Deep Learning and RNA structure

Computation Biology Seminar 2.14

Bingbing Wen

Caveats to deep learning approaches to RNA secondary structure prediction

Christoph Flamm[1], Julia Wielach[1], Michael T. Wolfinger[1,2], Stefan Badelt[1], Ronny Lorenz[1], and Ivo L. Hofacker[1,2,*]

[1]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria

[2]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria

Christoph Flamm

Associate Professor, Department of Theoretical
Chemistry, University of Vienna

Background

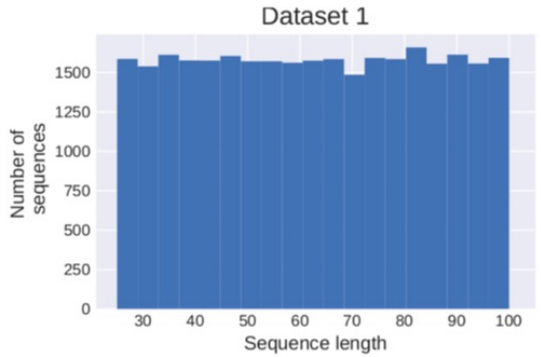
- Prediction of RNA secondary structure
- Most common approach -> Energy directed folding (Turner nearest-neighbor model)

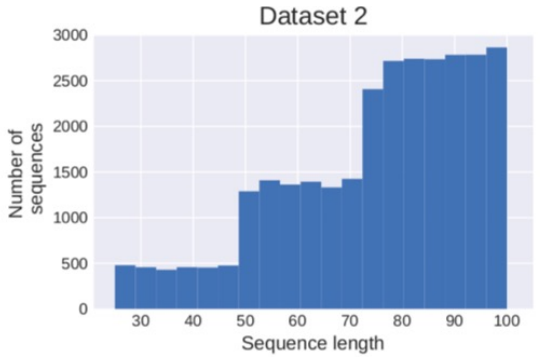
Background

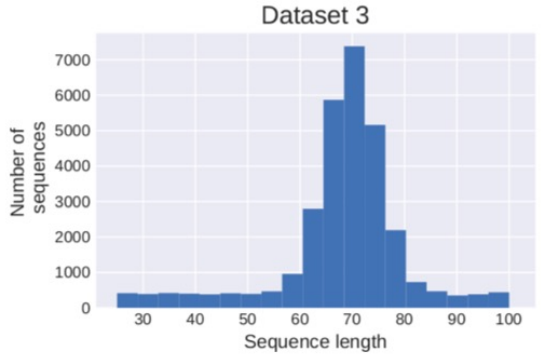
- Deep learning methods for the RNA secondary structure prediction
  - Data hungry problem
- Biased datasets (bpRNA set, Rfam database)
  - Limited Structural diversity of the data set
  - Uneven length distribution of sequences in bpRNA

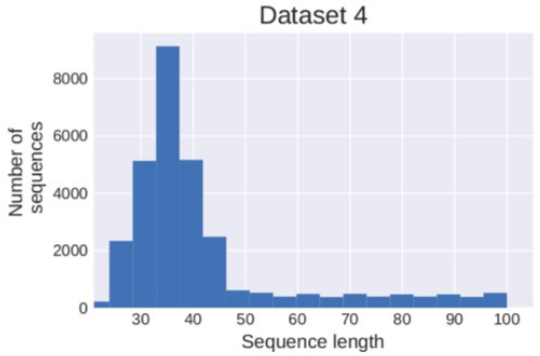
Synthetic data

- Pros: Generated in arbitrary amounts and guaranteed to be free of biases.
- Implementation: RNAfold from the ViennaRNA package to fold random sequences
- From same sequence length to four different length distributions
- equal A,U,C,G content



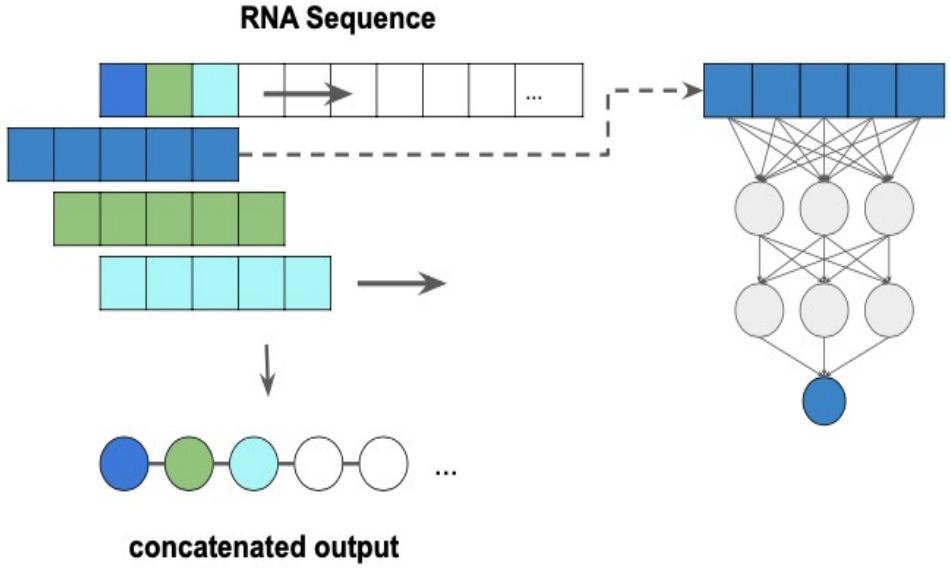


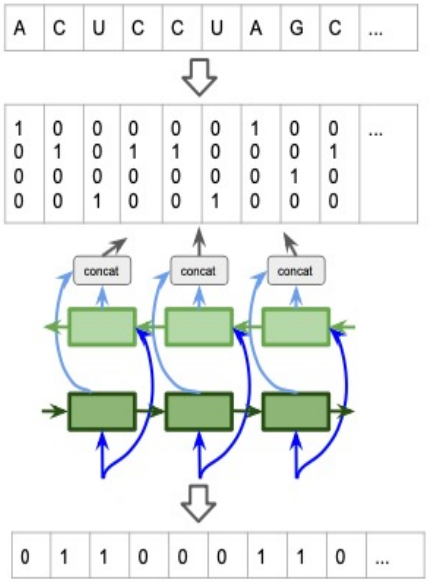




Predicting pairedness

- Predictors:
- A simple feed forward network (FFN)
- A more complex 1D convolutional neural network (CNN)
- A bi-directional long short term memory (BLSTM) network





Predicting pairedness

- BLSTM performed slightly better than the simpler sliding window approaches
- None of the predictors achieve a satisfactory performance

Modeltype	Parameters	Epochs	Accuracy	F1	Loss	MCC
BLSTM	1 Layer, 40 Neurons	43	0.667	0.594	0.609	0.166
	1 Layer, 80 Neurons	27	0.664	0.589	0.612	0.168
	3 Layers, 40 Neurons	38	0.676	0.609	0.604	0.207
Sliding Window	Window 15	89	0.654	0.559	0.623	0.120
	Window 35	94	0.659	0.559	0.620	0.118
	Window 71	59	0.661	0.569	0.618	0.118
CNN Sliding Window	Window 15	67	0.660	0.588	0.616	0.156
	Window 35	65	0.666	0.586	0.609	0.166
	Window 71	30	0.668	0.580	0.608	0.170

Predicting base pair matrices

- n is expanded to a n × n matrix, where each entry corresponds to a possible base pair.
- Method: SPOT-RNA network, a deep network employing ResNets (residual networks), fully connected layers and 2D BLSTMs. Implemented three variants, corresponding to Models 0, 1, and 3.

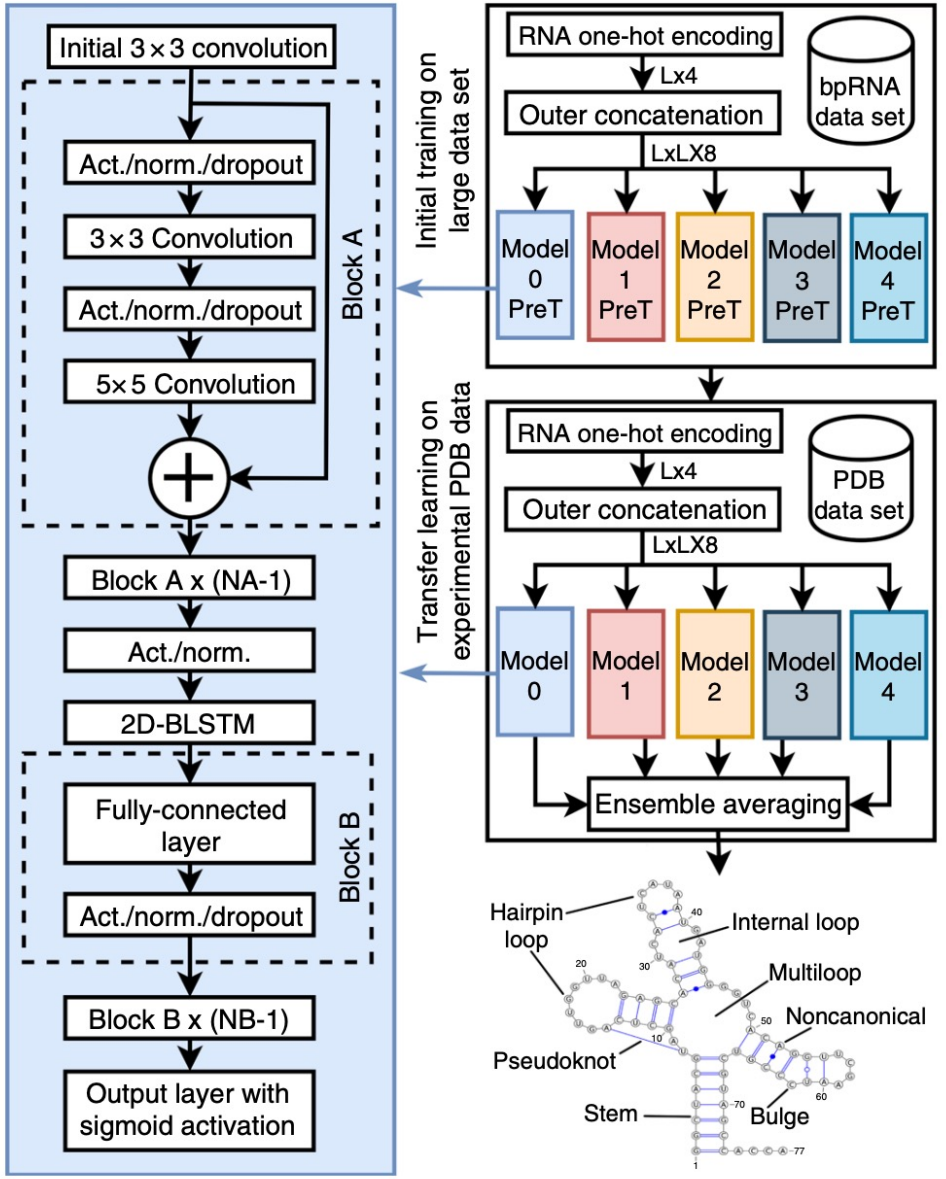
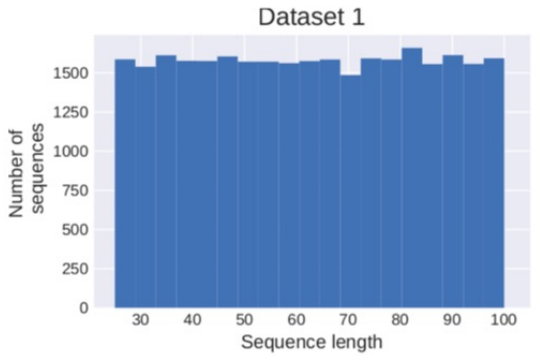
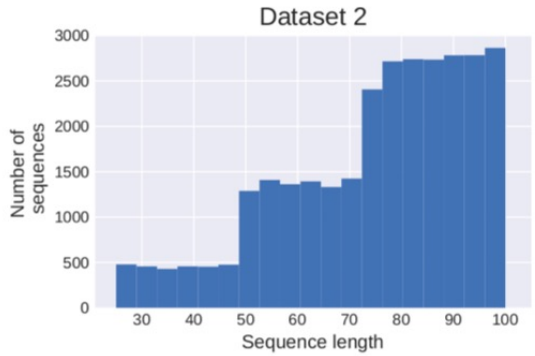
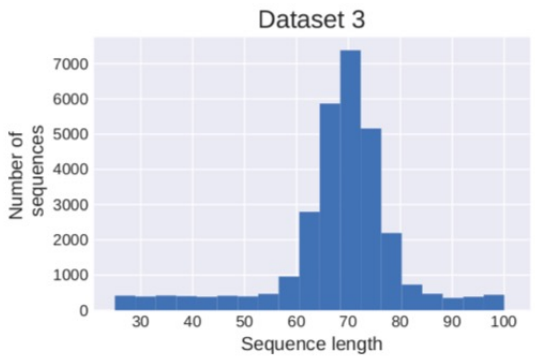


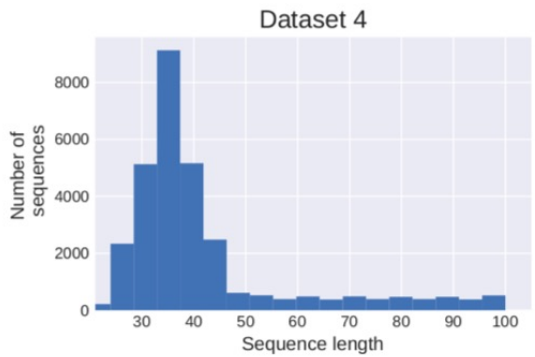
Fig. 1 Generalized model architecture of SPOT-RNA. The network layout of the SPOT-RNA, where *L* is the sequence length of a target RNA, Act. indicates the activation function, Norm. indicates the normalization function, and PreT indicates the pretrained (initial trained) models trained on the bpRNA dataset.

Predicting base pair matrices









Training set	Validation set				Performance
	1	2	3	4	(training set)
1	0.64	0.59	0.61	0.71	0.72
2	0.61	0.58	0.59	0.68	0.66
3	0.64	0.60	0.62	0.70	0.71
4	0.63	0.57	0.59	0.75	0.87

Table 2: The performances of all combinations of training and validation data sets for the four distributions shown in Figure 2. The diagonal in red shows the performance, when training and validation dataset have the same distribution.

Predicting base pair matrices

- The number of predicted base pairs scales quadratically with sequence length, even though a secondary structure can only accommodate a linear number of pairs.

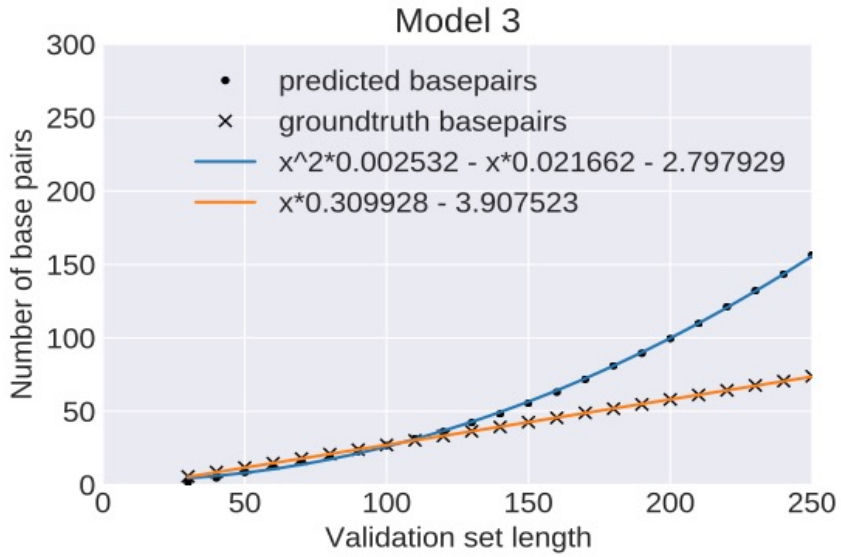


Figure 3: **Predicted number of base pairs:** Average number of base pairs predicted by model 3 (bullets) and in the ground truth data set (crosses) for 2000 sequence per length bin (30-250). The blue and orange curves are least-square regression fits of the data points. The ML-model predicts a wrong quadratic growth (blue curve) for the number of base pairs in contrast to a correct linear growth (orange line).

Predicting base pair matrices

- Network can reproduce many local features of RNA structures, such as the prevalence of different types of base pairs and loops.

Frequency of bases in context external loop (EL), bulge loop (BL), hairpin loop (HL), internal loop (IL), multi loop (ML), base pairs (bps)							
model / length	paired	EL	BL	HL	IL	ML	
VRNA / 70	0.508	0.176	0.033	0.156	0.114	0.014	
NN / 70	0.445	0.222	0.027	0.161	0.127	0.019	
VRNA / 100	0.541	0.123	0.031	0.143	0.126	0.035	
NN / 100	0.433	0.185	0.030	0.146	0.152	0.053	
Average number of structural element							
model / length	helix	EL	BL	HL	IL	ML	
VRNA / 70	4.825	0.992	1.112	1.754	1.841	0.118	
NN / 70	4.354	0.993	0.840	1.730	1.686	0.098	
VRNA / 100	7.132	0.991	1.586	2.314	2.889	0.343	
NN / 100	6.146	0.991	1.080	2.135	2.632	0.299	
Relative frequency of base pair types)							
model / length	GC	CG	AU	UA	GU	UG	NC
VRNA / 70	0.257	0.262	0.169	0.170	0.071	0.071	0.00
NN / 70	0.258	0.260	0.170	0.172	0.070	0.070	$9.63 \cdot 10^{-5}$
VRNA / 100	0.262	0.255	0.173	0.170	0.068	0.071	0.00
NN / 100	0.257	0.252	0.177	0.175	0.068	0.070	$2.30 \cdot 10^{-5}$

Table 3: **Predicted structural features** for RNAfold (VRNA) and Model 3 (NN) trained on sequences of length 70. The test sets consisted of 2000 sequences each of lengths 70 and 100.

Some takeaways

- Synthetic data might relieve the data hungry problem of deep networks
- Networks trained on synthetic data can reproduce many local features of RNA structures but struggle to correctly reproduce global properties and scaling behavior.

Discussion

- Algorithm bias and data bias
- Why the number of predicted base pairs scales quadratically with sequence length? How to improve the model design?
- Is the synthetic data really unbiased? Does it introduce new biases?