

John SantaLucia, Jr.¹

Douglas H. Turner²

¹ Department of Chemistry,
Wayne State University,
Detroit, MI 48202

² Department of Chemistry,
University of Rochester,
Rochester, NY 14627

Measuring the Thermodynamics of RNA Secondary Structure Formation

Received and accepted 22 October 1997

Abstract: *The thermodynamics of RNA secondary structure formation in small model systems provides a database for predicting RNA structure from sequence. Methods for making these measurements are reviewed with emphasis on optical methods and treatment of experimental errors. Analysis of experimental results in terms of simple nearest-neighbor models is presented. Some measured sequence dependences of non-Watson-Crick motifs are discussed.* © 1998 John Wiley & Sons, Inc. *Biopoly* 44: 309–319, 1997

Keywords: *RNA; secondary structure; thermodynamics; optical melting; measurement errors*

INTRODUCTION

Interest in RNA structure has been sparked by the discovery of catalytic functions for RNA,^{1–3} the targeting of RNA by therapeutics designed to bind by base pairing,^{3–6} and the explosion in available sequence information.⁷ Three-dimensional structures determined by x-ray crystallography and nmr reveal that RNA can form a myriad of different shapes.⁸ In principle, it is possible to predict the equilibrium shape of an RNA from thermodynamic principles. In practice, our limited knowledge of the details of RNA interactions and energetics prevents this. A major step toward generating the required knowledge base is an understanding of the thermodynamics of secondary structure formation. The secondary structure provides the largest constraint on possible three-dimensional foldings of an RNA chain. Determination of a secondary structure permits selection of motifs suitable for three dimen-

sional structure analysis, insights into structure–function relationships, and leads for designing potential therapeutics. In addition, the interactions determining secondary structure are probably important for determining three dimensional structure. Thus a major goal of thermodynamic studies of RNA is to provide a knowledge base able to predict secondary structure from sequence. Even this is a daunting task. As illustrated in Figure 1, there are many different motifs in an RNA secondary structure, and the number of possible sequence combinations for most motifs is large. Thus thermodynamic studies attempt to provide models both simple enough to use and good enough to provide reasonable approximations. Most thermodynamic studies on RNA have investigated secondary structure formation in oligonucleotides. Here we review the experimental details of these measurements, some recent results, and the concomitant progress toward structure prediction. We focus on free energy

Correspondence to: Douglas H. Turner or John SantaLucia, Jr.
Contract grant sponsor: National Institutes of Health
Contract grant number: GM22939

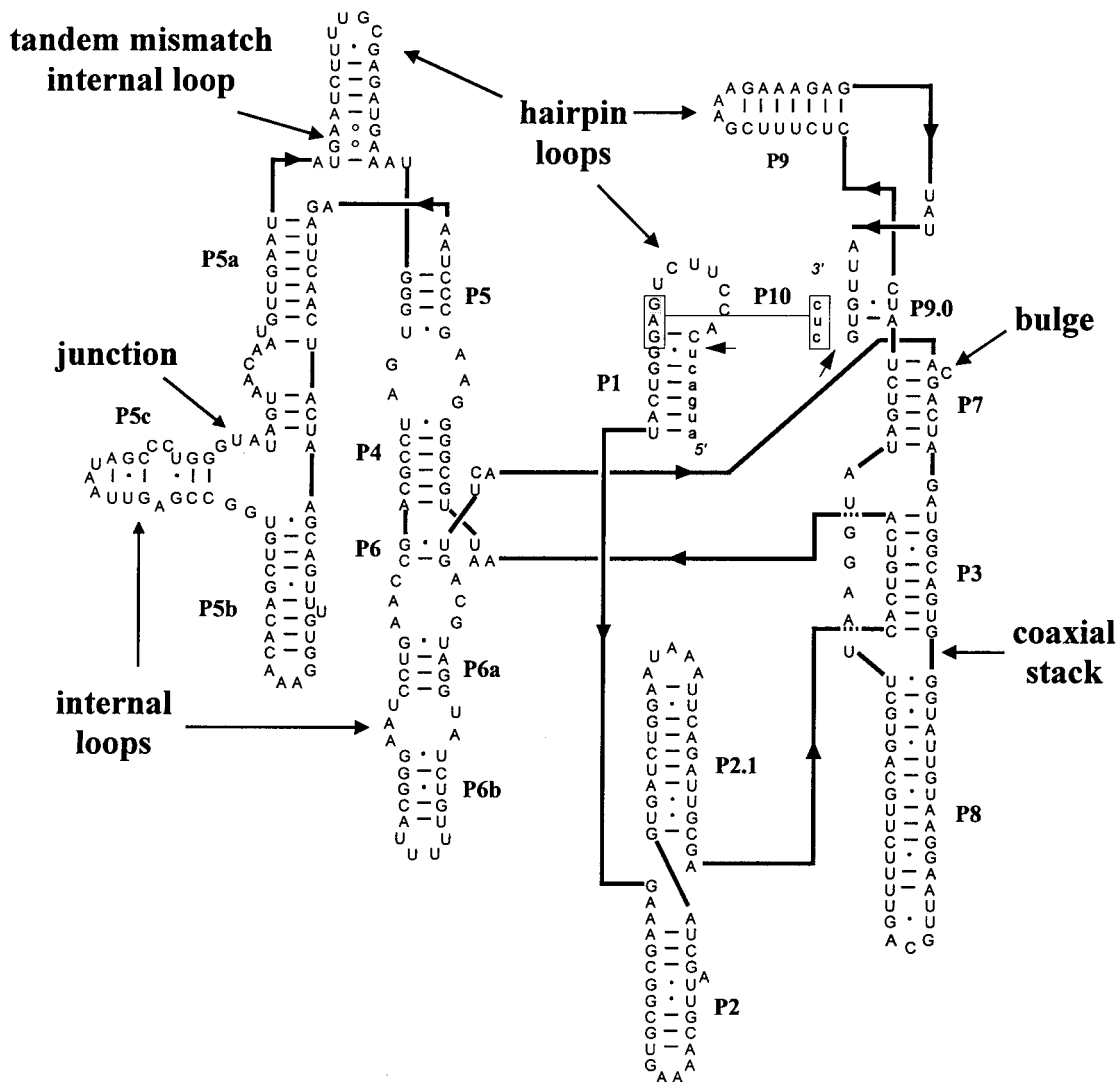


FIGURE 1 The secondary structure of the group I intron of mouse-derived *Pneumocystis carinii*.⁷⁶ Arrows indicate motifs that are found in the secondary structures of many different biologically important RNAs for which more thermodynamic data are required to improve secondary structure predictions.

changes associated with secondary structure formation at 37°C, ΔG_{37}° , since these are the most important parameters for structure prediction at the temperature of the human body. A review of various thermodynamic parameters for particular motifs along with examples of typical applications has recently been presented.⁹

MEASUREMENT OF THERMODYNAMIC PARAMETERS

In this section, methods for measuring thermodynamic parameters for secondary structure formation

are discussed, along with considerations important for error analysis. Emphasis is placed on measurements by optical melting methods since they provide accurate ΔG_{37}° parameters and are fast enough to allow studies of a large range of motifs and sequences.

Experimental Methods

Comparison of Optical Melting and Calorimetry. Two widely used methods for determining nucleic acid thermodynamics are absorbance melting curves^{10,11} and microcalorimetry, including differen-

tial scanning calorimetry (DSC) and isothermal titration calorimetry.¹² Both methods have distinct advantages and disadvantages and are complementary. Optical detection of thermal denaturation offers the advantages of high sensitivity so that small sample sizes are required; typically ~ 3 A260 units are required for a full set of measurements on a sequence. Values of ΔH° and ΔS° can be derived from a two-state van't Hoff analysis of optical melting data for oligonucleotide samples^{10,13–16} and the standard errors in ΔH° and ΔS° are typically about 5–8%. Due to compensating errors in ΔH° and ΔS° ,^{10,17} a van't Hoff analysis of optical melting curves provides very precise measurements of ΔG_{37}° (standard error ± 2 –5%) and melting temperature, T_M (± 0.5 –1°C; see below). Values of ΔG° are particularly accurate near the T_M , so that sequences are often designed to melt near 37°C. Even the ΔG° value for a duplex that does not melt in a two-state manner, (A₇U₇)₂, appears to be reasonably accurate near the T_M of the measurement.¹⁸

Microcalorimetry offers the advantages that transition enthalpy changes are directly measured, are model independent, and with recent improvements in instrument design, are accurate to 2–5%.¹⁹ Calorimetric methods, however, require substantially larger sample sizes, typically 20–50 A260 units for a full set of measurements on a sequence. In DSC, the excess heat capacity ΔC_p of the cell with nucleic acid minus the matched control cell with only buffer, is plotted vs the temperature and the area between this curve and the baseline provides the ΔH° for the unfolding of the nucleic acid. The same data can be used in a plot of $\Delta C_p/T$ vs T , and the area between this curve and the baseline provides ΔS° for the unfolding. While accurate ΔH° and ΔS° are generally obtained from DSC, the errors in ΔH° and ΔS° are apparently uncorrelated and therefore large errors in ΔG_{37}° and T_M are possible. Thus the calorimetric ΔH° is often used in conjunction with the T_M determined from calorimetry or optical melting and the oligonucleotide concentration to obtain the entropy for the reaction from

$$\Delta G_T^\circ = \Delta H^\circ - T\Delta S^\circ \quad (1)$$

where ΔG° at the T_M equals $RT_M \ln C_T$ for self-complementary duplexes and $RT_M \ln(C_T/4)$ for nonself-complementary duplexes. This amounts to applying the two-state approximation to the data.

Optical Melting. The simplest way to derive thermodynamic parameters from optical melting data is to apply a van't Hoff analysis to the data, though

more rigorous approaches can also be applied.^{20,21} Thermodynamic parameters for duplex formation are obtained from absorbance vs temperature melting curves by two methods: (1) the shape of individual curves are fit and (2) for self-complementary duplex formation, the T_M is measured at several different oligonucleotide concentrations C_T and fit to Eq. (2) by plotting reciprocal melting temperature (T_M^{-1}) vs $\ln C_T$.²²

$$T_M^{-1} = (R/\Delta H^\circ) \ln C_T + \Delta S^\circ/\Delta H^\circ \quad (2)$$

For nonself-complementary duplexes, C_T is replaced by $C_T/4$. Both methods assume that the equilibrium involves only two states, duplex and random coil. The difference in standard state heat capacities ΔC_p° of the duplex and random coil states is usually assumed to be zero.^{10,13,23} There is evidence that ΔC_p° is usually small for short oligonucleotides.²⁴ Recent work, however, suggests that the integrated van't Hoff equation should be applied in fitting curves and $1/T_M$ vs $\ln C_T$ plots, though the typical noise level in optical melting experiments may not justify this treatment.^{10,13,19,25} The details of one commonly used program for fitting optical melting curves was published recently.¹⁶

Criteria for Testing the Validity of the Two-State Approximation. The criterion typically used to indicate two-state behavior is agreement between the ΔH° obtained by different methods that depend differently on the two-state approximation.^{10,11,26} Agreement within 10% of ΔH° parameters from a $1/T_M$ vs $\ln C_T$ plot and from fits of the shapes of melting curves is generally regarded as required for a two-state transition,^{27,28} though caution is recommended. This criterion suggests the standard error in the van't Hoff ΔH° parameters from optical melting is about 5–8%. In most cases that have been studied by both optical melting and calorimetry, the ΔH° 's are in agreement if the two ways of analyzing the optical data give the same value within 10%.^{23,24,29}

Agreement between enthalpy changes determined by different methods is a necessary but not a sufficient criterion to definitively establish two-state behavior.^{27,30} For example, the ΔH° parameters from a $1/T_M$ vs $\ln C_T$ plot and from fits of the shapes of optical melting curves agree within 10% for the single strand to duplex transition for the self-complementary DNA sequence (CGTTGCGTAACG)₂, yet the temperature dependence of nmr spectra and comparison of this sequence with thermodynamics

for other sequences revealed that it actually melts through a hairpin intermediate.³⁰ An alternative criterion for two-state thermodynamics is to compare the enthalpy changes from optical melting and calorimetry.^{18,23,29} For transitions with large ΔC_p° , it appears that van't Hoff analyses and calorimetric data provide systematically different ΔH° parameters and the possible origins of these differences have been recently reviewed.^{19,25} Even agreement between van't Hoff analysis of optical melting and calorimetry does not guarantee a two-state transition. The ΔH° 's for the DNA duplex, (GCGTACGCATGCG · CGCATGTGTACGC), are in agreement, but the transition is not two state as evidenced by the melting of the individual single strands.^{30,32} Known exceptions of this type are rare, however.

In cases where the ΔH° parameters from different methods have marginal agreement (e.g., agreement of $\pm 20\%$ of ΔH° from fits of the curve shapes and from a $1/T_M$ vs $\ln C_T$ plot), the data from the $1/T_M$ vs $\ln C_T$ plot appear to be more reliable since these data generally agree with the calorimetrically determined ΔH° .²⁹ Previous work has indicated that the T_M is relatively insensitive to nontwo-state behavior and to the choice of baselines.^{10,29}

A van't Hoff analysis of optical melting data cannot be used to reliably measure the thermodynamics of molecules with nontwo-state transitions. Usually, however, it is possible to design oligonucleotides to minimize possibilities for alternative structures. Since long self-complementary sequences have a tendency to form hairpin intermediates, it is generally advisable to use short or nonself-complementary sequences whenever possible. In general, a van't Hoff analysis of optical melting data provides reliable ΔH° , ΔS° , and ΔG_{37}° parameters for carefully designed duplexes that are shorter than 16 base pairs.

Error Analysis for Experiments

Experimental Errors in ΔH° and ΔS° . The main sources of errors in optical melting are (1) signal-to-noise ratio of the data, (2) random errors due to fluctuations in sample preparation (e.g., oligonucleotide concentration, purity, volume, salt concentration, pH, errors in mixing, etc.), (3) systematic errors due to incorrect instrument calibration, (4) systematic errors introduced as a result of poor oligonucleotide design (e.g., a sequence that can form intermediates), (5) systematic errors due to incorrect assignment of baselines, and (6) systematic errors due to imposing the two-state approximation and from assuming that ΔC_p° is zero. An im-

portant distinction is the difference between precision, which reflects the experimental reproducibility of the data, and accuracy, which reflects how well the experimental measurement agrees with the real value if a perfect measurement were made.³³ The first two sources of error are easy to quantify by simply reproducing one's data or analyzing the sampling error in a $1/T_M$ vs $\ln C_T$ plot or the sampling error in the fitted data. The theory for determining sampling errors in the ΔG_{37}° , ΔH° , and ΔS° from the linear regression of the $1/T_M$ vs $\ln C_T$ plot using standard statistical analysis³⁴ has been previously described.¹⁷ The best method to quantify systematic errors (numbers 3–6 above) is to compare thermodynamic measurements on the same oligonucleotides that were independently determined from different laboratories utilizing different instrumentation and techniques. For DNA duplexes, an estimate for systematic errors 3 and 5 can be derived from results on three sequences (CGATATCG, GAA-GCTTC, and GGAATTCC) that have been independently measured by two groups.^{14,15,23} The average deviations for ΔG_{37}° , ΔH° , ΔS° , and T_M for these sequences are 3%, 6%, 6%, and 1.0°C, respectively. These differences provide reasonable expectations for systematic errors 3 and 5.

Propagation of errors in ΔH° and ΔS° to ΔG_{37}° and T_M . It is important to understand how errors in ΔH° and ΔS° propagate to give the error in ΔG_{37}° . Experimental ΔH° and ΔS° parameters are not independently determined, but instead are highly correlated, with a typical $R^2 > 0.99$.^{10,30,35} This enthalpy–entropy compensation results in errors in ΔG_{37}° that are much smaller than what would be expected if ΔH° and ΔS° were uncorrelated. Equation 4.8 of Bevington³³ provides the equation for error propagation for a general function $x = f(u, v)$:

$$(\sigma_x)^2 = (\sigma_u)^2(\partial x/\partial u)^2 + (\sigma_v)^2(\partial x/\partial v)^2 + 2(\sigma_{uv})^2(\partial x/\partial u)(\partial x/\partial v) \quad (3)$$

Performing the appropriate differentiation of Eq. (1) and substitution into Eq. (3) gives the equation for the propagation of error from ΔH° and ΔS° , $\sigma_{\Delta H^\circ}$ and $\sigma_{\Delta S^\circ}$, to give the error in ΔG_{37}° , $\sigma_{\Delta G_{37}^\circ}$ (Ref. 17):

$$(\sigma_{\Delta G_{37}^\circ})^2 = (\sigma_{\Delta H^\circ})^2 + T^2(\sigma_{\Delta S^\circ})^2 - 2T(\sigma_{\Delta H^\circ \Delta S^\circ})^2 \quad (4)$$

where $(\sigma_{\Delta H^\circ \Delta S^\circ})^2$ is the covariance between ΔH° and ΔS° , and T is 310.15 K. An alternative equation ex-

presses the covariance in terms of the correlation coefficient of a plot of ΔH° vs ΔS° , $R_{\Delta H^\circ \Delta S^\circ}$ (Ref. 36):

$$\begin{aligned} & (\sigma_{\Delta G_{37}^\circ})^2 \\ &= (\sigma_{\Delta H^\circ})^2 + T^2(\sigma_{\Delta S^\circ})^2 - 2T(R_{\Delta H^\circ \Delta S^\circ})\sigma_{\Delta H^\circ}\sigma_{\Delta S^\circ} \quad (5) \end{aligned}$$

Performing the appropriate differentiation of Eq. (2) and substitution into Eq. (3) gives the equation for the propagation of $\sigma_{\Delta H^\circ}$ and $\sigma_{\Delta S^\circ}$, to give the error in T_M , σ_{T_M} (Ref. 30):

$$\begin{aligned} \sigma_{T_M}^2 &= (\sigma_{\Delta H^\circ} T_M / \Delta H^\circ)^2 + (\sigma_{\Delta S^\circ} T_M^2 / \Delta H^\circ)^2 \\ &\quad - 2T_M^3 R_{\Delta H^\circ \Delta S^\circ} \sigma_{\Delta H^\circ} \sigma_{\Delta S^\circ} / (\Delta H^\circ)^2 \quad (6) \end{aligned}$$

This equation assumes that there is negligible error in the $\ln C_T$ term. This is reasonable because a 10% error in C_T propagates to a 1% error in $\ln C_T$ at oligonucleotide concentrations in the range of $10^{-5}M$. The enthalpy–entropy compensation effect is evident in the high quality of predictions made for ΔG_{37}° and T_M .^{13,30}

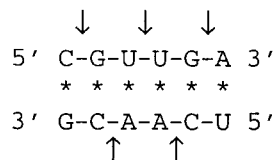
Sequence Dependence of Watson–Crick Pairs

In known secondary structures of RNA, the average length of helices containing all Watson–Crick pairs is roughly 6 base pairs,³⁷ but a wide range in length is found. With current technology, it is not possible to measure thermodynamic parameters for all possible helices containing only Watson–Crick base pairs. Thus models have been developed to predict the thermodynamics of helix formation from a limited set of measurements. A simple base-pairing model is not adequate. For example, at 37°C, the free energy changes of duplex formation for two helices which each contain six GC pairs, (GCC-GGCp)₂ and (CGCGCGp)₂, are -11.2 and -9.1 kcal/mol, respectively.³⁸ Thus, nearest-neighbor models have been developed in which the thermodynamic parameters are dependent on the base-pair doublets present in a sequence.^{22,30,39,40} Experimental tests of the nearest-neighbor model indicate that it is also an approximation. For example, for two duplexes with identical nearest-neighbor composition, (CAGCUG)₂ and (CUGCAG)₂, ΔG_{37}° 's of duplex formation are -6.7 and -7.1 kcal/mol, respectively.⁴¹ The average difference in ΔG_{37}° for nine such pairs was 6%, suggesting a nearest-neighbor model is often reasonable.⁴¹ Two such models are described below.

Model of Borer et al.²² The most popular nearest-neighbor model for predicting stabilities of RNA duplexes with only Watson–Crick pairs is that proposed by Borer et al.²² For any given property, e.g., ΔG_{37}° , there are 12 parameters in the model: one for each of the 10 different doublets of Watson–Crick pairs, one for duplex initiation when the helix has at least one GC pair, and one for duplex initiation when the helix has only AU pairs. The model is easy to apply, as illustrated in Figure 2A. Parameters for the 10 Watson–Crick doublets and for initiation with at least one GC have been determined based on optical melting studies of 45 duplexes.¹³ For these 45 sequences, the difference between measured and predicted ΔG_{37}° is 4% or 0.3 kcal/mol on average. This is similar to the agreement in experimental values for duplexes with the same nearest neighbors.⁴¹

In the model of Borer et al.,²² terminal and internal nearest neighbors are considered equivalent. Gray has pointed out that if terminal base pairs are different, then 14 parameters are required to model the sequence dependence of stability.⁴⁰ Due to sequence constraints, however, only 12 parameters can be determined experimentally.⁴⁰ Based on Gray's work, Allawi and SantaLucia³⁰ have revised the Borer et al.²² model so that the initiation parameter depends on the identities of the terminal base pairs. In this model, the initiation parameter includes both initiation and correction terms for any difference between internal and terminal base pairs (see Figure 2B). This model has been applied to duplex formation by DNA oligonucleotides.³⁰ Application to the RNA data set of Freier et al.¹³ gives an average difference between measured and predicted ΔG_{37}° 's of 3% or 0.2 kcal/mol, slightly better than the results with the Borer et al.²² model.¹³ The Freier et al.¹³ data set, however, contains no sequences with only AU pairs, so a comparison of the full Borer et al. model with 12 parameters to the 12 parameter model of Allawi and SantaLucia³⁰ is not yet possible.

Extensions from simple nearest neighbor to other models can also be envisaged. For example, with enough data, the model of Borer et al.²² can be extended to make all terminal doublets different from internal doublets. In such an expanded model, however, it is necessary to make at least 3 assumptions about the relative magnitudes of parameters. Extensions to a base triplet model (i.e., next nearest-neighbor model) are also possible.⁴² Such models would account for expected nonnearest-neighbor effects on base stacking in single strands. The simple doublet model, however, gives reasonable approximations for stabilities of duplexes containing only



A) Borer et al. (22)

$$\begin{aligned}
 \Delta G_{37}^{\circ}(\text{Total}) &= \Delta G_{37}^{\circ}(\text{CG/GC}) + \Delta G_{37}^{\circ}(\text{GU/CA}) + \Delta G_{37}^{\circ}(\text{UU/AA}) + \Delta G_{37}^{\circ}(\text{UG/AC}) \\
 &+ \Delta G_{37}^{\circ}(\text{GA/CU}) + \Delta G_{37}^{\circ}(\text{init at GC}) + \Delta G_{37}(\text{sym})
 \end{aligned}$$

B) Allawi & SantaLucia (30)

$$\begin{aligned}
 \Delta G_{37}^{\circ}(\text{Total}) &= \Delta G_{37}^{\circ}(\text{CG/GC}) + \Delta G_{37}^{\circ}(\text{GU/CA}) + \Delta G_{37}^{\circ}(\text{UU/AA}) + \Delta G_{37}^{\circ}(\text{UG/AC}) \\
 &+ \Delta G_{37}^{\circ}(\text{GA/CU}) + \Delta G_{37}^{\circ}(\text{init. w/ term. G-C}) \\
 &+ \Delta G_{37}^{\circ}(\text{init. w/ term. A-U}) + \Delta G_{37}(\text{sym})
 \end{aligned}$$

C) Gray (40)

$$\begin{aligned}
 \Delta G_{37}^{\circ}(\text{Total}) &= \Delta G_{37}^{\circ}(\text{ECGE}'/\text{E}'\text{GCE}) + \Delta G_{37}^{\circ}(\text{EGUE}'/\text{E}'\text{CAE}) + \Delta G_{37}^{\circ}(\text{EUUE}'/\text{E}'\text{AAE}) \\
 &+ \Delta G_{37}^{\circ}(\text{EUGE}'/\text{E}'\text{ACE}) + \Delta G_{37}^{\circ}(\text{EGAE}'/\text{E}'\text{CUE}) - 2 \Delta G_{37}^{\circ}(\text{ECE}'/\text{E}'\text{GE}) \\
 &- 2 \Delta G_{37}^{\circ}(\text{EUE}'/\text{E}'\text{AE}) + \Delta G_{37}(\text{sym})
 \end{aligned}$$

FIGURE 2 Application of three nearest-neighbor models for the prediction of ΔG_{37}° for the duplex CGUUGA · UCAACC. The arrows point to the middle of each nearest-neighbor dimer. In the equations, a slash, (/) indicates hydrogen bonding between strands in antiparallel orientation. (A) The model of Borer et al.²² with one initiation parameter per helix and 10 nearest neighbors. (B) The model of Allawi and SantaLucia³⁰ with two initiation parameters per helix and 10 nearest neighbors. (C) The model of Gray.⁴⁰ E and E' indicate the 5' and the 3' ends of the strands, respectively. Although based on different underlying physical models, the models of Allawi and SantaLucia³⁰ and Gray⁴⁰ are statistically equivalent and always make equal predictions. Note that $\Delta G_{37}(\text{sym})$ is zero for any nonself-complementary sequence, including the one shown.

Watson–Crick pairs. In contrast, as described later, good approximations are not available for many other secondary structure motifs.

Model of Gray.⁴⁰ Based on a nearest-neighbor model developed by Gray and Tinoco⁴³ for spectroscopic properties of nucleic acids, and extended to thermodynamic properties,^{39,44} Gray⁴⁰ has provided an alternative analysis for thermodynamic properties of duplex formation. This model uses 12 parameters associated with “independent short sequences.” These 12 parameters are combinations of the 14 parameters in the model, since only 12 parameters can be determined because sequence

constraints limit the number of possible independent equations.⁴⁰ The 12 available parameters can be partitioned in different ways.⁴⁰ For duplex formation, Gray has chosen a set consisting of base pairs and base pair doublets with ends, e.g., EGE'/ECE', a base pair with ends, and ECAE'/EUGE', a base-pair doublet with ends. An application of this model is shown in Figure 2C. Application of Gray's model to the RNA data of Freier et al.¹³ gives an average difference between measured and predicted ΔG_{37}° 's of 3% or 0.2 kcal/mol. Although based on different underlying physical models, the Allawi and SantaLucia³⁰ and Gray⁴⁰ models are statistically equivalent and always make equal predictions.

Propagation of Experimental Errors to the Nearest Neighbor Parameters for Watson–Crick Pairs.

It is important to consider how experimental errors in measured thermodynamics propagate to the nearest-neighbor (NN) parameters described above. Consider a 10 base pair duplex, with 9 nearest-neighbor interactions and an initiation parameter. The total error is thus:

$$(\sigma_{\text{Total}})^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_9^2 + \sigma_{\text{init}}^2$$

+ covariance terms among σ_{NN} and σ_{init} .

Our experience is that the different NN errors (σ_1 through σ_9) are usually similar in magnitude for a set of oligonucleotides with even representation of the different nearest neighbors, that the error from σ_{init} is essentially completely canceled by its covariation with the NN errors, and that covariances between the neighbors are negligible.¹⁴ Thus, we get $(\sigma_{\text{Total}})^2 = 9\sigma_1^2$ for a decamer duplex. Assuming that this 10 base pair duplex had a ΔG_{37}° of 10 kcal/mol and an error of 5%, then the expected average error contribution from each NN is $\sigma_{\text{NN}} = \pm 0.5$ kcal/mol/ $(9)^{1/2} = \pm 0.17$ kcal/mol. If we were to make measurements on 100 different oligonucleotides, then the average error would follow the expected $1/(N - v)^{1/2}$ dependence [see Ref. 33, Eqs. (5)–(13)], where N is the number of experimental measurements and v is the number of parameters determined from the data (in this case $v = 11$ because there are 10 NN and an initiation parameter). For a set of 100 measurements, the expected error for each nearest-neighbor parameter is $\sigma_{\text{NN}} = \pm 0.17$ kcal/mol/ $(100 - 11)^{1/2} = \pm 0.02$ kcal/mol. A real calculation of the propagated errors would weight each experimental value in the fit according to the actual experimental error and covariation among the NN would need to be accounted for. The method of singular value decomposition does all of this in a fashion that is transparent to the user, but is rigorously correct.⁴⁵ This example illustrates the value of performing a large number of measurements to reduce uncertainties in NN parameters. The magnitudes of propagated nearest-neighbor errors are verified from a resampling analysis of the data set, which directly assesses the error with minimum assumptions about the data.^{30,46} Note, however, that the accuracy of predictions is still limited by the approximations inherent in the nearest-neighbor model.

Sequence Dependence of Non-Watson–Crick Regions

Several motifs that involve non-Watson–Crick pairs are illustrated in Figure 1. These include hairpin loops,

internal loops, and multibranch loops or junctions. A huge number of different sizes and sequences are possible for these motifs, so once again it is necessary to develop models for predicting thermodynamics from relatively few experimental measurements. This is an active area of current research. Some recent results are discussed below.

GU Pairs. After GC and AU pairs, the next most common base pair in known RNA structures is GU. As with the Watson–Crick pairs, experiments show that a simple base-pairing model is not sufficient to explain the sequence dependence of thermodynamic stabilities for GU pairs. For example, a simple reversal of two adjacent GU pairs, from (GGA-UGUCC)₂ to (GGAGUCC)₂, changes the free energy of duplex formation from -8.4 to -6.4 kcal/mol, respectively.⁴⁷ Unlike Watson–Crick pairs, however, it appears that even a nearest-neighbor analysis does not reasonably approximate stabilities of helices when adjacent GU pairs are included in the data set.⁴⁷ In particular, the motif GGUC appears unusually stable when the nearest-neighbor model of Borer et al.²² is applied. This is surprising since nmr studies indicate that tandem GU mismatches with standard wobble hydrogen bonding⁴⁸ are all accommodated relatively easily into A-form RNA helices in GGUC, AGUC, and AUGC contexts.^{16,49} The thermodynamic results for RNA contrast with those for DNA, where the effects of internal single and double GT pairs on stability fit the nearest-neighbor model.³⁰

Hairpin Loops. Thermodynamic stabilities have been measured by optical melting for a relatively large number of RNA hairpin loops.^{50–53} Surprisingly, a rather simple equation for ΔG_{37}° is able to fit most of the results⁵¹:

$$\Delta G_{37}^\circ = \Delta G_{37}^\circ(\text{HL}) + \Sigma \Delta G_{37}^\circ(\text{NN}) \quad (7)$$

Here $\Sigma \Delta G_{37}^\circ(\text{NN})$ is the sum of the free energy increments for the nearest-neighbor Watson–Crick interactions found in the hairpin stem, as determined from duplex melting studies analyzed with the Borer et al. model,¹³ and $\Delta G_{37}^\circ(\text{HL})$ is the free energy increment for initiation of the helix upon loop formation. For the available data on hairpin loops with more than 3 nucleotides, this initiation term can be approximated in kcal/mol as⁵²

$$\Delta G_{37}^\circ(\text{HL}, n) = \Delta G_{37}^\circ(\text{iL}, n)$$

$$+ \Delta G_{37}^{\circ}(\text{MM}) \quad (8)$$

$$+ 0.6 \text{ (if loop is closed by AU or UA)}$$

where n is the number of nucleotides in the loop, $\Delta G_{37}^{\circ}(\text{iL}, n)$ is the free energy increment for initiation of a loop with n nucleotides, and $\Delta G_{37}^{\circ}(\text{MM})$ is the free energy increment for the first mismatch in the loop based on measurements of mismatches at the ends of short duplexes⁹ and is given a bonus of -0.7 kcal/mol if the first mismatch in the loop is GA or UU. Most of the time, these approximations appear able to predict the free energy of hairpin loop formation to within about 1 kcal/mol.⁵² An exception is the CUUCGG tetraloop, which appears to be stabilized by an amino to phosphate hydrogen bond within the loop.⁵⁴ It would not be surprising to find a limited number of other exceptions involving interactions of functional groups within a hairpin loop.

Internal Loops. Thermodynamic measurements have been made on a limited number of internal loops. Most of the available data is for tandem mismatches.^{35,55–57} For this motif, the thermodynamic stability clearly depends on the mismatches. For example, the free energy changes for duplex formation by (GCGGACGC)₂ and (GCGAACGC)₂ are -9.7 and -5.6 kcal/mol, respectively.^{55,56} Thus far, no simple model has proved adequate for approximating free energies of duplex formation when internal loops are formed. Nevertheless, some relatively general features have been uncovered. For both tandem mismatches^{35,55–57} and internal loops of 3 nucleotides,⁵⁸ GA and UU mismatches can be stabilizing relative to other mismatches. This enhanced stability probably arises from hydrogen bonding within these mismatches, as illustrated in Figure 3.^{55–60} Whether this hydrogen bonding potential is realized, however, depends on the context of the mismatch.³⁵ For example, ΔG_{37}° 's of duplex formation for GAGAGGAG·CUCAGCUC and GAGUUGAG·CUCUUCUC, which contain two GA and two UU mismatches, respectively, are -6.0 and -5.9 kcal/mol, respectively. The nmr spectra of these duplexes exhibit sharp resonances for the imino protons of the GA and UU mismatches, respectively, suggesting hydrogen bonding within the mismatches. In contrast, ΔG_{37}° of duplex formation for GAGAUGAG·CUCUUCUC, which contains one GA and one UU mismatch, is only -3.8 kcal/mol. The nmr spectrum has broad resonances for the UU mismatch and none for the AG mismatch, sug-

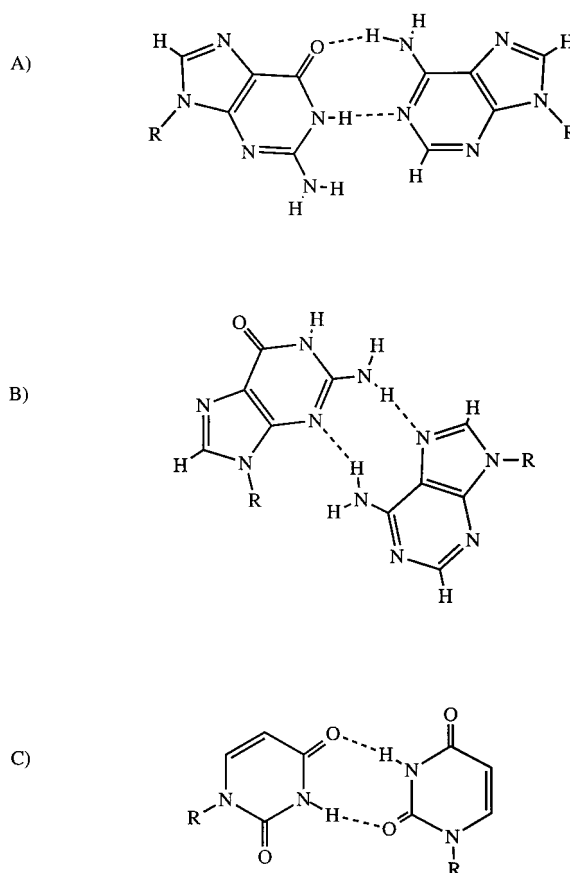


FIGURE 3 Mismatched hydrogen bonding commonly observed in G·A and U·U mismatches. Depending on the sequence context, the sheared G·A structure shown in B can involve two additional hydrogen bonds from the adenine amino to guanine 2' oxygen and from the guanine amino to the adenine nonbridging phosphate oxygen.⁵⁹

gesting weak or absent hydrogen bonding within the mismatches. Evidently, the stabilities and structures of internal loops depend on the hydrogen bonding potential of the constituent nucleotides and the steric fit of various hydrogen bonded conformations with each other and with the adjacent helices.³⁵

Gralla and Crothers⁶¹ have defined the free energy increment for an internal loop as illustrated with the following example:

$$\Delta G^{\circ}(\text{IL}) = \Delta G^{\circ}(\text{GCGGACGC}) \quad (9)$$

$$- \Delta G^{\circ}(\text{GCGCGC}) + \Delta G^{\circ}(\text{GC/CG})$$

Here $\Delta G^{\circ}(\text{GCGGACGC})$ and $\Delta G^{\circ}(\text{GCGCGC})$ are the free energy changes for duplex formation by the indicated sequences, and $\Delta G^{\circ}(\text{GC/CG})$ is the free energy increment for helix propagation by a GC/

CG nearest neighbor based on the model of Borer et al.²² as implemented by Freier et al.¹³ For this model, Xia et al.³⁵ have shown that it is possible to predict to within about 1 kcal/mol the $\Delta G_{37}^{\circ}(\text{IL})$ for tandem mismatches composed of different mismatches adjacent to GC pairs by averaging $\Delta G_{37}^{\circ}(\text{IL})$ values for symmetric tandem mismatches and adding a penalty dependent on the type and size of mismatches present.

Junctions. Essentially no thermodynamic data are available for RNA junctions. As shown in Figure 1, however, junctions are an important motif in known RNA structures. Thus free-energy parameters are required to allow predictions of secondary structure from sequence. This gap has been filled preliminarily by developing a model for junctions based on interactions known to be important for secondary structure formation in short oligonucleotides.^{9,62,63} In this model, the stability of a junction depends on the number of helices and unpaired nucleotides in the junction, stacking of unpaired nucleotides on helices, and coaxial stacking of helices. Free energy increments for stacking of nucleotides and helices are based on experimental studies of oligonucleotides.^{9,37,63} The dependence on numbers of helices and unpaired nucleotides is then estimated by optimizing predictions of known secondary structures by free energy minimization.^{62,63}

Prediction of Secondary Structure by Free Energy Minimization

In principle, the equilibrium structure of an RNA is the lowest free energy structure. In the few cases that have been tested, it appears that tertiary interactions are weaker than secondary structure interactions.^{64–67} This suggests that RNA secondary structure can be predicted by free energy minimization without considering tertiary interactions. As discussed above, however, much is unknown about the thermodynamics of RNA secondary structure. This forces the use of relatively crude models for the sequence dependence of free energy. Surprisingly, even these crude models appear sufficient to predict secondary structures that are useful for designing experiments.⁶⁸ For known secondary structures of about 500 nucleotides, the predicted lowest free energy structure is typically about 70% correct,⁶³ and recent preliminary results suggest this can be improved further.⁶⁹ It is particularly encouraging that the accuracy of prediction has steadily improved as more parameters have been measured and incorporated into folding algorithms.^{63,69,70} Recent advances

in solid phase synthesis and in developing methods for measuring duplex formation by large numbers of oligonucleotide arrays on silicon chips^{71–74} suggest that in the future it may be possible to rapidly increase the thermodynamic data base, allowing more rigorous testing of the energy minimization hypothesis. Algorithms have already been developed that allow the use of complex nonnearest-neighbor rules for stability, and these can be further modified to include tertiary interactions if necessary.^{63,69} In many cases, experimental data from chemical and enzymatic modification, site-directed mutagenesis, and sequence comparison can be combined with free energy minimization to further improve predictions.⁷⁵ These prospects and the rapid increase in available sequence information suggest that thermodynamics coupled with other approaches will facilitate the discovery of many structure–function relationships in the future.

We thank Hatim Allawi and Steve Testa for stimulating conversations and for preparing figures. This work supported by National Institutes of Health Grant GM22939 to DHT.

REFERENCES

1. Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. (1982) *Cell* **31**, 147–157.
2. Guerrier-Takada, C. & Altman, S. (1984) *Science* **223**, 285–286.
3. Eckstein, F. & Lilley, D. M. J. (1996) *Catalytic RNA*, Springer, Berlin.
4. Baserga, R. & Denhardt, D. T. (1992) *Antisense Strategies, Vol. 660, Annals of the New York Academy of Science*, New York.
5. Erickson, R. P. & Izant, J. G. (1992) *Gene Regulation: Biology of Antisense RNA and DNA*, Raven Press, New York.
6. Hertel, K. J., Herschlag, D. & Uhlenbeck, O. C. (1994) *Biochemistry* **33**, 3374–3385.
7. Williams, N. (1997) *Science* **275**, 301–302.
8. Uhlenbeck, O. C., Pardi, A. & Feigon, J. (1997) *Cell* **90**, 833–840.
9. Serra, M. J. & Turner, D. H. (1995) *Methods Enzymol.* **259**, 242–261.
10. Petersheim, M. & Turner, D. H. (1983) *Biochemistry* **22**, 256–263.
11. Puglisi, J. & Tinoco, I., Jr. (1989) *Methods Enzymol.* **180**, 304–325.
12. Breslauer, K., Freire, E. & Straume, M. (1992) *Methods Enzymol.* **211**, 533–567.
13. 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.
14. SantaLucia, J., Jr., Allawi, H. & Seneviratne, P. A. (1996) *Biochemistry* **35**, 3555–3562.
 15. Sugimoto, N., Nakano, S., Yoneyama, M. & Honda, K. (1996) *Nucleic Acids Res.* **24**, 4501–4505.
 16. McDowell, J. A. & Turner, D. H. (1996) *Biochemistry* **35**, 14077–14089.
 17. SantaLucia, J., Jr., Kierzek, R. & Turner, D. (1991) *J. Am. Chem. Soc.* **113**, 4313–4322.
 18. Freier, S. M., Petersheim, M., Hickey, D. R. & Turner, D. H. (1984) *J. Biomol. Struct. Dynam.* **1**, 1229–1242.
 19. Liu, Y. & Sturtevant, J. M. (1997) *Biophys. Chem.* **64**, 121–126.
 20. Wartell, R. M. & Benight, A. S. (1985) *Phys. Rep.* **126**, 67–107.
 21. Schmitz, M. & Steger, G. (1992) *Comput. Appl. Biosci.* **8**, 389–399.
 22. Borer, P. N., Dengler, B., Tinoco, I., Jr. & Uhlenbeck, O. C. (1974) *J. Mol. Biol.* **86**, 843–853.
 23. Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 3746–3750.
 24. Rentzeperis, D., Ho, J. & Marky, L. A. (1993) *Biochemistry* **32**, 2564–2572.
 25. Chaires, J. B. (1997) *Biophys. Chem.* **64**, 15–23.
 26. Jaeger, J. A., SantaLucia, J., Jr. & Tinoco, I., Jr. (1993) *Ann. Rev. Biochem.* **62**, 255–287.
 27. Marky, L. A. & Breslauer, K. J. (1987) *Biopolymers* **26**, 1601–1620.
 28. SantaLucia, J., Jr., Kierzek, R. & Turner, D. H. (1990) *Biochemistry* **29**, 8813–8819.
 29. Albergo, D., Marky, L., Breslauer, K. & Turner, D. (1981) *Biochemistry* **20**, 1409–1413.
 30. Allawi, H. T. & SantaLucia, J., Jr. (1997) *Biochemistry* **36**, 10581–10594.
 31. Longfellow, C. E., Kierzek, R. & Turner, D. H. (1990) *Biochemistry* **29**, 278–285.
 32. Plum, G. E., Grollman, A. P., Johnson, F. & Breslauer, K. J. (1995) *Biochemistry* **34**, 16148–16160.
 33. Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York.
 34. Meyer, S. L. (1975) *Data Analysis for Scientists and Engineers*, Wiley, New York.
 35. Xia, T., McDowell, J. A. & Turner, D. H. (1997) *Biochemistry* **36**, 12486–12497.
 36. Snedecor, G. W. & Cochran, W. G. (1971) *Statistical Methods*, The Iowa State University Press, Ames, IA.
 37. Turner, D. H., Sugimoto, N. & Freier, S. M. (1988) *Ann. Rev. Biophys. Biophys. Chem.* **17**, 167–192.
 38. Freier, S. M., Sinclair, A., Neilson, T. & Turner, D. H. (1985) *J. Mol. Biol.* **185**, 645–647.
 39. Goldstein, R. F. & Benight, A. S. (1992) *Biopolymers* **32**, 1679–1693.
 40. Gray, D. M. (1997) *Biopolymers* **42**, 783–793.
 41. Kierzek, R., Caruthers, M. H., Longfellow, C. E., Swinton, D., Turner, D. H. & Freier, S. M. (1986) *Biochemistry* **25**, 7840–7846.
 42. Doktycz, M. J., Morris, M. D., Dormady, S. J., Beattie, K. L. & Jacobson, K. B. (1995) *J. Biol. Chem.* **270**, 8439–8445.
 43. Gray, D. M. & Tinoco, I., Jr. (1970) *Biopolymers* **9**, 223–244.
 44. Vologodskii, A. V., Amirikyan, B. R., Lyubchenko, Y. L. & Frank-Kamenetskii, M. D. (1984) *J. Biomol. Struct. Dynam.* **2**, 131–148.
 45. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1989) *Numerical Recipes*, Cambridge University Press, New York.
 46. Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London.
 47. He, L., Kierzek, R., SantaLucia, J., Jr., Walter, A. E. & Turner, D. H. (1991) *Biochemistry* **30**, 11124–11132.
 48. Crick, F. H. C. (1966) *J. Mol. Biol.* **19**, 548–555.
 49. McDowell, J. A., He, L., Chen, X. & Turner, D. H. (1997) *Biochemistry* **36**, 8030–8038.
 50. Serra, M. J., Lyttle, M. H., Axenson, T. J., Schadt, C. A. & Turner, D. H. (1993) *Nucleic Acids Res.* **21**, 3845–3849.
 51. Serra, M. J., Axenson, T. J. & Turner, D. H. (1994) *Biochemistry* **33**, 14289–14296.
 52. Serra, M. J., Barnes, T. W., Betschart, K., Gutierrez, M. J., Sprouse, K. J., Riley, C. K., Stewart, L. & Temel, R. E. (1997) *Biochemistry* **36**, 4844–4851.
 53. Antao, V. P. & Tinoco, I., Jr. (1992) *Nucleic Acids Res.* **20**, 819–824.
 54. Varani, G., Cheong, C. & Tinoco, I., Jr. (1991) *Biochemistry* **30**, 3280–3289.
 55. Walter, A. E., Wu, M. & Turner, D. H. (1994) *Biochemistry* **33**, 11349–11354.
 56. Wu, M., McDowell, J. A. & Turner, D. H. (1995) *Biochemistry* **34**, 3204–3211.
 57. SantaLucia, J., Jr., Kierzek, R. & Turner, D. H. (1991) *Biochemistry* **30**, 8242–8251.
 58. Schroeder, S., Kim, J. & Turner, D. H. (1996) *Biochemistry* **35**, 16105–16109.
 59. SantaLucia, J. J. & Turner, D. H. (1993) *Biochemistry* **32**, 12612–12613.
 60. Wu, M. & Turner, D. H. (1996) *Biochemistry* **35**, 9677–9689.
 61. Gralla, J. & Crothers, D. M. (1973) *J. Mol. Biol.* **78**, 301–319.
 62. Jaeger, J. A., Turner, D. H. & Zuker, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7706–7710.
 63. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Muller, P., Mathews, D. H. & Zuker, M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 9218–9222.
 64. Crothers, D. M., Cole, P. E., Hilbers, C. W. & Schulman, R. G. (1974) *J. Mol. Biol.* **87**, 63–88.
 65. Hilbers, C. W., Robillard, G. T., Shulman, R. G., Blake, R. D., Webb, P. K., Fresco, R. & Riesner, D. (1976) *Biochemistry* **15**, 1874–1882.

66. Banerjee, A. R., Jaeger, J. A. & Turner, D. H. (1993) *Biochemistry* **32**, 153–163.
67. Jaeger, L., Westhof, E. & Michel, F. (1993) *J. Mol. Biol.* **234**, 331–346.
68. Jaeger, J. A., Turner, D. H. & Zuker, M. (1990) *Methods Enzymol.* **183**, 281–306.
69. Mathews, D. H., Andre, T. C., Kim, J., Turner, D. H., & Zuker, M. (1997) in *Molecular Modeling of Nucleic Acids*, Leontis, N. B. & SantaLucia, J., Jr., Eds., American Chemical Society, New York, pp. 246–257.
70. Turner, D. H., Sugimoto, N., Jaeger, J. A., Longfellow, C. E., Freier, S. M. & Kierzek, R. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 123–133.
71. Fodor, S. P. A., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P. & Adams, C. L. (1993) *Nature* **364**, 555–556.
72. Stimpson, D. I., Hoiyer, J. V., Hsieh, W., Jou, C., Gordon, J., Theriault, T., Gamble, R. & Baldeschieler, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 6379–6383.
73. O'Donnell-Maloney, M. J., Smith, C. L. & Cantor, C. R. (1996) *Trends Biotechnol.* **14**, 401–407.
74. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7912–7915.
75. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. (1997) *RNA* **3**, 1–16.
76. Testa, S. M., Haidaris, C. G., Gigliotti, F. & Turner, D. H. (1997) *Biochemistry* **36**, 15303–15314.