Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure

David H. Mathews[†], Matthew D. Disney[†‡], Jessica L. Childs[†‡], Susan J. Schroeder[‡], Michael Zuker[§], and Douglas H. Turner[†‡¶]

[†]Center for Human Genetics and Molecular Pediatric Disease, The Aab Institute of Biomedical Sciences, University of Rochester School of Medicine and Dentistry, 601 Elmwood Avenue, Box 703, Rochester, NY 14642; [‡]Department of Chemistry, University of Rochester, RC Box 270216, Rochester, NY 14627-0216; and [§]Department of Mathematics, 331 Amos Eaton Hall, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180-3590

Communicated by Ignacio Tinoco, Jr., University of California, Berkeley, CA, March 15, 2004 (received for review December 1, 2003)

A dynamic programming algorithm for prediction of RNA secondary structure has been revised to accommodate folding constraints determined by chemical modification and to include free energy increments for coaxial stacking of helices when they are either adjacent or separated by a single mismatch. Furthermore, free energy parameters are revised to account for recent experimental results for terminal mismatches and hairpin, bulge, internal, and multibranch loops. To demonstrate the applicability of this method, *in vivo* modification was performed on 5S rRNA in both *Escherichia coli* and *Candida albicans* with 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate, dimethyl sulfate, and kethoxal. The percentage of known base pairs in the predicted structure increased from 26.3% to 86.8% for the *E. coli* sequence by using modification constraints. For *C. albicans*, the accuracy remained 87.5% both with and without modification data. On average, for these sequences and a set of 14 sequences with known secondary structure and chemical modification data taken from the literature, accuracy improves from 67% to 76%. This enhancement primarily reflects improvement for three sequences that are predicted with <40% accuracy on the basis of energetics alone. For these sequences, inclusion of chemical modification constraints improves the average accuracy from 28% to 78%. For the 11 sequences with <6% pseudoknotted base pairs, structures predicted with constraints from chemical modification contain on average 84% of known canonical base pairs.

Recent discoveries have shown that RNA plays a larger role in biology than previously realized, e.g., in posttranscriptional regulation (1), development (2, 3), immunity (4, 5), and peptide bond formation (6, 7). It is necessary to determine the native structures of RNAs to understand their mechanisms of action, and determining secondary structure is a crucial step in this process.

RNA secondary structure can be predicted by free energy minimization with nearest neighbor parameters to evaluate stability (8–18). Previous studies demonstrated that nuclease cleavage data can be used to refine structure prediction and improve accuracy (8, 11). A predicted secondary structure can guide further experiments or comparative sequence analysis (19) and also aid in the design of RNA molecules (20, 21).

Chemical modification is a technique that reveals solvent accessible nucleotides (22). The nucleotides accessible to 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate, dimethyl sulfate, and kethoxal are unpaired, in A-U or G-C pairs at helix ends, in G-U pairs anywhere, or adjacent to G-U pairs. This limited specificity differs from that observed with nucleases, and an algorithm allowing constraints from such chemical modification has not been reported. Chemical modification is used extensively to test hypothesized RNA secondary structures (19, 23–28). Chemical modification can also be used to deduce possible tertiary contacts within an RNA (29), to probe RNA bound to protein (25, 26, 30–35), or to follow RNA folding pathways (36–38). The method can map RNA *in vivo* (39–43), which is not possible with nuclease mapping. This is an important advantage because much is not known about renaturing purified RNA into its native conformation.

In this study, a dynamic programming algorithm for prediction of RNA secondary structure has been revised to use experimentally determined chemical modification constraints. These constraints dramatically improve the accuracy of structure prediction when free energy minimization alone predicts <40% of known base pairs. The nearest-neighbor parameters for free energy are also revised on the basis of recent experiments, and the program RNASTRUCTURE now includes terms for the free energy of coaxial stacking of helices that are either adjacent or separated by a single mismatch in multibranch and exterior loops.

Methods

Nearest-Neighbor Parameters. Thermodynamic parameters are based on the set of Xia *et al*. (44–46) and Mathews *et al*. (8). Hairpin loop parameters (Tables 1 and 2) are revised on the basis of recent experimental results (47, 48) and the previous database of RNA hairpin stabilities (49–55).

Thermodynamic parameters for bulge loops of single nucleotides are revised on the basis of measurements by Znosko *et al.* (56) by using the model

$$\Delta G^\circ_{37\ \text{bulge}}(n=1) = \Delta G^\circ_{37\ \text{bulge initiation}}(n) + \Delta G^\circ_{37}(\text{special C bulge}) + \Delta G^\circ_{37\ \text{bp stack}} - RT\ \ln(\text{number of states}),$$

where the number of states is the number of secondary structures containing a bulge of identical sequence in slightly different positions because of bulge migration, such as observed by NMR (57). For example, an isoenergetic bulged C in 5′UGU/3′ACCA can occur in two positions. ΔG°_{37}(special C bulge), -0.9 ± 0.3 kcal (1 cal = 4.18 J)/mol, is an empirical bonus applied to bulged C residues adjacent to at least one C. The $\Delta G^\circ_{37\ \text{bulge initiation}}$ for single nucleotide bulges is 3.81 ± 0.08 kcal/mol.

Internal loop free energy parameters are revised on the basis of recent measurements (58–61) and the previously assembled database (8, 45, 46, 62–69). In the program described here, measured values are used when available for 1×1, 1×2, and

Abbreviation: RT, reverse transcription.

¶To whom correspondence should be addressed. E-mail: turner@chem.rochester.edu.

Table 1. Free energy parameters for hairpin loop formation

Parameter (number of nt or sequence)	ΔG°_{37}, kcal/mol
$\Delta G^\circ_{37\ \text{initiation}}(3)$	5.4 ± 0.2
$\Delta G^\circ_{37\ \text{initiation}}(4)$	5.6 ± 0.1
$\Delta G^\circ_{37\ \text{initiation}}(5)$	5.7 ± 0.2
$\Delta G^\circ_{37\ \text{initiation}}(6)$	5.4 ± 0.1
$\Delta G^\circ_{37\ \text{initiation}}(7)$	6.0 ± 0.2
$\Delta G^\circ_{37\ \text{initiation}}(8)$	5.5 ± 0.2
$\Delta G^\circ_{37\ \text{initiation}}(9)$	6.4 ± 0.2
$\Delta G^\circ_{37\ \text{bonus}}$(UU or GA first mismatch but not AG)	−0.9 ± 0.1
$\Delta G^\circ_{37\ \text{bonus}}$(GG first mismatch)	−0.8 ± 0.3
$\Delta G^\circ_{37\ \text{bonus}}$(special G-U closure)	−2.2 ± 0.2
$\Delta G^\circ_{37\ \text{penalty}}(C_3$ loop)	1.5 ± 0.5
$\Delta G^\circ_{37\ \text{penalty}}(C_n$ loop), A	0.3 ± 0.1
$\Delta G^\circ_{37\ \text{penalty}}(C_n$ loop), B	1.6 ± 0.9

Hairpin loop stabilities are estimated with the equation $\Delta G^\circ_{37\ \text{loop}}$ ($n > 3$) = $\Delta G^\circ_{37\ \text{initiation}}(n)$ + ΔG°_{37}(first mismatch stacking) + $\Delta G^\circ_{37\ \text{bonus}}$(UU or GA first mismatch but not AG) + $\Delta G^\circ_{37\ \text{bonus}}$(GG first mismatch) + $\Delta G^\circ_{37\ \text{bonus}}$(special G-U closure) + $\Delta G^\circ_{37\ \text{penalty}}$(oligo-C loops), where n is the number of unpaired nucleotides in the loop. ΔG°_{37}(first mismatch stacking) is derived from studies of terminal mismatch stability as compiled previously (45, 46). Terminal mismatch free energies for UU mismatches on both GU and UG pairs were updated from Dale *et al.* (47). The special GU closure bonus applies to GU closed hairpins in which a 5′ closing G is preceded by two G residues. The oligo-C penalty applies only to loops composed of all C residues. The penalty for oligo-C loops >3 nt is $\Delta G^\circ_{37\ \text{penalty}}$(oligo-C loops, $n > 3$) = An + B. In addition to the terms in the above equation, the AU/GU terminal pair penalty of 0.5 kcal/mol is also applied at the ends of helices closed by hairpin loops (8, 44). Hairpin parameters were derived from linear regression on the database, excluding the stable hairpins A**CAGUGC**U (where closing pairs are shown and unpaired nucleotides are in bold), A**CAGUGA**U, A**CAGUUC**U, A**CAGUAC**U, C**UACG**G, C**UCCG**G, and C**UUCG**G (47, 50, 86). The supporting information contains the complete database of hairpin loops used in the linear regression. Hairpin loops of lengths at 3, 4, and 6 unpaired nt with measured free energies that are either more or less stable by 0.9 kcal/mol when compared to prediction by the above model are included in a separate lookup table (Table 2). Hairpin loops of <3 nt are prohibited. ΔG°_{37}(first mismatch stacking) and terminal mismatch bonuses apply only to hairpin loops >3 unpaired nt. For hairpin loops >9 nt, initiation free energy is approximated (74) by $\Delta G^\circ_{37\ \text{initiation}}(n > 9) = \Delta G^\circ_{37\ \text{initiation}}(9) + 1.75RT\ln(n/9)$.

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

Hairpin	Ref(s).	$\Delta G^\circ_{37\ \text{loop}}$, kcal/mol
C**AACG**	87	6.8
G**UUAC**	87	6.9
C**AACGG**	48	5.5
C**CAAGG**	48	3.3
C**CACGG**	48	3.7
C**CCAGG**	48	3.4
C**CGAGG**	48	3.5
C**CGCGG**	48	3.6
C**CUAGG**	48	3.7
C**CUCGG**	48	2.5
C**UAAGG**	48	3.6
C**UACGG**	47, 50	2.8
C**UCAGG**	48	3.7
C**UCCGG**	47	2.7
C**UGCGG**	47	2.8
C**UUAGG**	48	3.5
C**UUCGG**	47, 50	3.7
A**CAGUAC**U	86	2.8
A**CAGUGC**U	86	2.9
A**CAGUGA**U	86	3.6
A**CAGUUC**U	86	1.8

For extra stable hairpins measured in 0.1 M Na^+ (48, 87), placement was determined by assuming that the relative stability of loops remains constant between 0.1 and 1 M Na^+. All values are based on experimental results rather than frequencies of occurrence as used in ref. 8. Unpaired nucleotides are shown in bold.

2 × 2 internal loops, but approximations are used for most internal loops. The range of measured free energies differs for different types of internal loops. For example, the range is roughly 2 and 6 kcal/mol for 1 × 3 and 2 × 2 loops, respectively. Evidently, different types of loops require different approximations. Table 3 gives the different approximations used.

The free energy increment for multibranch loop initiation is roughly approximated by

$$\Delta G^\circ_{37\ \text{multibranch initiation}} = a + c^*(\text{number of branching helices}).$$

On the basis of experiments (20, 70), a better approximation would include another term, b^*(average asymmetry), but this cannot be accommodated in a dynamic programming algorithm. Therefore, asymmetry in the location of unpaired nucleotides in the loop is neglected. The parameters a and c were optimized by finding the best set in the region suggested by the experimental values (70) $a = 9.3 \pm 0.9$ kcal/mol and $c = -0.6 \pm 0.2$ kcal/mol. The maximum accuracy of folding was found for $a = 9.3$ kcal/mol and $c = -0.9$ kcal/mol. Accuracy was not highly sensitive to the values of a and c within the region suggested by experiment.

Incorporating Coaxial Stacking in the Dynamic Programming Algorithm. The RNASTRUCTURE program (8) was extended to include the free energy increments of coaxial stacking for adjacent helices as previously implemented (14) and for helices separated by a single mismatch. The WM array, of size $N \times 3$, introduced to speed the multibranch loop calculation (8), is expanded to $N \times N$, where N is the number of nucleotides in the sequence. At any point in the algorithm where the end of a helix (defined by i and j) is being considered for a multibranch or exterior loop, coaxial stacking of two helices is now considered. This calculation requires a search of k, $i < k < j$, to divide the region into two helix ends. The supporting information, which is published on the PNAS web site, shows the required recursions.

Constraining Secondary Structure Prediction with Chemical Modification Data. Chemically modified nucleotides are unpaired, in A-U or G-C pairs at helix ends, in G-U pairs anywhere, or adjacent to G-U pairs. The dynamic programming algorithm uses large positive free energies to forbid conformations inconsistent with the data. The supporting information states the recursions.

***In Vivo* Chemical Modification of 5S rRNA.** Modification agents were added to exponentially growing *E. coli* or *C. albicans*, OD_{540} 0.4–0.6, at a concentration of 1% vol/vol or wt/vol. At specific times, 10-ml aliquots of the cultures were removed. Cells were isolated by centrifugation and washed three times with sterile water; pellets were immediately placed into a dry ice ethanol bath. Total RNA was isolated from the cells by treatment with Triazol reagent (Invitrogen) supplemented by vortexing the cells with 100 μl of glass beads.

Reverse transcription (RT) was used to determine positions of modification. RT was run by using standard manufacturer's conditions with Na acetate as described (25, 28, 71) with 10 μg of total RNA in each reaction. Two RT primers were used with each RNA. For *E. coli* 5S rRNA, primer sequences were d(ATGCCTGGCAGTTCCC) and d(CTACCATCGGCGCTACGGCG). For *C. albicans* 5S rRNA, primer sequences were

Table 3. Approximations for internal loop free energy parameters at 37 °C (in kcal/mol)

Specification	Free energy increments					
$\Delta G^\circ_{37\ \text{initiation}}(n)$	0.5 ± 0.1 (2)	1.6 ± 0.1 (3)	1.1 ± 0.1 (4)	2.1 ± 0.1 (5)	1.9 ± 0.1 (6)	1.9 + 1.08 ln(n/6) (>6)
$\Delta G^\circ_{37\ \text{AU/GU}}$	0.7 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.7 ± 0.1
$\Delta G^\circ_{37\ \text{asym}}$	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1
Type of loop/first pair:	5′RA/3′YG	5′YA/3′RG	5′RG/3′YA	5′YG/3′RA	GG	UU
1 × 1	NA	NA	NA	NA	−2.6 ± 0.2	−0.4 ± 0.1 if 5′RU/3′YU
1 × 2	0	−1.1 ± 0.2	−1.1 ± 0.2	−1.1 ± 0.2	−1.1 ± 0.2	−0.7 ± 0.2
1 × (n − 1), n > 3	0	0	0	0	0	0
2 × 3	0	−0.5 ± 0.2	−1.2 ± 0.1	−1.1 ± 0.1	−0.8 ± 0.2	−0.4 ± 0.1
Others, except 2 × 2	−0.8 ± 0.1	−0.8 ± 0.1	−1.0 ± 0.1	−1.0 ± 0.1	−1.2 ± 0.1	−0.7 ± 0.1

For $\Delta G^\circ_{37\ \text{initiation}}$, the total number of nts in the loop is n. Free energy increments for single noncanonical pairs (68, 69), i.e. 1 × 1 loops, are approximated by $\Delta G^\circ_{37\ \text{loop}}(1 \times 1) = \Delta G^\circ_{37\ \text{loop initiation}}(n = 2) + \Delta G^\circ_{37\ \text{AU/GU}}$ (per AU or GU closure) + $\Delta G^\circ_{37\ \text{GG}}(1 \times 1) + \Delta G^\circ_{37\ 5'\text{RU}/3'\text{YU}}(1 \times 1)$. Here, $\Delta G^\circ_{37\ \text{loop initiation}}(n = 2)$ is the free energy of initiation for a single noncanonical pair with adjacent GC pairs; $\Delta G^\circ_{37\ \text{AU/GU}}$ is the penalty for replacing a closing GC pair with an AU or GU pair and replaces the AU/GU terminal pair penalty used for helices (8, 44), $\Delta G^\circ_{37\ \text{GG}}(1 \times 1)$ is a bonus for a GG pair in a 1 × 1 loop; and $\Delta G^\circ_{37\ 5'\text{RU}/3'\text{YU}}(1 \times 1)$ is a bonus for a 5′RU/3′YU stack in a 1 × 1 loop, where R is A or G in an AU or GC pair. Free energy increments for symmetric 2 × 2 loops lacking a measured value are approximated by interpolation of measured increments for loops of similar sequence (supporting information). Increments for nonsymmetric 2 × 2 loops are approximated by $\Delta G^\circ_{37\ \text{loop}}$ (5′PXYS/3′ QWZT) = 0.5 [ΔG°_{37}(5′PXWQ/3′QWXP) + ΔG°_{37}(5′TZYS/3′SYZT)] + Δ_p + $\Delta G^\circ_{37\ \text{GG}}(2 \times 2)$. Here, PQ and ST are canonical base pairs and XW and YZ are noncanonical pairs. The Δ_p term (0.6 ± 0.2 kcal/mol) is applied to loops with an AG or GA pair adjacent to a UC, CU, or CC pair and to loops with a UU pair adjacent to an AA pair. The $\Delta G^\circ_{37\ \text{GG}}(2 \times 2)$ term (−1.3 ± 0.2 kcal/mol) is applied to loops with a GG pair adjacent to an AA or any noncanonical pair with a pyrimidine. Values for Δ_p and $\Delta G^\circ_{37\ \text{GG}}$ were obtained by linear regression on the 2 × 2 loop database (8, 58, 62, 64, 65, 68). Other internal loops are approximated by $\Delta G^\circ_{37\ \text{loop}}(n) = \Delta G^\circ_{37\ \text{loop initiation}}(n) + \Delta G^\circ_{37\ \text{AU/GU}} + |n1 - n2|\ \Delta G^\circ_{37\ \text{asym}} + \Delta G^\circ_{37\ \text{first noncanonical pairs}}$(except for 1 × ($n$ − 1) for n > 3). Here, $\Delta G^\circ_{37\ \text{loop initiation}}(n)$ is the free energy of initiation for a loop of n nucleotides, $\Delta G^\circ_{37\ \text{asym}}$ is a penalty for loops with unequal numbers of nucleotides on each side, with $n1$ and $n2$ the number of nucleotides on each side, $\Delta G^\circ_{37\ \text{first noncanonical pairs}}$ (except for 1 × (n − 1) for n > 3) is a parameter for the incremental free energy of the first noncanonical pair on each side of the loop; it is not applied to loops of the form 1 × (n − 1) with n > 3. Values for the parameters were obtained from a set of fits to available data for 1 × 1 (69), 1 × 2 (59, 62, 63), 1 × 3 (59, 62), 2 × 2 (8, 58, 62, 64, 65, 68), 2 × 3 (59, 61, 62), and 3 × 3 (ref. 62 and X. Jiao and D.H.T., unpublished results) loops (supporting information) and from theory (74) for n > 6. NA, not applicable to that type of loop. Identical values for adjacent parameters indicate that they were fit as a single parameter.

d(AGATTGCAGCACAATAC) and d(AATTGCAGCACAATAG). Products were separated on a denaturing 8% polyacrylamide gel and quantified with a Molecular Dynamics PhosphorImager and IMAGE QUANT 4.1 software. Each mapping experiment was run in at least triplicate, and modifications are only reported if the nucleotides were modified in each experiment. Strong hits are bands that had at least 10 times the integrated volume of the equivalent band in the control lane, and moderate hits are between 3 and 10 times the volume. Loading was normalized with an RT stop position in each lane that was unchanged by chemical modification.

Availability. RNASTRUCTURE for Microsoft Windows is available at the Turner laboratory homepage: http://rna.chem.rochester.edu. Source code is available from D.H.M. upon request. Thermodynamic parameters are also available at the Turner laboratory web site.

Results

Nearest-Neighbor Parameters. Nearest neighbor parameters for prediction of RNA conformational free energy at 37°C are revised on the basis of recent experiments on terminal mismatches (47) and hairpin (47, 48), bulge (56), internal (58–61, 72), and multibranch (20, 70) loops. RNASTRUCTURE for prediction of secondary structures is modified accordingly, and coaxial stacking of helices that are adjacent or separated by a single mismatch has been added. The algorithm remains $O(N^3)$ in time and $O(N^2)$ in memory. Although slower than without coaxial stacking, the calculation remains rapid. For example, the calculation time for a complete small subunit rRNA, 1,542 nt, is 20 min, and the memory requirement is 47.1 MB on a Pentium 4, 1.6 GHz machine with 512 MB of RAM and Microsoft Windows 2000. The supporting information includes a table of calculation time and memory use for RNA sequences ranging from 77 to 2,904 nt.

The average accuracy of secondary structure prediction is 73% ± 9% of known canonical base pairs for a database (8) of ≈150,000 nt of known RNA structures, divided into domains of <700 nucleotides (supporting information). The single best structure of a set of up to 750 predicted suboptimal structures contains, on average, 87% ± 8% of known base pairs. Finally, 97% ± 3% of known base pairs are found in at least one of the suboptimal structures.

Prior studies took secondary structures generated by a dynamic programming algorithm and revised the free energies with a second program, called EFN2 (8, 73), that added free energy increments for coaxial stacking and a logarithmic dependence for the penalty for the number of unpaired nucleotides in a multibranch loop (74). RNASTRUCTURE no longer uses EFN2 to revise free energies because coaxial stacking increments are now included in the dynamic programming algorithm. The accuracy of predictions is essentially identical to that obtained previously after EFN2 rearrangement (8).

Chemical Modification Data As Folding Constraints. The dynamic programming algorithm can now incorporate constraints from chemical modification data. Prior versions of the algorithm were unable to use these constraints because chemical modifications occur not just at unpaired nucleotides, but also in A-U or G-C pairs at the ends of helices, G-U pairs anywhere, or adjacent to G-U pairs. Previous studies used modification data by either searching suboptimal structures predicted directly (27, 37) or generated from motifs found in suboptimal structures (19). Neither approach is rigorous, because neither guarantees the lowest free energy structure.

When secondary structure is poorly predicted by free energy minimization alone, accuracy can be significantly improved by adding chemical modification constraints. Fig. 1 shows the secondary structure predicted for the *E. coli* 5S rRNA with and without constraints determined by chemical mapping *in vivo*. The accuracy of prediction improves from 26.3% of base pairs correctly predicted to 86.8% (Table 4). The chemical mapping is consistent with the secondary structure determined by comparative sequence analysis (75, 76). Constraints based on *in vitro* chemical mapping of *E. coli* 5 S rRNA (29) give identical accuracy.

BIOCHEMISTRY

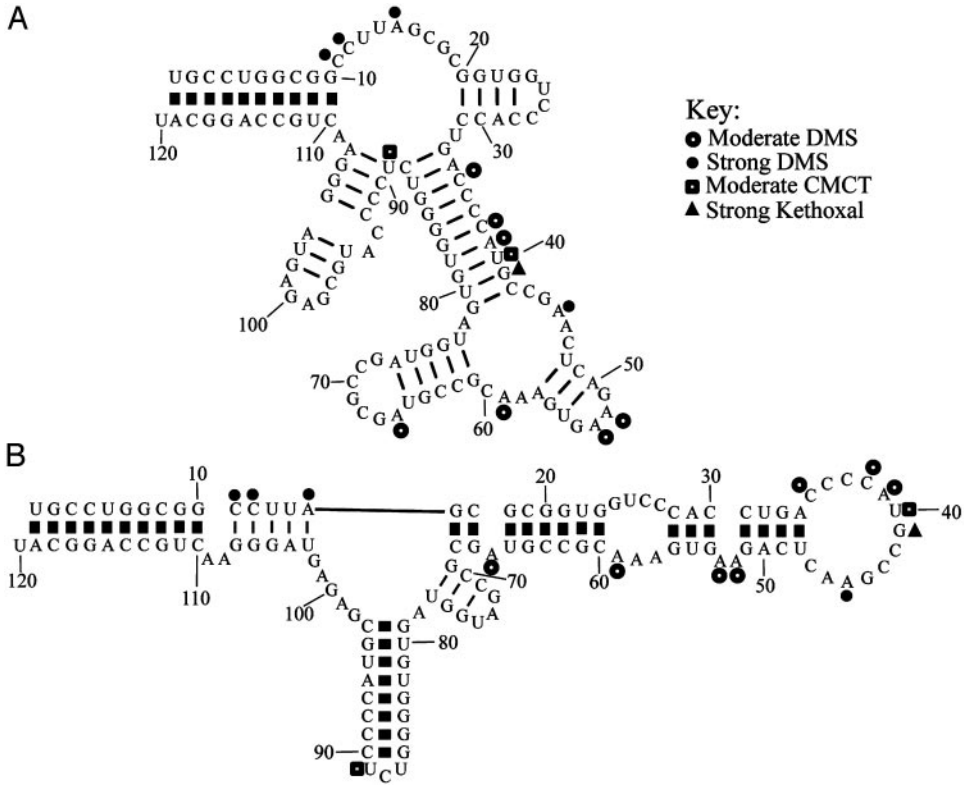

Fig. 1. The *E. coli* 5S rRNA secondary structure predictions and chemical modification. Heavy lines indicate base pairs in the known secondary structure (76, 88). (*A*) The predicted lowest free energy structure without experimental constraints. (*B*) The structure predicted with constraints from chemical modification data specified.

For *C. albicans* 5S rRNA, chemical mapping *in vivo* is consistent with the structure determined by comparative sequence analysis (supporting information). The predicted secondary structure is 90.6% accurate when calculated with or without chemical modification constraints (Table 4).

Table 4 also contains results for 14 RNA sequences of known

Table 4. The average accuracy of structure prediction with and without constraint with chemical modification data expressed as percentage of known canonical base pairs correctly predicted

RNA type	Ref(s).	Species	Pseudoknot basepairs, %	Unconstrained		Constrained	
				LFE	Best	LFE	Best
Signal recognition particle RNA	77, 81	Dog	0.0	18.2	97.7	84.1*	98.9*
5S rRNA *in vivo*	76	*E. coli*	0.0	26.3	86.8	86.8	97.4
Small subunit rRNA	25, 78	*E. coli*	1.6	39.0	49.0	63.3	73.2
RNase P	32, 80	*Chromatium vinosum*	10.5	53.5	81.6	53.5	81.6
RNase P	31, 80	*Bacillus subtilis*	7.1	56.3	70.5	56.3	68.8
RNase P	32, 80	*E. coli*	9.8	58.1†	73.4	64.5†	74.2
RNase P	30, 80	*Saccharomyces cerevisiae*	7.4	59.3	78.7	58.3	78.7
Telomerase RNA *in vivo*	43, 82, ‡	*Tetrahymena thermophila*	10.5	65.8	84.2	65.8	84.2
group I bI5	78, 89	*S. cerevisiae*	5.0	78.2	83.2	81.5	83.2
group I Intron *in vivo*	43, 78	*T. thermophila*	4.7	83.0	90.7	83.0	90.7
group II Intron aI5c	27, 79	Yeast	0.0	86.1	89.1	77.7	82.2
group I Intron L-21 Sca I	37, 78	*T. thermophila*	5.0	86.7	90.0	89.2	90.8
5S rRNA *in vivo*	76	*C. albicans*	0.0	90.6†	90.6	90.6†	90.6
Large subunit rRNA (domain 1)	26, 78	*E. coli*	0.4	88.9	90.5	88.9	91.3
group II Intron	79, 90	*Pylaiella littoralis*	0.0	90.3†	94.6	90.3†	94.6
5S rRNA	24, 76	Mouse	0.0	94.4	100.0	88.9	94.4
Average				67.2	84.4	76.4	85.9

Accuracies are reported for both the lowest free energy structure (LFE) and best suboptimal structure in a set of up to 750 structures, generated with a window size of zero.

*Results are reported for protein-bound RNA; when naked RNA chemical modification data are used, the accuracy is 64.8% for the lowest free energy structure and 89.8% for the best suboptimal structure.

†Best of three or four structures having identical free energies.

‡ten Dam, E., van Belkum, A. & Pleij, K. (1991) *Nucleic Acids Res.* **19,** 6951.

secondary structure with chemical modification data available in the literature (24–27, 30–32, 37, 43, 76–82, ‖). The average accuracy of secondary structure prediction without and with experimental constraints for the total database is 67% and 76%, respectively. This enhancement primarily reflects improvement for three sequences that are predicted with <40% accuracy on the basis of energetics alone.

The extent of chemical modification is generally graded into strengths. In this study, strong and moderate modifications are used as constraints. In most studies, weakly modified nucleotides can occur buried in helices and therefore are not suitable as constraints for secondary structure prediction. For example, 22 of 251 weak modifications in the small subunit rRNA are inconsistent with the accepted secondary structure (25, 78).

Discussion

Determination of RNA structure is important for understanding structure–function relationships and designing of therapeutics and diagnostics that target RNA. Free energy minimization is an important tool for elucidating RNA secondary structure because it can aid in the determination of a comparative sequence analysis model or suggest possible structures to test by site-directed mutagenesis or other methods. The accuracy of free energy minimization is limited, however, by lack of knowledge of the sequence and salt dependence of energetics and of the effects of tertiary and protein interactions. Constraints from chemical modification (22, 25, 28) can partially compensate for incomplete knowledge of all the factors determining RNA structure. Perhaps of most importance, *in vivo* studies (39–43) circumvent the difficulties (83) of finding *in vitro* conditions that mimic the native structure. This feature is an advantage of chemical modification as compared with nuclease mapping (84).

The *in vivo* chemical modification of *E. coli* 5S rRNA (Fig. 1) illustrates the impact of chemical modification constraints on secondary structure prediction. The results in Fig. 1 differ from previous *in vitro* mapping (29), largely because fewer nucleotides are accessible with *in vivo* mapping, probably because of protein binding. For example, nucleotides 73, 78, 99, and 104, which are accessible to modification *in vitro*, are not modified *in vivo*. In addition, the accessible nucleotides in the largest hairpin loop are shifted two nucleotides 5′ in the *in vivo* mapping so that four consecutive nucleotides are modified starting at position 38 *in vivo* as opposed to position 40 *in vitro*. Nevertheless, the chemical modification constraints increase the accuracy of secondary structure prediction from 26.3% to 86.8% (Table 4).

The results for dog signal recognition particle RNA further illustrate how chemical modification data can compensate for factors not included in structure prediction algorithms. The results from chemical modification change when the signal-recognition particle RNA is bound to protein (77). Presumably, the structure deduced by phylogenetic comparison (81) corresponds to the structure with protein bound. Thus, it is encouraging that the predicted structure is closest to the phylogenetic structure when chemical modification data for the RNA–protein complex (84.1%) rather than for the naked RNA (64.8%) are used as constraints (Table 4). Predictions of RNA secondary structure usually provide a number of possible structures with similar predicted free energies (8, 12, 15–18). In some cases, these predictions may reflect conformational switches that can be induced by binding of protein or other perturbations. Chemical modification constraints obtained in the presence and absence of protein and/or under different conditions may help reveal such dynamics.

The results in Table 4 fall into three general classes. Chemical modification constraints dramatically improve predictions of the three sequences that are <40% accurate on the basis of energetics alone. Sequences predicted with between 53% and 66% accuracy when unconstrained all have >7% of their nucleotides in pseudoknots. Pseudoknots are not allowed by the algorithm used in this study, and chemical modification restraints have little effect on the accuracy of prediction for these cases. The third class comprises eight sequences predicted with >78% accuracy on the basis of energetics alone. On average, the chemical modification results decrease the accuracy of predictions for these structures by 1% because of results for the yeast aI5c Group II intron and the mouse 5S rRNA. For the yeast Group II intron, 12 of the 127 moderate modifications violate the assumed rules; i.e., they are buried in helices and not in G-U pairs or adjacent to G-U pairs. This finding suggests that the Group II intron has more than one conformation in the mapping conditions used or that the structure from sequence comparison is not an equilibrium structure for naked RNA. For comparison, the *Phalacrocoracidae littoralis* Group II intron was mapped with a homogeneous sample and the constraints do not decrease accuracy. On the other hand, for mouse 5S rRNA, the modification data are consistent with the known secondary structure. For this case, a chemical modification is not consistent with the 94.4% accurate structure predicted by energetics alone, and an 88.9% accurate structure with two fewer correct base pairs is the predicted lowest free energy structure consistent with the modification data.

The approach described here for incorporating chemical modification constraints can be applied in essentially any dynamic programming algorithm for prediction of RNA secondary structure. In general, the constraints will reduce the number of structures generated. This should facilitate identification of pseudoknots by programs that allow them (14, 15). The inclusion of coaxial stacking in the dynamic programming algorithm of RNASTRUCTURE will also improve applications using dot plots because they now include the effects of coaxial stacking.

One difficulty in predicting RNA secondary structure is that the promiscuity of base pairing and the limited knowledge of the sequence dependence of loop energetics results in a large number of local free energy minima representing different secondary structures. The results presented here show that *in vitro* and *in vivo* chemical modification data can be used as constraints to limit predictions to those closely related to the structure of RNA in its true biological context. On average for sequences with <6% of nucleotides in pseudoknots, the structures predicted with constraints from chemical modification contain 84% of the known canonical base pairs. Such an accurate secondary structure model in conjunction with comparative sequence data can then be used to model tertiary contacts and therefore global folds (85). Development of more specific chemical modification reagents would allow tighter constraints and therefore even better deductions of secondary structures.

This work was supported by National Institutes of Health Grants GM22939 and GM54250 (to D.H.T and M.Z., respectively). D.H.M. was a trainee in the medical scientist training program, National Institutes of Health Grant 5T32 GM07356. J.L.C. was partially supported by National Institutes of Health Grant T32 DE07202.

‖ten Dam, E., van Belkum, A. & Pleij, K. (1991) *Nucleic Acids Res.* **19,** 6951.

1. Miranda-Rios, J., Navarror, M. & Soberón, M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 9736–9741.
2. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. (2001) *Science* **294,** 853–858.
3. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. (2001) *Science* **294,** 858–862.
4. Cullen, B. R. (2002) *Nat. Immunol.* **3,** 597–599.
5. McManus, M. T. & Sharp, P. A. (2002) *Nat. Rev. Genet.* **3,** 737–747.
6. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. (2000) *Science* **289,** 920–930.
7. Hansen, J. L., Schmeing, T. M., Moore, P. B. & Steitz, T. A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11670–11675.

8. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999) *J. Mol. Biol.* **288,** 911–940.
9. Gultyaev, A. P., van Batenburg, F. H. D. & Pleij, C. W. A. (1995) *J. Mol. Biol.* **250,** 37–51.
10. Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973) *Nat. New Biol.* **246,** 40–41.
11. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Res.* **9,** 133–148.
12. Zuker, M. (1989) *Science* **244,** 48–52.
13. Zuker, M. & Sankoff, D. (1984) *Bull. Math. Biol.* **46,** 591–621.
14. Rivas, E. & Eddy, S. R. (1999) *J. Mol. Biol.* **285,** 2053–2068.
15. Gaspin, C. & Westhof, E. (1995) *J. Mol. Biol.* **254,** 163–174.
16. Chen, S. & Dill, K. A. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 646–651.
17. Wuchty, S., Fontana, W., Hofacker, I. L. & Schuster, P. (1999) *Biopolymers* **49,** 145–165.
18. Ding, Y. & Lawrence, C. E. (2003) *Nucleic Acids Res.* **31,** 7280–7301.
19. Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. (1997) *RNA* **3,** 1–16.
20. Diamond, J. M., Turner, D. H. & Mathews, D. H. (2001) *Biochemistry* **40,** 6971–6981.
21. Pappalardo, L., Kerwood, D. J., Pelczer, I. & Borer, P. N. (1998) *J. Mol. Biol.* **282,** 801–818.
22. Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J. & Ehresmann, B. (1987) *Nucleic Acids Res.* **15,** 9109–9128.
23. Butcher, S. E. & Burke, J. M. (1994) *J. Mol. Biol.* **244,** 52–63.
24. Miura, K., Tsuda, S., Ueda, T., Harada, F. & Kato, N. (1983) *Biochim. Biophys. Acta* **739,** 281–285.
25. Moazed, D., Stern, S. & Noller, H. F. (1986) *J. Mol. Biol.* **187,** 399–416.
26. Egebjerg, J., Leffers, H., Christensen, A., Andersen, H. & Garrett, R. A. (1987) *J. Mol. Biol.* **196,** 125–136.
27. Kwakman, J. H. J. M., Konings, D. A. M., Hogweg, P., Patel, H. J. & Grivell, L. A. (1990) *J. Biomol. Struct. Dyn.* **8,** 413–430.
28. Inoue, T. & Cech, T. R. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 648–652.
29. Brunel, C., Romby, P., Westhof, E., Ehresmann, C. & Ehresmann, B. (1991) *J. Mol. Biol.* **221,** 293–308.
30. Tranguch, A. J., Kinderberger, D. W., Rohlman, C. E., Lee, J. & Engelke, D. R. (1994) *Biochemistry* **33,** 1778–1787.
31. Odell, L., Huang, V., Jakacka, M. & Pan, T. (1998) *Nucleic Acids Res.* **26,** 3717–3723.
32. LaGrandeur, T. E., Hüttenhofer, A., Noller, H. F. & Pace, N. R. (1994) *EMBO J.* **17,** 3945–3952.
33. Matsuura, M., Noah, J. W. & Lambowitz, A. M. (2001) *EMBO J.* **20,** 7259–7270.
34. Wakao, H., Romby, P., Westhof, E., Laalami, S., Grunberg-Manago, M., Ebel, J., Ehresmann, C. & Ehresmann, B. (1989) *J. Biol. Chem.* **264,** 20363–20371.
35. Melander, Y., Holmberg, L. & Nygård, O. (1997) *J. Biol. Chem.* **272,** 3254–3258.
36. Banerjee, A. R. & Turner, D. H. (1995) *Biochemistry* **34,** 6504–6512.
37. Banerjee, A. R., Jaeger, J. A. & Turner, D. H. (1993) *Biochemistry* **32,** 153–163.
38. Powers, T., Daubresse, G. & Noller, H. F. (1993) *J. Mol. Biol.* **232,** 362–374.
39. Harris, K. A., Jr., Crothers, D. M. & Ullu, E. (1995) *RNA* **1,** 351–362.
40. Doktycz, M. J., Larimer, F. W., Pastrnak, M. & Stevens, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14614–14621.
41. Wells, S. E., Hughes, J. M. X., Igel, A. H. & Ares, M., Jr. (2000) *Methods Enzymol.* **318,** 479–492.
42. Disney, M. D., Haidaris, C. G. & Turner, D. H. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 1530–1534.
43. Zaug, A. J. & Cech, T. R. (1995) *RNA* **1,** 363–374.
44. Xia, T., SantaLucia, J., Jr., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C. & Turner, D. H. (1998) *Biochemistry* **37,** 14719–14735.
45. Turner, D. H. (2000) in *Nucleic Acids*, eds. Bloomfield, V. A. Crothers, D. M. & Tinoco, I. (University Science Books, Sausalito, CA), pp. 259–334.
46. Xia, T., Mathews, D. H. & Turner, D. H. (2001) in *RNA*, eds. Söll, D. G., Nishimura, S. & Moore, P. B. (Elsevier, New York), pp. 21–48.
47. Dale, T., Smith, R. & Serra, M. (2000) *RNA* **6,** 608–615.
48. Proctor, D. J., Schaak, J. E., Bevilacqua, J. M., Falzone, C. J. & Bevilacqua, P. C. (2002) *Biochemistry* **41,** 12062–12075.
49. Antao, V. P., Lai, S. Y. & Tinoco, I., Jr. (1991) *Nucleic Acids Res.* **19,** 5901–5905.
50. Antao, V. P. & Tinoco, I., Jr. (1992) *Nucleic Acids Res.* **20,** 819–824.
51. Giese, M. R., Betschart, K., Dale, T., Riley, C. K., Rowan, C., Sprouse, K. J. & Serra, M. J. (1998) *Biochemistry* **37,** 1094–1100.
52. Groebe, D. R. & Uhlenbeck, O. C. (1988) *Nucleic Acids Res.* **16,** 11725–11735.
53. Serra, M. J., Lyttle, M. H., Axenson, T. J., Schadt, C. A. & Turner, D. H. (1993) *Nucleic Acids Res.* **21,** 3845–3849.
54. Serra, M. J., Barnes, T. W., Betschart, K., Gutierrez, M. J., Sprouse, K. J., Riley, C. K., Stewart, L. & Temel, R. E. (1997) *Biochemistry* **36,** 4844–4851.
55. Serra, M. J., Axenson, T. J. & Turner, D. H. (1994) *Biochemistry* **33,** 14289–14296.
56. Znosko, B. M., Silvestri, S. B., Volkman, H., Boswell, B. & Serra, M. J. (2002) *Biochemistry* **41,** 10406–10417.
57. Woodson, S. A. & Crothers, D. M. (1987) *Biochemistry* **26,** 904–912.
58. Burkard, M. E., Xia, T. & Turner, D. H. (2001) *Biochemistry* **40,** 2478–2483.
59. Schroeder, S. J. & Turner, D. H. (2000) *Biochemistry* **39,** 9257–9274.
60. Schroeder, S. J. & Turner, D. H. (2001) *Biochemistry* **40,** 11509–11517.
61. Schroeder, S. J., Fountain, M. A., Kennedy, S. D., Lukavsky, P. J., Krugh, T. R., and Turner, D. H. (2003) *Biochemistry*, **42,** 14184–14196.
62. Peritz, A. E., Kierzek, R., Sugimoto, N. & Turner, D. H. (1991) *Biochemistry* **30,** 6428–6436.
63. Schroeder, S., Kim, J. & Turner, D. H. (1996) *Biochemistry* **35,** 16105–16109.
64. Wu, M., McDowell, J. A. & Turner, D. H. (1995) *Biochemistry* **34,** 3204–3211.
65. Walter, A. E., Wu, M. & Turner, D. H. (1994) *Biochemistry* **33,** 11349–11354.
66. SantaLucia, J., Jr., Kierzek, R. & Turner, D. H. (1991) *J. Am. Chem. Soc.* **113,** 4313–4322.
67. SantaLucia, J., Jr., Kierzek, R. & Turner, D. H. (1991) *Biochemistry* **30,** 8242–8251.
68. Xia, T., McDowell, J. A. & Turner, D. H. (1997) *Biochemistry* **36,** 12486–12497.
69. Kierzek, R., Burkard, M. & Turner, D. H. (1999) *Biochemistry* **38,** 14214–14223.
70. Mathews, D. H. & Turner, D. H. (2002) *Biochemistry* **41,** 869–880.
71. Childs, J. L., Disney, M. D. & Turner, D. H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11091–11096.
72. Schroeder, S. J., Burkard, M. E. & Turner, D. H. (1999) *Biopolymers* **52,** 157–167.
73. Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Müller, P., Mathews, D. H. & Zuker, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 9218–9222.
74. Jacobson, H. & Stockmayer, W. H. (1950) *J. Chem. Phys.* **18,** 1600–1606.
75. Fox, G. E. & Woese, C. R. (1975) *Nature* **256,** 505–507.
76. Szymanski, M., Barciszewska, M. Z., Barciszewski, J. & Erdmann, V. A. (2000) *Nucleic Acids Res.* **28,** 166–167.
77. Andreazzoli, M. & Gerbi, S. A. (1991) *EMBO J.* **10,** 767–777.
78. Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Muller, K. M., et al. (2002) *BMC Bioinformatics* **3,** Available at www.biomedcentral.com/bmcbioinformatics. Accessed January 17, 2002.
79. Michel, F., Umesono, K. & Ozeki, H. (1989) *Gene* **82,** 5–30.
80. Brown, J. W. (1999) *Nucleic Acids Res.* **27,** 314.
81. Gorodkin, J., Knudsen, B., Zwieb, C. & Samuelsson, T. (2001) *Nucleic Acids Res.* **29,** 169–170.
82. Romero, D. P. & Blackburn, E. H. (1991) *Cell* **67,** 343–353.
83. Uhlenbeck, O. C. (1995) *RNA* **1,** 4–6.
84. Knapp, G. (1989) *Methods Enzymol.* **180,** 192–212.
85. Michel, F. & Westhof, E. (1990) *J. Mol. Biol.* **216,** 585–610.
86. Laing, L. G. & Hall, K. B. (1996) *Biochemistry* **35,** 13586–13596.
87. Shu, Z. & Bevilacqua, P. C. (1999) *Biochemistry* **38,** 15369–15379.
88. Speek, M. & Lind, A. (1982) *Nucleic Acids Res.* **10,** 947–965.
89. Chamberlin, S. I. & Weeks, K. M. (2003) *Biochemistry* **42,** 901–909.
90. Costa, M., Christian, E. L. & Michel, F. (1998) *RNA* **4,** 1055–1068.