Max-Margin Models for RNA Secondary Structure Prediction

Clara Fannjiang CLARAFJ@STANFORD.EDU

1. Introduction

RNA was first explored and understood as a messenger molecule, relaying DNA encodings of amino acids during protein synthesis. Beyond messenger RNA, however, a class of RNA known as non-coding RNA plays fundamental roles in transcriptional and translational gene regulation. As the biological mantra goes, form fits function, and these regulatory roles depend on the 3-D structure of the RNA molecule. The 3-D structure is largely induced by the secondary structure, or the base-pairs of the RNA sequence—unlike DNA, whose complementary strands are fully paired, RNA is single-stranded and displays complex patterns of base-pairs. As empirical methods for finding RNA structure, such as crystallography, are time-consuming and involve expensive equipment and expertise, computational methods for predicting RNA secondary structure are of great value to the study of non-coding RNA.

1.1. RNA Secondary Structure Prediction as a Structured Prediction Problem

Since each base in a sequence can only pair with one other base, it is natural to model RNA secondary structure as a graph matching. Each node of the graph corresponds to a base, and each edge corresponds to a base-pair that exists in the secondary structure. Given an RNA sequence, or the nodes of a complete graph where each edge corresponds to a possible base-pair, we approach secondary structure prediction as the problem of deducing the true matching. Biologically speaking, the true or most stable secondary structure is the one that minimizes free thermodynamic energy. To incorporate this concept into the graph model, we design some scoring function over all possible matchings, such that a higher score corresponds to a more stable secondary structure. Finding the true secondary structure is then be equivalent to finding the matching that maximizes the scoring function.

RNA secondary structure prediction is thus an instance of a *structured prediction* problem. Structured prediction arises when we predict not a single label, but rather a set of many interdependent labels. A structured prediction problem involves a scoring function over a set of combinatorial structures, as well as a tractable method for the finding the structure with the maximum score. In our problem, we predict a binary label for each edge in the complete graph, where a label of “1” means that base-pair exists in the secondary structure and a label of “0” means it does not. Label interdependence arises not only from the matching constraint, but also from local interactions between base-pairs that occur in common structural motifs. We hope to solve our structured prediction problem by learning a *max-margin model*, a generalization of the support vector machine (SVM).

1.2. Max-Margin Models for Structured Prediction (MMSP)

The intuition behind an SVM is to learn the decision boundary that maximizes the separation between different labels. As a generalization, a max-margin model learns a scoring function that maximizes the separation between the score of the true structure and the score of any other structure.

Let $\mathcal{X}$ be the space of RNA sequences and let $\mathcal{Y} = \bigcup_{\mathbf{x}\in\mathcal{X}} \mathcal{Y}(\mathbf{x})$ be the space of possible secondary structures, where $\mathcal{Y}(\mathbf{x})$ is the space of possible secondary structures (or graph matchings) for the sequence $\mathbf{x} \in \mathcal{X}$. We parameterize the scoring function $s(\mathbf{x}, \mathbf{y})$ as a linear combination of n features, such that $s(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top f(\mathbf{x}, \mathbf{y})$ where $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$ is the feature function. Given m training instances $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $i = 1, \ldots, m$, our constrained optimization problem is

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2$$
$$\text{s.t. } \forall i,\ \forall \mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)}),$$
$$\mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}) + \ell(\mathbf{y}^{(i)}, \mathbf{y})$$

where the constraints ensure that the “margin” of the score of $\mathbf{y}^{(i)}$ over the score of any other structure $\mathbf{y}$ is at least $\ell(\mathbf{y}^{(i)}, \mathbf{y})$, the loss of predicting $\mathbf{y}$ over the desired $\mathbf{y}^{(i)}$. By adding slack variables ζ to ensure that the constraints are feasible, and maximizing over the possible structures to get a single constraint per training instance, we have

$$\min_{\mathbf{w},\zeta} \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_i \zeta_i$$

Algorithm 1 Subgradient calculation for MMSP objective (from [1])

```
Input: $\mathbf{w}$, $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{m}$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$.
Initialize $\Delta_{\mathbf{w}} = 0$.
for $i = 1$ to $m$ do
    $\mathbf{y}^{*(i)} = \arg\max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} \mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}) + \ell(\mathbf{y}^{(i)}, \mathbf{y})$
    $\Delta_{\mathbf{w}} = \Delta_{\mathbf{w}} + f(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)}) - f(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$
end for
$\Delta_{\mathbf{w}} = \frac{1}{m}\Delta_{\mathbf{w}}$
```

s.t. $\forall i$,
$\mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)}) + \ell(\mathbf{y}^{(i)}, \mathbf{y}^{*(i)}) - \zeta_i$,
$\mathbf{y}^{*(i)} = \arg\max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} f(\mathbf{x}^{(i)}, \mathbf{y}) + \ell(\mathbf{y}^{(i)}, \mathbf{y})$,

where the regularization parameter λ tunes the emphasis on maximizing the margin over obeying the constraints. Since the constraints are active at the optimum, we substitute them into the objective to get the unconstrained optimization problem $\min_{\mathbf{w}} J(\mathbf{w})$ where

$$J(\mathbf{w}) = \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_i (\mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}^{*(i)}) + \ell(\mathbf{y}^{(i)}, \mathbf{y}^{*(i)}) - \mathbf{w}^\top f(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})).$$

2. Methods

2.1. Subgradient Method for MMSP

We minimize the MMSP objective $J(\mathbf{w})$ using the subgradient method, a generalization of gradient descent to non-differentiable convex functions. [3] gives Alg. 1 for calculating a subgradient of the MMSP objective. Note that the algorithm involves computing the max-score structure $\mathbf{y}^{*(i)}$, which becomes the bottleneck for implementation: we must design $f(\mathbf{x}, \mathbf{y})$ to be complex enough, such that the scoring function $s(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top f(\mathbf{x}, \mathbf{y})$ captures the many factors involved in structure stability, yet it must be simple enough that computing (or approximating) $\mathbf{y}^{*(i)}$ is tractable.

2.2. Single Base-Pair Matching

One of the simplest ways to parameterize $s(\mathbf{x}, \mathbf{y})$ is to define features $f_{edge}(\mathbf{x}, \mathbf{y}_j)$ for each edge $\mathbf{y}_j$ in the matching $\mathbf{y}$, and let the score of the matching simply be the sum of the scores of the edges in the matching: $s(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top f(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top \sum_j f_{edge}(\mathbf{x}, \mathbf{y}_j)$. We defined the features of an edge, or base-pair, as the distance between the two bases; the probability of the base-pair as computed by Mfold, a model that approximates the free energy of secondary structures using empirically measured energies of structural motifs; and indicators on the identity of the two bases, $\mathbf{1}$[bases are A-A], $\mathbf{1}$[bases are A-C], $\ldots$, $\mathbf{1}$[bases are U-U]. (Incorporating information from other predictors, such as Mfold, is not uncommon and is justified if the model outperforms the input predictors [2].) To calculate the max-score structure $\mathbf{y}^*$ for the MMSP objective subgradient in Alg. 1, we simply calculate the score of each edge $s_{edge}(\mathbf{x}, \mathbf{y}_j) = \mathbf{w}^\top f_{edge}(\mathbf{x}, \mathbf{y}_j)$ and run Edmonds' blossom algorithm to find the max-weighted matching $\mathbf{y}^*$.

We trained the single base-pair matching model on TrainSetB, a set of 1061 highly diverse RNA sequences and known structures designed in [4]. The model was tested on TestSetB, a set of 428 sequences and structures designed in the same paper. We compared the predictions to those made by Mfold, one of the oldest RNA secondary structure prediction methods, often seen as a baseline for performance. The mean Matthews correlation coefficient (MCC) of the predictions (0.204398) beats the mean MCC of the Mfold predictions (0.183955), where the MCC can be understood as the correlation between the true and predicted binary labels on the edges:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, and FN are the numbers of true positive edges, true negative edges, false positive edges, and false negative edges, respectively.

However, the single base-pair matching model is inherently limited as it neglects how interactions between base-pairs influence the stability of structures. To incorporate such interactions, the scoring function $s(\mathbf{x}, \mathbf{y})$ needs to capture the scores of local structural motifs, not just the scores of isolated base-pairs.

One strength of the single base-pair matching model that the successor model lacks is that the calculation of the max-score structure $\mathbf{y}^*$ is both tractable and exact. By making the scoring function more complex, we forgo our ability to calculate an exact subgradient as given in Alg. 1, and resort to a tractable approximation instead.

2.3. Greedy Stack Matching (GreedyStacks)

A prevalent motif in RNA secondary structures is *stacking* base-pairs, where base-pairs form adjacent to each other. Analogous to the pairing of complementary DNA strands, stacking base-pairs are key to the stability of RNA secondary structures, and are the next focus in enriching the scoring function $s(\mathbf{x}, \mathbf{y})$. We redefine the graph model of the RNA sequence, such that each node represents two adjacent bases and each edge represents a possible "stack", or two adjacent base-pairs. Note that this model cannot express isolated base-pairs, which are rarely observed due to the stability of stacking. Fig. 1 illustrates how the stack matching model expresses a sub-structure.

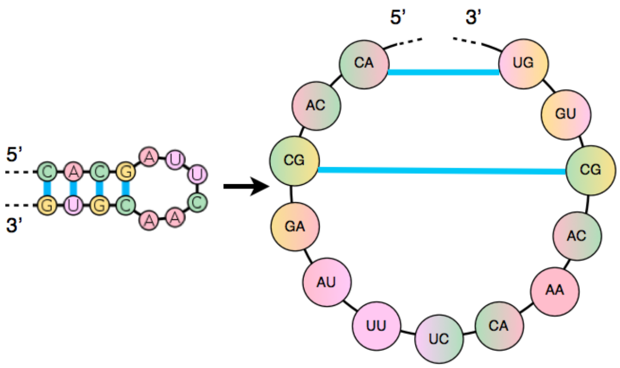

Figure 1. Example of a stack graph matching (right) modeling a sub-structure with stacking base-pairs (left).

Here, a matching $\mathbf{y}$ represents the set of stacks in the secondary structure. As in the single base-pair matching model, we then define the features $f_{stack}(\mathbf{x}, \mathbf{y}_j)$ for each edge or stack, and let the score of a matching be the sum of the scores of the edges in the matching: $s(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top f(\mathbf{x}, \mathbf{y}) = \mathbf{w}^\top \sum_j f_{stack}(\mathbf{x}, \mathbf{y}_j)$.

Note that the max-score structure $\mathbf{y}^*$ is no longer the max-weighted matching, since a matching of the stack graph may not be a valid matching of the single-base graph. Once the "CA-UG" stack exists in Fig. 1, for instance, the adjacent "AC" node cannot be paired since the "A" base has already been paired with the "U" base (and similarly for the "GU" node). In fact, as proven in [5], the calculation of the max-score structure $\mathbf{y}^{*(i)} = \arg\max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} \mathbf{w}^\top f(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^{(i)}, \mathbf{y})$ is NP-hard for $f(\mathbf{x}, \mathbf{y}) = \sum_j f_{stack}(\mathbf{x}, \mathbf{y}_j)$. We calculate the greedy approximation $\mathbf{y}^*_{greedy}$ described in Alg. 2 instead, and replace $\mathbf{y}^*$ with $\mathbf{y}^*_{greedy}$ in Alg. 1.

The features of an edge, or stack, are the distance between the bases of the inner base-pair; a length-dependent intercept $\frac{\#bases}{500}$, where 500 was the maximum sequence length observed in the training set; the probability of the inner base-pair as computed by CONTRAfold, currently one of the best-performing machine learning prediction methods; the probability of the outer base-pair as computed by CONTRAfold; and indicators on the identity of the bases in the stack, $\mathbf{1}$[bases are AA-UU], …, $\mathbf{1}$[bases are UU-GG]. For computational efficiency, we eliminated all edges in the stack graph involving non-canonical base-pairs, or base-pairs other than A-U, C-G, or G-U.

Algorithm 2 Greedy approximation of max-score structure

```
Input: $\mathbf{w}, (\mathbf{x}, \mathbf{y}), \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}, f_{stack}$.
Define a complete graph where each node represents two adjacent bases in $\mathbf{x}$, and the score of each edge $m$ is $s_{stack}(\mathbf{x}, m) = \mathbf{w}^\top f_{stack}(\mathbf{x}, m)$.
Run Edmonds' blossom algorithm for the max-weighted matching $\mathcal{M}$.
Sort the edges $m \in \mathcal{M}$ by decreasing score.
Initialize the greedy approximation $\mathbf{y}^*_{greedy} = \emptyset$.
for each edge $m \in \mathcal{M}$ do
  if $m$ does not conflict with edges in $\mathbf{y}^*_{greedy}$ then
    $\mathbf{y}^*_{greedy} = \mathbf{y}^*_{greedy} \cup m$
  end if
end for
```

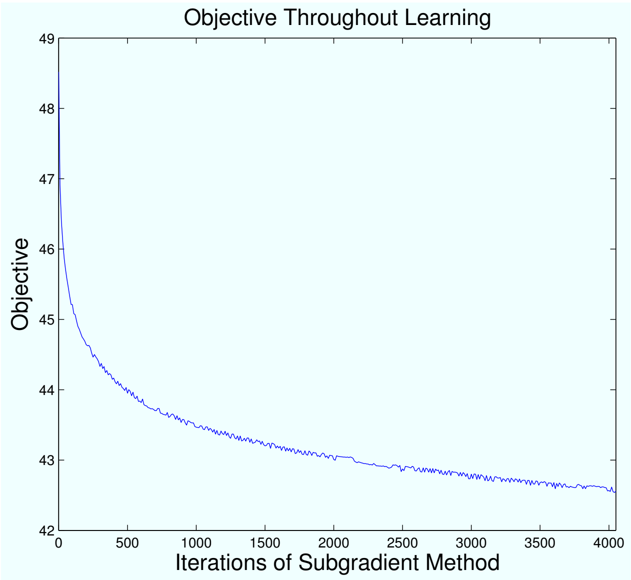

Figure 2. Value of the MMSP objective through iterations of the (approximate) subgradient method.

2.4. Pseudoknots

Here, we note that our graph model lends us a unique advantage in RNA secondary structure prediction. We focus on structures that involve *pseudoknots*, a structural motif where base-pairs are not properly nested

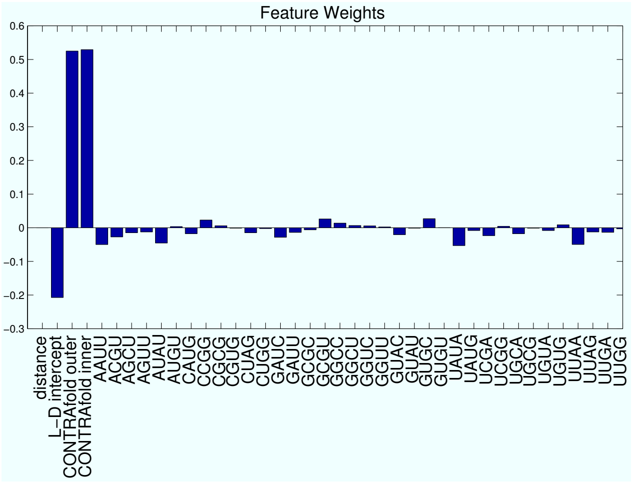

Figure 3. Learned weights of the GreedyStacks features.

and “cross over” one another (Fig. 4). CONTRAfold and most other RNA secondary structure prediction methods are unable to express pseudoknots, as they rely on the nested-ness of base-pairs to express structures recursively. Pseudoknots have thus become important special case of RNA secondary structure prediction. Unlike CONTRAfold, our stacked base-pair matching model make no assumptions about the nested-ness of base-pairs: there are no distinctions between pseudoknotted and non-pseudoknotted edges in the graph. To tune our model for pseudoknots, we train and test GreedyStacks only on pseudoknotted structures and compare the predictions to those made by IPknot, currently one of the best-performing prediction methods for pseudoknots [6].

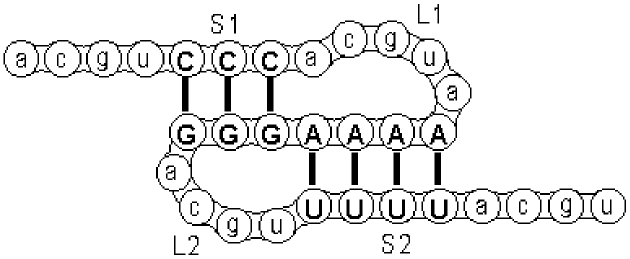

Figure 4. Example of a pseudoknot, where the A-U base-pairs “cross over” the C-G base-pairs.

We created the pseudoknotted TrainSetA and TestSetA by combining the three highly diverse sets cited in [6], and randomly selecting 100 sequences for the testing set.

3. Results

Fig. 5 compares the predictions made on TestSetA by IPknot and GreedyStacks, and reveal that GreedyStacks slightly outperforms IPknot in mean F-measure as well as MCC. The sensitivity vs. positive prediction value trade-off, shown in Fig. 6, shows that though GreedyStacks does slightly worse in PPV, it has an almost 15% advantage over IPknot in sensitivity.

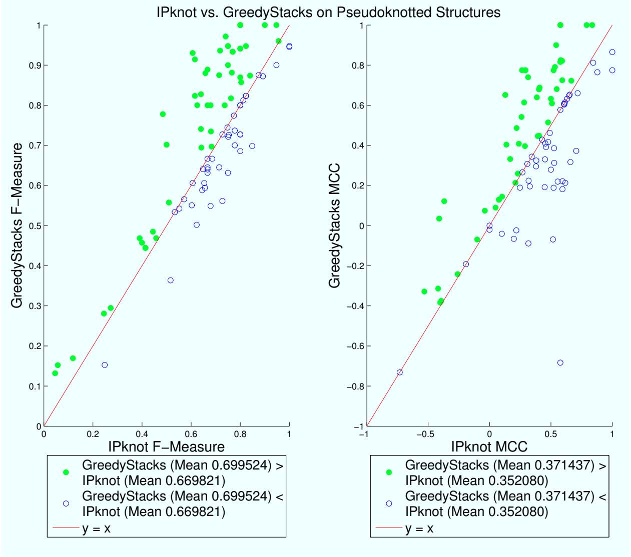

Figure 5. F-measure (left) and Matthews correlation coefficient (right) of TestSetA predictions made by IPknot and GreedyStacks.

4. Future Work

We are currently working on a model that incorporates scores over pseudoknots, not just scores over base-pair stacks. Again, there is no tractable algorithm like the Edmonds' blossom algorithm for finding the max-score structure in such a model. Instead, we define the scores over pseudoknots as higher-order potentials over a Markov random field, then implement dual decomposition as described in [1] to approximate the max-scoring structure.

The regularization parameter λ in GreedyMatch also has not been tuned. We are currently running 5-fold cross validation to tune this parameter, which may further increase the accuracy of its predictions.

5. Acknowledgements

Many thanks to my mentors Cristina Pop and Chuan-Sheng Foo at the Stanford AI Lab for their guidance.

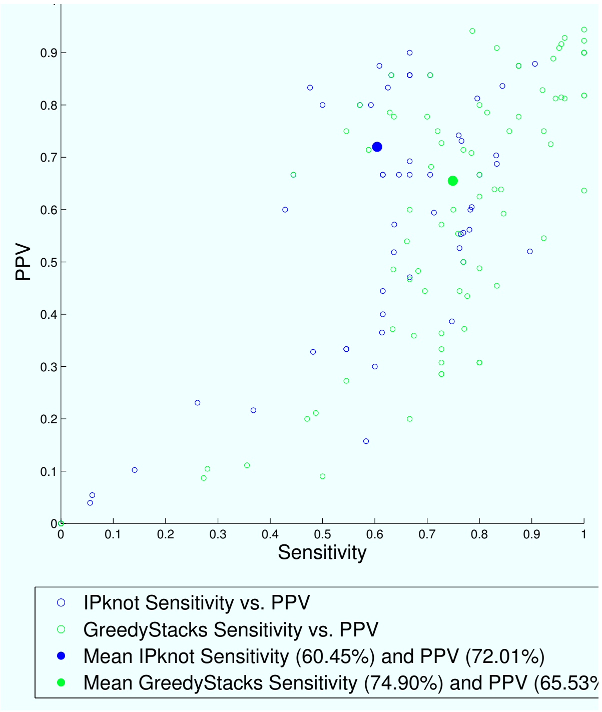

Figure 6. Sensitivity vs. PPV of TestSetA predictions made by IPknot and GreedyStacks.

References

[1] Komodakis, N. (2011). Efficient training for pairwise or higher order CRFs via dual decomposition. *Proc. of the 22th IEEE CVPR*, 1841-1848.

[2] Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for multiple sequence alignments. *JMB*, 302, 205-217.

[3] Ratliff, N., Bagnell, J., & Zinkevich, M. (2007). (Online) subgradient methods for structured prediction. *Proc. of the 11th AIStats*.

[4] Rivas, E., Lang, R., Eddy, S. (2012). A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18, 193-218.

[5] Sheikh, S.I., Backofen, R., and Ponty, Y. (2012) Impact of the energy model on the complexity of RNA folding with pseudoknots. *Proc. of the 2012 CPM*.

[6] Sato, K., et al. (2011). IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *ISMB*, 27, 85-93.

[7] Taskar, B., Chatalbashev V., Koller, & Guestrin, C. (2005). Learning structured prediction models: A large margin approach. *Proc. of the 22nd ICML*, 896-903.