Published online 00

Preprint, 2012, Vol. 00, No. 00 ***1–12***
doi:10.1093/nar/gkn000

RNA Folding with Soft Constraints: Reconciliation of Probing Data and Thermodynamic Secondary Structure Prediction

Stefan Washietl[1*], Ivo Hofacker[2], Peter F. Stadler[3,4,5], Manolis Kellis[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, 32 Vassar Street, Cambridge, MA 02139, USA
[2]Institute for Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[3]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[4]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[5]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

ABSTRACT

Thermodynamic folding algorithms and structure probing experiments are commonly used to determine the secondary structure of RNAs. Here we propose a formal framework to reconcile information from both prediction algorithms and probing experiments. The thermodynamic energy parameters are adjusted using "pseudo-energies" to minimize the discrepancy between prediction and experiment. Our framework differs from related approaches that used pseudo-energies in several key aspects. (i) The energy model is only changed when necessary and no adjustments are made if prediction and experiment are consistent (ii) Pseudo-energies remain biophysically interpretable and hold positional information where experiment and model disagree (iii) The whole thermodynamic ensemble of structures is considered thus allowing to reconstruct mixtures of suboptimal structures from seemingly contradicting data. (iv) The noise of the energy model and the experimental data is explicitly modeled leading to an intuitive weighting factor through which the problem can be seen as folding with "soft" constraints of different strength. We present an efficient algorithm to iteratively calculate pseudo-energies within this framework and demonstrate how this approach can be used in combination with SHAPE chemical probing data to improve secondary structure prediction. We further demonstrate that the pseudo-energies correlate with biophysical effects that are known to affect RNA folding such as chemical nucleotide modifications and protein binding.

INTRODUCTION

RNAs fulfill a large number of diverse biological functions in the cell (1). This wide functional spectrum of RNAs is made possible by the structural diversity of these highly flexible molecules. Studying the structure of a novel RNA thus is often the first step towards elucidating a possible biological function. Resolving the complete tertiary structure is a complex undertaking, however, so it is usually the secondary structure that is analyzed first. In a typical probing experiment, the RNA is enzymatically digested or chemically modified in a manner that is specific for structural context (2, 3). These experiments typically reveal which nucleotides are contained within a double stranded helix and which nucleotides form unpaired loops. In addition, solvent accessibility or local flexibility can be assessed, see (4) for a recent review. Structure probing experiments have been routinely used for many years. More recently, high-throughput methods have been introduced (5, 6, 7, 8) and next generation sequencing techniques have made it possible to perform probing experiments even on a genome-wide scale (9, 10).

All these experiments, however, only report partial information on the structure and even a perfect experiment does not reveal the actual base pairing patterns (11). Therefore, the results of probing experiments need to be combined with computational predictions. Most commonly, programs such as `mfold` (12), `RNAstructure` (13), or `RNAfold` (14) are employed that predict secondary structures by minimizing free energy. They are based on an empirical energy model (15), which is based on a very large set of thermodynamic measurements on small RNA oligonucleotides. In the simplest case, the predicted structure is manually adjusted to fit the measured constraints. To automatize this process, prediction programs allow the user to restrict the search space to only consider structures compatible with certain constraints observed in the probing data.

An alternative method is to include information from experiments as "pseudo-energies" in the energy model. This approach was introduced by Matthews *et al.* (16, 17) and is implemented in the program `RNAstructure` (13). On

*To whom correspondence should be addressed. Tel: +1-617 253 6284; Fax: +1-617-253-6652; Email: wash@mit.edu

chemical probing data generated by selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (5) it showed nearly perfect results on *E. coli* 16s rRNA (17) and it was successfully used to predict structure models for the complete HIV genome (18).

In this paper, we expand on the idea of incorporating experimental data as pseudo-energies into energy-based folding algorithms. Instead of adding *ad hoc* modifications to the minimum free energy calculation, we propose a formal method to reconcile experimental information with the theoretical prediction in the partition function over all possible structures. The partition function describes the entire ensemble of secondary structures in thermodynamic equilibrium and allows to calculate an intuitive matrix of base-pairing probabilities (19).

Our approach is based on the assumption that both experimental measurements and the thermodynamic energy parameters are imperfect, noisy approximations of the physical reality. In this setting it becomes natural to ask for a perturbation vector that minimizes a weighted sum of perturbation energies and discrepancies between measured and predicted base pairing probabilities. In the simplest case, this can be written as a least square approximation problem of the form

$$F(\vec{\epsilon}) = \sum_{\mu} \frac{\epsilon_\mu^2}{\tau_\mu^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (p_i(\vec{\epsilon}) - q_i)^2 \rightarrow \min \qquad \textbf{(1)}$$

Here, $\vec{\epsilon}$ is the perturbation vector and ϵ_μ the perturbation energy added for some structural element μ. $p_i(\vec{\epsilon})$ and q_i are the predicted and measured base pairing probabilities for position i, respectively. The estimated variances τ_μ^2 and σ_i^2 of energy parameters and measurements, respectively, serve as weighting factors. This optimization problem can be viewed as energy directed folding with soft constraints replacing the hard combinatorial constraints used before.

We show here (i) that the pseudo-energies $\vec{\epsilon}$ can be efficiently calculated by an iterative algorithm, (ii) that the approach combined with SHAPE data leads to improved secondary structure predictions (iii) that the algorithm also can successfully handle cases of RNAs with several alternative structures, and (iv) that the pseudo-energies have an interpretable meaning and indicate positions where experimental data and the thermodynamic energy model disagree.

MATERIALS AND METHODS

Minimization of the objective function using gradient descent

The objective function in eq. **1** and the motivation behind it is explained in more detail under "Rationale" in the Results section. Here, we show how to efficiently find the minimum of this function.

The minimum of the objective function, eq. **1**, satisfies $\partial F/\partial \epsilon_\mu = 0$ for all parameters, i.e.,

$$\epsilon_\mu = -\tau_\mu^2 \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (p_i(\vec{\epsilon}) - q_i) \frac{\partial p_i}{\partial \epsilon_\mu}(\vec{\epsilon}) \qquad \textbf{(2)}$$

Numerically, this can be solved by iteratively minimizing F. We use a gradient descent iteration of the form

$$\begin{aligned} \epsilon'_\mu &= \epsilon_\mu - a \frac{\partial F}{\partial \epsilon_\mu} \\ &= \left(1 - \frac{2a_t}{\tau_\mu^2}\right) \epsilon_\mu - 2a \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (p_i(\vec{\epsilon}) - q_i) \frac{\partial p_i}{\partial \epsilon_\mu}(\vec{\epsilon}) \end{aligned} \qquad \textbf{(3)}$$

with a step size $a < 0$. We chose this approach because it only depends on the first order derivatives of p_i with respect to ϵ_μ. In the following paragraphs we show that the required partial derivatives $\partial p_i/\partial \epsilon_\mu|_{\vec{\epsilon}}$ can be obtained analytically from constrained partition functions.

Analytic calculation of the gradient

Since ϵ_μ denotes the energy contribution that is added to all secondary structures that contain a particular "structural feature" μ, we can subdivide the structure ensemble into those structures that "have μ", and those that do not. This is possible for *any* parameter of the standard energy model and for any additional position-dependent term. Let $Z[i](\epsilon_\mu)$ be the partition function over all states with position i unpaired in the perturbed energy model, while $Z[i](0)$ is the corresponding partition function in the reference state. Similarly, $Z[\mu](.)$ and $Z[i,\mu](.)$ denote the partition functions over all structures that "have μ", and of those that both "have μ" and leave i unpaired, respectively, for each of the two energy models. The crucial observation is that the following identities hold for these constrained partition functions:

$$\begin{aligned} Z[i](\epsilon_\mu) &= Z[i](0) - Z[i,\mu](0) + Z[i,\mu](\epsilon_\mu) \\ Z(\epsilon_\mu) &= Z(0) - Z[\mu](0) + Z[\mu](\epsilon_\mu) \end{aligned} \qquad \textbf{(4)}$$

By construction, furthermore, we have

$$\begin{aligned} Z[\mu](\epsilon_\mu) &= Z[\mu](0) \exp(-\epsilon_\mu/RT) \\ Z[i,\mu](\epsilon_\mu) &= Z[i,\mu](0) \exp(-\epsilon_\mu/RT) \end{aligned} \qquad \textbf{(5)}$$

Since $p_i(.) = Z[i](.)/Z(.)$ we can express the partial derivatives in terms of restricted partition functions. We only need to compute the derivates at the reference energy model (which we take to be the energy model in each step of the

gradient iteration).

$$\begin{aligned}\left.\frac{\partial p_i}{\partial \epsilon_\mu}\right|_{\epsilon_\mu=0} &= \left.\frac{\partial}{\partial \epsilon_\mu}\frac{Z[i](0)-Z[i,\mu](0)\left(1-e^{-\epsilon_\mu/RT}\right)}{Z(0)-Z[\mu](0)\left(1-e^{-\epsilon_\mu/RT}\right)}\right|_{\epsilon_\mu=0} \\ &= \frac{1}{RT}\left[\frac{Z[i](0)}{Z(0)}\frac{Z[\mu](0)}{Z(0)}-\frac{Z[i,\mu](0)}{Z[i](0)}\frac{Z[i](0)}{Z(0)}\right] \\ &= \frac{1}{RT}p_i(0)\left[p[\mu](0)-p[\mu|i](0)\right]\end{aligned} \quad \textbf{(6)}$$

The probabilities of the structural patterns, $p[\mu]$, can be obtained by McCaskill's algorithm (19) provided μ is a base pair (k,l), an unpaired position j, or another feature that appears implicitly in the dynamic programming recursions.

Implementation for position specific perturbations

Here we consider the simplest case of perturbations that add positive or negative energy contributions to single positions. In that case case we can replace the generic dimension μ with an additional index $1 \leq j \leq n$, and the gradient takes the form

$$\left.\frac{\partial p_i}{\partial \epsilon_j}\right|_{\epsilon_j=0} = \frac{1}{RT}p_i(0)\left[p_j(0)-p[j|i](0)\right] \quad \textbf{(7)}$$

We have extended the implementation of McCaskill's algorithm in the Vienna RNA package (version 2.0 beta) to calculate the partition function of a sequence with additional position specific energy contributions. More precisely, during the energy evaluation step that calculates energies for different structural elements such as stacked pairs, hairpins, interior loops and multi-loops, we add ϵ_j if position j is unpaired for the particular structural element. By adding negative (favorable) perturbation energies we enforce a position to be unpaired, while adding positive perturbation energies will lead to a position be more likely to be paired. p_i can then directly be calculated from the partition function under the perturbed energy model.

In principle, also the term $p[j|i]$ can be easily calculated directly from the partition function. The conditional probability that j is unpaired given that i is unpaired as well can be obtained by constraining the dynamic programming recursion to structures in which i is unpaired. However, the partition function algorithm scales $\mathcal{O}(n^3)$ in CPU time with length n. Evaluating all n conditional probabilities renders the whole algorithm requiring $\mathcal{O}(n^4)$. This is too expensive in terms of computational resource for practical applications, however.

To overcome this problem, we estimate the term $p[j|i]$ by sampling structures from the thermodynamic ensemble. We use stochastic backtracking (20) to randomly generate structures proportional to their Boltzmann weight and empirically determine $p[j|i]$ from the random structures.

To get actual structure models from the base-pair probability matrix, we used the maximum-expected accuracy approach (21, 22) with a γ-parameter of 1.0.

Missing data, i.e. positions i for which no q_i is available, are handled transparently by setting $\sigma=\infty$ resulting in position i being effectively being ignored in the evaluation of the objective function (eq. **1**).

Analysis of SHAPE data

We used SHAPE reactivities for 23S and 16S rRNAs as reported by Deigan *et al.* (17). The 23S and 16S rRNAs were split in 6 and 4 domains, respectively, of a maximum length of 700nts as described before (21). We did not use domain 4 of the 16S rRNA because it was poorly covered by the SHAPE data and mainly consisted of missing data. As a reference structure we used the same phylogenetically derived structure as in (17). Accuracy was measured as *Sensitivity*=(*number of correctly predicted base-pairs*)/(*total number of known base-pairs*) and the positive predictive value *PPV* = (*number of correctly predicted base-pairs*)/(*total number of predicted base-pairs*). As a combined measure of sensitivity and positive predictive value we also used the Mathews correlation coefficient as described previously (23). Deigan *et al.* found 16.5% and 13.6% of the positions in the 16S and 23S rRNA, respectively, where the *in vitro* folded RNA as probed by the SHAPE data is different from the phylogenetic structure (corresponding the *in vivo* proteinized state). Deigan et al. removed these sites in their benchmark, while we kept it for the benchmarks reported here. The overall accuracies achieved in our benchmarks are therefore lower as reported by Deigan *et al.*

To use the SHAPE data with our algorithm we discretized the reactivities by classifying them in paired and unpaired positions using a cutoff of 0.25 (see also (11)). This cutoff corresponds to an error rate of about 25% of positions being incorrectly classified. We also tried a two cutoff approach and classified all positions with SHAPE reactivities <0.1 and >1.5 as paired and unpaired, respectively. This lowers the error rate to about 10% at the expense of a lower coverage of around 50%, i.e. more missing data. We did not see any significant advantage using this approach for any of the methods (data not shown). Furthermore we also tried more sophisticated machine learning methods to classify bases as "paired" and "unpaired" according to their SHAPE signal. Essentially, we face a machine learning problem to parse the continuous SHAPE signal as shown in Supplementary Fig. 1E into discrete states. In principle, this enables us to consider also the context of a base during classification. However, also here we did not find a significant improvement over the simple thresholding approach.

`RNAstructure` (version 5.3) was run with default values and with parameters of $m=2.6$ and $b=-0.8$ there were found to be optimal on this specific data-set (17). The "Sample + Select" strategy described in (11) was re-implemented using RNAfold, 10^5 structures were sampled and the structure with the lowest Manhattan distance to the discretized SHAPE vector was used. The results for hard constraints were calculated with `RNAfold` and the option `-C`.

Availability

All source code accompanying this paper can be downloaded here: https://github.com/wash/probing.

RESULTS

Rationale

Similar to previous approaches (16, 17) our algorithm modifies the folding energy parameters (15), which are used in `RNAfold`, `mfold`, and `RNAstructure`. In the following, we refer to these standard parameters as the "reference energy model" and the positive or negative pseudo-energies that change this model as "perturbations". Let $\vec{\epsilon}$ be a vector of perturbations of the reference energy model. In the most generic formulation, we consider a collection of structural elements whose contribution to the energy model can be perturbed. We use the index μ to refer to one of these degrees of freedom, which correspond to the coordinates of the vector of perturbation energies $\vec{\epsilon}$. Note that these degrees of freedom need not be structural elements that correspond to parameter of the reference energy model.

Our goal is to find a vector $\vec{\epsilon}$ that changes the standard energy model in the light of the experimental data. Deigan *et al.* (17) chose the perturbations proportional to the experimental signal. More precisely, for each position i they mapped the SHAPE reactivity $R(i)$ — the experimental signal for being unpaired (24) — to perturbation energies using the following relationship $a+m\ln[1+R(i)]$.

In practice, this strategy gave good results. Theoretically, however, this approach is poorly justified; in particular, there does not seem to be a meaningful biophysical interpretation of the energy model. Ideally, if experiment and the energy model agree perfectly, $\vec{\epsilon}$ should vanish. By setting $\vec{\epsilon}$ proportional to the experimental signal, however, the exact opposite is the case. Positions that show the highest signal in the experiment and already are predicted with high probability are assigned the highest perturbation energies.

Here, we regard both the experimental data and the structure prediction based on the energy model as a noisy approximation to the physical ground truth. Therefore, our goal is to find a perturbation vector that minimizes the discrepancy between the experimental measurement and computational prediction. In particular, we seek a perturbation vector ϵ that modifies the energy model only when necessary.

This is achieved by minimizing the total error of both energy model and measurements. We assume that the experimental data is given in form of a probabilistic signal as a vector q_i of the probabilities that position i is unpaired and an associated variance σ_i^2. Likewise, we assume a variance τ_μ^2 for the uncertainty of of the parameters of the standard energy model. Assuming, furthermore, that individual energy parameters as well as the measurements for each sequence position are independent, we obtain the error function

$$F(\vec{\epsilon}) = \sum_\mu \frac{\epsilon_\mu^2}{\tau_\mu^2} + \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(p_i(\vec{\epsilon}) - q_i\right)^2$$

Here, $p_i(\vec{\epsilon})$ is the *predicted* probability that nucleotide i is unpaired in the energy model perturbed by $\vec{\epsilon}$. The choice of the quadratic error function $F(\vec{\epsilon})$ is the most natural one from a mathematical point of view since all its terms have a natural interpretion as variances. The minimization problem thus evenly distributes the residual deviations between energy parameter set and measured data depending on their intrinsic variances σ_i^2 and τ_μ^2.

In principle, both the variances σ_i^2 and τ_μ^2 can be estimated from probing experiments and the experiments underlying the standard energy model, respectively. However, in this paper we will not use explicit estimates but rather treat them as parameters that control whether more weight is given on the experimental data or prediction of the energy model. If $\tau \gg \sigma$ the algorithm will find a solution closest to the experimental information, while in the other case the experimental information will be ignored and the solution will be essentially the same as the prediction of the unperturbed energy model. Note that the solution $\vec{\epsilon}_{\min}$ of the optimization problem depends only on the ratio τ/σ, i.e., on the relative accuracy of the energy model and measured probing data.

Iterative adaption of the energy parameters

So far we did not specify which parameters μ of the energy model are actually considered to be subject of perturbations. Since typical experiments only report data on whether a base is likely to be paired or unpaired, it is not useful to consider base-pairs or any higher order structural elements. We therefore concentrate on the simplest case and consider only position-specific perturbations ϵ_j (Methods).

We have implemented an efficient strategy to find the minimum of the objective function in this case (Methods). It is based on a gradient descent algorithm. The gradient for the objective function can be calculated analytically (see Methods).

We first tested the algorithm on an artificial sequence that can fold in two alternative structures. The one-stem-structure corresponding to the ground state of the unperturbed energy model, $\vec{\epsilon}=0$, is energetically highly favorable. The less stable alternative three-stem-structure is used here as the experimentally supported structure that our algorithm is supposed to recover. We considered the paired/unpaired probability profile of the target structure as perfect "experimental" data and set q_i to 0.0 or 1.0 for paired and unpaired positions i, respectively. Accordingly, we chose the associated variance σ^2 of $\vec{q}$ low and set $\sigma^2=0.01$ and $\tau^2=1.0$. This example is a hard test for our approach: since the two structures have very distinct pairing profiles, major refolding is required to correct the energy-based prediction.

We start with $\vec{\epsilon}=0$. Using the exact solution for the gradient (eq. **7**, Methods), we observe that the algorithm finds a minimum after about 150 iterations (Fig. 1, upper left diagram). This minimum is confirmed by the fact that norm of the gradient (Fig. 1, below) converges to zero and is <0.001 after 246 iterations. The corresponding base pairing probability matrices gradually change from the original one-stem-structure to the alternative three-stem structure. In the minimum, the structure is completely refolded and conforms to the desired target structure (Fig. 1, right).

We repeated the minimization calculation starting from five different random vectors $\vec{\epsilon}$. All five start points lead to

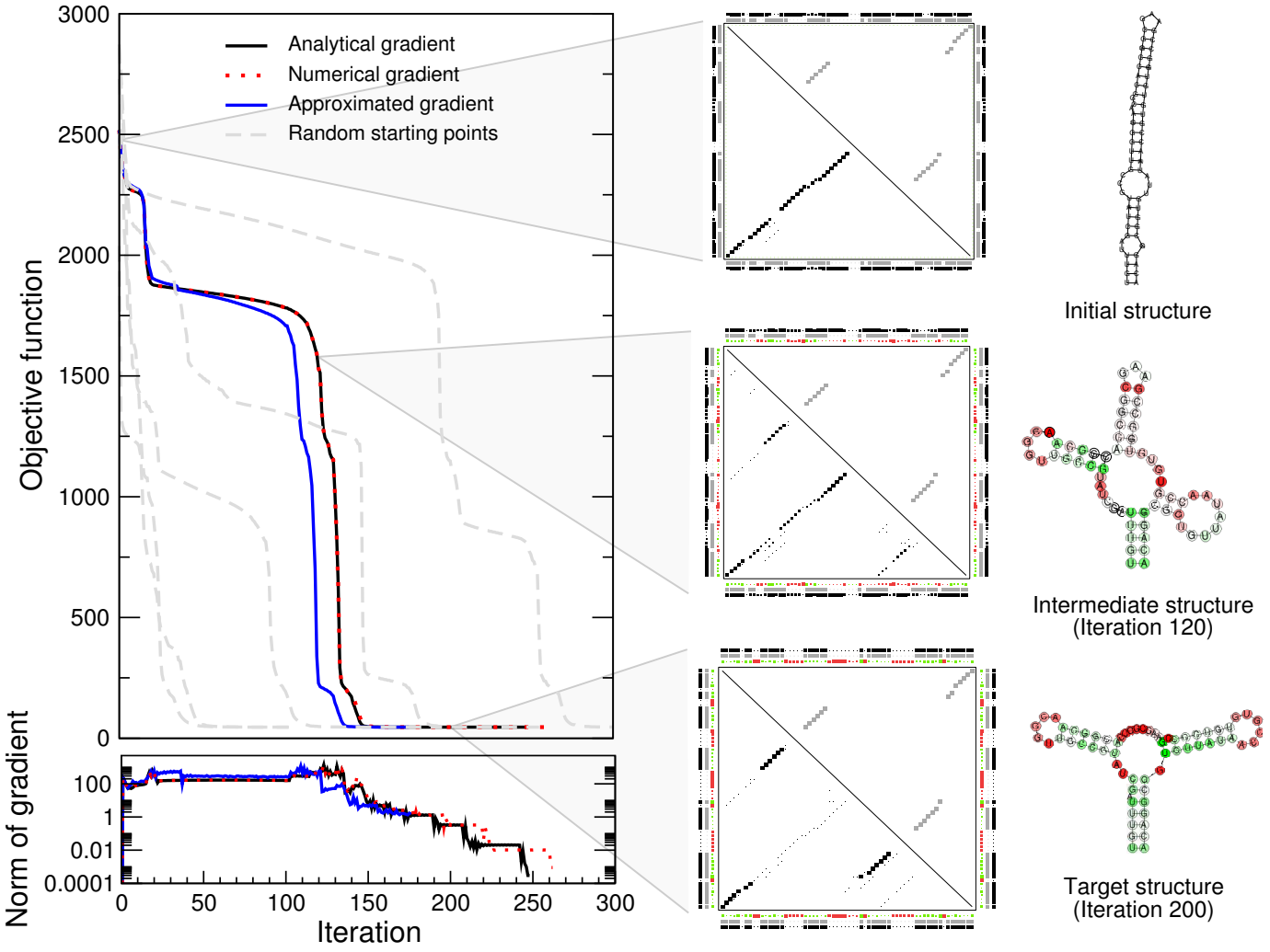

Figure 1. Iterative adaption of energy parameters. A sequence is re-folded from a single stem structure to a three stem structure. The diagram on the left shows the minimization path of the objective function (eq. **1**) and the associated norm of the gradient. The algorithm was run using different versions of the gradient (exact, approximated by numeric differentiation, approximated by sampling) and 5 different random initializations of the perturbation vector. The pair probabilities and associated structures are shown for three points in the iteration on the right. The upper right half of the matrix shows the base-pair matrix for the target structure, while the lower left matrix shows the matrix for the current prediction. The gray and black boxes in the margin, denote the probability of being paired for each position (gray: target structure, black: current prediction). The area of the boxes is proportional to the respective probability. The green/red boxes represent the values of the current perturbation vector ϵ (normalized between 0 and 1). Red represents a negative energy contribution for an unpaired position, i.e. supports a position to be unpaired. Green represents a positive energy contribution for an unpaired position, i.e. supports this position to form a base-pair.

the same minimum confirming that the procedure is robust and finds consistently the same solution. To further confirm the validity of the analytically derived gradient, we repeated the iteration with a numerically calculated gradient $\frac{\partial F(\vec{\epsilon})}{\partial \epsilon_i} = \frac{F(\vec{\epsilon},\epsilon_i+d)-F(\vec{\epsilon},\epsilon_\mu-d)}{2d}$. Setting $d=10^{-5}$, we observe that both solutions lead to exactly the same minimization path (Fig. 1).

Efficient solutions for long sequences

The exact analytical solution of the gradient as well as the numerical approximation scales as $\mathcal{O}(n^4)$ with the sequence length n (Methods). The form of the analytical solution given in eq. **7** (Methods), however, suggests that a major speedup can be achieved if the term $p[i|j]$ can be computed more efficiently. This can be done by random sampling from the thermodynamic equilibrium since an accurate estimate is needed only when position j is unpaired with a noticeable probability, so that fairly small samples are sufficient. We repeated the minimization using this approximation. Sampling the gradient from 10,000 random structures leads to the same minimum as the exact solution. In this particular example, we observe a slight deviation in the minimization paths after about 8 iterations (Fig. 1). In most other examples, however, we observed the paths to be identical. Only when the minimization reaches a point close to convergence, the approximated gradient fails to further improve the objective function. In this example, the calculation with the sampled gradient stopped after 113 iterations with the norm of the gradient in the order of 1.

To verify that our algorithm is capable of finding the solution also for longer sequences in reasonable time we ran the algorithm on RNAs of different lengths. We used the

same parameters as before and used the known secondary structure as "target" structure. To test if the sampled gradient gives the same solution as the exact gradient, we ran the minimization with the exact gradient until the norm of the gradient was <0.1 and with the sampled gradient until the objective function could not be improved any more. Using the sampling approach, the solution was found within seconds for small RNAs of about 100 nucleotides like tRNAs or the 5S rRNA and within minutes for longer RNAs of about 300 nucleotides (Tab. 1). Following previous work (17, 21), the longest sequence tested was 686 nucleotides long. Also for this length our algorithm using the sampled gradient could find a solution within an hour. In contrast, the exact solution took longer than six days. The objective function was minimized by more than 95% in all cases. Despite the extreme differences in running time, the sampling approach led to essentially the same rate of minimization as the exact approach.

Improved structure models using SHAPE probing data

We next demonstrate that our algorithm in the course of minimizing the objective function actually optimizes the secondary structure prediction. Following Deigan *et al.* (17), we used *E. coli* 23S and 16S rRNA to benchmark the structure models obtained by our algorithm.

First, we considered the limiting case of perfect data, i.e., we used the paired/unpaired profile of the reference structures (Methods) as input. In that case, the new iterative "soft constraint" algorithm should give the same results as the "hard" combinatorial constraints that can be applied to classical minimum free energy folding. We ran our algorithm with different combinations of τ/σ and compared the results to `RNAfold` without constraints and with hard constraints (Fig. 2, A). Combinations $\sigma \gg \tau$ essentially ignore the external data resulting in similar predictions as standard `RNAfold`. With increasing weight on the external data (i.e. $\sigma \ll \tau$) the accuracy increases and finally converges to the same level of `RNAfold` with hard constraints. This level represents the theoretical accuracy that can be achieved by the combination of thermodynamic folding with probing data.

We next used SHAPE data (17) to test our algorithm on real probing data. The SHAPE signal measures local nucleotide flexibility. The signal is generally higher in unpaired regions than in paired regions (Supplementary Fig. 1, A-C). It is important to note, however, that there is no simple relationship between nucleotide flexibility and base-pair probabilities and there are systematic differences between these two properties beyond statistical noise (Supplementary Fig. 1, D and E). For example, SHAPE signals have a typical peak structure with nucleotides in the middle of a loop being usually the most reactive. However, the probability of these nucleotides to be unpaired in the thermodynamic ensemble has generally not the same peak shape (Supplementary Fig. 1, E). We have tried various ways to map the SHAPE signal to the probability vector q_i. However, we found that converting the SHAPE signal into a discrete vector with $q_i = \{1.0, 0.0\}$ using a simple thresholding approach (Methods) gave the best results.

Again, we ran our algorithm with varying values of τ/σ (Fig. 2A). We observed an improvement in prediction accuracy over the standard RNAfold prediction with increasing weight on the SHAPE data. However, at around $\tau/\sigma = 0.5$ the improvement peaks for both the 23S and 16S rRNA, and due to the inherent noise in the SHAPE experiment, the accuracy drops again when more weight is given on the experimental data.

We also run RNAfold with hard constraints on the same data. Here, the accuracy does not improve and is generally worse than RNAfold without probing data. In contrast to the soft constraint algorithm, a small number of inaccurate constraints introduced by the noise in the data can almost completely destroy the prediction in this case.

We further compared to two other methods that were used in combination with SHAPE data before. We ran minimum free energy prediction augmented with pseudo energies as described in Deigan et al. ("RNAstructure+SHAPE") (17). We used the same parameters m and b that were found to be optimal on exactly the same data by Deigan *et al.* In addition, we also implemented the "Sample and Select" approach described in Quarrier et al. (11). This strategy samples a large number of random structure from the ensemble and chooses a structure with the minimum distance to the probing data under a simple distance metric (Methods). Fig. 2B summarizes the results for all methods averaged over all domains of both rRNAs. We found that all methods except RNAfold with hard constraints lead to improved predictions over RNAfold (and the equivalent RNAstructure implementation) of about 15%–20%. Our soft constrained algorithm achieves 0.70 ± 0.08 sensitivity and 0.71 ± 0.07 positive predictive value, while "RNAstructure+SHAPE" and the "Sample+Select" approach achieve $0.70 \pm 0.07 / 0.67 \pm 0.09$ and $0.67 \pm 0.07 / 0.65 \pm 0.08$, respectively.

Recovering the ensemble of a bistable structure

So far we only considered the case that the external pairing signal originates from a single target structure. However, RNA molecules typically are not present in single structure but form an ensemble in which very different structures can be present simultaneously. This is of biological significance in particular for riboswitches (25) and ribozymes (26, 27, 28). The signal measured in a probing experiment, therefore, will in general be a superposition of responses from structural alternatives. We tested, therefore, if our algorithm can recover the base-pairing matrix of more complex ensembles of alternative structures. We used a sequence that served as a starting point to design an effective thermoswitch (29). The sequence can fold into two alternative structures (a single hairpin or a two-stem structure). Folding with `RNAfold` at 37°C predicts that both alternatives are roughly equally probable in the ensemble (see base-pairing matrix labeled as "target ensemble" in Fig. 3). At low temperatures the single hairpin dominates. We asked if we can induce the mixed ensemble at low temperature by modifying the energy parameters using a perturbation vector. This represents a common situation where the experimental conditions such as temperature or salt concentration are different in the experiment and in the thermodynamic model.

First we tried the method from Deigan *et al.* (17) and set $\epsilon_i = b + m \ln[1 + q_i]$. For q_i we used the probability of being unpaired in the target ensemble at 37° and we set $m = 2.75$ and $b = -0.75$, a combination that generally worked well in our implementation and that is also close to the

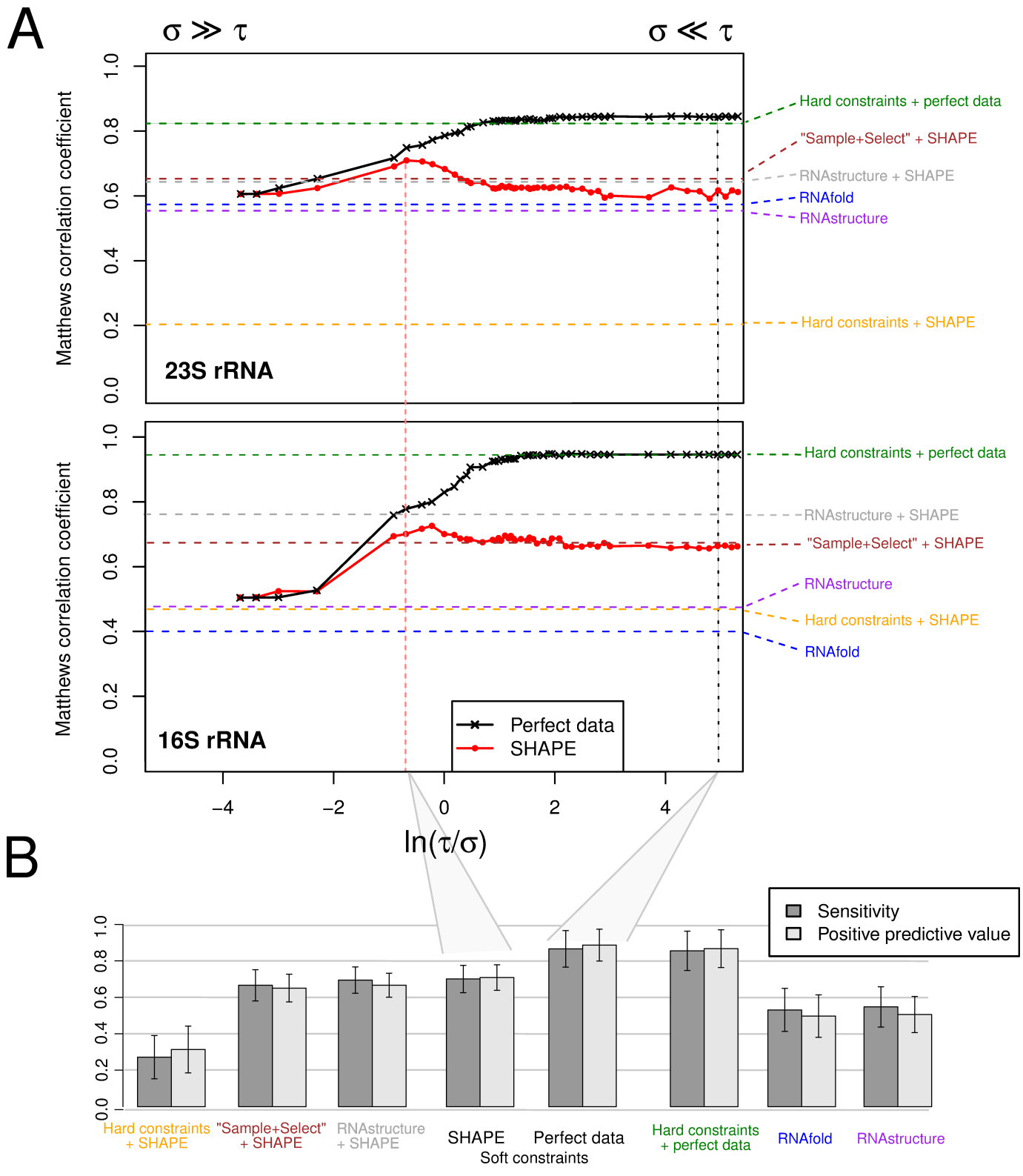

Figure 2. Structure prediction benchmark. (**A**) Prediction accuracy on 23S and 16S rRNAs as measured by the Matthews correlation coefficient (higher is better). Our iterative algorithm was run with different combinations of τ/σ on "perfect data" based on the reference structure and real SHAPE data. In comparison, the results of the "RNAstructure+SHAPE" from Deigan et al. (17) and "Sample+Select" from Quarrier et al. (11) are shown. As additional reference points, results from RNAfold with hard constraints and RNAfold/RNAstructure without any additional data is shown. (**B**) Sensitivity and positive predictive values averaged over all domains of the 23S and 16S rRNA are shown. For our algorithm ("RNAfold soft constraints") we used $\tau/\sigma = 200$ for perfect data and $\tau/\sigma = 0.5$ for the SHAPE data. The latter corresponds to the optimum found for the 23S rRNA in panel A. It was chosen to ensure a fair comparison to "RNAstructure+SHAPE" which was also run with parameters that were optimized for the 23S rRNA. Error bars show 95% confidence interval of the average.

Table 1. Optimization efficiency for RNAs of various length

		Exact gradient			Sampled gradient		
RNA	Length	No. iter.	Min. Rate.[a]	Time	No. iter.	Min. Rate[a]	Time
tRNA	78	15	0.99	23s	19	0.99	10s
5s rRNA	117	48	0.98	4m30s	57	0.98	49s
SRP RNA	301	104	0.98	5h8m1s	127	0.98	12m57s
23s rRNA (1)[b]	514	324	0.95	5d9h38m5s	104	0.94	43m55s
23s rRNA (2)[b]	686	136	0.96	6d22m7s	69	0.94	53m25s

Calculations were performed on six core AMD Opteron CPUs with 800 MHz

[a] Rate of minimization of the objective function after n iterations: $1 - D_1/D_n$

[b] Two sub-domains of the 23s rRNA were used

published parameters. Using this approach, the resulting base-pair matrix only shows one hairpin structure and not the expected ensemble of the two alternative structures (Fig. 3A). We also tried to systematically search for other parameter pairs m and b but also other combinations failed to recover the correct ensemble. We next used our iterative minimization algorithm and set the probability of being unpaired at 37° as our input vector q_i. Running our algorithm at 10° with σ^2=0.01 and τ^2=1.0, we could calculate a perturbation vector that gives exactly the expected results (Fig. 3A).

This simple example highlights a major advantage of the present approach over both hard constraints and simplistic bonus energies: since we consider the entire Boltzmann ensemble and model the observable experimental signal as a superposition of contributions from the individual members of the ensemble, we can also accommodate seemingly conflicting data that arise from different subsets of structures in the ensemble. The effect of our pseudo-energies is merely to distort the relative frequencies of structures within the ensemble.

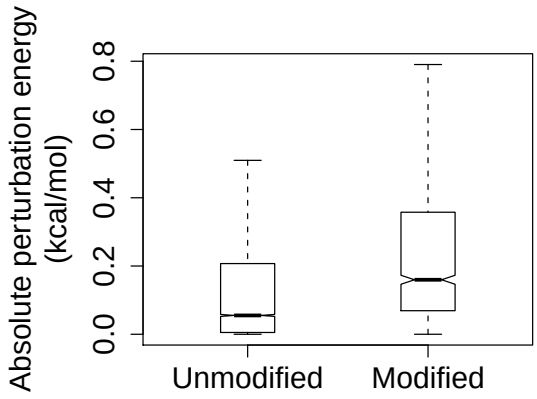

Figure 5. Distribution of perturbation energies for modified and non-modified nucleotides in tRNAs. 160 tRNAs with known modifications were forced to fold into the canonical tRNA structure and the values of the perturbation energies were analyzed. Non-zero perturbation energies indicate a discrepancy between the prediction under the standard energy model and the canonical structure.

Correlation of perturbation energies with nucleotide modifications

Another advantage of our algorithm is that it calculates position-specific perturbation energies that are non-zero only when they are required to reconcile the experimentally observed data with the energy model. The perturbations thus identify regions along the sequence where the energy model fails to accurately represent the observed data.

Chemically modified nucleotides are an important source of inaccuracies because they are not explicitly considered in the energy model. Such posttranscriptional modifications are common in several classes of noncoding RNAs. They are particularly well-studied for tRNAs (30, 31). Generally, tRNAs fold into the functional cloverleaf structure spontaneously *in vitro* without being modified (31). However, there is one well-known exception to this rule. The human mitochondrial tRNA-Lys was found to be misfolded in *vitro* while forming the canonical cloverleaf *in vivo*. One particular base methylation is sufficient to induce the correct folding also *in vitro* (32).

Theoretically, our approach should be able to identify nucleotides with modifications that influence their pairing behavior. In such a situation we expect a large perturbation energy localized at the modified nucleotide and possibly its pairing partner. We thus analyzed the behavior of the mito-tRNA-Lys *in silico*. Folding with `RNAfold` clearly does not result in the typical cloverleaf structure but rather yields an extended stem structure (Fig. 4). This is consistent with the *in vitro* results, which also did not show the canonical cloverleaf structure. We ran our algorithm on the sequence and imposed the cloverleaf structure as external constraint. The algorithm finds a minimum after 18 iterations and leads to a refold of the structure. The resulting perturbation vector shows two distinct peaks strongly suggesting that the base-pair stacks between positions 8,9 and 61,62 is the most critical for the molecule to fold into the correct structure. The high peak that suppresses this base-pair stack corresponds to the methylation that also was shown *in vitro* to be responsible for the refolding. It is important to note that a simple comparison of the misfolded prediction of the standard energy model to the reference structure will *not* give the same information (see the difference plot of p_i of the initial prediction and q_i of the reference structure in Fig. 4). Since the molecule undergoes re-folding the initial p_i of the misfolded structure is not informative and

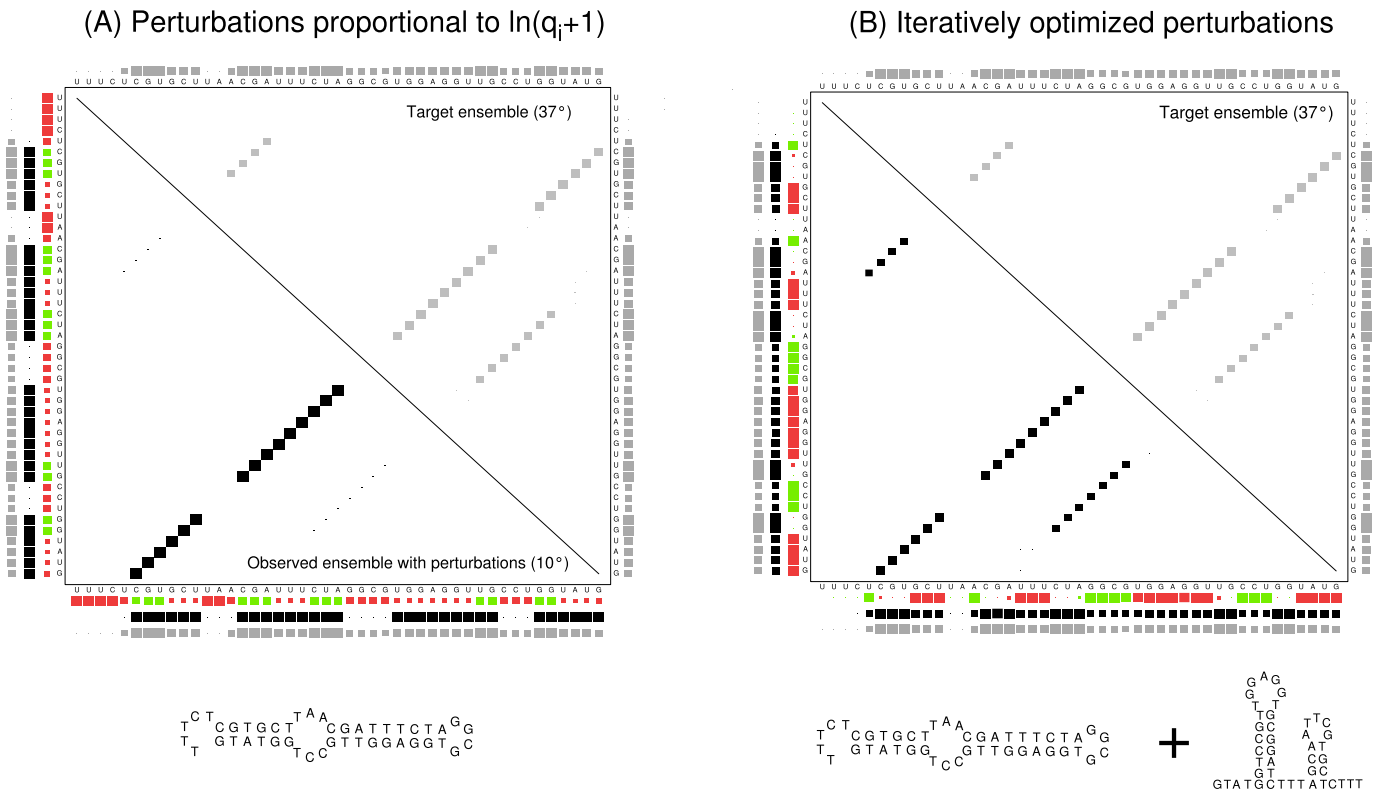

Figure 3. Recovering the correct structure ensemble of a bistable structure. A sequence that folds into a one and two-stem structure with equal probabilities at 37°C, folds predominantly into the one-stem structure at 10°C. Using the q_i vector of the probability of being unpaired of the ensemble at 37°C, the same bistable ensemble is attempted to be induced at 10°C. (**A**) If the perturbation vector is chosen proportional to the vector q_i ($\epsilon_i = b + m \ln[1+q_i]$), the correct solution cannot be found and the ensemble is still dominated by the one-stem structure. (**B**) Using the iterative optimization algorithm a perturbation vector can be found that recovers the bistable ensemble representing both the one-stem and two-stem structure.

only an iterative approach will identify the positions critical for the structure change.

We further asked if there is a general correlation of nucleotide modifications and perturbations calculated from our algorithm. To this end, we analyzed 160 tRNAs contained in the MODOMICS database (33) in exactly the same way as the human mito-tRNA-Lys example above. The MODOMICS database contains experimentally determined nucleotide modifications of various RNAs. Again, we used the canonical cloverleaf structure as the input of our algorithm. We found that the absolute value of the perturbations for modified bases (0.25 kcal/mol) is on average higher for modified bases than those for unmodified bases (0.17 kcal/mol). The difference (Fig. 5) is significant (Mann-Whitney test $p < 2 \times 10^{-16}$) and implies that discrepancies between the standard energy model and the canonical tRNA structure can be partly attributed to nucleotide modifications. However, we only found few candidates where a nucleotide modification seems to directly cause a complete refold. This confirms that the human mitochondrial tRNA-Lys described in the literature is an outstanding example and most other tRNAs fold into the cloverleaf shape spontaneously without modification (31).

Correlation of pseudo-energies and protein binding

RNA binding proteins are another reason that can cause differences between experimentally observed and thermodynamically predicted structures (34). The 5'-end of the sodB mRNA in *Escherichia coli* was found to change the structure upon binding of Hfq (35). Hfq acts as a chaperone and opens a region that forms a intermolecular interaction with the small RNA RyhB. We ran our algorithm applying the structure model proposed for the sodB mRNA by Geissmann *et al.* (35). We observed high energy perturbations in the second half of the analyzed region (Fig. 6) which corresponds exactly to the region that shows the protein-induced structure change.

DISCUSSION

The combination of thermodynamic folding and structure probing experiments is currently the standard method to establish secondary structure models. Probing experiments have seen rapid development over the past years leading to probing data for the complete HIV genome (18) and pilot studies of transcriptome wide probing in yeast (9) and mouse (10). Scaling the problem from individual RNAs to genome-wide data is not only an experimental challenge. The computational analysis of probing experiments to automatically generate reliable structure models seems

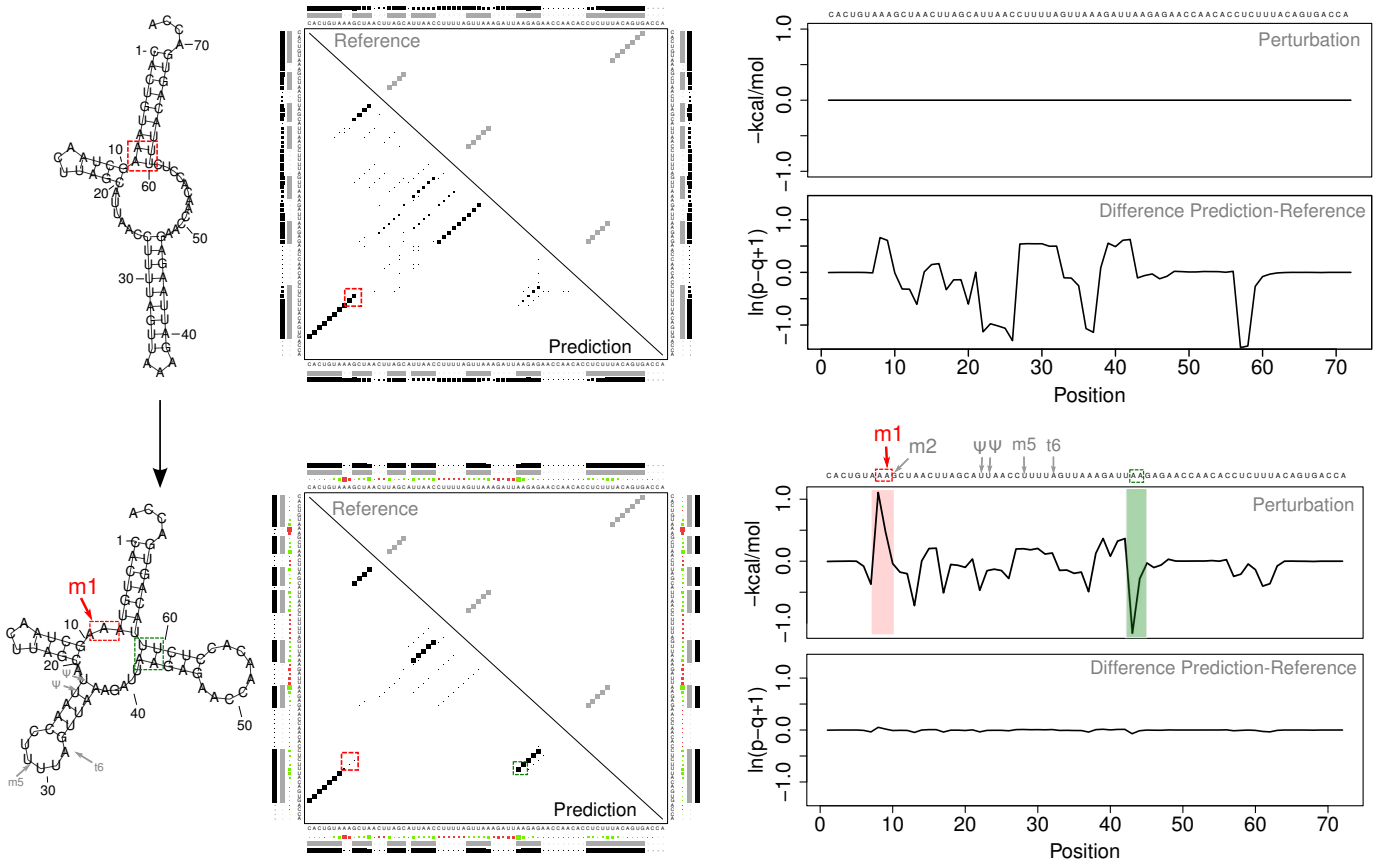

Figure 4. Perturbation energies correlate with nucleotide modifications in tRNAs. Human mitochondrial tRNA-Lys does not fold into the canonical cloverleaf using the standard energy model (top) but can be easily re-folded with perturbations calculated by our algorithm (bottom). The highest peak (favoring single strand formation, red) in the perturbation vector affects the same stack as the methylation of position 9 (red arrow) known to be necessary and sufficient for the correct folding *in vitro*. The lowest peak (favoring base-pairing, green) corresponds to the new base-pairing partners for the destroyed stack. Critical nucleotides for the re-folding are boxed. The dotplot coloring and annotation scheme is the same as in Figs. 1 and 3.

equally challenging. There are many steps involved and sophisticated methods to pre-process the data have been developed (8, 10). Here, we addressed the last step in this process, the actual folding step.

We proposed a novel way to incorporate experimental constraints into classical thermodynamic folding. Hard combinatorial constraints that have been used for long time only make sense when a model is manually built for an individual RNA but does not scale to automatic structure prediction from noisy data. Therefore, we introduced a "soft constraint" approach that is based on pseudo-energies that favor individual positions to be paired or unpaired. We formulated the problem using the partition function which offers the most flexible description of the thermodynamics properties of an RNA and allows for example to calculate pair probabilities or study suboptimal structures (19). Since previous pseudo-energy approaches cannot be easily applied in that case (see section "Rationale"), we introduced a formal framework to reconcile external constraints and thermodynamic predictions. In this framework, pseudo-energies have an interpretable meaning and the system shows some important properties such as the simple fact that in the case of experiment and thermodynamic model being in perfect agreement no pseudo-energies are applied. However, an iterative algorithm is required in practice to find the optimal pseudo-energies. We derived an analytic expression for the gradient of this optimization problem which allows for effective minimization.

We tested our method on a SHAPE data set of rRNAs that has been used for benchmarking previously. It provides of approximately 4000 probed positions (17) allowing for statistically relevant comparisons between methods. Unfortunately, similarly sized data sets are not available for other RNAs and it remains to be determined how our results generalize across various other classes of RNAs.

On the rRNA dataset, we found that our soft constraint approach with SHAPE data clearly improves structure prediction compared to normal thermodynamic folding. Varying the weight of the probing data used for the prediction identifies a maximum in accuracy, which, however, stays well below the best value theoretically possible with perfect data (Fig. 2). Although our algorithm performs well in this particular benchmark, it could not clearly outperform for example the much simpler method by Deigan *et al.* An important observation is that the difference between the observed and theoretically possible performance is much larger than the differences between the various methods. This

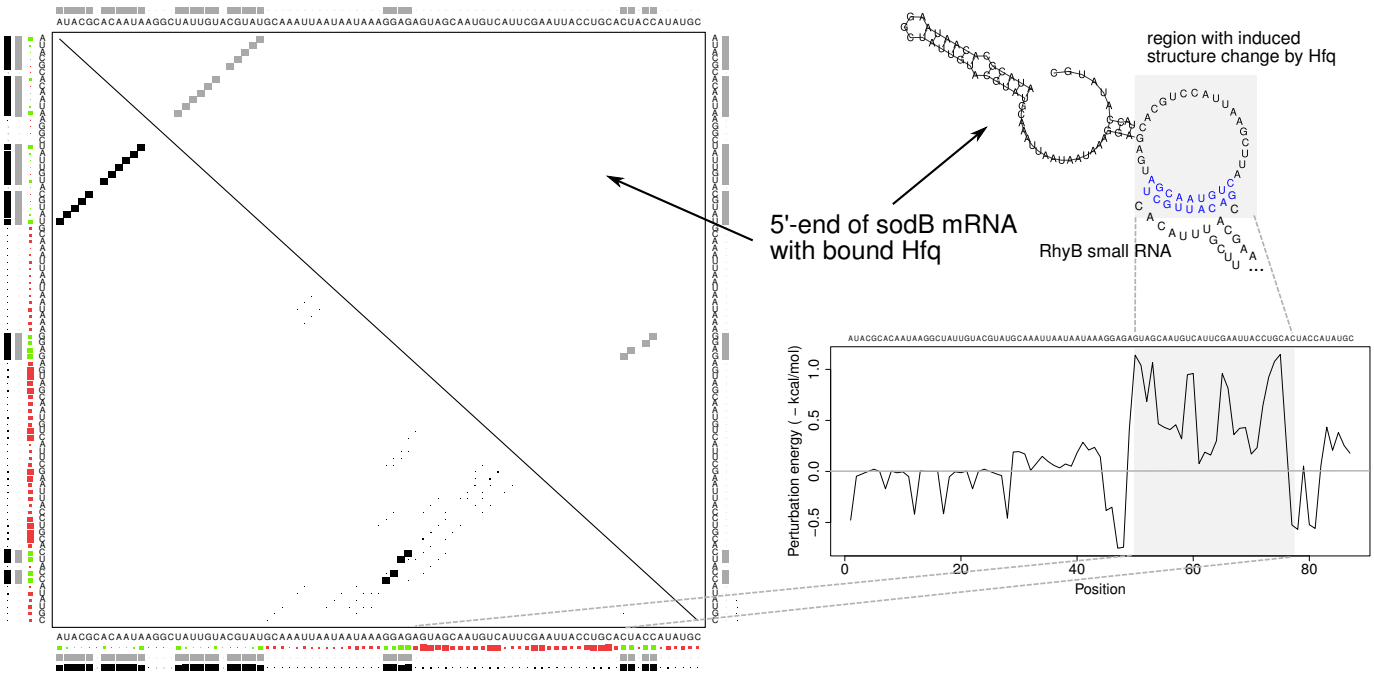

Figure 6. Energy perturbations correlate with Hfq induced structure changes in the sodB mRNA. A perturbation vector for the standard energy model was calculated to fit the experimentally established structure model by Geissmann *et al.* (35). The dotplot and color annotations are the same as in Figs. 1, 3 and 4.

suggests that substantial improvements cannot be achieved by improving the folding algorithm in a generic way but rather through more efficient noise filtering and pre-processing of the raw-data from the various experimental protocols. Although there is a clear correlation of SHAPE reactivities and pair probabilities, it is not straightforward to find a simple model to describe this relationship. The SHAPE reactivity measuring the local flexibility of a nucleotide seems to be dependent on the structural context, i.e. the type of loop (hairpin, bulge) and the position within the loop. It is also influenced by tertiary interactions. Systematic studies with different classes of RNAs will be necessary to understand this signal and the associated noise in more detail. Here, we used a simple tresholding method to convert the SHAPE reactivities into discrete states (paired or unpaired) as input for our algorithm.

We also studied the behavior of our algorithm on individual examples and found that it is capable of recovering the correct thermodynamic ensemble of a bistable RNA (29), identify the critical positions of nucleotide modifications required for correct *in vivo* folding of human mitochondrial tRNA-Lys (32) and the region of Hfq induced structure changes in sodB mRNA (35). These applications demonstrate the usefulness of our "soft constrained" partition function approach beyond pure structure prediction.

Finally, we have formulated the problem in a generic form such that the methodology presented in this paper is not limited to classical chemical or enzymatic probing data for individual positions. A new experimental procedure has been proposed that provides information on particular base-pairs on a short model RNA by extending classical probing with systematic mutation strategies (36). Our algorithm can be extended to any structural element for which the probability can be calculated from the partition function, including specific base-pairs which would allow one to analyze also the type of experiments presented by Kladwang *et al* (36).

ACKNOWLEDGEMENTS

This work was supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung [Schrödinger Fellowship J2966-B12 to S.W.]. We thank Michael Kertesz, Eran Segal and Howard Chang for helpful discussions, David Mathews for providing SHAPE data, Ronny Lorenz and Stephan Bernhart for help with the Vienna RNA package.

Conflict of interest statement. None declared.

REFERENCES

1. Amaral, P., Dinger, M., Mercer, T., and Mattick, J. (2008) The eukaryotic genome as an RNA machine. *Science*, **319**, 1787–9.
2. Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J., and Ehresmann, B. (1987) Probing the structure of RNAs in solution. *Nucleic Acids Res*, **15**, 9109–28.
3. Knapp, G. (1989) Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol*, **180**, 192–212.
4. Weeks, K. (2010) Advances in RNA structure analysis by chemical probing. *Curr Opin Struct Biol*, **20**, 295–304.
5. Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127**, 4223–4231.
6. Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A. (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res.*, **36**, e63.
7. Lucks, J., Mortimer, S., Trapnell, C., Luo, S., Aviran, S., Schroth, G., Pachter, L., Doudna, J., and Arkin, A. (Jun, 2011) Multiplexed RNA

structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proc Natl Acad Sci U S A,*.
8. Aviran, S., Trapnell, C., Lucks, J., Mortimer, S., Luo, S., Schroth, G., Doudna, J., Arkin, A., and Pachter, L. (Jun, 2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc Natl Acad Sci U S A,*.
9. Kertesz, M., Wan, Y., Mazor, E., Rinn, J., Nutter, R., Chang, H., and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–7.
10. Underwood, J., Uzilov, A., Katzman, S., Onodera, C., Mainzer, J., Mathews, D., Lowe, T., Salama, S., and Haussler, D. (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods*, **7**, 995–1001.
11. Quarrier, S., Martin, J., Davis-Neulander, L., Beauregard, A., and Laederach, A. (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*, **16**, 1108–17.
12. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406–15.
13. Reuter, J. and Mathews, D. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
14. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
15. Mathews, D., Sabina, J., Zuker, M., and Turner, D. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911–40.
16. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA*, **101**, 7287–7292.
17. Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci USA*, **106**, 97–102.
18. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess Jr, J. W., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, **460**, 711–716.
19. McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
20. Ding, Y. and Lawrence, C. E. (2003) A statistical sampling algorithm for rna secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.
21. Lu, Z., Gloor, J., and Mathews, D. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–13.
22. Do, C., Woods, D., and Batzoglou, S. (Jul, 2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–8.
23. Gardner, P. and Giegerich, R. (Sep, 2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **5**, 140.
24. Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M., and Giddings, M. C. (2008) ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, **14**, 1979–1990.
25. Garst, A. D., Edwards, A. L., and Batey, R. T. (2011) Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol.*, **3**, a003533.
26. Schultes, E. A. and Bartel, D. P. (2000) One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
27. Zhuang, X., Kim, H., Pereira, M. J. B., Babcock, H. P., Walter, N. G. W., and Chu, S. (2002) Correlating structural dynamics and function in single ribozyme molecules. *Science*, **296**, 1473–1476.
28. Huang, Z., Pei, W., Han, Y., Jayaseelan, S., Shekhtman, A., Shi, H., and Niu, L. (2009) One RNA aptamer sequence, two structures: a collaborating pair that inhibits AMPA receptors. *Nucleic Acids Res.*, **37**, 4022–4032.
29. Waldminghaus, T., Kortmann, J., Gesing, S., and Narberhaus, F. (2008) Generation of synthetic RNA-based thermosensors. *Biol Chem*, **389**, 1319–26.
30. Helm, M. (2006) Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res*, **34**, 721–33.
31. Motorin, Y. and Helm, M. (2010) tRNA stabilization by modified nucleotides. *Biochemistry*, **49**, 4934–44.
32. Helm, M., Brulé, H., Degoul, F., Cepanec, C., Leroux, J., Giegé, R., and Florentz, C. (1998) The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial tRNA. *Nucleic Acids Res*, **26**, 1636–43.
33. Czerwoniec, A., Dunin-Horkawicz, S., Purta, E., Kaminska, K., Kasprzak, J., Bujnicki, J., Grosjean, H., and Rother, K. (2009) MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res*, **37**, D118–21.
34. Mayer, O., Mayer, O., Windbichler, N., Wank, H., and Schroeder, R. Protein-Induced RNA Switches in Nature pp. 75–86 Springer (2006).
35. Geissmann, T. and Touati, D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*, **23**, 396–405.
36. Kladwang, W., Cordero, P., and Das, R. (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA*, **17**, 522–34.