BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i19–i28
doi:10.1093/bioinformatics/btm223

Efficient parameter estimation for RNA secondary structure prediction

Mirela Andronescu[1,*], Anne Condon[1], Holger H. Hoos[1], David H. Mathews[2] and Kevin P. Murphy[1]

[1]Department of Computer Science, University of British Columbia, Vancouver BC V6T 1Z4, Canada and [2]Department of Biochemistry & Biophysics and Department of Biostatistics & Computational Biology, University of Rochester Medical Center, Rochester NY 14642, USA

ABSTRACT

Motivation: Accurate prediction of RNA secondary structure from the base sequence is an unsolved computational challenge. The accuracy of predictions made by free energy minimization is limited by the quality of the energy parameters in the underlying free energy model. The most widely used model, the Turner99 model, has hundreds of parameters, and so a robust parameter estimation scheme should efficiently handle large data sets with thousands of structures. Moreover, the estimation scheme should also be trained using available experimental free energy data in addition to structural data.

Results: In this work, we present constraint generation (CG), the first computational approach to RNA free energy parameter estimation that can be efficiently trained on large sets of structural as well as thermodynamic data. Our CG approach employs a novel iterative scheme, whereby the energy values are first computed as the solution to a constrained optimization problem. Then the newly computed energy parameters are used to update the constraints on the optimization function, so as to better optimize the energy parameters in the next iteration. Using our method on biologically sound data, we obtain revised parameters for the Turner99 energy model. We show that by using our new parameters, we obtain significant improvements in prediction accuracy over current state of-the-art methods.

Availability: Our CG implementation is available at http://www.rnasoft.ca/CG/

Contact: andrones@cs.ubc.ca

1 INTRODUCTION

RNA molecules play essential roles in living cells. Many important and diverse functions of RNA molecules, including catalysis of chemical reactions and control of gene expression, have only recently come to light. Outside of the cell, novel nucleic acids have been selected using directed molecular evolution techniques *in vitro*, which can function as enzymes or aptamers with high binding specificity for target proteins (Breaker, 2002), with medical diagnostic or biosensing applications (Benenson *et al.*, 2004; Dirks and Pierce, 2004).

*To whom correspondence should be addressed.

Because of the importance of RNA molecules, and because structure is key to the function of RNA molecules in their diverse roles, there is a need to improve the accuracy of computational predictions of RNA structure from the base sequence. RNA tertiary structure is difficult to predict, but is significantly constrained by secondary structure (Tinoco and Bustamante, 1999) — i.e. the set of base pairs that forms when the molecule folds (see Fig. 1 for an example). Therefore, current RNA structure prediction methods are mostly focused on secondary structure. Given a sequence, the goal is to predict the structure with minimum free energy (MFE), relative to its unfolded state. There is considerable evidence that RNA secondary structures do indeed adopt their MFE configurations in their natural environments (Tinoco and Bustamante, 1999), and that in many cases these structures are pseudoknot-free (i.e. contain only hierarchically nested base-pairs).

Most models assume that the free energy of sequence x and structure y is given by an equation of the form

$$\Delta G(x, y, \boldsymbol{\theta}) = \mathbf{c}(x, y)^{\top}\boldsymbol{\theta} = \sum_{k=1}^{K} c_k(x,y)\theta_k \qquad (1)$$

where K is the number of features, $c_k(x, y)$ is the number of times feature k occurs in secondary structure y of sequence x, and θ_k is a parameter modelling the energy contribution of each occurrence of feature k. In this article, we use the features proposed by Mathews *et al.* (1999), which are widely accepted as biologically realistic, and are used in several software packages such as Mfold (Zuker, 2003), RNAstructure (Mathews, 2004), the Vienna RNA package (Hofacker *et al.*, 1994) and SimFold (Andronescu, 2003). We shall call this the Turner99 model. We will explain these features in more detail in Section 2 (see Fig. 1 for some examples).

Given a set of features, we are faced with the problem of estimating the model parameters $\boldsymbol{\theta}$—this is the focus of this work. Suppose we have a data set S consisting of a set of (x, y_x) pairs, where y_x is the true MFE structure of sequence x (as determined using trusted and highly accurate methods). We created such a data set using databases of known RNA structures (Cannone *et al.*, 2002; Sprinzl and Vassilenko, 2005, and other databases). One approach would be to estimate the parameter vector $\boldsymbol{\theta}$ that maximizes the likelihood of S, as used in the CONTRAfold algorithm (Do *et al.*, 2006). However, there are several problems with this approach. First, it is very

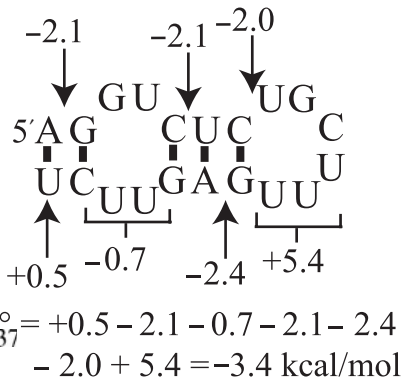

Fig. 1. Secondary structure of an RNA strand of length 20. An RNA molecule, or strand, is a sequence of Adenine (A), Cytosine (C), Guanine (G) or Uracil (U) bases, with two chemically distinct ends, known as the 5′ and 3′ ends. The secondary structure is the set of base pairs (indicated by black boxes) that form when the molecule folds, under fixed environmental conditions. Throughout, we consider only pseudoknot-free secondary structures. The base pairs give rise to *loops*. The depicted structure includes a hairpin loop (right end of diagram), as well as three base pair stacks and a 2×2 internal loop. In the Turner99 model, the total free energy change of a structure, relative to its unfolded state, is the sum of the free energy changes of its loops. The lower the free energy change, the more stable the structure. Generally, stacked base pairs tend to stabilize the RNA structure, whereas loops with unpaired bases are destabilizing. In the depicted structure, contributions to the total free energy change at 37°C, denoted by ΔG^0_{37} (measured in kcal/mol), include a +5.4 penalty for closing the hairpin loop, which is largely an entropic cost, a −2.4 favourable term for the rightmost (UA/CG) stacked pair, and a +0.5 penalty for an AU pair at the end of a helix (as well as other terms).

slow, which prevents us from applying it to large training sets. (For example, it took more than 80 h on a single reference processor to train CONTRAfold on 190 sequences of average length 100. However, a much larger training set is needed for accurate parameter estimation.) Second, it does not handle the fact that there may be label noise in the training set, i.e. y_x may not actually be the MFE structure for x, since the feature set is not perfect, and the structures may not be perfectly annotated.

We propose a novel algorithm that overcomes both of these problems: it is very fast (less than 20 min to train on 190 sequences of length 100), thus letting us train on large data sets, and it is robust to label noise. We show that the parameters learned using our algorithm yield 7% better prediction accuracy (as determined using the F-measure on base pairs) than the standard Turner99 parameters, and 5% better accuracy than the CONTRAfold predictions, when measured on a large structural data set.

In addition to predicting the secondary structure, to be of biological interest, a model must also accurately predict the free energy changes for structure formation. We therefore collected a second data set, the thermodynamic set T, comprised of triples (x, y_x, e_x), where x is an RNA sequence, y_x is the MFE secondary structure of x, and e_x is the free energy of structure y_x for sequence x, measured within some small experimental error. We compiled this data set from the results of thermodynamic experiments (Mathews *et al*., 1999, 2004; Xia *et al*., 1998). Not surprisingly, we find that our ability to accurately predict free energies is enhanced when we also train using T . Note that in contrast the scores produced by CONTRAfold have no intrinsic biological meaning.

2 THE TURNER99 MODEL

Turner and co-workers derived and refined an energy model, which we call the Turner99 model, over a period of more than two decades (Mathews *et al*., 1999; Xia *et al*., 1998). The model pertains to free energy changes at 37°C. Further refinements to the parameters were made by Mathews *et al*. (2004), based on new experimental data. The Turner99 features were carefully chosen to balance the goals of accurately modelling physical principles, and of ensuring that the resulting optimization problem of finding the MFE structure can be solved efficiently (using dynamic programming, in $O(n^3)$ time, where n is the sequence length). Some Turner99 free energy parameters were determined using reliable wet-lab experiments, while others were estimated from known structural data. However, estimation of parameter values was done in stages, with some values being fixed before others were determined, and parameter estimation did not take advantage of the large body of structural information available today. The Turner99 model achieves an average prediction accuracy (sensitivity) of 73% on a large set of biological RNAs of length shorter than 700 nucleotides with known secondary structures (Mathews *et al*., 1999).

The model features capture all types of stacked base pairs as well as loops, including hairpin loops, internal loops and multiloops. Non-canonical base pairs (i.e. base pairs other than CG, AU and GU) are not explicitly predicted; however, parameter values for internal loops do implicitly account for bonds between noncanonical base pairs. For larger loops, features include the number of branches, number of unpaired bases between branches, the closing base pairs and unpaired ('dangling') bases next to them. Thus, there are one or more features associated with each loop, as illustrated in Fig. 1.

Overall, the Turner99 model has tabulated energy values for about 7600 features; most of these can be determined by applying simple extrapolation rules to 363 free parameters. For computational efficiency, in this study, we assume the 3′ dangling end parameter values, used for multiloops and exterior loops (Mathews *et al*., 1999), are always lower than the respective values for 5′ dangling ends. To find improved values for the set of 363 free parameters is the goal of our work presented in the following.

3 PARAMETER ESTIMATION

Having defined the set of features, we now discuss some techniques for parameter estimation.

3.1 Maximum likelihood (ML) method

An obvious approach to parameter estimation is to use the maximum likelihood (ML) method, as in the CONTRAfold algorithm of Do *et al.* (2006). Specifically, we define the probability of an RNA structure y, given an RNA sequence x and parameter vector $\boldsymbol{\theta}$, using a conditional log-linear model (Boltzmann distribution) as follows:

$$p(y|x, \boldsymbol{\theta}) = \frac{1}{Z(x, \boldsymbol{\theta})} \exp\left(-\frac{1}{RT}\Delta G(x, y, \boldsymbol{\theta})\right).$$

Here, R is the gas constant, T is the absolute temperature, and $Z(x, \boldsymbol{\theta})$ is the partition function (McCaskill, 1990).

It is well known that $p(y|x,\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$ [see e.g. Lafferty *et al.* (2001)], and hence we can find the globally optimal parameter estimate of the log likelihood function $L_S(\theta) = \sum_{(x,y,x)\in S} \log p(y_x|x, \boldsymbol{\theta})$ using a gradient-based optimizer, such as the Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm, provided we can efficiently compute Z. Since we disallow pseudoknots, we can compute Z and the gradient of Z in $O(n^3)$ time using dynamic programming (McCaskill, 1990), where n is the length of x.

We can consider the thermodynamic set T as prior knowledge by assuming the observed energies e_x are noisy versions of the true energies. We can model this with a Gaussian distribution with precision τ and compute the maximum a posteriori (MAP) estimate of the posterior distribution $p(\boldsymbol{\theta}|S, T)$:

$$p(\boldsymbol{\theta}|\mathcal{S}, \mathcal{T}) \propto L_{\mathcal{S}}(\boldsymbol{\theta}) + \tau \sum_{(x,y_x,e_x)\in\mathcal{T}} (e_x - c(x, y_x)^T\boldsymbol{\theta})^2.$$

We implemented the objective function and its gradient in C++, and optimized it using an unconstrained and unbounded Matlab LBFGS implementation. Since our model assumes constraints on 48 parameters (namely dangling end parameters), in our current implementation we fix these values to the Turner99 values. A non-linear constrained optimization software would be needed to optimize for all 363 parameters.

However, in practice there are problems with using the ML approach (with or without prior). First, the method is computationally expensive, because evaluating the objective function and its gradient is slow, and this needs to be done many times. (For example, CONTRAfold took more than 80 h to train on a small set of 190 sequences, and our own implementation of ML took about 66 h on the same data.) Second, this approach does not gracefully handle the case where there is no parameter vector $\boldsymbol{\theta}$ such that y_x is the MFE structure for x with respect to $\boldsymbol{\theta}$ for all (x, y_x) in the structural set. This case can arise for two reasons: the feature set is not likely to be perfect, and the structures may not be perfectly annotated.

3.2 Constraint generation (CG) approach

An alternative approach to parameter estimation is to find a solution $\boldsymbol{\theta}$ for a system of constraints

$$\Delta G(x, y_x, \boldsymbol{\theta}) < \Delta G(x, y, \boldsymbol{\theta}),$$

where $(x, y_x) \in S$ and $y \in Y_x \setminus \{y_x\}$, and Y_x is the set of all secondary structures for sequence x; these constraints ensure that for each sequence x all non-optimal secondary structures y of sequence x have higher energy than the MFE structure y_x. (Throughout we assume there is no other structure which has the same MFE as the known structure, and thus use strict inequalities. This can be relaxed to non-strict inequalities.)

3.2.1 Handling infeasible constraints. Due to inaccuracies in the given MFE structures y_x (label noise) or inherent limitations of the given feature set, it may happen that this system of constraints is infeasible, i.e. no solution θ exists that satisfies all constraints simultaneously. To deal with infeasibility, we introduce slack variables $\delta_{x,y} \geq 0$ into the constraints, whose values are then minimized; this leads to relaxed constraints of the form:

$$\Delta G(x, y_x, \boldsymbol{\theta}) < \Delta G(x, y, \boldsymbol{\theta}) + \delta_{x,y}.$$

Considering the definition of the energy function ΔG (see Equation 1), these structural constraints can be expressed as a system of linear inequalities

$$(\mathbf{c}(x, y_x) - \mathbf{c}(x, y))^\top\boldsymbol{\theta} - \delta_{x,y} < 0$$

for all $(x, y_x) \in S$ and $y \in Y_x \setminus \{y_x\}$. This can be written more compactly in matrix form as

$$M_S\,\boldsymbol{\theta} - \boldsymbol{\delta} < 0$$

where each row of the matrix M_S is $(\mathbf{c}(x, y_x) - \mathbf{c}(x, y))^\top$ for some $(x, y_x) \in S$ and some $y \in Y_x \setminus \{y_x\}$, and $\boldsymbol{\delta}$ is the vector of slack values $\delta_{x,y}$. (The rows of M_S and the elements of δ are ordered consistently.)

This leads to the following formulation as a constrained optimization problem:

$$\begin{aligned} &\text{minimize } ||\delta||^2 \\ &\text{subject to} \\ &M_S\boldsymbol{\theta} - \boldsymbol{\delta} < 0 \\ &\boldsymbol{\delta} \geq 0. \end{aligned} \tag{2}$$

where $\|\boldsymbol{\delta}\|$ is the L2-norm of $\boldsymbol{\delta}$. (This system can get quite large, and we explain below how to address this issue.)

This is similar to the large margin approach proposed by Taskar *et al.* (2005) for learning connectivity parameters for disulfide bonds in protein structures. However, it is not quite the same. For our problem, we do not want to force a large distance between the known RNA secondary structures and other secondary structures. Our parameters are meant to have physical meaning, and there is evidence that there can be many low-energy folds of an RNA molecule that have energy close to the MFE (Uhlenbeck, 1995). Thus, margin approaches are not directly applicable to our problem.

3.2.2 Incorporating thermodynamic data. We incorporate the thermodynamic data by adding the following additional constraints:

$$\Delta G(x, y_x, \boldsymbol{\theta}) - \xi_x = \mathbf{c}(x, y_x)^\top\boldsymbol{\theta} - \xi_x = e_x. \tag{3}$$

where ξ_x is the error in predicting e_x. Again we can write this in vector form as

$$M_T\boldsymbol{\theta} - \boldsymbol{\xi} = \mathbf{e}$$

where each row of the matrix M_T is $\mathbf{c}(x, y_x)^\top$ for some $(x, y_x, e_x) \in T$. This leads to the following constrained optimization problem:

$$\begin{aligned} &\text{minimize } (1-\lambda)\cdot\frac{1}{|\mathcal{S}|}\|\mathrm{m}^\mathrm{T}\boldsymbol{\delta}\|^2 + \lambda\cdot\frac{1}{|\mathcal{T}|}\|\boldsymbol{\xi}\|^2 \\ &\text{subject to} \\ &M_S\boldsymbol{\theta} - \boldsymbol{\delta} < 0 \\ &M_T\boldsymbol{\theta} - \boldsymbol{\xi} = \mathbf{e} \\ &\delta \geq 0. \end{aligned} \tag{4}$$

where $|S|$ denotes the number of sequences in set S, m_x is 1 divided by the number of constraints in M_S for sequence x, and $\mathbf{m}$ is a vector of m_x.

The parameter λ controls the relative importance of T and S. The two extreme cases are: $\lambda = 0$, which means that we do not consider the thermodynamic set at all; and $\lambda = 1$, which causes those parameters which appear in the thermodynamic set to be fixed to the values which best fit the thermodynamic set, and the other parameters are unconstrained. Fig. 2 gives a schematic representation of T and S, and Fig. 3 motivates the use of inequality constraints.

One problem with the above objective is that if a certain feature does not occur in S or T, or if it appears only very few times, its corresponding parameter can become unbounded in magnitude. We therefore add an additional constraint that $\boldsymbol{\theta}$ should be bounded by the Turner99 parameters, plus or minus B kcal/mol, where we assume B is given to the algorithm. If the structural training data contains all features, we can even set B to infinity; however, in practice, a large value, such as 10 kcal/mol, should suffice. These bounds can be seen as the strength of a prior on the values of the Turner99 parameters.

3.2.3 Sequential CG algorithm. We have a quadratic objective subject to linear equality and inequality constraints, so we can find the global optimum. Unfortunately, the number of constraints can grow exponentially with the size of the input, since for each (x, y_x) in the structural data set S, there may be exponentially many structures in Y_x (Wuchty *et al.*, 1999). To circumvent this problem, we propose the following heuristic algorithm, similar to the cutting plane algorithm used by Tsochantaridis *et al.* (2005). The main idea is to iteratively estimate $\boldsymbol{\theta}$ using constraints $M_S\boldsymbol{\theta} - \boldsymbol{\delta} < 0$ for a matrix M_S that only includes rows for a manageable subset of sequences x and structures y.

Specifically, starting from an empty set of structures and the Turner99 parameters, in each iteration of our algorithm, for each sequence x from S, we predict its MFE structure using the current parameter vector θ and add the constraint

$$(\mathbf{c}(x, y_x) - \mathbf{c}(x, y'))^\mathrm{T}\boldsymbol{\theta}^{(i)} - \boldsymbol{\delta}^{(i)}_{x,y'} < 0,$$

where $y' \in Y_x$ is the MFE structure of x predicted using the parameter vector $\boldsymbol{\theta}^{(i-1)}$ from the previous iteration; this constraint enforces that the true structure y_x has lower energy (by margin $\delta^{(i)}_{x,y'}$) than the predicted structure y'. To avoid vacuous and redundant constraints, we never add constraints if $y' = y_x$ or if the new constraint is already in the system.

The intuition behind this sequential CG method is that most of the exponentially many constraints will not be active, since

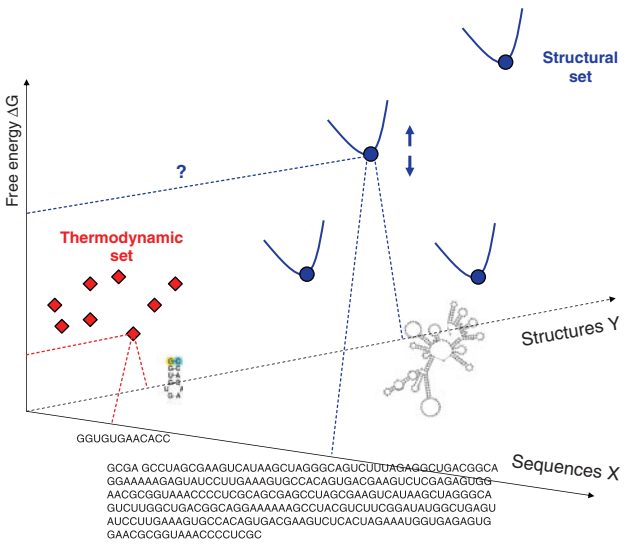

Fig. 2. Schematic representation of the structural and thermodynamic data sets we use in our CG algorithm. The X and Y axes represent RNA sequences and secondary structures, respectively. The diamonds on the left represent (x, y_x, e_x) triples that form the thermodynamic set, while the dots on the right represent (x, y_x) pairs forming the structural set. The curves depict the fact that the known y_x structures from the structural set have lower free energy change than any other structure into which x can fold, although we do not know where these points are situated on the vertical free energy axis.

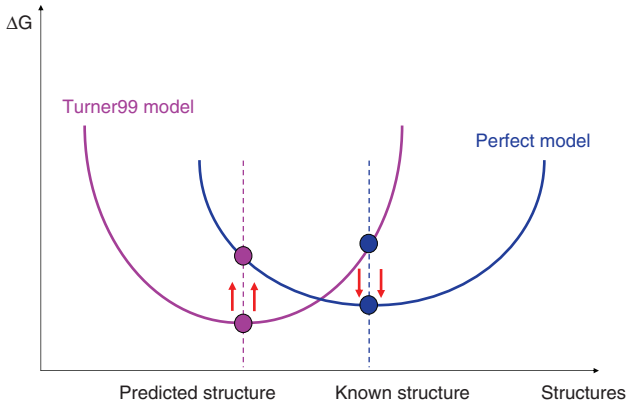

Fig. 3. Depiction of the motivation for the use of inequality constraints for a given sequence x. Secondary structures for x are represented on the X axis, and free energy changes on the Y axis. The left curve represents the free energy curve under the Turner99 model, which, when the prediction is incorrect, assigns a higher free energy to the known secondary structure than to the predicted secondary structure, although in the ideal model it should be lower (right curve). We wish to modify the parameters $\boldsymbol{\theta}$ so as to push up the free energy of the incorrectly predicted secondary structures (and of other structures), and to pull down the free energy of the known secondary structures.

procedure *CG* $(\mathcal{S}, \mathcal{T}, \lambda, B, K)$
  input: structural training set $\mathcal{S}$, thermodynamic set $\mathcal{T}$,
    parameter λ, bounds parameter B, number of iterations K;
  output: thermodynamic parameter vector θ^*, accuracy q^*;

  set $\theta^{(0)}$ to the Turner99 parameters;
  $\theta := \theta^{(0)}$; $M_S := [\,]$;
  $\theta^* := \theta$; $q^* := 0$;
  for $i := 1$ **to** K **do**
    for each $x \in \mathcal{S}$ **do**
      predict MFE structure y' of x using θ;
      add row $(\mathbf{c}(x, y_x) - \mathbf{c}(x, y'))^\top$ to M_S;
    end for;
    obtain new θ, ξ, δ by minimizing
      $(1-\lambda) \cdot \frac{1}{|\mathcal{S}|} \left|\left|\mathbf{m}^\top \delta\right|\right|^2 + \lambda \cdot \frac{1}{|\mathcal{T}|} ||\xi||^2$
      subject to
        $M_S\theta - \delta \leq 0$,
        $M_T\theta - \xi = \mathbf{e}$,
        $\delta \geq 0$,
        $\theta^{(0)} - B \leq \theta \leq \theta^{(0)} + B$;
    $q :=$ prediction accuracy obtained by using parameters θ on $\mathcal{S}$;
    if $(q > q^*)$ **then**
      $q^* := q$; $\theta^* := \theta$;
    end if;
  end for;
  return (θ^*, q^*);
end *CG*.

Fig. 4. Outline of the CG algorithm for RNA energy parameter optimization.

they refer to structures that are energetically very unfavourable. Assuming we start with a reasonable set of initial parameter values (here the Turner99 parameters), we can generate structures with more plausible (low) energies and effectively use constraints based on this much smaller set. The algorithm returns the θ values which give the best prediction accuracy on the training set. Fig. 4 summarizes our constraint generation algorithm, CG.

All secondary structure predictions are done using our SimFold software (Andronescu, 2003). Like the widely known Mfold algorithm (Zuker, 2003) and the RNAfold procedure from the Vienna RNA package (Hofacker *et al.*, 1994), SimFold is based on Zuker and Stiegler's dynamic programming algorithm and consequently has time complexity $O(n^3)$ and space complexity $O(n^2)$, where n is the sequence length. The constraint optimization problems are solved with ILOG CPLEX 9.1.

4 DATA SETS

In order to assess the improvement in prediction accuracy that can be achieved using our approach, we collected a large amount of structural and thermodynamic data. This data is summarized in Table 1.

The thermodynamic training set, T-Full, contains optical melting experimental data that we collected from 39 research papers, referenced by Mathews *et al.* (2004, 1999).

Table 1. Structural and thermodynamic sets

Set name	No. mols.	Avg. length	Used for
T-Full	946	17 ± 7	Training
T-Single	207	14 ± 4	Test
S-Processed	3439	178 ± 179	Training
S-Full	1660	295 ± 508	Test
S-151Rfam	151	136 ± 102	Training
S-A1	190	105 ± 28	Training
S-A5	836	105 ± 28	Training
S-A10	1531	103 ± 29	Training
S-A1′	193	106 ± 29	Test

We use structural sets and one thermodynamic set for training. For testing, we use one comprehensive structural set and one small thermodynamic set. In addition, we use three artificially created structural sets for training and one for testing.

Out of the 946 experiments, 739 are on RNA duplexes, which CONTRAfold cannot currently take as input for prediction. We therefore created a test set, T-Single, which contains the remaining 207 experimental results for single sequences.

The structural test set, S-Full, is a comprehensive RNA structural set that we assembled from databases of well-determined RNA secondary structures. Table 4 shows the RNA families included in this set, with their sizes and lengths. Several preprocessing steps have been applied, including removal of RNAs for archeae (which live in extreme environments), unannotated loops or unknown nucleotides. Non-canonical base pairs and a minimal number of bases to resolve any pseudoknots have been removed.

The training set, S-Processed, is similar to S-Full, but molecules longer than 700 nucleotides have been divided into shorter sequences, so that the MFE structure prediction step is reasonably fast. Unannotated branches or branches containing unknown base pairs have been truncated. For truncated structures, a *restriction string* that restricts the cut ends to pair has been added; of these structures, 66% have been included in S-Processed.

In addition to the above data sets that we collected, we used the structural set of Do *et al.*, which we call S-151Rfam. This contains one sequence-structure pair from each of 151 Rfam families collected from published papers. We have not included all of these families in S-Full because many of the structures have been predicted in the corresponding published papers (as opposed to measured), and are not biologically reliable.

Note that in biological data many features do not occur at all (see Fig. 5), making it hard to assess the potential for CG to estimate parameters for these features. Moreover, since we do not know what is the best accuracy achievable using the Turner99 feature set, even with a data set that covers all features we cannot know whether CG has found the best possible parameter values. For these reasons, we also created artificial data sets, generated by randomly choosing sequences

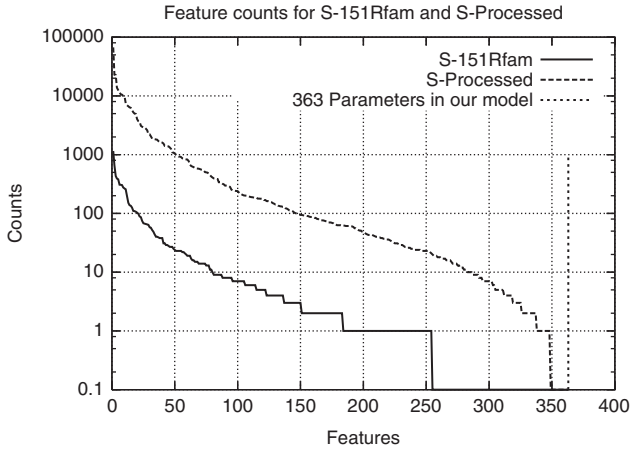

Fig. 5. The feature counts for various structural sets (features are ordered according to decreasing counts). Out of the 363 total parameters, only 254 appear at least once in S-151Rfam, and 348 appear at least once in S-Processed. The thermodynamic set T-Full contains 203 features out of all 363 features in the model.

x and then setting y_x to be the MFE secondary structure predicted using the Turner99 parameters. On this artificial data, we know that there exists a parameter setting (namely the Turner99 parameters) which gives perfectly accurate predictions. We sampled the data such that each feature occurs at least k times, for $k = 1, 5, 10$. (Six of the features are very unlikely to occur in MFE structures, and thus we fixed their parameter values to the Turner99 values). We call these sets S-A1, S-A5 and S-A10, and we call k the *feature coverage* of the set. We then checked that we could recover the Turner99 parameters using these training sets. We also measured performance on an artificial test set, S-A1′, which was obtained in exactly the same way as S-A1, but using a different random seed.

5 EXPERIMENTAL RESULTS

In this section, we report on several aspects of the performance of our CG method. First, using our artificially generated training sets, we show that CG runs much faster than CONTRAfold or ML; this is significant, because as a consequence, CG can be run on much larger training sets, for which running CONTRAfold or ML would be practically infeasible. Our analysis also indicates that CG can indeed find parameters that result in near-perfect predictions, when such parameters exist, and when the feature count is sufficiently high (10 for our artificial data). Next, we compare the accuracy of CG and CONTRAfold, when CG is trained on the S-151Rfam training set of Do *et al.*, both with and without the thermodynamic training set. While CG gives poor predictions when the thermodynamic set is not included, it matches or exceeds the prediction accuracy of CONTRAfold when the thermodynamic set is also included in training. Finally, we train CG on our large training set, S-Processed, and evaluate the accuracy of CG on our full structural data set, S-Full. We find that the parameter set found by CG achieves accuracy 7% better than that obtained with the Turner99 parameter set, and 5% better than that obtained by CONTRAfold. Following definitions of our accuracy measures, we first present our results on artificial data and then on biological data.

5.1 Performance measures

We use sensitivity and positive predictive value (PPV) as measures of structural prediction accuracy; a third measure, the F-measure (in short F), combines both sensitivity and PPV:

$$\text{Sensitivity} = \frac{\text{number of correctly predicted base pairs}}{\text{number of true base pairs}}$$

$$\text{PPV} = \frac{\text{number of correctly predicted base pairs}}{\text{number of predicted base pairs}}$$

$$\text{F-measure} = \frac{2 \times \text{sensitivity} \times \text{PPV}}{\text{sensitivity} + \text{PPV}}$$

Do *et al.* (2006) introduced a parameter called γ as a way to trade off sensitivity against PPV using their prediction algorithm. They found that setting $\gamma = 6$ gave the best overall performance. We could obtain a similar trade-off by computing the base pair probabilities and thresholding them, following Mathews (2004). However, in this work, we focus on MFE structure prediction, which does not support this trade-off.

5.2 Results on artificial data

In this section, we report on our runtime analysis, which we did primarily using our artificially generated sets. We then assess whether the CG method can robustly find an optimal parameter vector θ when one exists. Finally, we evaluate the sensitivity of the CG method to the feature count of the artificial training data.

5.2.1 Runtime comparison. We measured the run time of CG and CONTRAfold when trained on the artificial structural set S-A1, using a 2.4GHz Intel Xeon CPU with 512 KB cache size and 1GB RAM, running Linux 2.6.16 (SUSE 10.1). For CG training, we perturbed the Turner99 parameters by a number chosen uniformly at random between 0 and 1 kcal/mol, and we used this set as the initial set of parameters. The F-measures of this initial set are: 0.45, 0.42, 0.45 for S-A1, S-A5 and S-A10, respectively, and 0.43 for the test set S-A1′.

As Table 2 shows, when trained on S-A1, having 190 structures, CG took 4 min with $B = 1$, and 19 min with $B = 10$, whereas CONTRAfold took more than 80 h. Our ML implementation took 66 h.

Thus, CG is more than two orders of magnitude faster than conditional ML methods on our artificial data. On the artificial sets, CG always converges within 23 iterations. When trained on larger artificial sets, such as S-A5 and S-A10, CG's runtime was within 2 and 4 h on a single processor.

For the remaining experiments we parallelized the prediction step and ran it on 20 similar processors. When trained on S-151Rfam, the total runtime of CG was within 4 h, while the total runtime of ML was within 3 days. When trained

Table 2. Results when training on artificial data sets

Alg. and options	Set train	Train F-measure	Test (S-A1′) F-measure	Number iterations	Runtime
CG $B=1$	S-A1	1.00	0.90	9	4m
CG $B=10$	S-A1	1.00	0.80	23	19m
CG $B=1$	S-A5	1.00	0.96	9	24m
CG $B=10$	S-A5	1.00	0.95	13	1h35m
CG $B=1$	S-A10	1.00	0.98	9	49m
CG $B=10$	S-A10	1.00	0.98	13	4h
ML	S-A1	0.94	0.77	–	66h
CF $\gamma=6$	S-A1	0.83	0.64	–	> 80h

CG refers to constraint generation, CF refers to CONTRAfold [where we set $\gamma=6$, as recommended by Do *et al.* (2006)], and ML refers to maximum likelihood. All CG and ML runs were performed with $\lambda=0$ and $\tau=0$, respectively, so the thermodynamic set was not used.

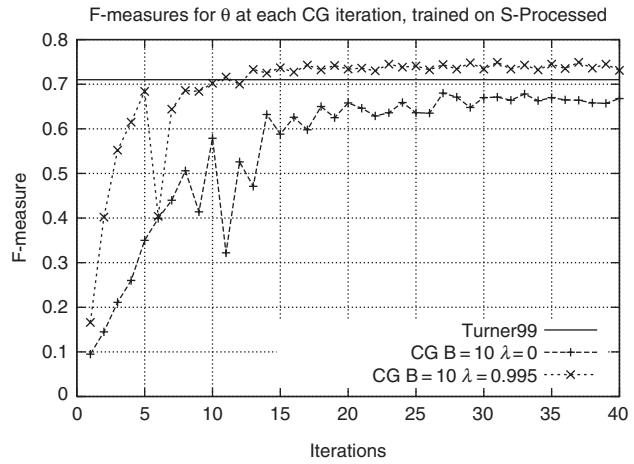

Fig. 6. F-measure when trained on S-Processed versus iteration number for the CG algorithm. Usually the accuracy at the first iterations is much lower than the accuracy of the initial parameter set used (i.e. the Turner99 set), because the number of inequality constraints is small. The algorithm usually converges in about 20 iterations.

on S-Processed, the runtime of CG was within 12 h. Moreover, the number of iterations it takes CG to converge remains low, even on our largest training set S-Processed, as shown in Fig. 6.

5.2.2 Accuracy of CG on artificial training data. When trained on the artificial sets, CG obtained $F=1$ on all training sets within 23 iterations (recall the initial set of parameters had F-measure no more than 0.45). CONTRAfold obtained $F=0.85$ on the training set, but the fact that CONTRAfold did not obtain $F=1$ is not surprising, since CONTRAfold uses a different set of features than does the Turner99 model.

5.2.3 Feature count and CG accuracy on artificial test data. Table 2 also shows that the accuracy of CG improves as the feature counts increase. On the test set S-A1′, the F score improves from $F=0.90$ to $F=0.98$, as the feature count k increases from 1 to 10. We also note that the accuracy of the CG parameters is sensitive to the choice of the bounds parameters B, which should be optimized to account for the size and feature counts of the training data set. In addition to improvements in accuracy, a higher feature count also improves the ability of CG to recover the true Turner99 parameters, as the correlation plots of Fig. 7 show. This indicates that CG is a consistent estimator.

5.3 Results on biological data

In order to compare CG with CONTRAfold, we first trained on S-151Rfam, which was used by Do *et al.* to train CONTRAfold. However, S-151Rfam does not include many of the solved secondary structures available today. Since CG is very efficient, we also trained it on the large structural data set S-Processed. Table 3 shows the results on the training sets, and the accuracy of the Turner99 parameters (columns 3 and 4). We test all three prediction methods on T-Single and S-Full (columns 5 and 6).

5.3.1 Results when training on S-151Rfam. When $\lambda=0.995$ CG performs 4% better than Turner99 and 1% worse than CONTRAfold on the training set. On the S-Full test set however, CG performs 4% better than Turner99 and 2% better than CONTRAfold ($F=$ 0.64 versus 0.60 and 0.62, see Table 3).

When the 48 dangling end parameters were fixed to the Turner99 values for both ML and CG, ML with prior ($\tau=1$) performed only 1% better than CG ($\lambda=0.995$, $B=10$) on the training set S-151Rfam and test set S-Processed. This clearly indicates that the accuracy of CG is comparable with the accuracy of ML when the same model is used. (ML without prior performed 7% worse than ML with prior on the test set, but better than CG with $\lambda=0$, and $B=1.5$ and $B=10$, respectively.)

5.3.2 Results when training on the large structural set S-Processed. Next we trained CG on S-Processed with $B=10$ and $\lambda=0.995$, and tested on S-Full. This resulted in a 3% improvement in prediction accuracy (F = 0.67 versus 0.64) compared to CG when trained on S-151Rfam, a 5% improvement compared to CONTRAfold trained on S-151Rfam ($F=0.67$ versus 0.62), and a 7% improvement compared to the Turner99 parameters (F = 0.67 versus 0.60, see Table 3).

Fig. 8 summarizes the sensitivity and PPV for the Turner99 parameters, CONTRAfold, CG trained on S-151Rfam with $B=1.5$ and $\lambda=0.995$, and CG trained on S-Processed with $B=10$ and λ = 0.995.

5.3.3 Feature counts. Fig. 5 shows that only 254 out of the 363 features underlying the Turner99 model appear at all in S-151Rfam. In fact, only about 170 of them appear more than once. Thus, it is not surprising that CG performs poorly (10% worse than the Turner99 parameters or CONTRAfold) when we train on this set and no thermodynamic data is used (i.e. $\lambda=0$), as seen in the first row of Table 3. When the thermodynamic set is considered, however, CG obtains higher

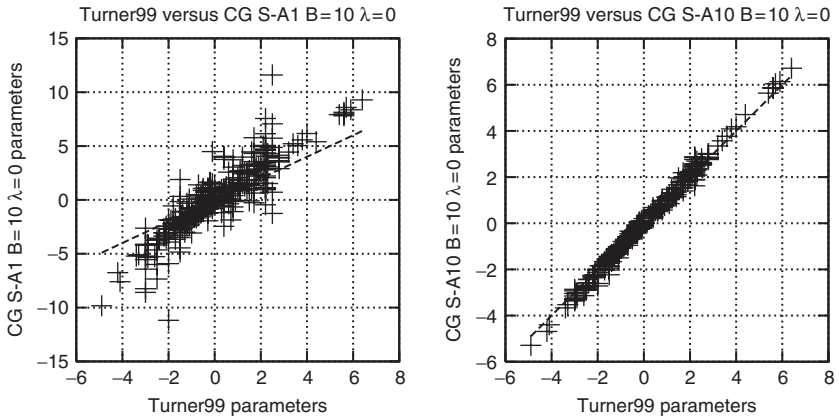

Fig. 7. Correlation between 'true' Turner99 parameters and estimated parameters on the artificial training set when the feature coverage k (minimum number of times each feature occurs in the set) is 1 (left) and 10 (right). When $k = 10$, the estimated parameters are very close to the 'true' ones.

Table 3. Prediction quality achieved by CG, CONTRAfold and the Turner99 parameters

Training sets used	Method	S-151Rfam F (sens/ppv) (training)	S-Processed F (sens/ppv) (training)	T-Single F (sens/ppv) (test)	S-Full F (sens/ppv) (test)	T-Single ΔG error (kcal/mol)
$(1-\lambda)\cdot$ S-151Rfam $+\ \lambda\cdot$ T-Full	CG $B=1.5$ $\lambda=0$	0.59 (0.56/0.62)	–	0.60 (0.44/0.95)	0.58 (0.55/0.61)	3.17
	CG $B=10$ $\lambda=0$	0.57 (0.54/0.60)	–	0.47 (0.31/**1.00**)	0.48 (0.45/0.51)	6.08
	CG $B=1.5$ $\lambda=0.995$	0.69 (0.73/0.65)	–	0.90 (0.85/0.96)	0.64 (0.65/0.63)	0.59
	CG $B=10$ $\lambda=0.995$	0.66 (0.69/0.63)	–	0.68 (0.53/0.96)	0.64 (0.65/0.63)	0.56
$(1-\lambda)\cdot$ S-Processed $+\ \lambda\cdot$ T-Full	CG $B=10$ $\lambda=0$	–	0.68 (0.69/0.67)	0.68 (0.53/0.96)	0.56 (0.57/0.54)	3.66
	CG $B=10$ $\lambda=0.995$	–	**0.75 (0.77/0.73)**	**0.95** (0.93/0.96)	**0.67 (0.70/0.64)**	**0.54**
S-151Rfam	CONTRAfold $\gamma=4$	**0.70** (0.73/**0.67**)	–	0.76 (0.64/0.93)	0.62 (0.62/0.61)	7.74
	CONTRAfold $\gamma=6$	0.69 (**0.75**/0.64)	–	0.84 (0.76/0.93)	0.62 (0.64/0.60)	
–	Turner99	0.65(0.72/0.60)	0.72 (0.75/0.70)	0.93(**0.97**/0.88)	0.60 (0.64/0.57)	0.96

Column 1 gives the training sets we used. Column 2 gives the method we are testing: CG (constraint generation) with various input parameters, CONTRAfold, and the Turner99 parameters. Columns 3 and 4 show the accuracy (F-measure, sensitivity and PPV) of CG and CONTRAfold, when tested on the training structural set used (S-151Rfam in Column 3 and S-Processed in Column 4); the last row of the table shows the accuracy of Turner99 on both training sets, for comparison. The closer the accuracy values are to 1.00, the better. Columns 5 and 6 show the prediction accuracy on our test sets. The last column gives the average error of the predicted free energy score, when compared with the measured free energy value for T-Single: $\sum_x |e_x - \hat{e}_x|/N$, where $N = 207$ is the size of T-Single (the smaller the average error, the better). Bold face values indicate cases where the corresponding parameter set performs best for that column.

average prediction accuracy than CONTRAfold on our large data set, S-Full.

Fig. 5 also shows that S-Processed contains almost all of the Turner99 features, missing only 15 of them. At the same time, the prediction accuracy on S-Full further increases when CG is trained on S-Processed using $\lambda = 0.995$.

The thermodynamic set T-Full contains 203 features out of all 363 features in the model. Note that, when $\lambda > 0$, one occurence of a feature in the thermodynamic set is sufficient to get a good estimate of the free energy value for that feature; this is different from the situation for the structural set, where it is beneficial to have several occurrences of a feature.

5.3.4 Bounds parameter B. The best setting of the bounds parameter B is correlated with the feature counts of the structural set used. If many of the features do not appear in this set, we need to set a tighter bound on the parameters. Thus, when we trained on S-151Rfam, a maximal deviation of $B = 1.5$ kcal/mol from the Turner99 parameters gave better prediction accuracy than $B = 10$. It is interesting however that, when $\lambda = 0.995$, the accuracy on S-Full is the same for both $B = 1.5$ and $B = 10$.

When we trained on S-Processed, we used $B = 10$. Experiments with $B = 30$ gave similar results, indicating that a larger value of B would not affect the quality of the parameters.

5.3.5 Weight of thermodynamic data set. As we already observed with the artificial data set, Table 3 shows clearly that the accuracy of prediction improves with increasing feature counts in the structural set. It also improves when strong weight λ is placed on the thermodynamic set. If many feature counts are zero, there is no absolute free energy information in the constraints of the quadratic program (i.e. no equality

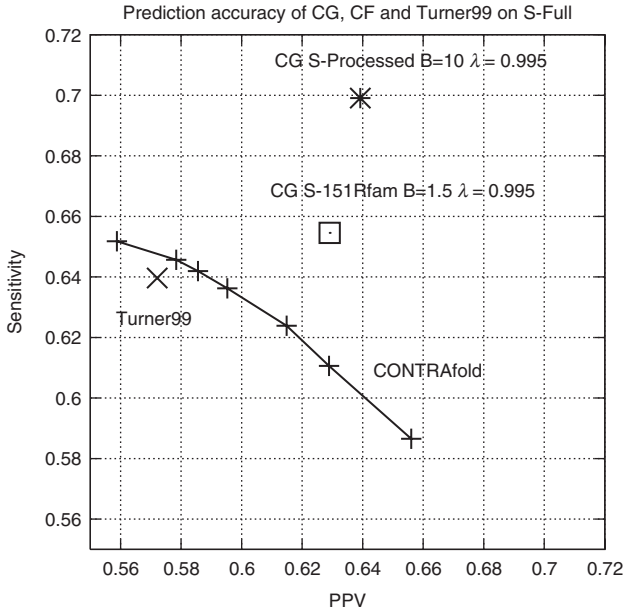

Fig. 8. RNA secondary structure prediction accuracy obtained when using the Turner99 parameters, CONTRAfold parameters ($\gamma \in \{2, 3, 4, 6, 8, 10, 20\}$) and CG parameters (trained on S-151Rfam and S-Processed). Tested on a wide range of biological RNA structures in set S-Full, the parameters obtained using CG give significantly better accuracy than those found by CONTRAfold and the Turner99 parameters.

constraints), and the feature counts cannot compensate for the lack of free energy information.

5.3.6 Free energy accuracy. In addition to measuring the accuracy of secondary structure prediction, we compare the average absolute difference between the experimentally measured free energy for the molecules in T-Single, and the predicted scores for the true structures. A good free energy estimation means this average error is low (rightmost column of Table 3). While CG with $\lambda = 0.995$ yields an average error lower even than the Turner99 parameters (which is 0.96 kcal/mol), CONTRAfold's score differs by 7.74. This clearly shows that the scores used by CONTRAfold lose the free energy physical meaning.

5.3.7 Prediction accuracy for different types of RNAs. Table 4 shows the F-measures of our best CG parameters (i.e. trained on S-Processed, with $B = 10$ and $\lambda = 0.995$), CONTRAfold and Turner99 parameters on various families of RNAs. On families such as transfer RNA, RNase P RNA or ribosomal RNA, CG performs best on average, between 2% and 16% better than Turner99, and between 1% and 14% better than CONTRAfold. Note that CONTRAfold performs particularly poorly on ribosomal RNAs (16S rRNAs and 23S rRNAs do not exist in the S-151Rfam set, however 5S rRNAs do), although it does perform 3% and 14% better than Turner99 on RNase P and transfer RNAs, respectively.

Table 4. Prediction accuracy on various classes of RNAs from S-Full

RNA class	No.	Length	CG (F)	Turner99 (F)	CF (best γ) (F)
tRNA	484	77 ± 5	**0.75**	0.59	0.73 ($\gamma = 3$)
RNase P RNA	379	333 ± 50	**0.57**	0.53	**0.57** ($\gamma = 3$)
5S rRNA	375	118 ± 2	**0.63**	0.61	0.51 ($\gamma = 10$)
16S rRNA	117	1326 ± 273	**0.50**	0.41	0.37 ($\gamma = 3$)
23S rRNA	36	2821 ± 443	**0.51**	0.44	0.45 ($\gamma = 10$)
SRP RNA	68	163 ± 96	0.60	0.69	**0.61** ($\gamma = 10$)
Ribozymes	63	56 ± 8	0.84	0.88	**0.86** ($\gamma = 2$)
Other	138	74 ± 270	**0.89**	0.88	0.87 ($\gamma = 4$)
S-Full	1660	295 ± 508	**0.67**	0.60	0.62 ($\gamma = 4$)

F-measures for our best parameters (CG trained on S-Processed, with $B = 10$ and $\lambda = 0.995$) and the prediction accuracy of CONTRAfold and Turner99 parameters, on various RNA families.
Bold face values indicate the parameter set which gives the highest F-measure for the corresponding RNA class.

On two families, namely SRP RNAs and ribozymes, CG performs 9% and 4% worse than Turner99, and 1% and 2% worse than CONTRAfold. The number of sequences in these families is smaller than for most of the other families.

6 RELATED WORK

As we have mentioned, Turner and his collaborators have refined their estimates of energy values for over 20 years, based in part on thermodynamic data, and in part on extrapolations from structural data, using genetic and grid search algorithms. However, estimation of parameter values was done in stages, with some values being fixed before others were determined, and were not able to take advantage of the large body of structural information available today. Do *et al.* (2006) also considered the problem of parameter estimation, using ML techniques. Using their method, they estimated parameters for a feature set that they constructed, using a small training data set (151 Rfam structures). They showed that, on their training set, predictions with their model have higher accuracy than predictions with the Turner99 model (using Mfold). However, their feature set is more than twice as large as that of Turner *et al.*, making it difficult to assess whether their success is due to their approach or to their set of features. Additionally, free energy values, which are valuable to biologists, cannot be predicted by their model. Finally, as our results show, the overall accuracy of their predictions is poorer on average than our predictions.

The idea of sequentially adding constraints to optimize a quadratic program was investigated by Tsochantaridis *et al.* (2005), although they used a different objective function and did not consider RNA structure prediction.

7 CONCLUSIONS AND FUTURE WORK

In this article, we present a constraint-based parameter estimation algorithm, CG, which efficiently combines structural and thermodynamic RNA secondary structure data.

Our method is substantially faster than a conditional ML method on relatively small training sets, and, unlike the ML approach, can be practically used on large training sets with thousands of structures.

We applied our method to derive new parameters for the Turner99 model, the most widely used energy model for RNA secondary structure prediction. The parameters obtained with our CG method are significantly better than the Turner99 parameters, in terms of prediction accuracy, both on a large structural set and on most families of RNAs, with a 7% average improvement in accuracy over a data set of 1660 structures. In contrast, CONTRAfold obtains a 2% accuracy improvement overall.

Our analysis to date indicates that both, high feature counts in the structural set, as well as thermodynamic data, contribute to the quality of the parameters obtained by the CG and ML algorithms, although ML is more robust when feature coverage is low.

In the future, we plan to combine the ML and CG methods; for example to use the ML method to optimize a small number of unreliable parameters, such as those pertaining to multi-loops, while using CG to optimize the remaining parameters. Finally, we will explore how the introduction of alternative features, such as co-axial base pair stacking and asymmetry in unpaired segments of multi-loops, can lead to improvements in RNA secondary structure prediction. We note that the CG method can easily be adapted to other feature sets with linear energy functions by replacing the secondary structure prediction procedure.

ACKNOWLEDGEMENTS

We thank Kevin Leyton-Brown for giving us access to CPLEX, and Mark Schmidt for providing us with his LBFGS implementation. We thank Romy Shioda, who suggested using the δ values to capture noise in CG. We thank Daniel G. Brown and colleagues for early suggestions on the CG algorithm. We thank Chuong B. Do for clarifications and help with CONTRAfold. Finally, we thank the funders of this research. Andronescu, Condon, Hoos and Murphy acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC), as well as from the Mathematics of Information Technology and Complex Systems (MITACS) Network of Centres of Excellence. Mathews is an Alfred P. Sloan Research Fellow and is supported by National Institutes of Health grant R01GM076485.

Conflict of Interest: none declared.

REFERENCES

Andronescu,M (2003) Algorithms for predicting the secondary structure of pairs and combinatorial sets of nucleic acid strands. *MSc Thesis*, University of British Columbia, Vancouver BC, Canada.

Benenson,Y *et al*. (2004) An autonomous molecular computer for logical control of gene expression. *Nature*, **429**, 423–429.

Breaker,RR (2002) Engineered allosteric ribozymes as biosensor components. *Curr. Opin. Biotechnol.*, **13**, 31–39.

Cannone,J *et al*. (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.

Dirks,RM and Pierce,NA (2004) Triggered amplification by hybridization chain reaction. *Proc. Natl Acad. Sci*, **101**, 15275–15278.

Do,CB *et al*. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

Hofacker,IL *et al*. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh.Chem.*, **125**, 167–188.

Lafferty,J *et al*. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*. 282–289.

Mathews,D (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.

Mathews,D *et al*. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D *et al*. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.

McCaskill,J (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Sprinzl,M and Vassilenko,K (2005) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **33** (Database issue), 139–140.

Taskar,B *et al*. (2005) Learning structured prediction models: a large margin approach. *Proceedings of the 22nd International Conference on Machine Learning*, 896–903.

Tinoco,I and Bustamante,C (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

Tsochantaridis,I *et al*. (2005) Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, **6**, 1453–1484.

Uhlenbeck,OC (1995) Keeping RNA happy. *RNA*, **1**, 4–6.

Wuchty,S *et al*. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.

Xia,T *et al*. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Zuker,M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.