Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 1239–1243, March 1987
Biophysics

Phase transitions in sequence matches and nucleic acid structure

MICHAEL S. WATERMAN*[†], LOUIS GORDON*, AND RICHARD ARRATIA*

Departments of *Mathematics and [†]Molecular Biology, University of Southern California, Los Angeles, CA 90089-1113

Communicated by Gian-Carlo Rota, October 1, 1986

ABSTRACT Analyses of phase transitions in biopolymers have previously been restricted to studies of average behavior along macromolecules. Extremal properties, such as longest helical region, can now be studied with a new family of probability distributions [Arratia, R., Gordon, L. & Waterman, M. S. (1986) *Ann. Stat.* 14, 971–993]. Not only is such extremal behavior analyzed with great precision, but new phase transitions are determined. One phase transition occurs when behavior of the free energy of the longest helical region abruptly changes from proportional to logarithm of the sequence length to proportional to sequence length. The annealing of two single-stranded molecules and the melting of a double helix are both considered. These results, initially suggested by studies of optimal matching of random DNA sequences [Smith, T. F., Waterman, M. S. & Burks, C. (1985) *Nucleic Acids Res.* 13, 645–656], also have importance for significance tests in comparison of nucleic acid or protein sequences.

A great deal of effort has been devoted to the study of helix-coil transitions in biopolymers, especially in nucleic acids (1). Analogy to the Ising model is often made. Recently, some new results have been obtained in the theory of probability that generalize a 1970 theorem of Erdos and Renyi (2) on the length of the longest run of successes in coin-tossing and that apply in a novel way to these biomolecular problems. The key difference is that earlier work on biopolymers dealt with the average behavior along a long, linear molecule. These new ideas allow us to study with great precision extremal properties, such as longest base-paired region, occurring somewhere along the linear sequence. The results apply to situations where two linear sequences can slide or shift along each other to achieve minimal free energy, as well as to the case of a double helix where the bases are initially base-paired. In addition, natural extensions to higher dimensional situations have been obtained (3) and describe a variety of other phenomena such as interactions between surfaces, the probability distribution of galactic clusters in space (3), and clumping of plants in a field.

Our original motivation was to give the probability distribution of the longest exact or approximate matching between two random DNA or protein sequences. The first result, here referred to as an example of the "log(n) law", was inferred from a data analysis of sequences from GenBank (4) with a dynamic programming algorithm (5). Since then extensions and generalizations have been obtained by ourselves and others (6, 7). Many of the results cited here have not been reported elsewhere.

We will first present probability results that have wide applicability. Then we will turn to two analogous problems: (*i*) helix-coil formation between two random single-stranded DNA chains and (*ii*) comparison of two random DNA sequences for sequence similarity. Both the cases of fixed positions and of possible shifts are treated and the phase transitions are described.

The celebrated Erdos–Renyi law (2) gave order-of-magnitude behavior for the longest run of heads in a sequence of n coin tosses. Their results actually include behavior of the longest head run containing $(1 - \alpha) \times 100\%$ tails, where $\alpha > P(H) = p$. For length R_n of pure head runs ($\alpha = 1.0$) their result is

$$R_n/\log_{1/p}(n) \rightarrow 1 \text{ with probability one,}$$

while for general $\alpha > p$ their result is

$$R_n/[\log(n)/H(\alpha, p)] \rightarrow 1 \text{ with probability one,}$$

where $H(\alpha, p) = \alpha \log(\alpha/p) + (1 - \alpha) \log[(1 - \alpha)/(1 - p)]$ is relative entropy. For $\alpha = 1$, $H(\alpha, p) = \log(1/p)$ and the results are consistent.

The log(n) law suggests that in $512 = 2^9$ tosses of a fair coin ($p = 1/2$) the longest run of pure heads should be approximately $\log_2(2^9) = 9$. Other work (8) has derived precise results for this law and gives

$$\text{mean } R_n = \log_{1/p}(n) + \frac{(0.577\ldots)}{\theta} - \frac{1}{2} + r_1(n)$$

and

$$\text{variance } R_n = \frac{\pi^2}{6\theta^2} + \frac{1}{12} + r_2(n),$$

where 0.577 ... is the Euler–Mascheroni constant, $\theta = \ln(1/p)$, and $r_1(n)$ and $r_2(n)$ are negligible for large n. For the 512 fair coin tosses, the mean ≈ 9.33 and the standard deviation ≈ 1.93.

The formulas above give useful results for longest base-paired region in a double helix or for longest match between two sequences with fixed alignment. At temperature T, the ensemble average of base pairs in a helix is $p = p(T)$ where $0 < p < 1$. For any $\alpha > p$, the longest region with at least $\alpha \times 100\%$ base pairs is given by R_n.

Two random DNAs in solution, however, are not already in a double helix and they will form the structure that has minimal free energy. Analogously in sequence matching, the two regions of sequence with best matching are often the object of search. In these cases there are usually unpaired or mismatched bases. An extension of the log(n) law to sequence matching is now given that incorporates mismatches. A corresponding statement for longest base-paired region interrupted by unpaired bases differs only in language.

Two sequences of length n are assumed to have bases chosen independently and identically with $p = P$ (two bases match) $= p_A^2 + p_C^2 + p_G^2 + p_T^2$. The mean length of the longest match with k mismatches becomes (6)

$$\log(qn^2) + k \log\log(qn^2) + k \log(q/p) - \log(k!) + k + \frac{(0.577\ldots)}{\theta} - \frac{1}{2} + r_1(qn^2),$$

where $q = 1 - p$ and all logarithms are taken to base $1/p$. The variance remains

$$\frac{\pi^2}{6\theta^2} + \frac{1}{12} + r_2(qn^2).$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

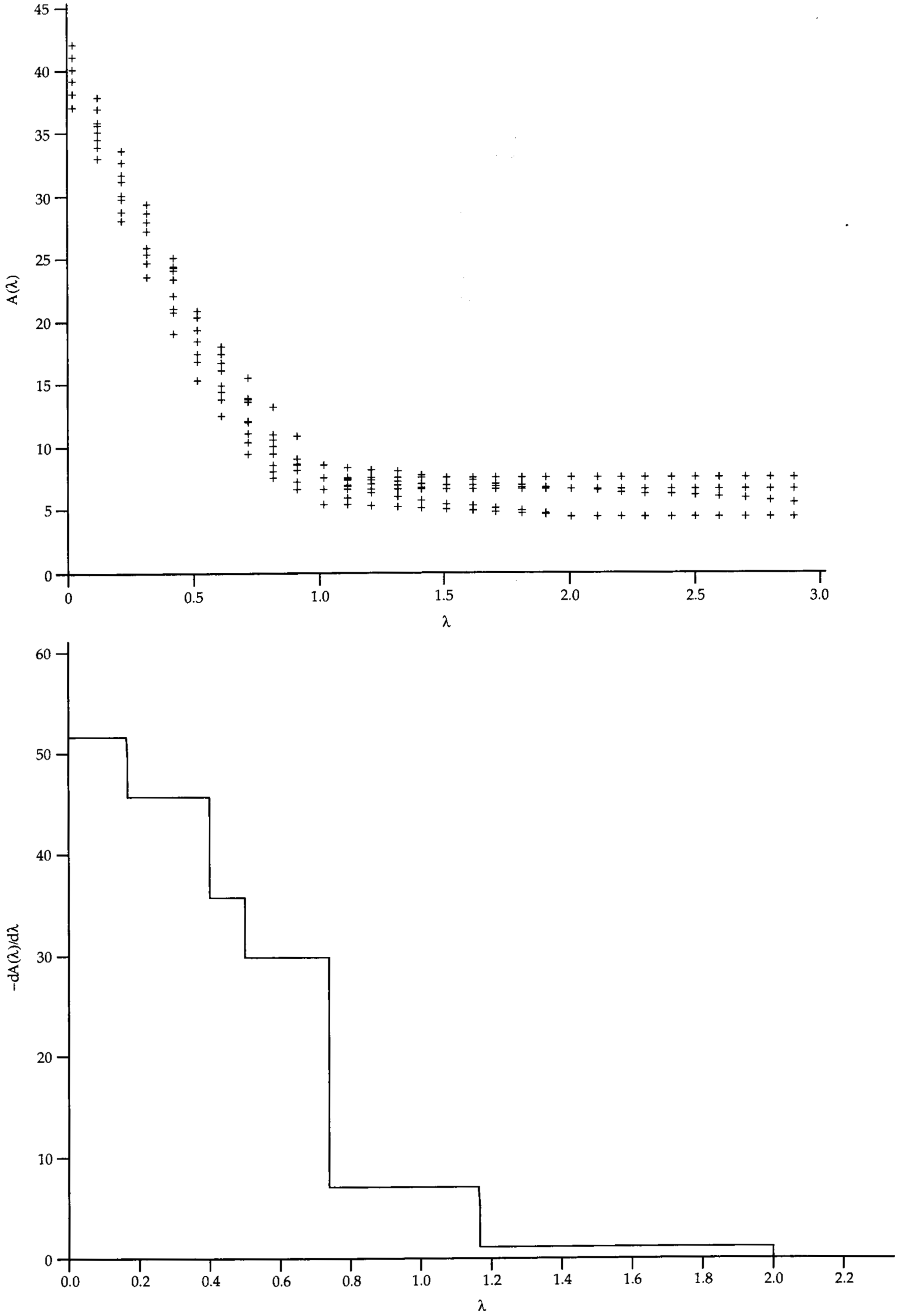

FIG. 1. (*Upper*) The values of $A(\lambda)$ as a function of λ for 10 pairs of random DNAs of length 64. The expected value of $A(\infty)$ is 5.71 . . . . The approximate value of λ_{cr} from this simulation is $0.75 \leqq \lambda_{cr} \leqq 0.86$. The function $A(\lambda)$, for each λ, is the maximum of over 10^{40} linear functions, each of which results from a distinct alignment. $A(\lambda)$ is found by a dynamic programming algorithm (5). (*Lower*) For the two length 64 sequences

CGTTGGTGTGAAAATATGGTGACATTCCATCAGGTCGCTGTTCTTGATGACATTCAAATCGAAT

and

CTCCCGAGTAACTGGAAGATCGTGGTACGGCCCGATACCCAAGCCCTAGAGTTAGTGAGGCCCT,

the graph shows $-A'(\lambda) = dA(\lambda)/d(\lambda)$. Each constant portion of the graph corresponds to a collection of alignments or structures that remain optimal as λ varies.

```
a      CGTT--G-GTG--TG-AA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGAA----T
       | |   | ||   || || | |  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||     |
       C-TCCCGAGTAACTGGAAGA-TCGTGGT-AC--GGCCCGATACCCAAG--C-C----CT--A-GA-GTT-A-GT-GAGGCCCT

b      CGTT--G-GTG--TG-AA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGA
       | |   | ||   || || | |  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||
       C-TCCCGAGTAACTGGAAGA-TCGTGGT-AC--GGCCCGATACCCAAG--C-C----CT--A-GA-GTT-A-GT-GA

c      TGTGAAAATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGA
       || ||| ||  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||
       TG-GAAGATCGTGGT-AC--GGCCCGATACCCAAG--C-C----CT--A-GA-GTT-A-GT-GA

d      TGTGAA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGC-T-G--TTCTTGA-TGA
       || ||| | |  |||| ||    ||  || |  | |  | | | | ||     | |||
       TG-GAAGA-TCGTGGT-AC--GGCCCGATACCCAAG--C-CCTAGAGTT----AGTGA

e      TGTGAA-AATA-TGGTGAC
       || ||| | |  |||| ||
       TG-GAAGA-TCGTGGT-AC

f      AA-ATCG
       || ||||
       AAGATCG

g      ATCG
       ||||
       ATCG
```

FIG. 2. For the sequences of Fig. 1 *Lower*, seven representative alignments are shown for the constant values of $-A'(\lambda)$. Alignment a, for example, is an alignment that is maintained from $\lambda = 0$ to $\lambda = 0.166$; b, from $\lambda = 0.166$ to $\lambda = 0.400$; c, from $\lambda = 0.400$ to $\lambda = 0.500$; d, from $\lambda = 0.500$ to $\lambda = 0.739$; e, from $\lambda = 0.739$ to $\lambda = 1.66$; f, from $\lambda = 1.66$ to $\lambda = 2.000$; and g, from $\lambda = 2.000$ to $\lambda = \infty$.

The Erdos–Renyi law for the length R_n of the longest 100% head run of n coin tosses then extends to a law for the length M_n of the longest match between two sequences (9). We have recently shown that the length of the longest run of matches containing $(1 - \alpha) \times 100\%$ mismatches satisfies

$$M_n/[\log(n^2)/H(\alpha, p)] \rightarrow 1 \text{ with probability one,}$$

which is the Erdos–Renyi law with n replaced by n^2. This last theorem has been obtained by use of the theory of large deviations.

Considering these results, it is not surprising that log(n) laws hold far beyond the longest exact head run or match. To quantify this for random sequences $\mathbf{X} = x_1, x_2 \ldots x_n$ and $\mathbf{Y} = y_1, y_2 \ldots y_n$, we study

$$A(\lambda) = \max_{\substack{I \subset \mathbf{X} \\ J \subset \mathbf{Y}}}\{(\text{no. of matched bases between } I \text{ and } J) - \lambda \text{ (no. of unmatched bases between } I \text{ and } J)\},$$

where $I(J)$ range over all contiguous regions of $\mathbf{X}(\mathbf{Y})$. For $\lambda = \infty$, mismatches are not allowed and the log(n) law holds. What about the other extreme, $\lambda = 0$, where nonmatched bases receive no penalty? Here $A(0)$ = the length of longest subsequence common to both sequences. The distribution $A(0)$ has been much studied, beginning with Chvatal and Sankoff (10). It is known that $A(0) \approx a{\cdot}n$, where a is a constant that has not yet been precisely determined. Except for $A(0)$ and $A(\infty)$ this function has not been studied. Our most important new result about $A(\lambda)$ concerns a phase transition that $A(\lambda)$ undergoes. This phase transition is described next.

The function $A(\lambda)$ can easily be shown to be continuous and decreasing. In Fig. 1 *Upper*, $A(\lambda)$ is plotted for 10 pairs of random DNAs of length 64. With some additional effort $A'(\lambda) = dA(\lambda)/d\lambda$ can be seen to be increasing, nonpositive, and piecewise constant with jump discontinuities. In Fig. 1 *Lower*, an example of $A'(\lambda)$ is shown for one of the sequence pairs from Fig. 1 *Upper*. The locations of these jump discontinuities for random sequences of length n tend to cluster as n becomes large. This behavior can be described as phase transitions. In particular, there is a value of λ, λ_{cr}, where behavior of $A(\lambda)$ abruptly changes from linear to logarithmic. This can be seen in Fig. 1 *Upper* where the curve changes from its sharp decline to almost horizontal. For n large enough,

$$0 \leqq \lambda < \lambda_{\text{cr}}: \quad A(\lambda) \approx a_\lambda{\cdot}n$$
$$\lambda_{\text{cr}} < \lambda \leqq \infty: \quad A(\lambda) \approx b_\lambda{\cdot}\log(n).$$

The above discussion gives $b_\infty = 2$ if the logarithm is to the base $1/p$, and it is known that $0.45 < a_0 < 0.77$ for the case of an equally probable four-letter alphabet. Examples of alignments from the sequences of Fig. 1 *Lower* are given in Fig. 2. In Fig. 3, the second sequence of Fig. 2 is complemented and the matches are converted to base pairs.

More detailed information can be obtained about these rates of growth. For $\lambda_{cr} < \lambda$, the function $A(\lambda) \approx b_\lambda \log(n) + c_\lambda \operatorname{loglog}(n)$. If only mismatches are allowed (no insertions or deletions), then b_λ can be found:

$$b_\lambda = 2/H(\alpha, p),$$

$$\text{with } \lambda = \frac{-H(\alpha, p) + \alpha dH(\alpha, p)/d\alpha}{H(\alpha, p) + (1 - \alpha)dH(\alpha, p)/d\alpha}$$

Essentially this is the λ that gives $\alpha \times 100\%$ matches.

The expected behavior of $A(\lambda)$ is of importance in evaluating sequence comparisons. If a located match is at or below that expected from random sequences of similar composition, then the match should not be further considered without additional biological information. Since these distributions

```
CGTT--G-GTG--TG-AA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGAA----T      a
| |   | ||   || || | |  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||     |
G-AGGGCTCATTGACCTTCT-AGCACCA-TG--CCGGGCTATGGGTTC--G-G----GA--T-CT-CAA-T-CA-CTCCGGGA

CGTT--G-GTG--TG-AA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGAAT          b
| |   | ||   || || | |  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||
G-AGGGCTCATTGACCTTCT-AGCACCA-TG--CCGGGCTATGGGTTC--G-G----GA--T-CT-CAA-T-CA-CTCCGGGA

      CGTTGGTGTGAAAATA-TGGTGACATT-CC--AT-C--A-GGTCGCTGTTCTTGATGACATTCAAATCGAAT       c
            || ||| ||  |||| ||    ||  || |  | |  | |    ||  | ||  || |  | ||
GAGGGCTCATTGAC-CTTCTAGCACCA-TG--CCGGGCTATGGGTTC--G-G----GA--T-CT-CAA-T-CA-CTCCGGGA

      CGTTGGTGTGAA-AATA-TGGTGACATT-CC--AT-C--A-GGTCGC-T-G--TTCTTGA-TGACATTCAAATCGAAT d
            || ||| | |  |||| ||    ||  || |  | |  | | | |  ||    | |||
GAGGGCTCATTGAC-CTTCT-AGCACCA-TG--CCGGGCTATGGGTTC--G-GGATCTCAA----TCACTCCGGGA

      CGTTGGTGTGAA-AATA-TGGTGACATTCCATCAGGTCGCTGTTCTTGATGACATTCAAATCGAAT             e
            || ||| | |  |||| ||
GAGGGCTCATTGAC-CTTCT-AGCACCA-TGCCGGGCTATGGGTTCGGGATCTCAATCACTCCGGGA

CGTTGGTGTGAAAATATGGTGACATTCCATCAGGTCGCTGTTCTTGATGACATTCAA-ATCGAAT                    f
                                                       || ||||
                                        GAGGGCTCATTGACCTTCTAGCACCATGCCGGGCTATGGGTTCGGGATCTCAATCACTCCGGGA

CGTTGGTGTGAAAATATGGTGACATTCCATCAGGTCGCTGTTCTTGATGACATTCAAATCGAAT                     g
                                                         ||||
                                       GAGGGCTCATTGACCTTCTAGCACCATGCCGGGCTATGGGTTCGGGATCTCAATCACTCCGGGA
```

FIG. 3. In this figure, the second sequence of Fig. 2 is complemented, converting matches to base pairs. The sequences become

CGTTGGTGTGAAAATATGGTGACATTCCATCAGGTCGCTGTTCTTGATGACATTCAAATCGAAT

and

GAGGGCTCATTGACCTTCTAGCACCATGCCGGGCTATGGGTTCGGGATCTCAATCACTCCGGGA.

The resulting structures appear in alignments a–g.

have been shown to fit biological sequences quite well (4), highly significant matching regions might merit additional study. Until now this paper has given only the simplest matching function. One in common use (5) extends $A(\lambda)$ to

$$B(\mu, \delta) = \max_{\substack{I\subset X\\ J\subset Y}}\{(\text{no. of matches}) - \mu(\text{no. of mismatches}) - \delta(\text{no. of deletions})\},$$

where μ is the mismatch penalty and δ the deletion penalty. We have also proven that $B(\mu, \delta)$ undergoes a phase transition. The $\log(n)$ and linear regions of this two-dimensional parameter space have been determined numerically in a Monte Carlo study to be published elsewhere. In Fig. 4 the log and linear regions are shown on a phase diagram. These results help those analyzing macromolecular sequences to proceed in a much less *ad hoc* manner.

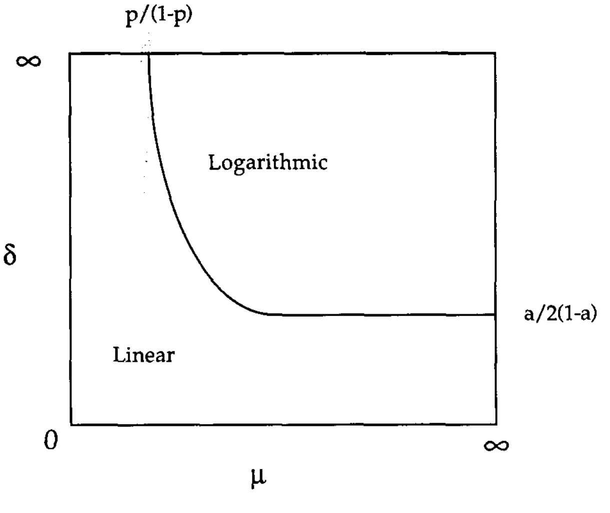

FIG. 4. The phase diagram for $B(\mu, \delta)$. The line $\mu = 2\delta$ corresponds to $\delta = \lambda$, so that $B(2\delta, \delta) = A(\delta)$. Notice that for $\mu > 2\delta$, $B(\mu, \delta) = B(2\delta, \delta)$. The constant a is the Chvatal–Sankoff constant discussed in the text.

As pointed out earlier, $A(\lambda)$ for sequence matching is analogous to the minimum free energy of the base-paired regions for helices (see Fig. 3). A simple thermodynamic measure of the free energy, $G(\lambda)$, of a region of base pairing is given by $-$(no. of base pairs) + λ(no. of unpaired bases). In this simple model, $G(\lambda) = -A(\lambda)$ and all results about $A(\lambda)$ carry over to the function $G(\lambda)$. Refining $G(\lambda)$ to have differing free energies for specific base pairs and to include bulge and interior loop destabilizations is easy conceptually and practical to carry out with an extension of the existing algorithm for DNA sequence comparisons and RNA secondary structure prediction (11). Such a numerical study has not yet been carried out, but the mathematical results described above can be extended to show the same array of phase transitions with change from linear to logarithmic behavior.

We thank Mark Eggert for programming and computer assistance and David Galas for comments on the manuscript. This work was supported by grants from the System Development Foundation (M.S.W.), the National Institutes of Health (M.S.W.), the National Science Foundation (R.A.), and the Institute for Mathematics and Its Applications (R.A.).

1. Poland, D. & Scheraga, H. A. (1970) *Theory of Helix-Coil Transitions in Biopolymers* (Academic, New York).
2. Erdos, P. & Renyi, A. (1970) *J. Anal. Math.* **22,** 103–111.
3. Darling, R. W. R. & Waterman, M. S. (1986) *SIAM J. Appl. Math.* **46,** 118–132.
4. Smith, T. F., Waterman, M. S. & Burks, C. (1985) *Nucleic Acids Res.* **13,** 645–656.
5. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.

6. Arratia, R., Gordon, L. & Waterman, M. S. (1986) *Ann. Stat.* **14,** 971–993.
7. Karlin, S., Ghandour, G., Ost, F., Tavare, S. & Korn, L. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5660–5664.
8. Gordon, L., Schilling, M. & Waterman, M. S. (1986) *Probab. Theor. Rel. Fields* **72,** 279–287.
9. Arratia, R. & Waterman, M. S. (1985) *Adv. Math.* **55,** 13–23.
10. Chvatal, V. & Sankoff, D. (1975) *J. Appl. Prob.* **12,** 306–315.
11. Zuker, M. & Sankoff, D. (1984) *Bull. Math. Biol.* **46,** 591–621.