
Jianping Feng
Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Repository Citation
https://corescholar.libraries.wright.edu/etd_all/494

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

By

Jianping Feng
B.E., Dalian Jiaotong University, 2009

2011
Wright State University
COPYRIGHT BY
JIANPING FENG
2011

______________________________
Shaojun Wang, Ph.D.
Thesis Director

______________________________
Mateen Rizki, Ph.D.
Chair, Department of Computer Science and Engineering

Committee on Final Examination

______________________________
Shaojun Wang, Ph.D.

______________________________
Keke Chen, Ph.D.

______________________________
Xinhui Zhang, Ph.D.

______________________________
Andrew Hsu, Ph.D.
Dean, Graduate School
ABSTRACT


A novel probabilistic discriminative model based on conditional random fields, CONTRAfold, has recently been proposed for single sequence RNA secondary structure prediction. By incorporating most of the features which closely mirror the local interaction terms of thermodynamics-based models, the CONTRAfold model has outperformed both probabilistic and physics-based techniques, and received the highest single sequence prediction accuracies. CONTRAfold, like most other RNA secondary structure prediction techniques, requires a collection of RNA sequences with known secondary structure to serve as training data for the algorithm. Manual annotation of RNA sequences is both expensive and time-consuming, and there remains a great deal more sequence data for which structure is not known than there are structurally annotated sequences. In this paper, we present a principled maximum entropy approach to train the same underlying model used in CONTRAfold using both structurally annotated RNA sequences and a large number of unlabeled RNA sequences. We propose a semi-supervised learning technique that using an entropy decomposition method to efficiently compute the gradient of the conditional entropy on unlabeled RNA sequences. Our experimental results show that the proposed maximum entropy semi-supervised learning technique significantly increases the F-value up to 3.5% when unlabeled RNA sequences are included in the training procedure.
# Table of Contents

1 Introduction ..............................................................................................................1

2 Supervised CONTRAfold .........................................................................................3

3 Semi-supervised CONTRAfold ................................................................................6

4 Experimental Results .............................................................................................13

5 Conclusions ..............................................................................................................21

References ..................................................................................................................22
List of Figures

1. An example of RNA sequence with length is 10.................................................................4
2. F-value vs number of unlabeled RNA sequence ...................................................................18
3. Sensitivity vs number of unlabeled RNA sequences..............................................................19
4. Specificity vs number of unlabeled RNA sequences..............................................................20
List of Tables

1. Baseline results: supervised case where N denotes the number of labeled RNA sequences.....15
2. Maximum entropy semi-supervised results when 100 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequences ...........................................15
3. Maximum entropy semi-supervised results when 200 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence........................................16
4. Maximum entropy semi-supervised results when 400 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence........................................16
5. Maximum entropy semi-supervised results when 800 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence........................................16
6. Maximum entropy semi-supervised results when 1122 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence.....................................17
7. Minimum entropy semi-supervised results when 1122 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequences .................................20
1 Introduction

Ribonucleic acid (RNA) molecules perform a wide variety of catalytic and regulatory functions in all living systems, some of which are only now beginning to be well understood. The enzymatic or regulatory role of a particular RNA macromolecule is a function of both the sequence and structure of the RNA. While RNA is synthesized as a linear chain of nucleotides, base-pairings among nucleotides result in complex secondary structure. Although the secondary structure of an RNA molecule is determined entirely by its sequence, there is as yet no known algorithm for reliably determining the secondary structure that will be adopted by an arbitrary RNA sequence. Experimental assays remain the most reliable method to determine secondary structure [1], though the cost in effort, equipment and reagents for these techniques are often prohibitory.

To date, the most successful computational secondary structure prediction techniques for single RNA sequences are those that rely on physical models of RNA structure. In these techniques, possible secondary structures for an RNA sequence are scored, and then optimized by free energy minimization via dynamic programming et al. [2] [3]. The parameters used in these energy-based methods are derived from empirical studies of RNA structural dynamics.

In response to these challenges, stochastic context-free grammars have emerged as an alternative methodology for RNA secondary structure prediction [4] [5] [6]. Even though the model parameters corresponding to the production rules in probabilistic context-free grammars (PCFGs) do not have direct physical interpretations, nevertheless, without the need for additional laboratory experiments, they can still be easily estimated by using a set of annotated RNA sequences with known secondary structures as training data. However, the accuracies of the best PCFG-based models haven’t matched those of the best physics-based models.

Recently, Do et al. [7] proposed an original data-driven secondary structure prediction method based on conditional log-linear models (CLLMs), the CONTRAfold model. This method generalizes both SCFGs and energy-based methods by using discriminative training and
incorporating features that mirror typical thermodynamic models. CONTRAfold outperforms currently available probabilistic and physics-based techniques and receives the highest accuracies for secondary structure prediction of single RNA sequences.

Because CONTRAfold is a supervised learning approach, it requires a large number of non-homologous RNA sequences to be manually annotated for structure. Since structural annotation of RNA sequences is both expensive and time-consuming, it would be advantageous to take advantage of the large amount of unannotated RNA sequence available in public repositories.

There has been a recent surge in the development of semi-supervised learning techniques within the machine learning community. These methods have the advantage of being able to exploit both labeled and unlabeled training data [8]. In the case of RNA secondary structure prediction, this is a significant benefit due to the sheer amount of unannotated training data available.

In this thesis, we demonstrate a novel semi-supervised machine learning technique employing a principled maximum entropy approach. We show that this method is able to exploit easily obtainable unlabeled RNA sequence to significantly improve upon the performance of the CONTRAfold model for RNA secondary structure prediction.
2 Supervised CONTRAfold

In this section, we give an overview of Do et al.'s CONTRAfold model [10] for RNA secondary structure prediction. We show how this model learns by maximizing conditional likelihood using quasi-Newton optimization algorithm.

In traditional energy-based computational approaches to RNA secondary structure prediction, the energy of a structure is modeled as the summation of local interaction terms. Each term describes a small portion of the global base-pairing energy, and the predicted RNA secondary structure is the one achieving the minimum free energy. Obtaining the free energies for each type of local interaction term that could occur in an RNA secondary structure is a difficult endeavor that usually involves carefully calibrated optical melting experiments.

Do et al. [7] adapt an existing probabilistic modeling technique, CLLMs, to the problem of modeling RNA secondary structure. Unlike previous applications of machine learning techniques to the problem of RNA secondary structure prediction, their model uses parameters which closely mirror the local interaction terms of thermodynamics-based models. They estimate these parameters directly from databases of RNAs with known structure using discriminative machine learning techniques, without relying on optical melting experiments.

Let $\Sigma = \{A, C, G, U\}$ be an alphabet of terminal symbols, and consider a string $x \in \Sigma^L$ with length $L$. Let $x$ to represent an unfolded RNA string, $x_i$ to refer to the $i$th character of $x$, for $i = 1, \ldots, L$, and $x^j_{i}$ to refer to the substring of characters from $x_i$ to $x_j$ in $x$. Since there is one position at each of the two ends of $x$, and there are $L - 1$ positions between consecutive nucleotides of $x$, $x$ has totally $L + 1$ positions. We use indices 0 to $L$ to denote these positions. Figure 4 shows an example of $x$ with length 10.
Let $\mathcal{Y}$ denote the set of all possible secondary structures of a RNA sequence $x$. Given an input RNA sequence $x$, we define the conditional probability of a secondary structure $y \in \mathcal{Y}$ as

$$p_0(y|x) = \frac{1}{Z_0(x)} \exp(\theta^T F(x, y))$$  \((2.31)\)

where $F(x, y) \in \mathbb{R}^k$ is a $k$-dimensional vector of feature counts for both $x$ and $y$, $\theta \in \mathbb{R}^k$ is an $k$-dimensional vector of parameters which measures the importance of corresponding features, and $Z_0(x)$, the partition function of a sequence $x$, is a normalization factor to ensure that $p_0(y|x)$ forms a proper conditional probability distribution over the space of all possible secondary structures $\mathcal{Y}$. Since the logarithm of the weight is a linear function of features $F(x, y)$, Equation (2.31) is typically known as the log-linear representation of a conditional random field (CRF) [9].

Do et al. [7] choose a set of features for the log-linear model, these features are base pairs, helix closing base pairs, free bases, helix lengths, hairpin lengths, internal loop asymmetry, internal loop lengths, bulge loop lengths, a full two-dimensional table of internal loop scores, internal loop asymmetry, helix base pair stacking interactions internal loop asymmetry, single (dangling) base stacking, and affine multi-branch loop scoring. These features closely reflect and mimic the local interaction terms of thermodynamics-based models but with a few key differences.

Given a set of labeled examples (i.e., RNA sequences with known secondary structure) $\mathcal{D} = \{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$, the standard supervised training procedure for CLLMs is to maximize the regularized log conditional likelihood of the labeled examples in $\mathcal{D}$.
\[
CL(\theta) = \sum_{i=1}^{N} \log p_{\theta}(y^{i}|x^{i}) - \lambda U(\theta)
\]  
(2.32)

where \( U(\theta) \) is typically chosen to be \( \| \theta \|^2 / 2 \) and \( \lambda \) is a regularization parameter to control the influence of \( U(\theta) \).

The gradient of the regularized log conditional likelihood is:

\[
\frac{\partial}{\partial \theta} CL(\theta) = \sum_{i=1}^{N} F(x^{(i)}, y^{(i)}) - \sum_{i=1}^{N} p_{\theta}(y^{(i)}|x^{(i)}) F(x^{(i)}, y) - \lambda \frac{\partial}{\partial \theta} U(\theta)
\]  
(2.33)

Do et al. have decomposed a secondary structure \( y \) into four basic types of substructures: hairpins, helices, single-branched loops and multi-branched loops. They derived an inside-outside algorithm in [10] to efficiently compute the features expectation, which is the second term in (2.33). This algorithm exhibits cubic order time complexity in terms of RNA sequence length. Once they obtain the gradient of the objective function (2.32), they use a quasi-Newton optimization algorithm [11], the so called limited-memory L-BFGS, to find the local maxima of the objective function (2.32). This technique has outperformed existing probabilistic and physics-based methods and achieved the best accuracies for secondary structure prediction of single RNA sequences.
3 Semi-supervised CONTRAfold

In this section, we describe a principled maximum entropy technique for semi-supervised learning, and apply this method to the CONTRAfold model using both fully labeled as well as unlabeled RNA sequences. We additionally describe an entropy decomposition method to efficiently compute the gradient of the conditional entropy on unlabeled RNA sequences.

Assume in addition to a set of labeled examples, $\mathcal{D}^l$, we also make use of a set of unlabeled RNA sequences, $\mathcal{D}^u = \{x^{(N+1)}, ..., x^M\}$. The objective is to build a CONTRAfold model $p_\theta(y|x)$ using both labeled and unlabeled RNA sequences, $\mathcal{D}^l \cup \mathcal{D}^u$. For ease of notation, we assume that there are no identical examples in $\mathcal{D}^l$ and $\mathcal{D}^u$.

As seen from Eq. (2.32), supervised CONTRAfold ignores the unlabeled RNA sequences in $\mathcal{D}^u$. In order to make use of both labeled and unlabeled RNA sequences, we propose a maximum entropy approach inspired by Jaynes’ maximum entropy principle [12] for density estimation. This approach is employed to train a semi-supervised CONTRAfold model that improves upon the performance of strictly supervised methods for RNA secondary structure prediction. We maximize conditional entropy (minimizing negative conditional entropy) of a CONTRAfold model over the set of unlabeled RNA sequences $\mathcal{D}^u$ subject to a constraint that the CONTRAfold model must remain consistent and predictive with respect to the set of labeled RNA sequences $\mathcal{D}^l$, i.e.,

$$\min_\theta \left( \sum_{x \in \mathcal{D}^u} p_{\theta}(x) H(p_{\theta}(y|x)) \right)$$  \hspace{1cm} (3.1)

s.t. $D(\tilde{p}_l(x, y), \tilde{p}_l(x)p_\theta(y|x)) + \lambda U(\theta) \leq d$  \hspace{1cm} (3.2)

where $\tilde{p}_l(x, y)$ denotes the empirical joint distribution of both $X$ and $Y$ on the set of labeled RNA sequences $\mathcal{D}^l$, $\tilde{p}_l(x)$ denotes the empirical distribution of $X$ on the set of labeled RNA...
sequences $D^1$, and $\overline{p}_u(x)$ denotes the empirical distribution of $X$ on the set of unlabeled RNA sequences $D^u$, and the objective is the negative conditional entropy of the CONTRAfold model over the set of unlabeled RNA sequences $D^u$

$$-\sum_{x \in D^u} \overline{p}_u(x) H(p_\theta(y|x))$$

$$= \sum_{x \in D^u} \overline{p}_u(x) \sum_y p_\theta(y|x) \log p_\theta(y|x)$$

$$= \frac{1}{M-N} \sum_{i=N+1}^{M} \sum_y p_\theta(y|x^{(i)}) \log p_\theta(y|x^{(i)})$$

and $D(p, q)$ denotes Kullback-Leiber distance between probability distributions $p$, and $q$

$$D(\overline{p}_l(x, y), \overline{p}_l(x)p_\theta(y|x))$$

$$= \sum_{(x, y) \in E^1} \overline{p}_l(x, y) \log \frac{\overline{p}_l(x, y)}{\overline{p}_l(x)p_\theta(y|x)}$$

$$\propto \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y^{(i)}|x^{(i)})$$

Following the standard procedure in optimization, we now convert the constrained optimization problem (3.1-3.2) into an unconstrained optimization problem which minimizes the following objective:

$$RL_{\text{maxCE}}(\theta) = -\sum_{x \in D^u} \overline{p}_u(x) H(p_\theta(y|x)) + \kappa \left(D(\overline{p}_l(x, y), \overline{p}_l(x)p_\theta(y|x)) + \lambda U(\theta)\right)$$

(3.3)

where $\kappa > 0$.

This unconstrained optimization problem again is equivalent to minimizing the following unconstrained optimization problem with $\gamma = \frac{1}{\kappa}$:

$$RL_{\text{maxCE}}(\theta) = D(\overline{p}_l(x, y), \overline{p}_l(x)p_\theta(y|x)) + \lambda U(\theta) - \gamma \sum_{x \in D^u} \overline{p}_u(x) H(p_\theta(y|x))$$

(3.4)

Using the same argument as in the minimum conditional entropy regularization case [13] [14], it is easy to verify that $H(p_\theta(y|x))$ is not convex. Thus (3.1-3.2) is not a convex optimization problem. Similarly there are generally local minima in (3.3) or (3.4) due to the non-convexity of its entropy regularization term.
Grandvalet and Bengio [13] and Jiao et al. [14] proposed a minimum conditional entropy based semi-supervised learning algorithm that exploits the unlabeled data. The objective they proposed to maximize is

$$R_{\text{minCE}}(\theta) = \sum_{i=1}^{N} \log p_{\theta}(y_i | x_i) - \lambda U(\theta) + \gamma \sum_{j=N+1}^{M} \sum_{y} p_{\theta}(y_j | x^{(j)}) \log p_{\theta}(y_j | x^{(j)})$$

(3.5)

where the first term is the log conditional likelihood on RNA sequences with known secondary structures, and the third term is the negative conditional entropy on unlabeled RNA sequences. The regularization parameters $\lambda$ and $\gamma$ control the influences of $U(\theta)$ and the unlabeled RNA sequences, respectively.

This is equivalent to minimizing the following objective (with different values of $\lambda$ and $\gamma$)

$$R_{\text{minCE}}(\theta) = D(\hat{p}_i(x, y), \bar{p}_i(x)p_{\theta}(y|x)) + \lambda U(\theta) + \gamma \sum_{x \in \mathcal{D}_n} \bar{p}_i(x) H(p_{\theta}(y|x))$$

(3.6)

When we compare the maximum conditional entropy approach with the minimum conditional entropy approach, we can see that there is only a sign change on the conditional entropy term. However, as our experimental results will show later, our proposed maximum conditional entropy approach gives much better results.

To optimize the objective function (3.4) or (3.6), we have to compute the gradient for the conditional entropy term. Jiao et al. [14] computed this gradient by the following equation

$$\frac{\partial}{\partial \theta} - H(p_{\theta}(y|x)) = \text{cov}_{p_{\theta}(y|x)}[F(x, y)]\theta$$

(3.7)

where the $(j, k)$-th term of covariance matrix of $\text{cov}_{p_{\theta}(y|x)}[F(x, y)]$ is

$$\text{cov}_{p_{\theta}(y|x)}[f_j(x, y)f_k(x, y)]$$

$$= \sum_{y} p_{\theta}(y|x) \left( f_j(x, y)f_k(x, y) \right) \left( \sum_{y} p_{\theta}(y|x) f_j(x, y) \right) \left( \sum_{y} p_{\theta}(y|x) f_k(x, y) \right)$$

(3.8)

It is easy to see that the second term of the covariance can be compute easily, however, the first term requires the computation of pair-wise features’ expectation, which is much harder to compute in the case of RNA secondary structure prediction.

We adopt the entropy decomposition approach proposed by Mann and McCallum [15] to efficiently compute the gradient of conditional entropy of unlabeled RNA sequences.
\[
\frac{\partial}{\partial \theta} - H(p_\theta(y|x))
= \sum_y p_\theta(y|x) \log p_\theta(y|x) F(x, y)
\]

\[
- \left( \sum_y p_\theta(y|x) \log p_\theta(y|x) \right) \left( \sum_y p_\theta(y|x) F(x, y) \right)
\]

Like the gradient obtained by Jiao et al. [14], there are two terms, and the second is easily computable. For a given RNA sequence the feature expectations and the entropy can be obtained by recursive inside/outside algorithms shown in [10]. However, unlike the previous method, now the first term can be calculated efficiently as well through the use of entropy decomposition technique, which exhibits the same cubic order of computational complexity as the inside/outside algorithm for feature expectations.

For notational and formal reasons, we consider a simple example PCFG that corresponds to the Nussinov folding algorithm [5] [4] [16], and we describe how to use the entropy decomposition technique to compute the first term. Denote \( S \) as the initial (start) nonterminal, \( T = \{a, c, g, u\} \) is a finite set of terminal symbols for RNA, and \( P \) is a finite set of production rules described below,

\[
S \rightarrow aSu \mid uSa \mid cSg \mid gSc
\]

\[
S \rightarrow aS \mid cS \mid gS \mid uS
\]

\[
S \rightarrow Sa \mid Sc \mid Sg \mid Su
\]

\[
S \rightarrow SS
\]

\[
S \rightarrow \epsilon
\]

Assume the \( k \)-th feature is \( S \rightarrow aSu \), then the first term corresponding to the \( k \)-th feature is

\[
\sum_y p_\theta(y|x) \log p_\theta(y|x) F(x, y)
= \sum_y p_\theta(y|x) \log p_\theta(y|x) \sum_{i<j} f_k(x, S \rightarrow aSu, y) \delta(x_i = a, x_j = u)
= \sum_{i<j} f_k(x, S \rightarrow aSu, y) \delta(x_i = a, x_j = u) \sum_{y: (S \rightarrow S_S S_S)} p_\theta(y|x) \log p_\theta(y|x)
\]
where $\delta(x_i = a, x_j = u)$ denotes the indicator function. Thus we need to efficiently compute

$$\sum_{y \rightarrow (S \rightarrow x_i S x_j | x)} p_\theta(y, S \rightarrow x_i S x_j | x) \log p_\theta(y, S \rightarrow x_i S x_j | x),$$

a feature constrained entropy. Fortunately this term can be recursively computed in an inside/outside manner through entropy decomposition as shown below

$$\sum_{y \rightarrow (S \rightarrow x_i S x_j | x)} p_\theta(y, S \rightarrow x_i S x_j | x) \log p_\theta(y, S \rightarrow x_i S x_j | x)$$

$$= \sum_{y \rightarrow (S \rightarrow x_i S x_j)} p_\theta(S \rightarrow x_i S x_j | x) p_\theta(y \rightarrow x_i S x_j, x) \left( \log p_\theta(S \rightarrow x_i S x_j | x) + \log p_\theta(y \rightarrow x_i S x_j, x) \right)$$

$$= p_\theta(S \rightarrow x_i S x_j | x) \log p_\theta(S \rightarrow x_i S x_j | x)$$

$$+ p_\theta(S \rightarrow x_i S x_j | x) H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j, x)$$

$$+ p_\theta(S \rightarrow x_i S x_j | x) H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j, x)$$

where $p_\theta(S \rightarrow x_i S x_j | x) \log p_\theta(S \rightarrow x_i S x_j | x)$ is feature-wise entropy, $H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j, x)$ and $H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j, x)$ are inside and outside conditional entropies and can be recursively computed in a way similar to inside and outside conditional probabilities. For example, the recursive formula for inside conditional entropy can be computed as

$$H(S \rightarrow x_i S x_j | S \rightarrow x_i S x_j)$

$$= p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1})$$

$$+ p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) + p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1})$$

$$+ p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) H(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) + p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1})$$

$$+ p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) + p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1})$$

We can easily see that the first term

$$p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) + p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1}) \log p_\theta(S \rightarrow x_i S x_j, S \rightarrow x_i S x_j_{j-1})$$

Follows the rule $S \rightarrow aS u \mid uS a \mid cS g \mid gS c$ and it can be shown as follow.
The second term

\[ p_0(S \rightarrow x_{i+1}S|x) \log p_0(S \rightarrow x_{i+1}|x) + p_0(S \rightarrow x_{i+1}S|x) H(S \rightarrow x_{i+1}^{-1}|x) \]

follows the rule \( S \rightarrow aS \ | \ cS \ | \ gS \ | \ uS \) and it can be shown as follow.

The third term

\[ p_0(S \rightarrow Sx_{j-1}|x) \log p_0(S \rightarrow Sx_{j-1}|x) + p_0(S \rightarrow Sx_{j-1}|x) H(S \rightarrow x_{j-1}^{-1}|x) \]

follows the rule \( S \rightarrow Sa \ | \ Sc \ | \ Sg \ | \ Su \) and it can be shown as follow.

The forth term

\[ p_0(S \rightarrow SS|x) \log p_0(S \rightarrow SS|x) + p_0(S \rightarrow SS|x) \left( \sum_{i+1 < k < j-1} H(S \rightarrow x_{i+1}^{k}|x) + H(S \rightarrow x_{k+1}^{-1}|x) \right) \]

follows the rule \( S \rightarrow SS \) and it can be shown as follow.

The outside entropy can be computed in a similar recursive fashion.
The above simple example illustrates that in order to compute the first term of the gradient of conditional entropy on unlabeled RNA sequences in Eqn. (3.9), we need to compute feature constrained conditional entropy. This is the sum of feature-wise entropy plus feature weighted inside and outside conditional entropies. Computing inside and outside conditional entropies can be easily performed using the probabilities obtained by inside/outside recursions through the entropy decomposition technique.

We can readily extend the above recursive computations to all of the features used in CONTRAfold model proposed by Do et al. [10]. The recursive formulas to compute the inside and outside conditional entropies are analogous to those for computing the inside and outside conditional probabilities that are fully described in the technical note for the CONTRAfold model [10].
4 Experimental Results

To validate the effectiveness of our proposed semi-supervised maximum entropy approach, we performed a series of cross-validation experiments. We use noncoding RNA families with known consensus secondary structures from the Rfam database[17][18]. Version 9.1 of Rfam contains seed multiple alignments for 1372 noncoding RNA families. The consensus secondary structures for each alignment are taken either from predicted using automated covariance based methods or from a previously published study in the literature. For each of these families, we projected the consensus family structure to every sequence in the alignment, and retained the sequence/structure pair with the lowest combined proportion of missing nucleotides and non-au, cg, gu base pairs. Thus finally we have a set of 1372 independent examples, each comes from a different RNA family.

Among 1372 independent examples, 250 are based upon previously published studies in the literature, and 1122 are predicted using automated covariance based methods. To establish “gold-standard” data for training and testing of semi-supervised learning, we treat the 1122 examples with predicted secondary structures as unlabeled data, and the remaining 250 families with secondary structures from the literature as labeled data which are used for training, cross-validation, and testing.

To compare the performance of different mechanisms, we compute sensitivity, specificity (PPV) and F-value defined as:

\[
\text{sensitivity} = \frac{\text{number of correct base pairings}}{\text{number of true base pairings}}
\]

\[
\text{specificity} = \frac{\text{number of correct base pairings}}{\text{number of predicted base pairings}}
\]
We use the supervised method as the baseline model. To evaluate the performance of the proposed semi-supervised learning approach, we adopt the same feature set used in the supervised method. Both supervised and semi-supervised training procedures for CONTRAfold were run with the same regularization function, $U(\theta) = \frac{\|	heta\|^2}{2}$, that is used in [7].

To evaluate the performance of the semi-supervised method in detail, we vary the ratio between the amount of labeled and unlabeled data, and we also vary the tradeoff parameter $\gamma$ where the optimal tradeoff parameter $\gamma$ is determined by cross-validation data.

Among 250 labeled RNA sequences, first we randomly choose 100 as training data and 50 for cross-validation data. The remaining 100 serve as test data. Second, among 100 labeled RNA sequences of training data, we randomly select 20, 40, 60 and 80 to create five levels of labeled training data, i.e., $N = 20, 40, 60, 80, 100$ respectively. Third, among 1122 unlabeled RNA sequences, we randomly select 100, 200, 400 and 800, to create 5 levels of unlabeled data, i.e., $M = 100, 200, 400, 800, 1122$, respectively. Thus, we conduct 25 experiments corresponding to a particular $N$ and $M$.

By way of example, assume the labeled training data size, $N$, is 20 and the unlabeled data size, $M$, is 100. First, we train the supervised CONTRAfold model with 20 labeled RNA sequences and test its performance using test data that has 50 sequences. Then, we initialize the parameters of the semi-supervised CONTRAfold model using those of the supervised CONTRAfold model. We then further train the model using the semi-supervised maximum entropy approach with 100 unlabeled RNA sequences as unlabeled training data. Since the performance is sensitive to the regularization parameter $\gamma$, we try 7 different gamma values, 0.005, 0.01, 0.03, 0.05, 0.08, 0.1, 0.12 respectively. Next, we test the performance of the semi-supervised model using cross-validation data comprising 100 RNA sequences, and pick the value of $\gamma$ that exhibits the highest performance in terms of F-measure. Finally, with this $\gamma$ value, we test the model using the test data that has 50 RNA
sequences to obtain the final results.

We repeat the above procedure 5 times.

Table 1 shows the average baseline results (over 5 repetitions) of the supervised CONTRAfold when N is 20, 40, 60, 80 and 100 respectively. Clearly as we increase the size of training data, all measures, sensitivity, specificity, and F-value, are generally monotonically increasing.

**Table 1** Baseline results: supervised case where N denotes the number of labeled RNA sequences

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.56292(0.35767-4e)</td>
<td>0.52526(1.10128-4e)</td>
<td>0.60724(5.87868-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.58076(3.01848-4e)</td>
<td>0.55030(9.27635-4e)</td>
<td>0.61564(0.82473-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.58678(2.01137-4e)</td>
<td>0.55874(5.85073-4e)</td>
<td>0.61608(1.80197-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.60148(2.57972-4e)</td>
<td>0.57314(9.18948-4e)</td>
<td>0.63302(4.28557-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.60200(3.27835-4e)</td>
<td>0.58250(6.55090-4e)</td>
<td>0.62334(2.95423-4e)</td>
</tr>
</tbody>
</table>

**Table 2** Maximum entropy semi-supervised results when 100 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequences

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.58068(2.36137-4e)</td>
<td>0.57180(21.6868-4e)</td>
<td>0.59356(9.17308-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.58838(4.08427-4e)</td>
<td>0.57770(10.2911-4e)</td>
<td>0.59998(1.60017-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.59710(5.74640-4e)</td>
<td>0.58576(23.8994-4e)</td>
<td>0.61076(0.57763-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.60700(5.91560-4e)</td>
<td>0.59958(13.8906-4e)</td>
<td>0.61546(3.85458-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.60950(3.50660-4e)</td>
<td>0.60362(8.80337-4e)</td>
<td>0.61628(3.54707-4e)</td>
</tr>
</tbody>
</table>

Tables 2-6 show the average results of maximum entropy semi-supervised CONTRAfold over five repetitions when we fix unlabeled data M to be 100, 200, 400, 800 and 1122 respectively, and vary labeled data N from 20 to 40, 60, 80 and 100 respectively. Clearly as we fix the size of unlabeled data, M and increase the size of labeled data, N, all measures, sensitivity, specificity, and F-value, are in general monotonically increasing.
Table 3 Maximum entropy semi-supervised results when 200 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.57438(1.28512-4e)</td>
<td>0.57782(17.2975-4e)</td>
<td>0.57454(11.8726-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.59044(1.26493-4e)</td>
<td>0.60552(4.15237-4e)</td>
<td>0.57678(3.87487-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.60524(6.14288-4e)</td>
<td>0.60826(26.4643-4e)</td>
<td>0.60406(0.83243-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.61802(7.76947-4e)</td>
<td>0.62550(27.3144-4e)</td>
<td>0.61228(2.18752-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.61828(7.21422-4e)</td>
<td>0.62062(24.6036-4e)</td>
<td>0.61760(2.73460-4e)</td>
</tr>
</tbody>
</table>

Table 4 Maximum entropy semi-supervised results when 400 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.57628(4.41587-4e)</td>
<td>0.58976(11.6877-4e)</td>
<td>0.56454(5.91993-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.59928(1.19207-4e)</td>
<td>0.63070(12.6522-4e)</td>
<td>0.57208(1.53747-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.61308(3.94307-4e)</td>
<td>0.62440(22.3316-4e)</td>
<td>0.60386(1.34968-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.62466(4.74218-4e)</td>
<td>0.65080(15.3498-4e)</td>
<td>0.60198(6.62437-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.63016(4.50453-4e)</td>
<td>0.64320(16.5966-4e)</td>
<td>0.61888(2.17567-4e)</td>
</tr>
</tbody>
</table>

Table 5 Maximum entropy semi-supervised results when 800 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.58394(2.37848-4e)</td>
<td>0.61492(12.6278-4e)</td>
<td>0.55720(3.49690-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.60114(2.00993-4e)</td>
<td>0.62702(6.64237-4e)</td>
<td>0.57816(4.58868-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.61254(5.31403-4e)</td>
<td>0.61980(22.2472-4e)</td>
<td>0.60884(1.24883-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.62292(5.78077-4e)</td>
<td>0.65978(24.8656-4e)</td>
<td>0.59182(4.95372-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.63466(2.57283-4e)</td>
<td>0.65532(13.5279-4e)</td>
<td>0.61632(1.93677-4e)</td>
</tr>
</tbody>
</table>
Table 6 Maximum entropy semi-supervised results when 1122 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequence

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.58776(1.78513-4e)</td>
<td>0.61792(16.3487-4e)</td>
<td>0.56272(6.59327-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.61160(1.79975-4e)</td>
<td>0.63134(5.36008-4e)</td>
<td>0.59344(1.67693-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.61318(3.67477-4e)</td>
<td>0.62226(16.4482-4e)</td>
<td>0.60564(1.84903-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.62266(4.15403-4e)</td>
<td>0.65438(24.5588-4e)</td>
<td>0.59588(3.34327-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.63476(1.90868-4e)</td>
<td>0.66250(8.31490-4e)</td>
<td>0.61000(2.57595-4e)</td>
</tr>
</tbody>
</table>

Finally, Tables 1-6 show the average results of maximum entropy semi-supervised CONTRAfold over five repetitions when we fix labeled data N from 20 to 40, 60, 80 and 100 respectively, and vary unlabeled data M from 100 to 200, 400, 800 and 1122 respectively. Figures 3 and 4 show the resulting sensitivity and specificity. As we can see, when N is fixed and as we increase the size of the unlabeled data set, M, from 100 to 800, sensitivity increases sharply. When we further add more unlabeled RNA sequences from 800 to 1122, the improvement on sensitivity is saturated, specially when N is large. However, when we have a small number of labeled RNA sequences, N = 20 or N = 40, specificity decreases sharply as we increase M. When we have relatively a large number of labeled RNA sequences, N = 60, 80 or 100, specificity decreases slightly as we increase M. Since the increment of sensitivity is much larger than the decrement of specificity. As a final result, when we fix the size of unlabeled data, M and increase the size of labeled data, N, F-measure is generally monotonically increasing. A marked improvement of F-value can be observed when M is 800. Figure 2 is the plot of F-values.
We conducted similar experiments using the widely accepted minimum entropy approach [13][14] for semi-supervised learning to train CONTRAfold. Comparing with our proposed principled semi-supervised maximum entropy approach, the only difference is a change in the sign of $\gamma$. We find that when using the semi-supervised minimum entropy approach, the specificity increases a little, while sensitivity decreases drastically. Consequently, the F-value becomes worse than the supervised CONTRAfold baseline result.

In the case of maximum entropy semi-supervised training, the specificity decreases a little, while sensitivity increases drastically. Overall, the results obtained from semi-supervised maximum entropy approach show clear improvement over the baselines obtained from supervised learning. In both situations, sensitivity changes inversely to specificity, and both methods are much more sensitive than specific.
Table 7 shows the average results over five repetitions of the minimum entropy semi-supervised approach to training the CONTRAfold model where we fix unlabeled data $M$ to be 1122, and vary labeled data $N$ from 20 to 40, 60, 80 and 100 respectively. The minimum entropy semi-supervised approach yields inferior accuracy as compared to its supervised counterpart.
Figure 4 Specificity vs number of unlabeled RNA sequences

Table 7 Minimum entropy semi-supervised results when 1122 unlabeled RNA sequences are used for semi-supervised training. N: the number of labeled RNA sequences

<table>
<thead>
<tr>
<th>N</th>
<th>F-value(variance)</th>
<th>Sensitivity(variance)</th>
<th>Specificity(variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.52436 (0.73328-4e)</td>
<td>0.45428 (3.14867-4e)</td>
<td>0.62240 (13.5299-4e)</td>
</tr>
<tr>
<td>40</td>
<td>0.57020 (9.42000-4e)</td>
<td>0.52616 (30.8990-4e)</td>
<td>0.62626 (2.43338-4e)</td>
</tr>
<tr>
<td>60</td>
<td>0.58254 (1.61573-4e)</td>
<td>0.54668 (4.88967-4e)</td>
<td>0.62426 (2.77993-4e)</td>
</tr>
<tr>
<td>80</td>
<td>0.59988 (3.77017-4e)</td>
<td>0.56318 (7.16817-4e)</td>
<td>0.64258 (1.09312-4e)</td>
</tr>
<tr>
<td>100</td>
<td>0.59906 (2.33903-4e)</td>
<td>0.57092 (2.29997-4e)</td>
<td>0.63058 (5.64377-4e)</td>
</tr>
</tbody>
</table>
5 Conclusions

In this paper, we propose a principled maximum entropy approach to train the CONTRAfold model that uses both manually labeled RNA sequences and a large amount of easily obtainable unlabeled RNA sequences. Our experimental results show that the proposed semi-supervised machine learning technique significantly improves upon the performance of CONTRAfold for secondary structure prediction.

As noted in [7], “To date, SCFGs and their extensions provide the foundation for many standard computational techniques for RNA analysis, ranging from modeling of specific RNA families to noncoding RNA detection to RNA structural alignment. In each of these cases, CLLMs provide principled alternatives to SCFGs which take advantage of complex features of the input data when making predictions.” Extending the CLLM methodology to these cases through maximum entropy semi-supervised learning technique provides a very promising direction for future research.
References


