The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure

J. S. McCASKILL

Max-Planck Institut für Biophysikalische Chemie, Nikolausberg am Faßberg D-3400, Göttingen, Federal Republic of Germany

SYNOPSIS

A novel application of dynamic programming to the folding problem for RNA enables one to calculate the full equilibrium partition function for secondary structure and the probabilities of various substructures. In particular, both the partition function and the probabilities of all base pairs are computed by a recursive scheme of polynomial order N^3 in the sequence length N. The temperature dependence of the partition function gives information about melting behavior for the secondary structure. The pair binding probabilities, the computation of which depends on the partition function, are visually summarized in a "box matrix" display and this provides a useful tool for examining the full ensemble of probable alternative equilibrium structures. The calculation of this ensemble representation allows a proper application and assessment of the predictive power of the secondary structure method, and yields important information on alternatives and intermediates in addition to local information about base pair opening and slippage.

The results are illustrated for representative tRNA, 5S RNA, and self-replicating and self-splicing RNA molecules, and allow a direct comparison with enzymatic structure probes. The effect of changes in the thermodynamic parameters on the equilibrium ensemble provides a further sensitivity check to the predictions.

INTRODUCTION

The prediction of folded RNA structure is a chemical problem of considerable biological importance. The sequence of ribonucleotide bases (C, G, U, and A), covalently linked in the chain polymer, determines the fold through specific base–base interactions (pairing and stacking) involving hydrogen bonds, ring system conjugation, and hydrophobic effects in addition to the more globally acting electrostatic interactions between the negatively charged phosphate groups and with ions in solution. Since the numbers of degrees of freedom in the polynucleotide chain exceeds that in polypeptides, the full structural prediction problem is intrinsically more difficult than the problem of protein folding. However, for RNA, it has proved permissible to focus initially on an intermediate level representation of the folding (secondary structure) in terms only of the bonds between pairs of nucleotide bases in the chain, relegating more detailed and additional folding information to a subsequent stage of analysis.[1] Furthermore, for this secondary structural level, an approximate decomposition of the free energy of a structure into a sum of terms involving adjacent stacked base pairs and the consequent constrained loops in the polynucleotide backbone has allowed the additive application of information gleaned from conformational transitions in small model molecules and duplices[2,3] to predict the optimal free energy secondary structure. The resulting structures have proven useful in the prediction of the full tertiary structure and the associated biological function of RNA sequences found in living organisms.[4]

In this work, the partition function for secondary structures of RNA is shown to be within computational reach by means of a novel applica-

CCC 0006-3525/90/051105-15 $04.00
Biopolymers, Vol. 29, 1105–1119 (1990)

tion of the dynamic programming technique introduced by Bellman.[5] In common with previous work (see the review in Ref. 6), the application of dynamic programming to the partition function here rests heavily on the constraint of the absence of knots (for further details see below). While a calculation of the partition function has in principle been possible since the direct enumeration of all possible bonding patterns (cf. Ref. 3), the computation time grows exponentially with the length N of the sequence, making this analysis of all but the shortest sequences intractable. Thus a major component in any theory for the partition function is the order of the solution algorithm. In recent years, the original application[7-9] of dynamic programming to the optimal structure prediction problem has been refined[6,10,11] to an order N^3 algorithm utilizing simplifying approximations for the free energies of large and multiple loops (see below). Recently,[12] Waterman proposed a backtracking method for obtaining all structures in the vicinity of the optimal structure (following a dynamic programming analysis). The difficulty in the application of this algorithm is the embarassing abundance of structures having a free energy near that of the optimum, for sequences of even moderate length (e.g., $N = 60$), as found in a recent implementation by the present author. This problem has been raised previously.[6]

The present work returns to the original statistical mechanical problem of first calculating the partition function, from which further specific quantities of thermodynamic interest may be derived. The most useful characteristic of the equilibrium ensemble of structures appears from the present work to be the matrix of equilibrium base pair binding probabilities between all the nucleotides in the RNA chain. To avoid any confusion at this point, it is essential to stress that these probabilities are not locally determined by the sequence—each reflects a sum over all equilibrium weighted structures in which the chosen binding pair occurs. In this way the matrix, which we shall display in terms of solid boxes whose size decreases as the logarithm of the binding probability, summarizes information about the global ensemble of structures in equilibrium. The display is particularly useful in allowing the reconstruction of complete alternative structures and kinetic folding pathways between these, in addition to a determination of the flexibility of the structures toward local changes in base pairing (e.g., base pair opening and slippage). This is possible because of the easy visual assessment of compatibility of pairs in a single structure. The present work should be distinguished from the so-called dot matrix methods,[2,13,14] in which all possible matching regions of a sequence are displayed. In common is the use of a matrix layout for displaying information about bound pairs, but the information displayed is radically different.

In addition to this analytical tool, the present work allows a full analysis of the unfolding transition of secondary structure as temperature is increased (and in principle as salt concentration is increased). Further, the equilibrium-summed probabilities of binding of various bases allows a direct comparison with enzymatic and chemical modification experiments designed to detect exposed bases in the structure.

The body of the article is organized as follows: The next section develops the partition function dynamic programming result, establishing its reduction to an N^3 computation under a linear extrapolation for multiple loops. In the third section, the backsumming calculation for pair binding probabilities is described, and the "box matrix" introduced. The fourth presents the results of this analysis for some representative RNA molecules of interest: a transfer RNA, a 5S RNA, the midivariant RNA MDV replicated by the $Q\beta$ viral replicase and a self-splicing RNA from *Tetrahymena thermophila*. In the fifth section, an unfolding transition is studied by temperature variation of the partition function. Comparison is made with calorimetric data for the melting of *Escherichia coli* 5S RNA and fair agreement is obtained considering the limited range of applicability of the thermodynamic parameters. The paper concludes with a discussion of the usefulness of the equilibrium secondary structure predictions for longer molecules and of the insights gained from the partition function approach.

DYNAMIC PROGRAMMING CALCULATION OF THE PARTITION FUNCTION

In describing the secondary structure, the notation of Ref. 6 has proved convenient. The sequence of length N is numbered from the phosphate (5′) end to the hydroxyl (3′) end. An admissable structure is defined as a set S of "compatible" bonded pairs of bases (i, j) with $i < j$. Two base pairs are compatible if they have no base in common and do not form a "knot." A knot occurs for two base pairs (i, j), (h, l) chosen in the order $i < h$ if $i < h < j$

$< l$. The formation of the base pair (i, j) results in a looped region $[i + 1, j - 1]$ in the RNA chain, but any additional base pairs (h, l) with $i < h < l < j$ form their own "loops" and these regions $[h + 1, l - 1]$ are then by convention excluded from the loop "closed by" (i, j). In general, a k: loop consists of u unpaired bases and $k - 1$ base pairs (excluding the closing base pair). Loops are traditionally classified as hairpins ($k = 1$), stacked pairs ($k = 2, u = 0$), bulges and interior loops ($k = 2$, $u > 0$ with the unpaired bases respectively all contiguous or not), and multiple loops ($k > 2$). Each base in the structure then belongs to one loop only or is "external" to all loops.

The free energy of a secondary structure is assumed additive in terms of its loops

$$F(S) = \sum_{L \in S} F_L \tag{1}$$

where the free energies F_L of the various k: loops L (for $k \leq 2$), especially including those of the sequence-specific stacked pairs that are usually called "stems" rather than "loops," have been obtained from experiments with model compounds. The decomposition into enthalpy and entropy terms is used in order to be able to compute the temperature dependence (cf. section five), and non-Watson–Crick pairs are allowed making use of the known mismatch free energies. More detail about these thermodynamic parameters is contained in the fourth section where the method is applied to representative RNA molecules. For multiple loops ($k > 2$), no detailed analysis of the free energy contribution has proved possible. It is known[15] that the asymptotic entropy contribution to closure of a loop of length u is

$$S_l = \tfrac{3}{2} A \ln u \tag{2}$$

where A is a constant and the simplest treatment of the effective length of a loop with multiple bonded pairs is to count only the u unpaired bases. For efficient programming it will prove necessary, at least for longer multiple loops, to replace the convex form Eq. (2) by a linear approximation.

The additivity of free energy implies a multiplicativity in the contributions to the partition function Q defined by

$$Q = \sum_S e^{-[F(S)/kT]} \tag{3}$$

where the sum over all structures S involves an exponentially increasing number of terms as the sequence length N increases. This factorization of terms may be made explicit by introducing the restricted partition function Q^b_{ij}, where the sum is now over all structures S_{ij} on the sequence segment $[i, j]$ for which the base pair $(i, j) \in S_{ij}$. This sum may be replaced by a sum over all possible loops closed by (i, j), including the summed contributions of their subloops:

$$Q^b_{ij} = \sum_L e^{-[F_L/kT]} \prod_{(h,l) \in L} Q^b_{hl} \tag{4}$$

with $Q^b_{ii} = 0$, $i < j$, $h \leq l$. Each term in the sum involves a product over a set of base pairs $\{(h, l)\}$ with $i < h_1 < l_1 < h_2 < l_2 < \cdots < h_{k-1} < l_{k-1} < j$ defining the k: loop L closed by (i, j). The full partition function Q_{ij} for the inclusive segment $[i, j]$ includes, in addition, all configurations with i not bound to j, and it too satisfies a recursive formula

$$Q_{ij} = 1.0 + \sum_{\substack{h,l \\ i \leq h < l \leq j}} Q_{i,h-1} Q^b_{hl} \tag{5}$$

for $h \leq l$ with $Q_{ii} = 1.0$ and $Q_{i+1,i} = 1.0$. Starting with the shortest segments one proceeds iteratively to calculate Q^b_{ij} and Q_{ij} until one reaches Q_{1N}, which is the full partition function Q.

The difficulty with Eq. (3) is that the number of sets of bound pairs in a loop L closed by (i, j) still increases exponentially with $n = j - i$, so unless there is some decomposition of F_L the recursive scheme remains intractable. Fortunately, the classification of loops introduced above restricts the need for such a decomposition: the problem only grows exponentially for multiple loops ($k > 2$) where a linear decomposition may be introduced[6] in the absence of experimentally determined free energies

$$F_L = a + b(k - 1) + cu \tag{6}$$

where u is again the number of unpaired bases in the k: loop L and where a, b, and c are constants. One can then write, instead of Eq. (4),

$$Q^b_{ij} = e^{-[F_1(i,j)/kT]} + \sum_{\substack{h,l \\ i<h<l<j}} e^{-[(F_2(i,j,h,l))/kT]} + \sum_{\substack{h,l \\ i<h<l<j}} Q^m_{i+1,h-1} Q^b_{hl} e^{-[(a+b+c(j-l-1))/kT]} \tag{7}$$

where

$$Q_{ij}^{m} = \sum_{h,l} \left(e^{-[c(h-i-1)/kT]} + Q_{i,h-1}^{m} \right) \times Q_{hl}^{b} e^{-[(b+c(j-l-1))/kT]} \quad (8)$$

with $Q_{ii}^{m} = 0$ and $Q_{i+1,i}^{m} = 0$. The dynamic programming solution in Eqs. (5), (7), and (8) grows as N^4 with increasing sequence length N.

A reduction to a cubic algorithm may be obtained if the free energy of long interior ($k = 2$) loops $u > u_m$ may be regarded as prohibitive. This restriction has been widely used for long sequences. The computation of 2 : loops is then only quadratic, with a constant factor proportional to u_m^2. Alternatively, and more realistically, if the free energy of large 2 : loops with $u > u_m$ only depends on u and not on the position of the pair (h, l) within the loop (excluding bulge loops whose calculation is N^3 in any case)

$$F_2(i, j, h, l) = F_2(i, j, u) \quad (9)$$

where $u = h - i + j - l - 2$ is the number of unpaired bases in the loop, then the 2 : loop term may be written in the order N^3 form:

$$\sum_{h,l} e^{-[F_2(i,j,h,l)/kT]} = \sum_{\substack{h,l \\ u \le u_m}} e^{-[F_2(i,j,h,l)/kT]} + \sum_{u > u_m} (u-1) e^{-[F_2(i,j,u)/kT]} \quad (10)$$

This is a much less restrictive constraint than disallowing all long 2 : loops. A similar result was reported by Waterman.[11]

The multiple loop partition functions, on the other hand, may be calculated in cubic time without further approximation by introducing the additional restricted partition function

$$Q_{ij}^{m1} = \sum_{l=i+1}^{j} Q_{il}^{b} e^{-[c(j-l)/kT]} \quad (11)$$

which enables Eqs. (7) and (8) to be written in the form

$$Q_{ij}^{b} = e^{-[F_1(i,j)/kT]} + \sum_{\substack{h,l \\ i<h<l<j}} e^{-[F_2(i,j,h,l)/kT]} + \sum_{i<h<j} Q_{i+1,h-1}^{m} Q_{h,j-1}^{m1} e^{-[(a+b)/kT]} \quad (12)$$

where the final sum is over the nonfixed index h and where

$$Q_{ij}^{m} = \sum_{i<h\le j} \left(e^{-[c(h-i)/kT]} + Q_{i,h-1}^{m} \right) Q_{hj}^{m1} e^{-[b/kT]} \quad (13)$$

The cubic order calculation of the full partition functions Q_{ij} for the segments $[i, j]$, replacing Eq. (5),

$$Q_{ij} = 1.0 + \sum_{i\le h\le j} Q_{i,h-1} Q_{hj}^{1} \quad (14)$$

is completed by performing and storing the l summations separately as the ancillary quantities Q_{ij}^{1} defined by

$$Q_{ij}^{1} = \sum_{i\le l\le j} Q_{il}^{b} \quad (15)$$

with initial conditions $Q_{ii}^{1} = 0$, $Q_{ii} = 1.0$, and $Q_{i+1,i} = 1.0$.

In summary, then, the partition function $Q = Q_{1N}$ can be calculated with a dynamic programming algorithm of cubic order by a recursive application of Eqs. (10)–(15). This is the same order as that achieved in the optimal structure computation of Zuker and Stiegler.[16]

EQUILIBRIUM PROBABILITIES FOR STRUCTURES AND SUBSTRUCTURES

The probability of a given structure S, while proportional to the free energy factor $\exp(-F/kT)$ in equilibrium, is not known in absolute magnitude until the partition function has been calculated. Following the previous section, such probabilities are now available:

$$P(S) = \frac{1}{Q} e^{-[F(S)/kT]} \quad (16)$$

However, what actually occurs, on the time scale of most enzymatic reactions relevant for biological function, is rather an ensemble of related structures interchanging more or less rapidly with one another. While an equilibrium calculation does not suffice to define these kinetically clustered objects, the probability of a single completely defined set of

base pairs is not the quantity of interest either. The solution to this dilemma is to focus on the equilibrium probabilities of substructures S_s, common to a whole class of related structures S:

$$P(S_s) = \sum_{S \supseteq S_s} P(S) \tag{17}$$

One such substructure is of fundamental importance as it allows the display of the most significant features of the equilibrium ensemble and the interpretation of features relevent to chosen kinetic ensembles. It is the binding of a given pair of bases in the sequence. In this section we calculate the equilibrium probability of occurrence for each possible base pair, making use of the partition function and its restricted forms calculated in the second section. A direct summation then further yields the probability that a given base in the sequence is bound. We postpone a discussion of the significance of these equilibrium structure-ensemble weighted probabilities until the details of the calculation have been elucidated.

We define P_{hl} as the probability that h is bound to l in the equilibrium ensemble of structures

$$P_{hl} = \sum_{S \ni (h,l)} P(S) \tag{18}$$

If (h, l) is always a nonenclosed (or exterior) base pair, then

$$P_{hl} = \frac{Q_{1,h-1} Q^b_{hl} Q_{l+1,N}}{Q_{1N}} \tag{19}$$

but there will be other contributions to P_{hl} from structures with exterior base pairs (i, j) with $i < h < l < j$. We proceed recursively, for $d = l - h$ decreasing from the starting value N, to construct

$$\begin{aligned} P_{hl} = {} & \frac{Q_{1,h-1} Q^b_{hl} Q_{l+1,N}}{Q_{1N}} \\ & + \sum_{\substack{i,j \\ i<h<l<j}} P_{ij} \frac{Q^b_{hl}}{Q^b_{ij}} e^{-[F_2(i,j,h,l)/kT]} \\ & + \sum_{\substack{i,j \\ i<h<l<j}} P_{ij} \frac{Q^b_{hl}}{Q^b_{ij}} e^{-[(a+b)/kT]} \\ & \times \left(e^{-[(h-i-1)c/kT]} Q^m_{l+1,j-1} + Q^m_{i+1,h-1} \right. \\ & \left. \times e^{-[(j-l-1)c/kT]} + Q^m_{i+1,h-1} Q^m_{l+1,j-1} \right) \end{aligned} \tag{20}$$

The first line gives the contribution of Eq. (19), and the second and third lines the contributions from structures involving (h, l) in bulge, interior, and multiple loops with closing pair (i, j), summed over all possible pairs (i, j), with the factor P_{ij} being previously calculated since $j - i > l - h$.

As formulated, the calculation of all P_{hl} is of order N^4, since summations over i, j, h, and l are required, and even a particular value P_{hl} requires an N^4 calculation, since in general of order N^2 other values must first be calculated owing to the recursivity. However, as in the previous section, it is possible to reduce the calculation to order N^3 by introducing ancillary quantities and a careful distortion of the natural order in which the sums are performed.

We first concentrate on the case where the 2 : loops are strictly truncated at $u = u_m$ so that only the multiple k : loops $k > 2$ constitute the strictly N^4 part of the calculation. The sum over i and j may be reduced to a single sum by introducing

$$P^m_{il} = \sum_{j>l} \frac{P_{ij}}{Q^b_{ij}} Q^m_{l+1,j-1} \tag{21}$$

and

$$P^{m1}_{il} = \sum_{j>l} \frac{P_{ij}}{Q^b_{ij}} e^{-[((j-l-1)c)/kT]} \tag{22}$$

These sums must be calculated at the appropriate point in the recursion, when P_{ij} has been defined, but being independent of h, they do not need recalculation if the values are stored. The expression for P_{hl} then becomes

$$\begin{aligned} P_{hl} = {} & \frac{Q_{1,h-1} Q^b_{hl} Q_{l+1,N}}{Q_{1N}} \\ & + \sum_{\substack{i,j \\ i<h<l<j \\ u<u_m}} P_{ij} \frac{Q^b_{hl}}{Q^b_{ij}} e^{-[F_2(i,j,h,l)/kT]} \\ & + \sum_{i<h} Q^b_{hl} e^{-[(a+b)/kT]} \left(P^{m1}_{il} Q^m_{i+1,h-1} \right. \\ & \left. + P^m_{il} \left(e^{-[((h-i-1)c)/kT]} + Q^m_{i+1,h-1} \right) \right) \end{aligned} \tag{23}$$

Similarly, the assumption of Eq. (9), rather than truncation, for large 2 : loops results in a cubic order summation in the second line in addition to the existing quadratic order term. In order for

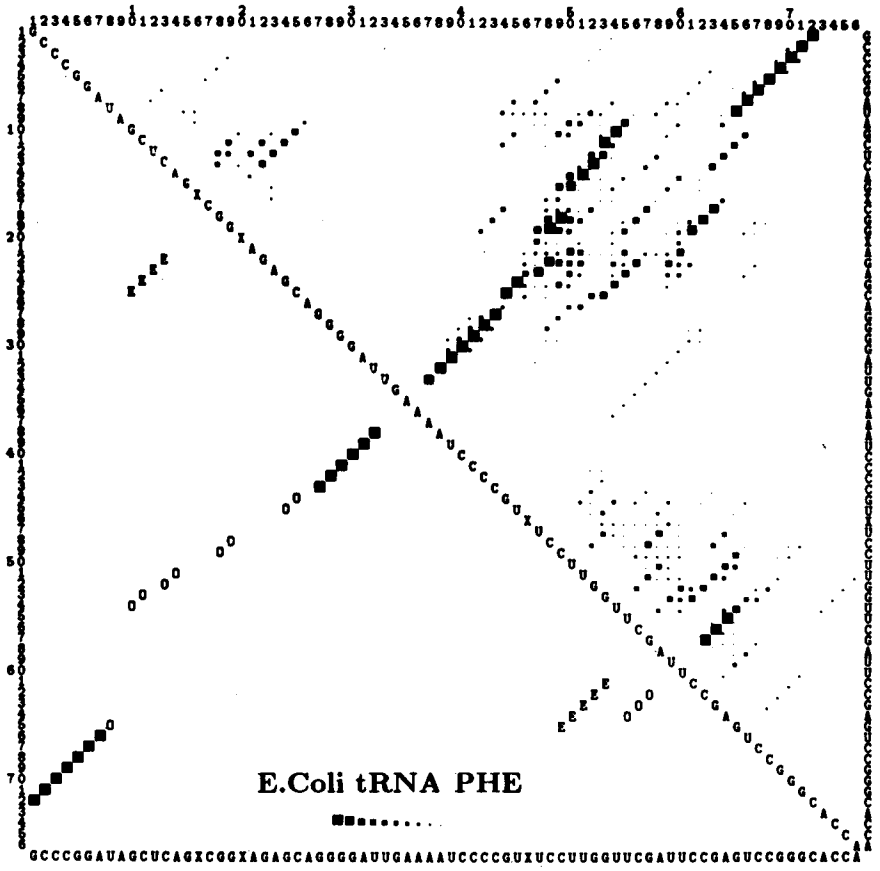

Figure 1. The box matrix for *E. coli* phenylalanine tRNA. The equilibrium ensemble probabilities of base pair bindings, using the Salser–Tinoco compilation of thermodynamic parameters at 25°C, are displayed in the upper right triangle. The probabilities P_{ij} are displayed on a logarithmic scale over 10 orders of magnitude by boxes of decreasing size at rows i and columns j in the matrix. In the lower left triangle, the predicted optimal structure is displayed by placing an O for each base pair (i, j) at row j and column i. The experimental structure is likewise denoted by the symbol E, and where the predicted optimal and experimental structures coincide, a box of size corresponding to unit probability is drawn. The linear sequence of decreasing size boxes is displayed near the base of the matrix. They correspond to logarithmic (base 10) intervals of width 1 (i.e., a factor of ten) centered at $-j$, $j = 0 \cdots 9$. The first interval is $[0, -0.5]$ and all probabilities P_{ij} with $\log_{10} P_{ij} < -9.5$ are displayed as blank. The sequence is displayed along the diagonal in 5′ to 3′ order and repeated with numbering along the edges of the plot. X indicates a nonbonding modified base.

this to occur, one further quantity must be introduced:

$$P_{lu}^{d} = \sum_{l<j\le l+u+1} \frac{P_{j-u-1,j}}{Q^{b}_{j-u-1,j}} e^{-[F_2(j-u-1,j,u)/kT]}$$

The extra term to the second line of Eq. (23) may then be written

$$P_{hl}^{>u_m} = \sum_{u>u_m} Q^{b}_{hl} P^{d}_{lu} \qquad (24)$$

The intermediate sums, P^{d}_{lu}, need only be calculated once for each d in the recursion for P_{hl}. They must be stored in the same array indexed by l and u for each d (to preserve the N^2 storage limitation). The resulting algorithm is order N^3.

The binding probabilities P_i, for the base at position i in the sequence to be bound (including all possible pairs), is simply

$$P_i = \sum_{j=i+1}^{N} P_{ij} + \sum_{j=1}^{i-1} P_{ji} \qquad (25)$$

The pair binding probabilities may be simultaneously displayed in the array of boxes, with size decreasing as the logarithm of the probability, $\{P_{ij}\}$ in the upper triangle $i < j$ (Figure 1). The sequence of bases is displayed along the diagonal and along two edges of the array, and is numbered along the remaining two. The optimal structure found with the same set of thermodynamic parameters is displayed in the lower triangle by setting $P_{ji} = 1.0$ when (i, j) belongs to the structure. It provides a useful mirror image (in the diagonal) comparison in the interpretation of the probability distribution in the upper triangle.

APPLICATIONS TO SPECIFIC RNA MOLECULES

In order to illustrate the new insight won by the present method, this section discusses the probability distributions of base pairing for four representative RNA molecules of general interest: a tRNA ($N = 76$), a 5S ribosomal RNA ($N = 120$), a self-replicating RNA ($N = 220$), and a self-splicing RNA ($N = 414$). The results serve to substantiate the claim that the present method removes the uncertainties about the predictive value of optimal structures obtained from secondary structures. In addition, the skeptics' unease about the huge number of near optimal structures found for sequences of moderate length is shown to be formally correct but practically exaggerated since the many alternatives group into rather well-defined ensembles of structures that can usually be identified as kinetically interchangeable, i.e., many alternative structures are near relatives that can interchange rapidly by "slippage" of several bonding pairs or by single base pairing changes.

A brief discussion of the detailed set of thermodynamic parameters used in the following calculations is necessary here, although the major import of this work is independent of the detailed empirical parameter determination process. The temperature-dependent data of Borer[17] collected in $1M$ NaCl near 25°C, giving enthalpy and entropy contributions for various sequence-dependent 2 : loops with $u = 0$ (stacked pairs), were collated with the Salser[18] collection of loop entropies and free energies of stacked pairs to fill out a complete set of temperature-dependent parameters.[19] The entropies of bulge loops have been corrected according to the subsequent revision.[20] This Borer–Salser–Tinoco data was recently subject to a major extension and review by experiments in the vicinity of 37°C.[21,22] This data, the Turner compilation, like the one above, is incomplete but contains further and corrected parameters. In what follows, one of these two data sets has been used.

One further refinement is necessary, in order to make proper use of the empirical thermodynamic parameters. Stems of length one, which occur as separate states in the above, actually should be lumped with the free chain states, i.e., disallowed in the partition function calculation. This is both because the oligonucleotide studies do not clearly distinguish these states in the optical measurements and because the loop entropy factors are not designed to correctly weight these states, becoming accurate only for stems of length greater than one base pair. Actually, this correction is straightforward with the introduction of an additionally restricted partition function Q_{ij}^{b2}, which sums the weights of all states on the segment $[i, j]$ with base pairs (i, j) and $(i + 1, j - 1)$. Then Q_{ij}^{b} may be used in the recursion to obtain further base pair stacking, but otherwise Q_{ij}^{b2} must be used. In the interests of clarity we omitted this detail from the presentation of sections two and three above, although it is included in the computations presented.

The predictions of optimal structures have been reported to be not very sensitive to small changes in the thermodynamic parameters.[23] For the transfer RNA for the amino acid phenylalanine from *E. coli* we show the comparison of the Salser–Tinoco data with the Turner data, extrapolated to 25°C, by displaying the box matrices of binding probabilities in Figures 1 and 2, respectively. While there is a significant difference in the optimal structure, the ensemble of structures has only minor weighting differences. Transfer RNA has been a testing ground for structural prediction because of its established common tertiary structure. The thermodynamic parameters have been tested on samples of 100 or more tRNA sequences where the failure to predict the common cloverleaf structure for all the sequences should expose any weaknesses in the thermodynamic parameterization. The original data compiled by Salser[18] predicts 67% in the correct conformation and the more recent turner compilation[21] 82%. It is consequently still of interest to examine the equilibrium ensemble of structures predicted by the present method for tRNA in a case where the optimal structure is noncruciform. A full analysis for all presently available tRNAs is beyond the scope of the present paper but would

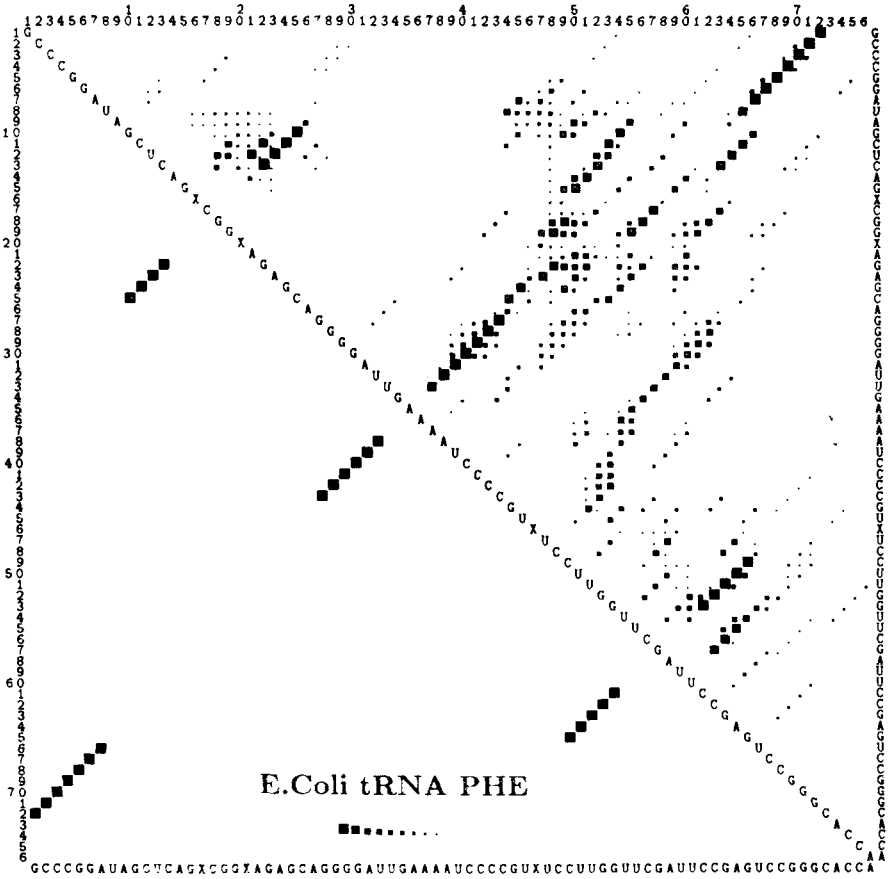

Figure 2. The box matrix for tRNA as in Figure 1, but using thermodynamic parameters from the Turner compilation extrapolated to 25°C.

provide an immediate application of the present method. In the box matrices, illustrated in Figures 1 and 2, it should be noted that, although the optimal structure predicted is not cruciform, the cruciform structure occurs with a moderate probability in the equilibrium ensemble. The strong, near end-to-end binding may be seen in the top right corner and the correct hairpins occur near the diagonal ($i = j$).

As in the case of tRNA, the 5S ribosomal RNA family has been predicted to possess a common secondary structure.[24] Application of an optimal folding algorithm has predicted 68% in the agreed consensus conformation.[25] The box matrix for the *E. coli* 5S RNA is displayed in Figure 3. It reveals the consensus structure[24] as dominant with a significant alternative for the first (upper) hairpin near the diagonal. The suggested kinetic equilibrium between these two forms[24] is seen to actually involve a wider ensemble of related structures. There are now more than 500 5S tRNA sequences[26] and the secondary structure is utilized in the sequence comparisons and evolutionary studies. Since the box matrix gives rather complete information about the structural biases in the sequence, we expect a box matrix analysis of this entire family to be useful in characterizing the important structural features of the molecule and in arriving at an independent measure of evolutionary distance between the parent organisms. In the fifth section a calculated melting curve for this molecule is compared with experiment.

Our interest in RNA folding actually arose in the study of self-replicating RNAs and the structural modification of replication rates for related sequences in the "quasispecies" of RNA,[27] which is found under specific chemical conditions *in vitro*. Work is in progress to identify common features of the known self-replicating RNAs for the $Q\beta$ viral enzyme complex. The midivariant MDV-1 is such an RNA whose sequence evolution has been studied *in vitro*[28] and whose secondary structure is still of interest.[29] The box matrix in Figure 4 reveals that both the structure in which the ends of the

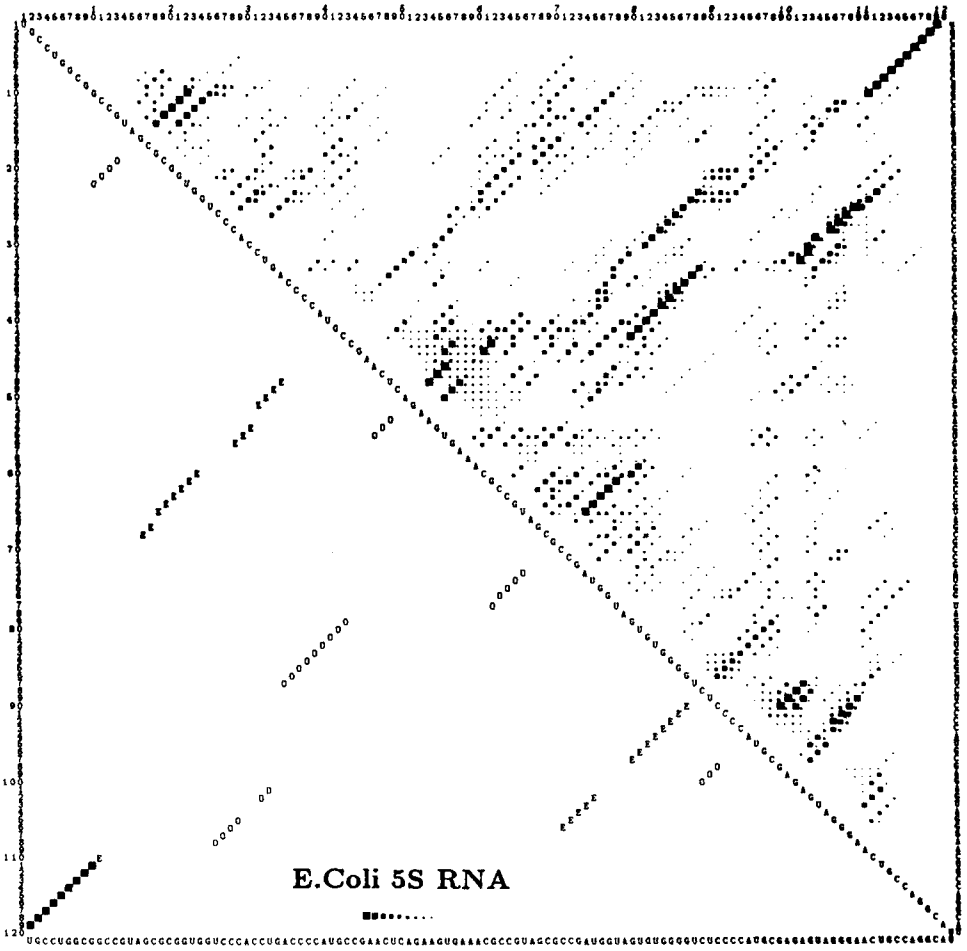

Figure 3. The box matrix for the *Drosophila melanogaster* 5S RNA equilibrium ensemble of binding probabilities using the Turner compilation at 25°C, but otherwise as described in the caption to Figure 1.

molecule are bound to one another and a structure consisting entirely of hairpin loops are significant in equilibrium. Intermediate are a spectrum of structures involving multiple loops, which compete to shorten some of the hairpin stems. The latter are in many cases seen to be not just a single substructure but a rather diffuse ensemble of slippage and single base pair transition related structures.

Finally, the self-splicing *Tetrahymena* IVS intervening sequence ($N = 414$) was investigated because of its biological interest, because it has been the object of several optimal structure prediction studies,[20,30,31] and because of the abundant cleavage data available.[32] The latter allows a correlation with the binding probabilities predicted in the equilibrium ensemble to be made. The pair binding probabilities are displayed in the box matrix of Figure 5. We see that some parts of the secondary structure are much more specifically determined than others and that the entire array of possibilities is not unmanageable. A comparison of the equilibrium binding probabilities P_i on either side of cleavage sites, calculated from Eq. (25), with the strengths of cleavage by double-(CV) and single-(T1, T2, S1) strand specific cleavage enzymes, yields a high correlation, as shown in Table I. We do not mean to suggest that the real structure is a mixture of all structures in the equilibrium ensemble—the equilibration time is likely to be much too long for this and tertiary effects may further bias the ensemble toward a specific structure—although it is certainly not a single structure as often suggested. Rather, we wish to argue here that the most information that can be extracted from the thermodynamic parameters about the binding properties of individual bases in the sequence is summarized in the binding probabilities.

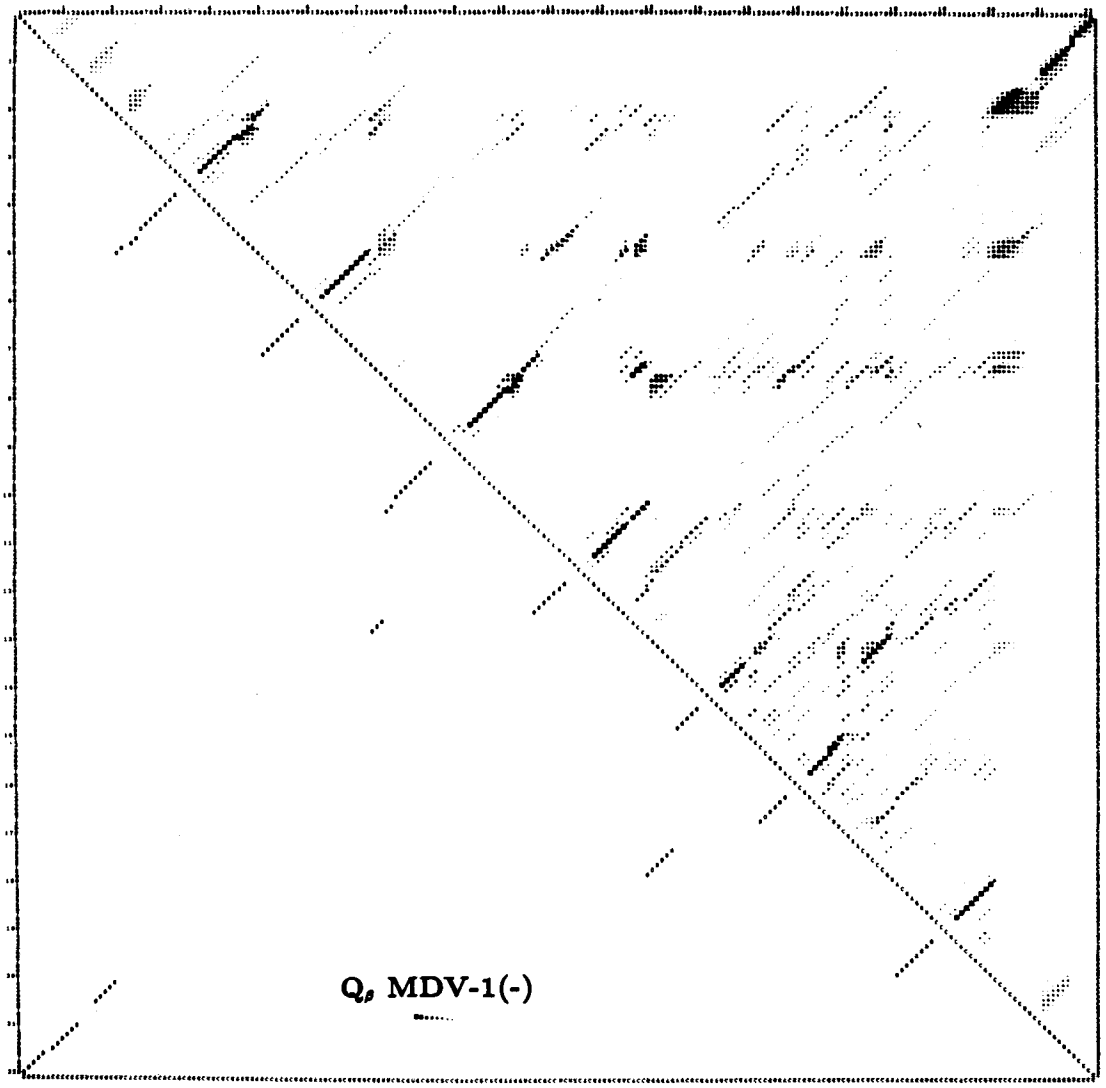

Figure 4. The box matrix for MDV-1, the midivariant RNA, which is self-replicating in the presence of the $Q\beta$ viral replicase *in vitro*, using the Turner compilation at 25°C. Details of the representation are the same as for Figure 1.

Although the implementation of our algorithm is not yet highly optimized, the computation times on a VAX 8650 are reported for the various lengths of RNA discussed above. These times are 1.2, 3.8, 17.3, and 85 min for molecules of length 76, 120, 220, and 414, respectively, and represent the combined computation time for the partition function, for the pair binding probabilities, and for the optimal folding comparison, which is displayed in the lower half of the box matrix. One can see the characteristic N^3 behavior, with the additive constant somewhat masked by page-faulting problems for the longer sequences. We expect a reduction by a further factor of five to be achievable with the same hardware.

THE PARTITION FUNCTION TEMPERATURE DEPENDENCE AND MELTING

Since the overwhelming majority of configurations are in the unfolded state, as reflected by the large entropy loss associated with base pairing and loop formation, the high temperature ensemble is unfolded and, according to our chosen reference point for the entropy of zero for an unfolded chain, the partition function must decrease toward one at high temperature. Differential scanning calorimetry is a widely used tool for investigating the melting of biopolymers.[33] The extraction of a series of two-state transitions from the temperature-dependent specific heat profile is used to gain

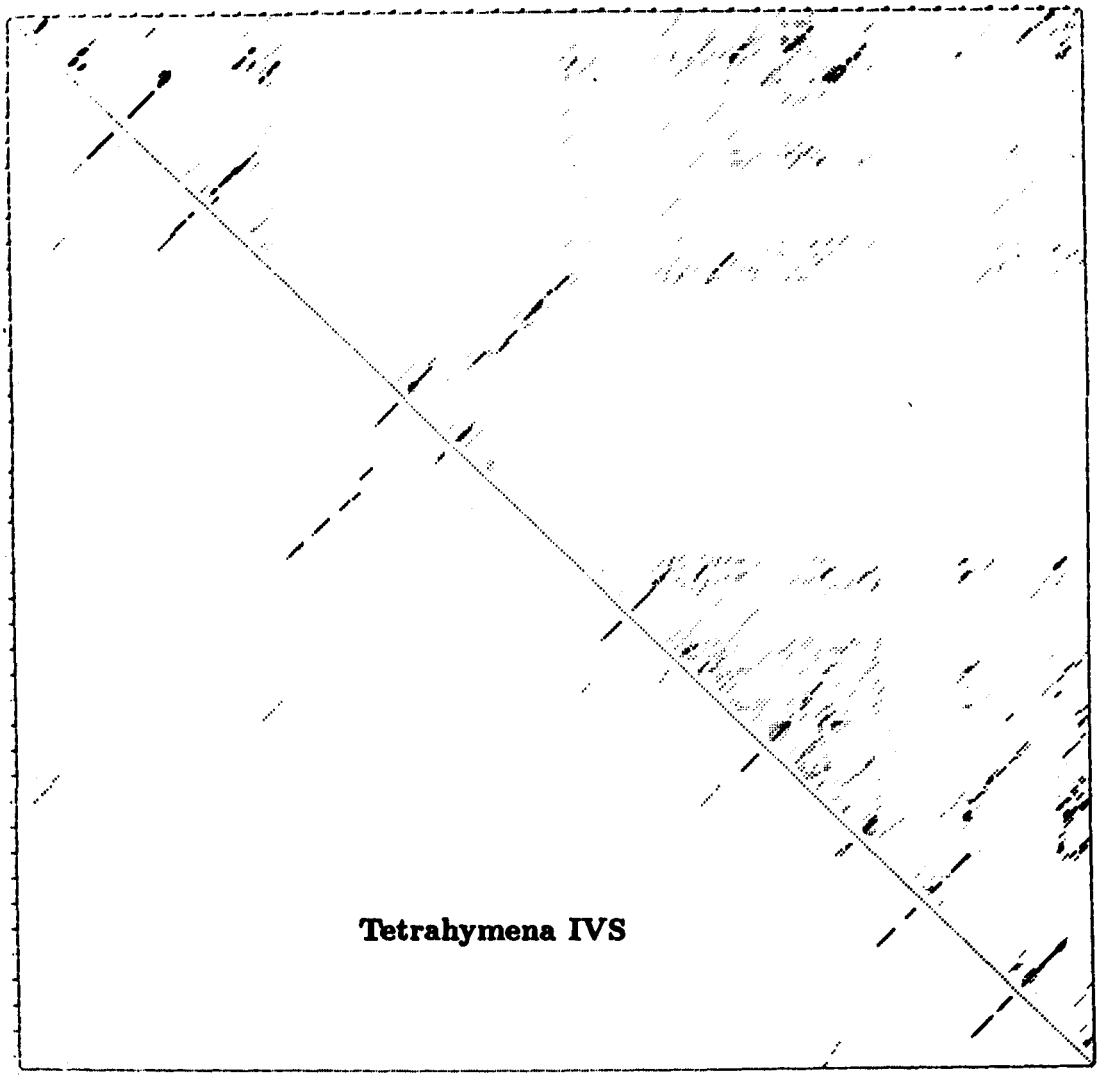

Figure 5. The box matrix for the self-splicing rRNA intervening sequence from *Tetrahymena*. The extreme compression of the picture has obscured the sequence and numbering, but the ensemble of structures remains highly restricted to specific bindings even for sequences of this length ($N = 414$). Details of the representation are the same as for Figure 1 but using the Turner compilation at 25°C.

structural information.[34,35] Here the experimental profile for the ribosomal 5S RNA from *E. coli*[36] are compared with the partition function calculated as a function of temperature. The specific heat C_p is derivable from the partition function *via* the effective Gibbs free energy of the ensemble G and the enthalpy H[37]

$$G = -kT \ln Q$$

$$H = kT^2 \frac{\partial \ln Q}{\partial T} = G - T\frac{\partial G}{\partial T} \qquad (26)$$

$$C = \frac{dH}{dT}$$

where the derivatives are at constant pressure since the empirical parameterization used to construct the partition function was also established at constant pressure. For *E. coli* 5S RNA, the calculated specific heat as a function of temperature is compared with the differential scanning calorimetry curve[36] in Figure 6. The restriction of the partition function to states involving no stems of length one, as discussed above, is necessary if there is a large initiation factor in addition to the apparent change in loop entropy once the stem is longer, and leads to the prediction of the right order of magnitude for the heat capacity. Further corrections, such as the temperature-dependent enthalpy arising from

Table I Cleavage vs Binding Probabilities[a]

i	C_1	C_2	C_3	P_i	P_{i+1}	i	C_1	C_2	C_3	P_i	P_{i+1}
8	S1			0.39	0.038	169	T1			0.0003	0.0003
9	S1			0.038	0.56	192	cv			0.9938	1.0000
12	CV			0.930	0.962	193	cv			1.0000	1.0000
13	CV			0.962	0.959	230	CV			0.9994	0.9994
14	CV			0.959	0.942	231	CV			0.9994	0.9999
17	t2			0.942	0.89	237	T2			0.0000	0.0000
18	t2	s1	cv	0.89	0.936	265	cv			0.985	0.988
19	S1	t2		0.936	0.66	266	cv			0.988	0.951
20	s1	T2	cv	0.66	0.36	269	t2			0.0090	0.054
21	s1	T2	cv	0.36	0.37	272	t1			0.927	0.063
22	T1			0.37	0.40	282	CV			0.985	0.987
23	T1			0.40	0.50	284	cv			0.987	0.9945
25	t1	CV		0.82	0.937	287	T2			0.0091	0.0080
26	t1			0.937	0.68	288	T1	T2		0.0080	0.013
27	t1			0.68	0.29	296	CV			0.988	0.965
36	CV			0.984	0.9998	310	CV			0.54	0.77
38	CV			0.9992	1.0000	317	CV			0.61	0.984
41	t2			0.0001	0.0000	319	CV			0.88	0.9999
42	T2			0.0000	0.0000	324	T2			0.23	0.23
43	t2	S1		0.0000	0.0000	325	T2			0.23	0.0085
44	S1	T1		0.0000	0.0000	343	CV			0.909	0.9980
45	t2			0.0000	0.0001	347	T2			0.0001	0.0000
47	CV			0.9998	1.0000	348	CV	T2		0.0000	0.0000
49	CV			0.9992	0.9998	349	T1			0.0000	0.0001
53	CV			0.9938	0.977	351	cv			0.0001	0.0017
54	CV			0.977	0.9982	353	cv			0.980	0.9965
64	CV			0.9982	0.9993	354	cv			0.9965	0.9963
65	CV			0.9993	0.9992	371	cv			0.9998	0.9996
67	CV			0.9999	0.9998	377	cv			0.035	0.90
68	CV			0.9998	0.0008	378	CV			0.90	0.9950
72	T2	S1		0.0003	0.11	379	CV			0.9950	0.9979
73	T1			0.11	0.11	380	cv			0.9979	0.9945
74	T2	S1		0.11	0.015	385	T2			0.015	0.0024
75	t2			0.015	0.0008	386	T2			0.0024	0.0024
76	T2			0.0008	0.0006	387	T2			0.0024	0.80
77	T1			0.0006	0.0003	390	CV			0.989	0.9981
88	CV			0.81	0.82	396	cv			0.967	0.9928
89	CV			0.82	0.82	397	cv			0.9928	0.9990
97	cv			0.81	0.9974	398	CV			0.9990	0.9988
98	cv			0.9974	0.9978	399	CV			0.9988	0.968
102	cv			0.982	0.965	401	CV			0.9920	0.964
103	cv			0.965	0.12	402	CV			0.964	0.67
117	CV			0.89	0.88	408	t1			0.52	0.51
118	CV			0.88	0.86	409	cv			0.51	0.088
127	CV			0.9997	1.0000	410	cv			0.088	0.29
128	CV			1.0000	1.0000	411	cv			0.29	0.46
144	cv			0.9992	0.9999	412	cv			0.46	0.55
145	cv			0.9999	0.9972	413	cv			0.55	0.40

[a] A comparison of the equilibrium binding probabilities on either side of cleavage sites with the strengths of cleavage by double-(CV) and single-(T1, T2, S1) strand specific cleavage enzymes taken from Ref. 23. Lower-case enzyme entries (cv, t1, t2, s1) refer to weak cleavage. The cleavage is between bases i and $i + 1$ in the sequence, and the calculated binding probabilities P_i and P_{i+1} for these positions are shown in the final columns. The consistency is excellent except at positions 348 and 351 with double (single) strand specific cleavage by CV occurring strongly only where the bond pairing probability is above (below) 75% (35%).

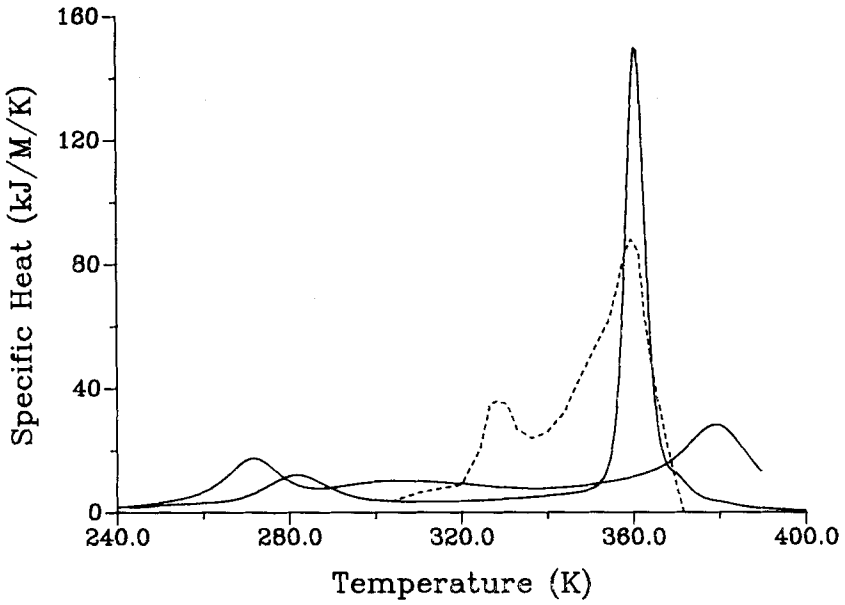

Figure 6. The specific heat for the ensemble of equilibrium structures as a function of temperature calculated from the partition function *via* Eq. (26) for the *E. coli* 5S RNA of figure 3. The temperatures are in degrees Kelvin (K) and the specific heat in kilocalories per mole of RNA per degree. Peaks in the function correspond to structural phase transitions. The dashed line shows the experimental curve for the A form of 5S RNA.[36] The first peak is absent for the B form obtained by renaturing. The lower full curve is that predicted including all helices of length 1 and the upper curve is obtained by disallowing helical stems of length 1. The Turner compilation of thermodynamic parameters was used.

stacking of bases to form single-stranded helices that compete with base pairing,[38] are possible, but the major import of the present article is that a complete calculation is now attainable. Tertiary structures are clearly not necessarily implicated by cooperative melting, which is predicted on the basis of secondary structure for this molecule.

In summary, the specific heat reflects the occurrence of any structural transitions as the temperature increases. It is to be hoped that experiments on model compounds over a wider temperature range will make such predictions quantitatively more reliable. Since these calculations are possible, following the method presented here, we expect the experimental melting curves to now be more useful in the determination of complex RNA secondary structures. A direct calculation of the fraction θ of the bases that are bonded in the equilibrium ensemble as a function of temperature would be useful for a comparison with optical melting data and may be achieved by means of derivation of the partition function with respect to a common base pairing energy multiplier (cf. Refs. 39 and 40), t say,

$$\theta = N^{-1} \frac{\partial \ln Q}{\partial \ln t}$$

where N and Q are the length of sequence and partition function, respectively, as above. On the other hand, the simple technique of summing P_i is not reliable because of the correlations between the pairing of different bases.

CONCLUSIONS

The brief assortment of applications reported here are intended to reveal the predictive capabilities of the partition function approach that the present work has made computationally tractable. All possible secondary structures are included in the calculation. The one major limitation is the restriction to pure secondary structures of the nonknotted type. This is necessary for the problem to be amenable to dynamic programming. Future work must incorporate the probabilistic information delivered by the box matrix into a more detailed tertiary structure analysis—using it to restrict the

search for an optimal structure. At present the box matrix provides an important tool for the assessment of the reliability of optimal structure predictions on the basis of secondary structures, for the consideration of alternative structures, and for establishing the sensitivity of the prediction to thermodynamic parameters. It portrays the equilibrium ensemble of structures and may assist in changing the heavy but misguided emphasis on single unique structures for biological macromolecules. With some knowledge of kinetic rates, it allows the kinetic ensemble of structures that interchange within significant time periods to be discerned. Furthermore, the structural transitions with increasing temperature may be predicted and the melting of the ensemble followed *via* the specific heat. We hope that further structural experiments will be encouraged by the tractability of this sequence-specific calculation of the equilibrium ensemble of structures for this important class of informational macromolecules.

The close relationship found in the present work between the structure superposed from the individually most probable bindings, which may be directly read from the box matrix, and the structure of minimum free energy (the most probable structure in the ensemble) is encouraging for the application of the superposition hypothesis to protein folding.[41]

The author wishes to acknowledge useful discussions with G. Bauer and with C. Biebricher concerning the representational tool employed, and the assistance of B. Lindemann in accessing a compilation of 5S RNA sequences. K. Nieselt, G. Koch, and especially T. Kneser provided assistance with the TEX[42] macro for converting line printer output to the box matrices, and M. Meyer performed the photoreduction of the figures. The comments of one of the referees proved helpful.

REFERENCES

1. Tinoco, I., Jr., Uhlenbeck, O. C. & Levine, M. D. (1971) *Nature* **230**, 362–367.
2. Tinoco, I., Jr., Borer, P. N., Dengler, B. & Levine, M. D. (1973) *Nature New Biol.* **246**, 40–41.
3. Pipas, J. M. & McMahon, J. E. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 2017–2021.
4. Auron, P. E., Rindone, W. P., Vary, C. P. H., Celantano, J. J. & Vournakis, J. N. (1982) *Nucleic Acid. Res.* **10**, 403–419.
5. Bellman, R. & Kaluba, R. (1960) *J. SIAM* **8**, 582–588.
6. Zuker, M. & Sankoff, D. (1984) *Bull. Math. Biol.* **46**, 591–621.
7. Waterman, M. S. (1978) in *Studies in Foundations and Combinatorics, Advances in Mathematics Supp. Studies*, Vol. 1, Academic Press, New York, pp. 167–212.
8. Waterman, M. S. & Smith, T. F. (1978) *Math. Biosci.* **42**, 257–266.
9. Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. (1978) *SIAM J. Appl. Math.* **35**, 68–82.
10. Sankoff, D. (1985) *SIAM J. Appl. Math.* **45**, 810–825.
11. Waterman, M. S. & Smith, T. F. (1986) *Adv. Appl. Math.* **7**, 455–464.
12. Waterman, M. S. & Byers, T. H. (1985) *Math. Biosci.* **77**, 179–188.
13. Gibbs, A. J. & McIntyre, G. A. (1970) *Eur. J. Biochem.* **16**, 1–11.
14. Trifonov, E. N. & Bolshoi, G. (1983) *J. Mol. Biol.* **169**, 1–13.
15. Jacobson, H. & Stockmayer, W. H. (1950) *J. Chem. Phys.* **18**, 1600–1606.
16. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Res.* **9**, 133–148.
17. Borer, P. N., Dengler, B. & Tinoco, I. Jr. (1974) *J. Mol. Biol.* **86**, 843–853.
18. Salser, W. (1977) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 985–1002.
19. McCaskill, J. S. (1984) unpublished material.
20. Cech, T. R., Tanner, N. K., Tinoco, I., Jr., Weir, B. R., Zuker, M. & Perlman, P. S. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3903–3907.
21. Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 9373–9377.
22. Freier, S. M. & Turner, D. H. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 123–133.
23. Nussinov, R., Tinoco, I., Jr. & Jacobson, A. B. (1982) *Nucleic Acids Res.* **10**, 341–349.
24. De Wachter, R., Chen, M. W. & Vandenburghe, A. (1982) *Biochimie* **64**, 311–329.
25. Papanicolou, C., Gouy, M. & Ninio, J. (1984) *Nucleic Acids Res.* **12**, 31–44.
26. Wolters, J. & Erdmann, V. A. (1988) Compilation of 5S rRNA and 5S rRNA gene sequences, preprint.
27. Eigen, M., McCaskill, J. S. & Schuster, P. (1989) *Adv. Chem. Phys.* **75**, 149–263.
28. Mills, D. R., Peterson, R. L. & Spiegelmann, S. (1967) *Proc. Natl. Acad. Sci. USA* **58**, 217–224.
29. Biebricher, C. K., Diekmann, S. & Luce, R. (1982) *J. Mol. Biol.* **154**, 629–648.
30. Michel, F. & Dujon, B. (1983) *EMBO J.* **2**, 33.
31. Waring, R. B., Scazzocchio, T. A., Brown, T. A. & Davies, R. W. (1983) *J. Mol. Biol.* **167**, 595.
32. Been, M. D., Barfod, E. T., Burke, J. M., Price, J. V. Tanner, N. K., Zaug, A. J. & Cech, T. R. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 147–157.
33. Sturtevant, J. M. (1987) *Ann. Rev. Phys. Chem.* **38**, 463–488.

34. Kidokoro, S. & Wada, A. (1987) *Biopolymers* **26**, 213–229.
35. Chang, L. H. & Marshall, A. G. (1986) *Biopolymers* **25**, 1299–1313.
36. Matveev, S. V., Filimonov, V. V. & Privalov, P. L. (1982) *Mol. Biol.* **16**, 990–999.
37. Poland, D. & Scheraga, H. A. (1970) *The Theory of Helix-Coil Transitions in Biopolymers*, Academic Press, New York, pp. 283ff.
38. Pörschke, D., Uhlenbeck, O. C. & Martin, F. H. (1973) *Biopolymers* **12**, 1313–1335.
39. Zimm, B. H. (1960) *J. Chem. Phys.* **33**, 1349–1356.
40. Appleby, D. W. & Kallenbach, N. R. (1973) *Biopolymers* **12**, 2093–2120.
41. Pain, G. H. & Scheraga, H. A. (1985) *Biopolymers* **24**, 1391–1436.
42. Knuth, D., (1984) *The Textbook*, Addison-Wesley.

Received April 29, 1988
Accepted November 30, 1988