Annu. Rev. Biophys. Biomol. Struct. 2004. 33:415–40
doi: 10.1146/annurev.biophys.32.110601.141800

THE THERMODYNAMICS OF DNA STRUCTURAL MOTIFS

John SantaLucia, Jr.[1,2] and Donald Hicks[2]
[1]*Department of Chemistry, Wayne State University, Detroit, Michigan 48202; email: jsl@chem.wayne.edu*
[2]*DNA Software, Inc., Ann Arbor, Michigan 48104; email: don@dna-software.com*

Key Words secondary structure, prediction, hybridization, oligonucleotides, nucleic acid folding

■ **Abstract** DNA secondary structure plays an important role in biology, genotyping diagnostics, a variety of molecular biology techniques, in vitro–selected DNA catalysts, nanotechnology, and DNA-based computing. Accurate prediction of DNA secondary structure and hybridization using dynamic programming algorithms requires a database of thermodynamic parameters for several motifs including Watson-Crick base pairs, internal mismatches, terminal mismatches, terminal dangling ends, hairpins, bulges, internal loops, and multibranched loops. To make the database useful for predictions under a variety of salt conditions, empirical equations for monovalent and magnesium dependence of thermodynamics have been developed. Bimolecular hybridization is often inhibited by competing unimolecular folding of a target or probe DNA. Powerful numerical methods have been developed to solve multistate-coupled equilibria in bimolecular and higher-order complexes. This review presents the current parameter set available for making accurate DNA structure predictions and also points to future directions for improvement.

CONTENTS

INTRODUCTION

Biological Importance of DNA Secondary Structure

Any time that DNA is single stranded it can fold back upon itself to form unimolecular folded structures in a fashion similar to that routinely observed for RNA (81). In biology, DNA is partially single stranded during replication (20, 81), transcription, recombination, and DNA repair. In most cases DNA secondary structure results in aberrant biological function; for example, triplet repeat expansion causes a number of neurological disorders (20). On the other hand, there is a whole class of single-stranded DNA viruses in which DNA secondary structure plays an essential role in protein recognition and defining the origin of replication (7, 19, 27, 31, 52, 59, 67, 78). In retroviruses and other RNA viruses, secondary structure in the single-strand DNA intermediates is important for mediating strand jumping and other activities (13, 46, 80). Thus, understanding the physical basis of DNA secondary structure contributes significantly to elucidating biological function.

Molecular Biology and Biotechnology Applications of DNA Secondary Structure

When genomic DNA is taken out of its biological context and used in molecular biology techniques, it becomes single stranded upon heat denaturation and can fold upon cooling. Such structure inhibits primer/probe hybridization needed for PCR, cDNA expression profiling, and a variety of genotyping and other genomic diagnostics. Formation of secondary structure by target DNAs is well documented to inhibit probe/primer hybridization (47, 55, 56, 79). Formation of hairpins by probe DNAs inhibits hybridization, causing false-negative results in various assays. In contrast, undesired bimolecular cross-hybridzation reactions between different probe DNAs and undesired hybridization to mismatch sites can cause false-positive signals in assays. The folding potential of DNA suggests that DNA can also fold into compact three-dimensional structures that possess catalytic activity similar to that observed for ribozymes. Although no DNA catalysts have been observed in biology to date, a large number of “deoxyribozymes” and ligand binding DNA aptamers have been discovered by in vitro selection experiments (10). A number

of biotechnology techniques that exploit the three-dimensional folding potential of DNA have also been demonstrated including DNA nanotechnology (75) and DNA computing (21).

The DNA Folding Problem

Similar to the protein and RNA folding problems, there is a corresponding "DNA folding problem" in which it is desired to predict the structure and folding energy of the DNA given its sequence. Fortunately, several features of DNA and RNA make them especially amenable to structure prediction. Notably, DNA and RNA secondary structures result from strong Watson-Crick pairing interactions, and tertiary interactions are a weaker second-order effect (81). Thus, to an excellent approximation, tertiary interactions may be neglected and accurate secondary structure prediction is possible. The strong pairing rules also allow for the DNA secondary structure to be reduced to discrete interactions in which two positions in a sequence are either paired or not. Even with the neglect of tertiary interactions such as pseudoknots, however, the number of possible secondary structures is approximately 1.8^N, where N is the sequence length (95). Fortunately, with the discrete pairing approximation, DNA and RNA are suitable for powerful dynamic programming algorithms, which were described in a previous review (83). Dynamic programming algorithms guarantee that for a given set of rules, the minimum energy structure (i.e., optimal) will be found in computation time order N^3 with memory order N^2, thereby allowing predictions of sequences with fewer than 10,000 nucleotides with currently available computers. Dynamic programming algorithms also predict suboptimal structures within user-defined energy and distance windows (94). This is important because the energy rules are not perfect and tertiary interactions are neglected (as are interactions with proteins and the specific interactions with magnesium or other cofactors). Thus, one of the few structures near the free-energy minimum is likely to be correct. It is important to note the important difference between selected functional sequences and random sequences of DNA or RNA. Random sequences have a low probability of folding into compact three-dimensional structures stabilized by tertiary interactions; thus random sequences are most amenable to secondary structure prediction because the neglect of tertiary interactions is appropriate. On the other hand, selected sequences (selected either by evolution or by in vitro selection, or rationally designed) are more likely to contain tertiary interactions, which compromise the reliability of the secondary structure prediction algorithms. This difference makes DNA folding much easier to predict (for random sequences) than corresponding biologically selected RNAs. Note that dynamic programming algorithms also neglect kinetically trapped structures and assume structures are populated according to an equilibrium Boltzmann distribution; thus the structures close to minimum free energy are most probable. Recently, we have also extended the dynamic programming algorithm to predict bimolecular optimal and suboptimal structures so that match and mismatch hybridizations of a short probe to long-target DNA may be readily identified on

the basis of thermodynamic rules rather than sequence similarity (J. SantaLucia, unpublished results).

Overview of the DNA Thermodynamic Database

Dynamic programming algorithms for DNA secondary structure prediction require a database of thermodynamic parameters for various DNA motifs, which is the main subject of this review. Figure 1 shows the structural motifs that occur in unimolecular folded DNAs as well as bimolecular hybridization. We have accumulated a nearly complete database of parameters for base pairs, mismatches, terminal dangling ends, terminal mismatches, coaxial stacking, and a variety of loop motifs including hairpins, bulges, internal loops, and multibranched loops. Methods for measurement of the thermodynamic parameters have been reviewed elsewhere (71, 72). Because it is not possible to measure all possible sequences for all the motifs, extrapolations with appropriate theories are used as an approximation. To make the database useful for a variety of solution conditions, empirical sodium and magnesium dependence equations have been developed. Tables of the parameters are provided and examples of their proper use are given so that researchers may utilize the database in their own work and also criticize our approach and improve upon them in the future. We note that the database presented is not appropriate for partition function computations (50; J. SantaLucia, unpublished results). The reliability of the parameters and the directions of future research are also discussed.

Software Implementations

We have incorporated the DNA database presented here into the DNA-MFOLD server (collaboration with Dr. Michael Zuker; http://www.bioinfo.rpi.edu/applications/mfold/), the HYTHER server (http://ozone.chem.wayne.edu), as well as our commercial software Visual OMP (Oligonucleotide Modeling Platform; DNA Software Inc., http://www.dnasoftware.com/). The parameters have also been provided to Dr. Ivo Hofacker for use in the Vienna package (http://www.tbi.univie.ac.at/~ivo/RNA/).

THERMODYNAMIC PARAMETER DATABASE

Watson-Crick Base Pair Nearest Neighbors.

Table 1 presents the thermodynamic nearest neighbor (NN) parameters for Watson-Crick base pairs in 1 M NaCl. These parameters were derived from multiple linear regression of 108 sequences solving for 12 unknowns (10 NN propagation parameters, 1 initiation parameter, and 1 correction for terminal AT pairs). Because the dataset originated from a variety of labs, the parameters are referred to as the unified NN (70). Detailed comparisons of the unified set to those of

previously published NN parameters have been critically reviewed (57, 70). This is discussed further below, but the essential point is that we have great confidence in the reliability of the Watson-Crick NN parameters. This is an important point because the Watson-Crick parameters form the foundation by which the rest of the thermodynamic database must be derived, namely, by measurement of thermodynamics of a motif in a larger sequence and then reliably subtracting Watson-Crick contribution. Equation 1 shows an example of the application of the unified NN parameters:

$$\Delta G^\circ_{37}(\text{total}) = \Delta G^\circ_{37\,\text{initiation}} + \Delta G^\circ_{37\,\text{symmetry}} + \Sigma \Delta G^\circ_{37\,\text{stack}} + \Delta G^\circ_{\text{AT terminal}} \quad 1.$$

$$\begin{aligned} &5'\text{-CGTTGA-}3' = \Delta G^\circ_{37\,\text{initiation}} + \Delta G^\circ_{37\,\text{symmetry}} \\ &3'\text{-GCAACT-}5' \qquad + \underset{\text{GC}}{\text{CG}} + \underset{\text{CA}}{\text{GT}} + \underset{\text{AA}}{\text{TT}} + \underset{\text{AC}}{\text{TG}} + \underset{\text{CT}}{\text{GA}} + \text{AT}_{\text{terminal}} \end{aligned}$$

$$\Delta G^\circ_{37}\,(\text{predicted}) = 1.96 + 0 - 2.17 - 1.44 - 1.00 - 1.45 - 1.30 + 0.05$$

$$\Delta G^\circ_{37}\,(\text{predicted}) = -5.35\ \text{kcal mol}^{-1}.$$

Note that no symmetry penalty is applied because the duplex is nonself-complementary. The ΔH° and ΔS° are calculated analogously with the corresponding parameters in Table 1. Equation 2 is used to predict the ΔG°_T at a different temperature, T:

$$\Delta G^\circ_T = \Delta H^\circ - T\Delta S^\circ, \qquad 2.$$

where T is in Kelvin, ΔH° is in cal mol^{-1}, and ΔS° is in units of cal K^{-1} mol^{-1} (entropy units, e.u.). Note that Equation 2 assumes that ΔCp° is zero, which means that ΔH° and ΔS° are assumed to be temperature independent; this is an excellent approximation for nucleic acids (62, 71). Equations for computation with nonzero ΔCp° are published (62, 71). Note that predictions of ΔG° are most accurate at temperatures near 50°C, even though the ΔG° is traditionally given at 37°C because that is the temperature of the human body, and get worse as the temperature deviates from 50°C (see References 71 and 72 for discussion of error extrapolation). The two-state melting temperature (T_M) may be calculated with Equation 3:

$$T_M = \Delta H^\circ \times 1000/(\Delta S^o + R \times \ln(C_T/x)) - 273.15, \qquad 3.$$

where C_T is the total molar strand concentration, R is the gas constant 1.9872 cal/K-mol, and x equals 4 for nonself-complementary duplexes and equals 1 for self-complementary duplexes. For a nonself-complementary duplex with $\Delta H^\circ = -43.5$ kcal mol^{-1}, $\Delta S^\circ = -122.5$ e.u., and strand concentrations of 0.2 mM for each strand, Equation 3 gives:

$$T_M = -43.5 \times 1000/(-122.5 + 1.9872 \times \ln(0.0004/4)) - 273.15 = 35.8^\circ\text{C}.$$

Note that many duplexes have competing single-strand structure, and this compromises the validity of the two-state approximation and results in systematically

TABLE 1 Nearest-neighbor thermodynamic parameters for DNA Watson-Crick pairs in 1 M NaCl[a]

Propagation sequence	ΔH° (kcal mol^{-1})	ΔS° (e.u.)	ΔG°_{37} (kcal mol^{-1})
AA/TT	−7.6	−21.3	−1.00
AT/TA	−7.2	−20.4	−0.88
TA/AT	−7.2	−21.3	−0.58
CA/GT	−8.5	−22.7	−1.45
GT/CA	−8.4	−22.4	−1.44
CT/GA	−7.8	−21.0	−1.28
GA/CT	−8.2	−22.2	−1.30
CG/GC	−10.6	−27.2	−2.17
GC/CG	−9.8	−24.4	−2.24
GG/CC	−8.0	−19.9	−1.84
Initiation	+0.2	−5.7	+1.96
Terminal AT penalty	+2.2	+6.9	+0.05
Symmetry correction	0.0	−1.4	+0.43

[a]The slash indicates the sequences are given in antiparallel orientation. (e.g., AC/TG means 5′-AC-3′ is Watson-Crick base paired with 3′-TG-5′). The symmetry correction applies to only self-complementary duplexes. The terminal AT penalty is applied for each end of a duplex that has a terminal AT (a duplex with both end closed by AT pairs would have a penalty of +0.1 kcal/mol for ΔG°_{37}).

lower T_Ms than would be predicted by Equation 3. The issue of multistate-coupled equilibria is discussed below.

Figure 2 shows the reliability of the unified parameters for predicting a dataset of 264 sequences ranging in length from 4 to 16 bp. This is a good test of the model because the dataset is much larger than the set of 108 sequences from which the parameters were derived. In addition, the parameters were optimized for prediction of the ΔG°, ΔH°, and ΔS°, not the T_M. The average deviation between experimental and predicted is 1.6°C (corresponding to a standard deviation of 2.3°C). This level of prediction accuracy is sufficient for most applications of nucleic acids, and no other model has yet been devised that performs better. Importantly, the Watson-Crick NN parameters cannot be significantly improved even if a method were to become available for measuring millions of sequences with infinite accuracy. The only way to improve the predictions would be to change the model, for example, to a next-nearest-neighbor model, but we and others (58) have data to suggest that even the NNN model will not improve predictions significantly over the NN model.

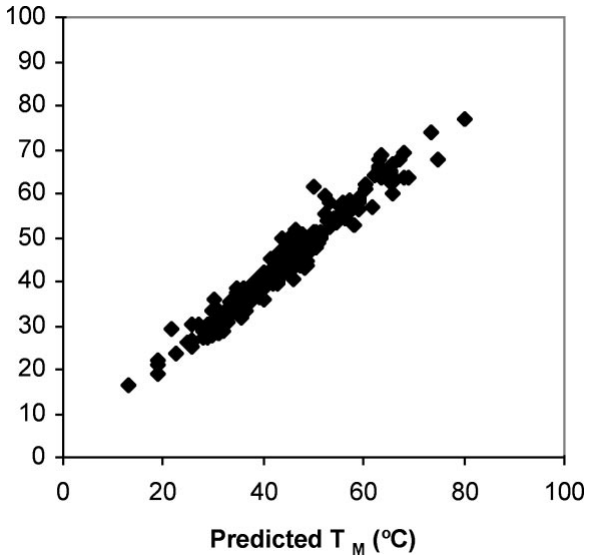

Figure 2 Experimental T_M versus predicted T_M for 264 duplexes of length 4 to 16 bp dissolved in 1 M NaCl. Linear regression gives a slope of 0.96, intercept of 1.58, and $R^2 = 0.96$. The average absolute deviation is 1.6°C.

Several software packages (24, 41, 51, 65, 68) use outdated thermodynamic parameters (12) that give average T_M deviation of 6.8°C (corresponding to a standard deviation of 8.8°C) for the same dataset shown in Figure 2 and perform even worse when extrapolated for different sodium concentrations (see below). Equations that compute T_M using %G + C content (16) work well for polymer duplexes, but perform badly for oligonucleotide duplexes, particularly since these equations do not account for bimolecular initiation and the effect of strand concentration. The accuracy level is important when using the parameters for high-throughput design and for complicated assays that have many interacting oligonucleotides. For example, poor thermodynamic predictions may be tolerated for single target PCR primer design, because even if the predicted T_M is 10°C inaccurate, one has the luxury of experimentally optimizing the annealing temperature. In more complicated assays, such as multiplex PCR, however, all the amplifications must occur under the same conditions and inaccuracies in T_M predictions result in poor primer designs that cause failed amplifications. For a standard deviation of 2°C in T_M, one expects 5% of the sequences will be predicted worse than 4°C, which is still good enough for many applications and usually would not result in complete failure of an assay. In contrast, for an 8°C standard deviation in T_M, one expects 5% of the sequences will be predicted worse than 16°C, which would likely result in complete failure for many assays.

Sodium Dependence

To make the database useful at a variety of solution conditions, empirical salt correction equations have been derived (70) and are given in Equations 4 and 5:

$$\Delta G^\circ_{37}[Na^+] = \Delta G^\circ_{37}[1\,M NaCl] - 0.114 \times N/2 \times \ln[Na^+], \quad 4.$$

$$\Delta S^\circ[Na^+] = \Delta S^\circ[1\,M NaCl] + 0.368 \times N/2 \times \ln[Na^+], \quad 5.$$

where N is the total number of phosphates in the duplex, and $[Na^+]$ is the total concentration of monovalent cations from all sources (the same equation works for sodium, potassium, and ammonium; J. SantaLucia, unpublished experiments). The ΔH° is assumed to be independent of $[Na^+]$, which is valid for nucleic acids for total sodium concentrations above 0.05 M and below 1.1 M. Equations 4 and 5 were derived from measurements on 26 duplexes, where only the single parameter in front of the natural logarithm (i.e., 0.114) was allowed to float. Applying Equation 4 to the duplex given in Equation 1 at 0.1 M NaCl, 10 mM sodium phosphate, pH 7 (gives a total of 0.115 M Na^+ because at pH 7 there are 1.5 equivalents of sodium for each phosphate) gives:

$$\begin{aligned}\Delta G^\circ_{37}\,[0.115\,M\,Na^+] &= -5.35\ \text{kcal mol}^{-1} - 0.114 \times 10/2 \times \ln(0.115)\\ &= -4.12\ \text{kcal mol}^{-1}.\end{aligned}$$

The 6-bp duplex in Equation 1 does not have 5′-terminal phosphates; thus the total number of phosphates in the duplex is 10. To calculate the two-state T_M at the desired $[Na^+]$, the salt-corrected ΔS° from Equation 5 is plugged into Equation 3. Note that the NN parameters themselves (Table 1) may be corrected for salt (70) by setting $N = 1$ in Equations 4 and 5. Equation 4 applies over a range of monovalent concentration of 0.05 to 1 M Na^+ (the same equation works for sodium, potassium, and ammonium; J. SantaLucia, unpublished experiments). The equation begins to break down for duplexes longer than 16 bp. In the section on hairpins below, we describe how to apply the duplex salt dependence for unimolecular transitions. For polymers the coefficient in front of the natural logarithm changes to 0.175, presumably owing to counterion condensation effects (70). A salt dependence function that accounts for all lengths has not yet been derived. In addition, the equation applies only to duplexes that melt in a two-state fashion, which often is not the case for longer duplexes where single-strand folding can compete with duplex formation and where slow dissociation kinetics can inhibit equilibration. This approach provides much more accurate predictions than the previously published empirical equations for directly correcting the T_M (57) for $[Na^+]$. This is because Equations 4 and 5 capture the essential physics of the salt effect, namely, the entropic effects that are due entirely to the geometry of the phosphates. Figure 3 shows the validation set of 81 oligonucleotides in different $[Na^+]$, which provides ample evidence that the salt effects are sequence independent within 2°C in T_M. This set of oligonucleotides is also an excellent test of the NN model itself, since none of the data were used to derive the NN parameters at 1 M NaCl.

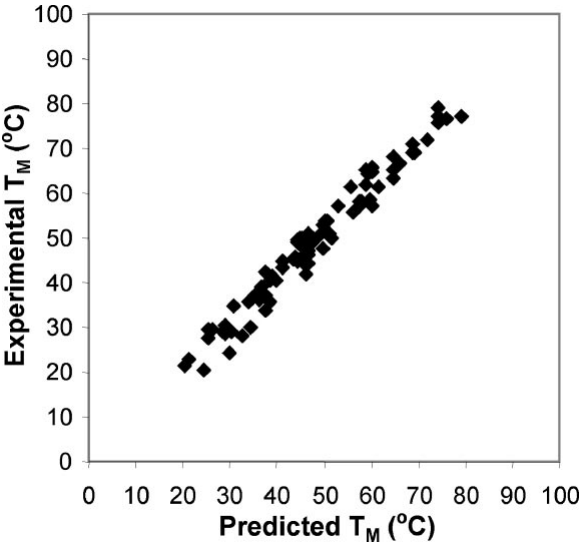

Figure 3 Experimental T_M versus predicted T_M for 81 duplexes 6 to 24 bp in length in solutions ranging from 0.01 to 0.5 M NaCl. Linear regression gives a slope of 1.02, intercept of 0.11, and $R^2 = 0.97$. The average absolute deviation is 2.3°C.

Internal Single Mismatches

The nearest-neighbor model can be extended beyond the Watson-Crick pairs to include parameters for interactions between mismatches and neighboring base pairs (1–4, 25, 64). Table 2 provides the complete thermodynamic database for internal single mismatches, which was derived from UV melting experiments on 174 sequences and solved for 44 unknowns (see References 1–4 and 64 for further explanation of the number of unique unknowns).

An example of the application of the parameters in Table 2 is shown below (underlined residues are mismatched):

$$\begin{array}{l} 5'\text{-GGAC}\underline{\text{T}}\text{GACG-}3' = \text{initiation} + \text{symmetry} + \text{GG} + \text{GA} + \text{AC} \\ 3'\text{-CCTG}\underline{\text{G}}\text{CTGC-}5' \qquad\qquad\qquad\qquad\qquad\quad \text{CC} \quad \text{CT} \quad \text{TG} \\ \qquad\qquad\qquad\qquad + \text{C}\underline{\text{T}} + \text{C}\underline{\text{G}} + \text{GA} + \text{AC} + \text{CG} \\ \qquad\qquad\qquad\qquad\quad \text{G}\underline{\text{G}} \quad \text{G}\underline{\text{T}} \quad \text{CT} \quad \text{TG} \quad \text{GC} \end{array}$$

$$\begin{aligned} \Delta G^\circ_{37}\,(\text{predicted}) &= +1.96 + 0 - 1.84 - 1.30 - 1.44 - 0.32 - 0.47 \\ &\quad - 1.30 - 1.44 - 2.17 \\ &= -8.32 \text{ kcal mol}^{-1}. \end{aligned} \qquad 6.$$

TABLE 2 Nearest-neighbor ΔG°_{37} increments (kcal mol^{-1}) for internal single mismatches next to Watson-Crick pairs in 1 M NaCl[a]

Propagation sequence	X	Y			
		A	C	G	T
GX/CY	A	0.17	0.81	−0.25	WC
	C	0.47	0.79	WC	0.62
	G	−0.52	WC	−1.11	0.08
	T	WC	0.98	−0.59	0.45
CX/GY	A	0.43	0.75	0.03	WC
	C	0.79	0.70	WC	0.62
	G	0.11	WC	−0.11	−0.47
	T	WC	0.40	−0.32	−0.12
AX/TY	A	0.61	0.88	0.14	WC
	C	0.77	1.33	WC	0.64
	G	0.02	WC	−0.13	0.71
	T	WC	0.73	0.07	0.69
TX/AY	A	0.69	0.92	0.42	WC
	C	1.33	1.05	WC	0.97
	G	0.74	WC	0.44	0.43
	T	WC	0.75	0.34	0.68

[a]WC indicates a Watson-Crick pair, which is given in Table 1. Error bars and ΔH° and ΔS° parameters are provided in the original references.

The observed ΔG°_{37} for this sequence is −8.37 kcal mol^{-1} (1). The mismatch NN thermodynamic parameter database is as reliable as the Watson-Crick database and T_M predictions are within 1.6°C, on average. The mismatch parameters in Table 2 have been independently validated for a large set of oligonucleotides (73, 86–88). The salt dependence given in Equations 4 and 5 also apply equally well to mismatches at pH 7 (the salt dependence of A·C and C·C mismatches at low pH may be significantly different). Figure 4 shows graphically the data in Tables 1 and 2 and demonstrates a clear trend in order of decreasing stability: G-C > A-T > G·G > G·T ≥ G· A > T·T ≥ A·A > T·C ≥ A· C ≥ C·C. "G" is the most promiscuous base, since it forms the strongest base pair and the strongest mismatches. On the other hand, "C" is the most discriminating base, since it forms the strongest pair and the three weakest mismatches. In addition, the closing base pair context plays an important role, with closing GC pairs being more favorable than closing AT pairs. The stabilities of the triplets range from −2.22 for G<u>G</u>C/C<u>G</u>G to +2.66 kcal mol^{-1} for A<u>C</u>T/T<u>C</u>A, indicating strong sequence dependence for mismatches. This 4.88 kcal mol^{-1} range corresponds to a factor of over 2700 in equilibrium constant at 37°C—clearly it is not appropriate to neglect the sequence dependence of mismatches. A commonly used heuristic for computing mismatch stability is to decrease the T_M by 1°C for every 1% mismatch in the duplex

regardless of the mismatch type or context (69). This results in huge inaccuracy in the T_M (typically $>10^\circ$C) and thus is not recommended. Also note that with the exception of the terminal and penultimate positions (see below), the thermodynamics of a given mismatch in a given context is independent of its position in a duplex, contrary to common opinion.

Terminal Mismatches

We have completed the database of measurements for terminal mismatches (S. Varma & J. SantaLucia, manuscript in preparation). The results indicate a large dependence on the identity of the mismatch, its orientation, and the closing Watson-Crick pair. The NN stabilities at 37°C range from −1.23 to −0.21 kcal mol^{-1} for C<u>G</u>/G<u>A</u> and A<u>C</u>/T<u>C</u>, respectively (64). Interestingly, all terminal mismatches are stabilizing, whereas internal mismatches may be either stabilizing or destabilizing; presumably, the destabilizing internal mismatches are due to unfavorable helical constraints that prevent the formation of the optimal stacking and H-bond geometry. This difference in trends between internal versus terminal mismatches has an interesting consequence for mismatches at the penultimate and sometimes even the pen-penultimate positions, particularly when the terminal base pair is AT. Consider the following self-complementary duplex structures:

```
                                    A           T
AGTGAGCTCATT    versus      GTGAGCTCAT
TTACTCGAGTGA                TACTCGAGTG
                                T           A
```

Our thermodynamic database predicts the structure on the right, without terminal A-T hydrogen bonding, is approximately 2.5 kcal mol^{-1} more stable than the structure on the left, which has terminal A-T hydrogen bonding. Indeed, NMR studies of this duplex indicate a lack of hydrogen bonding between either the terminal AT pairs or penultimate GT mismatches (64).

Dangling Ends

Table 3 shows the complete thermodynamic database for unpaired 5′- and 3′-dangling ends (9). Unlike A-form RNA duplexes, which show a strong stability preference for 3′ dangling ends over 5′ dangling ends, in B-form DNA there does not appear to be an obvious preference for one end over the other. The average 5′-dangling end contributes −0.45 kcal mol^{-1}, while the average 3′-dangling end contributes −0.29 kcal mol^{-1}. There is a large stability range, however, from +0.48 to −0.96 kcal mol^{-1} for <u>A</u>C/G and <u>G</u>T/A, respectively (compare with Table 3). The few positive dangling end contributions (<u>A</u>C/G and A<u>C</u>/T) are unusual, but were experimentally confirmed (9), and contrast with RNA where all ΔG°_{37} dangling ends are favorable or zero (17, 18). Dangling end contributions are important to account for when a short oligonucleotide hybridizes to a longer target DNA (Figure 1).

TABLE 3 Nearest-neighbor ΔG°_{37} increments (kcal mol^{-1}) for terminal dangling ends next to Watson-Crick pairs in 1 M NaCl[a]

Dangling end sequence	X = A		X = C		X = G		X = T	
	ΔH°	ΔG°_{37}	ΔH°	ΔG°_{37}	ΔH°	ΔG°_{37}	ΔH°	ΔG°_{37}
5′-dangling ends								
XA/T	0.2	−0.51	0.6	−0.42	−1.1	−0.62	−6.9	−0.71
XC/G	−6.3	−0.96	−4.4	−0.52	−5.1	−0.72	−4.0	−0.58
XG/C	−3.7	−0.58	−4.0	−0.34	−3.9	−0.56	−4.9	−0.61
XT/A	−2.9	−0.50	−4.1	−0.02	−4.2	0.48	−0.2	−0.10
3′-dangling ends								
AX/T	−0.5	−0.12	4.7	0.28	−4.1	−0.01	−3.8	0.13
CX/G	−5.9	−0.82	−2.6	−0.31	−3.2	−0.01	−5.2	−0.52
GX/C	−2.1	−0.92	−0.2	−0.23	−3.9	−0.44	−4.4	−0.35
TX/A	−0.7	−0.48	4.4	−0.19	−1.6	−0.50	2.9	−0.29

[a]The slash indicates the sequences are given in antiparallel orientation. (e.g., XA/T means that the A of 5′-XA-3′ is Watson-Crick base paired with T, and X is unpaired). Error bars and ΔS° parameters are provided in the original reference. ΔS° parameters may also be calculated with Equation 3.

Note that in some cases (e.g., AC/G and GA/C) the dangling ends can contribute as much as a full AT base pair to duplex stability; thus neglect of dangling ends can significantly compromise the accuracy of hybridization predictions (9, 15, 23, 76). Some reports have suggested that dangling nucleotides beyond the first nucleotide can contribute to duplex stability (15, 76). Our work, however, indicates that nearly all of the dangling end contribution comes from the first dangling end and the additional nucleotides contribute less than 0.2 kcal mol^{-1}, unless they interfere with hybridization because of the formation of intramolecular hairpin structures. Such long-range dangling end stacking may be important at temperatures below 25°C, but it is unlikely to contribute significantly above 25°C.

Loop Database

Table 4 shows the ΔG°_{37} increments for different lengths of DNA hairpin, bulge, and internal loops, published here for the first time. Application of the loop parameters is different for each motif and thus each is described separately. Unlike for base pairs, mismatches, dangling ends, and terminal mismatches, where an exhaustive determination of all possible sequence variants was performed, for loop motifs the number of possible sequence combinations is enormous and thus a simplifying theory was applied. In general, we have determined or gathered from the literature a large number of measurements on short loops and a few on longer loop lengths.

TABLE 4 ΔG°_{37} increments (kcal mol^{-1}) for length dependence of loop motifs in 1 M NaCl[a]

Loop size[b]	Internal loops[c]	Bulge loops[d]	Hairpin loops[e]
1	—	4.0	—
2	(f)	2.9	—
3	3.2	3.1	3.5
4	3.6	3.2	3.5
5	4.0	3.3	3.3
6	4.4	3.5	4.0
7	4.6	3.7	4.2
8	4.8	3.9	4.3
9	4.9	4.1	4.5
10	4.9	4.3	4.6
12	5.2	4.5	5.0
14	5.4	4.8	5.1
16	5.6	5.0	5.3
18	5.8	5.2	5.5
20	5.9	5.3	5.7
25	6.3	5.6	6.1
30	6.6	5.9	6.3

[a]A dash indicates that the loop length is not allowed. All loop ΔH° parameters are assumed to equal zero. The loop ΔS° increment may be calculated from: $\Delta S^{\circ} = \Delta G^{\circ}_{37} \times 1000/310.15$.

[b]The increments for loop lengths not shown may be calculated with Equation 7 (see text).

[c]For asymmetric internal loops an additional correction must be applied (see text).

[d]For bulge loops with one nucleotide, the intervening base pair stack must be added.

[e]For hairpin loops of length 3 or 4, special sequence dependent triloop and tetraloop corrections must be applied (see supplementary material).

[f]Internal loops of two are calculated using the mismatch nearest neighbor parameters (see Table 2).

A Jacobson-Stockmayer entropy extrapolation is then used to fill in the gaps and provide parameters for closure of long loops (32) according to Equation 7.

$$\Delta G^{\circ}_{37}(\text{loop-n}) = \Delta G^{\circ}_{37}(\text{loop-x}) + 2.44 \times R \times 310.15 \times \ln(n/x), \qquad 7.$$

where ΔG°_{37} (loop-n) is the free energy increment of a loop of length n, ΔG°_{37} (loop-x) is the free-energy increment of the longest loop of length x for which there are experimental data, and R is the gas constant. Note that the coefficient 2.44 is based on recent kinetics measurements in DNA (22), and thus it is used in preference to the older theoretically derived value of 1.75 (48).

Hairpin Loops

Hairpins with lengths of 3 and 4 are treated differently than longer hairpin loops because certain sequences are particularly stable. Importantly, these stable triloop and tetraloop sequences have a significant probability of occurring by random chance, in probes, primers, and targets, and they can significantly inhibit hybridization in various assays. Most software packages to date, however, have not properly accounted for this important effect. Hairpin loops with lengths shorter than 3 are sterically prohibited.

For hairpins of length 3 Equation 8 is applied:

$$\Delta G^\circ_{37}(\text{total}) = \Delta G^\circ_{37}(\text{Hairpin of 3}) + \Delta G^\circ_{37}(\text{triloop bonus}) + \text{closing AT penalty}, \quad 8.$$

where ΔG°_{37} (Hairpin of 3) is $+3.5$ kcal mol^{-1} (Table 4) and the closing AT penalty is $+0.5$ kcal mol^{-1} and is applied only to hairpin sequences that are closed by AT. The ΔG°_{37} (triloop bonus) values are given in the supplementary material (follow the Supplemental Material link from the Annual Reviews home page at http://www.annualreviews.org) and are meant to account for the known special stability of hairpins of the form GNA, where N is any nucleotide (29).

For hairpins of length 4 Equation 9 is applied:

$$\Delta G^\circ_{37}(\text{total}) = \Delta G^\circ_{37}(\text{Hairpin of 4}) + \Delta G^\circ_{37}(\text{triloop bonus}) + \Delta G^\circ_{37}(\text{terminal mismatch}), \quad 9.$$

where ΔG°_{37} (Hairpin of 4) is $+3.5$ kcal mol^{-1} (Table 4) and ΔG°_{37} (terminal mismatch) is the increment for terminal mismatches (S. Varma & J. SantaLucia, unpublished). The tetraloop bonus energies are present to account for known examples of sequences that are exceptionally stable such as GNRA and GNAB (5, 6, 8, 28, 54, 77, 85), where R is a purine and B is C, G, or T. Also included are sequences of length 4 of which good measurements are available. For the sequence C<u>GCAA</u>G, the total hairpin $\Delta G^\circ_{37} = +3.5 - 1.6 - 1.23 = +0.67$ kcal mol^{-1}.

For hairpin loops with lengths longer than 4, Equation 10 is applied:

$$\Delta G^\circ_{37}(\text{total}) = \Delta G^\circ_{37}(\text{Hairpin of N}) + \Delta G^\circ_{37}(\text{terminal mismatch}), \quad 10.$$

where ΔG°_{37} (Hairpin of N) is given in Table 4. To compute the stability of a complete hairpin + stem, one simply adds the salt-corrected base pair NN contributions (Table 1; Equation 3) to the loop energy from Equations 8–10. The thermodynamic contributions of loop nucleotides of a hairpin are assumed to be salt concentration independent. We note that there is room for refinement of this hairpin salt dependence model. The two-state T_M for hairpins is calculated from Equation 11:

$$T_M = \Delta H^\circ \times 1000/\Delta S^\circ - 273.15, \quad 11.$$

where hairpin loop ΔH° and ΔS° are computed with equations analogous to Equations 8–10.

Equations 8–10 have been validated on a series of 61 hairpin sequences of lengths 3 to 8 from the literature (5, 6, 8, 28, 54, 60, 77, 84, 85) and proprietary data (DNA Software Inc.). The results show that for such short hairpins the T_M is predicted within 4°C on average. In a collaboration between DNA Software Inc. and Gorilla Genomics, a series of 859 measurements on 320 molecular beacons with loop lengths from 10 to 35 and stems from 5 to 9 bp were synthesized and melted in 1 to 5 different salt conditions. The results show that the standard deviation between experiments and predicted T_M with Equations 10 and 11 is 3.9°C. This is remarkably good considering that hairpin T_Ms are extremely sensitive to inaccuracies in ΔG°_{37}, and the model for salt and sequence dependence is quite crude yet apparently effective (see Future Directions, below).

Internal Loops

Table 4 gives the length dependence of internal loops. Parameters for loops of lengths 3 to 8 are based on unpublished measurements (J. SantaLucia, unpublished results). Parameters for internal loops longer than 8 were calculated from the Jacobson-Stockmayer equation (previously presented in Equation 7). Like RNA, asymmetric internal loops are significantly less stable than symmetric internal loops of the same length. Thus, an asymmetry penalty is applied in addition to the length penalty given in Table 4. The terminal mismatches in internal loops are assumed to have the same salt dependence as base pairs (Equation 3), whereas the stability of the remainder of the internal loop nucleotides are assumed to be salt independent. Thus, internal loop stability is calculated according to Equation 12:

$$\begin{aligned}\Delta G^\circ_{37}(\text{Loop total}) = {} & \Delta G^\circ_{37}(\text{Internal Loop of N}) + \Delta G^\circ_{37}(\text{asymmetry}) \\ & + \Delta G^\circ_{37}(\text{left terminal mismatch}) \\ & + \Delta G^\circ_{37}(\text{right terminal mismatch}), \qquad 12.\end{aligned}$$

where $\Delta G^\circ_{37}(\text{asymmetry}) = |\text{length A} - \text{length B}| \times 0.3$ kcal mol^{-1} and A and B are the lengths of both sides of the internal loop. The DNA internal loop asymmetry penalty has not yet been fully tested. Note that single mismatches are formally considered symmetric internal loops of 2, but they are calculated using the mismatch NN parameters (Table 2) rather than the sequence-independent approximation that is commonly used in RNA structure predictions (48). For RNA, a huge database of symmetric and mixed tandem mismatches has been measured (91). In DNA, on the other hand, parameters for tandem mismatches are available only for tandem GT (1), and other 2×2 internal loop sequences remain to be determined and thus are approximated by Equation 12.

Bulges

There are few systematic studies of bulges in DNA (34, 39, 82, 90, 93). These studies were used to derive the parameters for lengths 1 to 4 in Table 4. Bulges of longer lengths were calculated using the Jacobson-Stockmayer equation (Equation 7).

Bulges of length 1 are calculated assuming that they are "flipped out," and thus the intervening base pair stack is added (the same approximation is used in RNA) (48):

$$\Delta G^\circ_{37}(\text{Loop total}) = \Delta G^\circ_{37}(\text{Bulge Loop of 1}) + \Delta G^\circ_{37}\ (\text{intervening NN}) + \text{closing AT penalty}, \quad 13.$$

For example, the sequence 5′-C<u>A</u>T-3′ paired with 5′-AG-3′ would be calculated as:

$$\Delta G^\circ_{37}(\text{Loop total}) = +4.0 - 1.28 + 0.5 = +3.22\ \text{kcal mol}^{-1}.$$

For bulge loops longer than 1, Equation 13 is applied, but the intervening NN term is not added. The large destabilizing contribution of bulges means that they are relatively rare in DNA secondary structures of random sequences. Nonetheless, bulges play important roles as intermediates in insertion and deletion mutagenesis and occasionally result in artifacts in genotyping assays, and thus they are included as part of the database for completeness. We have begun a systematic study of the sequence dependence of bulges and find that A bulges are significantly more stable than C-containing bulges, whereas G- and T-containing bulges are intermediate in stability (N. Watkins & J. SantaLucia, unpublished results).

Coaxial Stacking Parameters

Coaxial stacking parameters are important for accurately predicting the stability of multibranched loops (49) for various assays (36, 38, 66) and self-assembling systems (74). We have determined a complete database of thermodynamic parameters of coaxial stacking of helices (63). Coaxial stacking occurs when two oligonucleotides hybridize at adjacent locations on a template or when a probe DNA binds next to a unimolecular hairpin of a template (49, 66). Alternatively, coaxial stacking may be thought of as occurring as the result of a strand "nick" (38). Consider the two structures below:

GGTCGCTCTG	GGTCG–CTCTG
CCAGCGAGAC	CCAGC/GAGAC
$\Delta G^\circ_{37}(\text{Total}) = -12.19$	$\Delta G^\circ_{37}(\text{Total}) = -11.34$

The structure on the left is a normal 10-bp bimolecular duplex, and the structure on the right is a trimolecular coaxially stacked complex of two 5-mers bound to one 10-mer. The "nick" site is indicated by the slash, "/"; the dash "–" indicates the covalently continuous strand. The method for computing the total stability of the coaxially stacked complex is shown in Equation 14:

$$\begin{aligned}\Delta G^\circ_{37}(\text{w/nick}) &= \Delta G^\circ_{37}(\text{without nick}) - \Delta G^\circ_{37}(\text{GC/CG}) \\ &\quad + \Delta G^\circ_{37}(\text{G-C + C/G coaxial}) + \text{extra initiation}, \quad 14. \\ &= -12.19 - (-2.24) + (-3.35) + 1.96 = -11.34\ \text{kcal mol}^{-1},\end{aligned}$$

where -3.35 kcal mol^{-1} is the measured coaxial stacking contribution (63). The extra initiation penalty is required because another bimolecular event must take place to form the coaxially stacked complex. Importantly, the formation of a trimolecular coaxial stacking complex brings up an important concept, namely, that the T_M of such a structure is defined as the temperature at which half of the template strands are simultaneously bound by both probe molecules (if the template is stoichiometrically limiting). Such a reaction is inherently non-two-state and requires the multistate-coupled equilibrium approach described below.

Multibranched Loops

There have been several systematic studies of DNA multibranched loops (33, 37, 40, 42, 44, 45). The stability of multiloops depends on (*a*) the number of helices in the loop, (*b*) the number of unpaired nucleotides in the loop, (*c*) coaxial stacking in the loop, (*d*) terminal mismatch contributions, and (*e*) base composition of the unpaired nucleotides. In RNA, it is commonly assumed that the penalty for multiloops is a linear function of the number of helices and unpaired nucleotides (48). While it has been known for some time (83) that the multiloop length dependence should follow a Jacobson-Stockmayer logarithmic dependence on length, most current dynamic programming algorithms have not been able to accommodate multiloops with a logarithmic dependence (94). Recently, however, DNA Software Inc. developed a novel modification of the dynamic programming algorithm that allows for arbitrary rules for multiloops to be applied [including logarithmic dependence, and the novel length dependences observed in the literature (33)]. Our preliminary multiloop length dependence is given in the supplementary material. The parameters for larger multiloops are calculated with Equation 7. Multiloops remain the least verified parameters in our model for DNA and future work is clearly needed in this area.

QUALITY OF SECONDARY STRUCTURE PREDICTIONS

For RNA, comparative sequence analysis has yielded a huge database of secondary structures that can be used to test the quality of secondary structure prediction algorithms (48). For DNA, however, the database of secondary structures determined by physical means or by comparative sequence analysis is much smaller (7, 10, 11, 14, 19, 27, 30, 31, 35, 43, 61, 78). Table 5 shows the secondary structure prediction results for the currently available DNA database. The results indicate a relatively high degree of accuracy compared to the 73% accuracy currently observed for RNA (48). The high quality of the DNA structure predictions may be an artifact of the small size of the database. Alternatively, DNA secondary structures may be inherently easier to predict than RNA because of fewer interactions with proteins and because of fewer tertiary interactions. Prediction of the correct secondary structure is not the only goal; the accurate prediction of the energy required to unfold a portion of a long DNA so that an oligonucleotide can bind is also important.

TABLE 5 Accuracy of DNA secondary structure predictions for the optimal structure from OMP

Molecule name	Length (nts)	Predicted bp/total bp	Percent predicted	Reference(s)
msDNA-Sa163	163	55/55	100	(19)
tDNAPhe	76	16/21	76	(35, 43)
tDNAHis	118	20/20	100	(30)
tDNALys	76	20/20	100	(35)
tDNAMet	75	21/21	100	(61)
antitDNAMet	75	21/21	100	(61)
67-mer	67	11/17	69	(61)
RNase	62	7/15	47	(11)
Ligase	80	30/31	97	(14)
M13Gori1	334	87/87	100	(31)
F1	670	83/90	92	(27)
M13	450	84/90	93	(78)
parvovirus 3′-end	125	41/51	80	(7)
Total		496/539	92	

MULTISTATE MODELING OF DNA FOLDING AND HYBRIDIZATION

The parameter database presented in this review was derived from model sequences that were rationally designed to melt in a two-state fashion (see References 71 and 72 for design principles). On the other hand, "real" single-stranded target DNA sequences are folded molecules, and this folding must be broken before a primer or probe oligonucleotide can bind. Such folding in the target or probe DNAs inhibits hybridization and causes false-negative assays. Alternatively, mismatch hybridization can give undesired signal that results in false-positive assays. The result of these observations is that computations with a two-state model can be misleading. Further, many researchers focus on the two-state T_M parameter as determining the success of their assay. However, we would like to encourage molecular biologists to change their attention to what really matters in their assays, namely, how much of a target is correctly bound by an oligonucleotide (sensitivity) and how little signal results from undesired hybridizations (selectivity or specificity). To answer such questions requires numerical solution of the multistate-coupled equilibrium equations for the concentrations of all the species in the solution at any desired temperature or salt condition, as shown in Figure 5. Previously, simple multistate equilibrium equations involving competition of folded single strands and bimolecular duplexes have been solved analytically (1, 9a, 47). To account for large numbers of competing unimolecular and bimolecuar reactions requires

numerical simuation. The first use of such a numerical simulation to solve coupled equilibria of nucleic acids was used to simulate the competition between a nonself-complementary heteroduplex and the self-complementary duplex formed by one of the strands (43a). Dr. Nicolas Peyret has described a generalization of the numerical simulation methods in his PhD thesis (63). DNA Software Inc. has further generalized the numerical approach in the software package Visual OMP so that systems of hundreds of competing species may have their equilibrium concentrations calculated as a function of temperature. Such an approach is possible now that the individual equilibrium constants can be accurately predicted as described in this review. An important concept is that of "net T_M"—the temperature at which half of a template is bound by probe—which must be calculated taking the competitive equilibria into account in a multistate model (63). An important application of the concept of net T_M is for molecular beacons. For a molecular beacon, a simple duplex T_M does not accurately reflect what is measured in a normal experiment due to the fact that probe signal generation is the result of competition between probe folding and probe-target duplex formation. Thus, a beacon's experimental T_M should actually be considered as a net T_M. Table 6 presents the net T_M predictions for four molecular beacons. These are stringent tests of our methodology, since they require accurate modeling of base pairs, salt dependence, mismatches, hairpins, and solution of the coupled equilibria. The results show that our multistate model is based on firm scientific principles; in contrast, the two-state model fails dramatically to predict complex assays. The accuracy of numerical computations also opens the possibility of in silico simulation and optimization of various molecular biology assays.

FUTURE DIRECTIONS

There are several avenues of investigation necessary to further improve the quality of the database of parameters for DNA secondary-structure prediction. First, there is a need to refine the thermodynamic parameters that characterize the various structural motifs. Although length-based approximations have yielded improved predictions, sequence-dependent rules and parameters for hairpins, bulges, internal loops, and multibranch loops are needed. Many molecular biology assays contain enzymes that require the use of magnesium in the buffer, which significantly affects DNA hybridization. Magnesium and calcium are also present in vivo, and these ions affect in situ hybridization applications. Thus, there is a need to develop empirical corrections for mixtures of sodium and magnesium and to incorporate these into software that can be used by nonexperts. Other areas that require further study include modified nucleotides, effects of terminal and internal fluorophores, and effects of added denaturants such as DMSO, formamide, glycerol, and urea. Systematic studies of the topics discussed above are currently underway in the SantaLucia laboratory at Wayne State University and at DNA Software, Inc. There is also a need to expand the database of known complex DNA secondary structures determined experimentally or by comparative sequence analysis so that we may better test the quality of our predictions.

TABLE 6 Experiments vs OMP predictions for molecular beacon duplex hybridization[a]

Target sequence	Central pair[b] X–Y	ΔG°_{37} (kcal mol^{-1})[c]			T_M (°C)		
		Expt.	Pred. multistate	Pred. two-state	Expt.	Pred. multistate	Pred. two-state
GGTTTTTTTTTTTGG	A-T	−10.49	−10.69	−13.24	42	43.4	47.9
GGTTTTTATTTTTGG	A-A	−6.66	−7.39	−9.94	27	30.1	39.1
GGTTTTTCTTTTTGG	A-C	−6.72	−6.48	−9.03	23	24.8	36.0
GGTTTTTGTTTTTGG	A-G	−7.62	−7.81	−10.36	28	32.3	40.4

[a]Data are from Reference 9a. The molecular beacon used for all experiments was F-CGCTCCCAAAAAAAAAAACCGAGCG-Q, where F is the fluorophore and Q is the quencher and the underlined residues are in the hairpin loop and bind to the target shown in the table. The experiments were performed with 0.105 M NaCl, 1.0 mM $MgCl_2$, [beacon] $= 5 \times 10^{-8}$ M, [target] $= 3 \times 10^{-7}$ M.

[b]X and Y correspond to the central "A" in the beacon, and the central nucleotide in the target (underlined) corresponds to "X" in the X-Y pair.

[c]Expected (Expt.) columns are the experimental results from Reference 9a. Multistate columns are the OMP predictions using the multistate-coupled equilibrium model (see text). Two-state columns are the OMP prediction using the two-state model.

Every day, thermodynamic calculations take on more and more importance in molecular biology applications. There is a need to incorporate thermodynamics principles and assay-specific heuristics into optimization algorithms so that assays may be automatically designed to improve their reliability. For example, a simple PCR reaction is quite difficult to simulate in detail, and thermodynamics of primer hybridization and competition with template folding and primer dimerization are important factors. However, the enzyme in the reaction also plays an essential role and thus known heuristics include avoiding runs of guanines, and ensuring that proper annealing specificity of the 3′ end of the primer is more important than the 5′ end. As described above for molecular beacons, simulating the thermodynamic competition is essential for accurately simulating their behavior. An additional concept is to consider how various design considerations (e.g., stability of the hairpin stem versus stability of the hybridized duplex versus desired specificity for different alleles) should be weighed in an overall calculated "figure of merit" for the simulated behavior of a molecular beacon. Incorporation of such a figure of merit with an algorithm for trying different sequences and solution conditions could be used to make an algorithm that would automatically design assays with optimal performance, thereby saving significant time and money for the development of new assays. Full simulation and optimization of nucleic acid–based assays would benefit significantly by borrowing concepts from the operations research community and integrating these concepts with concepts from the molecular biology and biophysical chemistry communities. This approach has been taken with the commercial software Visual OMP from DNA Software, Inc.

Microarray applications rely on DNA hybridization as the phenomenon underlying their specificity and sensitivity. Clearly, improved thermodynamic calculations will result in greater precision in the hypotheses tested on this powerful high-throughput platform. Initial investigations have indicated that solution parameters are relevant, but not fully predictive of hybridization on a surface (26, 92). Systematic studies are necessary to parameterize new models of hybridization on microarray surfaces, and new algorithms are needed to numerically simulate the hybridization process with proper accounting of surface electrostatic and steric effects, oligomer synthesis quality, and competition with the unimolecular folding that occurs in bulk solution above the surface.

Recently, there have appeared several reports (22, 53, 89) of studies on hybridization and hairpin folding kinetics, which raises the possibility of simulating DNA folding processes in the time domain. Finally, now that the accurate prediction of DNA secondary structure appears to be within sight, we should begin to focus our efforts on the second half of the nucleic acid folding problem, namely, three-dimensional structure prediction, which ought to keep us busy for at least the near future.

ACKNOWLEDGMENTS

We gratefully acknowledge NIH grant HG02020 and Michigan Life Sciences Corridor Grant LSC1653 to JSL and NIH SBIR grant R44 HG/GM02555 to DNA Software Inc. We also acknowledge the years of work done by previous and present

graduate students who made this review possible, namely, Hatim Allawi, Rostem Irani, Svetlana Morosyuk, Nicolas Peyret, Shikha Varma, and Norm Watkins. We thank Nana Lee, Svetlana Morosyuk, and Bob Royce for preparing figures. We thank Michael Zuker for his fruitful collaboration on making the DNA-MFOLD server. Finally, we thank Ignacio Tinoco Jr. and Douglas H. Turner for their enduring contributions to the field of RNA folding, which inspired much of the work presented here.

The *Annual Review of Biophysics and Biomolecular Structure* is online at http://biophys.annualreviews.org

LITERATURE CITED

1. Allawi HT, SantaLucia J Jr. 1997. Thermodynamics and NMR of internal G−T mismatches in DNA. *Biochemistry* 36:10581–94
2. Allawi HT, SantaLucia J Jr. 1998. Nearest-neighbor thermodynamics of internal A-C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* 37:9435–44
3. Allawi HT, SantaLucia J Jr. 1998. Nearest-neighbor thermodynamics parameters for internal GA mismatches in DNA. *Biochemistry* 37:2170–79
4. Allawi HT, SantaLucia J Jr. 1998. Thermodynamics of internal C-T mismatches in DNA. *Nucleic Acids Res.* 26:2694–701
5. Antao VP, Lai SY, Tinoco I Jr. 1991. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* 19:5901–5
6. Antao VP, Tinoco I Jr. 1992. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.* 20:819–24
7. Astell CR, Chow MB, Ward DC. 1985. Sequence analysis of the termini of virion and replicative forms of minute virus of mice DNA suggests a modified rolling hairpin model for autonomous parvovirus DNA replication. *J. Virol.* 54:171–77
8. Blommers MJJ, Walters ALI, Haasnoot CAG, Aelen JMA, van der Marel GA, et al. 1989. Effects of base sequence on the loop folding in DNA hairpins. *Biochemistry* 28:7491–98
9. Bommarito S, Peyret N, SantaLucia J Jr. 2000. Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic Acids Res.* 28:1929–34

9a. Bonnet G, Tyagi S, Libchaber A, Kramer FR. 1999. Thermodynamic basis of the enhanced specificity of structured DNA probes. *Proc. Natl. Acad. Sci. USA* 96: 6171–76

10. Breaker RR. 1997. DNA aptamers and DNA enzymes. *Curr. Opin. Chem. Biol.* 1:26–31
11. Breaker RR, Joyce FG. 1994. A DNA enzyme that cleaves RNA. *Chem. Biol.* 1:223–28
12. Breslauer KJ, Frank R, Blocker H, Marky LA. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83:3746–50
13. Buiser RG, DeStefano JJ, Mallaber LM, Fay PJ, Bambara RA. 1991. Requirements for the catalysis of strand transfer synthesis by retroviral DNA polymerases. *J. Biol. Chem.* 266:13103–9
14. Cuenoud B, Szostak JW. 1995. A DNA metalloenzyme with DNA ligase activity. *Nature* 375:611–19
15. Doktycz MJ, Paner TM, Amaratunga M, Benight AS. 1990. Thermodynamic stability of the 5′ dangling-ended DNA hairpins formed from sequences 5′-$(XY)_2$

GGATAC(T)$_4$GTATCC-3′, where X, Y = A, T, G, C. *Biopolymers* 30:829–45
16. Frank-Kamenetskii MD. 1971. Simplification of the empirical relationship between melting temperature of DNA, its GC content and concentration on sodium ions in solution. *Biopolymers* 10:2623–25
17. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* 83:9373–77
18. Freier SM, Sugimoto N, Sinclair A, Alkema D, Neilson T, et al. 1986. Stability of XGCGCp, GCGCYp, and XGCGCYp helices: an empirical estimate of the energetics of hydrogen bonds in nucleic acids. *Biochemistry* 25:3214–19
19. Furuichi T, Inouye S, Inouye M. 1987. Biosynthesis and structure of stable branched RNA covalently linked to the 5′ end of multicopy single-stranded DNA of *Stigmatella aurantiaca*. *Cell* 48:55–62
20. Gacy AM, Goellner G, Juranic N, Macura S, McMurray CT. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* 81:533–40
21. Gibbons A, Amos M, Hodgson D. 1997. DNA computing. *Curr. Opin. Biotechnol.* 8:103–6
22. Goddard NL, Bonnet G, Krichevsky O, Libchaber A. 2000. Sequence dependent rigidity of single stranded DNA. *Phys. Rev. Lett.* 85:2400–3
23. Guckian KM, Schweitzer BA, Ren RX-F, Sheils CJ, Paris PL, et al. 1996. Experimental measurement of aromatic stacking affinities in the context of duplex DNA. *J. Am. Chem. Soc.* 118:8182–83
24. Haas S, Vingron M, Poustka A, Wiemann S. 1998. Primer design for large scale sequencing. *Nucleic Acids Res.* 26:3006–12
25. He L, Kierzek R, SantaLucia J Jr, Walter AE, Turner DH. 1991. Nearest neighbor parameters for GU mismatches: GU/UG is destabilizing in the contexts CGUG, UGUA, and AGUU, but stabilizing in GGUC. *Biochemistry* 30:11124–32
26. Held GA, Grinstein G, Tu Y. 2003. Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci. USA* 100:7575–80
27. Higashitani N, Higashitani A, Horiuchi K. 1993. Nucleotide sequence of the primer RNA for DNA replication of filamentous bacteriophages. *J. Virol.* 67:2175–81
28. Hilbers CW, Haasnoot CAG, de Bruin SH, Joordens JJM, van der Marel GA, van Boom JH. 1985. Hairpin formation in synthetic oligonucleotides. *Biochimie* 67:685–95
29. Hirao I, Nishimura Y, Tagawa Y, Watanabe K, Miura K. 1992. Extraordinary stable mini-hairpins: electrophoretical and thermal properties of the various sequence variants of d(GCGAAAGC) and their effect on DNA sequencing. *Nucleic Acid Res.* 20:3891–96
30. Holmes CE, Hecht SM. 1993. Febleomycin cleaves a transfer RNA precursors and its “transfer DNA” analog at the same major site. *J. Biol. Chem.* 268:25909–13
31. Ikoku AS, Hearst JE. 1981. Identification of a structural hairpin in the filamentous chimeric phage M13Gori1. *J. Mol. Biol.* 151:245–59
32. Jacobson H, Stockmayer WH. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* 18:1600–6
33. Kadrmas JL, Ravin AJ, Leontis NB. 1995. Relative stabilities of DNA three-way, four-way and five-way junctions (multi-helix junction loops): Unpaired nucleotides can be stabilizing or destabilizing. *Nucleic Acids Res.* 23:2212–22
34. Ke S-H, Wartell RM. 1995. Influence of neighboring base pairs on the stability of single base bulges and base pairs in a DNA fragment. *Biochemistry* 34:4593–600
35. Khan AS, Roe BA. 1988. Aminoacylation of synthetic DNAs corresponding to *Escherichia coli* phenylalanine and lysine tRNAs. *Science* 241:74–79

36. Kieleczawa J, Dunn JJ, Studier FW. 1992. DNA sequencing by primer walking with strings of contiguous hexamers. *Science* 258:1787–91
37. Ladbury JE, Sturtevant JM, Leontis NB. 1994. The thermodynamics of formation of three-strand, DNA three-way junction complex. *Biochemistry* 33:6828–33
38. Lane MJ, Paner T, Kashin I, Faldasz BD, Li B, et al. 1997. The thermodynamic advantage of DNA oligonucleotide 'stacking hybridization' reactions: energetics of a DNA nick. *Nucleic Acids Res.* 25:611–17
39. LeBlanc DA, Morden KM. 1991. Thermodynamic characterization of deoxyribooligonucleotide duplexes containing bulges. *Biochemistry* 30:4042–47
40. Leontis NB, Kwok W, Newman JS. 1991. Stability and structure of three-way DNA junctions containing unpaired nucleotides. *Nucleic Acids Res.* 19:759–66
41. Li P, Kupfer KC, Davies CJ, Burbee D, Evans GA, Garner HR. 1997. PRIMO: a primer design program that applies base quality statistics for automated large-scale DNA sequencing. *Genomics* 40:476–85
42. Lilley DMJ, Hallam LR. 1984. Thermodynamics of the ColE1 cruciform: comparison between probing and topological experiments using single topoisomers. *J. Mol. Biol.* 180:179–200
43. Lim AC, Barton JK. 1993. Chemical probing of tDNA with transition metal complexes: a structural comparison of RNA and DNA. *Biochemistry* 32:11029–34
43a. Longfellow CE, Kierzek R, Turner DH. 1990. Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry* 29:278–85
44. Lu M, Guo Q, Kallenbach NR. 1991. Effect of sequence on the structure of three-arm DNA junctions. *Biochemistry* 30:5815–20
45. Lu M, Guo Q, Marky LA, Seeman NC, Kallenbach NR. 1991. Thermodynamics of DNA branching. *J. Mol. Biol.* 223:781–89
46. Luo G, Taylor J. 1990. Template switching by reverse transcriptase during DNA synthesis. *J. Virol.* 64:4321–28
47. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH. 1999. Predicting oligonucleotide affinity to nucleic acid targets. *RNA* 5:1458–69
48. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–40
49. Mathews DH, Turner DH. 2002. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry* 41:869–80
50. McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–19
51. McKay SJ, Jones SJ. 2002. AcePrimer: automation of PCR primer design based on gene structure. *Bioinformatics* 18:1538–39
52. Morozov SY, Chernov BK, Merits A, Blinov VM. 1994. Computer-assisted predictions of the secondary structure in the plant virus single-stranded DNA genome. *J. Biomol. Struct. Dyn.* 11:837–47
53. Muragan R. 2002. Revised theory of DNA renaturation kinetics and its experimental verification. *Biochem. Biophys. Res. Commun.* 293:870–73
54. Nakano M, Moody EM, Liang J, Bevilacqua PC. 2002. Selection for thermodynamically stable DNA tetraloops using temperature gradient gel electrophoresis reveals four motifs: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc). *Biochemistry* 41:14281–92
55. Nazarenko I, Pires R, Lowe B, Obaidy M, Rashtchian A. 2002. Effect of primary and secondary structure of oligodeoxyribonucleotides on the fluorescent properties of conjugated dyes. *Nucleic Acids Res.* 30:2089–195

56. Okumoto Y, Ohmichi T, Sugimoto N. 2002. Immobilized small deoxyribozyme to distinguish RNA secondary structures. *Biochemistry* 41:2769–73
57. Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS. 1997. Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers* 44:217–39
58. Owczarzy R, Vallone PM, Goldstein RF, Benight AS. 1999. Studies of DNA dumbbells. VII. Evaluation of the next-nearest neighbor sequence dependent interactions in duplex DNA. *Biopolymers* 52:29–56
59. Ozawa K, Kurtzman G, Young N. 1986. Replication of the B19 parvovirus in human bone marrow cell cultures. *Science* 233:883–86
60. Paner TM, Amaratunga M, Doktycz MJ, Benight AS. 1990. Analysis of melting transitions of the DNA hairpins formed from the oligomer sequence d[GGATAC(X)4GTATCC] (X = A, T, G, C). *Biopolymers* 29:1715–34
61. Paquette J, Nicoghosian K, Qi G, Beauchemin N, Cedergren R. 1990. The conformation of single-stranded nucleic acids tDNA versus tRNA. *Eur. J. Biochem.* 189:259–65
62. Petersheim M, Turner DH. 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: 256–63
63. Peyret N. 2000. *Prediction of nucleic acid hybridization: parameters and algorithms*. PhD thesis. Wayne State University, Detroit, MI. 352 pp.
64. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J Jr. 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A-A, C-C, G-G, and T-T mismatches. *Biochemistry* 38:3468–77
65. Proutski V, Holmes EC. 1996. PrimerMaster: a new program for the design and analysis of PCR primers. *CABIOS* 12:253–55
66. Riccelli PV, Merante F, Leung KT, Bortolin S, Zastawny RL, et al. 2001. Hybridization of single-stranded DNA targets to immobilized complementary DNA probes: comparison of hairpin versus linear capture probes. *Nucleic Acids Res.* 29:996–1004
67. Russel M. 1994. Mutants at conserved positions in gene IV, a gene required for assembly and secretion of filamentous phages. *Mol. Microbiol.* 14:357–69
68. Rychlik W, Rhoads ER. 1989. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing, and in vitro amplification of DNA. *Nucleic Acids Res.* 17:8543–51
69. Sambrook J, Fritsch EF, Maniatis T. 1989. Site-directed mutagenesis of cloned DNA, In *Molecular Cloning: A Laboratory Manual*, pp. 15.51–80. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press
70. SantaLucia J Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95:1460–65
71. SantaLucia J Jr. 2000. The use of spectroscopic techniques in the study of DNA stability. In *Spectrophotometry and Spectrofluorimetry: A Practical Approach*, ed. MG Gore, pp. 329–56. Oxford, UK: Oxford Univ. Press
72. SantaLucia J Jr, Turner DH. 1997. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* 44:309–19
73. Schutz E, von Ahsen N. 1999. Spreadsheet software for thermodynamic melting point prediction of oligonucleotide hybridization with and without mismatches. *Biotechniques* 27:1218–22
74. Seeman NC. 1999. DNA engineering and its application to nanotechnology. *Trends Biotechnol.* 17:437–43
75. Seeman NC. 2003. Biochemistry and structural DNA nanotechnology: an

evolving symbiotic relationship. *Biochemistry* 42:7259–69
76. Senior M, Jones RA, Breslauer KJ. 1988. The influence of dangling thymidine residues on the stability and structure of two DNA duplexes. *Biochemistry* 27:3879–85
77. Senior MM, Jones RA, Breslauer KJ. 1988. Influence of loop residues on the relative stabilities of DNA hairpin structures. *Proc. Natl. Acad. Sci. USA* 85:6242–46
78. Specthrie L, Bullitt E, Horiuchi K, Model P, Russel M, Makowski L. 1992. Construction of a microphage variant of filamentous bacteriophage. *J. Mol. Biol.* 228:720–24
79. Stull RA, Taylor LA, Szoka FC Jr. 1992. Predicting antisense oligonucleotide inhibitory efficacy: a computational approach using histograms and thermodynamic indices. *Nucleic Acids Res.* 20: 3501–8
80. Summers J, Mason WS. 1982. Replication of the genome of a hepatitis B-like virus by reverse transcription of an RNA intermediate. *Cell* 29:403–15
81. Tinoco I, Bustamante C. 1999. How RNA folds. *J. Mol. Biol.* 293:271–81
82. Turner DH. 1992. Bulges in nucleic acids. *Curr. Opin. Struct. Biol.* 2:334–37
83. Turner DH, Sugimoto N, Freier SM. 1988. RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.* 17:167–92
84. Vallone PM, Paner T, Hilario J, Lane MJ, Faldasz BD, Benight AS. 1999. Melting studies of short DNA hairpins: influence of loop sequence and adjoining base pair identity on hairpin thermodynamic stability. *Biopolymers* 50:425–42
85. Varani G. 1995. Exceptionally stable nucleic acid hairpins. *Annu. Rev. Biophys. Biomol. Struct.* 24:379–404
86. von Ahsen N, Oellerich M, Armstrong VW, Schutz E. 1999. Application of a thermodynamic nearest-neighbor model to estimate nucleic acid stability and optimize probe design: prediction of melting points of multiple mutations of apolipoprotein B-3500 and factor V with a hybridization probe genotyping assay on the LightCycler. *Clin. Chem.* 45:2094–101
87. von Ahsen N, Oellerich M, Schutz E. 2000. DNA base bulge vs. unmatched end formation in probe-based diagnostic insertion/deletion genotyping: genotyping the UGTIAI (TA)(n) polymorphism by real-time fluorescence PCR. *Clin. Chem.* 46:1939–45
88. von Ahsen N, Wittwer CT, Schutz E. 2001. Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg^{2+}, deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin. Chem.* 47:1956–61
89. Wang J-Y, Drlica K. 2003. Modeling hybridization kinetics. *Math. Biosci.* 183:37–47
90. White SA, Draper DE. 1989. Effects of single-base bulges on intercalator binding to small RNA and DNA hairpins and a ribosomal RNA fragment. *Biochemistry* 28:1892–97
91. Wu M, McDowell JA, Turner DH. 1995. A periodic table of tandem mismatches. *Biochemistry* 34:3204–11
92. Zhang L, Miles MF, Aldape KD. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.* 21:818–21
93. Zieba K, Chu TM, Kupke DW, Marky LA. 1991. Differential hydration of dA-dT base pairing and dA and dT bulges in deoxyoligonucleotides. *Biochemistry* 30:8018–26
94. Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52
95. Zuker M, Sankoff D. 1984. RNA secondary structures and their prediction. *Bull. Math. Biol.* 46:591–621

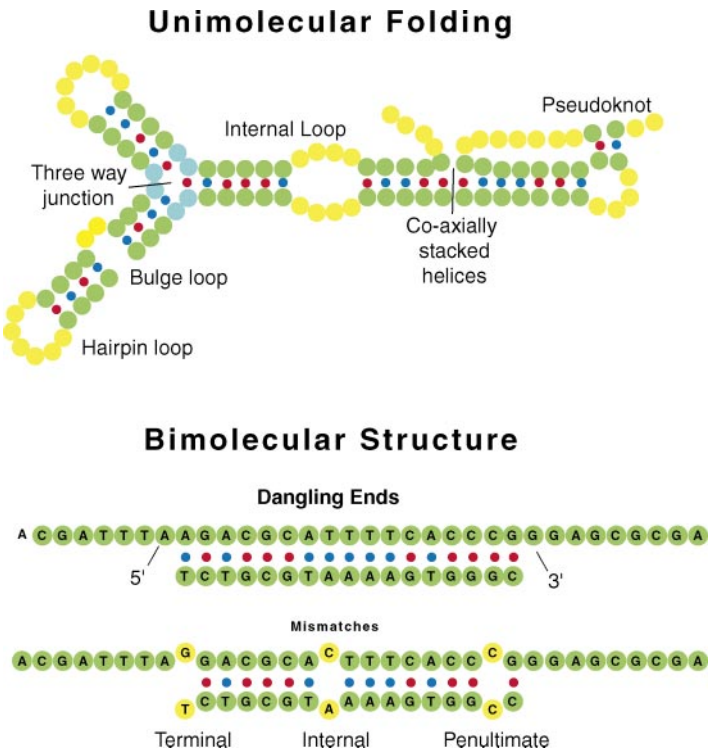

Figure 1 DNA structural motifs.

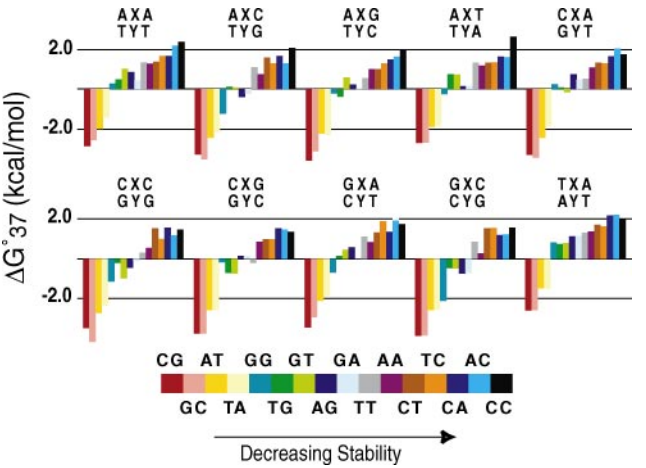

Figure 4 Thermodynamic stabilities for all possible X-Y pairs in all 10 different triplet contexts closed by Watson-Crick pairs. The figure was generated by adding the appropriate NN in Tables 1 and 2.

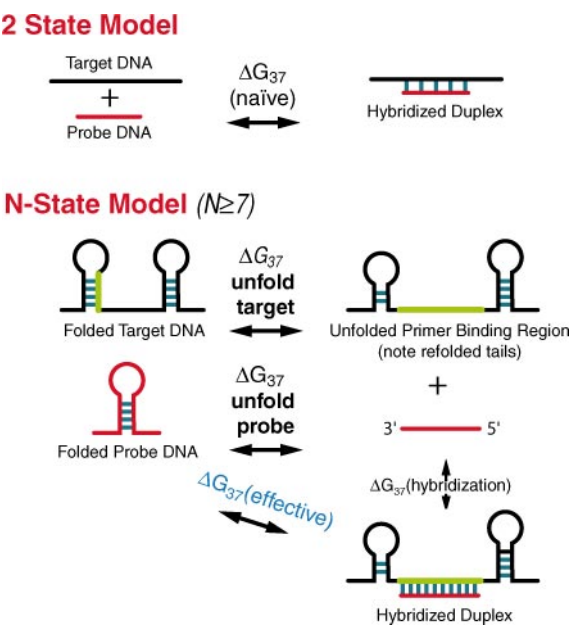

Figure 5 Multistate-coupled equilibrium model for DNA hybridization. Note that OMP is capable of including many other species including suboptimal structures and mismatch hybridizations.