

# **Deciphering the Information Encoded in RNA Viral Genomes**

Christine E. Heitsch

Genome Center of Wisconsin and Mathematics Department  
University of Wisconsin – Madison

Clay Mathematics Institute

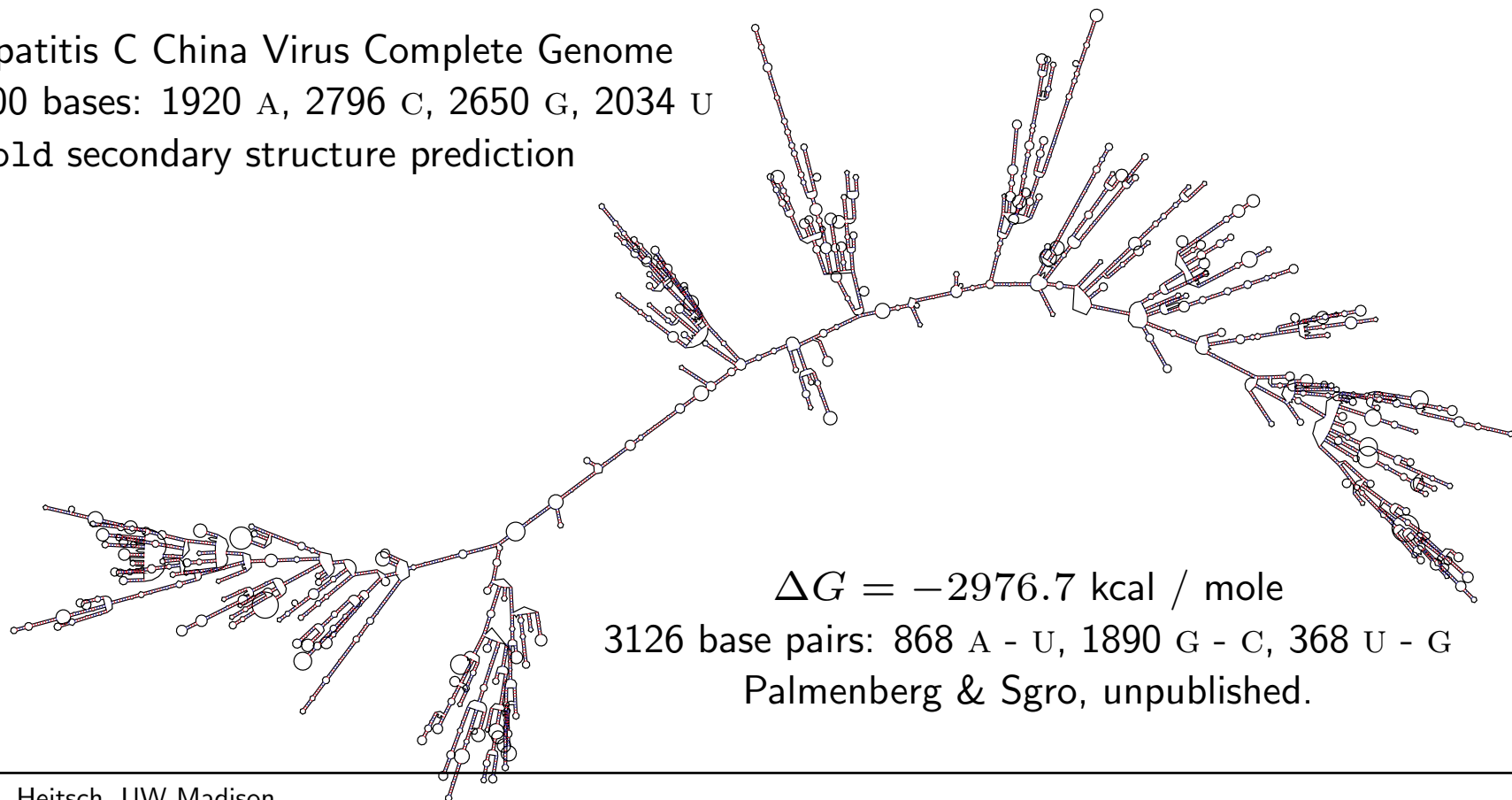
Workshop on Algebraic Statistics and Computational Biology

November 14, 2005

# RNA Base Pairing Encodes Biological Information

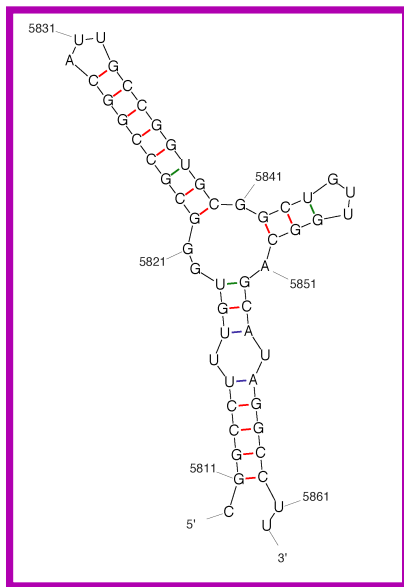
Single-stranded RNA sequences form molecular structures.  
Selective **base pair** hybridization  $\longleftrightarrow$  **structure** and **function**.

Hepatitis C China Virus Complete Genome  
9400 bases: 1920 A, 2796 C, 2650 G, 2034 U  
Mfold secondary structure prediction

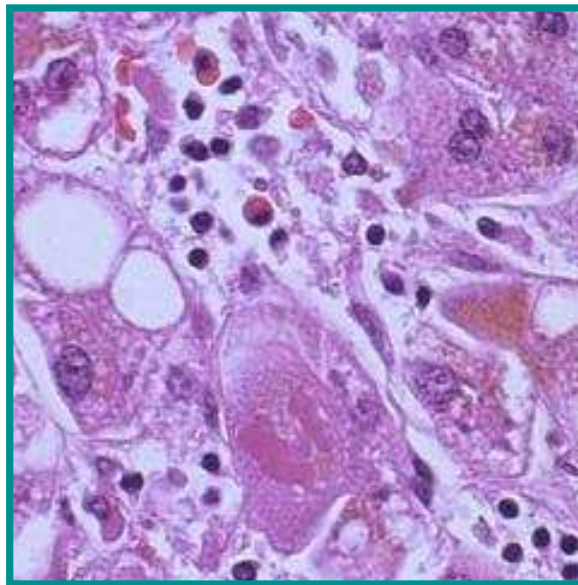


# A Fundamental Challenge

How are **structure** and **function** encoded in biological **sequences**?



Hepatitis C genome  
**structural** fragment  
Palmenberg & Sgro, UW Madison



Liver cells **infected** with  
Hepatitis C virus  
Hepatic Pathology, Florida State

```
.....CGGCC  
UUUGUGG  
GCGCCGG  
CAUUGCC  
GGUGCGG  
CUGUUGG  
CAGCAUA  
GGCCUU...
```

Segment of Hepatitis C  
viral RNA **genome**  
Genbank Accession No. L02836

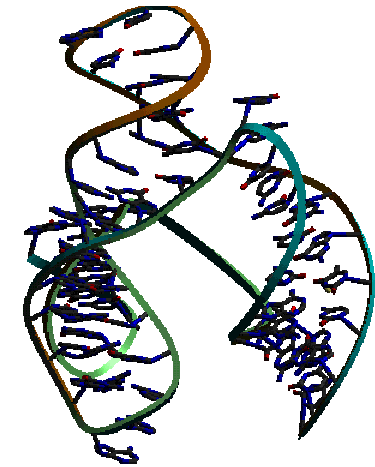
# Biological Function Follows Form

## Three Dimensional RNA Molecular Structure

Tertiary: all other  
intra-molecular  
interactions

Secondary: set of base pairs  
induced by self-bonding

Primary: linear sequence of nucleotide bases

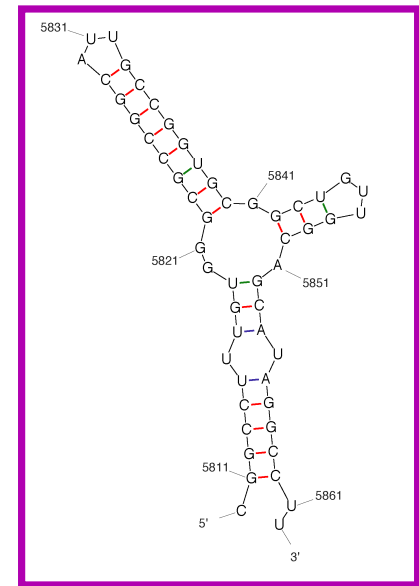
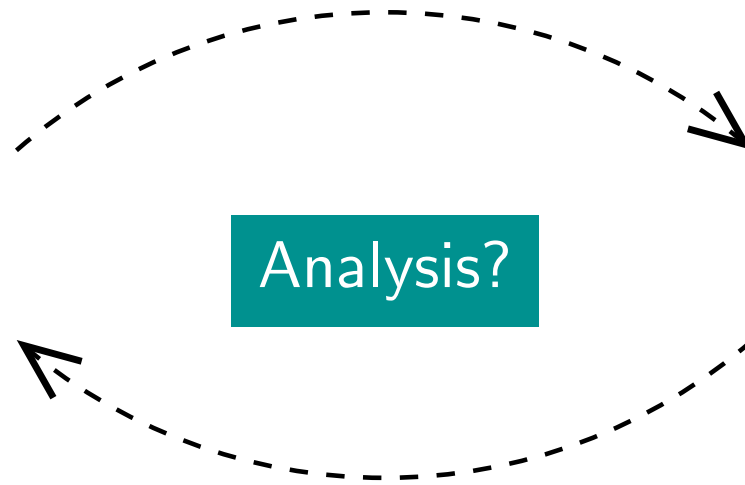


Hammerhead Ribozyme  
[www.bioinfo.rpi.edu](http://www.bioinfo.rpi.edu)

# Important Mathematical, Computational, and Biological Questions

.....CGGCC  
UUUGUGG  
GCGCCGG  
CAUUGCC  
GGUGCGG  
CUGUUGG  
CAGCAUA  
GGCCUU...

Prediction?



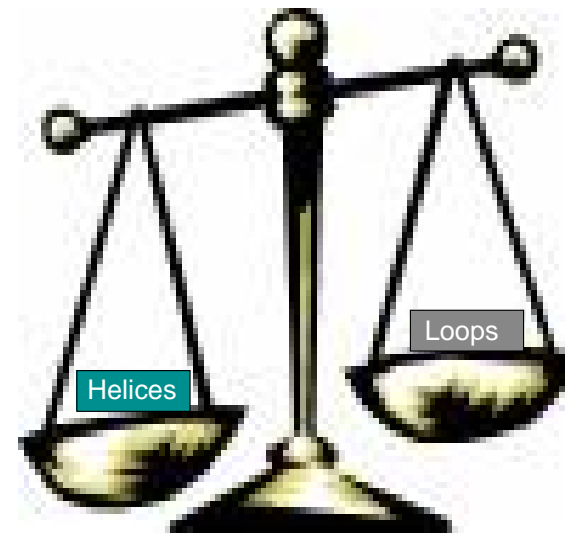
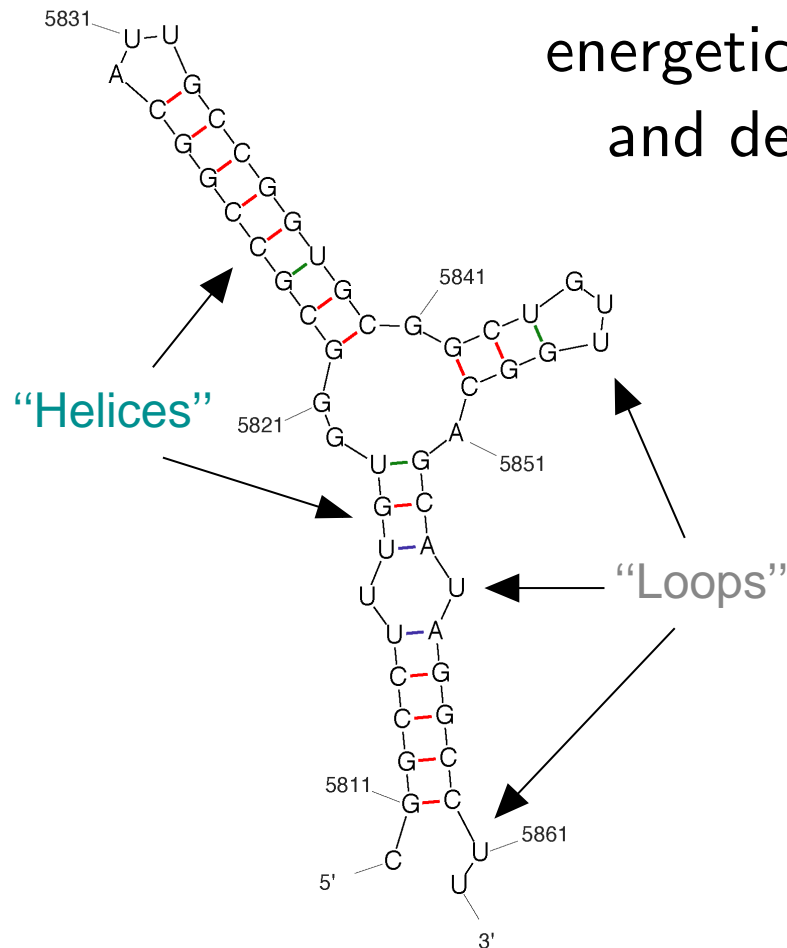
Design?

For a primary **sequence**  $R$  and predicted secondary **structure**  $S(R)$ , can we identify crucial **functional** motifs in RNA viral genomes?

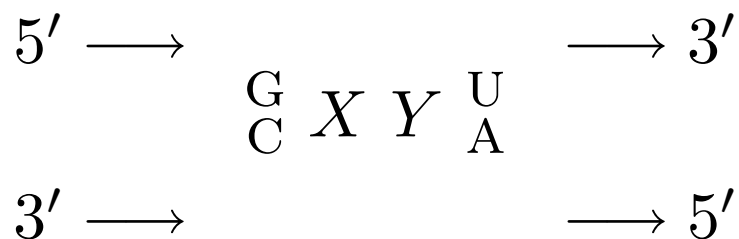


# Thermodynamics of RNA Folding

RNA secondary structures are balanced between energetically favorable **helices** (stacked base pairs) and destabilizing **loops** (single-stranded regions).



# A 2 x 2 Interior Loop Energy Table



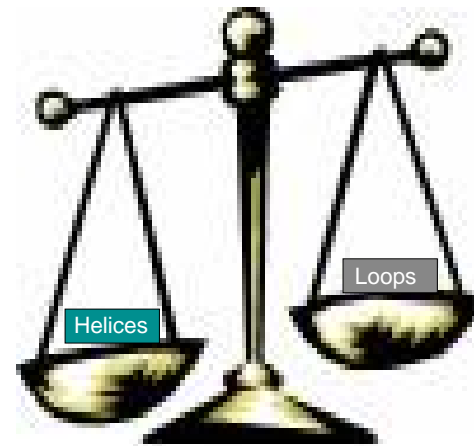
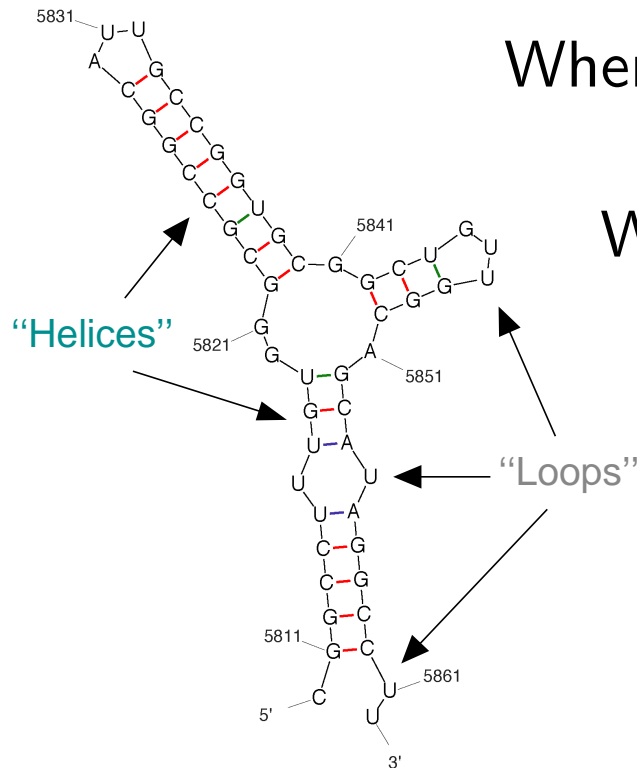
A	A	A	A	C	C	C	C	G	G	G	G	U	U	U	U
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
2.10	1.90	0.90	2.00	2.00	2.40	2.00	2.40	0.90	2.00	1.40	0.70	2.00	2.40	1.70	2.00
1.80	1.60	0.60	2.00	1.70	1.50	2.00	1.50	0.60	2.00	1.10	-0.60	2.00	1.50	0.40	0.50
0.70	0.50	-0.50	2.00	0.60	1.40	2.00	1.40	-0.50	2.00	0.00	0.10	2.00	1.40	1.10	1.50
2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1.90	1.60	0.60	2.00	1.70	1.60	2.00	1.60	0.60	2.00	1.20	-0.50	2.00	1.60	0.40	0.30
2.50	1.60	1.60	2.00	1.70	1.60	2.00	1.60	1.60	2.00	1.80	0.50	2.00	1.60	1.40	0.50
2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
2.50	1.70	1.70	2.00	1.80	1.60	2.00	1.60	1.70	2.00	1.80	0.50	2.00	1.60	1.50	0.50
0.10	-0.10	-1.10	2.00	0.00	0.80	2.00	0.80	-1.10	2.00	-0.60	-0.50	2.00	0.80	0.50	0.30
2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1.80	1.50	0.50	2.00	1.60	2.10	2.00	2.10	0.50	2.00	1.10	0.40	2.00	2.10	1.30	1.50
0.40	-0.90	-0.10	2.00	-0.80	0.10	2.00	0.10	-0.10	2.00	-0.30	-3.10	2.00	0.10	-2.10	0.50
2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1.50	0.60	0.70	2.00	0.70	0.60	2.00	0.60	0.70	2.00	0.80	-0.50	2.00	0.60	0.50	-0.50
0.00	-1.30	-0.50	2.00	-1.20	-0.30	2.00	-0.30	-0.50	2.00	-0.70	-3.50	2.00	-0.30	-2.50	-0.50
2.10	0.90	1.70	2.00	1.00	0.80	2.00	0.80	1.70	2.00	1.40	0.50	2.00	0.80	1.50	0.50

# Analyzing the Impact of Helices and Loops

For a primary sequence  $R$  and predicted secondary structure  $S(R)$ , can we identify crucial functional motifs in RNA viral genomes?

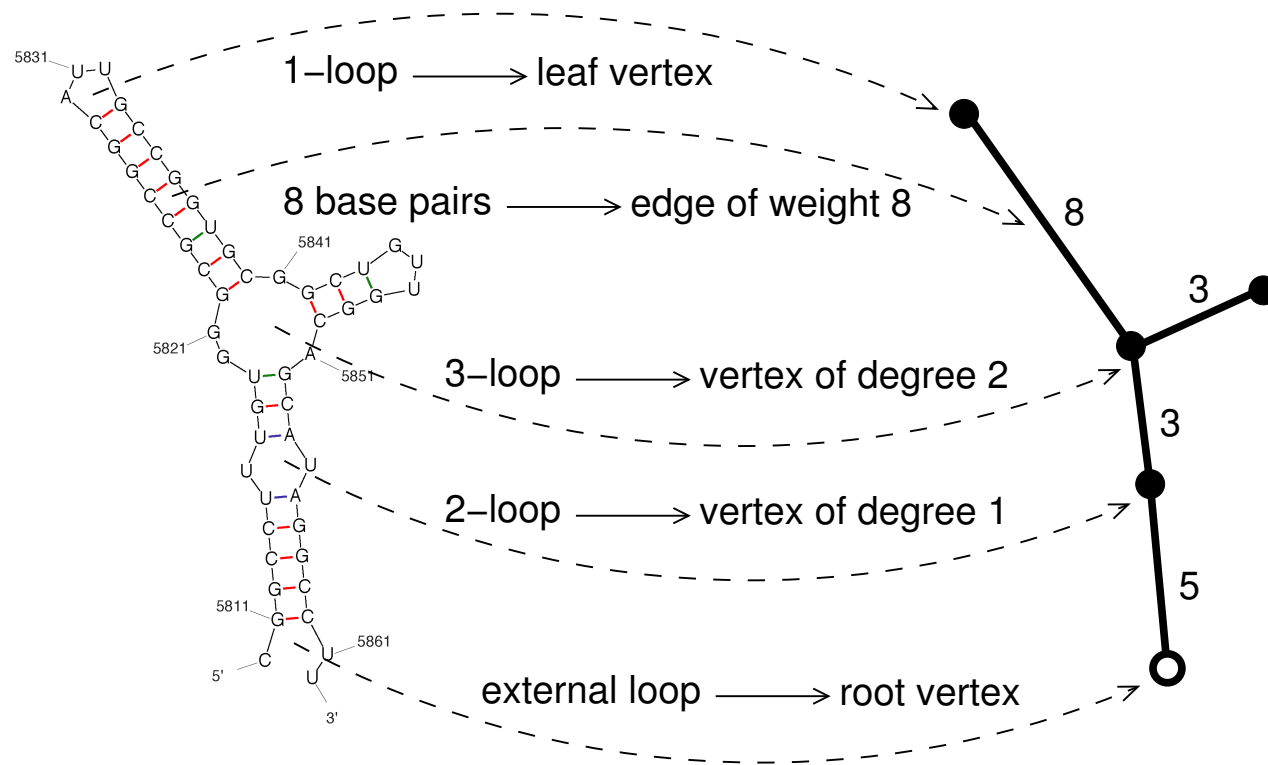
When do helices encode a unique configuration?

What configurations minimize loop energies?



# Representing RNA Secondary Structures by Trees

Abstract folded **sequence** to its graphical “skeleton” *T*:

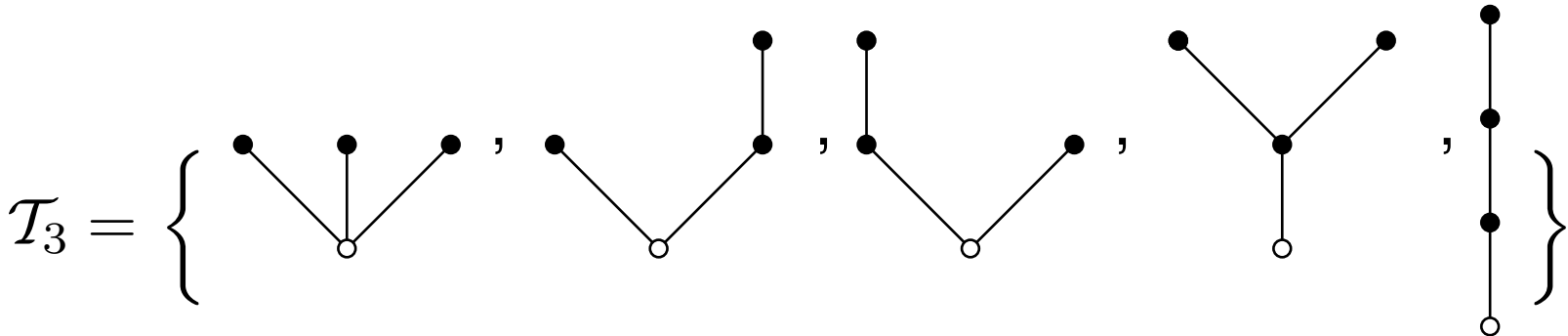


stacked base pairs  $\longrightarrow$  **edges**, single-stranded regions  $\longrightarrow$  vertices.

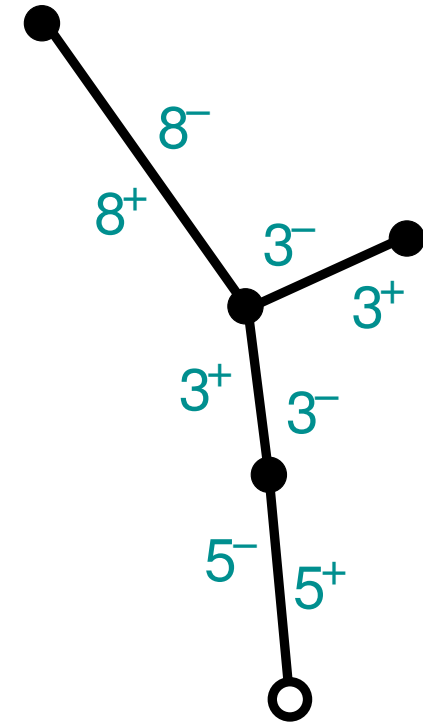
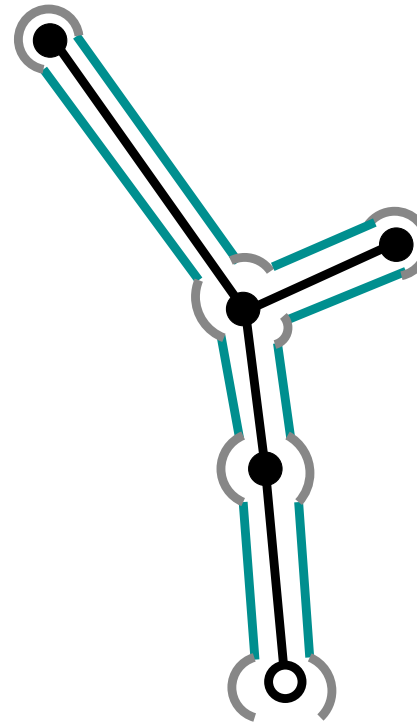
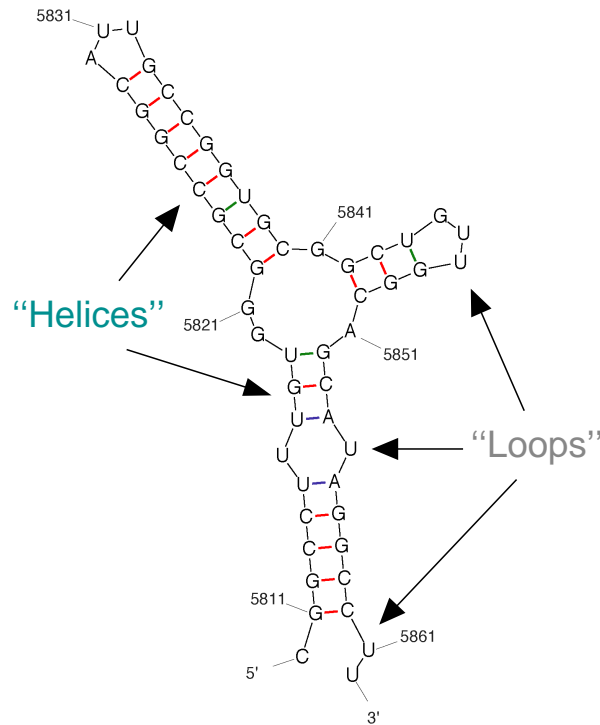
# RNA Configurations as Plane Trees

**Definition.** A **plane tree**  $T$  is a rooted tree whose subtrees at any vertex are linearly ordered. A vertex with  $k$  children has degree  $k$ .

$$\mathcal{T}_n = \{T \mid n \text{ edges (and } n + 1 \text{ vertices)}\}$$



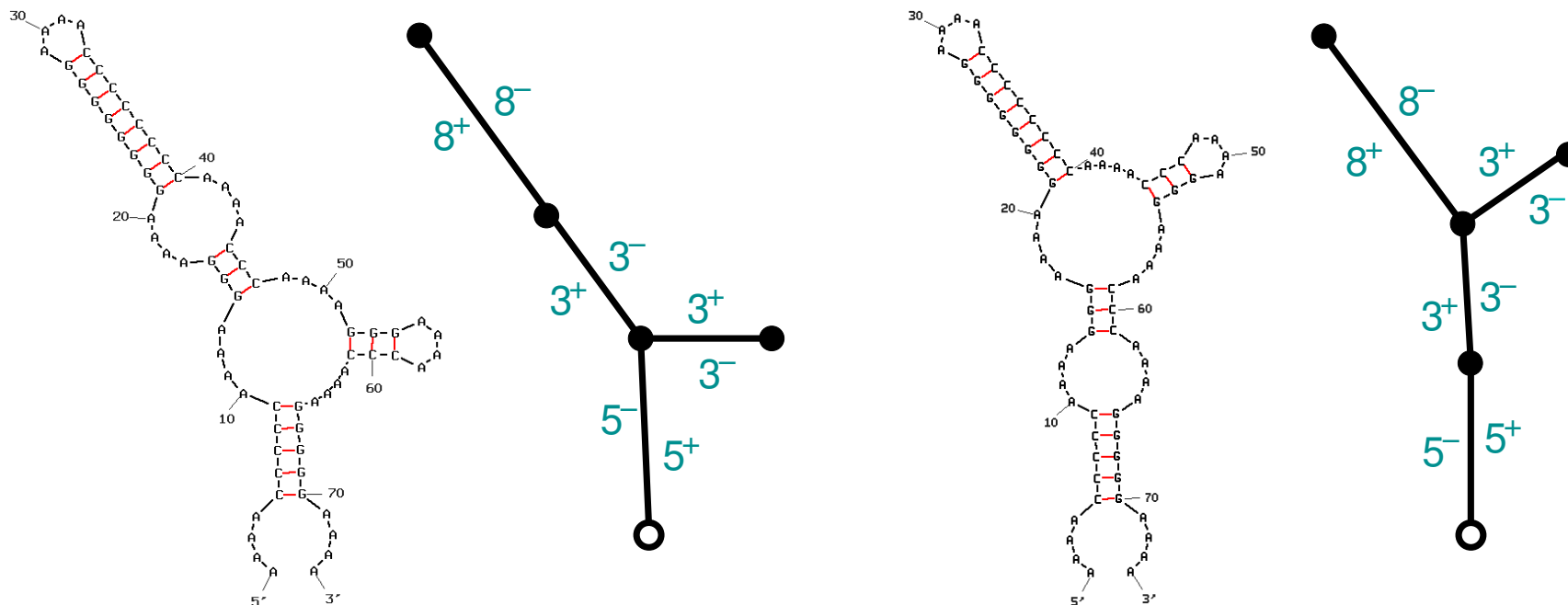
# A Combinatorial Model of RNA Folding



Let  $\rho(k^+) = \overbrace{G \dots G}^k A^4 = G^k A^4$  and  $\rho(k^-) = C^k A^4$  for  $k \in \mathbb{N}$ .  
 Consider  $R = A^4 \rho(s)$  for strings like  $s = 5^- 3^+ 8^+ 8^- 3^- 3^+ 3^- 5^+$ .

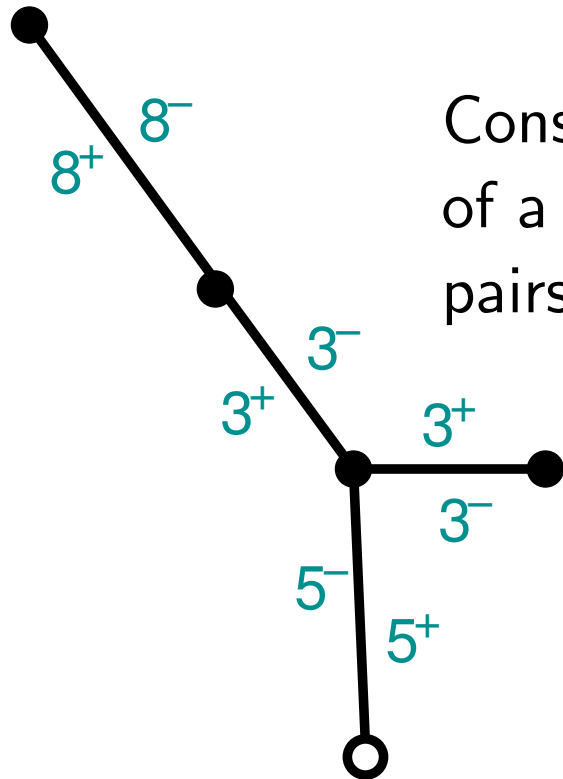
# When Do Helices Encode a Unique Configuration?

Fold  $R = A^4\rho(s)$  for string  $s = 5^- 3^+ 8^+ 8^- 3^- 3^+ 3^- 5^+$ .



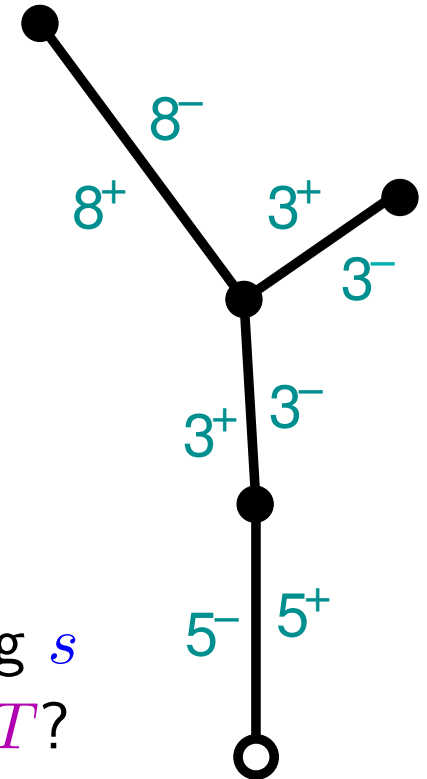
Two distinct  $T$  for two different secondary structures of  $R$ !

# Analyzing Strings Encoding Trees



Consider labeling the boundary of a plane tree  $T$  with integer pairs  $k^+, k^-$  to form a string  $s$ .

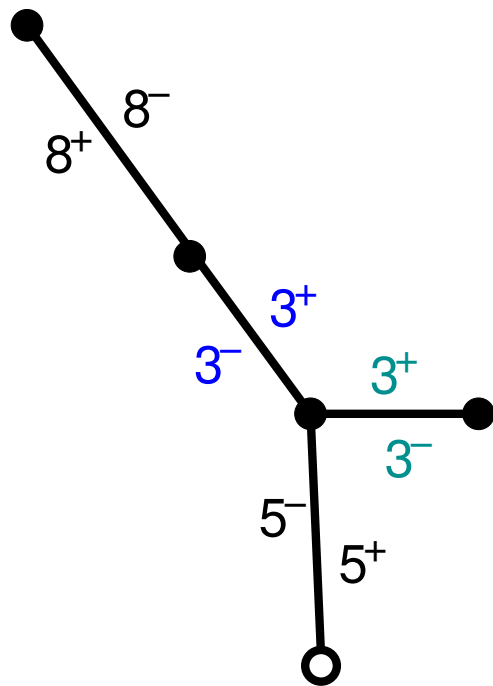
How many distinct integers  $k$  are needed to produce a string  $s$  that “folds” uniquely to tree  $T$ ?



**Idea.** Local constraints are provably sufficient for global structure.

# Local Constraints Give Global Structure

**Theorem. [H]** *Let  $m$  be the maximum number of edges incident on a vertex in a plane tree  $T$ . Then  $\lceil \frac{m}{2} \rceil$  distinct  $k$  are necessary and sufficient to produce a string  $s$  which folds uniquely to  $T$ .*



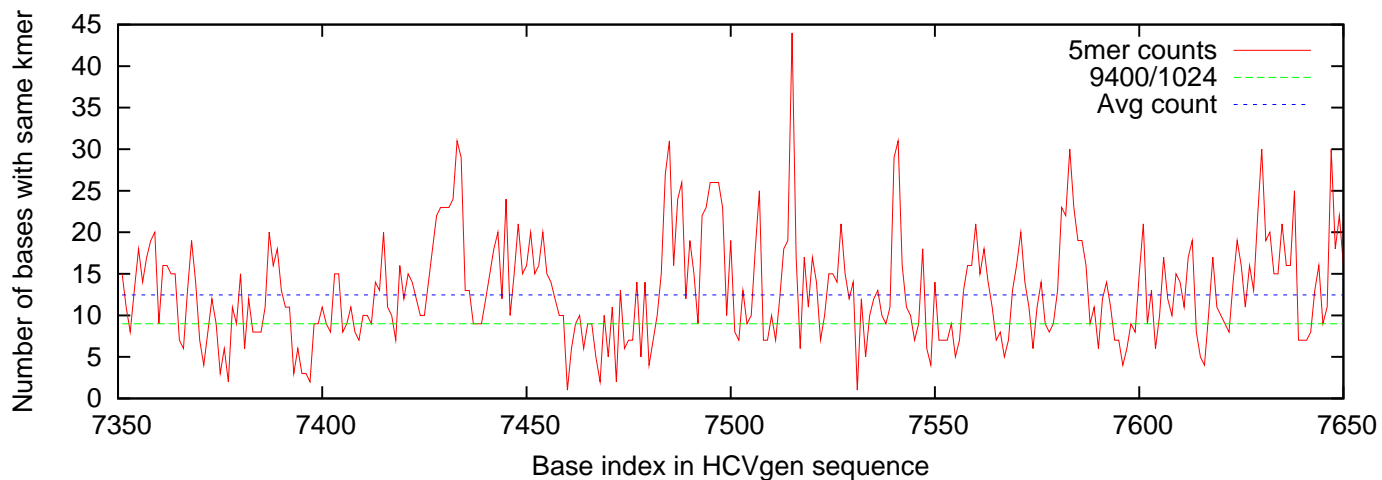
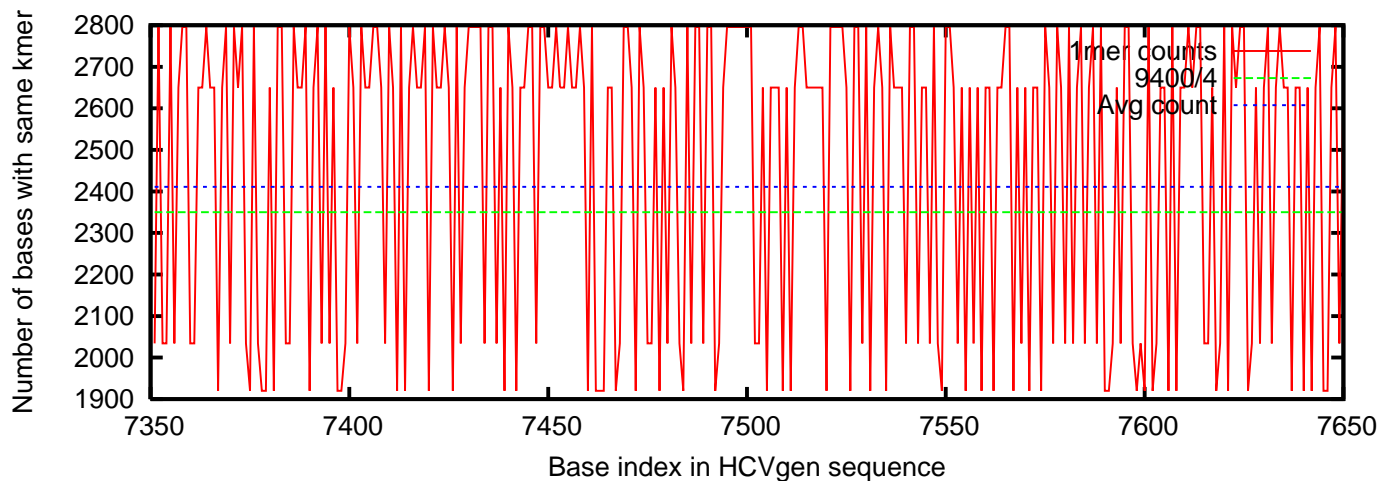
Consider  $s' = 5^- 3^- 8^+ 8^- 3^- 3^+ 3^+ 5^+$   
instead of  $s = 5^- 3^+ 8^+ 8^- 3^- 3^+ 3^- 5^+$ .

With at least  $\lceil \frac{m}{2} \rceil$  distinct integers  $k$ ,  
label the boundary of  $T$  using  
each  $k^-, k^+$  symbol at most once  
in entering (resp. exiting) a vertex.

Necessary as well as provably sufficient.

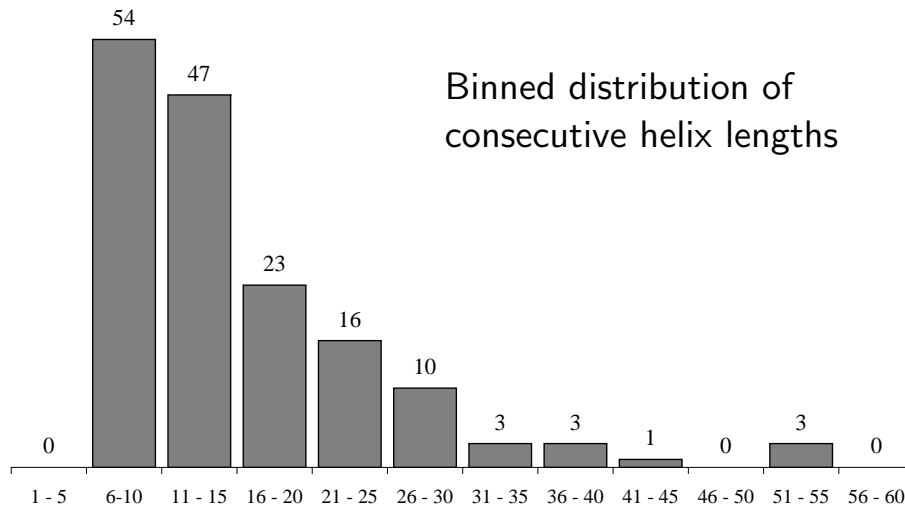
# Analyzing Helical Encodings in Hepatitis C

Annotate bases with the count of repeated kmers,  $k = 1, \dots, 5$ .



# Helical Lengths and Kmer Repeats

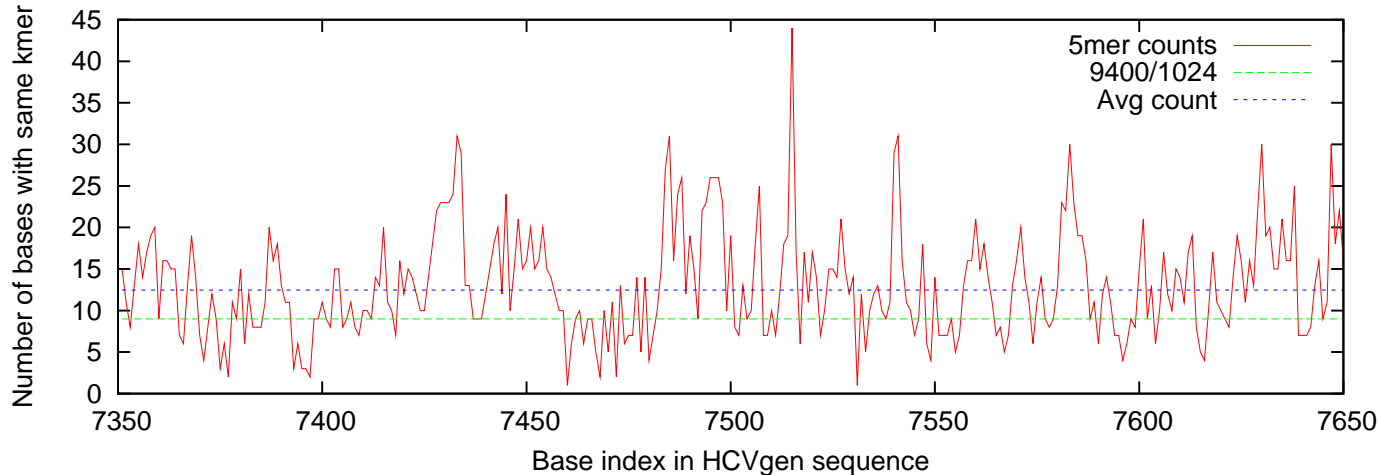
istem#	istart	jstart	#bp	Kcal	jstem#	Floop	Bloop	closing	Hnum
552	7434	7574	4	-6.9	146	1	2	548	190.8
553	7439	7569	2	-2.1	147	1	1	552	208.5
554	7443	7566	7	-16.1	148	1	1	553	29.6
555	7451	7557	2	-1.9	149	1	1	554	29
556	7454	7553	4	-6.8	150	1	1	555	22.8
557	7459	7547	3	-4.4	151	1	1	556	18
558	7464	7543	8	-13.8	152	1	1	557	4.8
559	7473	7534	7	-10.2	153	1	1	558	5.9
560	7480	7526	3	-4.3	154	1	1	559	13
561	7486	7520	6	-10.9	155	1	1	560	27.5
562	7494	7513	2	-3.4	156	1	1	561	28
563	7496	7510	1	0	157	1	1	562	28
564	7497	7508	3	-5.8	158	0	1	563	27.7



$L > 1$	Helix	Consec. Helix
#	643	160
avg $L$	4.84	15.89
std $L$	2.26	9.07

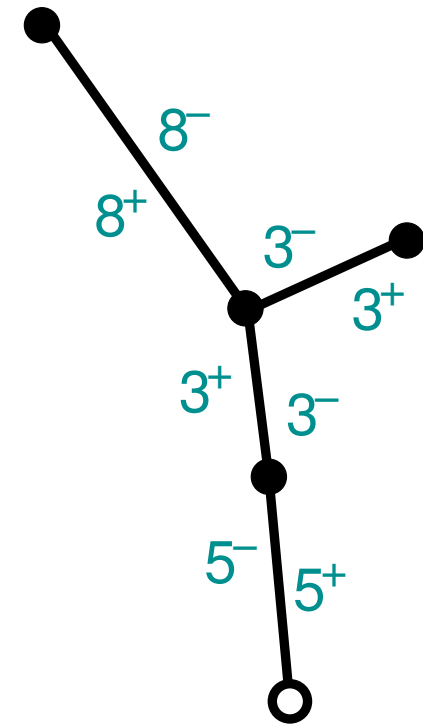
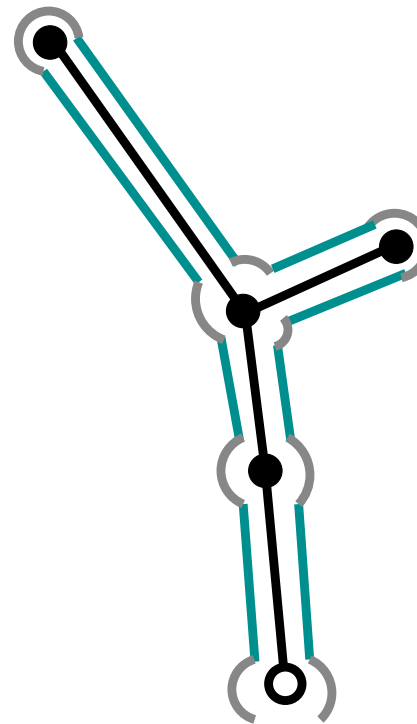
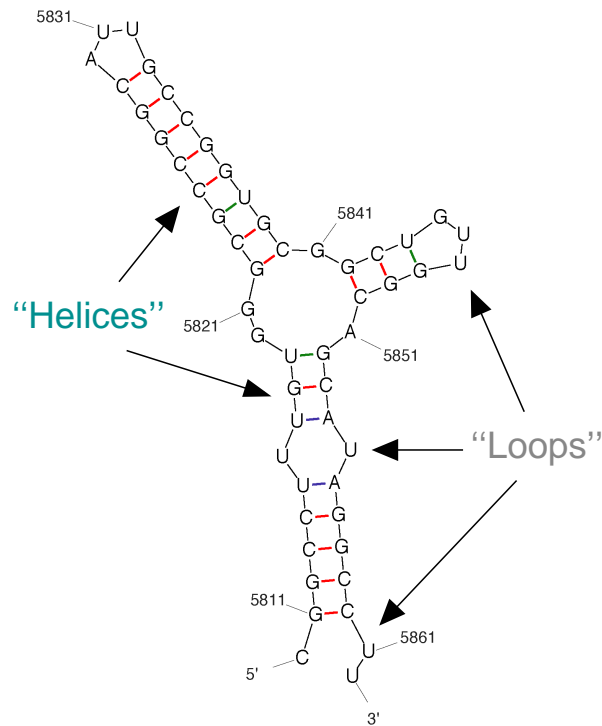
# Towards Understanding Helix Encodings

**Combinatorial Result.** Local helical constraints are necessary and sufficient for the folding of global structure.



**Computational Analysis.** A large helical substructure in HCVgen sequence was identified by the uniqueness of its kmer parsing.

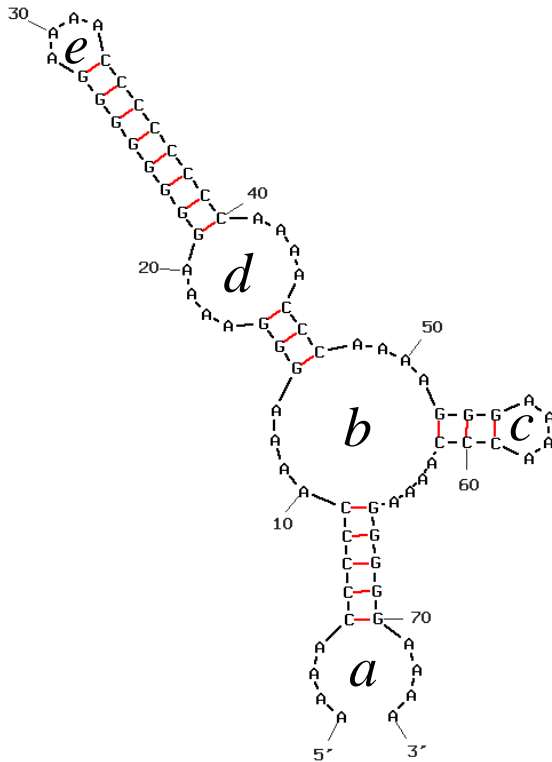
# A Combinatorial Model of RNA Folding



Let  $\rho(k^+) = \overbrace{G \dots G}^{k \text{ times}} A^4 = G^k A^4$  and  $\rho(k^-) = C^k A^4$  for  $k \in \mathbb{N}$ .  
 Consider  $R = A^4 \rho(s)$  for strings like  $s = 5^- 3^+ 8^+ 8^- 3^- 3^+ 3^- 5^+$ .

# What Configurations Minimize Loop Energies?

Loop free energy decomposition calculated by Zuker's `mfold` algorithm using Turner thermodynamic values.



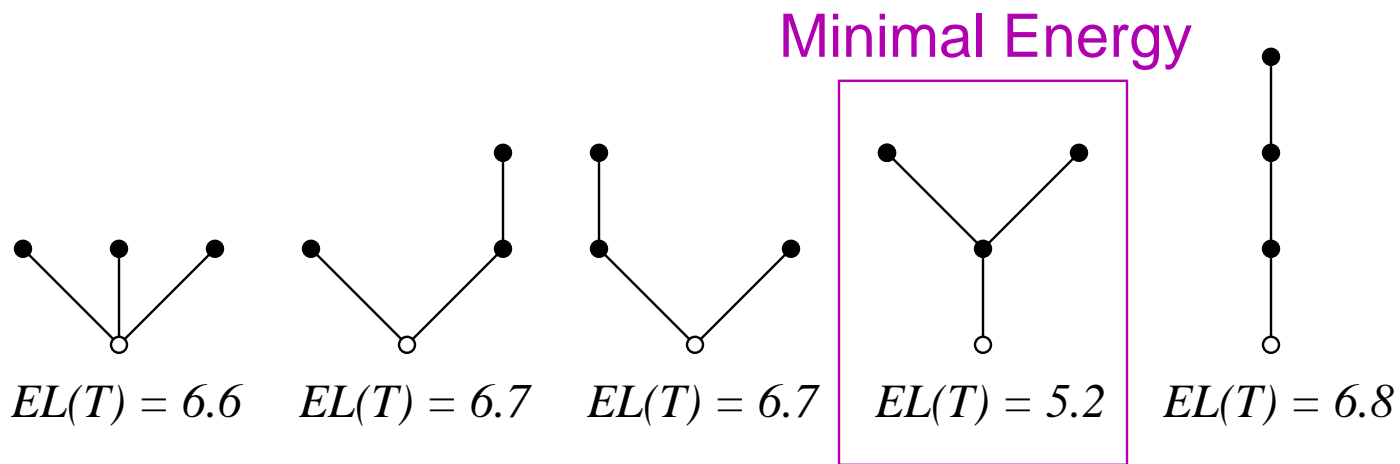
Loop	Type	# bp	$\Delta G$
<i>a</i>	external	1	-1.60
<i>b</i>	multibranch	3	-1.10
<i>c</i>	hairpin	1	4.50
<i>d</i>	interior	2	2.30
<i>e</i>	hairpin	1	4.50

An internal  $k$ -loop contains  $k$  base pairs.

# Analyzing Vertex Energies

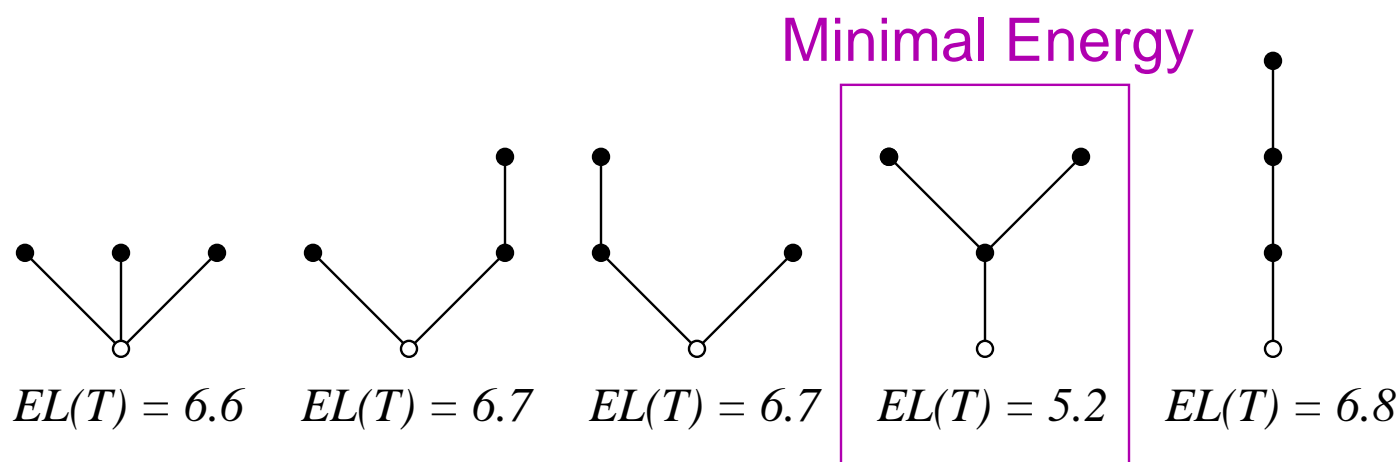
For a tree  $T$ , let  $EL(T)$  be the sum of the vertex energies.

Vertex degree	0	1	$k - 1, k \geq 3$	root, degree $j \geq 1$
Related loop	1-loop	2-loop	$k$ -loop, $k \geq 3$	external
Minimal energy	4.10	2.30	$3.40 - 1.50 k$	$-1.90 j$



Plane trees  $T$  with 3 edges and their total loop energies  $EL(T)$ .

# Minimal Loop Energy Configurations



Plane trees  $T$  with 3 edges and their total loop energies  $EL(T)$ .

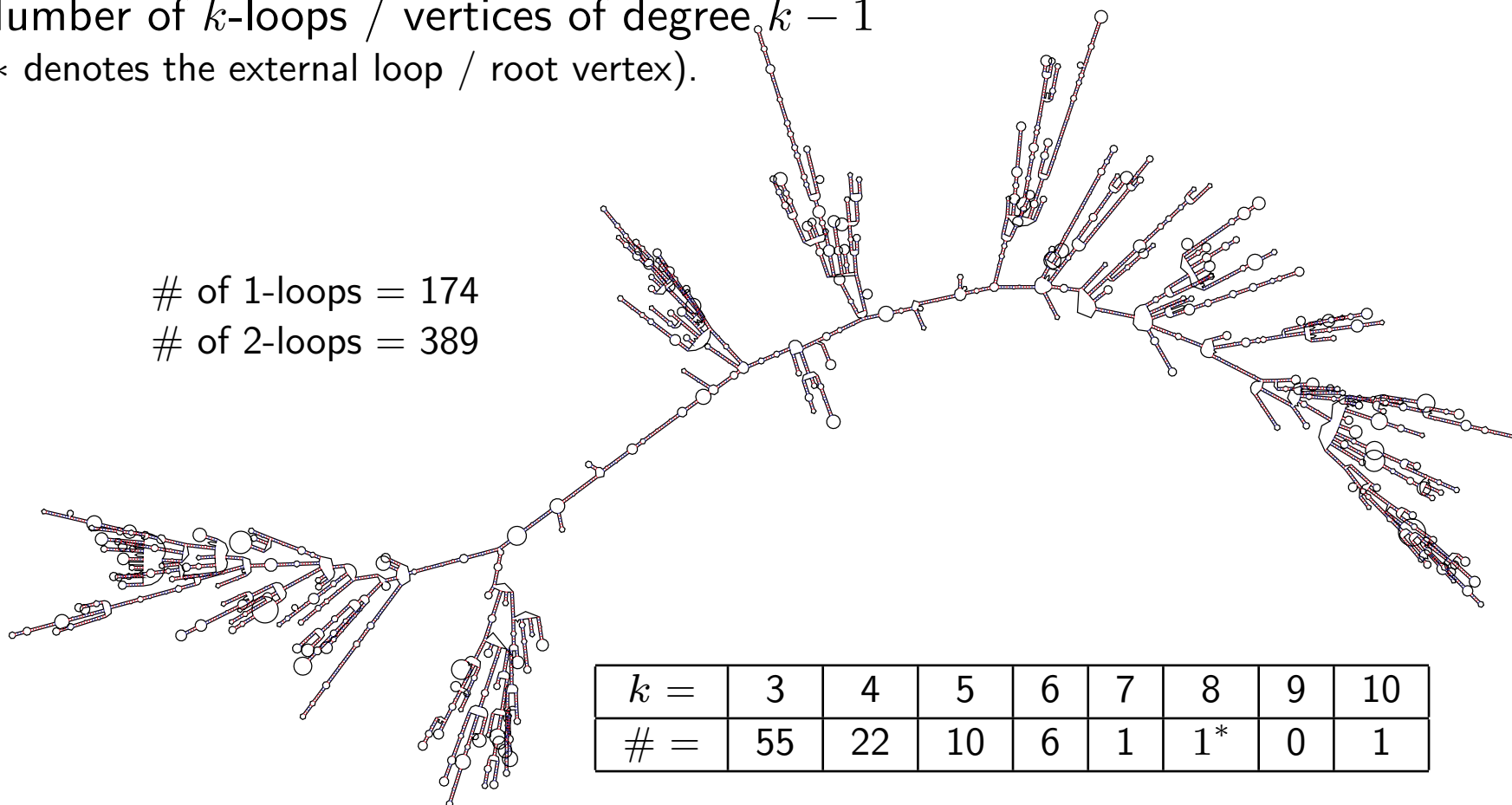
**Theorem. [H]** For plane trees  $T$  with  $n$  edges, the total loop energy  $EL(T)$  is minimal when  $T$  has the maximal number of vertices with degree 2. (When  $n$  is odd, the root has degree 1.)

# Branching Degree in Hepatitis C

For Hepatitis C, the majority of branching loops do have degree 2!

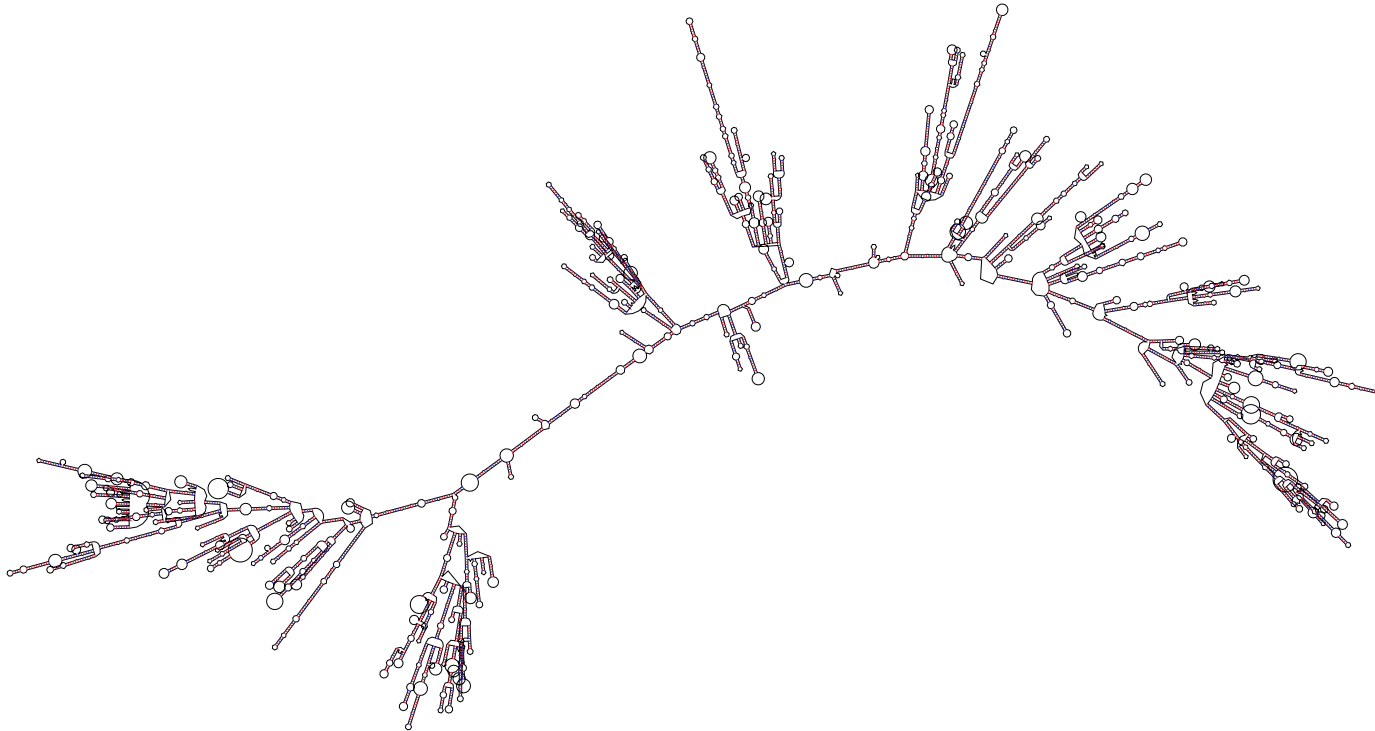
Number of  $k$ -loops / vertices of degree  $k - 1$   
(\* denotes the external loop / root vertex).

# of 1-loops = 174  
# of 2-loops = 389



# Towards Understanding Loop Impact

**Combinatorial Result.** Associated loop energies are minimized by maximizing vertices with three edges.

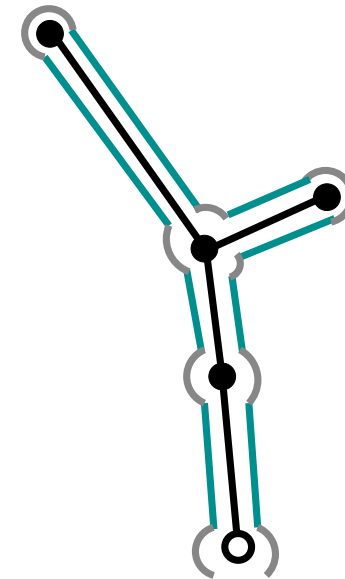
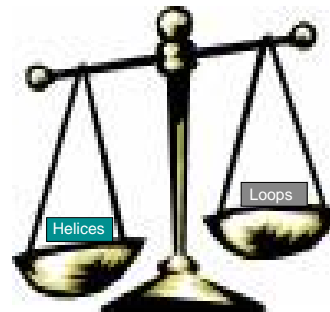
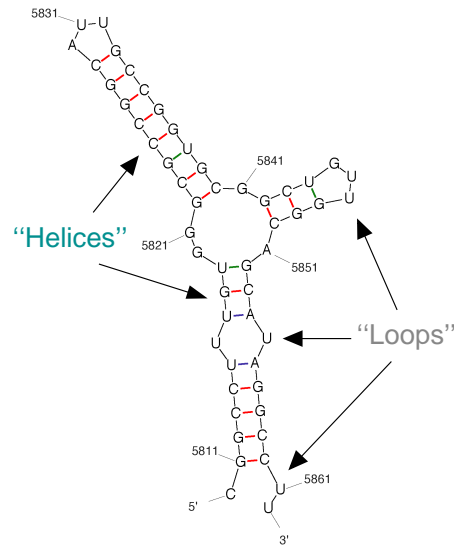


**Computational Analysis.** HCVgen predicted secondary structure is consistent with an overall energetic cost for loop branching.

# Current & Future Research: Analyzing RNA Viral Functional Motifs

**Result 1.** Local helical constraints are necessary and sufficient for folding of global structure.

**Hypothesis 1.** Well-determined viral RNA substructures have high helical encoding quality.

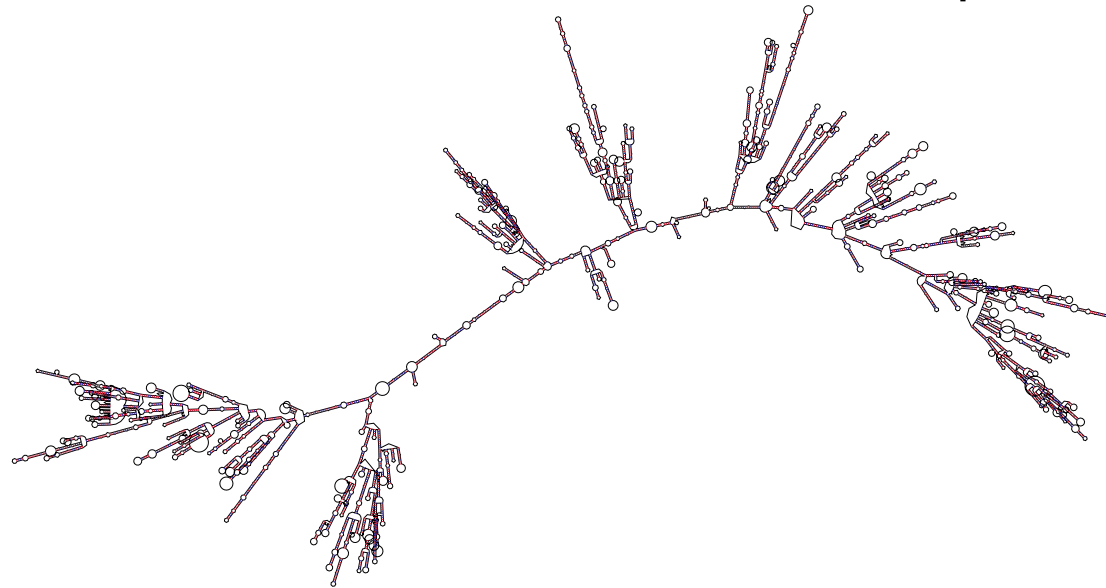


**Result 2.** Associated loop energies are minimized by maximizing vertices with three edges.

**Hypothesis 2.** Branching degree in viral RNA loops correlates with functional significance.

# In Conclusion

Understanding the **structure** and **function** of RNA viral **genomes** is a fundamental scientific question.



Combinatorial analysis suggests new insights into the folding of RNA **secondary structures** and new directions for the computational identification of **functional motifs**.

# Acknowledgments

- Lior Pachter, Bernd Sturmfels, Seth Sullivant, and James Carlson.
- Hepatitis C Secondary Structure Prediction data made publicly available by Prof. Ann C. Palmenberg and Dr. Jean–Yves Sgro, UW Madison:  
[http://virology.wisc.edu/acp/RNAFolds/RNAFolds\\_hep.html](http://virology.wisc.edu/acp/RNAFolds/RNAFolds_hep.html).
- Predicted RNA foldings courtesy of Michael Zuker's mfold algorithm.
- RNA analysis performed using MySQL and CocoaMySQL.
- Currently supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.
- Previously supported by NLM grant #T15 LM07359, "Computation and Informatics in Biology and Medicine."
- Anne Condon and Holger Hoos, University of British Columbia.