
10. Studnicka, G.M. Rahn, G.M., Cummings, I.W., and Salser, W.A. (1978) Nucl. Acids Res. 5, 3365-3387.
11. Fox, G.E. and Woese, C.R. (1975) Nature 256, 505-507.
12. Noller, H.F. and Woese, C.R. (1981) Science 212, 403-411.
13. Waterman, M.S. and Smith, T.F. (1978) Math Biosci. 42, 257-266.
14. van Wezenbeek, P.M.G.F. et al. (1980) Gene 11, 129-148.
15. Fiers, W. et al. (1976) Nautre 260, 500-507.
16. Brosius, J. et al. (1980) Proc. Nat Acad. Sci. USA 77, 201-204.
17. Brosius, J. et al. (1978) Proc. Nat. Acad. Sci. USA 75, 4801-4805.
18. Sanger, F. et al. (1978) J. Mol. Biol. 125, 225-246.
19. Godson, G.N. et al. (1978) Nature 276, 236-247.
20. Gralla, J. and DeLisi, C. (1974) Nature 248, 330-332.
21. Proudfoot, N.J. and Maniatis, T. (1980) Cell 21, 537-544.
22. Lawn, R.M. et al. (1980) Cell 21, 647-651.
23. Spritz, R.A. et al. (1980) Cell 21, 639-646.
24. Baralle, F.E. et al. (1980) Cell 21, 621-626.
25. Slightom, J.L. et al. (1980) Cell 21, 627-638.
26. Nishioka, Y. and Leder P. (1979) Cell 18, 875-882.
27. Konkel, D.A. et al. (1978) Cell 15, 1125-1132.
28. van Ooyen, A. et al. (1979) Science 206, 337-344.
29. Lomedico, P. et al (1979) Cell 18, 545-558.
30. Robertson, M.A. et al. (1979) Nature 278, 370-372.
31. Bentley, D.L. and Rabbitts, T.H. (1980) Nature 288, 730-733.

32. Matthyssens, G. and Rabbitts, T.H. (1980) Proc. Nat. Acad. Sci. USA 77, 6561-6565.
33. Honjo, T. et al. (1979) Cell 18, 559-568.
34. Yamawaki-Kataoka, Y. et al. (1980) Nature 283, 786-789.
35. Bernard, O. et al. (1978) Cell 15, 1133-1144.
36. Tsujimoto, Y. and Suzuki, Y. (1979) Cell 18, 591-600.
37. Ng, R. and Abelson, J. (1980) Proc. Nat. Acad. Sci. USA 77, 3912-3916.
38 Bonitz, S.T. et al. (1980) J. Biol Chem. 255, 11927-11941.
39. Knapp, G. et al. (1978) Cell 14, 221-236.