[20] Computer Prediction of RNA Structure

By MICHAEL ZUKER*

Introduction

The biological activity of a RNA molecule is determined by its three-dimensional conformation. This structure is achieved by the molecule bending back on itself and forming helical regions stabilized by hydrogen bonds between complementary bases. Base pairing can be of three types: G with C, A with U, and the weaker G with U. This chapter does not deal with three-dimensional structure or with the complications introduced by interactions with other biomolecules. Instead, this chapter deals with computer methods to predict which hydrogen bonds will occur. It is the collection of hydrogen bonds that comprise the secondary structure, or folding, of the RNA molecule.

RNA structure is essential to the functioning of transfer RNA (tRNA) and to the assembly and functioning of ribosomal RNA (rRNA). The structure of yeast $tRNA^{Phe}$ has been determined by crystallographic means,[1] and several hundred other tRNAs, which have been sequenced, possess the same cloverleaf folding potential. Common foldings have been determined for many 5 S RNA molecules, and closely related structures have been computed for the small and large subunits of rRNA. In both cases, the models are well supported by phylogeny. Much of the interest in RNA secondary structure prediction comes from the desire to fold messenger RNA (mRNA). The structure of mRNA controls translation and splicing of introns in eukaryotes and transcriptional regulation in bacterial systems. Thus, mRNA structure controls whether some proteins are expressed and the level of expression as well. It has been proposed that mRNA structure might even affect the structure of the expressed protein.

The usual criterion for computing a RNA secondary structure is to minimize the free energy of the folded molecule. This model is a great simplification of reality. Three-dimensional effects are ignored, and the energy rules used to assign free energies to structures are derived from melting data on small oligonucleotides. In addition, the model itself, as a mathematical entity, has a disturbing property. In general, many different foldings are possible close to the minimum energy. If energy minimization

* National Research Council of Canada.

[1] S. H. Kim, F. L. Suddath, G. J. Quigley, A. McPherson, J. L. Sussman, A. H. J. Wang, N. C. Seeman, and A. Rich, *Science* **185**, 435 (1974).

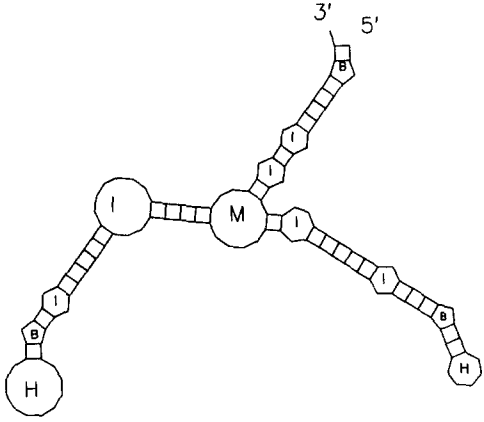

FIG. 1. Minimum energy folding of a 126-base sample RNA. The minimum free energy using Turner's rules is −35.5 kcal/mol. All the different kinds of loops occur and are designated by letters: B, Bulge loop; I, interior loop; H, hairpin loop; M, multiloop or multibranched loop. The unmarked areas are stacking regions between consecutive hydrogen bonds.

is the only folding criterion, then one must accept multiple solutions and find ways to compute them. Methods to determine the reliability of foldings become important. If additional information is available, this must be incorporated into the folding algorithm to narrow the range of solutions.

Definitions and Nomenclature

The ribonucleotides of a RNA molecule of length N will be denoted by $r_1, r_2, \ldots, r_N$. A base pair between r_i and r_j will be written as $r_i \cdot r_j$. The topology of RNA secondary structure has been well discussed.[2-4] A region of consecutive hydrogen bonds is called a helix or a stack. Single-stranded regions of the molecule comprise various types of loops, as shown in Fig. 1. A collection of helices interrupted only by bulge or interior loops is called a stem. A stem which ends in a hairpin loop is called a hairpin.

Only a few rules are imposed. The molecule cannot bend back too sharply on itself, so that at least three bases are required in hairpin loops.

[2] G. M. Studnicka, G. M. Rahn, I. W. Cummings, and W. A. Salser, *Nucleic Acids Res.* **5**(9), 3365 (1978).

[3] D. Sankoff, J. B. Kruskal, S. Mainville, and R. J. Cedergren, *in* "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, eds.), p. 93. Addison-Wesley, Reading, Massachusetts, 1983.

[4] M. Zuker and D. Sankoff, *Bull. Math. Biol.* **46**(4), 591 (1984).

A base can pair with at most one other base. Two base pairs $r_i \cdot r_j$ and $r_{i'} \cdot r_{j'}$ which satisfy

$$i < i' < j < j'$$

are said to form a pseudoknot. Although structures with pseudoknots have been shown to occur,[5] are one of two universal foldings for 5 S RNA molecules,[6] and have been invoked in intron splicing models,[7] structures with pseudoknots have been excluded by all RNA folding algorithms. Some algorithms would break down altogether if pseudoknots were permitted. Other algorithms simply become more complicated and slower. Another reason for their exclusion is that the pseudoknots would create complicated loops for which energy assignment would be very difficult. Complicated rules would have to be developed to predict which pseudoknots were possible and which were not.

Energy Rules

RNA foldings are usually computed to minimize free energy. At a finite temperature, the molecule is in some sort of equilibrium, and a minimum energy solution must be regarded as a most probable folding. Energy is assigned not simply to hydrogen bonds, but to the stacking of one hydrogen bond over another. Various loops are given destabilizing energies. A principle of additivity is assumed, so that the overall free energy of a folding is the sum of the energies of the stacked base pairs and the loops. This principle is certainly not correct, but the additivity assumption is essential to some prediction algorithms, and better experimental data are not yet available to produce more sophisticated rules.

Free-energy data come from melting studies performed on small oligonucleotides.[8-14] This information was summarized by Salser.[15] Standard base pairs of the form G–C and A–U are allowed. In addition, the non-

[5] J. D. Puglisi, J. R. Wyatt, and I. Tinoco, Jr., *Nature (London)* **331,** 283 (1988).
[6] E. N. Trifonov and G. Bolshoi, *J. Mol. Biol.* **169,** 1 (1983).
[7] R. W. Davies, R. B. Waring, J. A. Ray, T. A. Brown, and C. Scazzocchio, *Nature (London)* **300,** 719 (1982).
[8] I. Tinoco, Jr., O. C. Uhlenbeck, and M. D. Levine, *Nature (London)* **230,** 362 (1971).
[9] T. R. Fink and D. M. Crothers, *J. Mol. Biol.* **66,** 1 (1972).
[10] O. C. Uhlenbeck, P. N. Borer, B. Dengler, and I. Tinoco, Jr., *J. Mol. Biol.* **73,** 483 (1973).
[11] J. Gralla and D. M. Crothers, *J. Mol. Biol.* **73,** 497 (1973).
[12] J. Gralla and D. M. Crothers, *J. Mol. Biol.* **78,** 301 (1973).
[13] I. Tinoco, Jr., P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla, *Nature (London) New Biol.* **246,** 40 (1973).
[14] P. N. Borer, B. Dengler, I. Tinoco, Jr., and O. C. Uhlenbeck, *J. Mol. Biol.* **86,** 843 (1974).
[15] W. Salser, *Cold Spring Harbor Symp. Quant. Biol.* **42,** 985 (1977).

```
BASE PAIRING ENERGIES IN TENTHS OF A KCAL/MOL
SALSER'S DATA
STACKING ENERGIES (UG = GU)
     |  GU  |  AU  |  UA  |  CG  |  GC  |
  GU     -3     -3     -3    -13    -13
  AU     -3    -12    -18    -21    -21
  UA     -3    -18    -12    -21    -21
  CG    -13    -21    -21    -48    -43
  GC    -13    -21    -21    -30    -48
BULGE LOOP DESTABILIZING ENERGIES BY SIZE OF LOOP
             1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10| 12| 14| 16| 18| 20| 25| 30|
            28  39  45  50  52  53  55  56  57  58  59  61  62  63  64  65  67
HAIRPIN LOOP DESTABILIZING ENERGIES BY SIZE OF LOOP
             1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10| 12| 14| 16| 18| 20| 25| 30|
CG CLOSING  999 999  84  59  41  43  45  46  48  49  50  52  53  54  55  57  59
AU CLOSING  999 999  80  75  69  64  66  68  69  70  71  73  74  75  76  77  79
INTERIOR LOOP DESTABILIZING ENERGIES BY SIZE OF LOOP
CLOSED BY    1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10| 12| 14| 16| 18| 20| 25| 30|
  CG-CG     999   1   9  16  21  25  26  27  28  29  31  32  33  34  35  37  39
  CG-AU     999  10  18  25  30  34  35  36  37  38  39  40  41  42  43  45  47
  AU-AU     999  18  26  33  38  42  43  44  45  46  48  49  50  51  52  54  56
```

FIG. 2. Reproduction of the Salser energy input file for the author's RNAFOLD program discussed in the text. Energies are in tenths of a kilocalorie per mole so that integer arithmetic can be used. For the stacking energies, the column base pairs are 5′–3′, while the row base pairs are 3′–5′. The "999" energies are simply large numbers used to prevent hairpin loops which are too small, or impossible interior loops.

standard G–U base pair is allowed, although the nonstandard G–U base pair has sometimes not been allowed to occur at either end of a helix.[2,16] Loop destabilizing energies depend on the size (number of single-stranded bases) of the loop and on the nature of the base pairs which close them. There are no measured energies for multiloops. Multiloops can be treated as interior loops of the same size[16] or in a slightly more complicated way.[17] These so-called Salser's rules are given in Fig. 2. The folding temperature implicit in these rules is 25°.

Various alternatives and modifications to the above rules have appeared. Tinoco altered the rules slightly in 1982,[18] eliminating the dependence on closing base pairs in hairpin and interior loop destabilizing energies. Ninio adjusted the energy rules so that the phylogenetically determined structure would also be a minimum energy structure as frequently as possible. This was done first for tRNAs[19] and later extended to 5 S RNAs.[17] The set of nonstandard pairs is expanded to cover G–G,

[16] M. Zuker and P. Stiegler, *Nucleic Acids Res.* **9**(1), 133 (1981).
[17] C. Papanicolaou, M. Gouy, and J. Ninio, *Nucleic Acids Res.* **12**(1), 31 (1984).
[18] T. R. Cech, N. K. Tanner, I. Tinoco, Jr., B. R. Weir, M. Zuker, and P. S. Perlman, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3903 (1983).
[19] J. Ninio, *Biochimie* **61**, 1133 (1979).

Base pairing energies in tenths of a kcal/mol
Ninio/Turner loops + Soo NN at 37°
Stacking Energies

	GU	AU	UA	CG	GC
GU	-5	-5	-7	-15	-13
AU	-5	-9	-11	-18	-23
UA	-7	-9	-9	-17	-21
CG	-19	-21	-23	-29	-34
GC	-15	-17	-18	-20	-29

Bulge loop destabilizing energies by size of loop

1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
32	52	60	67	74	82	91	100	105	110	118	125	130	136	140	150	158

Hairpin loop destabilizing energies by size of loop

	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
CG CLOSING	999	999	74	59	44	43	41	41	42	43	49	56	61	67	71	81	89
AU CLOSING	999	999	74	59	44	43	41	41	42	43	49	56	61	67	71	81	89

Interior loop destabilizing energies by size of loop

closed by	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
CG-CG	999	8	13	17	21	25	26	28	31	36	44	51	56	62	66	76	84
CG-AU	999	8	13	17	21	25	26	28	31	36	44	51	56	62	66	76	84
AU-AU	999	8	13	17	21	25	26	28	31	36	44	51	56	62	66	76	84

FIG. 3. Turner's energy rules at 37° in the same format as Fig. 2. A computer program exists for creating these energy files for folding at arbitrary temperatures.

U–U, C–C, C–A, A–A, A–G, and U–C base pairs. A variety of special rules are introduced, which make the overall energy assignment nonadditive. In addition, these Ninio rules distinguish between (for example)

5′-UGC-3′ and 5′-UUC-3′
3′-AUG-5′ 3′-AGG-5′

when considering stacking involving nonstandard base pairs, while the Salser rules do not.

More recently, new experimental data on RNA duplex stability, made possible by breakthroughs in oligoribonucleotide synthesis, have resulted in energy rules which supercede Salser's rules.[20] Fig. 3 contains some of these new data. Note that 37° is the folding temperature. The only nonstandard base pair is G–U, but energy does depend on whether G or U is the 5′ base in a stack. As with the Ninio rules, single-stranded, terminal nucleotides, or dangling ends, are given free-energy increments. A new rule adjusts the energies of hairpin and interior loops depending on the nature of the terminal mismatched pair(s). These rules will be referred to as Turner's rules. Work continues to refine them.

Thermodynamic calculations allow one to alter energy rules for folding at different temperatures. This was done by Steger *et al.*,[21] starting

[20] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9373 (1986).

[21] G. Steger, H. Hofmann, J. Förtsch, H. J. Gross, J. W. Randles, H. L. Sänger, and D. Riesner, *J. Biomol. Struct. Dyn.* **2**(3), 543 (1984).

TABLE I
ADJUSTABLE PARAMETERS OF SEQL PROGRAM: IDEAS SEQUENCE ANALYSIS SYSTEM FOR RNA LOCAL SECONDARY STRUCTURE

SEQL parameters	Meaning (default value)
GMAX	Hairpins with lower free energies are printed (−10.0)
MODE	0, Do not print sequence; 1, print sequence; 2, print and annotate sequence (0)
LWID	Linewidth for terminal or output file (80)
LHMAX	Maximum size of a hairpin loop (20)
LIMAX	Maximum size of an interior loop (10)
LBMAX	Maximum size of a bulge loop (5)
LEN	Maximum size of any hairpin structure (100)

from most of the original data used by Salser. Because of differences in theoretical treatments, the Steger energy rules do not agree with Salser's rules even at 25°. A computer program written by Jaeger[22] adjusts the Turner rules for folding at arbitrary temperatures. Folding algorithms usually read in stacking and loop energies from external files. These files are easily changed for folding at alternate temperatures.

Types of Folding Programs

Four types of RNA folding programs are discussed. No single method is sufficiently precise to yield a definitive answer, and so different approaches are necessary to complement one another.

The first type of folding program will be called basic. This includes methods that predict hairpin structures or simply helices. Energy considerations may or may not enter. Programs which compute hairpins fall short of full secondary structure prediction by excluding the prediction of multibranched loops and by limiting the distance along the sequence between base pairs. These programs are good for finding local folding motifs. A good example of such a program is SEQL,[23] now part of the IDEAS analysis package of the Laboratory of Mathematical Biology at the National Cancer Institute of the National Institutes of Health (Bethesda, MD). This FORTRAN program has been written for VAX/VMS. The user can adjust seven parameters, which are explained in Table I. A sample output is shown in Fig. 4. The output includes hairpins, which may be incompatible with one another in a single global folding either

[22] J. A. Jaeger, personal communication (1987).
[23] M. I. Kanehisa and W. B. Goad, *Nucleic Acids Res.* **10**(1), 265 (1982).

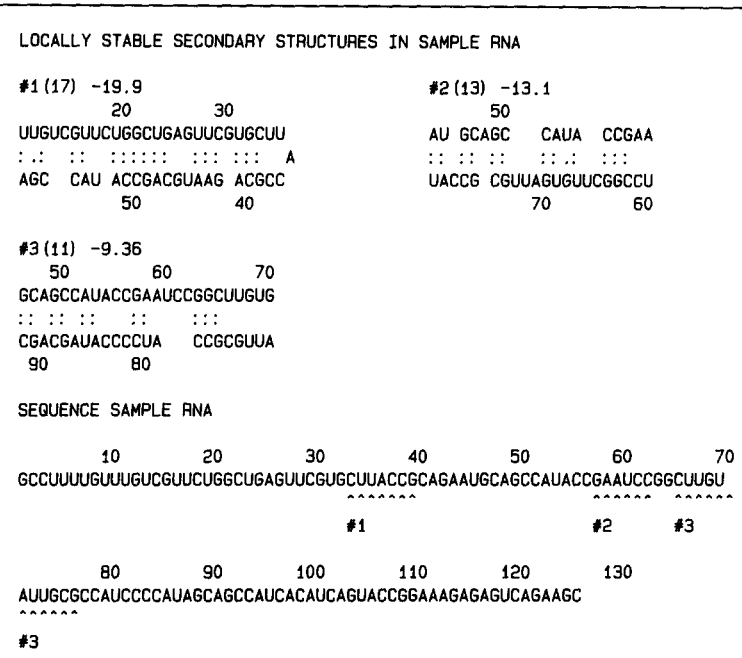

FIG. 4. SEQL analysis of the sample sequence folded in Fig. 1. The energy threshold for display is −5.0 kcal/mol. LHMAX, LIMAX, LBMAX, and LEN are 10, 8, 3, and 50, respectively. The annotated sequence shows the hairpin loops (including the closing base pairs) that were found.

because they involve hydrogen bonding of the same bases or because pseudoknots are created. The energies are according to Salser.[15]

Another kind of secondary structure program can be called combinatorial. A list of all possible helices is generated. These helices are then pieced together in all possible ways to form secondary structures. The best such program in use today was developed by Ninio, Dumas, Papanicolaou, and Gouy.[17,19,24,25] Called CRUSOE, it is written in FORTRAN 77 and is available from Gouy.[26] The computer memory required by the program is proportional to the number of helices used, and this grows as the square of the sequence size. It is computation time, not memory, which limits this algorithm. Although the program uses a clever tree search procedure to limit the amount of searching, computation time increases exponentially with sequence size, and the practical limit for a thorough

[24] J.-P. Dumas and J. Ninio, *Nucleic Acids Res.* **10**(1), 197 (1982).

[25] M. Gouy, P. Marliere, C. Papanicolaou, and J. Ninio, *Biochimie* **67**, 523 (1985).

[26] M. Gouy, *in* "Nucleic Acid and Protein Sequence Analysis: A Practical Approach" (M. J. Bishop and C. J. Rawlings, eds.), p. 259. IRL Press, Washington, D.C., 1987.

analysis is about 150 nucleotides. Larger molecules can be folded by increasing the minimum helix size from two and by eliminating the nonstandard base pairs. Both of these options are available. Pseudoknots are not allowed. This program has several advantages. Strict additivity of locally assigned energy is not necessary. Multiple foldings are easily predicted.

Recursive programs build optimal foldings one base at a time. They comprise two distinct parts. The first part, called the *fill,* computes and stores minimum folding energies for all fragments based on minimum folding energies of smaller fragments, starting with pentanucleotides. The final part, called the *traceback,* computes a structure by searching through the matrix of folding energies. The fill algorithm requires the bulk of computing time, while the time for a single traceback is negligible. If multiloops are treated in a simple manner, such algorithms can be much faster than the combinatorial type, executing in time proportional to the cube of sequence length. Computer storage requirements grow as the square of the sequence length and it is usually memory, not time requirements, which limit this kind of algorithm. Pseudoknots cannot be handled, and the additivity assumption for free-energy contributions is necessary. A recursive algorithm was first used by Nussinov *et al.*[27] to maximize base pairing and later extended to free-energy minimization.[28] Zuker and Stiegler developed a recursive algorithm independently, emphasizing the need for imposing constraints.[16] The algorithm was reprogrammed in 1983 by the author and has been adapted for a variety of different computers. Recursive algorithms of this kind are also called dynamic programming algorithms. By their nature, recursive algorithms predict the optimal folding of all subfragments. This means, for example, that optimal foldings of a growing RNA sequence can be simulated without any additional cost, once the entire sequence is folded. Recursive algorithms are designed to yield a single solution, but with some effort, the traceback algorithm can be extended to yield multiple solutions. This has been achieved by Williams and Tinoco.[29]

Another kind of folding algorithm can be called dynamic, because this algorithm simulates the folding of a RNA molecule in time. A good example of such an algorithm is the MONTECARLO program of Martinez, written in C.[30,31] The program first compiles a list of stems that can form.

[27] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, *SIAM J. Appl. Math.* **35,** 68 (1978).

[28] R. Nussinov and A. B. Jacobson, *Proc. Natl. Acad. Sci. U.S.A.* **77,** 6309 (1980).

[29] A. L. Williams, Jr. and I. Tinoco, Jr., *Nucleic Acids Res.* **14**(1), 299 (1986).

[30] H. M. Martinez, *Nucleic Acids Res.* **12**(1), 323 (1984).

[31] H. M. Martinez, *Nucleic Acids Res.* **16**(5), 1789 (1988).

These stems may contain small bulge or interior loops. The free energies of these stems are then computed, and the stems are then given weights, or Boltzmann probabilities, in proportion to $\exp(-\Delta G/RT)$, where ΔG is the stem energy, R is the gas constant per mole, and T is the temperature in Kelvin. One stem is then chosen at random, using the assigned probabilities. After $(i - 1)$ stems have been chosen, all stems in the list incompatible with the ones already chosen are removed, and the ith stem is added at random. The process terminates when no additional stems can be added. Pseudoknots are not permitted, although pseudoknots could easily be allowed. The only problems are how to assign energies and how to draw the resulting structures. The algorithm requires a modest amount of storage, similar to combinatorial algorithms, and is also very fast. Large RNA molecules can be folded quickly, but there is no guarantee of achieving a minimum energy folding. The folding procedure often terminates far from the global energy minimum as computed by a recursive algorithm. The strategy in MONTECARLO is to refold the molecule over and over again and to compute statistics on which stems tend to occur repeatedly. Thus, multiple foldings are very much part of the algorithm.

Zuker–Stiegler Algorithm and Associated Programs

The original algorithm, described by Zuker and Stiegler,[16] was written in FORTRAN 66 and ran in batch mode on an IBM 3032 processor under an earlier operating system (TSS), which is no longer in use. Several versions of the program were created for special purposes: one version forced known or desired base pairs to occur, while another version did not allow multibranched loops. These were all combined into a single, new program in the spring of 1983, which was designed to be interactive and to incorporate all the special features of the earlier programs. The programming language and computer system were the same.

This "generic" folding program will be called RNAFOLD for the purposes of this chapter. This program was adapted to the CRAY supercomputer by Michael Ess, a programmer at Cray Research, Inc. (Mendota Heights, MN). This program resides in the CRAY program library as RNAFOLD, version OCT83, but will be called CRAYFOLD in this chapter. The author adapted the program for use on the BIONET[32] DEC 2060 computer, using the TOPS 20 operating system, in which the program is called BIOFLD. When the IBM AT and associated clones appeared, the program was adapted for these microcomputers. This version, known as

[32] D. Roode, R. Liebschutz, S. Maulik, T. Friedemann, D. Benton, and D. Kristofferson, *Nucleic Acids Res.* **16**(5), 1857 (1988).

PCFOLD, also runs on the XT microcomputer and is completely menu-driven, with help files available on-line. A version was included in the University of Wisconsin Genetics Computer Group (UWGCG) (Madison, WI) package,[33] but this version would fold only a given fragment without allowing constraints or the repeat folding of subfragments. Called FOLD internally, this implementation will be called GCGFOLD in this chapter. Version 5 of the UWGCG package allows more options in FOLD. Other adaptations to the Cyber, to Unix machines, and to IBM VM/CMS have appeared. The author's personal versions now reside on a VAX 11/750 running VMS. The program closest to the original is called FOLD_VAX. The most up-to-date version is called FOLD_VAX_WISC, because this version is the basis of FOLD in version 5 of the UWGCG package. Both of these are distributed by the author along with notes, energy files, and sample runs.

Although computation time grows as the cube of sequence length, the real limits to RNAFOLD are space requirements. The program requires $3N^2/2$ half-integers (16-bit integers) plus assorted other fixed and variable arrays, which grow linearly with N. Many computers have a virtual memory capacity, which makes the working memory larger by using disk space temporarily. BIOFLD is limited to folding 280 bases, because half-integers are not available in FORTRAN on the DEC 2060, and there is no virtual memory. PCFOLD is limited to 425 bases. In fact, PCFOLD only does this well, because PCFOLD packs three integers into the space of two, utilizing only N^2 storage. PCFOLD will fail when as few as 14 base pairs are forced, because the extra energies used to force the folding must be stored in limited space. For this reason only, an "unpacked" version of PCFOLD, called PCFOLD2, has been retained. PCFOLD2 will not fail with a modest number of forced base pairs, but PCFOLD2 can fold only 345 bases. CRAYFOLD could handle about 2500 bases on the CRAY XMP. A more recent adaptation on the CRAY 2, using the vector processing capability of the machine, folded the entire human immunodeficiency virus (HIV) (9718 bases) in about 7 hr.[34] The present VAX version can easily fold 2000 bases on an 11/750. The program has been written to minimize the exchange of information between the central processing unit (CPU) and the disk. The "Wisconsin" version uses only as much space as is required for folding the given molecule. The size limit in GCGFOLD has been arbitrarily set at 1200, although the size limit could easily be increased.

The first prompt by the program is for an energy file name. The usual

[33] J. Devereux, P. Haeberli, and O. Smithies, *Nucleic Acids Res.* **12**(1), 387 (1984).
[34] J. V. Maizel, Jr., personal communication (1988).

energy file, currently called FOLD_VAX.ENE, uses Salser's energies. Tinoco's modifications[18] to Salser's rules are also available in a file named FOLD_VAX.NEWEN. A blank energy file called FOLD_VAX.NOEN is also distributed, which gives a zero energy to all stackings and loops. This blank energy file allows the user to set constant loop and stack energies within the program. Most recently, a new energy file using Turner's rules has been created. The fact that these energy rules distinguish between such stacks as

5′-AGG-3′	and	5′-AUG-3′
3′-UUC-5′		3′-UGC-5′

meant that the program had to be slightly altered. This modification has been done in FOLD_VAX_WISC and in GCGFOLD. The other versions of RNAFOLD remain unchanged. Nevertheless, the new energy file can still be used with the unaltered versions, since the energy discrepancies are slight. Single-base stacking (dangling ends) have not been incorporated into the algorithm.

The next prompt is for a SAVE, CONTINUATION, or regular run. The SAVE option creates a large file containing the energies of the time-consuming *fill* algorithm. This option can be done in batch mode. A subsequent CONTINUATION run allows the user to compute an optimal folding of the chosen fragment or of subfragments. This feature must be chosen as a command line option with GCGFOLD and is not available in PCFOLD. The S or C for SAVE or CONTINUATION must be entered in uppercase, as must all other commands.

The sequence file format has changed over the years. Sequences must be in uppercase, and T is treated the same as U. This could be a problem in using GenBank data, which is in lowercase. The original sequence format originated in the author's group. This format allowed many sequences in a single file. The sequences could only be accessed sequentially. The VAX versions use a subroutine called FORMID, which examines a sequence file and decides which format is being used. The earlier format is allowed, as well as the IntelliGenetics format used on BIONET,[32] the Protein Identification Resource format of the National Biomedical Research Foundation,[35] the GenBank[36] format, and the EMBL[37] format. The program reads and displays all sequence titles in the file. The user may select any sequence by number or name. GCGFOLD requires a single sequence in a file in UWGCG format, as do other pro-

[35] K. E. Sidman, D. G. George, W. C. Barker, and L. T. Hunt, *Nucleic Acids Res.* **16**(5), 1869 (1988).

[36] H. S. Bilofsky and C. Burks, *Nucleic Acids Res.* **16**(5), 1861 (1988).

[37] G. N. Cameron, *Nucleic Acids Res.* **16**(5), 1865 (1988).

grams in this package. PCFOLD will read a file in IntelliGenetics format or in the format used by the Pustell programs licensed by International Biotechnologies, Inc. (IBI, New Haven, CT).[38] Only one sequence per file is allowed.

After a sequence has been selected, the program queries the user to choose between terminal or file output and whether to create a CT (coordinate table) file for later use in plotting. The next step is the selection of a fragment for folding. The sequence itself is numbered from 1 to N. There is no way to have the numbering begin at a negative integer, so that, for example, the -35 and -10 regions of a mRNA leader sequence show up as such. However, the fragment, which is selected, retains the "historical numbering" of the full sequence from which the fragment was selected. At this point, the user may enter the command F to begin folding, T to terminate execution, or a variety of other commands to constrain or otherwise alter the folding. These commands and their syntax will be discussed later. The F (fold) command is always used to begin folding. After the folding, a refolding of the entire fragment or of any subfragment may be obtained by entering the 5′ and 3′ ends in historical numbering. This refolding is very quick. If the number one is added after the 5′ and 3′ subfragment ends, then the refolding forces the ends to pair with one another if possible. This refolding procedure, which can also be used to compute the stability of local regions, is demonstrated in Fig. 5. At this stage, various output parameters can be altered. Program execution may then be terminated, or another fragment may be selected from the same or another sequence.

Output and Display

There are three criteria for judging the output of an RNA folding program. The output should be quickly and easily produced without requiring special devices. The results should be visually appealing. Finally, the display should be designed so that alternative foldings of the same sequence can be visually inspected for similarity. In practice, these criteria often conflict with one another.

All the programs mentioned in this chapter produce a line printer output, which either lists helices or else draws a picture of the secondary structure including all single- and double-stranded regions. This is the "quick and easy" output, which is not suitable for publication quality representations of secondary structures or for comparative studies.

Better quality displays are produced by graphics programs, which read base-pair lists created by folding programs and which output files for

[38] J. Pustell and F. C. Kafatos, *Nucleic Acids Res.* **14**(1), 479 (1986).

```
FOLDING BASES    1 TO  126 OF sample RNA
ENERGY  =     -35.5

         10             20         30
-  C     U  G  --    UC       A    G   UU
 GC UUUUG UU UC    GU  UGGCUG GUUC UGC
 CG AAGAC GA AG    CA  ACCGAC UAAG ACG  A
C  -     U  G  AA    -U       G    -   CC
     120                50          40

                      60             70          80
                    CGAA      ----CU       U  -  CAUC
                        UCCGG       UGUGAU GC GC    C
                        AGGCC       ACACUA CG CG    C
                    ----      AUGACU       C  A  AUAC
                      110        100         90

ENTER: T TERMINATE, NS NEW SEQUENCE, NF NEW FRAGMENT,
O OUTPUT PARAMETER DEFINITION, OR THE ENDPOINTS OF A
SUBFRAGMENT BETWEEN     1 AND   126.
END WITH 1 TO FORCE ENDS TO BASEPAIR.
14 113 1 ←
FOLDING BASES   14 TO  113 OF sample
ENERGY  =     -28.7

          20         30
 --     UC       A    G   UU
C     GU  UGGCUG GUUC UGC
G     CA  ACCGAC UAAG ACG  A
 AA     -U       G    -   CC
           50          40

         60             70          80
      CGAA      ----CU       U  -  CAUC
          UCCGG       UGUGAU GC GC    C
          AGGCC       ACACUA CG CG    C
      ----      AUGACU       C  A  AUAC
        110        100         90
```

FIG. 5. Example of repeat foldings and the line printer output of RNAFOLD. The first structure is the minimum energy folding plotted in Figs. 1, 6, and 7. The second folding of bases 14–113 forces the C-14–G-113 base pair, which would not occur in an optimal folding of this segment. The energy difference between the two foldings gives the stability of the stem from G-1–C-125 to C-14–G-113 (−6.8 kcal/mol). The arrow points to input from the user.

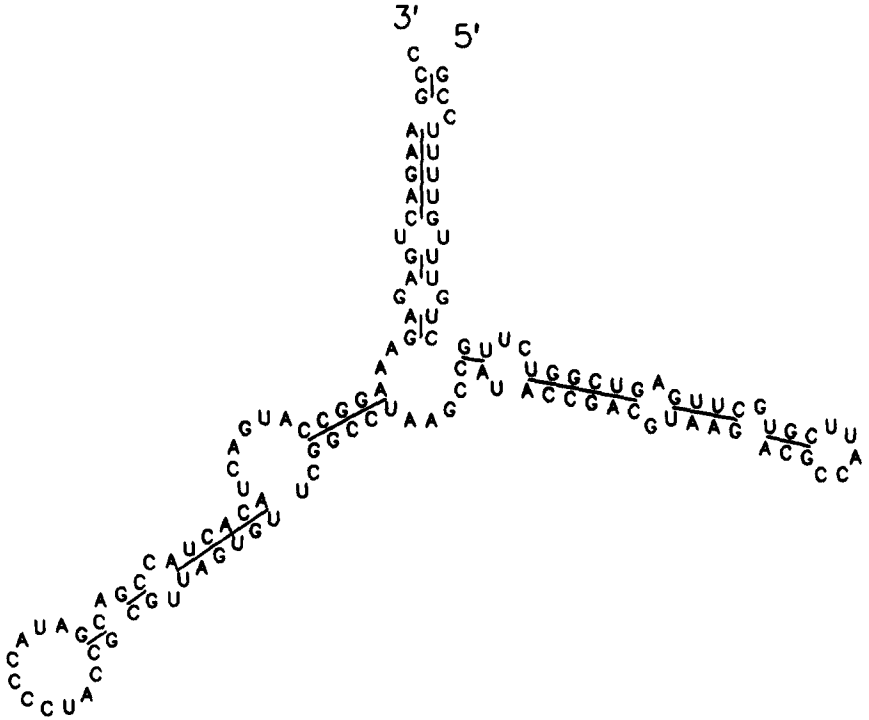

FIG. 6. Plot of same folding given in Fig. 1. The DRAW program was used to generate both. The automatic untangling feature was used, and the base letter display option was chosen.

use on standard plotting devices. One such program, originally developed by Osterburg and Sommer,[39] has been adapted for the UWGCG package and is called SQUIGGLES. Shapiro *et al.*[40] wrote the DRAW program in the SAIL language for use on the VAX. The DRAW program was soon translated into Pascal. Figures 1 and 6 were drawn using a version of this program adapted to the IBM VM/CMS computing environment. The program is interactive. The user can use the cross-hair feature of Tektronix-type terminals to designate pivot points for rotating portions of the structure to eliminate overlaps. This program produces a very pleasing output, although some effort is usually required. There is an automatic untangling feature, which eliminates most of the need for user intervention, but the resulting output is often not as visually satisfying. However, by placing

[39] G. Osterburg and R. Sommer, *Comput. Programs Biomed.* **13,** 101 (1981).

[40] B. A. Shapiro, J. V. Maizel, Jr., L. E. Lipkin, K. Currey, and C. Whitney, *Nucleic Acids Res.* **12**(1), 75 (1984).

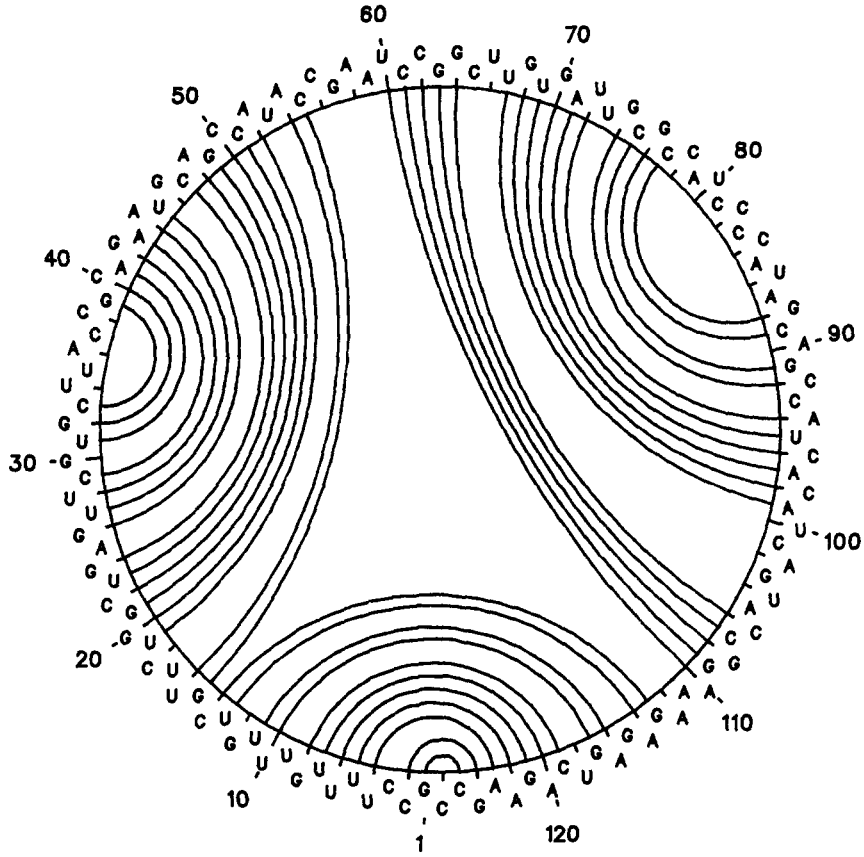

FIG. 7. A circle plot of the folding displayed more conventionally in Figs. 1, 5, and 6.

stems at prescribed angles to one another, this option is useful for the visual comparison of repeated foldings of the same molecule. Bruccoleri and Heinrich report an improvement in this algorithm, programmed in C for VAX/VMS systems.[41] Perhaps the best way to draw structures for comparative purposes is the circle representation first introduced by Nussinov *et al.*[27] The sequence is mapped along the perimeter of a circle, and hydrogen bonds are represented by circular arcs, which cut the circle at right angles. This presentation is one of the most abstract ways to display RNA foldings. An example is shown in Fig. 7. J. R. Thompson has created a graphics program called MOLECULE written in Pascal for use on

[41] R. E. Bruccoleri and G. Heinrich, *CABIOS* 4(1), 167 (1988).

REGION table

(1)	1	125	2	-3.4
(2)	4	123	5	-3.7
(3)	10	117	2	-0.5
(4)	13	114	2	-2.3
(5)	15	55	2	-2.1
(6)	19	52	6	-11.6
(7)	26	45	4	-3.9
(8)	31	41	3	-5.2
(9)	60	110	5	-10.1
(10)	67	99	6	-8.9
(11)	74	92	2	-3.4
(12)	76	89	2	-3.4

CT file

126	ENERGY =	-35.5			sample RNA
1	G	0	2	125	1
2	C	1	3	124	2
3	C	2	4	0	3
4	U	3	5	123	4
5	U	4	6	122	5
6	U	5	7	121	6
7	U	6	8	120	7
8	G	7	9	119	8
.	.	and so on .		...	.
125	C	124	126	1	125
126	C	125	0	0	126

FIG. 8. REGION table and part of the CT file for the optimal folding of the sample RNA sequence displayed in Figs. 1, 5, 6, and 7. The REGION table contains no sequence information. This information must be supplied in another file. Each row describes a single helix. The first column is the number of the helix. The next two columns give the exterior 5'- and 3'-closing bases of the helix. The fourth column is the length of the helix. The final column is the helix energy. The first row of the CT file contains the total number of bases in the folded sequence, the folding energy, and the sequence name. Subsequent lines contain the base number (in the fragment), the base letter, the number of the 5'-connecting base, the number of the 3'-connecting base, the number of the hydrogen-bonded base (0 if the base is single stranded), and the historical number of the base in the original sequence. The REGION table is far more compact.

the IBM AT and XT. MOLECULE is based on an original program by Lapalme *et al.*[42] The output is similar to that of DRAW, although the program is not interactive. MOLECULE is distributed along with PC-FOLD on a single diskette by the author.

Two formats have been widely used for base-pair output from folding programs. The first is the CT file introduced by Feldmann.[43] All the UWGCG graphics programs for RNA structure display accept this format, as does the MOLECULE program. The REGION table was introduced by Shapiro *et al.*[40] for use with DRAW. REGION does not contain any sequence information. Examples of these files are given in Fig. 8.

The author's RNAFOLD program will produce a line printer output, as well as CT files and REGION tables. The program has six output parameters, three of which can be set by the user to control output. These parameters can be set before folding begins or before a repeat folding is requested. The command syntax before folding is

$$\text{PO } i_1, j_1 \qquad i_2, j_2 \ldots$$

[42] G. Lapalme, R. J. Cedergren, and D. Sankoff, *Nucleic Acids Res.* **10**(24), 8351 (1982).

[43] R. J. Feldmann, "Manual for Programs NUCSHO and NUCGEN of Nucleic Acid Structure Synthesis and Display." Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland, 1981.

TABLE II
THREE DIFFERENT KINDS OF OUTPUT FROM RNAFOLD PROGRAM CONTROLLED BY SINGLE PARAMETER

Parameter value	Output parameter number 2 from RNAFOLD		
	Line printer output	REGION table output	CT file output
1	Yes	No	No
2	No	Yes	No
3	No	No	Yes
4	Yes	Yes	No
5	Yes	No	Yes
6	No	Yes	Yes
7	Yes	Yes	Yes

and has the effect of setting output parameter i_1 equal to j_1, parameter i_2 equal to j_2, and so on. All three variable parameters may be set in a single command. Before refolding, the command syntax is

$$\text{O } i_1, j_1 \quad i_2, j_2 \ldots$$

Although PCFOLD is completely menu-driven, the above command is still valid at the refolding stage and makes the program a bit more flexible. Parameter number 2 controls which kinds of output will be produced by the program. Table II shows the possibilities. This parameter is 1 by default and is automatically set to 5 if a CT file is requested. Setting this parameter to 4, 6, or 7 is the only way of getting a REGION table, which will show up as file FOR025.DAT or FILE.REG on the VAX, depending on which version of RNAFOLD is used. With PCFOLD, the user is prompted by the operating system for a file name. Parameter number 5 controls the number of columns in the line printer output. Parameter number 5 is 132 by default on the VAX and 79 in PCFOLD. Parameter number 6 is not normally used. Parameter number 6 is the FORTRAN unit number used for the line printer output. This parameter is 6 (terminal output) normally and is set to 24 when file output is selected. If a file is named for the output, the user will not see the output on the screen. After folding, this parameter can be set to 6, and a refolding can be generated, which will be sent to the terminal. Setting this parameter back to 24 will send further output to the file. Setting the parameter to 7, for example, would create the new file FOR007.DAT on the VAX (user-defined name with PCFOLD), which would collect all further line printer output. These

features are not available in GCGFOLD, in which the line printer output and CT file are created by default.

Experimental Data and Constrained Folding

Because free-energy minimization is not sufficient to determine a folding with confidence, it is essential to utilize whatever other information is available. This information is usually about specific single- or double-stranded regions, which are believed to occur. It is best to fold a RNA molecule first without constraints and see just how much of the folding conflicts with data on single- or double-stranded regions. This is called a free folding. In general, the energy of the refolded molecule (using constraints) will be greater than the energy of the free folding. If this energy difference is of the order of 10%, there is no problem. When the energy difference rises to 50%, this is an indication that something is wrong. Although the examples given refer almost entirely to the author's program, the principles are general and could be applied to other methods.

Various kinds of data indicate single-stranded regions. Some chemically modified bases lose their ability to base pair. A ribosome or protein, which binds to a RNA molecule, may prevent many bases in a row from pairing. Some enzymes cleave a RNA molecule preferentially in single-stranded regions. Analysis of digestion fragments leads to predictions that some bases are single stranded. One way to prevent these bases from pairing in computer foldings is to assign large destabilizing energies to loops or stacks which involve paired bases that should be single stranded. This is what is done in RNAFOLD, in which single-stranded regions can be designated in two different ways. The first way is to modify the sequence. An X can be used to mark a base which cannot pair, or, better still, lowercase a, c, g, and u can be used. This will not work with GCGFOLD. There is also a command within the program that can be used. Before folding begins, the command

$$\text{AP } i,\ 0,\ k$$

will force $r_i, r_{i+1}, \ldots, r_{i+k-1}$ to be single stranded. In PCFOLD, this can be accomplished using the menu system. In GCGFOLD, the option

$$\text{/PREVENT} = \text{i},\ 0,\ k$$

on the command line will have the same effect, although this is not stated in the program manual. The CRUSOE program allows single-stranded regions to be designated by using an X. This has the effect of reducing execution time.

It may be desirable to prevent certain base pairs or entire helices from

forming. This can be especially useful when folding regulatory RNA.[44] When one set of base pairs is disrupted, another interesting folding may appear. As with forced single-stranded regions, penalty energies can be used to prevent unwanted base pairs. In RNAFOLD, the command

$$\text{AP } i, j, k$$

issued before folding begins will prevent the formation of the base pairs $r_i \cdot r_j$, $r_{i+1} \cdot r_{j-1}$, ..., $r_{i+k-1} \cdot r_{j-k+1}$. This is accomplished in GCGFOLD by the option

$$\text{/PREVENT} = i, j, k$$

on the command line and through the menu system in PCFOLD.

Double-stranded-specific enzymes, such as cobra venom RNase, may suggest that certain bases are in helices. A nonenzymatic double-stranded-specific probe has been reported.[45] Such data must be used carefully, since the data really indicate probable double-stranded regions. If a number of consecutive bases are thought to be double stranded, it is safer to force only a few of them to be paired. In RNAFOLD, the command

$$\text{AF } i, 0, k$$

issued before folding begins will force r_i, r_{i+1}, ..., r_{i+k-1} to be double stranded if possible. PCFOLD uses a menu system, while in GCGFOLD the option

$$\text{/FORCE} = i, 0, k$$

on the command line will have the same effect, although this is not reported in the notes. The forced pairing is accomplished by giving a bonus energy to stacks or loops closed by a designated base. The value of the bonus energy is -50.0 kcal/mol by default (-20.0 in PCFOLD). This quantity is totally arbitrary; all that is important is that the bonus be sufficient to cause the desired base to be double stranded. The bonus energy number is actually an adjustable parameter and can be set before folding commences by the command

$$\text{PE } 9, e$$

where e is the new bonus energy. The bonus energy is stored in tenths of a kilocalorie per mole, so that -50.0 is stored as the integer number -500. This energy cannot be altered in PCFOLD or GCGFOLD, except by altering the value of EPARAM(9) in the source code. A total of 66 forced base pairs will cause a computer memory overflow, because half-integer

[44] C. Yanofsky, *Nature (London)* **289**, 751 (1981).
[45] C. P. H. Vary and J. N. Vournakis, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6978 (1984).

variables are used and the largest (absolute) energy that can be stored is −3276 kcal/mol. In such cases, fewer bases should be forced to pair or else the bonus energy could be set to −30.0 or −20.0 kcal/mol, as long as the base pairing still occurs. The bonus energies are subtracted by the traceback algorithm so that the output contains the correct energy. PCFOLD will overflow with as few as 14 forced base pairs. One solution is to force fewer base pairs. Another is to use PCFOLD2, which can tolerate larger energies, although PCFOLD2 is limited to smaller molecules.

Cross-linking experiments can be valuable in indicating the existence of specific base pairs. Phylogenetic comparison between homologous RNA sequences may also point to conserved helices, which are thought to occur in a RNA folding. Enzymatic digestion data may also suggest the existence of certain base pairs. In this case, specific base pairs can be forced to occur. In RNAFOLD, the command

$$\text{AF } i, j, k$$

issues before folding begins will force the base pairs $r_i \cdot r_j$, $r_{i+1} \cdot r_{j-1}$, ..., $r_{i+k-1} \cdot r_{j-k+1}$. This is handled by the menu in PCFOLD and by the command line option

$$\text{/FORCE} = i, j, k$$

in GCGFOLD. Forced base pairs between noncomplementary bases will not occur. Bonus energies are used to achieve these base pairs, and the same comments given above hold. In case of overflow, there is one extra trick that can be used. The command

$$\text{AF } i, j, k$$

which forces k base pairs causes k bonus energies to be added to the overall energy. The commands

$$\text{AF } i, j, 1$$
$$\text{AF } i + k - 1, j - k + 1, 1$$

force just the first and last base pairs of the helix. This adds just two bonus energies, and the intervening base pairs will form automatically. The CRUSOE program allows for forced helices, and this option greatly reduces execution time. In RNAFOLD, execution time does not improve.

It is desirable to be able to excise fragments from a sequence and fold the remaining sequence. In RNAFOLD, the command

$$\text{AX } i, j$$

excises r_i through to r_j and covalently links r_{i-1} with r_{j+1}. This command may be repeated to cut out other segments. The most obvious application is to intron splicing. The program retains the historical numbering of

nucleotides in the processed sequence, making it easier to compare foldings of a mRNA with or without introns. Another feature, which seems to be unique to RNAFOLD, is what the author calls the closed excision. The command

$$\text{AE } i, j$$

excises r_{i+1} through to r_{j-1} and links r_i and r_j together by a hydrogen bond, not by a covalent bond. The base pair $r_i \cdot r_j$ must be possible. Multiple, closed excisions are allowed. If closed excisions are made at i_1, j_1; i_2, j_2; $\ldots, i_k, j_k$; then the condition $i_1 < j_1 < i_2 < j_2, \ldots, < i_k < j_k$ must hold. The folding of the "processed" sequence and the foldings of all the excised fragments with their 5′ and 3′ ends forced to base pair yield a composite folding of the entire sequence with base pairs $r_{i_1} \cdot r_{j_1}$ to $r_{i_k} \cdot r_{j_k}$ forced. This composite folding is much faster than using the AF i, j, k option above, and because the folding is split into smaller parts, this procedure might turn a very large folding problem into a tractable one. The negative side to this procedure is that the output is fragmented into $k + 1$ separate files. If REGION table output is produced, these files could all be appended into a single file for immediate use by DRAW or another plotting program. PCFOLD handles excisions with the menu system, while GCGFOLD offers only the regular excision with the option

$$\text{/REMOVE} = i, j$$

on the command line.

RNAFOLD has five parameters which are called folding parameters. The correct syntax for setting the folding parameters is

$$\text{PF } i_1, j_1; i_2, j_2 \ldots$$

issued before folding commences. One up to all five parameters can be set in a single command. The effect of this command is to set parameter i_1 equal to j_1 and so on. These parameters are set using the menu feature of PCFOLD and are not available in GCGFOLD. The use of the first three parameters can greatly restrict the range of allowable base pairs, so that a large increase from the free folding energy is to be expected in general.

The first folding parameter is the minimum size of a hairpin loop and equals three by default. This parameter can be set larger to force base pairs between distant regions of the molecule. The second parameter is the maximum value of $j - i$ allowed in a base pair $r_i \cdot r_j$ and is equivalent to the LEN parameter in SEQL. Normally set to infinity so that there is no constraint, this parameter can be set to 100, 50, or even 30 to force foldings with only short-range base pairs. When this parameter is small, folding is faster. The third parameter is either 0 or 1. When 0, normal

folding occurs. When set to 1, multiloops are not allowed, and the output is just a sequence of hairpin structures, called an open folding. This parameter is often used in conjunction with the second parameter. The program runs much faster, with execution time growing as the square of the fragment size, rather than as the cube with the regular algorithm. The fourth parameter takes effect only when an open folding is selected using parameter 3. This parameter is normally 1. In this case, repeat foldings of subfragments will yield correct results only when the 5′ end is included. If the 5′ end is not included, only a single optimal hairpin structure will be produced. When this parameter is set to 2, repeat foldings may omit the 5′ end, but repeat foldings must include the 3′ end to produce an optimal open folding. If this parameter is set to 0, then arbitrary subfragments may be chosen for repeat foldings, but the price paid is that improvements in program speed are lost.

The fifth folding parameter offers another way to deal with single-stranded-specific information. RNase T1, for example, cuts the covalent linkage of a G nucleotide with its 3′ neighbor. The crudest way to handle such a situation is to prevent the G from pairing. A more subtle approach is to allow G or its 3′ neighbor to pair, but not both. In RNAFOLD, bases are called accessible (to RNase attack) if their 3′-phosphodiester bond is cleaved by a RNase. The accessible bases are designated using the letters B, Z, H, and V for A, C, G, and U, respectively. The default value of this fifth parameter is 0. In this case, B, Z, H, and V are treated as bases that cannot pair. When the parameter equals 1, B, Z, H, and V are treated just like A, C, G, and U, respectively. When the parameter equals 2, accessible bases are allowed to pair only if their 3′ neighbor is not paired. This is a minimum constraint for nuclease accessibility. This constraint is achieved using large penalty energies. When single-stranded-specific probes are used, initial cuts may allow the molecule to unfold somewhat, exposing further sites to attack. For this reason, such data must be used conservatively, with only the very best sites designated as accessible.

Another feature of RNAFOLD is that G–U base pairs at the ends of helices are either allowed or not allowed. The parameter that controls this is set within the program, so the user has no choice. However, the line

$$\text{EPARAM}(1) = 1$$

in the main program, which prevents these external G–U base pairs, can be changed to

$$\text{EPARAM}(1) = 0$$

to allow the external G–U base pairs. The earlier versions of RNAFOLD, including PCFOLD, do not allow the external G–U base pair. The newer

versions, including GCGFOLD, do allow the external G–U base pairs. The author's latest VAX version of RNAFOLD offers the user a choice. There no longer seems to be any reason to exclude the external G–U base pairs.

Confidence and Reliability

When a free folding and a folding using constraints give the same answer, the user's confidence is increased. The same is true when phylogenetic data support a folding. However, even when data are available to constrain a folding, it is desirable to know how robust the computer prediction is. How can one determine which parts of a structure are well determined and which are not? One solution is to predict a number of alternative foldings close to the minimum energy and to observe what motifs, if any, are shared by these foldings. Another method is to compare the minimum folding energy of a molecule with what would be expected from folding a random sequence with the same A, C, G, and U content.

By their nature, both CRUSOE and MONTECARLO produce a variety of foldings. With MONTECARLO, there is no guarantee that any of the solutions will be close to the minimum energy, but this is not the case with CRUSOE. Unfortunately, this program can only fold small RNA molecules, and so its usefulness is limited. Recursive programs, such as RNAFOLD, do not naturally yield multiple solutions. Williams and Tinoco[29] extended the traceback algorithm in RNAFOLD, so that multiple solutions not far from the energy minimum are predicted. One cannot expect to predict all foldings within, for example, 10% of the energy minimum, because there may be an astronomical number of foldings, and in any case, there may be many foldings that are similar to one another. The author has extended the RNAFOLD program so that all base pairs that can occur in foldings close to the energy minimum are predicted.[46–48] Individual suboptimal foldings can be predicted, but it is the collection of all possible base pairs that shows some motifs to be well defined and other motifs to be poorly defined.

Virologists, folding very large viruses using RNAFOLD, have been forced to break up the problem into the folding of overlapping segments. Structural agreement in overlapping areas increases confidence in the overall folding. This technique is in general use.

[46] A. B. Jacobson, M. Zuker, and A. Hirashima, *in* "Molecular Biology of RNA: New Perspectives" (M. Inouye and B. S. Dudock, eds.), p. 331. Academic Press, Orlando, Florida, 1987.

[47] M. Zuker, *in* "Mathematical Methods for DNA Sequences" (M. S. Waterman, ed.), p. 154. CRC Press, Boca Raton, Florida, 1989.

[48] M. Zuker, *Science* **244,** 48 (1989).

Because recursive methods compute optimal foldings on subfragments, the folding of a RNA sequence as the sequence grows from the 5′ end can be simulated. Motifs appear and disappear as the simulated molecule grows into its final configuration. Substructures which form and are not destroyed as the molecule continues to grow may be regarded as significant. Modelevsky and Akers[49] monitor the percentage of bases in fixed windows (continuous segments of RNA) that are double stranded in a growing RNA sequence. The simulation uses RNAFOLD. Modelevsky and Akers correlate this statistic with the level of gene expression.

Le *et al.*[50] fold overlapping segments of RNA and then randomly permute the bases in that segment a number of times, refolding at each stage. This Monte Carlo method gives the mean and standard deviation of the folding energy which could be expected in that segment by chance alone. The segments used are not long; 100 bases is a typical length. This technique gives a basis for deciding which local folding motifs are significant. The size of these segments are varied, so that significant local structures of optimal size can be predicted. The method does not assess the significance of long-range interactions.

Another way to produce alternative structures is to perturb the folding rules and to repeat secondary structure prediction. This method is unnecessary with combinatorial or Monte Carlo folding programs for reasons already discussed, but this method is useful with traditional recursive programs, such as RNAFOLD, which have not been adapted to produce multiple solutions. The usual perturbation is to change the energy rules slightly and to refold the molecule a number of times. This method is as costly as the Monte Carlo assessment of folding significance developed by Le *et al.*[50] The reason is that the time-consuming fill algorithm must be performed over and over again. RNAFOLD has six "energy" parameters, which allow some energy assignments or some other folding rule changes to be made by the user without altering any input file. The energy parameters are set before folding commences with the command

$$\text{PE } i_1, j_1; i_2, j_2 \ldots$$

which has the effect of setting parameter i_1 equal to j_1, parameter i_2 equal to j_2, and so on. One or all parameters may be set with a single command. PCFOLD allows these parameters to be changed with the menu system, and GCGFOLD does not allow these parameters to be changed. Parameter 1 controls the existence of G–U base pairs at the ends of helices and has already been discussed. Parameters 2 through 5 inclusive are the extra energies, in tenths of a kilocalorie per mole, which are to be added to all

[49] J. L. Modelevsky and T. G. Akers, *CABIOS* 4(1), 161 (1988).
[50] S.-Y. Le, J.-H. Chen, K. M. Currey, and J. V. Maizel, Jr., *CABIOS* 4(1), 153 (1988).

base-pair stackings, bulge or interior loops, hairpin loops, or multiloops, respectively. These parameters are 0 by default. Parameter 6 is the maximum size allowed for bulge or interior loops. Parameter 6 is 30 by default. Although RNAFOLD has an automatic feature which causes the program to give up looking for arbitrarily large bulge or interior loops, this parameter does speed up program execution as the value is lowered. Thirty is a safe value for folding at 37°, but this value should be increased for folding at higher temperatures. Parameter 7 is the maximum value for |SIZE1–SIZE2|, where SIZE1 and SIZE2 are the number of single-stranded bases on the two sides of an interior or bulge loop. Parameter 7 can be called the maximum lopsidedness of an interior loop. Parameter 7 is also the maximum size of a bulge loop. Equal to 30 by default, this parameter takes effect only when this parameter is smaller than parameter 6. It is worthwhile to note that Ninio's CRUSOE program has a gradually increasing penalty for lopsided interior loops. This is a good idea and has already been incorporated into a trial version of RNAFOLD.

Altering the energy parameters does not produce random folding perturbations. The parameter changes bias the folding in predictable ways. Adding an extra 1–2 kcal/mol to all bulge or interior loops will usually produce a refolding with fewer bulge and interior loops. The addition of energy will tend to eliminate weakly stable helices and to consolidate single-stranded regions into fewer but larger loops. Adding a few extra kilocalories per mole to hairpin loops, say ε kcal/mol, will often result in a folding with fewer hairpin loops. If it does not, then the conclusion is that there are no alternative foldings with fewer hairpin loops within ε kcal/mol from the minimum energy. The same sort of analysis applies to multiloops. Unlike the situation with bonus energies used to force base pairs, these extra energies are included in the overall folding energy. Decreasing parameters 6 or 7 by one or two from the maximum size found in a first folding is another good way to perturb the original folding. Setting parameter 6 to 0 when folding a tRNA is an excellent way to force the desired cloverleaf structure, which is often not the minimum energy structure in the folding model. This brief description is meant to give the reader the idea of "playing" with the rules while refolding a sequence of interest. An example is given in Fig. 9.

Concluding Remarks

Three-dimensional prediction of RNA structure will not be possible for the foreseeable future. The problem is still unsolved for proteins, and in that case, there is at least a database of over 400 solved proteins, so that predictions can be compared with structures deduced from X-ray diffrac-

```
ENTER: T TERMINATE, NS NEW SEQUENCE, NF NEW FRAGMENT,
O OUTPUT PARAMETER DEFINITION, OR THE END POINTS OF A
SUBFRAGMENT BETWEEN      1 AND    126.
END WITH 1 TO FORCE ENDS TO BASE PAIR.
NF  ←
ENTER END POINTS OF FRAGMENT TO BE FOLDED. (DEFAULT  =  1,  126 )
1 126  ←
ENTER: F BEGIN FOLDING, T TERMINATE, A AUXILIARY INFORMATION,
P PARAMETER DEFINITION
PE 3, 10  ←
F  ←
   10   20   30   40   50   60   70   80   90  100  110  120

FOLDING BASES    1 TO  126 OF sample RNA
ENERGY  =     -27.9

                  10          20         30
--------  -      GUUUG     UUC       A    G   UU
        GC CUUUU        UCG   UGGCUG GUUC UGC
        UG GAGAA        AGC   ACCGAC UAAG ACG  A
CCGAAGAC  A      -----      CAU       G    -   CC
    120                          50          40

                        60             70          80
                        A      ----CU       U  -  CAUC
                         UCCGG        UGUGAU GC GC    C
                         AGGCC        ACACUA CG CG    C
                        -      AUGACU        C  A  AUAC
                      110         100          90
```

FIG. 9. Perturbation folding of the sample RNA shown folded in Figs. 1, 5, 6, and 7. A penalty of 1 kcal/mol is given to all bulge and interior loops using the command PE 3, 10. The result is an alternate folding, which differs in energy from the optimal folding by only 0.6 kcal/mol after the energy correction is made (7 kcal/mol). User input is indicated by the arrows.

tion experiments. Nevertheless, some three-dimensional aspects of RNA structure should be considered. The most important of these is the prediction of pseudoknots. In the absence of energy rules to deal with the complicated loops created by pseudoknots, and without ways of knowing whether hypothetical structures are stereochemically possible, the most sensible course of action is to use a two-stage procedure. First, predict an ordinary secondary structure and then look for tertiary interactions, which might induce pseudoknots at a final stage in the folding process. These tertiary interactions would be potential helices between single-

stranded regions leftover after the molecule is folded. The single-stranded regions could be enlarged by opening up short or weak helices in the secondary structure.

The treatment of enzymes and other chemicals, which indicate single- or double-stranded regions, has been crude to date. The method of forcing a helix is valid when there is phylogenetic evidence or cross-linking data. Some chemically modified bases cannot pair with certainty. However, very often the available data are saying that some bases are probably single stranded or probably double stranded. Forcing the folding one way or another, or even using the nucleotide accessibility option in RNAFOLD, is too heavy handed. What is needed is a multiple folding procedure in which bases likely to be single stranded would be so most of the time, or else a certain fraction of the time corresponding to a given probability. The same would hold for bases likely to be double stranded.

It is often desirable to predict a common folding of the leader sequences or coding regions of homologous mRNAs. The existence of a common folding greatly increases confidence in the predicted structure and can be used to explain the similar regulation of the related genes. Although Sankoff[51] reports an algorithm for simultaneously aligning and folding homologous RNA sequences, the method is not practical. Even if the related sequences are already aligned, there is no reliable and automatic way of predicting a common folding. The usual procedure is to use a program, such as SEQL, to find lists of potential helices in each sequence. These lists can be scanned for common helices. At this point, each sequence can be folded separately using an ordinary folding program, with the common helices forced. Depending on how many helices are forced, how similar the sequences are, and chance, the results will vary from a common folding of all sequences to dissimilar structures with some shared motifs.

Acknowledgment

M. Zuker is a Fellow of the Canadian Institute for Advanced Research.

[51] D. Sankoff, *SIAM J. Appl. Math.* **45**, 810 (1985).