On Finding All Suboptimal Foldings of an RNA Molecule

MICHAEL ZUKER

An algorithm and a computer program have been prepared for determining RNA secondary structures within any prescribed increment of the computed global minimum free energy. The mathematical problem of determining how well defined a minimum energy folding is can now be solved. All predicted base pairs that can participate in suboptimal structures may be displayed and analyzed graphically. Representative suboptimal foldings are generated by selecting these base pairs one at a time and computing the best foldings that contain them. A distance criterion that ensures that no two structures are "too close" is used to avoid multiple generation of similar structures. Thermodynamic parameters, including free-energy increments for single-base stacking at the ends of helices and for terminal mismatched pairs in interior and hairpin loops, are incorporated into the underlying folding model of the above algorithm.

THE RNA SECONDARY STRUCTURE MODEL HAS BEEN IN existence since Fresco *et al.* (*1*) first showed that single-stranded RNA folds back onto itself in structures stabilized by hydrogen bonds between complementary bases. This model is not concerned with three-dimensional aspects of structure, but focuses solely on which hydrogen bonds form. This approach is appropriate, because while detailed three-dimensional structure data exists only for transfer RNA (*2*), three-dimensional modeling is premature for general RNA molecules.

This folding model is an example of what mathematicians call a discrete model. There are no continuously varying parameters such as bond lengths, angles, or interatomic distances. Instead, either a optimal folding (*3*). The uncertainties inherent in the model and in the thermodynamic data on which folding is based can be mitigated if a means of predicting suboptimal foldings is available.

Two types of RNA folding algorithms have the ability to find a minimum energy secondary structure. The "combinatorial" method, first introduced by Pipas and McMahon (*4*), forms structures by combining all potential helices in all possible ways. By their nature, combinatorial algorithms predict alternative foldings. The program developed by Ninio and co-workers (*5–7*) is based on a time-saving tree search method, but it does not escape from combinatorial reality. The number of possible foldings, and hence the computation time, grow exponentially with the size of the sequence (*8*), and it is not surprising that this and similar programs are limited to folding about 150 to 200 bases.

Minimum energy foldings can also be computed with recursive, or dynamic programming, algorithms. They were first used in the RNA folding problem by Nussinov *et al.* (*9*) to maximize base pairing. This method was subsequently extended to energy minimization (*10, 11*). These programs work in two stages. The first part, called the fill algorithm, computes and stores minimum folding energies for all fragments of the sequence. The process begins with all pentanucleotides and builds up to larger fragments in a recursive fashion. The second algorithm, called the traceback, computes a minimum energy structure by searching systematically through the matrix of stored energies. The main advantages over combinatorial algorithms are speed and the ability to fold relatively large molecules. By examining possible base pairs in the context of what neighboring base pairs might be, the algorithm escapes the tyranny of an exponentially growing number of structures. If the treatment of multibranched loops is sufficiently simple (*12*), a recursive folding algorithm can execute in time proportional to the cube of the sequence length. My own algorithm (*11*) can fold about 2000 bases

[redacted]

The first step toward this multiple folding algorithm came with attempts to extend the algorithm to fold circular RNA such as viroids (*15*). In a circular RNA, the choice of an origin is arbitrary. The key observation is that, in a circular molecule composed of ribonucleotides $r_1, r_2, \ldots, r_n$, a base pair linking r_i and r_j divides the secondary structure into two parts. There is a folding of the "included fragment" from r_i to r_j, and another folding of the "excluded fragment" from r_j through the origin to r_i. In a linear molecule, this symmetry is lost since the "excluded fragment" is broken into two linear segments, r_1 to r_i and r_j to r_n. The additivity assumption characteristic of recursive algorithms implies that the total folding energy is the sum of the energies of the two foldings. Steger *et al.* (*16*) extend the algorithm of Zuker and Stiegler (*11*) by computing additional numbers $V(j,i)$, analogous to $V(i,j)$, but

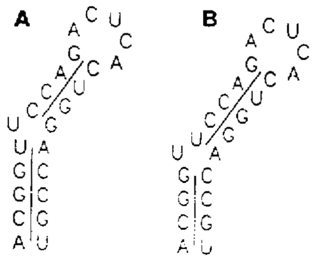

Fig. 1. Two foldings of the same oligonucleotide fragment that are a distance of 1 apart. The base pair U^5-A^{20} of (**A**) does not occur in (**B**), but its base numbers (*5, 20*) are both within 1 of the base numbers of G^4-C^{21} in (B). Similarly, U^6-A^{20} of (B) is equally close to C^7-G^{19} of (A). All other base pairs are common.

5 or 10 percent from an optimal folding of −100 kcal/mole would correspond to rare events with probabilities 2×10^{-4} and 6×10^{-8}, respectively. These large energy increments are chosen

[redacted]

replaced by U^{130}-A^{191} and in which U^{129} is single-stranded. The distance criterion introduced earlier was designed to eliminate the prediction of two such close structures. Thirty separate runs were made with the automatic feature to select foldings (Table 1). It is remarkable that so many trivially different 10-optimal foldings are found. When the distance between these foldings is forced to be greater than 2, the number falls dramatically. The 96 10-optimal foldings with $d = 10$ were examined in some detail. All of the structural motifs in the model of Burke *et al.* (*26*) occur in this collection, as do the structural elements contained in the model of Cech *et al.* (*24*). The P3 region is found without the two base pairs after the U-U mismatch. This entire motif appears only when $d \leq 2$. The reason is that those two base pairs are energetically unfavorable even when the rest of the motif forms. The entire P3 motif occurs in an 8.2-optimal structure (Fig. 5).

In the 5- and 10-optimal energy dot plots for the IVS (Fig. 6, B and C), the added lines create three triangular regions above the diagonal, corresponding to base pairs within the segments from 1 to 105, 106 to 213, and 214 to 413. In the 5-optimal plot, there are very few dots outside these triangular regions, which means that, within 5 percent of the minimum energy, base pairing between the three segments is unlikely. Alternative structures most likely occur from alternative foldings within these segments. The third and largest triangle is the most cluttered, implying that the greatest variability is in the last segment. In the 10-optimal dot plot, the number of possible long-range base pairs is considerable. However, the rectangles above and to the right of the middle triangle contain relatively few dots, which means that the segment from nucleotides 106 to 213 is likely to base pair only with itself in 10-optimal middle triangular region. Selecting a base pair (such as G^{109}-U^{321}) in the region results in a 4.6-optimal folding that eliminates 13 base pairs in the stem region of the branched motif (Fig. 7). Similar analyses of the energy dot plots for the IVS show that motif A (Fig. 4) is also well determined, as is the hairpin on A^{30} to U^{55}. In contrast, bases 75 to 105 of the IVS can participate in many alternative structures within 10 percent of the minimum energy.

This qualitative image analysis can be made more precise by introducing a new kind of plot. If r_i is the *i*th ribonucleotide in a sequence, then *P*-Num(*i*) can be defined as the total number of different base pairs in which r_i can participate in all *P*-optimal foldings. Thus *P*-Num(*i*) is the number of points in the *i*th row and column of the *P*-optimal energy dot plot. In plots of 5-Num and 10-Num for the IVS (Fig. 8), the 10-Num plot forms a trough in the region of the branched motif (U^{106} to A^{213}), indicating a relatively well-defined structure. The average value of 10-Num is 15.9 for this segment. However, the P-Q base pairing at the base of this region is not well defined, with average 10-Num values of 44.5 and 29.0 for P and Q, respectively. At the 10-percent-level of suboptimality, P and Q can take part in numerous alternative foldings. Thus a study of the 10-Num plot leads to the more conservative prediction that only the middle portion of the branched motif (bases 115 to 204) is well defined. The average value of 10-Num for the segment from 75 to 105 is high (41.1), confirming the earlier observation based on a visual inspection of the 10-optimal dot plot. The best defined regions are the A^{30} to U^{55} hairpin and the A motif (bases 226 to 246), with 10-Num averages of 4.8 and 6.2, respectively. At the 5 percent level, more precise statements can be made. There are 20 bases that are always single-stranded and 42 base pairs that always

[redacted]

occur in 5-optimal structures. In particular, the hairpin G^{31} to C^{54}

of combinatorial argument can be used to increase the output of the