*4912–4924* *Nucleic Acids Research, 2006, Vol. 34, No. 17* *Published online 18 September 2006*
*doi:10.1093/nar/gkl472*

# A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation

**Zhi John Lu, Douglas H. Turner[1] and David H. Mathews***

Department of Biochemistry & Biophysics and Center for Pediatric Biomedical Research, University of Rochester Medical Center, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA and [1]Department of Chemistry, University of Rochester, Box 0216, Rochester, NY 14627, USA

Received March 27, 2006; Revised and Accepted June 20, 2006

## ABSTRACT

**A complete set of nearest neighbor parameters to predict the enthalpy change of RNA secondary structure formation was derived. These parameters can be used with available free energy nearest neighbor parameters to extend the secondary structure prediction of RNA sequences to temperatures other than 37°C. The parameters were tested by predicting the secondary structures of sequences with known secondary structure that are from organisms with known optimal growth temperatures. Compared with the previous set of enthalpy nearest neighbor parameters, the sensitivity of base pair prediction improved from 65.2 to 68.9% at optimal growth temperatures ranging from 10 to 60°C. Base pair probabilities were predicted with a partition function and the positive predictive value of structure prediction is 90.4% when considering the base pairs in the lowest free energy structure with pairing probability of 0.99 or above. Moreover, a strong correlation is found between the predicted melting temperatures of RNA sequences and the optimal growth temperatures of the host organism. This indicates that organisms that live at higher temperatures have evolved RNA sequences with higher melting temperatures.**

## INTRODUCTION

RNA is more than a simple single-stranded sequence carrying genetic information as in the Central Dogma of Biology. For example, it can form tertiary structures that, such as proteins, can be catalytic. Natural and engineered RNA molecules are widely used as functional tools in enzymatic catalysis and genetic control (1–5). One current problem is how to predict the structures of functional RNA sequences.

Secondary structure, the sum of canonical base pairs, is stronger (6–9) and forms faster (10) than tertiary structure. Therefore, secondary structure can largely be determined without knowledge of tertiary structure. Comparative sequence analysis is a standard technique for determining the secondary structure of homologous RNA sequences (11–13). When only a few or even a single sequence is available, the secondary structure at 37°C can be predicted by free energy minimization algorithms (14–17) using a set of empirical free energy parameters, determined from optical melting experiments (17–21). Each parameter only depends on the sequence identity of nucleotides in the motif and in adjacent base pairs and the total free energy is the sum of nearest neighbor terms. The average sensitivity (the percentage of known base pairs that are correctly predicted) of free energy minimization prediction has been benchmarked as high as 72.8 ± 9.4% for a diverse database of sequences having fewer than 800 nt (17). Furthermore, experimentally determined constraints can improve this accuracy of prediction up to 84% (17,18) for sequences with <6% pseudoknotted (non-nested) base pairs (17). Partition function prediction of base pair probabilities can be used to identify base pairs in the predicted lowest free energy structure that are much more likely than average to be in the known secondary structure (22,23). For example, 91.0% of base pairs in the lowest free energy structure with pairing probability of 0.99 or higher are contained in the known structure, on average (22). The high accuracy of thermodynamic structure prediction (17) demonstrates that many RNA secondary structures can be determined from sequences, without knowledge of any tertiary contacts or protein interactions.

The current set of free energy nearest neighbor parameters for predicting the free energy of RNA secondary structure, however, is limited to application at 37°C. Many organisms, thermophiles and psychrophiles, live at temperatures far from 37°C and many experiments are conducted at other temperatures. The prediction of secondary structure of RNA at arbitrary temperature would expand our knowledge of structure and evolution in the RNA world. Moreover, it would

*To whom correspondence should be addressed. Tel: 1 585 275 1734; Fax: 1 585 275 6007; Email: david_mathews@urmc.rochester.edu

facilitate studying and designing functional RNA molecules at temperatures other than 37°C. The enthalpy nearest neighbor parameters can be used in conjunction with available free energy nearest neighbor parameters for 37°C to determine free energy nearest neighbors at other temperatures. But the most recent enthalpy parameters were derived in 1995 using a simple model (24). At that time, no themes had emerged for the sequence-dependent stability of internal loops. Subsequently, the nearest neighbor model for free energy change at 37°C was significantly improved (17) using experimental results. Therefore, we applied the principles of the current free energy nearest neighbor model (17,18) to determine a complete set of enthalpy nearest neighbor parameters using the available optical melting data.

## MATERIALS AND METHODS

### Database of experiments

The database of experimental data for derivation of enthalpy parameters is included in Supplementary Data. It includes 130 hairpin loops (25–31), 37 bulge loops (32,33), 337 internal loops (17,18,34–49) (99 of which are 2 × 2 internal loops), 74 multibranch loops (50,51) and 43 coaxial stacking models (52–55).

### Derivation and refinement of enthalpy parameters

*Canonical base pairs*. The enthalpies of Watson–Crick and GU base pairs were derived by Xia *et al*. (21) and Mathews *et al*. (18), respectively.

*Dangling ends and terminal mismatches*. Dangling ends are unpaired nucleotides adjacent to canonical pairs and their enthalpy parameters were compiled previously (24). Dangling ends on terminal GU pairs are treated similar to dangling ends on terminal AU pairs. Terminal mismatches are non-canonical pairs at the end of helixes. The enthalpy parameters of terminal mismatches are taken from another compilation (20), with the exception of mismatches on terminal GU pairs, which were measured recently (30).

If a terminal mismatch has the potential to pair canonically, the values of A–C and C–A mismatches are used for the purine–pyrimidine mismatch and pyrimidine–purine mismatches, respectively. This is important for partition function calculations, where all possible secondary structures are considered.

*Hairpin loops*. The experimental enthalpies of hairpin loop formation are calculated from published experimental data (25–31) with the following equation:

$$\Delta H^{\circ}_{\mathrm{loop}} = \Delta H^{\circ}_{\mathrm{stem-loop}} - \Delta H^{\circ}_{\mathrm{stem}},$$

where $\Delta H^{\circ}_{\mathrm{stem-loop}}$ is the experimental value for unfolding the hairpin loop with stem, $\Delta H^{\circ}_{\mathrm{stem}}$ is calculated by the INN-HB parameters (18,21), without an intermolecular initiation term.

The hairpin loop enthalpy parameters are estimated by linear regression using the same model as free energy nearest neighbor parameters (17), except that the GG first mismatch bonus observed for free energy does not apply for enthalpy because the bonus was not statistically significant for enthalpy. The GG stability bonus is therefore entropic in nature, consistent with the observation that GG mismatches are dynamic (56), i.e. they sample more than one single microstate on short timescales.

The enthalpies of hairpin loops are estimated by the following equation:

$$\begin{aligned}\Delta H^{\circ}_{\mathrm{loop}}(n>3) = {} & \Delta H^{\circ}_{\mathrm{initiation}}(n) + \Delta H^{\circ}(\text{first mismatch stacking}) \\ & + \Delta H^{\circ}_{\mathrm{bonus}}(\text{UU or GA first mismatch but not AG}) \\ & + \Delta H^{\circ}_{\mathrm{bonus}}(\text{special G-U closure}) \\ & + \Delta H^{\circ}_{\mathrm{penalty}}(\text{oligo-C loops}),\end{aligned}$$

where $n$ is the number of unpaired nucleotides in the loop. Hairpins with fewer than 3 unpaired nucleotides are not allowed by the model. When $n = 3$, only the initiation term is considered without any bonus and penalty terms, except a penalty for hairpin loops with three Cs. When $n > 3$, the special GU closure bonus applies to GU closed hairpins in which a 5′ closing G is preceded by two G residues; and $\Delta H^{\circ}_{\mathrm{bonus}}$ (UU or GA first mismatch but not AG) is applied to loops with first mismatches of UU or GA (G on the 5′ side and A on 3′ side of loop). The oligo-C penalty applies only to loops composed of all C residues and, if $n > 3$, is calculated with $\Delta H^{\circ}_{\mathrm{penalty}}$ (oligo-C loops, $n > 3$) = A$n$ + B. For hairpin loops composed entirely of 3 C residues, the $\Delta H^{\circ}_{\mathrm{penalty}}$ (oligo-C loops, $n = 3$) is applied.

The enthalpy parameters are listed in Table 1 and the database of measured loop enthalpies is available as Supplementary Data. In the absence of data, for hairpin loops longer than 9 nt, the initiation enthalpy is approximated with the initiation term for a hairpin of 9 nt. This assumes that additional instability of hairpin loops as the loop lengthens derives from the entropy (57).

The measured free energies at 37°C of some special hairpin loops of 3, 4 or 6 unpaired nucleotides (30,31,34–36) are either more or less stable by 0.9 kcal/mol than the model predicts. The enthalpies for each of these sequences are listed in a separate lookup table (Table 2), to be consistent with the free energy parameters.

**Table 1.** Hairpin loop enthalpy parameters[a]

| Parameter | Condition | $\Delta H^{\circ}$ (kcal/mol) | SE (kcal/mol) |
|---|---|---|---|
| $\Delta H^{\circ}_{\mathrm{initiation}}(n)$ | $n = 3$ | 1.3 | 1.79 |
| | 4 | 4.8 | 1.31 |
| | 5 | 3.6 | 1.61 |
| | 6 | −2.9 | 1.01 |
| | 7 | 1.3 | 1.73 |
| | 8 | −2.9 | 1.72 |
| | 9 | 5.0 | 2.16 |
| | >9 | 5.0 | — |
| $\Delta H^{\circ}_{\mathrm{bonus}}$ | UU or GA first mismatch but not AG | −5.8 | 1.27 |
| $\Delta H^{\circ}_{\mathrm{bonus}}$ | Special GU closure | −14.8 | 2.35 |
| $\Delta H^{\circ}_{\mathrm{penalty(oligo\text{-}C\,loops)}}$ | $n = 3$ | 18.6 | 5.66 |
| $\Delta H^{\circ}_{\mathrm{penalty(oligo\text{-}C\,loops)}}$ | A | 3.4 | 1.48 |
| $\Delta H^{\circ}_{\mathrm{penalty(oligo\text{-}C\,loops)}}$ | B | 7.6 | 9.57 |

[a]Hairpin loops of <3 nt are prohibited. $\Delta H^{\circ}$(first mismatch stacking) and terminal mismatch bonuses apply only to hairpin loops with >3 unpaired nucleotides.

**Table 2.** Lookup table for unstable triloops and stable tetraloops and hexaloops

| Hairpin[a] | Ref(s)[b] | $\Delta H^{\circ}_{loop}$ (kcal/mol) |
|---|---|---|
| CaacG | A | 23.7 |
| GuuaC | A | 10.8 |
| CaacgG | B | 6.9 |
| CcaagG | B | −10.3 |
| CcacgG | B | −3.3 |
| CccagG | B | −8.9 |
| CcgagG | B | −6.6 |
| CcgcgG | B | −7.5 |
| CcuagG | B | −3.5 |
| CcucgG | B | −13.9 |
| CuaagG | B | −7.6 |
| CuacgG | C, D | −10.7 |
| CucagG | B | −6.6 |
| CuccgG | C | −12.9 |
| CugcgG | B | −10.7 |
| CuuagG | B | −6.2 |
| CuucgG | C, D | −15.3 |
| CuuugG | D | −6.8 |
| AcaguacU | E | −16.8 |
| AcagugcU | E,C | −12.8 |
| AcagugaU | C | −11.4 |
| AcaguguU | E | −15.4 |

Closing pairs are included and unpaired nucleotides are shown in lower case.
[a]For extra stable hairpins measured in 0.1 M $Na^+$ (A, B), placement was determined by assuming that the relative enthalpy of loops remains constant between 0.1 and 1 M $Na^+$(30,34).
[b]A, Ref. (34); B, Ref. (35); C, Ref. (30); D, Ref. (31); E, Ref. (36).

**Table 3.** Bulge loop initiation enthalpy parameters[a]

| Bulge length | $\Delta H^{o}_{initiation}$(kcal/mol) | SE (kcal/mol) |
|---|---|---|
| 1 | 10.6 | 1.2 |
| 2 | 7.1 | 4.3 |
| 3 | 7.1 | 11.7 |
| $n \geqslant 4$ | (7.1) | — |

[a]Note that the nearest neighbor parameter for stacking of adjacent base pairs is added for bulges with 1 nt. For bulges with >1 nt, calculation of the stabilities of adjacent helices includes the terminal AU penalty terms for AU or GU pairs adjacent to the bulge.

*Bulge loops*. RNA secondary structure is destabilized by bulge loops, which are an interruption of helical structure in one strand only (32,37,38). The initiation terms, $\Delta H^{\circ}_{bulge\ initiation}(n)$ for bulge loops of 1–3 nt, are listed in Table 3. They are the average values of experimental data (32,33), calculated using the following equation:

$$\Delta H^{o}_{bulge\ initiation} = \Delta H^{o}(\text{duplex with bulge}) - \Delta H^{o}(\text{duplex without bulge}) + \Delta H^{o}_{bp\ stack}(n > 1),$$

where the enthalpy of the duplex without bulge is the experimental value of the sequence of the duplex without the bulge or as calculated with INN-HB parameters (21) if the experimental values were not available. $\Delta H^{\circ}_{bp\ stack}$ is the stacking enthalpy of the base pairs in the duplex without the bulge that flank the bulge loop in the duplex with the bulge. Because the difference of initiation enthalpies between 2 and 3 nt bulges is almost zero, it is assumed that the increasing instability for longer bulges ($n \geqslant 4$) comes from the entropy of the loop closure (39,57). Thus, the initiation enthalpy for bulges longer than 3 nt is approximated as the 3 nt bulge enthalpy.

Assuming that helical stacking is continuous between the adjacent helices for single bulges, but is interrupted by longer bulges (39,40), the enthalpies of bulge loops are calculated with the following equation:

$$\Delta H^{o}_{bulge}(n) = \Delta H^{o}_{bulge\ initiation}(n) + \Delta H^{o}_{bp\ stack}(\text{only applied to 1 nt loops}).$$

The calculation of enthalpies for the adjacent helices would include the terminal AU/GU penalty (21) for AU/GU pairs adjacent to the bulge loops that are longer than 1 nt. $\Delta H^{\circ}_{bp\ stack}$ is the canonical helix stacking enthalpy applied for the two closing base pairs as though the helix was not interrupted by the bulge loop.

*Internal loops*. Internal loop enthalpies were calculated from experimental data (17,18,34–49) using the following equation:

$$\Delta H^{o}_{internal\ loop} = \Delta H^{o}(\text{entire sequence with internal loop}) - \Delta H^{o}(\text{reference sequence without internal loop}) + \Delta H^{o}_{bp\ stack}.$$

The range of measured enthalpies differs for internal loops of different size and symmetry; therefore, different enthalpy models are used to predict different loop types. The models are similar to those used to model free energies (17).

### 1 × 1 Internal loops (single mismatches)

For single non-canonical pairs (1 × 1 internal loops), the loop enthalpies are approximated by the following equation:

$$\Delta H^{o}_{loop}(1\times1) = \Delta H^{o}_{loop\ initiation}(n=2) + \Delta H^{o}_{AU/GU}(\text{per AU or GU closure}) + \Delta H^{o}_{GG}(1\times)1 + \Delta H^{o}_{5'RU/3'YU}(1\times1),$$

where $\Delta H^{\circ}_{loop\ initiation}(n=2)$ is the enthalpy of initiation for a single non-canonical pair; $\Delta H^{\circ}_{AU/GU}$ is the penalty for each AU or GU closing base pair; $\Delta H^{\circ}_{GG}(1\times1)$ is a bonus for a GG pair in a 1 × 1 loop; and $\Delta H^{o}_{5'RU/3'YU}(1\times1)$ is a bonus for a 5′RU/3′YU stack in a 1 × 1 loop, where R is a purine and Y is a pyrimidine.

### 2 × 2 Internal loops (tandem mismatches)

The 2 × 2 internal loops, also called tandem mismatches, interrupt helical RNA with two opposing unpaired nucleotides on each strand. Many of the sequence-symmetric 2 × 2 loops have been studied experimentally (17,18, 34–49) and their enthalpies are assembled in a 'periodic table' (Table 4). Symmetric sequences that have not been measured are approximated by averaging the most adjacent columns that have been measured. For asymmetric

**Table 4.** The periodic table of tandem mismatch (2 × 2 internal loop) enthalpy[a]

| Closing BP | Mismatch<br>GA<br>AG | AG<br>GA | UU<br>UU | GG<br>GG | CA<br>AC | CU<br>UC | UC<br>CU | CC<br>CC | AC<br>CA | AA<br>AA |
|---|---|---|---|---|---|---|---|---|---|---|
| GC | −31.0[c]<br>−15.9[c]<br>−28.4[c]<br>**−25.1** | −15.6[c] | −14.4[b] | −22.8 | −10.3[b] | −29.4[b]<br>−12.4[b]<br>**−20.9** | (−14.7) | (−14.7) | −8.6[b] | −1.3[b] |
| CG | −8.9[d]<br>−16.5[g]<br>**−14.6** | −12.7[d] | −17.5[d] | −22.8[d] | −10.8[d] | −0.6[d] | −2.8[d] | −1.8[d] | −1.7[d] | −4.2[d]<br>−5.0[e]<br>**−4.6** |
| UA | −13.4[c] | −19.4 | −6.7[b] | 2.7 | 9.1[b] | 3.3[b] | 9.5[b] | (12.1) | (12.1) | 14.7[b] |
| AU | −17.9[c]<br>−11.0[c]<br>**−14.4** | −10.8 | −12.2[b] | −1.0 | 7.2[b] | (7.4) | (7.4) | (7.4) | 7.5[b] | 13.4[b] |
| UG | −15.3[c] | −18.7[f] | (−8.5) | (−8.5) | (−8.5) | (−8.5) | (−8.5) | (−8.5) | (−8.5) | 1.7[b] |
| GU | −19.9[c] | −16.1[f] | (−19.6) | (−19.6) | (−19.6) | (−19.6) | (−19.6) | (−19.6) | (−19.6) | −23.2[f] |

[a]Boldface numbers are averages of multiple measurements on the same internal loops and numbers in parentheses are predicted by average of the nearest numbers to the left and right. The enthalpies of reference helices were taken from Ref. (21).
The enthalpies (kcal/mol) are drawn from b, Ref. (76); c, Ref. (52); d, Ref. (55,77); e, Ref. (78); f, (79); g, Ref. (80).

**Table 5.** Approximations for internal loop enthalpy parameters at 37°C (in kcal/mol)

| Length (nt) | 2 | 3 | 4 | 5 | 6 | >6[a] | |
|---|---|---|---|---|---|---|---|
| $\Delta H^{o}_{initiation}$ | −10.5 ± 1.4 | 0.3 ± 1.2 | −7.2 ± 1.1 | −6.8 ± 1.8 | −1.3 ± 1.2 | (−1.3) | |
| $\Delta H^{o}_{AU/GU}$ | | | | 5.0 ± 0.7 | | | |
| $\Delta H^{o}_{asym}$ | | | | 3.2 ± 0.7 | | | |
| Type of loop (first pair): | 5′RA<br>3′YG | 5′YA<br>3′RG | 5′RG<br>3′YA | 5′YG<br>3′RA | G<br>G | U<br>U | 5′RU<br>3′YU |
| 1 × 1 | NA[b] | NA | NA | NA | −7.9 ± 3.7 | NA | −3.4±1.7 |
| 1 × 2 | 0 | −5.8 ± 1.5 | −5.8 ± 1.5 | −5.8 ± 1.5 | −5.8 ± 1.5 | −10.1 ± 1.7 | NA |
| 1 × (*n* − 1), *n* > 3 | 0 | 0 | 0 | 0 | 0 | 0 | NA |
| 2 × 3 | 0 | −5.7 ± 3.8 | −10.9 ± 2.7 | −8.6 ± 1.9 | −9.0 ± 4.6 | −6.4 ± 2.5 | NA |
| Others (except 2 × 2) | −3.4 ± 1.3 | −3.4 ± 1.3 | −7.6 ± 1.0 | −7.6 ± 1.0 | 2.8 ± 2.4 | −5.8 ± 1.1 | NA |

The parameters were obtained from a set of linear regressions of experimental data for 1 × 1, 1 × 2, 1 × 3, 2 × 2, 2 × 3 and 3 × 3 loops. $\Delta H^{o}_{initiation}(n = 2,4,5,6)$, $\Delta H^{o}_{AU/GU}$, $\Delta H^{o}_{asym}$, $\Delta H^{o}_{first\ noncanonical\ pairs}$ (others: AG), $\Delta H^{o}_{first\ noncanonical\ pairs}$ (others: GA), $\Delta H^{o}_{first\ noncanonical\ pairs}$ (others:GG), $\Delta H^{o}_{first\ noncanonical\ pairs}$ (others: UU), $\Delta H^{o}_{first\ noncanonical\ pairs}$ (1 × 1: GG), $\Delta H^{o}_{first\ noncanonical\ pairs}$ (1 × 1: UU) are determined with linear regression of all the loops excluding the 2 × 2 and 1 × 2 loops. $\Delta H^{o}_{first\ noncanonical\ pairs}$ (2 × 3: GG) was specified and separated in the regression to make the standard errors of $\Delta H^{o}_{first\ noncanonical\ pairs}$ (others: GG) smaller. Some parameters of 1 × 2 and 2 × 3 were specified by refitting of 1 × 2 and 2 × 3 loops respectively, supposing that $\Delta H^{o}_{first\ noncanonical\ pairs}$ (5′RA/3′YG) was zero.
[a]When the internal loop is large ($n > 6$) the increase of free energy is assumed to be derived from entropy (57), so the initiation term, $\Delta H^{o}_{initiation}(n > 6)$is the same as $\Delta H^{o}_{initiation}$ (6).
[b]NA, not applicable to that type of loop.

2 × 2 loops, the enthalpies are approximated using the following equation:

$$\Delta H^{o}_{loop}(2\times 2)(5'\mathrm{PXYS}/3'\mathrm{QWZT}) = [\Delta H^{o}_{37}(5'\mathrm{PXWQ}/3'\mathrm{QWXP}) + \Delta H^{o}_{37}(5'\mathrm{TZYS}/3'\mathrm{SYZT})]/2 + \Delta H^{o}_{GG} + \Delta_{p},$$

where $\Delta H^{o}_{GG}$ (12.5 ± 2.7 kcal/mol) is applied to loops with a GG pair adjacent to an AA or any non-canonical pair with a pyrimidine and $\Delta_p$ (2.4 ± 3.1 kcal/mol) is applied to loops with an AG or GA pair adjacent to a UC, CU or CC pair or with a UU pair adjacent to an AA pair.

### Other internal loops

The enthalpies of other internal loops are approximated using the following equation:

$$\Delta H^{o}_{loop}(n) = \Delta H^{o}_{loop\ initiation}(n) + \Delta H_{AU/GU} + |n_1 - n_2|\,\Delta H^{o}_{asym} + \Delta H^{o}_{first\ non\text{-}canonical\ pairs} \times[\text{except for } 1\times(n-1) \text{ for } n>3],$$

where $\Delta H^{\circ}_{loop}(n)$ is the enthalpy of initiation for a loop of $n$ nucleotides; $\Delta H^{\circ}_{asym}$ is a penalty for loops with unequal numbers of nucleotides on each side, with $n_1$ and $n_2$ the number of nucleotides on each side; $\Delta H^{\circ}_{first\ non\text{-}canonical\ pairs}$ is applied for each sequence-specific first mismatch (Table 5), but it is not applied to loops of the form 1 × (*n* − 1) with $n > 3$ ($n$ is the total number of unpaired bases). Special first mismatch bonuses were determined for 2 × 3 and 1 × 2 internal loops with separate linear regressions.

Moreover, the free energy parameters (Table 6) were updated for internal loops based on recent experimental measurements. The free energy parameters were obtained using the method of Mathews *et al*. (17). The recent data include the 3 × 3 loops from Chen *et al*. (41), but excluding the 3 × 3 loops with a middle GA pair. The middle GA pair is shown to enhance stability and this extra stability cannot be predicted by the nearest neighbor parameter set used in this work (41).

*Coaxial stacking*. Coaxial stacking, which is a favorable interaction of two helices stacked end to end, occurs in

**Table 6.** Updated internal loop free energy parameters at 37°C (in kcal/mol)

| Length (nt) | 2 | 3 | 4 | 5 | 6 | n>6 | |
|---|---|---|---|---|---|---|---|
| $\Delta G^{o}_{37\text{initiation}}$ | 0.5 ± 0.1 | 1.7 ± 0.1 | 1.1 ± 0.1 | 2.0 ± 0.1 | 2.0 ± 0.1 | 2.0 + 1.08 ln($n$/6) | |
| $\Delta G^{o}_{37\text{AU/GU}}$ | | | | 0.7 ± 0.1 | | | |
| $\Delta G^{o}_{37\text{asym}}$ | | | | 0.6 ± 0.1 | | | |
| Type of loop (first pair): | 5′RA | 5′YA | 5′RG | 5′YG | G | U | 5′RU |
| | 3′YG | 3′RG | 3′YA | 3′RA | G | U | 3′YU |
| 1 × 1 | NA[a] | NA | NA | NA | −2.6 ± 0.3 | NA | −0.4 ± 0.1 |
| 1 × 2 | 0 | −1.2 ± 0.2 | −1.2 ± 0.2 | −1.2 ± 0.2 | −1.2 ± 0.2 | −0.8 ± 0.2 | NA |
| 1 × ($n$ − 1), $n$ > 3 | 0 | 0 | 0 | 0 | 0 | 0 | NA |
| 2 × 3 | 0 | −0.5 ± 0.2 | −1.2 ± 0.1 | −1.1 ± 0.1 | −0.7 ± 0.2 | −0.4 ± 0.1 | NA |
| Others (except 2 × 2) | −0.8 ± 0.1 | −0.8 ± 0.1 | −1.0 ± 0.1 | −1.0 ± 0.1 | −1.0 ± 0.2 | −0.6 ± 0.1 | NA |

The parameters were obtained from a set of linear regression of the same experimental data as Mathews *et al*. (17), except for some updated data of 3 × 3 loops from Chen *et al*. (41).
[a]NA, not applicable to that type of loop.

multibranch loops and exterior loops. Stability increments for coaxial stacking were measured with a structure composed of a short oligonucleotide bound to a single-stranded end of a stem–loop structure, creating a helical interface (52–55). The enthalpy of coaxial stacking is quantified as follows:

$$\Delta H^{o}_{\text{coaxial}} = \Delta H^{o}(\text{duplex in context of stem-loops structure}) - \Delta H^{o}(\text{duplex without stem-loop structure, predicted}) + \Delta H^{o}(\text{correction}),$$

where $\Delta H^{\circ}$(correction) is the enthalpy for displacing a 3′ dangling end on the stem–loop structure if one is present.

When the helixes have no intervening mismatches, the enthalpy bonus is approximated by the nearest neighbor parameter (21) of a base pair in a helix. The excess enthalpy above the helical stacking nearest neighbor from Xia *et al*. (21), $\Delta H^{\circ}_{\text{coaxial}} - \Delta H^{\circ}_{\text{NN}}$, for each measured interface was calculated. With flush interfaces, i.e. with no intervening mismatch, and no strand extensions beyond the interface, the average excess enthalpy is −1.53 ± 1.45 kcal/mol. For interfaces followed by strand extensions, the excess enthalpy is 1.82 ± 1.13 kcal/mol. As the excess enthalpy changes are not statistically significant, coaxial stacking of helices with no intervening nucleotides is modeled with the enthalpy parameter in a helix.

With one intervening nucleotide from each strand, two helices can stack with an intervening mismatch between them. There are two stack increments: one is the mismatch stack at the end of one helix with continuous backbone, which is equal to the mismatch stacking parameter on a helix, and the other is the mismatch stack with discontinuous backbone, which is modeled as sequence independent. The average enthalpy of sequence independent stacks is −8.46 ± 2.75 kcal/mol. In addition to this, an enthalpy bonus of −0.4 or −0.2 kcal/mol are applied to intervening mismatches composed of nucleotides that could form a Watson–Crick or a GU base pair, respectively. These bonuses are identical to free energy increments that are used and are empirically found to improve structure prediction accuracy.

*Multibranch loops*. The parameters are determined by linear regression of experimental data for three- and four-way multibranch loops (50,51). In a nearest neighbor model, the bimolecular enthalpy ( $\Delta H^{\circ}_{\text{bimol}}$) for the formation of the duplex with a multibranch loop is given by the following equation:

$$\Delta H^{o}_{\text{bimol}} = \Delta H^{o}_{\text{helix1}} + \Delta H^{o}_{\text{helix2}} + \Delta H^{o}_{\text{bimol init}} + \Delta H^{o}_{\text{MBL}} - \Delta H^{o}_{\text{product mm}},$$

where helix 1 and helix 2 are the intermolecular paired helices with $\Delta H^{\circ}$ predicted from nearest neighbor parameters for Watson–Crick pairs (without including bimolecular initiation so that $\Delta H^{\circ}_{\text{bimol init}}$ appears only once). The $\Delta H^{o}_{\text{product mm}}$ is a term that accounts for the stacking enthalpy increment of the nucleotides that can stack on the hairpin loop stems to form a modified motif after the two strands have dissociated. This is the most favorable configuration with coaxial stacking of helixes (in the case of four-way multibranch loops) or of the stacking of unpaired nucleotides. $\Delta H^{\circ}_{\text{bimol}}$ is the experimental value which is taken from $T_{\text{M}}^{-1}$ versus ln($C_{\text{T}}$/4) plots. The multibranch loop enthalpy initiation term ($\Delta H^{\circ}_{\text{MBL init}}$) can be calculated from the above equation. The enthalpy of multibranch loops ( $\Delta H^{\circ}_{\text{MBL}}$) is then modeled as the sum of two terms, initiation and stacking:

$$\Delta H^{o}_{\text{MBL}} = \Delta H^{o}_{\text{MBL initiation}} + \Delta H^{o}_{\text{MBL stacking}}.$$

The stacking term is the favorable enthalpy of coaxial stacking, terminal mismatch and/or dangling end stacking. It is determined from the stacking conformation that gives the lowest free energy, as determined by free energy nearest neighbors (50). The initiation term can be approximated by the following equation:

$$\Delta H^{o}_{\text{MBL initiation}} = a + b \times \text{asym} + c \times h + \Delta H^{o}_{\text{strain}}(\text{three-way loops with fewer than two unpaired nucleotides}),$$

where $a$, $b$ and $c$ are parameters determined from linear regression (Table 7) and $h$ is the number of branching helices. $\Delta H^{\circ}_{\text{strain}}$ is a strain enthalpy that only applies to three-way multibranch loops with fewer than two unpaired nucleotides. The asym term is the average asymmetry that reflects the

**Table 7.** Enthalpy parameters for multibranch loop initiation

| Parameter[a] | Value (kcal/mol) | SE (kcal/mol) |
|---|---|---|
| $a$ | 38.9 | 14.2 |
| $b$ | 12.9 | 2.9 |
| $c$ | −11.9 | 3.7 |
| $\Delta H^{o}_{strain}$ | 27.1 | 6.8 |

[a]The $b$ term is excluded in the dynamic programming algorithm prediction of secondary structure. And the parameters a and c were optimized to be $a = 30.0$ kcal/mol and $c = -2.2$ kcal/mol in the dynamic programming calculation to achieve the highest prediction sensitivity.

distribution of unpaired nucleotides, which is defined by the following equation:

$$\text{asym} = \min\left[2.0, \frac{\left(\sum_1^h |\ \text{unpaired nucleotides } 5'\text{-unpaired nucleotides } 3'\ |\ \right)}{h}\right].$$

The average asymmetry is limited to 2.0, following the rules suggested by free energy parameters. Asymmetry cannot be applied, however, by dynamic programming algorithms for secondary structure prediction (17,22). Thus, the $b$ term was excluded for secondary structure prediction and the parameters $a$ and $c$ were optimized by finding the parameters that lead to the highest average sensitivity of secondary structure prediction by free energy minimization. The maximum sensitivity of prediction was found with $a = 30.0$ kcal/mol and $c = -2.2$ kcal/mol.

### Database of RNA secondary structures

The revised enthalpy nearest neighbor model was tested with RNA sequences with known secondary structure from organisms with known optimal growth temperature. The structures were taken from comparative analysis databases (42–49,58,59). Small (16S) subunit rRNA sequences are divided into domains as defined by Jaeger *et al*. (39). Large (23S) subunit rRNA sequences are divided into domains of fewer than 700 nt each (18). The optimal growth temperatures of different organisms were taken from the Prokaryotic Growth Temperature Database (http://pgtdb.csie.ncu.edu.tw/) and the DSMZ German Collection of Microorganisms and Cell Cultures website (http://www.dsmz.de/). Only the RNA sequences of mesophiles (organisms living at temperatures between 10 and 60°C, but with organisms living at 37°C excluded) were chosen to test the sensitivity and positive predictive value (PPV) of secondary structure prediction. Considering that posttranscriptional modification (60) and high pressure (61) in the thermophiles and hyperthermophiles (organism living above 60°C) would change the thermodynamics of secondary structure formation, sequences from these organisms were excluded. A list of sequences and optimal growth temperatures used are available in Supplementary Data.

### Accuracy of secondary structure prediction

The accuracy of structure prediction is determined by the sum of the canonical base pairs correctly predicted. A base pair is considered correctly predicted even if it is shifted by 1 nt on one side. For example a base pair between nucleotides $i$ and $j$ is considered to be correctly predicted if any of these base pairs is predicted: $i$ to $j$, $i$ to $j - 1$, $i$ to $j + 1$, $i - 1$ to $j$ or $i + 1$ to $j$. The predicted base pair between $i - 1$ and $j + 1$, however, is not considered to be correct. This scoring scheme reflects the uncertainty of exact base pair matches in comparative sequence analysis and the possibility for dynamics in base pairing. The values of sensitivity and PPV of this scoring scheme are ~2–3% higher than when determined with exact base pairing only, where only the $i$ to $j$ base pair is considered to be correct. The prediction accuracies are shown in Supplementary Tables 11 and 12. Each table includes accuracies determined when pairs can be shifted and when pairs must be an exact match.

### Availability of parameters

Machine-readable tables of the enthalpy parameters are available on the Mathews lab website (http://rna.urmc.rochester.edu/).

## RESULTS

### Nearest neighbor model parameters

In the nearest neighbor model of free energy (17,18), the parameters for Watson–Crick base pairs are well determined at 37°C with errors <10%, or ~0.1–0.2 kcal/mol (21). For other motifs such as loops and GU base pairs, individual nearest neighbor free energy increments are often determined with an error <0.5 kcal/mol (17,18). In order to extend the current model to predict free energy at temperatures other than 37°C, enthalpy parameters consistent with the current nearest neighbor model are required. The free energy at arbitrary temperature for each parameter is then

$$\Delta G^{o}(T) = \Delta H^{o} - T\Delta S^{o} = \Delta H^{o} - T[\Delta H^{o} - \Delta G(37\,^{\circ}\mathrm{C})]/310.15, \quad \mathbf{1}$$

where the enthalpy ($\Delta H^{\circ}$) and entropy ($\Delta S^{\circ}$) are assumed to be temperature independent. As described in Materials and Methods, parameters for enthalpy prediction, compatible with the free energy model, were determined using available experimental data from optical melting experiments.

Experimental studies consistently demonstrate that enthalpy and entropy measurements have considerably larger percent error than free energy measurements. Free energy at 37°C is determined with greater precision because of correlation between errors in enthalpy and entropy (21). The larger experimental errors in enthalpy result in larger percent errors for enthalpy nearest neighbor parameters than free energy parameters. The enthalpy of RNA secondary structure is known to be a function of temperature. A linear model for heat capacity change predicts the following:

$$\Delta H^{o}(T) = \Delta H^{o}(T_0) + \Delta C^{o}_{p}(T - T_0), \quad \mathbf{2}$$

$$\Delta S^{o}(T) = \Delta S^{o}(T_0) + \Delta C^{o}_{p}\ln(T/T_0), \quad \mathbf{3}$$

where $\Delta C^{\circ}_{p}$ is a constant heat capacity change and $T_0$ is a chosen reference temperature. It is hypothesized that the heat capacity change arises from the extent of stacking

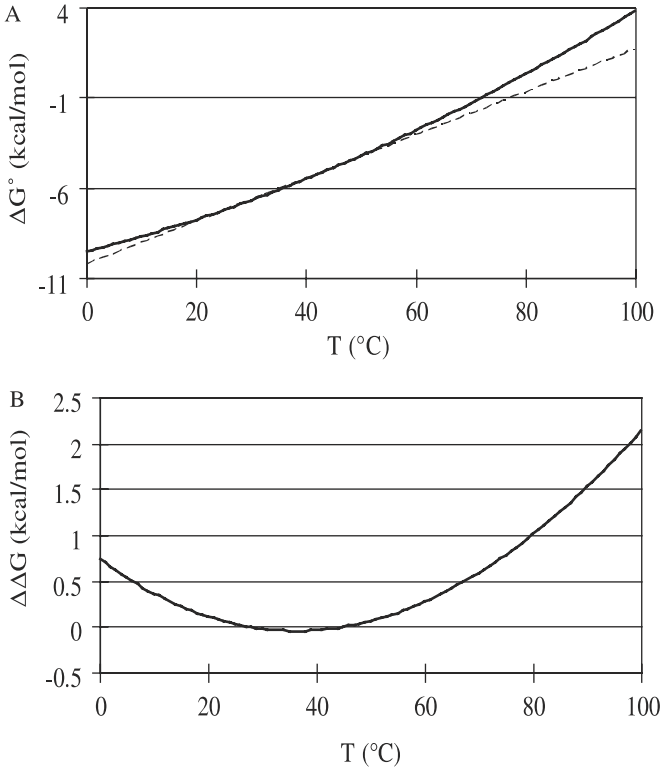


**Figure 1.** (**A**) Free energy difference of RNA duplex CCGGUp. $\Delta G^\circ$ (dashed line) was derived from Equation 3, where enthalpy and entropy were averaged from the optical melting curve fits, assuming that they were independent of the temperature. $\Delta G^\circ_T$ (solid line) was calculated from Equations 1–3, where the heat capacity was accounted. (**B**) Free energy difference is $\Delta\Delta G^\circ = \Delta G^\circ_T - \Delta G^\circ$ (62).

**Table 8.** Free energy differences of RNA duplexes

| Sequence | $\Delta G^\circ$(39°C)[a] (kcal mol$^{-1}$) | $\Delta C^\circ_p$ (cal K$^{-1}$ mol$^{-1}$) | $\Delta\Delta G^\circ$ (kcal mol$^{-1}$)[b] 0°C | 10°C | 60°C | 75°C | 100°C |
|---|---|---|---|---|---|---|---|
| CCGG | −4.36 | −382 | 0.5 | 0.2 | 0.9 | 1.4 | 3.1 |
| CCGGAp | −6.58 | −263 | 0.9 | 0.5 | 0.2 | 0.4 | 1.3 |
| CCGGUp | −5.56 | −355 | 0.8 | 0.4 | 0.4 | 0.8 | 2.1 |
| ACCGGp | −5.39 | −393 | 0.8 | 0.4 | 0.5 | 0.9 | 2.5 |
| ACCGGUp | −8.17 | −434 | 1.7 | 1.0 | −0.1 | 0.2 | 1.3 |

[a]Experimental results of total free energy at 39°C.
[b]Free energy difference: $\Delta\Delta G^\circ = \Delta G^\circ_T - \Delta G^\circ$, where $\Delta G^\circ$ is derived from Equation 3, assuming that the enthalpy and entropy were independent of the temperature and $\Delta G^\circ_T$ is calculated from Equations 1–3, including the non-zero heat capacity (73).

increasing with decreasing temperature. Thus, $\Delta C^\circ_p$ is negative because single strands are more organized at low rather than high temperature (62–67). The $\Delta C^\circ_p$ can be estimated by linear fits of enthalpy and entropy changes as a function of melting temperature (50,51,62) or determined by isothermal titration calorimetry at multiple temperatures (68,69). However, the effects of heat capacity change on enthalpy and entropy are antagonistic in terms of free energy change:

$$\Delta G^\circ(T) = \Delta H^\circ(T) - T\Delta S^\circ(T), \quad \textbf{4}$$

Therefore, for certain $\Delta T$ ($\Delta T = T - T_0$), $\Delta C^\circ_p$ can be neglected because the effects are compensated in terms of free energy. To calculate the compensation for a set of RNA duplexes (62), the free energy, $\Delta G^\circ$, was derived directly from Equation 4 assuming that the entropy and enthalpy were independent of temperature. Then the temperature-dependent free energy, $\Delta G^\circ_T$, was calculated with the measured non-zero $\Delta C^\circ_p$ from Equations 2–4. The free energy difference, $\Delta\Delta G^\circ = \Delta G^\circ_T - \Delta G^\circ$, increases with the deviation of temperature from $T_0$ (37°C) (Figure 1). The exact $\Delta\Delta G^\circ$ for each duplex is shown in Table 8 for different temperatures. The experimental error in individual loop free energy nearest neighbor parameters at 37°C is as large as 0.5 kcal/mol (17), which corresponds to roughly a factor of 2 in equilibrium constant. Thus, the small $\Delta\Delta G^\circ$ for helices suggests that the approximation of $\Delta C^\circ_p = 0$ is reasonable for predictions from ~10 to 60°C. Therefore, the enthalpy parameters derived here assume $\Delta C^\circ_p = 0$ and are most accurate at predicting free energy change close to 37°C.

## Dynamic programming algorithm for RNA secondary structure prediction

RNAstructure is a program for RNA secondary structure prediction and analysis. It includes prediction of secondary structure by free energy minimization (17), prediction of base pair probabilities using a partition function (22), the efn2 function for predicting the free energy change of folding given a sequence and secondary structure (18), and the Dynalign algorithm for finding the secondary structure common to two sequences (70). RNAstructure was revised to make predictions at user-defined temperature. Because large internal loops are more likely at high temperature, the previous limitation on internal loop size (fewer than 30 unpaired nucleotides) (17,18,22) was removed by implementing the method of Lyngsø *et al*. (71). This provides an $O(N^3)$ algorithm that can predict internal loops of arbitrary size. Benchmarks for calculation time and memory requirement with and without this revision are shown in Table 9.

## Sensitivities and PPVs of structure predictions

The enthalpy nearest neighbor parameters were compared with the previous parameters and model for enthalpy and free energy assembled by Serra and Turner (24) by predicting the secondary structures of RNA sequences with known secondary structures. Sensitivities, the percent of known base pairs that are correctly predicted, using both sets of parameters are shown in Figure 2 (detailed numbers are in Supplementary Table 11A) for different types of structural RNA sequences. The known structures of these sequences were taken from comparative analysis databases (42–49,58,59). The average sensitivity is improved from 65.2 to 68.9% using the new parameters assembled here. Sensitivities are improved for most types of the RNA. The exceptions are 5S rRNA and Group II introns.

To test the enthalpy parameters, the accuracy of secondary structure prediction at optimal growth temperature was compared to the accuracy of structure prediction at 37°C for organisms that do not grow optimally at 37°C for several types of RNAs (Table 10). The comparison of predictions was shown in different groups divided by optimal growth

**Table 9.** Calculation time and memory size of dynamic programming for sequences of different length

| Sequence | Length (nt) | $O(N^4)$ Time (h:min:s) | | Memory (MB) | $O(N^3)$ Time (h:min:s) | | Memory (MB) |
|---|---|---|---|---|---|---|---|
| *E.coli* arginine tRNA | 77 | 00:00:00.3 | (00:00:00.3) | 13 | 00:00:00.2 | (00:00:00.3) | 13 |
| *Bacillus stearo thermophilus* SRP | 268 | 00:00:21 | (00:00:03) | 13 | 00:00:04 | (00:00:03) | 14 |
| *Tetrahymena thermophila group I intron* | 433 | 00:02:27 | (00:00:12) | 15 | 00:00:14 | (00:00:11) | 16 |
| *S.cerevisiae A5 group II intron* | 631 | 00:11:59 | (00:00:35) | 16 | 00:00:46 | (00:00:34) | 19 |
| *E.coli* small subunit rRNA | 1542 | 06:09:03 | (00:06:47) | 31 | 00:13:42 | (00:07:41) | 45 |
| *E.coli* large subunit rRNA | 2904 | 67:00:43 | (00:47:00) | 73 | 01:41:00 | (01:03:54) | 121 |

Calculation size and time on a computer with Pentium 4, 3.2 GHz, processor and 1 GB of RAM using the gcc (version 3.2.3) compiler on Red Hat Enterprise Linux 3. The algorithm was improved from $O(N^4)$ to $O(N^3)$ in time complexity. In parentheses are the results with a limitation of internal loop size set at fewer than or equal to 30 unpaired nucleotides. The $O(N^3)$ algorithm is the implementation of the Lyngsø *et al.* (71) algorithm.

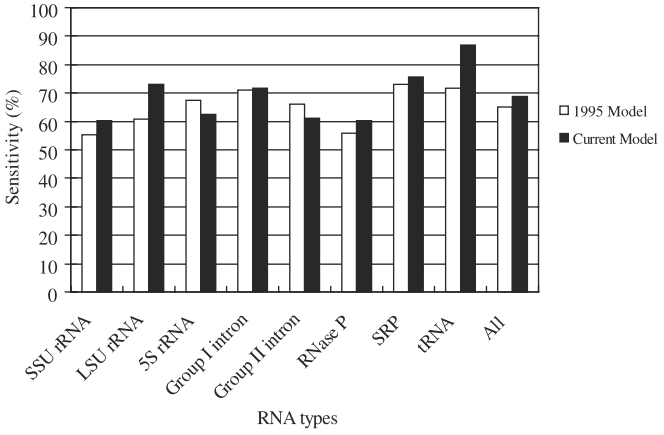


**Figure 2.** Improvement of prediction at optimal growth temperatures. The sequences are those from mesophiles (optimal growth temperature from 10 to 60°C) without organisms with optimal growth at 37°C. The lowest free energy secondary structures were predicted at the organims' optimal growth temperatures using two models. The previous model and parameters are those of Serra and Turner (24), which are widely used. The improved prediction uses the model and parameters presented in this work. The small and large subunits of rRNA sequences are divided into domains of <700 nt. The total sensitivity is the average of sensitivities of different types of RNA.

**Table 10.** Prediction sensitivities of the lowest free energy structure[a]

| Organisms' optimal growth temperature[b] (°C) | Nucleotides | Average sensitivity (%)[c] Prediction at 37°C | Prediction at optimal growth temperature |
|---|---|---|---|
| ⩽21 | 5536 | 79 ± 19.4 | 62.8 ± 28.4 |
| 22–26 | 7459 | 70.6 ± 13.4 | 71.0 ± 12.6 |
| 27–31 | 20 877 | 66.8 ± 10.4 | 67.7 ± 9.6 |
| 32–36 | 3124 | 64.9 ± 15.9 | 72.4 ± 21.8 |
| 38–42 | 1471 | 79.8 ± 2.2 | 79.8 ± 2.2 |
| 43–47 | 6268 | 78.3 ± 16.3 | 75.6 ± 20.2 |
| 48–52 | 1255 | 75.4 ± 14.8 | 71.3 ± 19.3 |
| 53–57 | 385 | 87.7 ± 8.6 | 90.8 ± 13.0 |
| 58–62 | 2937 | 84.5 ± 15.2 | 84.1 ± 15.7 |
| ⩾63 | 12 395 | 76.9 ± 11.2 | 48.6 ± 11.3 |

[a]The sequences are those from organisms with optimal growth temperature from 10 to 90°C, excluding 37°C.
[b]The prediction at 37°C and optimal growth temperature for the organisms growing in different range of temperatures, using the current model in Materials and Methods.
[c]Sensitivity equals the number of correctly predicted base pairs divided by the total number of known base pairs. The average sensitivity is the average of sensitivities of available types of RNA at different range of temperatures.

temperature. The organisms in each group grow optimally in a certain range of temperatures. Compared to the prediction at 37°C, structure prediction at optimal growth temperature performs better for the organism living at temperatures between 22 and 37°C, but is worse at other optimal growth temperatures. This suggests that when enthalpy parameters are assumed to be temperature independent, their utility as a tool for deriving free energy parameters for use in predicting the lowest free energy structure is limited to a narrow temperature range. Small errors in enthalpy change parameters have a larger effect on free energy change parameter determination (Equation 1), the farther the temperature is from 37°C.

Figure 3 shows the PPV for base pairs from the lowest free energy structure for base pairs with different pairing probabilities (see detailed numbers in Supplementary Table 12A). They are predicted using a partition function calculation at optimal growth temperature (22). PPV is the percentage of predicted base pairs that are found in the known structure. The average PPV of all pairs in the lowest free energy structures is only 62.0%, which is lower than the sensitivity (68.9%). This suggests that the model over-predicts base pairs and/or that the base pairs may not be annotated completely in the structures from comparative analysis (22). For example, if a base pair is completely conserved, then it is sometimes not annotated by comparative analysis (42–49,58,59). Base pair probabilities for all possible pairs are calculated with a partition function and grouped by different thresholds. The PPV is significantly higher for predicted base pairs in the lowest free energy structure with higher pairing probability. The average PPV is up to 90.4% for those known base pairs having probability of 0.99 or above. It has been demonstrated previously that base pair probabilities predicted at 37°C can be used to find pairs with high PPV (22). The fact that this holds true at other temperatures shows that the enthalpy parameters are robust for base pair probability prediction.

The fact that the accuracy of secondary structure prediction is sensitive to the accuracy of the nearest neighbor parameters, but the base pair probabilities remain a robust measure of confidence for a wide variety of temperatures is consistent with a previous work. Layton and Bundschuh (72) demonstrated that the predicted lowest free energy structure was often changed in repeated structure predictions after random adjustments of the nearest neighbor parameters within the limits of their error. Base pair probabilities, however, were

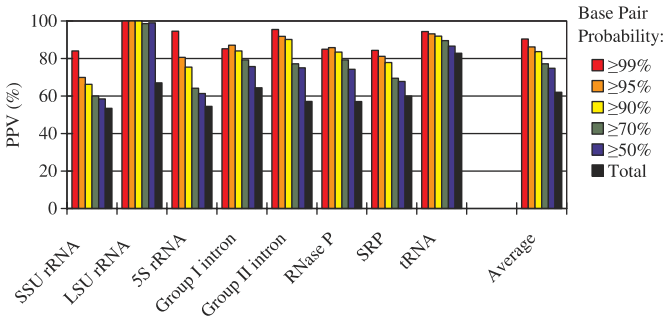


**Figure 3.** PPV for optimal structure and base pairs with different pairing probabilities. PPV equals the number predicted base pairs in that are in the known structure divided by total number of predicted base pairs. Pairs in the optimal structures are grouped by different thresholds of pairing probabilities. The pairing probabilities were calculated with a partition function calculation (22) at organisms' optimal growth temperatures, using the model and parameters presented in Materials and Methods. The small and large subunits of rRNA sequences are divided into domains of <700 nt. The sequences of different type of RNA are those from mesophiles (living from 10 to 60°C) without organisms living at 37°C.

less perturbed by changes in the parameters (72). With the extrapolation of nearest neighbor parameters to temperatures far from 37°C, the accuracy of the predicted lowest free energy structure is often reduced as compared to structure prediction at 37°C. The ability of the partition function predicted base pair probabilities to determine base pairs predicted with a higher confidence is unchanged with secondary structure prediction at temperatures far from 37°C. This is because the determination of base pair probabilities is not as perturbed by errors in the nearest neighbor parameters.

An example of secondary structure prediction at 37°C and at optimal growth temperature of 30°C is shown in Figure 4 for a tRNA sequence. The base pairs with higher predicted pairing probability (color annotated according to pairing probability in Figure 4B and C) are pairs predicted with greater confidence. For this sequence, secondary structure prediction is more accurate and the fidelity of structure prediction (as judged by the percent of high probability pairs) is improved at optimal growth temperature.

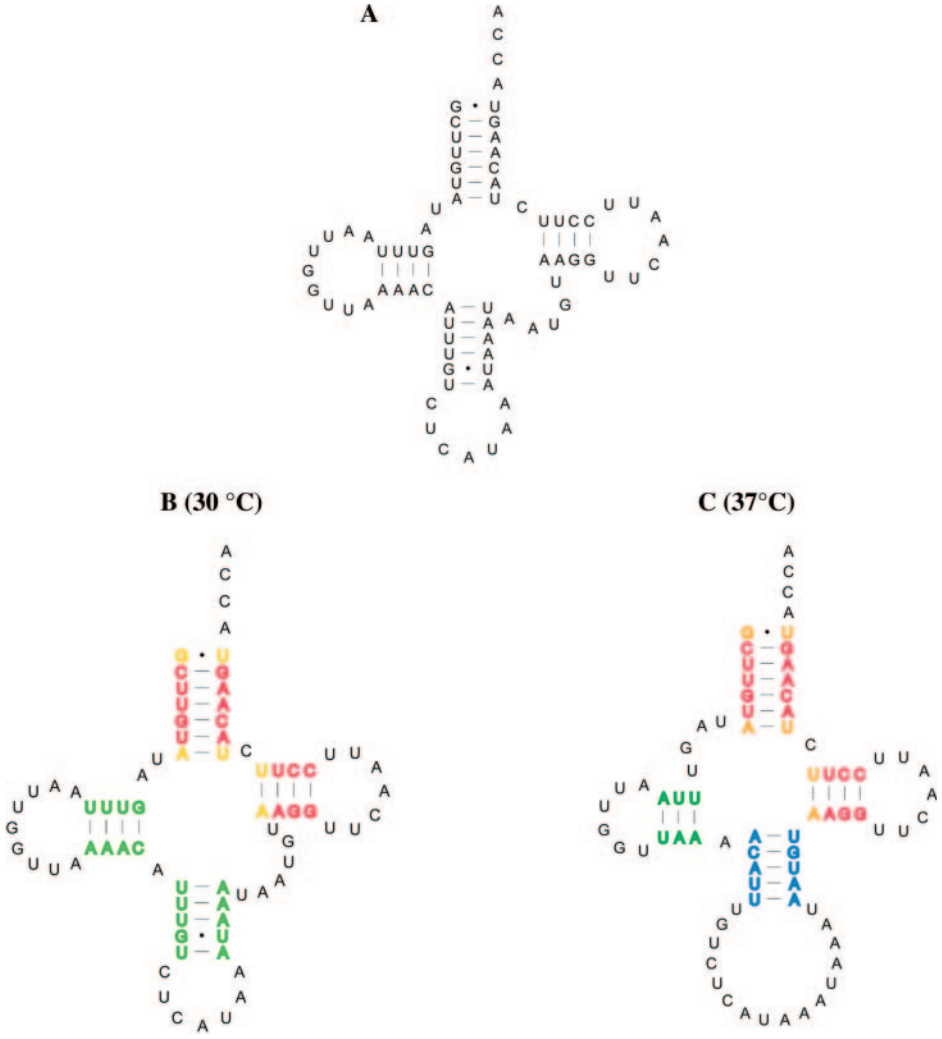


**Figure 4.** Secondary structure prediction of *Saccharomyces cerevisiae* tRNA (RM4000) at optimal growth temperature (30°C) (**B**) and at 37°C (**C**) with the presented nearest neighbor parameters. Base pairs in the original structure (**A**) are derived from the comparative analysis database (42–49,58,59). Structures are also color annotated to indicate predicted base pair probabilities ($P_{bp}$) for each helix: red, $P_{bp} \geqslant 0.95$; yellow, $0.95 > P_{bp} \geqslant 0.7$; green, $0.7 > P_{bp} \geqslant 0.3$; blue, $0.3 > P_{bp}$. The structures were drawn with XRNA (http://rna.ucsc.edu/rnacenter/xrna/xrna.html) and Adobe Illustrator.

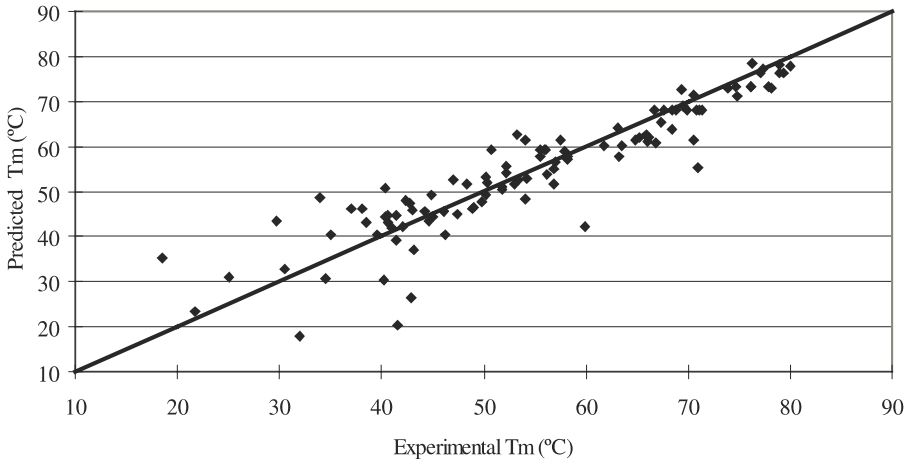


**Figure 5.** Experimental (Supplementary Data) (25–31) versus predicted ($T_m = \Delta H^\circ/\Delta S^\circ - 273.15$) melting temperatures of hairpin stem–loop structures. The line shows the ideal location of points, predicted $T_m$ = measured $T_m$. The root mean squared deviation (r.m.s.d.) of prediction compared to experiment is 5.86°C. The new enthalpy parameters provide improved $T_m$ prediction compared to the previous compilation of parameters (24), which have an r.m.s.d. of 7.58°C as compared to experiment for this dataset.

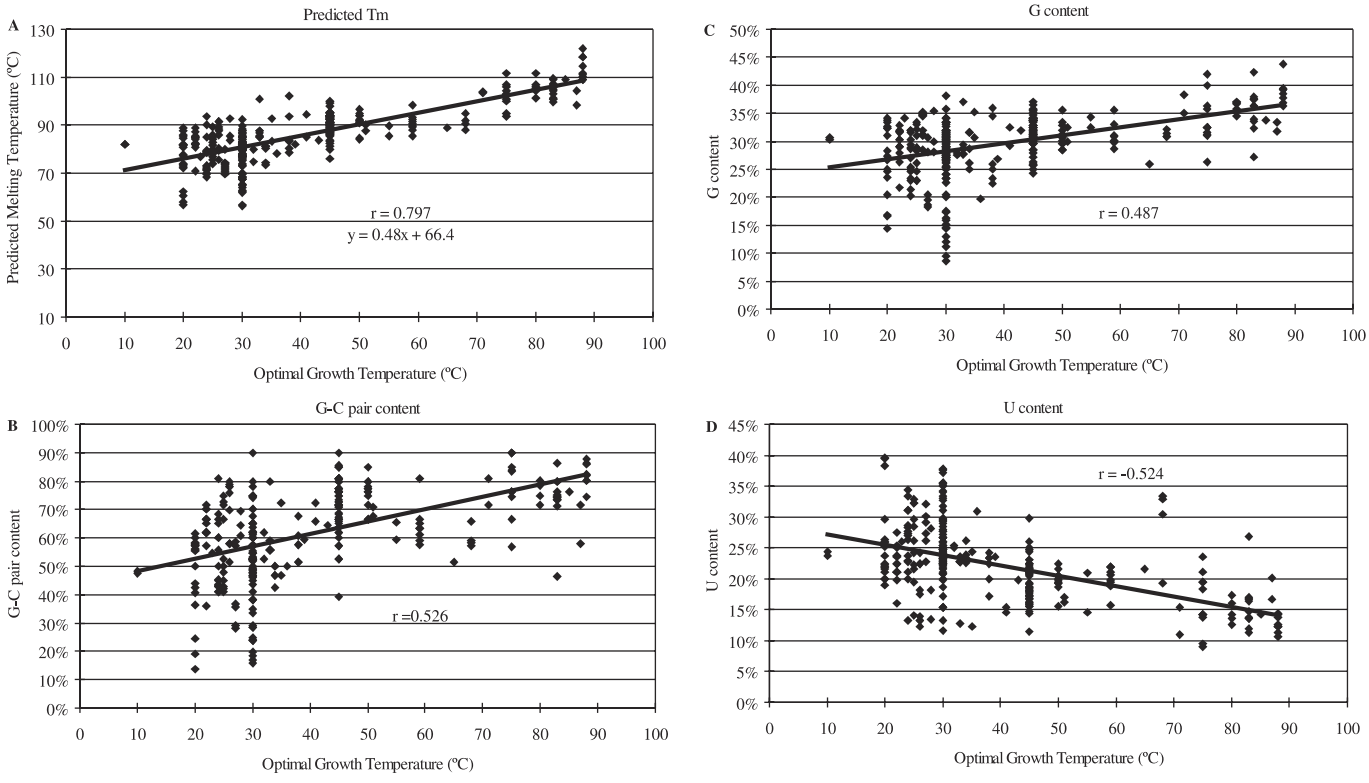


**Figure 6.** Relationships of melting temperatures, nucleotide contents and optimal growth temperatures of different types of RNA in different organisms with optimal growth temperature from 10 to 90°C: (**A**) Predicted melting temperature; (**B**) G–C pair content; (**C**) G content; and (**D**) U content versus optimal growth temperature. Melting temperatures are predicted for different types of RNA sequences from comparative analysis databases (42–49,58,59) with a two-state transition assumption.

## Correlation between melting temperature and optimal growth temperature

Melting temperature, $T_m$, is defined as the temperature at which half of strands are unpaired. Assuming that an RNA melts with a two-state transition, the melting temperature (in Kelvins) of a single-stranded RNA structure can be predicted by $T_m = \Delta H^\circ/\Delta S^\circ$ (73). For example, the predicted melting temperatures (°C) for all hairpins in the database of optically melted sequences (Supplementary Data) (25–31) are plotted in Figure 5 as a function of experimentally determined $T_m$. This shows that the parameters adequately reflect the thermal stabilities of RNA sequences with known $T_m$. Better

correlation was found at higher temperatures. This is expected because most hairpins were measured with high melting temperatures in experiments (25–31).

Melting temperature reflects the thermal stability of a structure. Therefore RNA structures in organisms living at higher temperature are expected to have higher melting temperatures. Figure 6A shows a plot of predicted melting temperatures of the lowest free energy structure versus organism optimal growth temperature (10–90°C). A strong correlation (linear correlation coefficient of 0.797) is found between the melting temperature and the optimal growth temperature for different types of RNA structures. On the other hand, there appears to be less correlation between nucleotide content and optimal growth temperature (Figure 6B–D) for diverse types of RNA, although uracil content of 16S rRNA of thermophiles and psychrophiles were found recently to correlate inversely with their optimal growth temperatures (74). Evidently, the thermal stability of RNA structure is not simply controlled by base content. Organisms that grow at high temperature have apparently evolved RNA secondary structures with a combination of motifs that provide thermal stability.

## DISCUSSION

The nearest neighbor parameters for enthalpy were derived here using similar rules as for free energy nearest neighbor parameters at 37°C (17). This makes these parameters useful for determining free energy parameters at arbitrary temperature that are compatible with dynamic programming algorithms for secondary structure prediction. Some of the enthalpy parameters have large percent standard errors as compared with the parameters of free energy. This reflects the larger errors in the experimental results of enthalpy than free energy, but it also suggests that enthalpy may be more sequence dependent than free energy. This sequence dependence cannot be determined using the currently available database of optical melting experiments and suggests a need for further optical melting experiments on model RNA systems.

Another source of error comes from the assumption that the enthalpy and entropy are independent of the temperature in both the model and in the analysis of optical melting experiments. When the temperature is too far from 37°C, the sensitivity of prediction is expected to be worse than 68.9% on average because of the approximation of $\Delta C_p^\circ = 0$. For example, experiments demonstrate cold denaturation of RNA (68,69), but the nearest neighbor model does not reproduce those results. Further experiments by isothermal titration calorimetry would be needed to provide the data for a model that can include a non-zero heat capacity change.

There are common error sources that should be considered for the prediction of base pairs. Free energy minimization assumes that the secondary structure is at equilibrium. The nearest neighbor model is an incomplete representation of structural free energy. The parameters average some sequence-specific effects and were derived from a limited set of experiments. Some RNA sequences, in particular mRNA, may sample multiple structures at equilibrium. The parameters are derived from experimental data at 1 M NaCl, whereas the salt concentration in different organisms may be very different.

In spite of all these limitations, the nearest neighbor model predicts secondary structures with a 72.8% average sensitivity (17). Recent experimental results on the self-folding of the 16S rRNA 5′ domain (75) support the assumption of thermodynamic control of folding pathway. Moreover, the base pair prediction with the partition function can be used to determine pairs predicted with greater confidence (22).

In spite of the fact that the enthalpy parameters have larger percent errors than the free energy parameters for 37°C, the enthalpy parameters are able to predict optical melting temperatures for small model sequences. Predicted melting temperatures for structural RNA sequences correlate well with optimal growth temperature, suggesting that these parameters capture many of the sequence-dependent features of RNA folding enthalpy change.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors thank Rahul Tyagi and Andrew V. Uzilov for helpful discussions. D.H.M. is an Alfred P. Sloan Research Fellow. This work was supported by National Institutes of Health Grants GM22939 to D.H.T. and GM076485 to D.H.M. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Nelson,P., Kiriakidou,M., Sharma,A., Maniataki,E. and Mourelatos,Z. (2003) The microRNA world: small is mighty. *Trends Biochem. Sci.*, **28**, 534–540.
2. Doudna,J. and Cech,T. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
3. Walter,P. and Blobel,G. (1982) Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**, 691–698.
4. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
5. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
6. Onoa,B. and Tinoco,I.,Jr (2004) RNA folding and unfolding. *Curr. Opin. Struct. Biol.*, **14**, 374–379.
7. Crothers,D.M., Cole,P.E., Hilbers,C.W. and Schulman,R.G. (1974) The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J. Mol. Biol.*, **87**, 63–88.
8. Mathews,D.H., Banerjee,A.R., Luan,D.D., Eickbush,T.H. and Turner,D.H. (1997) Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA*, **3**, 1–16.
9. Banerjee,A.R., Jaeger,J.A. and Turner,D.H. (1993) Thermal unfolding of a group I ribozyme: the low temperature transition is primarily a disruption of tertiary structure. *Biochemistry*, **32**, 153–163.
10. Woodson,S.A. (2000) Recent insights on RNA folding mechanisms from catalytic RNA. *Cell Mol. Life Sci.*, **57**, 796–808.

11. James,B.D., Olsen,G.J. and Pace,N.R. (1989) Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol.*, **180**, 227–239.
12. Pace,N.R., Thomas,B.C. and Woese,C.R. (1999) Probing RNA structure, function, and history by comparative analysis. In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World,* 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 113–141.
13. Woese,C.R., Gutell,R.R., Gupta,R. and Noller,H.F. (1983) Detailed analysis of the higher order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.*, **47**, 621–669.
14. Andronescu,M., Aguirre-Hernandez,R., Condon,A. and Hoos,H.H. (2003) RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.
15. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
16. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
17. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.
18. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA Secondary Structure. *J. Mol. Biol.*, **288**, 911–940.
19. Turner,D.H. (2000) Conformational changes. In Bloomfield,V., Crothers,D. and Tinoco,I. (eds), *Nucleic Acids.* University Science Books, Sausalito, CA, pp. 259–334.
20. Xia,T., Mathews,D.H. and Turner,D.H. (1999) Thermodynamics of RNA secondary structure formation. In Soll,D.G., Nishimura,S. and Moore,P.B. (eds), *Prebiotic Chemistry, Molecular Fossils, Nucleosides, and RNA.* Elsevier, NY, pp. 21–47.
21. Xia,T., SantaLucia,J.,Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick pairs. *Biochemistry*, **37**, 14719–14735.
22. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
23. McCaskill,J.S. (1990) The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
24. Serra,M.J. and Turner,D.H. (1995) Predicting thermodynamic properties of RNA. *Methods Enzymol.*, **259**, 242–261.
25. Giese,M.R., Betschart,K., Dale,T., Riley,C.K., Rowan,C., Sprouse,K.J. and Serra,M.J. (1998) Stability of RNA hairpins closed by wobble base pairs. *Biochemistry*, **37**, 1094–1100.
26. Serra,M.J., Lyttle,M.H., Axenson,T.J., Schadt,C.A. and Turner,D.H. (1993) RNA hairpin loop stability depends on closing pair. *Nucleic Acids Res.*, **21**, 3845–3849.
27. Serra,M.J., Axenson,T.J. and Turner,D.H. (1994) A model for the stabilities of RNA hairpins based on a study of the sequence dependence of stability for hairpins of six nucleotides. *Biochemistry*, **33**, 14289–14296.
28. Serra,M.J., Barnes,T.W., Betschart,K., Gutierrez,M.J., Sprouse,K.J., Riley,C.K., Stewart,L. and Temel,R.E. (1997) Improved parameters for the prediction of RNA hairpin stability. *Biochemistry*, **36**, 4844–4851.
29. Groebe,D.R. and Uhlenbeck,O.C. (1988) Characterization of RNA hairpin loop stability. *Nucleic Acids Res.*, **16**, 11725–11735.
30. Dale,T., Smith,R. and Serra,M.J. (2000) A test of the model to predict unusually stable RNA hairpin loop stability. *RNA*, **6**, 608–615.
31. Antao,V.P. and Tinoco,I.,Jr (1992) Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.*, **20**, 819–824.
32. Longfellow,C.E., Kierzek,R. and Turner,D.H. (1990) Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, **29**, 278–285.
33. Znosko,B.M., Silvestri,S.B., Volkman,H., Boswell,B. and Serra,M.J. (2002) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*, **41**, 10406–10417.
34. Shu,Z. and Bevilacqua,P.C. (1999) Isolation and characterization of thermodynamically stable and unstable RNA hairpins from a triloop combinatorial library. *Biochemistry*, **38**, 15369–15379.
35. Proctor,D.J., Schaak,J.E., Bevilacqua,J.M., Falzone,C.J. and Bevilacqua,P.C. (2002) Isolation and characterization of stable tetraloops with the motif YNMG that participates in tertiary interactions. *Biochemistry*, **41**, 12062–12075.
36. Laing,L.G. and Hall,K.B. (1996) A model of the iron responsive element RNA hairpin loop structure determined from NMR and thermodynamic data. *Biochemistry*, **35**, 13586–13596.
37. Groebe,D.R. and Uhlenbeck,O.C. (1989) Thermal stability of RNA hairpins containing a four-membered loop and a bulge nucleotide. *Biochemistry*, **28**, 742–747.
38. Fink,T.R. and Crothers,D.M. (1972) Free energy of imperfect nucleic acid helices,I. The bulge defect. *J. Mol. Biol.*, **66**, 1–12.
39. Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.
40. Weeks,K.M. and Crothers,D.M. (1993) Major groove accessibility of RNA. *Science*, **261**, 1574–1577.
41. Chen,G., Znosko,B.M., Jiao,X. and Turner,D.H. (2004) Factors affecting thermodynamic stabilities of RNA $3 \times 3$ internal loops. *Biochemistry*, **43**, 12865–12876.
42. Gutell,R.R. (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res.*, **22**, 3502–3507.
43. Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Res.*, **21**, 3055–3074.
44. Schnare,M.N., Damberger,S.H., Gray,M.W. and Gutell,R.R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J. Mol. Biol.*, **256**, 701–719.
45. Szymanski,M., Specht,T., Barciszewska,M.Z., Barciszewski,J. and Erdmann,V.A. (1998) 5S rRNA data bank. *Nucleic Acids Res.*, **26**, 156–159.
46. Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A. and Steinberg,S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.
47. Larsen,N., Samuelsson,T. and Zwieb,C. (1998) The signal recognition particle database (SRPDB). *Nucleic Acids Res.*, **26**, 177–178.
48. Brown,J.W. (1998) The ribonuclease P database. *Nucleic Acids Res.*, **26**, 351–352.
49. Damberger,S.H. and Gutell,R.R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res.*, **22**, 3508–3510.
50. Mathews,D.H. and Turner,D.H. (2002) Experimentally derived nearest neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869–880.
51. Diamond,J.M., Turner,D.H. and Mathews,D.H. (2001) Thermodynamics of three-way multibranch loops in RNA. *Biochemistry*, **40**, 6971–6981.
52. Walter,A.E., Wu,M. and Turner,D.H. (1994) The stability and structure of tandem GA mismatches in RNA depend on closing base pairs. *Biochemistry*, **33**, 11349–11354.
53. Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Miller,P., Mathews,D.H. and Zuker,M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.
54. Kim,J., Walter,A.E. and Turner,D.H. (1996) Thermodynamics of coaxially stacked helices with GA and CC mismatches. *Biochemistry*, **35**, 13753–13761.
55. SantaLucia,J.,Jr, Kierzek,R. and Turner,D.H. (1991) Functional group substitutions as probes of hydrogen bonding between GA mismatches in RNA internal loops. *J. Am. Chem. Soc.*, **113**, 4313–4322.
56. Burkard,M.E. and Turner,D.H. (2000) NMR structures of $r(GCAGGCGUGC)_2$ and determinants of stability for single guanosine–guanosine base pairs. *Biochemistry*, **39**, 11748–11762.
57. Jacobson,H. and Stockmayer,W.H. (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.*, **18**, 1600–1606.
58. Michel,F., Umesono,K. and Ozeki,H. (1989) Comparative and functional anatomy of group II catalytic introns—a review. *Gene*, **82**, 5–30.

59. Waring,R.B. and Davies,R.W. (1984) Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing—a review. *Gene*, **28**, 277–291.
60. Kowalak,J.A., Dalluge,J.J., McCloskey,J.A. and Stetter,K.O. (1994) The role of posttranscriptional modification in stabilization of transfer RNA from hyperthermophiles. *Biochemistry*, **33**, 7869–7876.
61. Dubins,D.N., Lee,A., Macgregor,R.B.,Jr and Chalikian,T.V. (2001) On the stability of double stranded nucleic acids. *J. Am. Chem. Soc*., **123**, 9254–9259.
62. Petersheim,M. and Turner,D.H. (1983) Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry*, **22**, 256–268.
63. Holbrook,J.A., Capp,M.W., Saecker,R.M. and Record,M.T.,Jr (1999) Enthalpy and heat capacity changes for formation of an oligomeric DNA duplex: interpretation in terms of coupled processes of formation and association of single-stranded helices. *Biochemistry*, **38**, 8409–8422.
64. Suurkuusk,J., Alvarez,J., Freire,E. and Biltonen,R. (1977) Calorimetric determination of the heat capacity changes associated with the conformational transitions of polyriboadenylic acid and polyribouridylic acid. *Biopolymers*, **16**, 2641–2652.
65. Pörschke,D., Uhlenbeck,O.C. and Martin,F.H. (1973) Thermodynamics and kinetics of the helix-coil transition of oligomers containing GC base pairs. *Biopolymers*, **12**, 1313–1335.
66. Appleby,D.W. and Kallenbach,N.R. (1973) Theory of oligonucleotide stabilization. I. The effect of single-strand stacking. *Biopolymers*, **12**, 2093–2120.
67. Freier,S.M., Hill,K.D., Dewey,T.G., Marky,L.A., Breslauer,K.J. and Turner,D.H. (1981) Solvent effects on the kinetics and thermodynamics of stacking in poly(cytidylic acid). *Biochemistry*, **20**, 1419–1426.
68. Takach,J.C., Mikulecky,P.J. and Feig,A.L. (2004) Salt-dependent heat capacity changes for RNA duplex formation. *J. Am. Chem. Soc*., **126**, 6530–6531.
69. Mikulecky,P.J., Takach,J.C. and Feig,A.L. (2004) Entropy-driven folding of an RNA helical junction: an isothermal titration calorimetric analysis of the hammerhead ribozyme. *Biochemistry*, **43**, 5870–5881.
70. Mathews,D.H. (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics*, **21**, 2246–2253.
71. Lyngsø,R., Zuker,M. and Pederson,C. (1999) Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, **15**, 440–445.
72. Layton,D.M. and Bundschuh,R. (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res*., **33**, 519–524.
73. Borer,P.N., Dengler,B., Tinoco,I.,Jr and Uhlenbeck,O.C. (1974) Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol*., **86**, 843–853.
74. Khachane,A.N., Timmis,K.N. and dos Santos,V.A. (2005) Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. *Nucleic Acids Res*., **33**, 4016–4022.
75. Adilakshmi,T., Ramaswamy,P. and Woodson,S.A. (2005) Protein-independent folding pathway of the 16S rRNA 5′ domain. *J. Mol. Biol*., **351**, 508–519.
76. Wu,M., McDowell,J.A. and Turner,D.H. (1995) A periodic table of symmetric tandem mismatches in RNA. *Biochemistry*, **34**, 3204–3211.
77. SantaLucia,J.,Jr, Kierzek,R. and Turner,D.H. (1991) Stabilities of consecutive A·C, C·C, G·G, U·C, and U·U mismatches in RNA internal loops: evidence for stable hydrogen-bonded U·U and $C{\cdot}C^{+}$ pairs. *Biochemistry*, **30**, 8242–8251.
78. Znosko,B.M., Burkard,M.E., Krugh,T.R. and Turner,D.H. (2002) Molecular recognition in purine-rich internal loops: thermodynamic, structural, and dynamic consequences of purine for adenine substitutions in 5′ (rGGCAAGCCU)2. *Biochemistry*, **41**, 14978–14987.
79. Schroeder,S.J. and Turner,D.H. (2001) Thermodynamic stabilities of internal loops with GU closing pairs in RNA. *Biochemistry*, **40**, 11509–11517.
80. Xia,T., McDowell,J.A. and Turner,D.H. (1997) Thermodynamics of nonsymmetric tandem mismatches adjacent to G·C base pairs in RNA. *Biochemistry*, **36**, 12486–12487.