Almost all RNAs can fold to form extensive base-paired secondary structures. Many of these structures then modulate numerous fundamental elements of gene expression. Deducing these structure–function relationships requires that it be possible to predict RNA secondary structures accurately. However, RNA secondary structure prediction for large RNAs, such as a single predicted structure for a single sequence reliably represents the correct structure, has remained an unsolved problem. Here, we demonstrate that quantitative, nucleotide-resolution information from a SHAPE experiment can be interpreted as a pseudo-free energy change term and used to determine RNA secondary structure with high accuracy. Free energy minimization, by using SHAPE pseudo-free energies, in conjunction with nearest neighbor parameters, predicts the secondary structure of deproteinized *Escherichia coli* 16S rRNA (>1,300 nt) and a set of smaller RNAs (75–155 nt) with accuracies of up to 96–100%, which are comparable to the best accuracies achievable by comparative sequence analysis.

---

**Results**

Structure of *Escherichia coli* 16S rRNA, as Predicted by a Best-of-Category Algorithm. We focused on 16S ribosomal RNA (rRNA) because its structure is known and it contains numerous typical RNA motifs (14, 15). We predicted the secondary structure of 16S rRNA by using the program RNAstructure (11), whose algorithm finds the lowest free energy structure by using empirical thermodynamic parameters fit against a large database of model structures with known stability (11, 16). We also implemented a maximum allowable distance between base pairs of 600 nt, because 99% of base pairs in rRNAs involve pairings of less than this distance (12, 17). Throughout this work, we only consider the lowest free energy structure output by RNAstructure because, even if more accurate structures are predicted at higher folding free energies, there is no general way to identify these as improved structures.

Prediction errors can be of 2 classes. Either known base pairs are missed or base pairs are predicted that do not exist in the accepted target structure. These errors are reported by 2 prediction accuracy measures, sensitivity and positive predictive value (PPV; the percentage of predicted base pairs in the known structure). By using thermodynamic information alone, prediction sensitivity and PPV for *E. coli* 16S rRNA are 49.7% and 46.2%, respectively (errors are illustrated with red x’s and lines; Fig. 1).

A critical objective of RNA secondary structure prediction is to create models useful for developing biological hypotheses regarding RNA function. This objective can be well met by defining the overall topology of an RNA in terms of the constituent helices and their connectivity. Thus, we also calculate the prediction sensitivity for helices. There are 69 helices in the covariation structure for 16S rRNA, defined as a continuous stack of 3 or more canonical base pairs interrupted by no more than a single nucleotide bulge. Overall, 52% of helices in 16S rRNA are predicted in the lowest-energy structure. Errors are distributed unevenly throughout the RNA and, for example, 71% (15 of 21) of helices in the 3’ major domain are not predicted correctly (Fig. 1). All 3 metrics, sensitivity of base pairs, PPV, and sensitivity of helices, support the same conclusion. For 16S rRNA, the predicted secondary structure is correct in some regions; whereas, in other regions, the structure is completely wrong (Fig. 1 and Table 1).

The structure of 16S rRNA has been assessed by using conventional chemical modification reagents (DMS, kethoxal, and CMCT) (18). Prediction accuracies using RNAstructure improve when positions judged to have strong or moderate reactivities are prohibited from participating in Watson–Crick base pairs except at the end of helices or adjacent to GU pairs: the resulting sensitivity and PPV are 71.8% and 67.4%, respectively; 75% of helices are predicted correctly [Table 1 and supporting information (SI) Fig. S1]. However, predictions at 75% sensitivity are still characterized by many regions with large errors (Fig. S1). An alternate, widely used, 2-criterion approach for interpreting chemical modification data, prohibiting sites of chemical modification from forming internal base pairs and forcing sites of strong reactivity to be single-stranded, actually reduces accuracy: sensitivity and PPV
decrease to 66.7% and 64.2%, only 70% of helices are predicted correctly (Table 1).

In sum, these calculations emphasize the persistent and unmet challenges in secondary structure prediction. Neither thermodynamic-based prediction nor prediction constrained by conventional chemical mapping data yield an accurate structure for 16S rRNA. Developing useful biological hypotheses by using RNA secondary structures predicted at even 75% sensitivity is difficult. Moreover, widespread prediction of elements that are not in the accepted structure, as reflected in a poor PPV, underscores the difficulty, or impossibility, of designing instructive experiments guided by this level of accuracy.

Redefining the RNA Secondary Structure Prediction Problem. Current thermodynamic parameters are spectacularly useful for predicting the stability of individual helices and hairpins (7, 19). However, several factors make it difficult to predict large RNA structures. First, many structures have predicted folding free energies within 1 kcal/mol of that for the most stable structure. Second, kinetic processes and protein–RNA interactions may modulate RNA folding. Third, local interactions exhibit complex sequence-dependent interactions (20, 21) and it may not be possible to account for all interactions with a tractable number of parameters.

Local nucleotide flexibility can be measured at the vast majority of positions in any RNA by use of SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) chemistry (22, 23). SHAPE is approaching conventional DNA sequencing in terms of the facility and straightforwardness with which it can be performed (24–27). In a SHAPE experiment, RNA is treated with an electrophile that reacts selectively, but sparsely, with the 2'-hydroxyl position at conformationally flexible nucleotides to form a 2'-O- adduct. 2'-O-adducts are then detected by primer extension. SHAPE reactivities report the extent to which a nucleotide is constrained by base pairing or other interactions (22, 24, 27–29). We therefore sought to redefine the RNA secondary structure prediction problem to use quantitative, nucleotide-resolution SHAPE information in concert with thermodynamic parameters for RNA folding.

SHAPE Analysis of E. coli 16S and 23S rRNAs. Total RNA was purified from E. coli bacteria by using a nondenaturing protocol, equili-
brated under conditions that stabilize native RNA structure (Fig. 2A), and treated with 1-methyl-7-nitroisatoic anhydride (1M7) (24). Sites of adduct formation were detected by a high-throughput SHAPE approach in which the primer extension reactions, performed by using color-coded fluorescently labeled DNA primers, are resolved by capillary electrophoresis (Fig. 2B) (24, 25). SHAPE reactivities for each primer read, covering 350–600 nt, were normalized by using model-free statistics to a scale spanning 0 to ≈2, where 1.0 is the average intensity for highly reactive positions (Fig. 2C). Nucleotides with normalized SHAPE reactivities >0.7 or 0.3–0.7 are considered highly and moderately reactive, respectively, and are colored red and yellow. Unreactive nucleotides, with SHAPE reactivities <0.3, are black (Fig. 2D).

We analyzed 91% and 95% of the nucleotides in E. coli 16S and 23S rRNAs (1,542 and 2,904 nt, respectively) in many regions, including domain II of 23S rRNA, agreement between SHAPE reactivities and the secondary structure determined by comparative sequence analysis is essentially perfect (Fig. 3). Nucleotides that participate in canonical base pairs are unreactive; whereas, nucleotides in loops, bulges, and other connecting regions are reactive (compare black with red and yellow nucleotides; Fig. 3). In a few regions, nucleotides expected to be base paired are scored as reactive by SHAPE (blue boxes, Fig. 3); these positions apparently reflect regions in which evolutionarily supported base pairs do not form when rRNA is isolated from bacteria. The number of such nucleotides is small, ~9% of all nucleotides in the 16S and 23S rRNAs. SHAPE thus provides comprehensive, direct, and quantitative information regarding the structure of large RNAs.

\[ \Delta G_{\text{SHAPE}}(i) = m \ln(\text{SHAPE reactivity}(i) + 1) + b \]  

This model has 2 free parameters, the intercept b and slope m. The intercept is negative and represents a favorable energy free energy increment for pairing nucleotides at which the SHAPE reactivity is low. The slope is positive and penalizes base pairing at nucleotides with high SHAPE reactivities. The \( \Delta G_{\text{SHAPE}} \) term was integrated into the dynamic programming algorithm in RNAstructure (11) as an additional nearest neighbor free energy change term (16).

The slope and intercept were parameterized against 23S rRNA by using the secondary structure determined by comparative sequence analysis (15) as the target structure (Fig. 4). 23S rRNA is a good choice for parameterization because this single RNA encompasses a large database of diverse and nonredundant RNA motifs. In this analysis, we excluded nucleotides (14%) where SHAPE shows that base pairs in the comparative structure do not form or for which no SHAPE reactivity information was obtained (blue boxes and gray nucleotides, Figs. 3 and S2). In the absence of the \( \Delta G_{\text{SHAPE}} \) term, base pairs in 23S rRNA are predicted with a sensitivity and PPV of 72% and 60% (0.0 point; Fig. 4). As the absolute values of the intercept and slope increase, prediction accuracy improves to produce a large “sweet spot” corresponding to >89% sensitivity (in red, Fig. 4).

The optimal parameter regions for both sensitivity and PPV are large. Good predictions are therefore obtained even if the \( \Delta G_{\text{SHAPE}} \) parameters are varied by large increments (Fig. 4). As general parameters for folding large RNAs, we selected a slope and intercept of 2.6 and −0.8 kcal/mol, respectively, because this point corresponds to a high prediction sensitivity, is adjacent to other
10S rRNA Structure Determination. We chose a point close to the origin to impose the points in the sweet spot, and is as close as possible to the origin. We selected parameters centrally located in the optimal region to accommodate RNAs whose folding properties might differ from 23S rRNA. We chose a point close to the origin to impose the smallest bias in the nearest neighbor free energy calculation consistent with high prediction accuracy. The estimate of >89% correctly predicted base pairs in 23S rRNA (Fig. 4) is a conservative, lower limit because some regions in the deproteinized rRNA do not actually fold to the phylogenetically accepted structure.

16S rRNA Structure Determination. Use of ΔGSHAPE free energies, optimized against 23S rRNA, dramatically increase the prediction accuracy for E. coli 16S rRNA (compare Figs. 1 and 5). We considered 3 target structures when quantifying the overall prediction accuracy.

1. The structure determined by comparative sequence analysis. This is a conservative approach and assumes that all base pairs showing evolutionary covariation are maintained in the free RNA in the absence of ribosomal proteins.
2. The comparative structure after omitting regions (i) that clearly do not fold to this structure as judged by SHAPE or (ii) for which no structural data could be obtained. These “omit” regions are emphasized with blue boxes and gray nucleotide lettering, respectively (Fig. 5).
3. A structure that allows for local RNA refolding. Although we purified 16S and 23S rRNAs from cells under nondenaturing conditions (Fig. 2A), the deproteinized 16S rRNA clearly rehocks in some regions (Fig. 5 and ref. 18). Many base pairs predicted by our algorithm are, in fact, strongly supported by SHAPE data. For this target, we thus allow alternative base pairings in regions where a well-defined local RNA refolding is more consistent with the experimental SHAPE reactivity than are the base pairs in the comparative structure. There are 43 such base pairs, corresponding to 6% of the nucleotides in 16S rRNA (in green, Fig. 5). We also allow local refolding at the 4-helix junction spanning positions 139–224 because direct experimental analysis supports the alternate model (Fig. S3).

Taking the secondary structure model established by comparative sequence analysis as the target structure (target 1), sensitivity and PPV for SHAPE-directed prediction of E. coli 16S rRNA are 84.2% and 80.9% (Table 1). The overall topology is also good: 90% of all helices are identified correctly.

If regions for which SHAPE reactivities are clearly incompatible with the comparative structure or for which no data could be obtained are omitted (target 2), sensitivity and PPV are 91.1% and 83.1%, respectively (Table 1). Moreover, the topology is almost exactly right: 95% of helices outside of the omit regions are predicted correctly.

Allowing for experimentally supported refoldings (target 3; identified with green dots and boxes, Fig. 5), sensitivity and PPV are 97.2% and 95.1%. Sixty-eight of the 69 helices are predicted correctly and thus the topology of the RNA is correct (Table 1 and Fig. 5).

Structure Determination for Nonribosomal RNAs. To assess the generality of the SHAPE-directed approach, we also determined
secondary structures for 3 smaller, pseudoknot free, RNAs: yeast tRNA<sup>Ap</sup>, domain II of the HCV internal ribosome entry sequence (HCV IRES), and the P546 domain of the bI3 group I intron. Inclusion of SHAPE constraints yields accurate structures in all cases. The structure of tRNA<sup>Ap</sup> is well predicted by thermodynamics parameters alone (95% sensitivity), but SHAPE data still provide sufficient information to yield a perfect prediction (100% sensitivity). The HCV IRES and bI3 intron RNAs are, like 16S rRNA, poorly predicted by thermodynamic information alone; critically, inclusion of SHAPE information results in nearly perfect predictions (Table 2; structures are provided in Fig. S4).

**Discussion**

By incorporating experimental SHAPE information as a pseudo-free energy change term in RNAstructure, we determine the structures of *E. coli* 16S rRNA and of 3 smaller RNAs almost perfectly (Fig. 5, Tables 1 and 2). Differences between the SHAPE-directed structures and the accepted target structures are usually small and short-range. At this level of difference, it is not clear whether the error lies in the predicted structure or in the accepted structure. SHAPE-directed secondary structure determination also gives excellent results for wide choices in the 2 free ΔG<sub>SHAPE</sub> parameters and is thus tolerant of experimental and procedural variability (Fig. 4).

16S rRNA is among the most comprehensive structure prediction challenges available. The secondary structure for 16S rRNA

**Table 2. Prediction accuracies for nonribosomal RNAs**

<table>
<thead>
<tr>
<th>RNA</th>
<th>Nucleotides</th>
<th>No constraints Sensitivity</th>
<th>SHAPE Sensitivity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast tRNA&lt;sup&gt;Ap&lt;/sup&gt;</td>
<td>75</td>
<td>95.2</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>HCV IRES domain II</td>
<td>95</td>
<td>56.5</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>P546 domain, group I intron</td>
<td>155</td>
<td>42.9</td>
<td>96.4</td>
<td>98.2</td>
</tr>
</tbody>
</table>
was established by comparative sequence analysis and 97% of the predicted base pairs are visualized in the crystal structure of intact 30S ribosomal subunits (15). This modeling accuracy required analysis of 7,800 sequences and refinement over 20 years. The 97.7% sensitivity obtained here for deproteinized 16S rRNA based on a single SHAPE analysis is comparable to that achieved by covariation analysis. We find that SHAPE-directed folding also yields excellent results for RNAs whose structures cannot be determined by covariation analysis such as folding intermediates (27–29) and intact viral genomes (25).

The simplicity of SHAPE chemistry and the availability of appropriate data analysis tools (this work and refs. 11 and 26) make this technology amenable to a wide variety of problems. There remain 2 major, addressable challenges. First, none of the 5 RNAs studied here form pseudoknots in their deproteinized forms and our algorithm does not allow this structure. In the future, experimentally based ΔGSHAPE pseudo-free energy approaches can clearly be incorporated into algorithms that predict secondary structures with pseudoknots. Second, extensions of the current experimental approach will be required for RNA regions in which base pairs either form only in context of higher-order tertiary interactions (24) or are so tightly constrained by such interactions that few nucleotides are reactive.

The high level of confidence demonstrated by SHAPE-directed RNA structure determination now makes it possible to analyze the plurality of RNA secondary structures that cannot be gleaned from comparative sequence analysis or that are changing in response to dynamic cellular processes. Such RNAs include authentic viral genomes, intact messenger RNAs, and noncoding RNAs in distinct functional states.

Supporting Information

Deigan et al. 10.1073/pnas.0806929106

Materials and Methods

Isolation of Escherichia coli Protoplasts. Cell cultures (E. coli DH5α) were grown to midlog phase (A600 = 0.8) under shaking in Luria broth, aliquoted (1.5 ml), collected by centrifugation (5 min, 4 °C, 17,000 × g), and resuspended in 1 ml of 15 mM Tris (pH 8.0), 450 mM sucrose, and 8 mM EDTA. Lysozyme (400 μg) was added and the solution was incubated at 22 °C for 5 min, and on ice for 10 min. Protoplasts were collected by centrifugation (5 min, 4 °C, 5,000 × g).

Recovery of Native RNA. Protoplast pellets were resuspended in 120 μl of 50 mM Heps (pH 8.0), 200 mM NaCl, 5 mM MgCl2 and 1.5% (wt/vol) SDS and incubated at 22 °C for 5 min and on ice for 5 min. SDS was precipitated by adding 30 μl of 50 mM Heps (pH 8.0), 1 M potassium acetate, and 5 mM MgCl2. The precipitate was collected and discarded by centrifugation (5 min, 4 °C, 17,000 × g) and the buffer of the RNA-containing solution exchanged by gel filtration (G-50, 400-μl column) prequilibrated in 1× folding buffer [50 mM Heps (pH 8.0), 200 mM potassium acetate (pH 8.0), and 5 mM MgCl2]. RNA was eluted in the same solution. The RNA-containing eluent was extracted 3 times with phenol [(pH 8.0):chloroform:isoamyl alcohol; 25:24:1] and 3 times with chloroform; 1.5 ml of bacterial culture was diluted to 6.0 μl with water and incubated at 65 °C for 5 min, 45 °C for 5 min, and placed on ice. Primer extension was initiated by addition of enzyme mix [3 μl; 250 mM KCl; 167 mM Tris-HCl (pH 8.3), 1.67 mM each deoxynucleotide, 10 mM MgCl2; 52 °C, 1 min], SuperScript III (0.5 μl, 100 units, Invitrogen), and incubation at 45 °C for 1 min, 52 °C for 7 min, and 65 °C for 5 min. Sequencing reactions were identical, except that they used unmodified nonnative rRNA (4.8 μg), the RNA-primer solution was diluted to 6.0 μl in deionized water, and 0.5 μl of a 10 mM ddNTP solution was added immediately before SuperScript III. Primers used for sequencing were typically labeled with either 6-TAMRA and 5-ROX (Anaspec) or NED and PET (Applied Biosystems) fluorophores. Appropriate reactions (+) and (−) reagent, 2 sequencing extensions; Each extension by using the same primer sequence but labeled with a different fluorophore] were quenched by precipitation with ethanol, washed 3 times with 70% ethanol, dried under vacuum, and redissolved in 10 μl of deionized formamide. The cDNA samples were resolved on an Applied Biosystems 3130 capillary electrophoresis DNA sequencer.

SHAPE Data Processing. Raw electropherograms of fluorescence intensity versus elution time (Fig. 2B) were analyzed by using ShapeFinder (2, 3). Data processing steps included baseline and mobility shift corrections and a correction for signal decay as a function of primer extension length. Peaks for the (+) and (−) reagent channels were aligned with each other and also with the RNA sequence. The area under each peak was quantified by whole-channel Gaussian integration. After subtracting background, SHAPE reactivities from each primer read were placed on a normalized scale by dividing by the average intensity of the 10% most highly reactive nucleotides, after first excluding highly reactive outliers. Outliers in each primer dataset were identified by using a model-free box plot analysis as reactivities >1.5 × the interquartile range (4). For small RNA datasets (<~100 nt), the maximum number of outliers is capped at 5%. Use of model-free statistics is important because SHAPE reactivities do not exhibit a normal distribution. These calculations place all SHAPE reactivities on a scale spanning 0 to ~2 (Fig. 2C). Reactivity data for each primer were processed and normalized independently. Final SHAPE data for different, overlapping, primers for the 16S and 23S rRNAs consistently fell on the same scale, without the need for additional normalization.

SHAPE Analysis of tRNAAsp, HCV IRES Domain II, and bI3 Group I Intron P546 Domain. Data for tRNAAsp and the bI3 intron were reported previously (5, 6). Data for the HCV IRES will be reported elsewhere. Accepted target structures for these RNAs were taken from refs. 7–9.
Incorporation of SHAPE Pseudo-Free Energy Change Terms into a Dynamic Programming Algorithm. All structure calculations were performed by using RNAstructure (10). The maximum allowed base-pairing distance for all structure calculations was 600 nt. \( \Delta G_{\text{SHAPE}}(i) \) is added to the free energy change for each nucleotide \( i \) in a base-pairing stack (11). For the total folding free energy change, nucleotide \( i \) therefore contributes \( \Delta G_{\text{SHAPE}} \) if it is involved in a stack at the end of helix, \( 2 \times \Delta G_{\text{SHAPE}} \) if it is paired and adjacent to 2 stacked pairs or a single bulged nucleotide, or zero if it is single-stranded.

Calculation of Prediction Sensitivity, PPV, and Fraction Correct Helices. When calculating sensitivity and PPV, base pairs between nucleotides \( i \) and \( j \) are considered correctly predicted pairs if the known structure contains a pair between \( i \) and \( j \), \( i + 1 \) or \( i - 1 \) and \( j \), or \( i \) and \( j + 1 \) or \( j - 1 \). We allow these slipped pairs because of the difficulty in conclusively determining the exact pairing and because these predictions are qualitatively consistent with correct pairs (12). We define a helix as a continuous stack of 3 or more canonical base pairs interrupted by no more than a single nucleotide bulge. By this definition, there are 69 helices in \( E. coli \) 16S rRNA. Helices were considered correct if \( >50\% \) of the constituent base pairs were predicted correctly. Helices were not included in the calculation if \( >50\% \) of their nucleotides lie in omit regions or have no SHAPE data. For the calculation of prediction accuracy after allowing local refolding (target 3), we required (i) single additional base pairs to stack on an existing helix (for example, the 587–754 base pair; Fig. 5) and (ii) multiple base pairs be separated from a phylogenetically supported helix by a bulge of 2 nucleotides or fewer (for example, the 980 loop; Fig. 5).

Fig. S1. Accuracy of secondary structure prediction for *E. coli* 16S rRNA using conventional chemical reagents. Reagent data included DMS, kethoxal and CMCT (13). Nucleotides judged to show strong or moderate reactivity toward chemical probes (13) were prohibited from forming Watson–Crick base pairs at internal positions in a helix unless they are adjacent to a GU pair. Missed base pairs are represented by red x's; incorrectly predicted base pairs are represented by purple lines.
Fig. S2. Secondary structure prediction for the 5\' (A) and 3\' (B) regions of E. coli 23S RNA using SHAPE constraints. SHAPE data are superimposed on the structure determined by comparative sequence analysis (14); nucleotides are colored by their SHAPE reactivities. Nucleotides with no data are gray. Missed base pairs are represented by red x’s; incorrectly predicted base pairs are represented by purple lines. ΔG_{SHAPE} parameters were an intercept of −0.8 kcal/mol and a slope of 2.6 kcal/mol; the maximum base pairing distance was 600 nt.
Prediction accuracies for 23S rRNA as a function of SHAPE constraints

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Target</th>
<th>Sensitivity</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>comparative sequence model</td>
<td>65.4</td>
<td>60.6</td>
</tr>
<tr>
<td>SHAPE</td>
<td>with omit regions</td>
<td>78.7</td>
<td>76.7</td>
</tr>
<tr>
<td>SHAPE</td>
<td></td>
<td>89.8</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Fig. S2. continued.
Comparative SHAPE analysis of 16S RNA at the four-helix junction at nucleotides 139–224. Two possible conformations for this region were tested using in vitro transcripts and comparison with the structure in the intact 16S rRNA. The in vitro transcripts (spanning nucleotides 126–235) were imbedded within 5' and 3' structure cassette sequences (15) to facilitate analysis by SHAPE. Three RNA transcripts were tested: (i) a native sequence RNA (NS), (ii) a mutant that strengthens the proposed alternate structure (M1), and (iii) a mutant that strengthens the conventional structure proposed based on sequence covariation (M2).

(A) Integrated SHAPE data for 16S RNA, the NS RNA, and M1. Overall SHAPE reactivities for these three RNAs are similar. (B) Integrated SHAPE data for 16S RNA and M2. There are clear differences in SHAPE reactivities between the two RNAs, most notably at positions 158 and 177. (C) SHAPE data superimposed on proposed alternative and covariation-based secondary structures. Boxes indicate mutations; arrows point from the native sequence nucleotide to the mutated nucleotide. In sum, these experiments strongly support the interpretation that the 3-helix junction in deproteinized 16S rRNA does not fold to the structure derived from covariation analysis but, instead, folds to the alternate structure emphasized by the box in C.
Fig. S4. Accuracy of SHAPE-directed prediction for 3 small, nonribosomal, RNAs using ΔG_{SHAPE} parameters (of −0.8, 2.6 kcal/mol). Left and Right illustrate prediction accuracies in the absence and presence of SHAPE constraints, respectively.