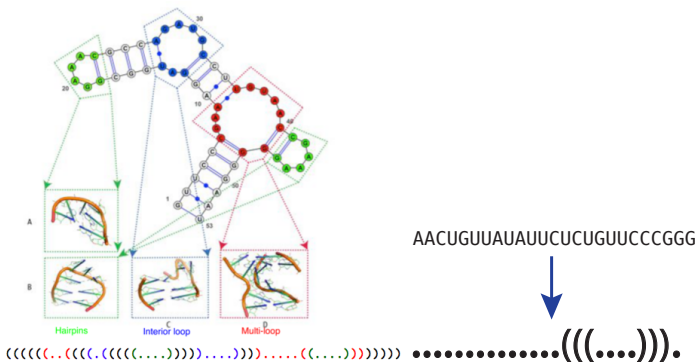


# Learning the RNA inverse folding problem.

Carlos G. Oliver

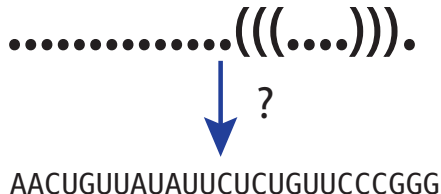
April 19, 2017

# Structure Prediction a.k.a. Folding



- ▶ **Input:** RNA sequence  $x_1, x_2, \dots, x_N$  where  $x_i \in \{A, U, C, G\}$
- ▶ **Output:** RNA 2D structure  $\omega = \{(i, j), \dots\}$  where  $i$  and  $j$  are indices in the sequence  $x$  that are paired.
- ▶ **Solution:** Dynamic Programming + physics based rules.
- ▶ Current approaches achieve high accuracy in  $\mathcal{O}(n^3)$  time for sequences of length  $< 200$ .

# Structure Design a.k.a. Inverse Folding



- ▶ **Input:** 2D Structure
- ▶ **Output:** sequence whose minimum energy fold corresponds to the input.
- ▶ Key problem for synthetic biology and drug design.
- ▶ Current solutions: local search strategies
- ▶ **No linear time algorithms exist to solve this problem.**

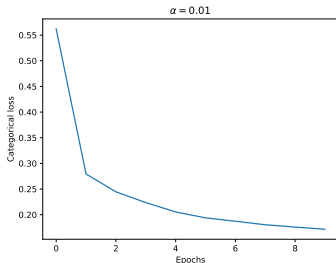
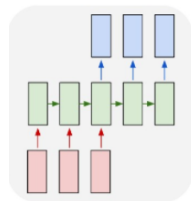
# Data

Two types of sequence-structure data:

- ▶ Real world
  - ▶ **Rfam** database (hand curated sequences for structural families):  
 $\mathcal{O}(100K - 1M)$  sequences per family
- ▶ Artificial
  - ▶ Local search software for design: unlimited data size

# Approach: One model $\rightarrow$ one structure

- ▶ Generate sequences belonging to one of 5 **Rfam** structural families.
- ▶ Given set of member sequences, generate novel likely members.
- ▶ Model: RNN + LSTM
- ▶ Evaluation: Covariance models, GC content, base pair distance, many-to-one discriminator NN.



**training:** > --CUUGAC-GA-U-C-AU-AGA----GC-G-U-U-G--GA-----A-CC-A-----  
**RNN:** > GAUG----GUACUGCG--UCU--CAA-G-ACGUG-GGA--G--AGUA--GG-U-CA-CC--  
> UCU--CAA-G-ACGUG-GGA--G--AGUA--GG-U-CA-CC--G-CCGGUC--GUUG--  
> AUU--ACCUAGUA-----AAA-----AGUUG-GA-----GGG-----AAC-----CC  
> --GGCCCUA-G-UUC--AU-CCAAG--AGUU-----A-----GUUA--AAUCCCC--

## Approach: One model $\rightarrow$ all structures

- ▶ Goal: generate a likely sequence for a given structure.
- ▶ **Input:** vector representations  $s_i \in \{0, 1\}^{|\omega|}$  for each index  $i$  in structure  $\omega$  where

$$s_i[j] = \begin{cases} 1 & i \text{ paired with } j \\ 0 & \text{else} \end{cases}$$

- ▶ **Output:** 1 of 4 encoding of the nucleotide in  $\{A, U, C, G\}$  belonging to position  $i$  in structure.
- ▶ Approach: sequence to sequence RNN, LSTM. Recursive NN.
- ▶ Long term goal

