ADVANCES IN APPLIED MATHEMATICS **18,** 111-132 (1997)
ARTICLE NO. AM960502

A Phase Transition for the Minimum Free Energy of Secondary Structures of a Random RNA*

Momiao Xiong and Michael S. Waterman

Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113

Received July 14, 1996

The free energy of a single-stranded RNA can be calculated by adding the free energies of the components: basepairs, bulges, and loops. Basepairs receive negative free energy while the unpaired bases receive positive free energy. The minimum free energy of a random RNA secondary structure with one domain has value F_n where the sequence length is n. Under simplifying assumptions, we show that for "small" values of bulge and loop penalties F_n has linear growth in n, while for "large" values of these parameters F_n has logarithmic growth in n. This phase transition generalizes results obtained for the local-alignment score of two random sequences. The random variable F_n is conjectured to have a Poisson approximation. The multi-domain secondary structure minimum free energy E_n has linear growth in n for all values of the penalty functions. Nothing more is known about the distributional properties of E_n.

1. INTRODUCTION

A ribonucleic acid (RNA) molecule is a chain of covalently bound molecules called ribonucleotides. There are four ribonucleotides, determined by their bases: A (adenine), C (cytosine), G (guanine), or U (uracil). For our purposes, an RNA is a word over this four-letter alphabet. An RNA is copied from a strand of DNA where a T in DNA corresponds to a U in RNA. RNA molecules are single-stranded and fold onto themselves to form basepairs. Structures for tRNA, 5SRNA, and 16SRNA are well known. The folded structure that is assumed in the cell determines the biological function of the molecule so that the structure assumed by a molecule is important. In addition, predicting the two- or three-dimen-

* Supported by grants from the National Institute of Health (GM 36230) and the National Science Foundation (DMS 90-05833).

0196-8858/97 $25.00

sional structure from the sequence of nucleotides is far from routine. First we shall discuss structure in more detail.

Let the single-stranded RNA be represented as $\mathbf{A} = A_1 \cdots A_n$ (for example, **A** = CAUAUGUUUACAAAUG), which is called the primary structure. Of course each $A_i \in \{A, C, G, U\}$. These bases can form basepairs, where conventionally A pairs with U and C pairs with G. In addition, the pairing of G and U is frequently allowed. If A_i pairs with A_j, then $|i - j| > 1$. Under normal physiological conditions, a ribonucleotide chain can fold back on itself, and the basepairs then form. We define *secondary structure* to be a planar graph (where vertices are bases and edges are basepairs) that satisfies the following condition: If A_i pairs with A_j and A_k is paired with A_l with $i < k < j$, then $i < l < j$ (Waterman, 1978). The secondary structure may also be represented by a list **P** of pairs, where (i, j) is in **P** if and only if A_i and A_j form a basepair. The pair itself will sometimes be referred to as $i \cdot j$. The secondary structure for the RNA sequence **A** is implied by **P** and can be described as being composed of substructures of the following types: helices, end loops, bulges, interior loops, multi-loops, and external single-stranded regions. The secondary structure assumed in the solution is one of those that has minimum free energy. Free energy is a thermodynamic constant that gives the amount of energy required for or released by a reaction. Structures such as loops and bulges that require energy have a positive value. Structures such as basepairs that release energy have negative value. We assume the following functions give the free energy associated with substructures:

$\xi(k)$	destabilization free energy of an end loop of k bases,
$\beta(k)$	destabilization free energy of a bulge of k bases,
$\gamma(k)$	destabilization free energy of an interior loop of k bases,
$\rho(k)$	destabilization free energy of k unpaired bases in a multi-branch loop,
$s(a_i, b_j)$	free energy of basepair (a_i, b_j).

To simplify our discussion in this paper, we assume that the destabilization free energy functions are non-negative and have the following forms:

$$\xi(k) = \tau k,$$
$$\beta(k) = \lambda k,$$
$$\rho(k) = \phi k,$$
$$\gamma(k) = \psi k.$$

For the compatibility of the terminology with DNA sequence alignment, the free energy of an RNA secondary structure will hereafter be called simply by its score or free-energy score.

Figure 1 gives a simple example of RNA secondary structure. Assume that we score the matched pair of GC or AU by -1 and that the free energy of the various elements of RNA secondary structure are given by the above linear functions. Then the score of this RNA secondary structure is $-23 + \lambda + 6\psi + 3\phi + 11\tau$.

Experimental determination of RNA structure is extremely difficult so scientists often predict structure from the linear sequence $\mathbf{A} = A_1 A_2 \cdots A_n$. One of the most popular methods for predicting secondary structure is dynamic programming, first presented by Waterman (1978), Waterman and Smith (1978), and Nussinov et al. (1978). Zuker and Sankoff (1984) provide an excellent review. Waterman and Smith (1986) propose some speedups of this method. Sankoff (1985) considers simultaneous alignment and secondary-structure prediction. Dynamic programming is still a method of choice for secondary-structure prediction although computation time can be limiting. In Section 5, we use dynamic programming to produce the minimum free-energy scores F_n for simulated sequences.

Because computer programs are used to predict biological structures, there are very natural questions about their reliability. After all, a program produces a structure for any sequence, real or not. We are studying only one aspect of this general question here: How does the computed minimum free-energy score F_n compare with that from a random RNA sequence? Gralla and DeLisi (1974) first pointed out how much secondary structure exists in a random RNA; the implication is that it is easy to be fooled into thinking a folded RNA is the result of natural selection and therefore real. In the years since Gralla and Delisi's work not much progress has been made on the problem of finding the statistical distribu-

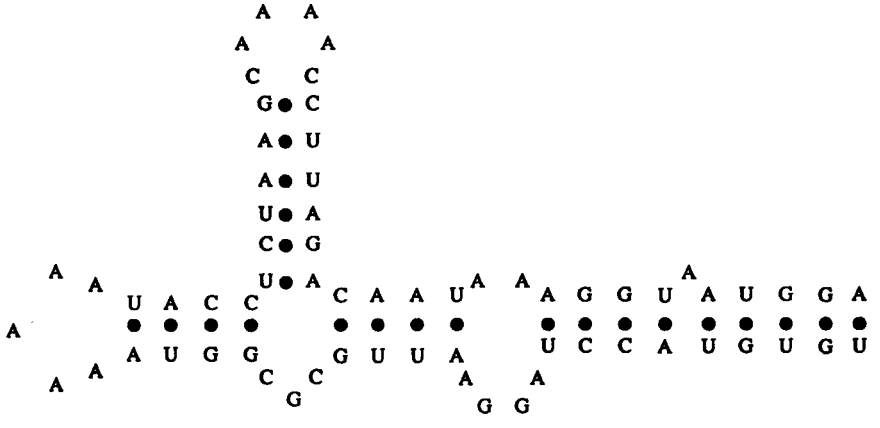


FIG. 1. An example of RNA secondary structure.

tion of the minimum free energy secondary structure of a random RNA. Maizel and collaborators (Le *et al.*, 1988) have a heuristic approach to determining the statistical significance of F_n, but their approach has serious flaws. For each interval of length W, $A_i A_{i+1} \cdots A_{i+W-1}$, they compute the $F(i)$ = minimum free energy of $A_i A_{i+1} \cdots A_{i+W-1}$. The mean and standard deviation of F is found by simulation and the "statistical significance" of $F(i)$ is estimated by the number of standard deviations $F(i)$ is above or below the mean.

There are several difficulties with this approach. $F(i)$ is assumed to be normally distributed which it almost surely is not because it is the result of taking the minimum over all secondary structures. (For a fixed structure, the free energy of a random sequence is of course approximately normal by the central limit theorem.) In addition there is the multiple-hypothesis testing fallacy: If you test 100 hypotheses at the 5% level, you should expect five hypotheses to be rejected *under the null hypothesis being true*. The same objection holds with Maizel's approach, and the dependence of overlapping windows makes a theoretical analysis challenging.

For problems of estimating statistical significance such as we have just described, the powerful method of Chen–Stein approximation has recently been developed (Arratia *et al.*, 1989). There have been applications to alignment scores where the asymptotic behavior of alignment scores of global and local sequence-comparisons have been studied. Large deviation results for local DNA sequence-comparisons and Poisson approximations were obtained, for example, by Arratia and Waterman (1985a), Karlin and Ost (1987), Arratia *et al.* (1990), Karlin and Dembo (1992), Arratia and Waterman (1994), Goldstein and Waterman (1994), Waterman (1994), Waterman and Vingron (1994ab), and Neuhauser (1994).

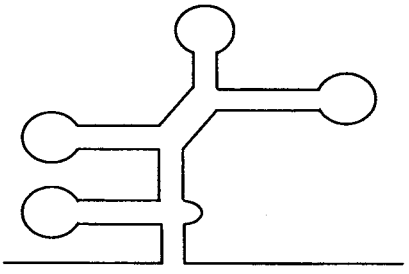
These results depend on positive local-alignment scores having small probability. The phenomenon of phase transitions of local-alignment scores between linear score growth in n, when the penalty parameters are small, and logarithmic growth in n, when the penalties are large, was announced by Waterman *et al.* (1987) and rigorously proved by Arratia and Waterman (1994). The logarithmic region is the realm of large deviations. It is conjectured that Poisson approximation is valid in the logarithmic region of parameters, and numerical results are presented in Waterman and Vingron (1994a, b) to support this conjecture.

In this paper we generalize the Arratia–Waterman result to the case of free energy for RNA. In Section 2 we establish subadditivity of a free energy score S_n and then in Section 3 we show that for "large" values of bulge and loop penalties F_n has logarithmic growth in n. In Section 4 we prove the phase transition result for a special case of F_n. In Section 5 we give a numerical estimate of the phase transition curve and some conjectures. Generally our method of proof follows Arratia and Waterman

(1994), but it is necessary to carefully check the details as there are some key differences between alignment and free energy. As we point out in Section 5, there is reason to believe that a rigorous proof of Poisson approximation in the logarithmic region will be easier for the free-energy score than for local-alignment score.

2. SUBADDITIVE THEORY

In this section we establish some facts that are basic to the proof of the phase transition. Our techniques will only allow us to prove a phase transition for a restricted definition of F_n, the minimum free-energy score over all structures with "one domain". By this we mean that there is an $i \cdot j$ pair where $A_i \cdots A_{i-1}$ and $A_{j+1} \cdots A_n$ have no basepairs. This includes the structure



but not structures such as

Therefore we will prove subadditivity with an energy function for our one-domain case. Let $A_{g+1} \ldots A_{g+i}$ be an RNA sequence with $1 \leq g + 1 \leq g + i \leq n$. In this i-letter sequence, let u be the total number of bases in the end loops; of course, this total of u bases in end loops is the sum of the bases in individual end loops, $u = \sum u_j$. Then because $\xi(u) = \tau u$,

$\xi(u) = \sum \xi(u_j)$. Similarly let w be the total number of bases in bulges, with $w = \sum w_j$ and $\beta(w) = \sum \beta(w_j)$; let m be the total number of bases in interior loops, with $m = \sum m_j$ and $\rho(m) = \sum \rho(m_j)$; and finally let v be the total number of unpaired bases in multi-branch loops with $v = \sum v_j$ and $\gamma(v) = \sum \gamma(v_j)$. Usually, the total free energy of the secondary structure is the sum of free energy of substructures but here there is a small modification. The score function $S(A_{g+1} \dots A_{g+i})$ of an RNA sequence $A_{g+1} \dots A_{g+i}$ is defined as the minimum free-energy score of its folded secondary structures, i.e.,

$$S(A_{g+i} \dots A_{g+i}) = \min\left\{ pen(\Delta) + \xi(u) + \beta(w) + \gamma(m) + \rho(v) + \sum_{k=1}^{l} s(A_{a(k)}, A_{b(k)}) \right\}, \quad (1)$$

where $a(k) < a(k+1)$ (so $b(k+1) < b(k)$) for all i, $\Delta = i - u - w - m - v - 2l$, $pen(\Delta) = \max\{\xi(\Delta), \beta(\Delta), \gamma(\Delta), \rho(\Delta)\}$. This definition puts the Δ bases not accounted for into the least-favorable energy conformation.

Let

$$S_k = S(A_1 \dots A_k),$$

and

$$S_{k+l} = S(A_1 \dots A_k A_{k+1} \dots A_{k+l}).$$

Because the secondary structures assumed by sequences $A_1 \dots A_k$ and $A_{k+1} \dots A_{k+l}$ are contained in the possible secondary structures of the sequence $A_1 \dots A_k A_{k+1} \dots A_{k+l}$, and the score function S_{k+l} is the minimum over all possible secondary structures assumed by $A_1 \dots A_{k+l}$, we have

$$S_{k+l} \le S_k + S(A_{k+1} \dots A_{k+l}).$$

Due to the assumption of iid letters, from this equation it follows that the expectation S_{k+l} is subadditive:

$$E[S_{k+l}] \le E[S_k] = E[S_l],$$

and in addition

$$P(S_{k+l} \le q(k+l)) \ge P(S_k \le qk) P(S_l \le ql). \quad (2)$$

Subadditivity implies the deterministic limit of the expectations exists and equals the infimum

$$a = \lim_{k \to \infty} \frac{ES_k}{k} = \inf_{k \ge 1} \frac{ES_k}{k}.$$

Furthermore, Kingman's subadditive ergodic theorem (Kingman, 1973) implies the stochastic limit holds with probability 1 and in L_2:

$$a = \lim_{k \to \infty} \frac{S_k}{k}. \tag{3}$$

In order to study this problem in the simplest setting without much loss of generality, we proceed as follows. Score a letter in bulges by δ, and all other unpaired letters except bulges by μ, and finally $G \cdot C$ and $A \cdot U$ pairs by -1. The parameter space is $(\mu, \delta) = [0, \infty]^2$ Then our RNA secondary structure alignment score takes the following form:

$$S(A_{g+1}, \ldots, A_{g+i}) = \min\left\{ \delta(i - 2l - m) + m\mu + \sum_{k+1}^{l} s(A_{a(k)}, A_{b(k)}) \right\}, \tag{4}$$

where m is the number of unpaired letter not in bulges and $a(k)$ and $b(k)$ are defined as in (1). This corresponds to a global sequence alignment score. We have reduced the number of parameters to two, for simplicity. Since the score function $S_k = S(A_1 A_2 \ldots A_k)$ is now a function of the parameters δ and μ, we denote a by $a(\mu, \delta)$.

Next we show that $\{(\mu, \delta) : a(\mu, \delta) = 0\}$ defines a curve that separates the positive region $\{a(\mu, \delta) > 0\}$ and negative region $\{a(\mu, \delta) < 0\}$. Later we will show that this curve is a phase transition curve.

LEMMA 1. *The set* $\{(\mu, \delta) : a(\mu, \delta) = 0\}$ *defines a line in the parameter space* $[0, \infty]^2$, *separating the negative and positive regions* $\{a < 0\}$ *and* $\{a > 0\}$.

Proof. The proof of this lemma proceeds by showing that $a(\mu, \delta)$ is continuous and strictly monotone in the $(1, 1)$ direction. Let $\delta_1 > \delta_2$ and $\mu_1 > \mu_2$. Let $M_k(\mu, \delta)$, $D_k(\mu, \delta)$ and $U_k(\mu, \delta)$ be the number of pairs, of letters in bulges, and of unpaired letters not in bulges, respectively, in an optimal alignment for $S_k(\mu, \delta)$. Apparently,

$$\begin{aligned} S_k(\mu, \delta) &= \delta D_k(\mu, \delta) + \mu U_k(\mu, \delta) - M_k(\mu, \delta); \\ k &= D_k(\mu, \delta) + U_k(\mu, \delta) + 2M_k(\mu, \delta). \end{aligned} \tag{5}$$

Then

$$\begin{aligned} S_k(\mu, \delta_1) &= \delta_1 D_k(\mu, \delta) + \mu U_k(\mu, \delta_1) - M_k(\mu, \delta_1) \\ &\geq \delta_2 D_k(\mu, \delta_1) + \mu U_k(\mu, \delta_1) - M_k(\mu, \delta_1) \\ &\geq S_k(\mu, \delta_2). \end{aligned}$$

It follows from this that

$$a(\mu, \delta_1) = \lim_{k\to\infty} \frac{ES_k(\mu, \delta_1)}{k} \geq \lim_{k\to\infty} \frac{ES_k(\mu, \delta_2)}{k} = a(\mu, \delta_2).$$

Similarly we have

$$a(\mu_1, \delta) \geq a(\mu_2, \delta). \tag{6}$$

This shows that $a(\mu, \delta)$ is non-decreasing in each of its parameters. It is easy to see that

$$\begin{aligned} S_k(\mu, \delta) + \epsilon k &\geq (\delta + \epsilon)D_k(\mu, \delta) + (\mu + \epsilon)U_k(\mu, \delta) - M_k(\mu, \delta) \\ &\geq S_k(\mu + \epsilon, \delta + \epsilon). \end{aligned}$$

After taking expectation and limits on both sides of this equation, we obtain

$$a(\mu, \delta) + \epsilon \geq a(\mu + \epsilon, \delta + \epsilon). \tag{7}$$

Now we show that $a(\mu, \delta)$ is continuous. Let $\epsilon \equiv |\mu - \acute{\mu}| + |\delta - \acute{\delta}|, Q \equiv (\mu, \delta), \acute{Q} = (\acute{\mu}, \acute{\delta}), R = (\mu_0, \delta_0) = (\mu \wedge \acute{\mu}, \delta \wedge \acute{\delta})$ and $P = (\mu_0 + \epsilon, \delta_0 + \epsilon)$. From Eqs. (6) and (7) it follows that

$$a(R) \leq a(Q) \leq a(P) \leq a(R) + \epsilon,$$

Similarly, we have

$$a(R) \leq a(\acute{Q}) \leq a(P) \leq a(R) + \epsilon.$$

Thus,

$$|a(Q) - a(\acute{Q})| \leq \epsilon, \forall |\mu - \acute{\mu}| + |\delta - \acute{\delta}| \leq \epsilon,$$

and we have proved that $a(\mu, \delta)$ is continuous.

Although $a(\mu, \delta)$ might not be strictly monotone in each parameter in the whole space, in the neighborhood of the line $a(\mu, \delta) = 0$ we can prove that $a(\mu, \delta)$ is strictly monotone in the $(1, 1)$ direction.

To see this, let $\gamma = \max(\mu, \delta)$. Observe from Equation (5) that

$$\begin{aligned} &S_k(\mu + \epsilon, \delta + \epsilon) \\ &\quad \leq (\gamma + \epsilon)(D_k(\mu + \epsilon, \delta + \epsilon) + U_k(\mu + \epsilon, \delta + \epsilon)) \\ &\qquad - M_k(\mu + \epsilon, \delta + \epsilon) \\ &\quad = \left(\gamma + \epsilon + \frac{1}{2}\right)(D_k(\mu + \epsilon, \delta + \epsilon) + U_k(\mu + \epsilon, \delta + \epsilon)) - \frac{k}{2}, \end{aligned}$$

which implies

$$\frac{S_k(\mu+\epsilon,\delta+\epsilon)+\frac{k}{2}}{\gamma+\epsilon+\frac{1}{2}} \le D_k(\mu+\epsilon,\delta+\epsilon)+U_k(\mu+\epsilon,\delta+\epsilon).$$

On the other hand, because an optimal alignment for $S_k(\mu,\delta)$ may not be an optimal alignment for $S_k(\mu+\epsilon,\delta+\epsilon)$, we have

$$\begin{aligned} &S_k(\mu+\epsilon,\delta+\epsilon)\\ &\quad=(\delta+\epsilon)D_k(\mu+\epsilon,\delta+\epsilon)+(\mu+\epsilon)U_k(\mu+\epsilon,\delta+\epsilon)\\ &\qquad-M_k(\mu+\epsilon,\delta+\epsilon)\\ &\quad\ge S_k(\mu,\delta)+\epsilon\big(D_k(\mu+\epsilon,\delta+\epsilon)+U_k(\mu+\epsilon,\delta+\epsilon)\big). \end{aligned}$$

Combining the last two equations, we obtain

$$S_k(\mu+\epsilon,\delta+\epsilon) \ge S_k(\mu,\delta)+\frac{\epsilon\big(S_k(\mu+\epsilon,\delta+\epsilon)+\frac{k}{2}\big)}{\gamma+\epsilon+\frac{1}{2}}.$$

Dividing by k on both sides of this equation and taking limits yields

$$\begin{aligned} a(\mu+\epsilon,\delta+\epsilon) &\ge a(\mu,\delta)+\frac{\epsilon\big(a(\mu+\epsilon,\delta+\epsilon)+\frac{1}{2}\big)}{\gamma+\epsilon+\frac{1}{2}}\\ &> a(\mu,\delta). \end{aligned}$$

The last inequality follows because $a(\mu+\epsilon,\delta+\epsilon)$ is in the neighborhood of the line $a(\mu,\delta)=0$.

This completes the proof of Lemma 1. ∎

3. LOGARITHMIC GROWTH

In this section we study the behavior of the tail probabilities $P(S_k \le qk)$, where $q<0$. Recall Equation (2):

$$P(S_{k+l} \le q(k+l)) \ge P(S_k \le qk)P(S_l \le ql).$$

Taking logarithms,

$$-\log P(S_{k+l} \le q(k+l)) \le -\log P(S_k \le qk) - \log P(S_l \le ql),$$

and therefore we can define the rate function $r(q)$:

$$r(q) = \lim_{k\to\infty} -\frac{1}{k}\log P(S_k \le qk) = \inf -\frac{1}{k}\log P(S_k \le qk).$$

We want to study scores that are more extreme (that is, smaller) than the average behavior $a(\mu, \delta)k$. The next theorem shows the large deviation behavior of such scores.

THEOREM 1. *Let $A_1A_2\ldots$ be iid with $q < a(\mu, \delta)$. Then*

$$0 < r(q) = \lim_{k\to\infty} -\frac{1}{k}\log P(S_k \le qk).$$

The proof depends heavily on the following Azuma–Hoeffding inequality. See Alon and Spencer (1992).

LEMMA 2 Azuma-Hoeffding. *Let X_i be a martingale with $X_0 = 0$ such that for some sequence $c_i, i \ge 1$ of positive constants*

$$|X_{i-1} - X_i| \le c_i.$$

Then, for $x > 0$,

$$P\left(\sup_{i\le k} X_i \ge x\right) \le \exp\left\{-\frac{x^2}{2}\Big/\sum_{i=1}^{k} c_i^2\right\}.$$

Proof of Theorem 1. We first define a martingale whose increments are bounded. Let $\mathcal{F}_i$ be a σ-field generated by the sequence of letters $A_1\ldots A_i$, denoted by $\sigma(A_1\ldots A_i)$, and define $X_i = E[S_k - E[S_k]|\mathcal{F}_i]$. It is clear that X_i is a martingale (and $X_0 = 0$). Because S_k is $\mathcal{F}_k$ measureable, by the property of conditional expectation $X_k = S_k - E[S_k]$.

To bound the martingale increments, we first derive a deterministic bound

$$S_k - S'_k \le c = \max(1 + 2\delta, 1 + 2\mu),$$

where $S_k = S(A_1, \ldots, A_{i-1}, A_i, A_{i+1}, \ldots, A_k)$ is the free-energy score for k letters, and $S' = S(A_1, \ldots, A_{i-1}, \acute{A}_i, A_{i+1}, \ldots, A_k)$ is the score with the ith letter changed.

Begin with a particular optimal alignment for S_k and assume that letter A_i is paired to A_j. Alignments for S' are given by (1) placing A'_i and A_j in bulges, so that $S'_k = S_k + 1 + 2\delta$, and by (2) placing A'_i and A_j in loops so that $S'_k = S_k + 1 + 2\mu$.

Thus,

$$S'_k \leq S_k + \max(1 + 2\delta, 1 + 2\mu) \tag{8}$$

As in Alon and Spencer (1992) or Arratia and Waterman (1994), we obtain $|X_{i-1} - X_i| \leq c_i = c = \max\{1 + 2\delta, 1 + 2\mu\}$. Now the Asuma–Hoeffding inequality can be invoked. Because $q < a = a(\mu, \delta), \epsilon = q - a < 0$. By subadditivity, $E[S_k] \geq ka$. Hence,

$$\begin{aligned} P(S_k \leq qk) &\leq P(S_k - E[S_k] \leq (q - a)k) \\ &= P(X_k \leq \epsilon k) \\ &= P(-X_k \geq -\epsilon k) \\ &\leq P\left(\sup_{i \leq k}(-X_i) \geq -\epsilon k\right). \end{aligned}$$

Let $x = -\epsilon k > 0$ and apply the Azuma-Hoeffding inequality:

$$\begin{aligned} P\left(\sup_{i \leq k}(-X_i) \geq -\epsilon k\right) &\leq \exp\left\{-\frac{\epsilon^2 k^2}{2} \Big/ \sum_{i=1}^{k} c_i^2\right\} \\ &= \exp\left\{-\frac{(q - a)^2 k}{2c^2}\right\}. \end{aligned}$$

Thus, combining the last two equations,

$$-\frac{1}{k} \log P(S_k \leq qk) \geq \frac{(q - a)^2}{2c^2} > 0.$$

This completes the proof. ∎

The relevant function is F_n, because free-energy does not penalize for unpaired bases "outside" of secondary structure, defined by

$$F_n = \min\left\{\min_{1 \leq i < j \leq n} S(i, j), 0\right\},$$

where $S(i, j) = S(A_i \ldots A_j)$.

Intuitively, the quantity $a(\mu, \delta)$ represents the average score per pair of letters. If we assume $a(\mu, \delta) > 0$, then we can consider negative values of

F_n to be rare events. For $a(\mu, \delta) > 0$ and $q \leq 0$, by Theorem 1, $r(q) \geq 0$. Then, we can define

$$b = \min_{q \leq 0} \frac{q}{r(q)}.$$

Because for all $\epsilon > 0, r(-\frac{1}{2} + \epsilon) \leq -\frac{1}{2} \log P(A_1 \text{ pairs } A_2) < \infty$, we have $b < 0$.

The following lemma shows that under the assumption that $a(\mu, \delta) > 0$, the growth of F_n is a constant b times the logarithm of the total length n of the RNA sequence. This is tighter than the corresponding result for local-alignment scores in (Arratia and Waterman, 1994) which has been improved in (Zhang, 1995).

LEMMA 3. *The sequence $A_1A_2 \ldots A_n$ is made up of iid letters. For all $(\mu,\delta) \in [0,\infty]^2$, if $a(\mu, \delta)$, then for all $\epsilon > 0$,*

$$P\left((1 + \epsilon)b < \frac{F_n}{\log n} < (1 - \epsilon)b\right) \to 1. \tag{9}$$

Proof. We first prove the upper bound

$$P\left(\frac{F_n}{\log n} < (1 - \epsilon)b\right) \to 1,$$

that is equivalent to

$$P(F_n > (1 - \epsilon)b \log n) \to 0 \qquad \text{as } \to \infty.$$

Let $t = (1 - \epsilon)b \log n$ and $k = \lceil t/q \rceil$. To approximate the probability $P(F_n > t)$, we divide the whole sequence into non-overlapping blocks of length $k + 1$. The subsequences in the blocks are independent. The probability $P(F_n > t)$ then will be approximated by products of the probability that the score of subsequences in the block exceeds $q(k + 1)$. Now we calculate this probability.

Because $b = \min_{q \leq 0}(q/r(q))$, given $\epsilon > 0$, we can chose small $\delta > 0$ and q such that

$$(1 - \epsilon)b\left(\frac{r(q) + \delta}{q}\right) < 1 - \frac{\epsilon}{2}.$$

For sufficiently large n, $k = \lceil t/q \rceil$ is sufficiently large so that

$$-\frac{1}{k} \log P(S_k \leq qk) \leq r(q) + \delta$$

Thus,

$$\begin{aligned} P(S_k \le qk) &\ge \exp\{-k(r(q)+\delta)\} \\ &\ge \exp\left\{-t\frac{r(q)+\delta}{q}\right\} \\ &= \exp\left\{-(1-\epsilon)b\frac{r(q)+\delta}{q}\log n\right\} \\ &\ge \exp\{-(1-\epsilon/2)\log n\} \\ &= n^{-(1-\epsilon/2)}. \end{aligned}$$

With blocks of length $k+1$, the total number of blocks is approximately n/k. Below we show that $P(F_n > t = (1-\epsilon)b\log n)$ approaches zero as $n \to 0$. Let $j = k+1$. Then $t > qj$. Thus,

$$\begin{aligned} P(F_n > t) &\le P(F_n > qj) \\ &\le P\Big(\min_{1\le i<j\le n} S(i,j) > qj\Big) \\ &\le P\Big(\bigcap_{0\le i\le [n/j]-1} S(A_{ij+1}\cdots A_{ij+j}) > qj\Big) \\ &= P(S_j > qj)^{[n/j]} \\ &\le (1-n^{-1+\epsilon/2})^{[n/j]} \text{ (for sufficiently large n)} \\ &\to 0 \qquad (\text{as } n \to \infty). \end{aligned}$$

This proves the upper bound.

The lower bound we now prove is:

$$P(F_n > (1+\epsilon)b\log n) \to 1.$$

In order to prove this, we will show that $P(F_n \le (1+\epsilon)b\log n) \to 0$. Let $t = (1+\epsilon)b\log n$. The event $\{F_n \le t\}$ is contained in a union of about n^2 events, by choosing the starting and ending points for the high-scoring regions. This union of n^2 events can be decomposed further into two sub-unions: one sub-union consisting of order $n\log n$ events that contribute most of the probability and another sub-union containing remaining events that have less significant contribution.

Formally, let $c = 3/r(0)$; by Theorem 1, $r(0) > 0$. Because $t < 0$, then $\{F_n \le t\} = \{\min_{1 \le i < j \le n} S(i, j) \le t\}$. The events $\{\min_{1 \le i < j \le n} S(i, j) \le t\}$ are contained in the two unions, i.e.,

$$\{F_n \le t\} \subseteq \bigcup_{\substack{1 \le i_0 \le n \\ 1 \le j \le c \log n}} \{S(A_{i_0+1} \dots A_{i_0+j}) \le t\}$$

$$\cup \bigcup_{\substack{1 \le i_0 \le n \\ c \log n < j}} \{S(A_{i_0+1} \dots A_{i_0+j}) \le 0\}.$$

Let $t = qj$ where $q < 0$. Then we have

$$t\frac{r(q)}{q} = (1 + \epsilon)(\log n)b\frac{r(q)}{q}$$

$$\ge (1 + \epsilon)\log n. \tag{10}$$

The last step follows because $br(q)/q \ge 1$ which is implied by the definition of b and the assumption $q < 0$. Because the sequence $A_1 \dots$ has i.i.d. letters, in the first union each event has probability at most

$$P(S(A_{i_0+1} \dots A_{i_0+j}) \le t) = P(S_j \le t)$$

$$= P(S_j \le qj).$$

For all k, we have

$$-\frac{1}{k} \log P(S_k \le qk) \ge r(q).$$

Thus,

$$P(S_j \le qj) \le e^{-jr(q)}$$

$$= e^{-t(r(q)/q)}$$

$$\le e^{-(1+\epsilon)\log n}$$

$$= n^{-(1+\epsilon)}.$$

As mentioned above, the first union consists of at most $n(c \log n)$ events, hence the probability of the first union satisfies

$$P\left(\bigcup_{\substack{1 \le i_0 \le n \\ 1 \le j \le c \log n}} S(A_{i_0+1} \dots A_{i_0+j}) \le t\right) \le (nc \log n)n^{-(1+\epsilon)}$$

$$= c\frac{\log n}{n^{\epsilon}} \to 0.$$

The second union involves at most n^2 events of the form $\{S_j \le 0\}$. Because the length j of each sequence in the second union is larger than $c \log n$ and $c = 3/r(0)$, the probability of each of these events satisfies

$$\begin{aligned} P(S_j \le 0) &\le e^{-jr(0)} \\ &\le e^{(-c \log n)r(0)} \\ &= e^{-3 \log n}. \end{aligned}$$

Therefore, the second union has probability at most

$$P\left(\bigcup_{\substack{1 \le i_0 \le n \\ c \log n < j}} \{S(A_{i_0+1} \dots A_{i_0+j}) \le 0\} \right) \le n^2 e^{-3 \log n}$$

$$= \frac{1}{n} \to 0.$$

This completes the proof that

$$\frac{F_n}{\log n} \to b \qquad \text{in } \textit{probability}. \quad \blacksquare$$

For sequence alignment scores, the corresponding even $\{M \ge t\}$ is expressed as a union of n^4 events. This explains why the upper bound for the coefficient of b in sequence matching has a factor of 2, but for RNA free-energy scores it has a factor of 1. From the above discussion, we know that, if $a(\mu,\delta) > 0$, the score of the optimal subregions will grow like $b \log n$, where b is defined as $b = \min_{q \le 0} (q/r(q))$.

4. LINEAR GROWTH

In this section we show that if $a(\mu,\delta) > 0$, then both M_n/n and S_n/n converge to $a(\mu,\delta)$ with probability 1; that is, they grow linearly.

LEMMA 4. *If $a(\mu,\delta) < 0$, then both M_n and S_n grow linearly. More precisely, the following limits hold with probability* 1:

$$\frac{F_n}{n} \to a(\mu, \delta), \tag{11}$$

$$\frac{S_n}{n} \to a(\mu, \delta). \tag{12}$$

Proof. In the previous section we have shown that S_k is subadditive and thus the subadditive ergodic theory implies Eq. (12):

$$\frac{S_n}{n} \xrightarrow{a.s.} a(\mu, \delta). \tag{13}$$

Now we establish Eq. (11). Because $F_n \le S_n$, the event $[F_n \ge (1 - \epsilon)na]$ implies the event $[S_n \ge (1 - \epsilon)na]$. Hence we obtain that

$$\begin{aligned} P(F_n \ge (1 - \epsilon)na) &\le P(S_n \ge (1 - \epsilon)na) \\ &\to 0. \end{aligned}$$

Next we prove that

$$P(F_n \le (1 + \epsilon)na) \to 0.$$

Let $k = j - i + 1$, $t = (1 + \epsilon)na$. Then for all $i, j \le n$, because $k < n$, we have that

$$\begin{aligned} P(S_{ij} \le t) &= P(S_{ij} \le (1 + \epsilon)na) \\ &\le P(S_{ij} \le (1 + \epsilon)ak). \end{aligned} \tag{14}$$

Because

$$r = r((1 + \epsilon)a) = \inf - \frac{1}{k} \log P(S_{ij} \le (1 + \epsilon)ak)$$

it follows that

$$P(S_{ij} \le (1 + \epsilon)ak) \le e^{-rk}. \tag{15}$$

Because $a(1 + \epsilon) < a$, Theorem 1 implies $r > 0$.

Because each basepair scores -1, $\{S_{ij} \le t\}$ implies that $k \ge -2t$. From Eqs. (14) and (15), it follows that

$$P(S_{ij} \le t) \le e^{2rt}.$$

Therefore,

$$\begin{aligned} P(F_n \le t) &\le P\left(\min_{1 \le i < j \le n} S_{ij} \le t\right) \\ &= P\left(\bigcup_{1 \le i < j \le n} \{S_{ij} \le t\}\right) \\ &\le n^2 e^{2rt} \to 0. \end{aligned}$$

This proves that Eq. (11) holds in probability. The Azuma–Hoeffding inequality in Lemma 2 applies to F_n as well as S_n. Then as in Arratia and Waterman, it follows that F_n converges to a almost surely. ∎

Combining Lemma 1, Lemma 3, and Lemma 4, we obtain the following phase transition theorem for RN secondary structure alignment scores.

THEOREM 2. *For i.i.d. letters $A_1, A_2, \ldots$, the optimal RNA secondary structure alignment score $F_n = \min\{\min_{1 \le i < j \le n} S(i,j), 0\}$, with penalty parameters δ per letter in bulges and μ per remaining unpaired letters, has a phase transition between linear growth with n for small μ and δ, and logarithmic growth with n for large μ and δ. More precisely, if $a(\mu, \delta) < 0$ then $F_n/n \to a(\mu, \delta)$ and if $a(\mu, \delta) > 0$ then $F_n/(\log n) \to b$.*

5. SIMULATION OF THE PHASE TRANSITION CURVE

Our theorem shows that there is a phase transition between linear growth of the minimum free-energy score in n with "small" values of bulge and loop penalties and logarithmic growth in n with "large" values. We know little theoretically about the location of the phase transition curve in $[0, \infty]^2$. To obtain more information about the shape of the phase transition curve, we use simulation to study the free-energy score of a random RNA. We begin with calculation of minimum free energy. For more about the logarithm for computing minimum free energy, we refer reader to Zuker and Sankoff (1984) and Waterman (1995). Under our simple free-energy model, we give a dynamic-programming algorithm for computing minimum free energy.

First we define some notation necessary for describing the algorithm. Let $g(i,j)$ be the minimum free energy of the RNA sequence $A_i \ldots A_j$ with A_i and A_j paired, $e(i,j)$ be the free energy for an end loop with A_i and A_j paired, $b(i,j)$ be the free energy for a bulge with penalty parameter δ, $b_1(i,j)$ and $b_2(i,j)$ be the free energy for left- and right-bulges with parameter δ, $t(i,j)$ be the free energy for the interior loop, and $l(i,j)$ be the minimum free energy for a multi-branch loop. For the convenience of discussing the algorithm, we also define $b^{\mu}(i,j)$ and $b_{\mu}(i,j)$ to be the free energy for the left- and right-bulges but with the penalty parameter μ instead of δ.

Now we give an algorithm to compute the free energy of $A_1 A_2 \ldots A_n$. Recall that this is defined to be the minimum of $g(i,j)$ free-energy score when i and j are paired ($i \cdot j$), or zero:

$$F_n = \min\left\{ \min_{1 \le i < j \le n} g(i,j), 0 \right\}.$$

Also basepairs are scored -1, bulged (unpaired) letters are scored $+\delta$ per letter and all other unpaired letters are scored $+\delta$ per letter.

We begin with end-loops where $i \cdot j$ and the letters $(i+1)\cdots(j-1)$ are unpaired. Define

$$e(i,j) = -1 + \mu(j-i+1).$$

Bulges require a little more work. There are two cases. $b_1(i,j)$ is the minimum free energy of all structures where $i \cdots (i+k-1)$ is bulged and $(i+k)\cdot j$ is a basepair:

$$\begin{aligned} b_1(i,j) &= \min_{k\geq 1}\{k\delta + g(i+k,j)\} \\ &= \min\Big\{\delta + g(i+1,j), \min_{k\geq 2}\{k\delta + g(i+k,j)\}\Big\} \\ &= \min\Big\{\delta + g(i+1,j), \min_{l\geq 1}\{\delta(l+1) + g(i+1+l,j)\}\Big\} \\ &= \min\Big\{\delta + g(i+1,j), \delta + \min_{l\geq 1}\{\delta l + g(i+1+l,j)\}\Big\} \\ &= \delta = \min\{g(i+1,j), b_1(i+1,j)\} \end{aligned}$$

Similarly, when $b_2(i,j)$ is the minimum energy structure with $i\cdot(j-k)$ a basepair and with the letters $(j-k+1)\cdots j$ unpaired,

$$b_2(i,j) = \delta + \min\{g(i,j-1), b_2(i,j-1)\}.$$

Then

$$\begin{aligned} b(i,j) &= \min\{b_1(i,j), b_2(i,j)\} \\ &= \delta + \min\{g(i+1,j), g(i,j-1), b_1(i+1,j), b_2(i,j-1)\}. \end{aligned}$$

It is useful to have these quantities scoring the "bulge" with μ per letter rather than δ per letter:

$$b^{\mu}(i,j) = \mu + \min\{g(i+1,j), b^{\mu}(i+1,j)\}.$$

and

$$b_{\mu}(i,j) = \mu + \min\{g(i,j-1), b_{\mu}(i,j-1)\}.$$

By definition, the minimum free energy $t(i,j)$ for interior loops on $i \cdots j$ is

$$t(i,j) = \min_{k_1\geq 1, k_2\geq 1}\{\mu(k_1+k_2) + g(i+k_1, j-k_2)\}$$

We will decompose this minimum into four terms:

$$\{k_1 = k_2 = 1\}, \{k_1 \geq 2, k_2 = 1\}, \{k_1 = 1, k_2 \geq 2\}, \{k_1 \geq 2, k_2 \geq 2\},$$

and then simplify them.

$$\begin{aligned} t(i,j) = \min\Big\{ & 2\mu + g(i+1, j-1), \min_{k_1 \geq 2} \{\mu(k_1 + 1) + g(i + k_1, j - 1)\}, \\ & \min_{k_2 \geq 2} \{\mu(1 + k_2) + g(i + 1, j - k_2)\}, \\ & \min_{k_1 \geq 2, k_2 \geq 2} \{\mu(k_1 + k_2) + g(i + k_1, j - k_2)\}\Big\} \\ = 2\mu + \min\Big\{ & g(i+1, j-1), \min_{l_1 \geq 1} \{l_1 \mu + g(i + 1 + l_1, j - 1)\}, \\ & \min_{l_2 \geq 1} \{l_2 \mu + g(i + 1, j - 1 - l_2)\}, \\ & \min_{l_1 \geq 1, l_2 \geq 1} \{\mu(l_1 + l_2) + g(i + 1 + l_1, j - 1 - l_2)\}\Big\} \\ = 2\mu + \min\{ & g(i+1, j-1), b^{\mu}(i+1, j-1), \\ & b_{\mu}(i+1, j-1), t(i+1, j-1)\}. \end{aligned}$$

Now we consider the free energy $l(i, j)$ of multi-branch loop structures. These are loops that have one or more helices extending from them. The unpaired letters in the loop are scored μ per letter. At the left side of the "loop" is A_i which is in a basepair or not. This implies

$$l(i,j) = \min\Big\{\mu + l(i+1, j), \min_{i<k\leq j} \{g(i,k) + l(k+1, j)\}\Big\}.$$

Finally the minimum free energy $g(i, j)$ on $i \cdots j$ with $i \cdot j$ paired, is given by

$$\begin{aligned} g(i,j) = \min\{ & e(i,j), -1 + g(i+1, j-1), -1 + b(i+1, j-1), \\ & -1 + t(i+1, j-1), -1 + l(i+1, j-1)\}. \end{aligned}$$

The computation will be performed on line of $j - i = c$, for constant $c = m, m + 1, \ldots$. For details on organizing the computation of the minimum free energy the reader is referred Waterman (1995).

To see how quickly the average free energy S_n/n converges to $a(\mu, \delta)$, we plot Fig. 2, which shows S_n/n against the length n of the sequence for $(\mu, \delta) = (0.1, 0.2)$. It can be seen that after $n = 300$, S_n/n fluctuates

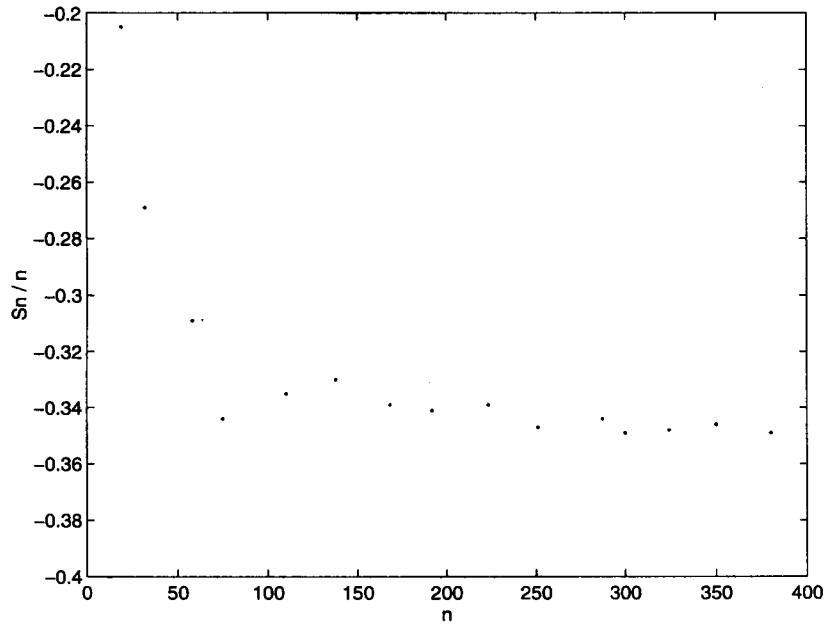


FIG. 2. Score per letter, S_n/n as a function of n, for parameter values $(\mu, \delta) = (0.1, 0.2)$.

around $a(\mu, \delta) = a(10, 1) = -0.35$. So, to simulate the shape of the phase transition curve, we studied

$$\{(\mu, \delta) : S(A_1 \cdots A_{300}) = 0\}$$

The study is motivated by the definition of the phase transition curve, $\{(\mu, \delta) : \lim_{k \to \infty} ES_k/k = 0\}$. The simulated curve appears as Fig. 3.

In case (μ, δ) is in the logarithmic region, we conjecture by analogy to results for local alignment that there is a valid Poisson approximation (Arratia and Waterman, 1994; Waterman and Vingron, 1994). By this conjecture, we mean that for large positive values of t,

$$P(F_n \leq -t) \approx 1 - e^{-\xi n p^t}$$

where $0 < p < 1$. There are two parameters to estimate, ξ and p. We computed the minimum free energy with the parameters $(\mu, \delta) = (10, 1)$. Our estimates of the distribution parameters are $\xi = 1.6 \times 10^{-3}$ are $p = 0.7$. The Poisson approximation provides a good fit to the data. In sequence matching, it is conjectured that the coefficient of $log(n)$ is $2b$ but Arratia and Waterman (1994) were only able to show that the coefficient was between b and $2b$. In the present case for RNA we have shown in Lemma 3 using the same methods that the coefficient is b. (By more

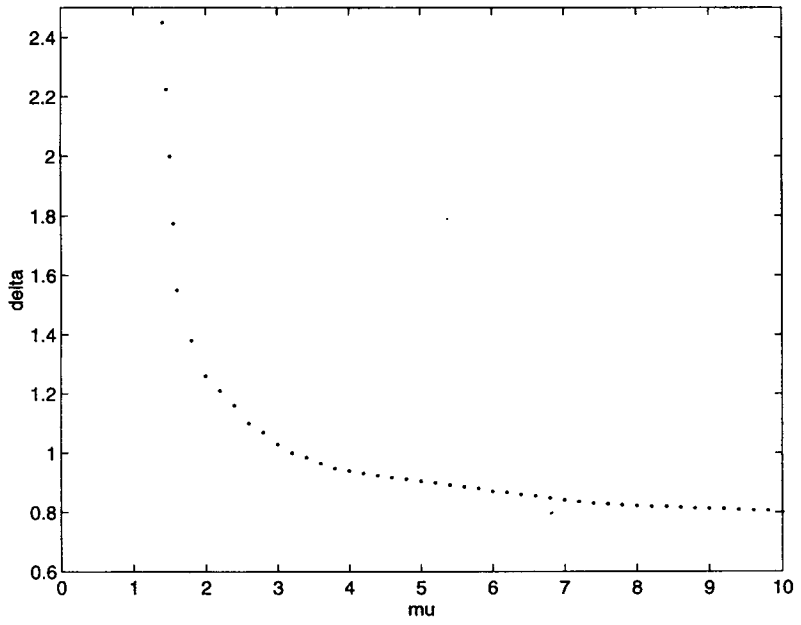


FIG. 3. Location of the phase transition curve $a(\mu, \delta) = 0$ in the (μ, δ) plane.

complex methods, Zhang (1995) has proven that for sequence matching the coefficient is 2b.) For these reasons, we feel that a rigorous proof of Poisson approximation will be less difficult for F_n than for local alignment scores for sequence matching.

Programs for free energy often compute a more complex function that allows multiple domains of folding:

$$E_n = \min_m \{S(i_1, j_1) + S(i_2, j_2) + \cdots + S(i_m, j_m):$$
$$1 \le i_1 < j_1 < i_2 < j_2 < \cdots < i_m < j_m \le n\}$$

Under any reasonable assignment of free-energy values, sequences such as *GGAAACC*, for example, have free energy, $S(GGAAACC) = e < 0$.) (It is possible to increase the number of $G \cdot C$ pairs until $e < 0$. Then by the strong law of large numbers, we expect to find $E_n < (n/7)P(GGAAACC)e$ for large n. This shows that E_n has linear growth. Its asymptotic distribution remains an important open question.

ACKNOWLEDGMENT

The authors thank Richard Arratia for helpful conversations about RNA free energy and phase transitions.

REFERENCES

N. Alon and J. H. Spencer, "The Probabilistic Method," Wiley, New York, (1992).
R. A. Arratia, L. Goldstein, and L. Gordon, Two moments suffice for Poisson approximation: The Chen-Stein method, *Ann. Probab.*, **17**, (1989), 9–25.
R. A. Arratia, L. Gordon, and M. S. Waterman, The Erdős-Rényi law in distribution, for coin tossing and sequence matching, *Ann. Statist.*, **18**, (1990), 539–570.
R. A. Arratia and M. S. Waterman, An Erdős-Rényi law with shifts, *Adv. in Math.*, **55**, (1985a), 13–23.
R. A. Arratia and M. S. Waterman, Critical phenomena in sequence matching, *Ann. Probab.*, **13**, (1985b), 1236–1249.
R. A. Arratia and M. S. Waterman, A phase transition for the score in matching random sequences allowing deletions, *Ann. Appl. Probab.*, **4**, (1994), 200–225.
L. Goldstein and M. S. Waterman, Approximations to profile score distributions, *J. Comp. Biol.*, **1**, (1994), 93–104.
J. Gralla and C. DeLisi, mRNA is expected to form stable secondary structures, *Nature*, **248**, (1974), 330–332.
S. Karlin and A. Dembo, Limit distributions of maximal segmental score among Markov dependent partial sums, *Adv. Appl. Probab.*, **24**, (1992), 113–140.
S. Karlin and F. Ost, Maximal length of common words among random letter sequences, *Ann. Probab.*, **16**, (1988), 535–563.
S. Le, J-H. Chen, K. M. Currey, and J. Maizel, Jr., A program for predicting significant RNA secondary structures, *CABIOS*, **4**, (1988), 153–159.
C. Neuhauser, A Poisson approximation for sequence comparisons with insertions and deletions, *Ann. Statist.*, (1994).
R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman, Algorithms for loop matching, *SIAM J. Appl. Math.*, **35**, (1978), 68–82.
D. Sankoff, Simultaneous solution of the RNA folding, alignment and protosequence problems, *SIAM J. Appl. Math.*, **45**, (1985), 810–824.
M. S. Waterman, Secondary structure of single-stranded nucleic acids, Studies in Foundations and Combinatorics, *Adv. Math. Suppl. Stud.*, **1**, (1978), 167–211.
M. S. Waterman, Estimating statistical significance of sequence alignments, *Philos. Trans. R. Soc. London B.*, **344**, (1994), 383–390.
M. S. Waterman, L. Gordon, and R. Arratia, Phase transitions in sequence matches and nucleic acid structure, *Proc. Natl. Acad. Sci.*, **84**, (1987), 1239–1243.
M. S. Waterman and T. F. Smith, RNA secondary structure: A complete mathmatical analysis, Math. Biosci., **42**, (1978), 257–266.
M. S. Waterman and T. F. Smith, Rapid dynamic programming algorithms for RNA secondary structure, *Adv. Appl. Math.*, **7**, (1986), 455–464.
M. S. Waterman and M. Vingron, Rapid and accurate estimates of statistical significance for sequence data base researches, *Proc. Natl. Acad. Sci. USA*, **91**, (1994a), 4625–4628.
M. S. Waterman and M. Vingron, Sequence comparison significance and Poisson approximation, *Statist. Sci.*, **9**, (1994b), 367–381.
Y. Zhang, A limit theorem for matching random sequences allowing deletions, *Ann. Appl. Probab.* **5**, (1995), 1236–1240.
M. Zuker and D. Sankoff, RNA secondary structures and their prediction, *Bull. Math. Biol.*, **46**, (1984), 591–621.