Rich Parameterization Improves RNA Structure Prediction

Shay Zakov, Yoav Goldberg, Michael Elhadad, Michal Ziv-Ukelson

Yakup Korkmaz

Outline

- Introduction
- Preliminaries and Problem Definition
- Feature Representations
- Learning Algorithm
- Experiments
- Conclusion

Introduction

- RNAs functionalities depend on its structural features
  - Number of known RNA structures is still limited

- Secondary structure or folding of RNA sequence: set of formed base-pairs (A,G,C,U)
  - tertiary structure: actual three dimensional molecule structure

- RNA folding: optimization problem, choosing the folding with the maximum score after giving a score for every possible folding of a RNA sequence
  - Standard scoring approach: sum of scores of local structural elements (basic: Nussinov&Jacobson, complex: Turner99 model)

Introduction

- The parameter values (i.e. scores of each local element) traditionally obtained from wet-lab experiments
  - fine-tuned parameter estimation based on machine-learning (ML) techniques possible using known RNA structures
- Today model parameterization remained fairly constant
  - Having few parameters corresponding score of one particular local configuration

- Contribution: much richer parameterizations (≈70.000)
  - models based on the structural elements defined by Turner99
  - score of each structural element is composed of the sum of scores of many fine-grained local features

Introduction

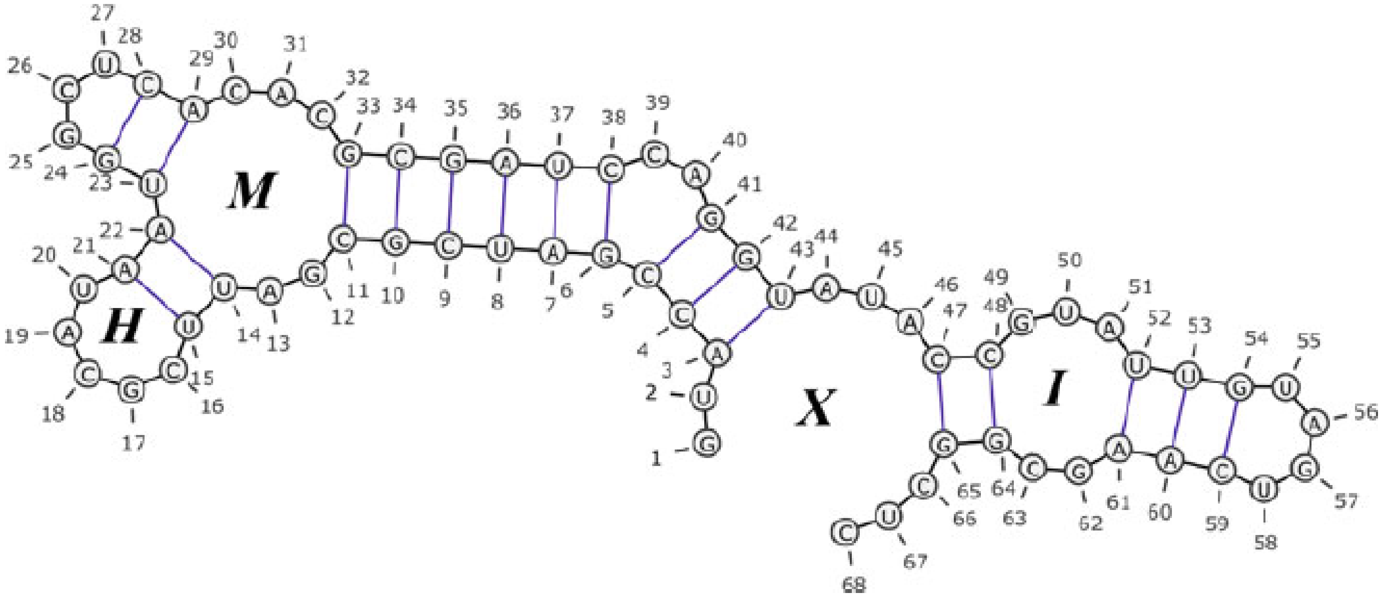


Fig. 1. RNA secondary structure. The figure exemplifies a *secondary structure* of an RNA sequence. Consecutive bases in the sequence are connected with (short) black edges, where base-pairs appear as blue (longer) edges. The labels within the loops stand for loop types, where *H* denotes a *hairpin*, *I* denotes an *internal-loop*, *M* denotes a *multi-loop*, and *X* denotes an *external-loop*. Drawing was made using the

Preliminaries and Problem Definition

- Problem: given an RNA sequence x, find a folding ˆy ∈ Y_x s.t. G(x, ˆy) is maximal
  - index-pairs of the form (i, j), i < j
  - sequence-folding pair (*x*, *y*), where x is an RNA sequence and y is the folding of x
  - scoring model G, function that assigns real-values to sequence-folding pairs (x, y)

- f_G: Folding prediction algorithm

$$\hat{y} = f_G(x) = \mathrm{argmax}_{y \in \mathcal{Y}_x} \{G(x, y)\}$$

Preliminaries and Problem Definition

- Linear model

$$G(x,y) = \sum_{\phi_i \in \Phi(x,y)} \phi_i \mathbf{w}_i = \Phi(x,y)^T \cdot \mathbf{w}$$

  - Φ, the set of different features
  - Φ(x, y) feature representation of (x, y)
    - φ_i corresponds to the ith feature in Φ.
  - Each feature in Φ is associated with a score (or a weight), w
    - w_i is the weight of the ith feature in Φ

Feature Representations

- Two kinds of features (for examples, refer slide 5)
  - Binary features
    - occurrence values are always 1, thus the scores of such occurrences are simply the corresponding feature weights
    - Example: hairpin_base_0=G_+1=C_-2=U (pos. 17 and 25 in slide 5)
      - □ unpaired-base of type G inside a hairpin at a sequence position i, while positions i + 1 and i − 2 contain bases of types C and U respectively
  - Real-valued features
    - set of real-valued length features
    - Example: intervals of unpaired bases within hairpins (interval 16-20)
    - In this work, value of an occurrence of a length feature is log of the interval length

Learning Algorithm

- Goal of the learning algorithm:
  - find a set of parameter values **w** such that the expected cost over unseen sequences x and their true foldings y is minimal
  - Updating weight vector, w

$$\mathbf{w}^i = \begin{cases} \mathbf{w}^{i-1}, & \rho(y,\hat{y}) = 0, \\ \mathbf{w}^{i-1} + \tau_i \Phi(x,y) - \tau_i \Phi(x,\hat{y}), & \text{otherwise,} \end{cases}$$

$$\tau_i = \min\left(1, \frac{\Phi(x,\hat{y})^T \cdot \mathbf{w}^{i-1} - \Phi(x,y)^T \cdot \mathbf{w}^{i-1} + \sqrt{\rho(y,\hat{y})}}{||\Phi(x,\hat{y}) - \Phi(x,y)||^2}\right)$$

  - Decrease the weights of features appearing only in the predicted structure, and
  - increase the weights of features appearing only in the correct structure

Experiments

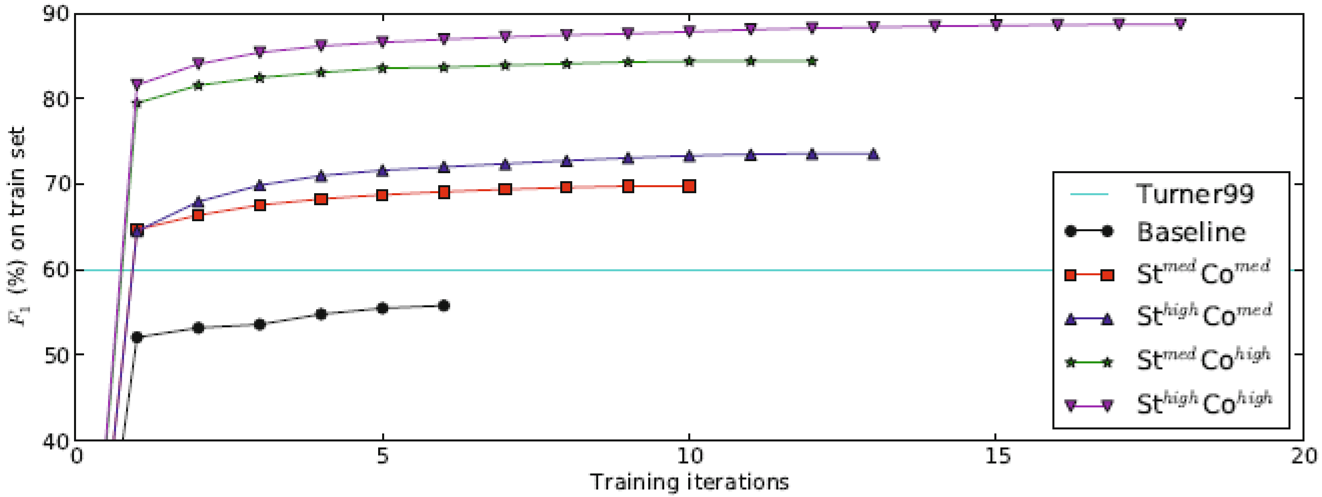
- Data set: (S-Full) is based on the RNA-Strand dataset
  - contains known RNA secondary structures for a diverse set of RNA families across various organisms.

- Models: $St^{med}Co^{med}$, $St^{high}Co^{med}$, $St^{med}Co^{high}$ and $St^{high}Co^{high}$
  - basic model enriched with varying amounts of structural (St) and contextual (Co) information
  - Also baseline model (Baseline) which includes a trivial amount of contextual information

- Measures: sensitivity, positive predictive value (PPV), and F_1-measure

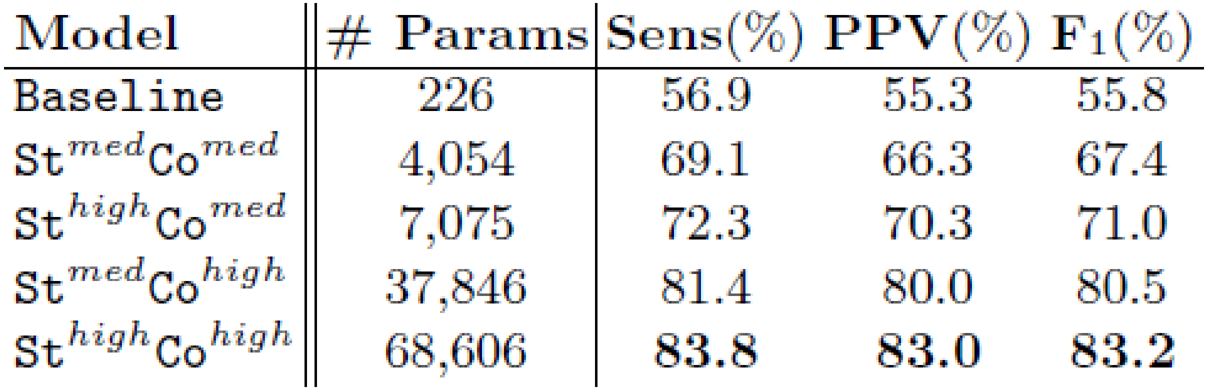
Experiments

- Performance on S-AlgTrain as a function of the number of training iterations



Experiments

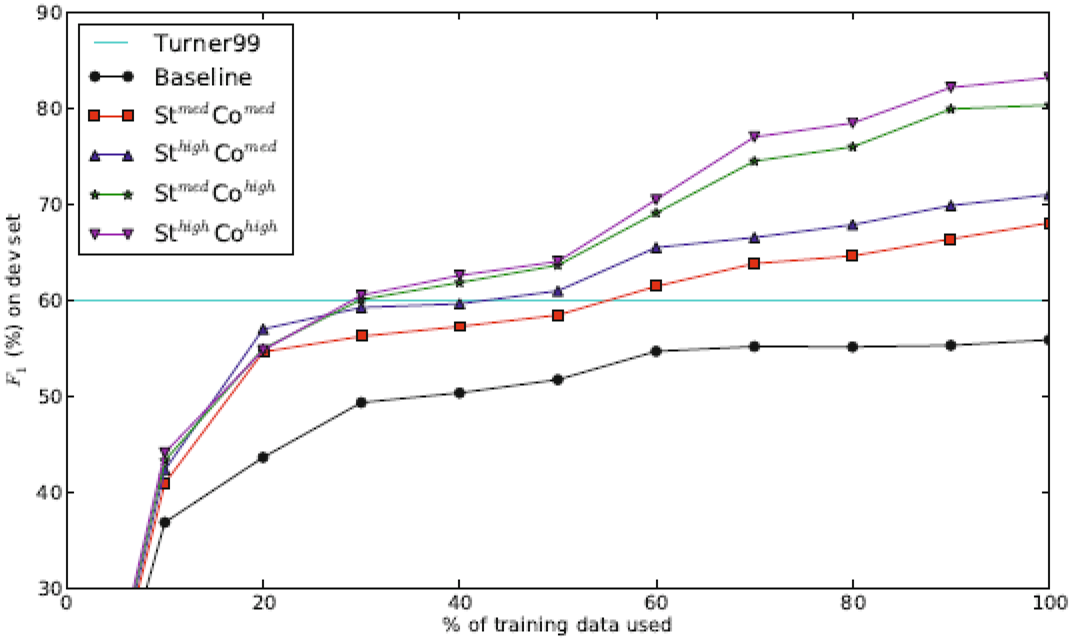
- Performance of final models on the dev set S-AlgTest



Model	# Params	Sens(%)	PPV(%)	F_1(%)
Baseline	226	56.9	55.3	55.8
St^{med}Co^{med}	4,054	69.1	66.3	67.4
St^{high}Co^{med}	7,075	72.3	70.3	71.0
St^{med}Co^{high}	37,846	81.4	80.0	80.5
St^{high}Co^{high}	68,606	**83.8**	**83.0**	**83.2**

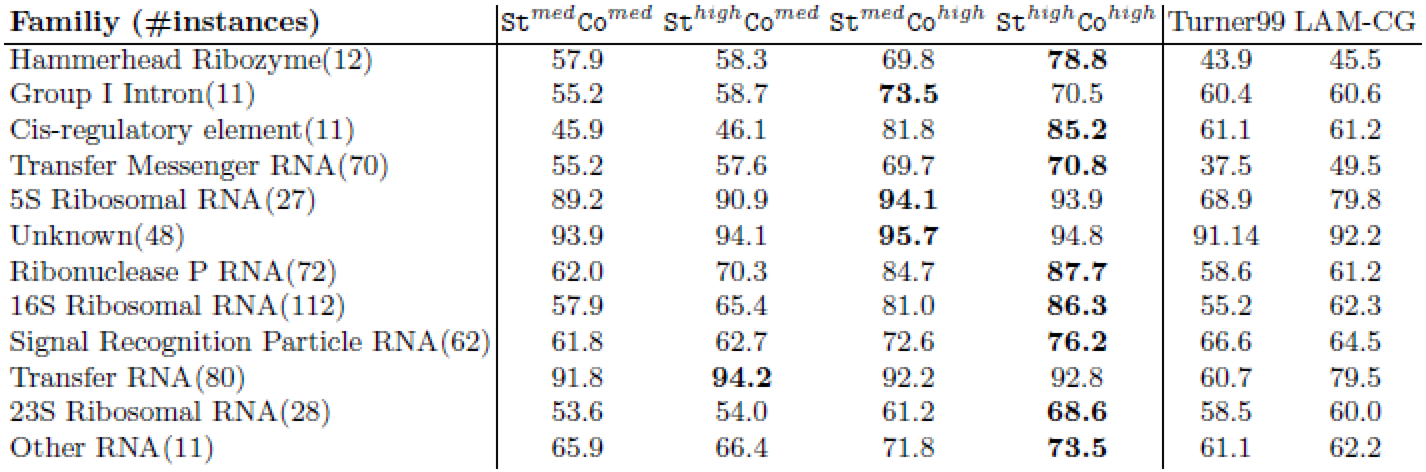
Experiments

- Effect of training set size on validation-set accuracies



Experiments

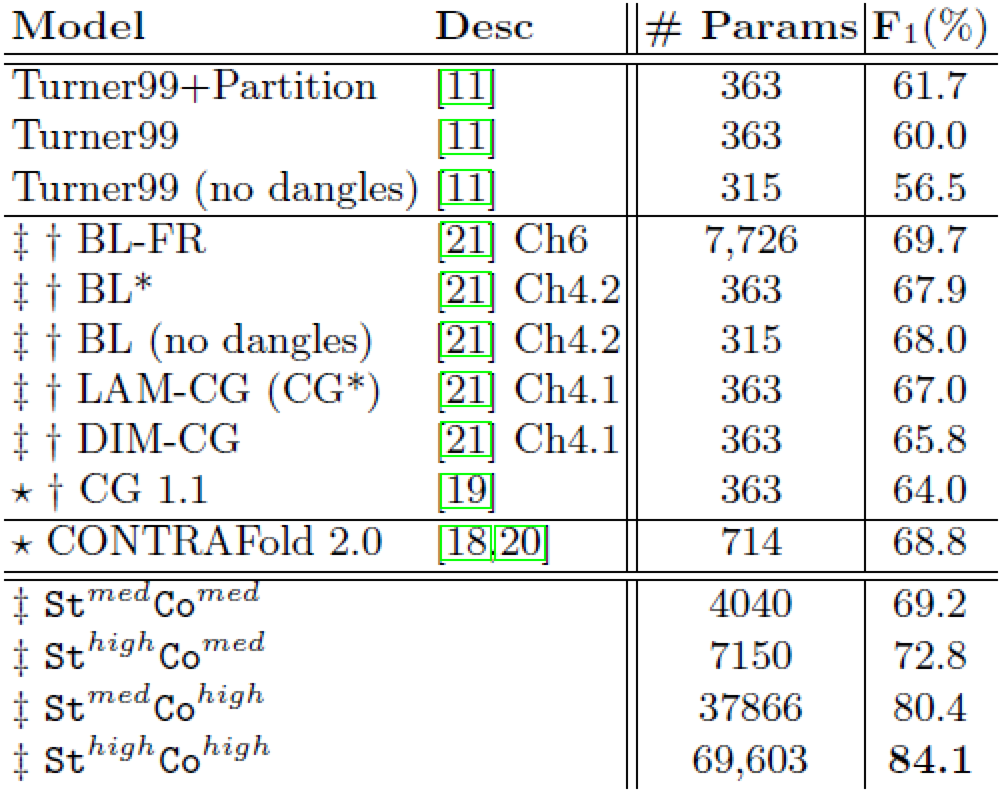
- F_1 scores (in %) of on the development set, grouped by RNA family



Familiy (#instances)	$\mathtt{St}^{med}\mathtt{Co}^{med}$	$\mathtt{St}^{high}\mathtt{Co}^{med}$	$\mathtt{St}^{med}\mathtt{Co}^{high}$	$\mathtt{St}^{high}\mathtt{Co}^{high}$	Turner99	LAM-CG
Hammerhead Ribozyme(12)	57.9	58.3	69.8	**78.8**	43.9	45.5
Group I Intron(11)	55.2	58.7	**73.5**	70.5	60.4	60.6
Cis-regulatory element(11)	45.9	46.1	81.8	**85.2**	61.1	61.2
Transfer Messenger RNA(70)	55.2	57.6	69.7	**70.8**	37.5	49.5
5S Ribosomal RNA(27)	89.2	90.9	**94.1**	93.9	68.9	79.8
Unknown(48)	93.9	94.1	**95.7**	94.8	91.14	92.2
Ribonuclease P RNA(72)	62.0	70.3	84.7	**87.7**	58.6	61.2
16S Ribosomal RNA(112)	57.9	65.4	81.0	**86.3**	55.2	62.3
Signal Recognition Particle RNA(62)	61.8	62.7	72.6	**76.2**	66.6	64.5
Transfer RNA(80)	91.8	**94.2**	92.2	92.8	60.7	79.5
23S Ribosomal RNA(28)	53.6	54.0	61.2	**68.6**	58.5	60.0
Other RNA(11)	65.9	66.4	71.8	**73.5**	61.1	62.2

Experiments

- Final results on the test set



Model	Desc	# Params	F_1(%)
Turner99+Partition	[11]	363	61.7
Turner99	[11]	363	60.0
Turner99 (no dangles)	[11]	315	56.5
‡ † BL-FR	[21] Ch6	7,726	69.7
‡ † BL*	[21] Ch4.2	363	67.9
‡ † BL (no dangles)	[21] Ch4.2	315	68.0
‡ † LAM-CG (CG*)	[21] Ch4.1	363	67.0
‡ † DIM-CG	[21] Ch4.1	363	65.8
⋆ † CG 1.1	[19]	363	64.0
⋆ CONTRAFold 2.0	[18,20]	714	68.8
‡ $\texttt{St}^{med}\texttt{Co}^{med}$		4040	69.2
‡ $\texttt{St}^{high}\texttt{Co}^{med}$		7150	72.8
‡ $\texttt{St}^{med}\texttt{Co}^{high}$		37866	80.4
‡ $\texttt{St}^{high}\texttt{Co}^{high}$		69,603	**84.1**

Conclusion

- Richer parameterizations is beneficial to ML-based RNA structure prediction
  - Best model yields an error reduction of 50% over the previously best published results

- Limitations with respect to the physics-based models
  - does not provide estimates of free energies of secondary structures
  - cannot compute the partition function, base-pair binding probabilities and centroid structures derived from them
  - learned parameter weights are currently not interpretable

Q&A

- Thanks for listening