Supporting Text

MiRNA Secondary Structure Statistical Test. The statistical significance of the folding of the miRNA precursor candidates was assessed by using a randomization test (1-3). Some have used similar methods to find significant secondary structures in RNA sequences (4-7). In our case, the value of the minimum free energy (MFE) of the sequence is compared with a distribution of values inferred from randomized sequences. Residues of the native sequence are randomized while preserving the dinucleotide distribution (5). The number of values smaller or equal to the MFE of the native sequence gives the probability that the free energy value can be obtained by chance. Under this model, no assumption is made on the nature of the distribution and the probability is straightforward to compute, as detailed below.

1. Compute the MFE of the secondary structure inferred from the original sequence.

2. Randomize the order of the nucleotides in the original sequence and compute the MFE for the inferred structure based on the shuffled sequence.

3. Repeat step 2 a great number of times (1,000) to build the distribution of MFE values.

4. If N is the number of iterations and R the number of randomized sequences that have an MFE value less or equal to the original value, then p is defined here as:

$$p = \frac{R}{N+1}$$

A test study was performed by using this procedure on the 506 miRNA of the RFAM registry (8, 9). The result was that >90% of the precursor sequences have a p value below 0.01, thus demonstrating that miRNA sequences clearly exhibit a folding free energy that is considerably lower than that for randomly shuffled sequences, indicating a high bias in the sequence toward a stable secondary structure. This result contrasts with other noncoding RNAs such as tRNAs and rRNAs that were proved to have folding free energies not significantly different from folding free energies of randomly generated sequences (5, 6). Therefore, computation of free energies of miRNAs was used to discard false-positive secondary structures in our computational pipeline by removing all putative precursor sequences that have a p value >0.01.

1. Efron, B. (1979) *Soc. Ind. Appl. Math.* **21,** 460-480.

2. Eddington, E. S. (1995) *Randomization Tests* (Dekker, New York).

3. Manly, B. F. J. (1997) *Randomization, Bootstrap, and Monte Carlo Methods in Biology* (Chapman & Hall, London).

4. Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V., Jr. (1988) *Comput. Appl. Biosci.* **4,** 153-159.

5. Workman, C. & Krogh, A. (1999) *Nucleic Acids Res.* **27,** 4816-4822.

6. Rivas, E. & Eddy, S. R. (2000) *Bioinformatics* **16,** 583-605.

7. Katz, L. & Burge, C. B. (2003) *Genome Res.* **13,** 2042-2051.

8. Griffiths-Jones, S. (2004) *Nucleic Acids Res.* **32,** D109-D111.

9. Bonnet, E. Wuyts, J., Rouzé, P. & Van de Peer, Y. (2004) *Bioinformatics*, in press.