



ALGORITHMS OF BIOINFORMATICS

8

RNA Structure Prediction

29 January 2026

Prof. Dr. Sebastian Wild

Outline

8 RNA Structure Prediction

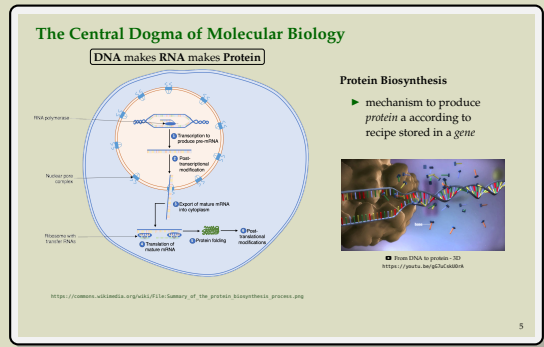
- 8.1 Noncoding RNA
- 8.2 RNA Secondary Structure
- 8.3 Pseudoknot-free secondary structures
- 8.5 Refined Models
- 8.6 Grammar-based Approaches

8.1 Noncoding RNA

RNA

RNA (Ribonnucleic acid)

- ▶ similar to DNA: polymer of *nucleotides*
- ↪ sequence of *nitrogenous bases*
Adenine, and *Cytosine*, *Guanine*, *Uracil*
- ▶ unlike DNA, typically *single-stranded*
- ▶ more “sticky” backbone
- ▶ mostly known as *messenger RNA (mRNA)*
 - ▶ including *mRNA vaccines*!
 - ▶ mRNA is a coding RNA since they encode a protein



Noncoding RNA

But RNA serves many other roles!



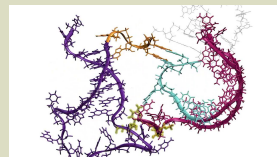
Introduction to Non-Coding RNA
<https://youtu.be/KIohfQsRRdQ>

- ▶ ironically, *ribosomes* (protein factories) themselves are mostly made of RNA
- ▶ for noncoding RNA, structure (3D folding form) crucial for function
- ▶ indeed, sequence often highly variable between species, but structure is similar!

RNA Secondary Structure Prediction

- ▶ Unfortunately, 3D shape hard and expensive to determine experimentally (*X-ray crystallography*)
 - ▶ Available (diverse) data much smaller than for proteins
 - ~> May **not** soon see successful machine-learning solutions similar to AlphaFold
- Rhiju Das, https://youtu.be/XqFq_zYx7Vo

- ▶ To make matters worse, often not a single static structure



▶ RNA folding in action
<https://youtu.be/2XTi9LG9NnU>

- ~> study *de-novo* approaches
- ~> and use simplified models of chemistry and shape to make progress

8.2 RNA Secondary Structure

Model of RNA Structure

- ▶ *RNA sequence / primary structure* $R[0..n) \in \Sigma^n$ $\Sigma = \{A, C, G, U\}$
- ▶ *RNA secondary structure*: matching of indices
 $S \subset [0..n)^2$ of pairs (i, j) that are
 - ▶ ordered $i \leq j$
 - ▶ disjoint: $(i, j), (k, l) \in S \wedge (i = k \vee j = l) \implies (i, j) = (k, l)$
 - ▶ not too close $(i, j) \in S \implies j - i \geq 4$ backbone can't bend more
↖ min. length of hairpin loop
- ▶ secondary structure S is valid for sequence R if
$$(i, j) \in S \implies (R[i], R[j]) \in \mathcal{C} = \{(A, U), (U, A), (C, G), (G, C), (G, U), (U, G)\}$$
- ▶ \mathcal{C} are the *canonical base pairs*: can form *hydrogen bonds* to stabilize RNA

Optimal RNA Structure – Attempt 1

- ▶ Since base pairs provide stability

Try to maximize $|S|$ (# pairs) among all valid secondary structures for $R[0..n]$.

↪ **maximum matching** in graph of all bases

- ▶ possible in polynomial time

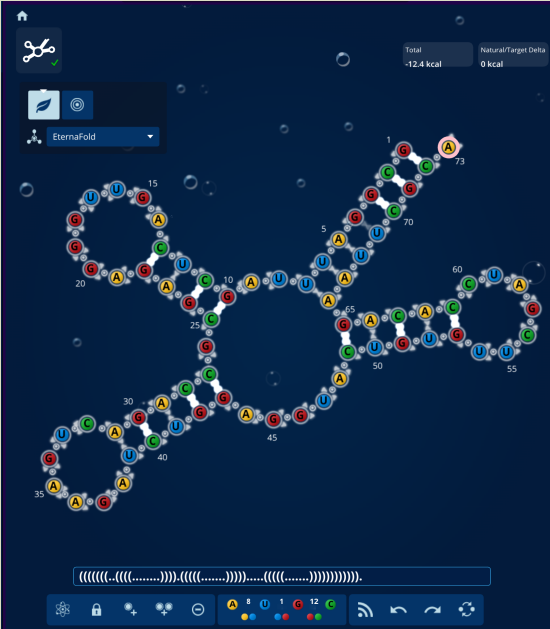
- ▶ actually, ignoring minimum hairpin length, trivial greedy approach is optimal:

1. form arbitrary C – G pairs (until we run out of Cs or Gs)
2. form arbitrary A – U pairs (until we run out)
3. form arbitrary G – U pairs (until we run out)

- ▶ unfortunately, useless predictions!

- ▶ number of pairs dictated by base counts
 - ▶ many equally good options exist
 - ▶ many “optimal” solutions force entire molecule crowd up in one place

Let's play a game!



phenylalanine transfer RNA from *Saccharomyces*
<https://rnacentral.org/rna/URS000011107D/4930>

EteRNA

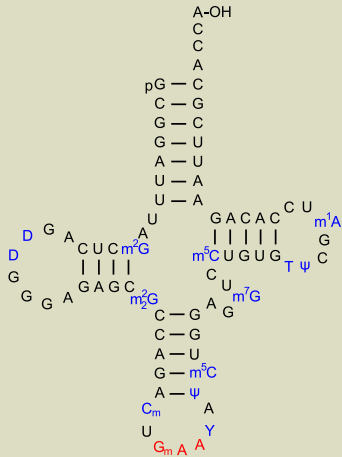
eternagame.org

- ▶ Eterna is a citizen scientist computer game running since 2010 lead by Rhiju Das (Stanford University School of Medicine)
 - ▶ You have to design an RNA sequence that folds into a given *target secondary structure*.
 - ▶ The game uses the best available simulation of RNA folding.
 - ▶ Simulation, prediction, and RNA design algorithms are co-evolving
 - ▶ RNA design crowdsourced to players
 - ▶ top designs synthesized and structure determined
- ~> growing dataset for RNA structures

2D Approximation

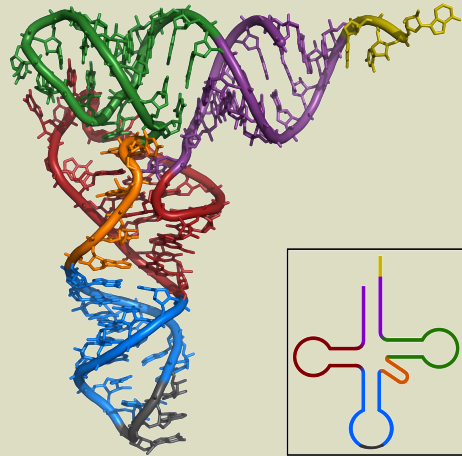
- As in Eterna, RNA secondary structure often drawn as “*roadkill diagrams*”

Roadkill diagram of yeast Phe tRNA

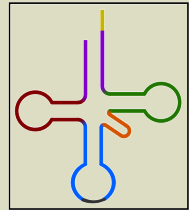


https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_blanco.svg

3D Structure of yeast Phe tRNA



https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_lehz.png

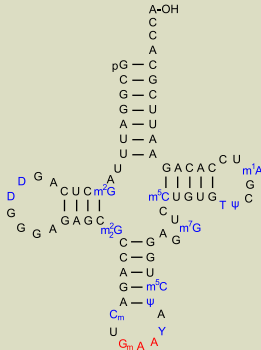


Stacks

Key Observation: Stable structures have many **adjacent pairs**

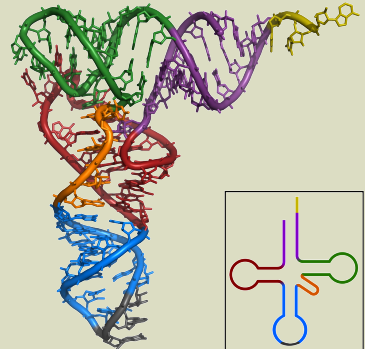
- ▶ “*stacked*” *pairs* forming a *stem* (the “ladder” regions)
- ▶ in 3D, stems form into a double helix (similar to DNA!)
- ▶ only reverse complement stems are stable

Roadkill diagram of yeast Phe tRNA



https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_blanco.svg

3D Structure of yeast Phe tRNA

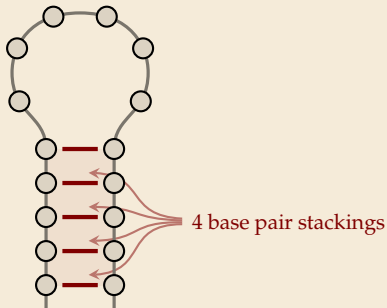


https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png

Optimal RNA Structure – Attempt 2

- ▶ Recall: $S \subset [0..n)^2$ set of indices of paired bases
- ▶ instead of maximizing $|S|$ (# pairs), let's maximize number of **base pair stackings**!

$$\text{BPS}(S) = \left| \left\{ (i, j) \in S : (i + 1, j - 1) \in S \right\} \right|$$



General Secondary Structure Prediction

- ▶ **Given:** Sequence $R \in \{A, C, G, U\}^n$
- ▶ **Goal:** *Valid* secondary structure S with maximal $\text{BPS}(S)$

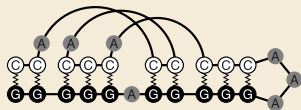
Hardness

Unfortunately, General Secondary Structure Prediction is **NP-hard**.

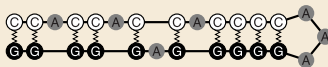
► reduction from BINPACKING



Lyngsø: Complexity of Pseudoknot Prediction in Simple Models, ICALP 2004



(a) An optimum structure for the RNA sequence constructed from an instance of BIN PACKING with four items of sizes 2, 2, 3, and 3, and two bins of capacity 5.



(b) An optimum structure for the RNA sequence constructed from an instance of BIN PACKING with four items of sizes 2, 2, 2, and 4, and two bins of capacity 5.

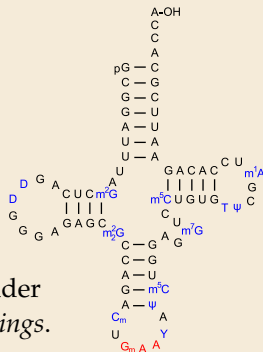
Fig. 3. Illustration of how the number of helices can be kept to one per item for an RNA sequence constructed from a ‘yes’ instance of BIN PACKING, while the base pairs of at least one substring corresponding to an item have to be split over at least two helices if the RNA sequence is constructed from a ‘no’ instance of BIN PACKING.

8.3 Pseudoknot-free secondary structures

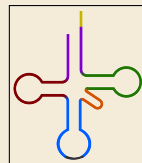
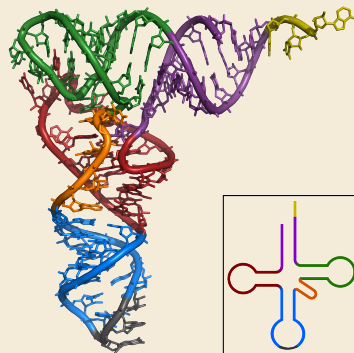
Flat Structures

Recall example tRNA structure

Roadkill diagram of yeast Phe tRNA



3D Structure of yeast Phe tRNA



⇒ Seems reasonable to only consider roadkill diagrams *without crossings*.

https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_blanco.svg

https://commons.wikimedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png

“Correct” formalization seems to be:

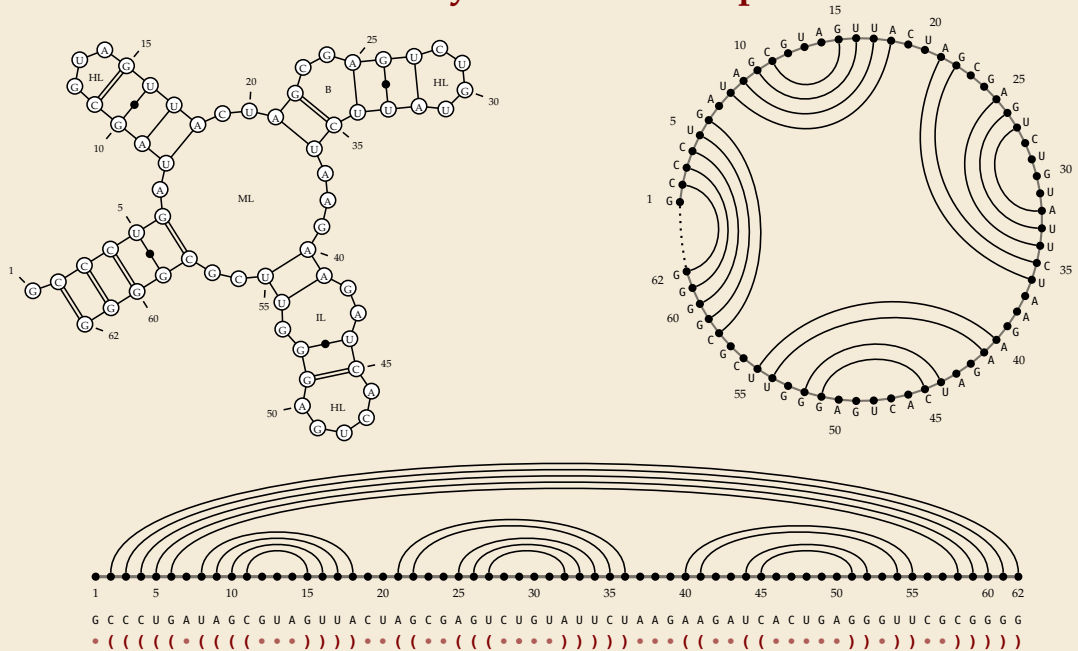
Require graph of pairs bases and backbone edges to be *outerplanar*.

Any other secondary structure is called a *pseudoknot*.

Pseudoknot-free secondary structures

- ▶ planar secondary structure (pairs) cover most of *free energy* of folding
- ▶ “coarse graining” of 3D structure biochemically useful
- ▶ natural intermediate step on folding pathway
- ▶ often well conserved between related species
- ▶ computationally tractable

Pseudoknot-free secondary structures – Representations



Nussinov's Algorithm

Idea: Maximize total number of valid pairs among all pseudoknot-free structures.

- ▶ back to maximum matching, but subject to outerplanar constraint . . .
- ▶ key insight: *decomposability!* see arc diagram / dot-bracket representation

~> Apply dynamic programming
on subproblems $R[i..j]$



$D(i, j) = \max$ valid pairs in pseudoknot-free structure for $R[i..j]$

Nussinov's Algorithm – DP

$D(i, j) = \max$ valid pairs in any pseudoknot-free structure for $R[i..j]$

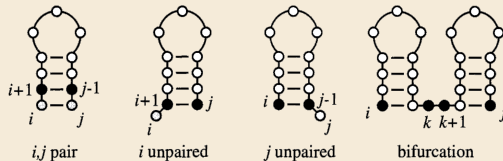


Figure 10.7 from Durbin et al. 1998

$$\rightsquigarrow D(i, j) = \begin{cases} 0, & \text{if } j - i \leq 4; \\ \max \begin{cases} D(i + 1, j - 1) + [(R[i], R[j-1]) \in \mathcal{C}], \\ D(i + 1, j), \\ D(i, j - 1), \\ \max_{k \in [i..j)} D(i, k) + D(k + 1, j) \end{cases} & \text{else.} \end{cases}$$

$\rightsquigarrow O(n^3)$ time, $O(n^2)$ space

8.5 Refined Models

Back to Base Pair Stackings

- ▶ While maximum outerplanar matching is well-defined and tractable, it doesn't usually yield natural structures.
- ▶ already know that we should count base pair stackings!
- ▶ We can extend the DP solution to count those instead!

Graphical notation for DP recursions

$$\text{gray arc} = \min \left\{ \text{empty arc}, \text{arc with 1 sub-arc}, \text{arc with 2 sub-arcs} \right\}$$

Key

- ▶ dots bases; if touching, neighbors on backbone
- ▶ horizontal line RNA backbone
- ▶ wiggly arcs base pair
- ▶ dashed arcs boundary; could be paired or not
- ▶ white area no arcs here
- ▶ gray area potentially further arcs

Counting Base Pair Stackings

Idea: Need to remember whether outermost bases paired.

$$\text{Diagram}(i, j) = \min \left\{ \begin{array}{l} \text{Diagram}(i, j) \\ \text{Diagram}(i, i_1) + \text{Diagram}(j_1, j) \\ \text{Diagram}(i, i+1) + \text{Diagram}(p, p+1) + \text{Diagram}(j-1, j) \end{array} \right\}$$

► In the middle case, if $(i_1, j_1) = (i, j)$, count stacked base pair for (i, j)

$$\text{Diagram}(i, j) = \min \left\{ \begin{array}{l} \text{Diagram}(i, j) \\ \text{Diagram}(i, i+1) + \text{Diagram}(j-1, j) \\ \text{Diagram}(i, p) + \text{Diagram}(p+1, j) \end{array} \right\}$$

↪ Same $O(n^3)$ time, $O(n^2)$ space complexity

Turner Energy Model

- ▶ Simply counting base pair stackings is still a **very crude approximation**
- ▶ Which bases are paired influences bonding strength
- ▶ Which bases are adjacent in stems influences stabilization contribution of stem
- ▶ Which bases form first unpaired base in hairpin loop influences stability
- ▶ ... (play Eterna a bit for more 😊)

↪ More refined models to compute free energy (\approx instability) of structure

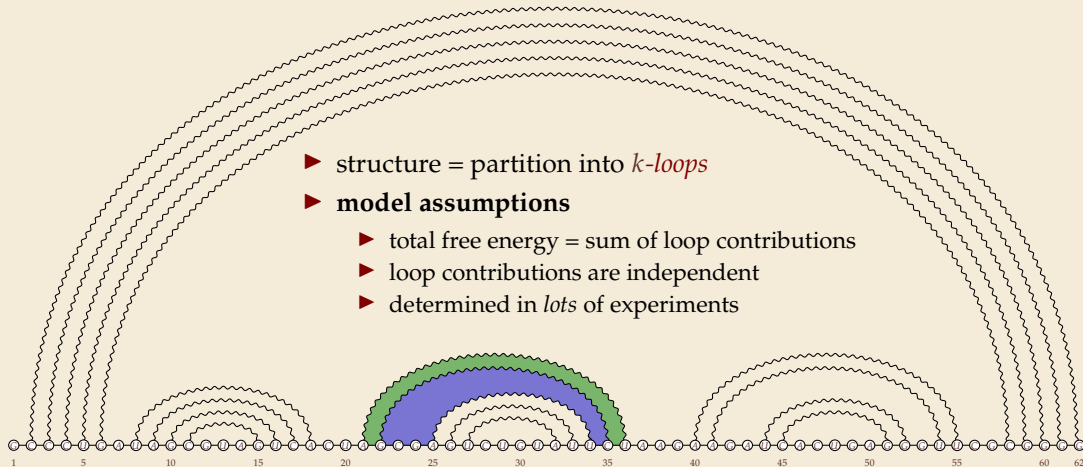


Mathews, Sabina, Zuker, Turner: *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*, J Molecul. Biolog. 1999



Mathews, Disney, Childs, Schroeder, Zuker, Turner: *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*, PNAS 2004

Turner Energy Model [2]



Conceptually unbounded sum

$$\text{shaded loop} = \min \left\{ \text{empty loop}, \text{shaded loop}, \text{shaded loop}, \text{shaded loop}, \dots \right\} \quad \text{⚡ too many variables!}$$

Zuker's Algorithm

- ▶ Only compute exactly up to 2-loops (2 enclosed pairs)
- ▶ additive approximation for bigger multiloops

↪ *same* mutually recursive cost as for pair stackings

8.6 Grammar-based Approaches

Can't machine learning help?

- ▶ free-energy models are great *ab initio* methods
 - ▶ however, they remain limited in accuracy
 - ▶ with growing datasets, tempting to improve structure prediction using machine learning
 - ▶ but: available data much too few for blackbox learning
- ↪ statistical learning with curated probabilistic model

Probabilistic Context-Free Grammars

Recap from your formal languages intro course ...

Context-free grammars (CFG)


$$G = (N, T, R, S)$$

- ▶ nonterminals N
- ▶ terminals T
- ▶ rules $R \subseteq N \times (N \cup T)^*$
- ▶ start symbol $S \in N$

Applying rules to replace nonterminals

$$S \Rightarrow^* w \rightsquigarrow w \in \mathcal{L}(G)$$

Example

- ▶ $N = \{E, I, V, C, C'\}$
- ▶ $T = \{x, y, 0, \dots, 9, +, \cdot\}$
- ▶ $E \rightarrow (E + E) \mid (E \cdot E) \mid I$
 $I \rightarrow C \mid V$
 $V \rightarrow x \mid y$
 $C \rightarrow 0 \mid 1C' \mid \dots \mid 9C'$
 $C' \rightarrow \varepsilon \mid 0C' \mid \dots \mid 9C'$


empty string

Probabilistic Context-Free Grammars (PCFG)

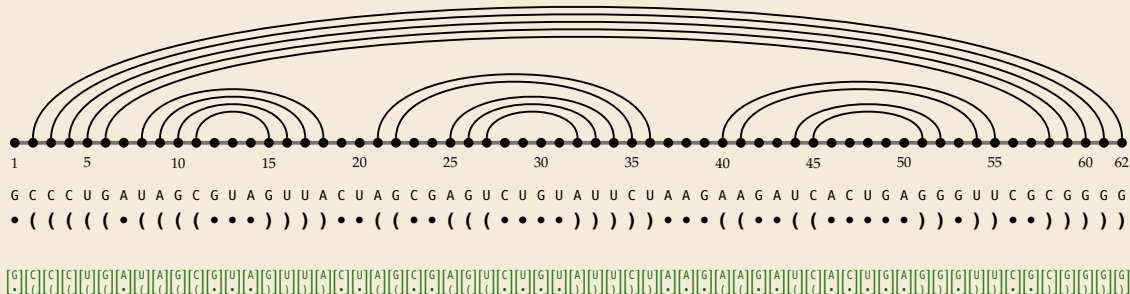
- ▶ For each nonterminal, assign *probabilities* to right-hand sides.

\rightsquigarrow prob of a derivation in G = product of rule probabilities.

RNA Sequence and Structure

Pseudoknot-free secondary structures \approx correctly parenthesized strings
 \rightsquigarrow representable by context-free grammar

Idea: Represent RNA sequence and structure as string of pairs



\rightsquigarrow Consider grammars with terminals

$$T = \left\{ [A], [A], [A], [C], [C], [C], [G], [G], [G], [U], [U], [U] \right\}$$

RNA Grammars

Example Grammar to produce RNA sequence and structure pairs

(with some fictitious rule probabilities)

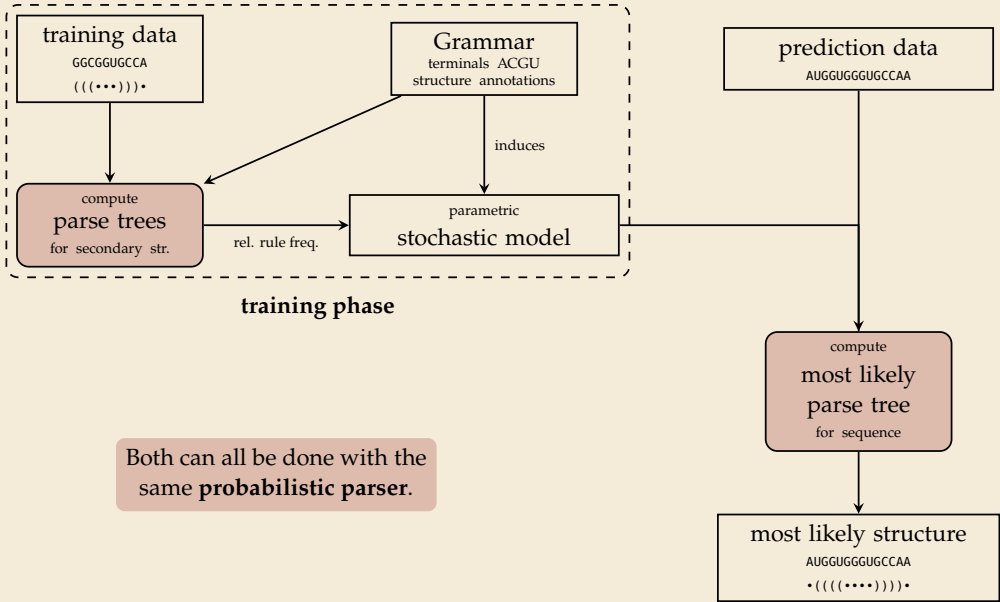
| rule | prob. | rule | prob. | rule | prob. |
|---|-------|---|-------|--|-------|
| $S \rightarrow LS$ | 0.65 | $L \rightarrow \begin{bmatrix} C \\ \cdot \end{bmatrix} S \begin{bmatrix} G \\ \cdot \end{bmatrix}$ | 0.10 | $L \rightarrow \begin{bmatrix} A \\ \cdot \end{bmatrix}$ | 0.10 |
| $S \rightarrow \varepsilon$ | 0.35 | $L \rightarrow \begin{bmatrix} G \\ \cdot \end{bmatrix} S \begin{bmatrix} C \\ \cdot \end{bmatrix}$ | 0.05 | $L \rightarrow \begin{bmatrix} U \\ \cdot \end{bmatrix}$ | 0.15 |
| $L \rightarrow \begin{bmatrix} A \\ \cdot \end{bmatrix} S \begin{bmatrix} U \\ \cdot \end{bmatrix}$ | 0.05 | $L \rightarrow \begin{bmatrix} U \\ \cdot \end{bmatrix} S \begin{bmatrix} G \\ \cdot \end{bmatrix}$ | 0.05 | $L \rightarrow \begin{bmatrix} C \\ \cdot \end{bmatrix}$ | 0.10 |
| $L \rightarrow \begin{bmatrix} U \\ \cdot \end{bmatrix} S \begin{bmatrix} A \\ \cdot \end{bmatrix}$ | 0.15 | $L \rightarrow \begin{bmatrix} G \\ \cdot \end{bmatrix} S \begin{bmatrix} U \\ \cdot \end{bmatrix}$ | 0.10 | $L \rightarrow \begin{bmatrix} G \\ \cdot \end{bmatrix}$ | 0.15 |

Can prove: Grammar has unique parse tree *iff* valid RNA sequence structure pair.

Key idea for structure prediction:

- Given only sequence of bases, cannot distinguish $\begin{bmatrix} A \\ \cdot \end{bmatrix}$, $\begin{bmatrix} A \\ \cdot \end{bmatrix}$, and $\begin{bmatrix} A \\ \cdot \end{bmatrix}$
- ↪ Many possible parse trees compatible with sequence
- ↪ Predict structure corresponding to *most likely parse tree*

Stochastic Approach to Secondary Structure Prediction



Probabilistic Parsing