Bulletin of Mathematical Biology Vol. 51, No. 1, pp. 167–171, 1989.
Printed in Great Britain.

0092-8240/89$3.00+0.00
Pergamon Press plc

A CONTINUOUS ANALOG FOR RNA FOLDING

■ VINCENT FERRETTI and DAVID SANKOFF*
Centre de recherches mathématiques,
Université de Montréal,
C.P.6128, Succursale "A",
Montréal, Canada H3C 3J7

A linear segment in which a number of pairs of intervals of equal length are identified as potential stems is the subject of a folding problem analogous to inference of RNA secondary structure. A quantity of free energy (or equivalently, energy per unit length) is associated with each stem, and the various types of loops are assigned energy costs as a function of their lengths. Inference of stable structures can then be carried out in the same way as in RNA folding. More important, perturbation of stem lengths and energy densities (modelling various mutational processes affecting nucleotide sequences) allows the delineation of domains of stability of various foldings, through the explicit calculation of their boundaries, in a low-dimensional parameter space.

Introduction. The study of how RNA secondary structure is dynamically related to primary structure is complicated by the discrete nature of the RNA molecule and the necessity of taking into account all possible Watson–Crick pairings at the level of the individual nucleotides. A way of simplifying this problem is suggested by some common methods for inferring RNA secondary structure from knowledge of primary structure. These methods, exemplified by the early work of Pipas and McMahon (1975), take into account the detailed nucleotide sequence only in the first steps of the algorithm in order to construct the possible "stems" or base-paired regions, and to evaluate their potential energetic contribution to the secondary structure. The determination of which of these regions are compatible with each other and the final choice of regions in the optimum structure can then be made with little or no reference to the precise nucleotide sequence.

In this note we propose using continuous analogs to discrete nucleotide sequences as a way of investigating the dynamic relationship between key parameters of primary and secondary structures without the difficulties of working in discrete spaces and avoiding the complexities of discrete optimization.

Thus we would hope to be able to examine what changes in secondary structure are provoked by small changes in primary structure, whether a given secondary structure is stable under small changes in primary structure and whether a given molecule can shift back and forth easily between two

* Author to whom correspondence should be addressed.

configurations. The small changes in question would be realized by changes in a few parameters rather than the many combinatorial possibilities in the corresponding discrete problems.

There are a number of ways of going about this program, but here we will confine ourselves to the simplest approach we have been able to devise. First, we look at the inference problem.

The Model. For the molecule of fixed length L, we are given a number of possible stems $s_i=(x_i, y_i, l_i, e_i)$, $i=1, \ldots, n$, where x_i and y_i identify the mid-points of two intervals, both of length $l_i>0$, which can be paired to each other along their entire lengths, thus releasing free energy e_i.

In each stem $l_i/2 \leqslant x_i$, $y_i-x_i>l_i+h_1$ and $y_i \leqslant L-l_i/2$, where $h_1>0$ is the minimum length of a "hairpin loop"; also $E^* \leqslant e_i/l_i<0$, where E^* is some negative constant representing the largest amount of free energy possible per base pair in a stem. In practice, $h_1=4$ and $E^* \approx -5$ kcal mol^{-1}. In addition we are given $E_1(.)$, $E_2(.)$ and $E_3(.)$, positive functions on $[h_1, \infty)$, $[h_2, \infty)$ and $[h_3, \infty)$, indicating the cost of hairpin loops, interior loops (including bulges) and multiple loops, respectively. The argument of each of these functions is the length of the loop, analogous to the number of unpaired bases in discrete secondary structures. For example, $E_j(t)=a_j+b_j \log t$, for $t \geqslant h_j, j=1, 2, 3$, where $h_2=1$ and $h_3=0$. We adopt the convention that $E_j(t)=\infty$ for $t<h_j$. Note that this is a great simplification since E_1 is known to be a decreasing function of t for small t and bulge energies are known to be much higher than that of other interior loops.

A secondary structure is any sub-set S of the n stems satisfying the usual assumptions that for any s_i and s_j in S where $x_i<x_j$, the intervals $(x_i-l_i/2, x_i+l_i/2)$, $(y_i-l_i/2, y_i+l_j/2)$, $(x_j-l_j/2, x_j+l_j/2)$ and $(y_j-l_j/2, y_j+l_j/2)$ are all disjoint ("no tertiary interactions") and either $y_i<x_j$ or $y_i>y_j$ ("no knots"). Furthermore a valid secondary structure must be stable, as explained in the following section.

Inferring Secondary Structure. The inference problem is to pick some sub-set S of the n stems which minimizes:

$$\sum_S e_i+\sum_B E_j(H_r), \tag{1}$$

where B is the set of loops determined by S and H_r is the length of the r^{th} loop, which is of type j, defined as follows.

If $s_i \in S$ and for no other $s_j \in S$ is $x_i<x_j<y_i$, then S determines a hairpin loop of length $y_i-x_i-l_i$ as in Fig. 1a.

If s_i and $s_j \in S$ where $x_i<x_j$ and $y_j>y_i$ but for all other $s_k \in S$ neither $x_i<x_k<x_j$ nor $y_j<y_k<y_i$, then S determines an interior loop if $h=x_j-x_i+y_i-y_j-l_i-l_j \geqslant h_2$, in which case h is its length, as in Fig. 1b.

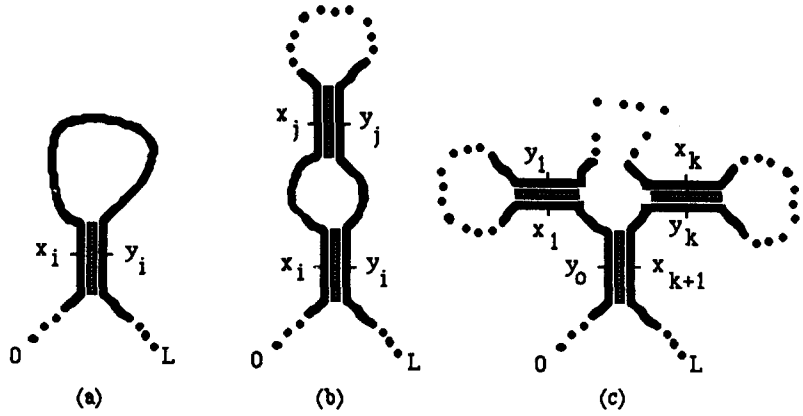

Figure 1. Types of loop in continuous secondary structure. Hatching indicates pairing of intervals in stems.

Finally if $s_1, \ldots, s_k$ and $s=(y_0, x_{k+1}, l_0, e_0)$ are all in S, where $k \geqslant 2$, such for each $i, 0 \leqslant i \leqslant k$, we have $y_i < x_{i+1}$ but for no other $s_j \in S$ is $y_i < x_j < x_{i+1}$, then S defines a multiple loop of length $\sum_0^k x_{i+1} - y_i - l_i$, as in Fig. 1c.

Note that each loop is associated with a particular stem, which is said to close the loop. In Figs 1a and b, this is the stem s_i, while in Fig. 1c it is s. For a structure to be stable, each loop it contains must be stable. If $E(H)$ is the energy of the loop closed by stem s, then the condition for stability is that $e(s)+E(H)$ be negative.

The optimal sub-set S may be found in a number of ways. One is to set up an $n \times n$ compatibility matrix M where $M(i,j)=M(j,i)=0$ unless s_i and s_j (or s_j and s_i) satisfy the two conditions (no tertiary interactions and no knots), in which case $M(i,j)=M(j,i)=1$. Following Pipas and McMahon (1975) we then generate all possible cliques (sub-sets C of $\{1, \ldots, n\}$ such that if i and j are both in C, $M(i,j)=1$), and then choose the one which minimizes equation (1).

Another approach is to use some version of dynamic programming (*cf.* Zuker and Sankoff, 1984) over the set of $2n$ paired interval endpoints.

Stability. There is no essential difference between the inference problems in the discrete and continuous cases, but in the latter setting it is much easier to address the problem of stability.

In the most elementary case, it suffices to multiply each l_i by $(1+\alpha)$ and each e_i/l_i by $(1+\beta)$, so that the energy of the stem becomes $e_i(1+\alpha)(1+\beta)$, and to see whether the optimizing structure changes or not. Since the energy may be written analytically as a function of α and β, the boundary between two structures in parameter space can be calculated explicitly.

This exercise is meaningful for those α for which $y_i - x_i \geqslant (1+\alpha)l_i + h_1$ and for those β for which $e_i(1+\beta)/l_i \geqslant E^*$.

Example. Consider a simple structure determining one interior loop and one hairpin loop as in Fig. 2. For the hairpin loop to be stable, we require:

$$-(1+\alpha)(1+\beta)e_2 > a_1 + b_1 \log[y_2 - x_2 - (1+\alpha)l_2].$$

For the interior loop to be stable, we require:

$$-(1+\alpha)(1+\beta)e_1 > a_2 + b_2 \log[x_2 - x_1 + y_1 - y_2 - (1+\alpha)(l_1 + l_2)].$$

For the stem s_1 to define a stable hairpin loop if s_2 is not present, we require:

$$-(1+\alpha)(1+\beta)e_1 > a_1 + b_1 \log[y_1 - x_1 - (1+\alpha)l_1\}.$$

Values $a_1 = 38$, $b_1 = 9$, $a_2 = 7$ and $b_2 = 14$ were estimated by a least squares fit to the Salser data cited by Zuker and Sankoff (1984), e_1/l_1 and e_2/l_2 were both set equal to -5.

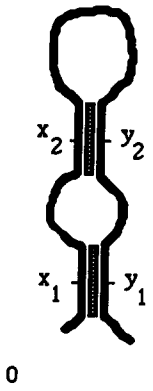

Figure 2. $L = 934$; $l_1 = 140$; $l_2 = 90$; $x_1 = 85$; $x_2 = 417$; $y_1 = 849$; $y_2 = 517$.

The three conditions listed above are summarized in Fig. 3. The boundaries between the different "phases" are found by replacing each of the three inequalities above by an equation.

Discussion. The elementary parametrization of the secondary structure energy we have given is of limited practical interest, since all stems are constrained to act alike. Our goal, however, was to demonstrate the type of consideration which becomes tractable in a continuous analog. The explicit calculation of the boundary between the stability domains of two structures exemplifies this sort of result.

In further research, it would be of interest to allow interval lengths and perhaps average energy levels to vary independently for each stem, for small n, and to try to characterize the type of "phase space" thus obtained.

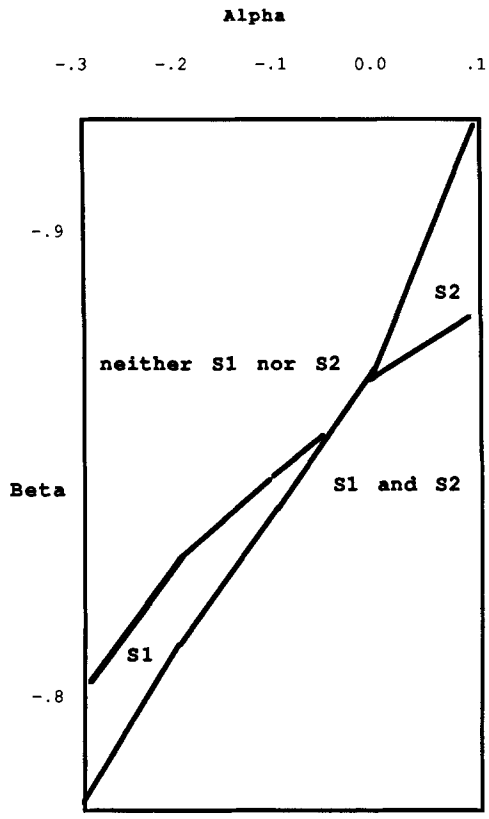

Figure 3. Stability domains for various secondary structures in parameter space.

REFERENCES

Pipas, J. M. and J. E. McMahon. 1975. "Method for Predicting RNA Secondary Structure." PNAS **72**, 2017–2021.

Zuker, M. and D. Sankoff. 1984. "RNA Secondary Structures and their Prediction." *Bull. Math. Biol.* **46**, 591–621.

Received for publication 1 July 1988