Volume 12 Number 1 1984 Nucleic Acids Research

Energy directed folding of RNA sequences

P.Hogeweg and B.Hesper

Bioinformatica, Padualaan 8, de Uithof, Utrecht, The Netherlands

Received 16 August 1983

ABSTRACT

A modification of Nussinov's algorithm (1) for (planar) secondary structure generation is described. Our algorithm postpones descisions on matches involving destabiling loops until they prove to be energetically more favourable than more local matches. We present, moreover, an alternative way of representing secondary structures which avoids unwarranted suggestions on higher order neighbourhood, can be automated easily, allows for any amount of annotation of the sequences, makes comparison of alternate foldings easy and is pleasing to the eye. 5S RNA sequences are used to illustrate the methods.

INTRODUCTION

Nussinov et al (1,2,3,4) proposed an efficient algorithm for the generation of maximum match foldings of strings under the constraint of planarity. The algorithm presupposes, however, a monotonic non-decreasing optimality function for growing strings. Nussinov uses her algorithm for energy directed folding of RNA sequences solving the non-monotonicity in the free energy caused by destabilising loops in an ad hoc manner (2,3,4), using the length of non bound adjacent areas.

Zuker and Stiegler (5) recognised the problem of the non monotonicity and proposed a modification of the algorithm in which the problem is solved but at the expense of doubling of storage and time requirements. The algorithm which we describe in this paper can be seen as a reconciliation of both these algorithms; it needs only the original amount of storage and very little more computing time.

Conventions for the representation of secondary structures of RNA are discussed and an alternative representation is proposed.

ENFOLD, ENERGY DIRECTED FOLDING ALGORITHM

Nussinov's algorithm (1) for maximal match folding sets up an n x n matrix (n length of string) henceforward called MATCH. The lower triangular

half of this matrix is used to store the maximal match values of all possible substrings; the upper triangular part is used to store the position within the substring to which its last position of this substring is matched (in case of maximal match folding of this substring (see fig. 1).

For increasing lengths, the algorithm considers all possible substrings of the given length in turn and constructs the maximal match folding for the relevant substring by considering the previosly calculated (and stored) maximal match foldings of its substrings: the algorithm checks all possible matches of the last symbol (i.e. the one added to the substring at this step) to the other symbols and calculates the resulting match value as the sum of the maximal match values of the two substrings formed by inclusion of this match:

$$Mk = MATCH(j,k)+MATCH(k-1,i)$$

for i: first position of substring, j: last position and k: position to which j is matched. ($i < k < j$). Only the case of k=i is not previously calculated as this length was not considered before. Its value is:

$$(MATCH(j-1,i+1) + B(str(i),str(j))$$

when in matrix B the match values of all pairs of symbol values are stored. The maximal match value of the current substring:

$$Max(Mk\ (i < k < j),\ MATCH\ (j-1,\ i))$$

MATCH(j-1, i) being the value of j is not bound to any symbol.
The maximal match folding of the total string is obtained by backtracking through the upper half of the matrix after it is all filled, the maximum match value is stored in MATCH (N,1).

The problem with appying the Nussinov algorithm to the energy directed folding of RNA sequences is that in the RNA case, unlike the case of maximal match folding of arbitrary strings, the inclusion of a new bond does not always increase the match value (here free energy) because the inclusion of such a bond may form a destabilising loop; therefore loops and forks are never formed with the algorithm in its present form, because the inclusion of the initial bond which forms them is always less optimal then its non-inclusion. Whether or not this bond should be included for a maximal free energy solution depends on the extent of stacking 'below' this bond, i.e. how many bonds are formed for successive larger substring-lengths. However, these stackings are calculated only later in the algorithm. Therefore it seems appropriate also to postpone the decision on loop inclusion until the free energy of the whole stacking region exceeds that of the sequence without this destabilising loop.

The postponement of the decision on loop inclusion can be included in the algorithm in the following way: a destabilising loop is only considered if no increase of the match value can be reached by binding position j to any position k, i.e. when the maximal match algorithm would enter 'no match'. In that case no new information is added to MATCH, only an earlier value is stored in MATCH(j,i) and a $\emptyset$ in MATCH(i,j). Thus without loss of information we can use these positions to store the destabilising loop, marking it as 'provisional', i.e. the match value (= free energy) for inclusion of this bond is entered in the lower half of the matrix and the value of k (always i) is entered in the upper half as -i thus indicating the non optimality of this match.

In the calculation of the maximal match of subsequent substrings of increasing length, the negative entries are treated in a special way, to retrieve the maximal free energy one has to backtrack to the first positive value. The algorithm can operate in two different modes:

1. the local stabilisation mode: Destabilising bonds are not piled on top of locally unstable regions, i.e. they are only considered relative to stable substructures.
2. the global optimisation mode: Several destabilising loops can be piled on top of each other if it results in minimization of the final (global) free energy. Thus two small internal loops can occur adjacent to each other although the internal bond is unstable because such a structure can, as a whole, be more stable than a single large loop (cf. energy rules of Salzer (6)).

In the first mode negative marked entries are simply ignored in all cases except for stabilising bonds (stacking, small bulges). In the second case the algorithm considers all preliminary (i.e. negative) entries while backtracking to the first definite (i.e. positive) one, and chooses the entry with minimal free energy. In the second mode the backtracking algorithm to obtain the final folding becomes more complicated: the match value has to be recalculated in order to chose the subfolding relative to which it was formed. So far we mainly used the first mode because we do not like the local instability allowed in the second mode although it produces higher optimality values.

The program is written in PL/I and uses recursive subroutines. Moreover it is integrated in BIOPAT, our programsystem for bioinformatic pattern analysis. If necessary, however, it should be possible to construct a stand-alone version. Its storage requirements depend on the length of sequence as n^2, the time requirements as n^3. A sequence of length 1000

needs 20 minutes CPU on AMDAHL V7b

SECONDARY STRUCTURE REPRESENTATION

The conventional representation of secondary structure of RNA molecules is semi-pictorial: local proximity relationships are only qualitatively conserved and the global spatial relationships not conserved at all, but the representation gives the impression of depicting spatial relationships directly. Disadvantages of such a semipictorial representation are:

1. The human visal system is inclined to give more weight to overall form (which is here of course meaningless) than to local structure which here contains the relevant information. Therefore visual estimates of similarity of secondary structures represented semi-pictorially are often wrong.
2. The representation suggests higher order proximity relations which are meaningless.
3. The comparison of sequences or of subsequences is tedious in this representation because the correspondence of parts cannot easily be established visually. This works against the use of Nussinov's valuable proposal of generating the minimal energy foldings of the successively longer RNA sequences (her (and our) algorithm produces these foldings for 'free').
4. It is not easy to produce the semi pictorial representation automatically because overlap of substructures should be avoided in the otherwise arbitrary layout. Although this problem can be overcome, it imposes a large overhead on secondary structure algorithms and is therefore little used.

Nevertheless we think that an attractive secondary structure representation is important because of the heuristic nature of the folding problem: comparison of foldings produced under different constraints is required to obtain meaningfull results. A visual representation does not necessarily show spatial proximity relations directly but can use devices such as e.g. colour or connecting lines.

An alternative representation of secondary structures which overcomes the above mentioned problems is as follows:

1. Only primary structure induced proximity is represented as proximity. Thus the representation is essentially linear: sothat different sequences, different foldings of the same sequence or foldings of subsequences can be easily aligned in order to identify corresponding parts.

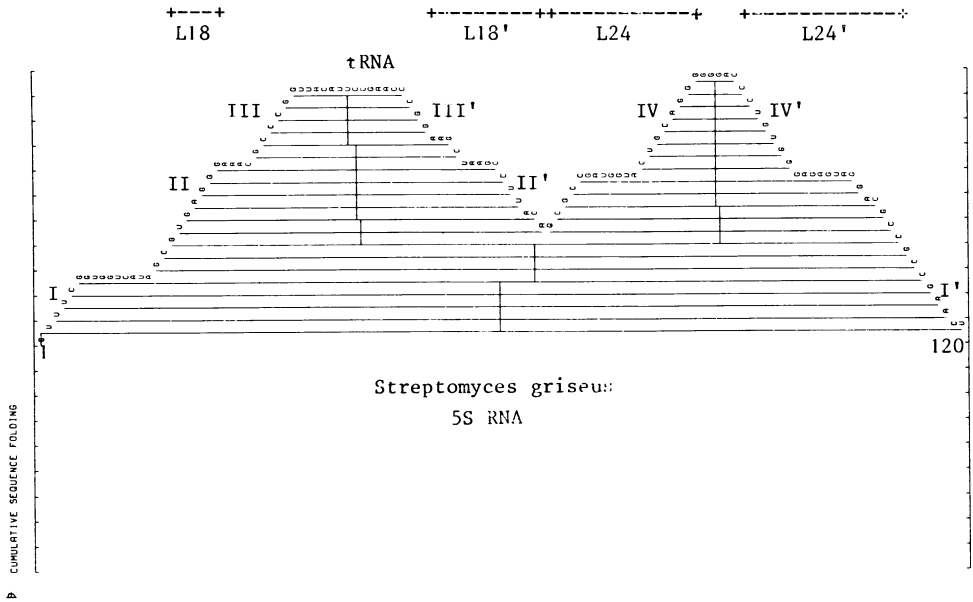

fig 1. Minimal energy folding of Streptomyces griseus 5S RNA. Calculated free energy: -56.9; I - IV : concensus helices; L18 and L24 : enzym interaction regions; tRNA : interaction site with tRNA (sequence CCGAAC).

2. Secondary bonds are represented by connecting lines; in order to avoid overlap of lines, stacked regions are expanded vertically while retaining the horizontal position that is dictated by the primary structure.
3. Single stranded regions are not extended vertically but are positioned at the 'level' reached by their adjacent double stranded regions. They form 'plateaus' and are therefore easily identified as single stranded regions.
4. Vertical lines link the middle of the horizontal connecting lines. This brings out the underlying tree like structure of planar foldings. In addition it fascilitates the recognition of uninterrupted stacking regions (seen as straight vertical line), the bulges (seen as interrupted sidewise displaced line) and the branching points (seen as initiation of multiple lines).
5. The representation is unique for a folding. Except for level changes it retains the same representation of substructures regardless of how they are embedded in the global structure.
6. No spatial constraints hamper the representation, which is easily imple-

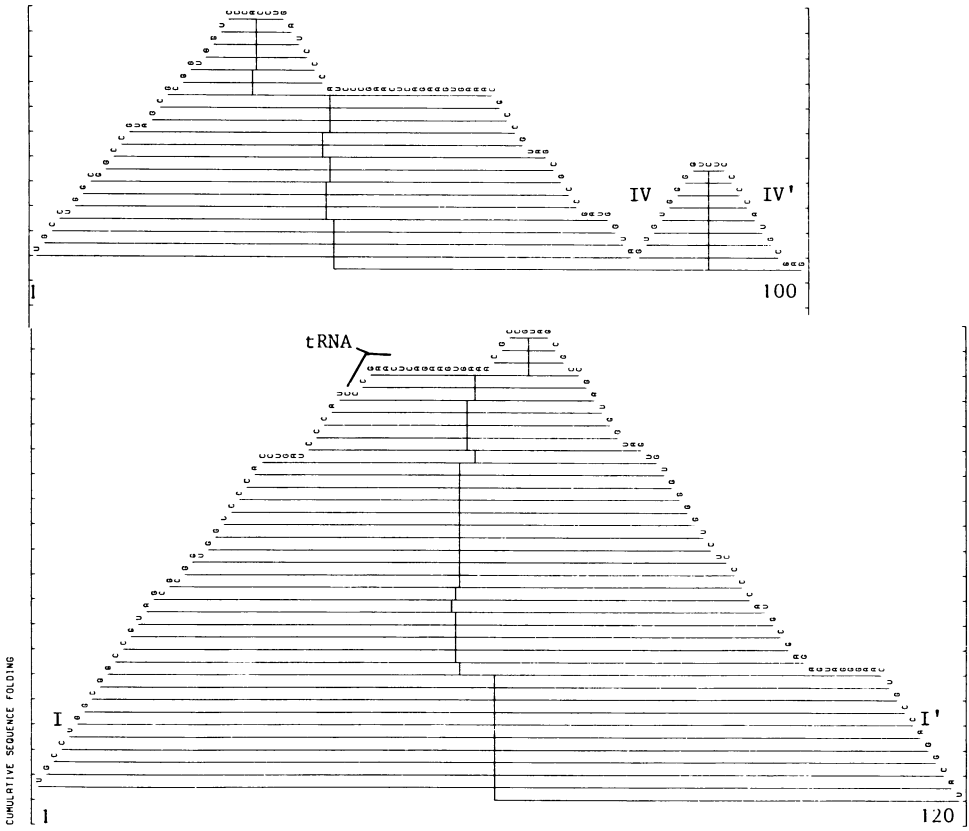

fig. 2. Minimal energy folding of Escherichia coli 5S RNA.
Calculated free energy: -69.8;
top: sub-sequence 1 - 100; concensus helic IV is formed.
bottom: complete sequence; only the main stem (I) corresponds with the in vivo secondary structure.

mented and allows for annotation of the sequences.

EXAMPLE: 5S-RNA SECONDARY STRUCTURES

The secondary structure of 5S-RNA is relatively well known. Comparative research has revealed which potential helices and which single stranded regions are preserved over most known sequences (7,8,9). Experimental research has confirmed that this consensus secondary structure is indeed the structure which occurs in vivo; the function of some of the regions is elucidated (10).

It is interesting to see that this consensus secondary structure coincides with the minimal free energy secondary structure in some species whereas in other species the minimal free energy structure does not form the consensus loops and helices. For example Streptomyces griseus (fig. 1) forms spontaneous the consensus structure including the main stem, two branches of which one has a helix near its base and a second helix closing off the hairpin loop with the t-RNA interaction site CCGAAC, and of which the other has a sharp (4) base) hairpin loop which is closed off with a GC-rich helix. The intermediate, supposedly single stranded loops which interact with enzymes do, however show some stacking with bulges; these stacking regions are less stable than the concensus helices sothat the minimal energy structure is consistent with the experimental findings.

Contrarywise, Escherichia coli 5S-RNA is capable of a more stable secondary structure than the biologically active structure (fig. 2). In this structure the t-RNA binding site is part of a helix and the molecule obviously cannot function in this structure. Thus in Escherichia coli other molecules interacting with 5S RNA should actively force the latter into the correct configuration, whereas Streptomyces griseus obtains the correct configuration spontaneously.

DISCUSSION

It is by no means evident that the minimal free energy configuration of molecules is meaningful for their biological function. In many cases a molecule will be forced into a particular configuration by its interaction with other molecules. The sequential generation of the molecule can trap the configuration in a local optimum. Alternative secondary structures play a role in the regulation of the transcription of operons (the attenuation mechanism, cf 11). Nevertheless the minimal free energy configuration provides us with a base line with which the structure of the in vivo molecule can be compared. Moreover the energy directed folding approach can be used to obtain the minimal free energy configuration of the molecule after the fixation of the part of its structure which is supposedly controlled by the interaction with other molecules. Thus we can investigate which interactions are actually necessary to form a particular configuration and which interactions are made possible by the prevailing configuration.

REFERENCES

1. Nussinov, R., Pieczenik, G, Griggs, J.R. and Kleitman, D.J. (1978) SIAM J.Appl. Math. 35 (I), 68-82.

2. Nussinov, R. and Jacobson, A.B. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 6309-6313.
3. Nussinov, R. and Tinoco, I. (1981) J. Mol. Biol. 151, 519-533.
4. Nussinov, R. and Tinoco, I. (1982) Nucl. Acids. Res. 10, 341-349.
5. Zuker, M. and Striegler, P. (1981) Nucl. Acids. Res. 9, 133-148.
6. Salzer, W. (1977) Cold Spring Harbor Symp. on Quant. Biol. 42, 985-1002.
7. Fox, G.E. and Woese, C.R. (1975) Nature 256, 505-507.
8. Woese, C.R., H. Luehrsen, C.D.P. and Fox, G.E. (1976) J. Mol. Evol. 8, 143-153.
9. Hori, H. (1976) Molec. gen. Genet. 145, 119-123.
10. Garett, R.A., Douthwaite, S. and Noller, H.F. (1981) TIBS 6, 137-139.
11. Watson, M.D. (1981) TIBS 6, 180-182.