Deciphering the Information Encoded in RNA Viral Genomes

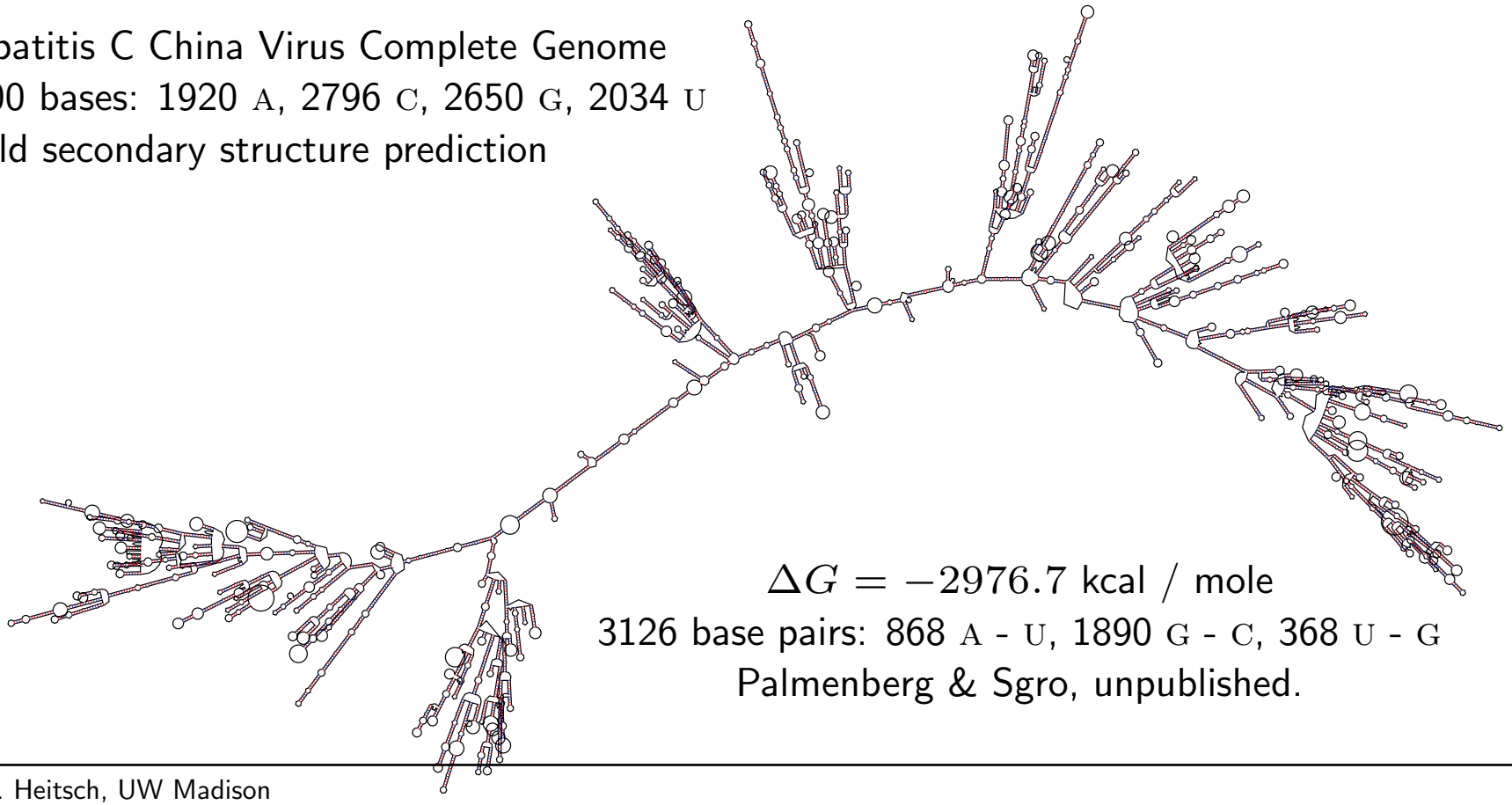
Christine E. Heitsch
Genome Center of Wisconsin and Mathematics Department
University of Wisconsin – Madison

Detecting and Processing Regularities in High Throughput Biological Data
June 21, 2005

RNA Sequences Encode Molecular Structures

Single-stranded RNA viral genomes form secondary structures.
Selective base pair hybridization $\Longleftrightarrow$ structure and function.

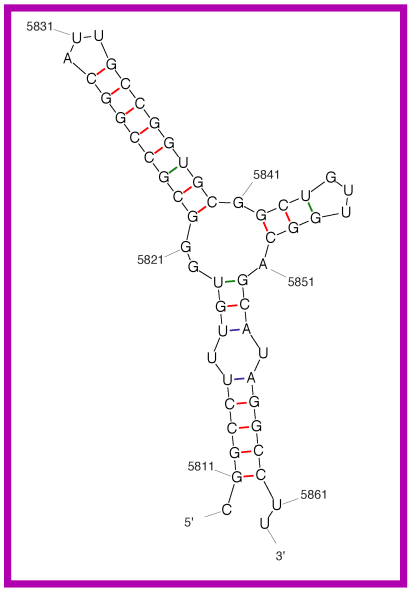
Hepatitis C China Virus Complete Genome
9400 bases: 1920 A, 2796 C, 2650 G, 2034 U
Mfold secondary structure prediction



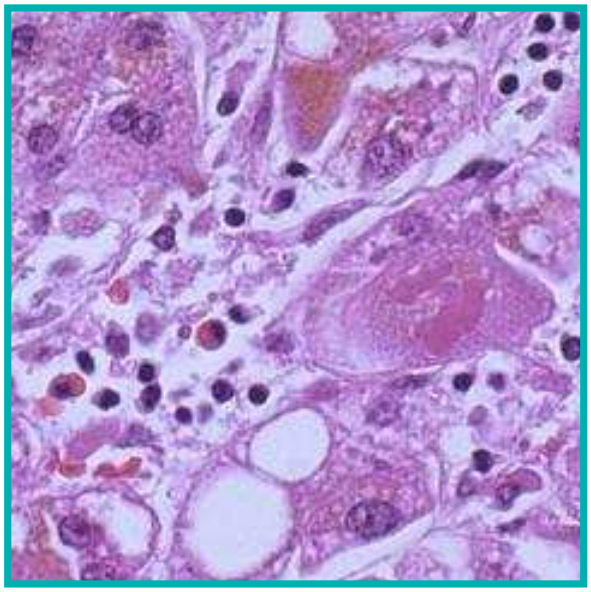
$\Delta G = -2976.7$ kcal / mole
3126 base pairs: 868 A - U, 1890 G - C, 368 U - G
Palmenberg & Sgro, unpublished.

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...

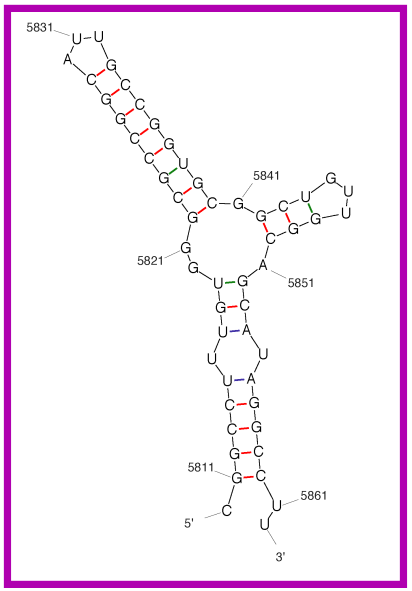
Segment of Hepatitis C
viral RNA genome
Genbank Accession No. L02836

RNA Base Pairing Yields Structure

Primary sequence R

Secondary structure $S(R)$

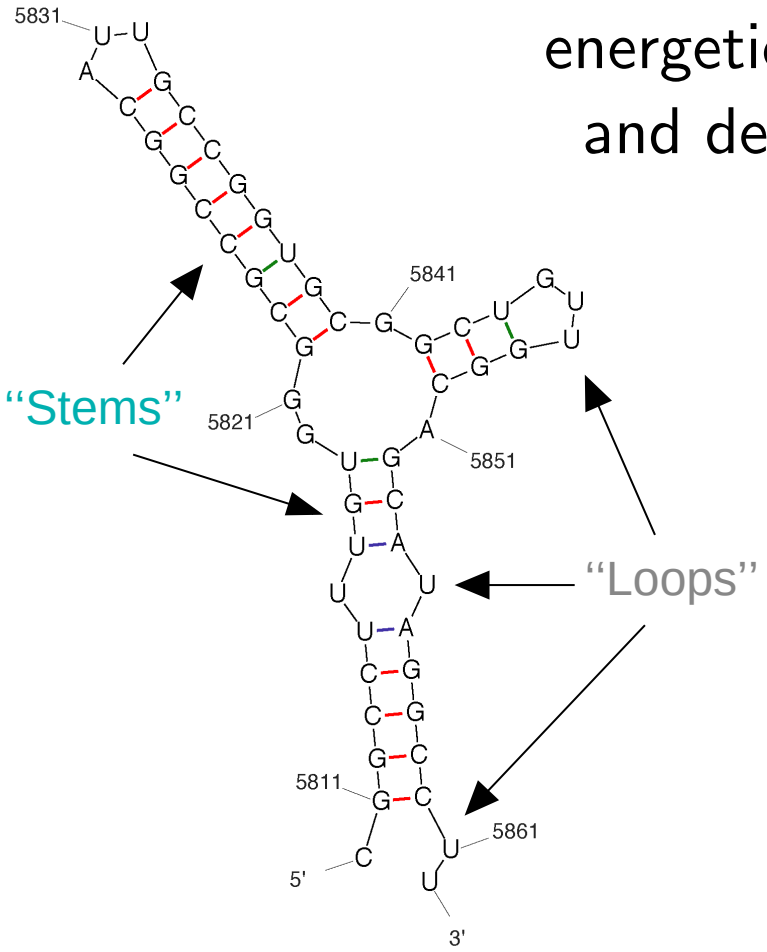
.....CGGCC
UUUGUGG
GCGCCGG
CAUUGCC
GGUGCGG
CUGUUGG
CAGCAUA
GGCCUU...

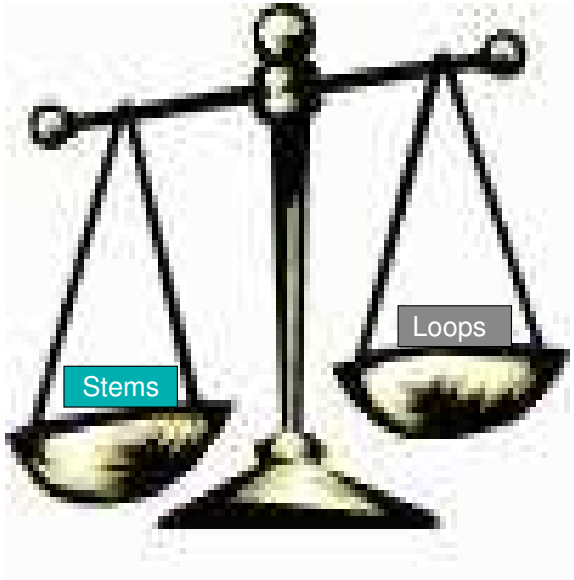


For a primary sequence R and predicted secondary structure $S(R)$, can we identify crucial functional motifs in RNA viral genomes?

Thermodynamics of RNA Folding

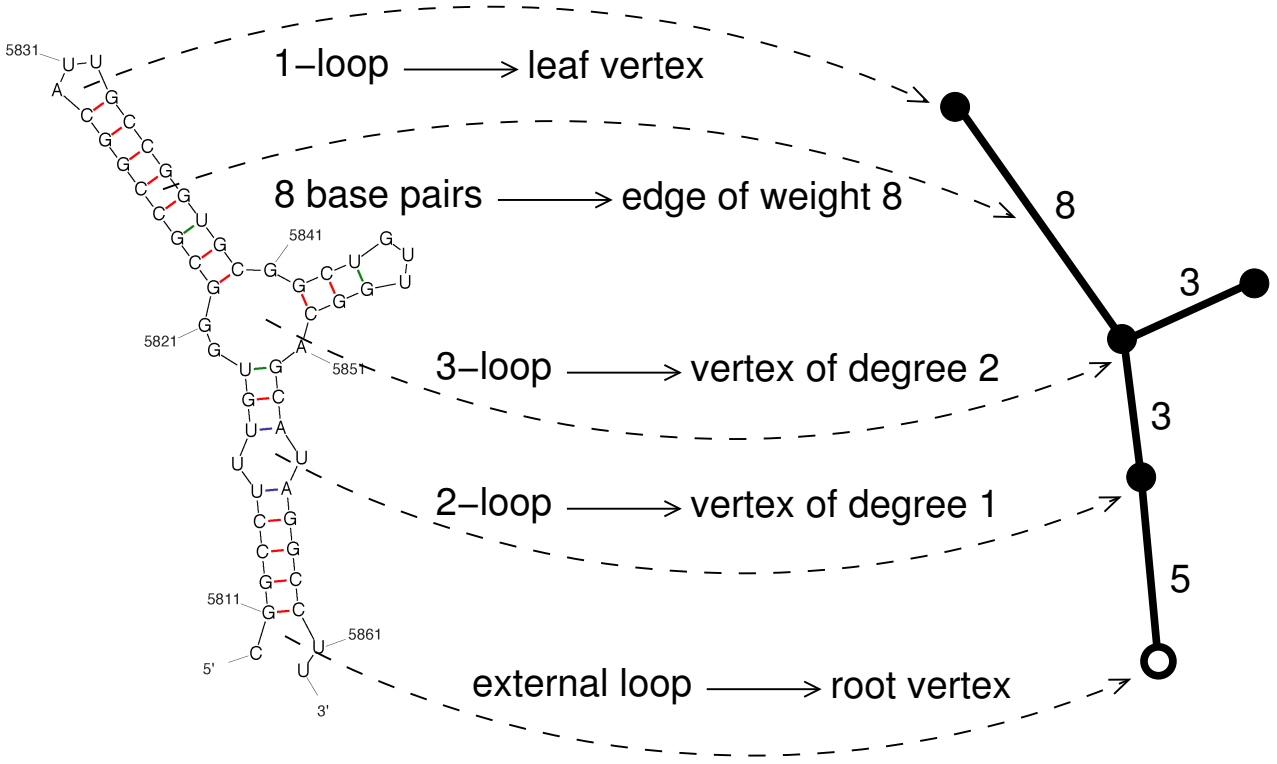
RNA secondary structures are balanced between energetically favorable stems (stacked base pairs) and destabilizing loops (single-stranded regions).





Representing RNA Secondary Structures by Trees

Abstract folded sequence to its graphical "skeleton" T:

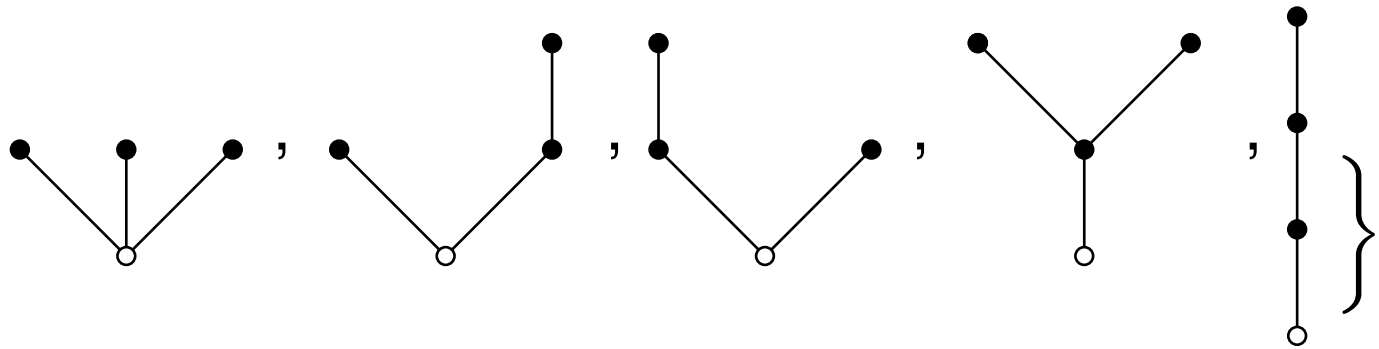


stacked base pairs ⟶ edges, single-stranded regions ⟶ vertices.

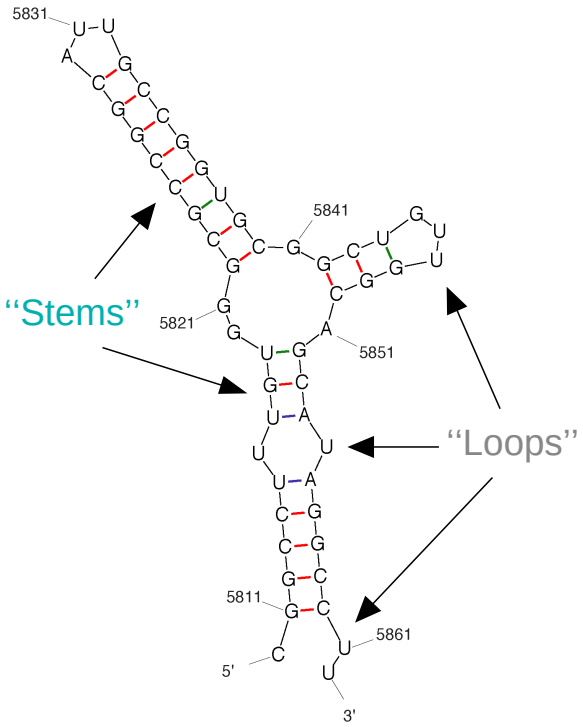
Plane Trees

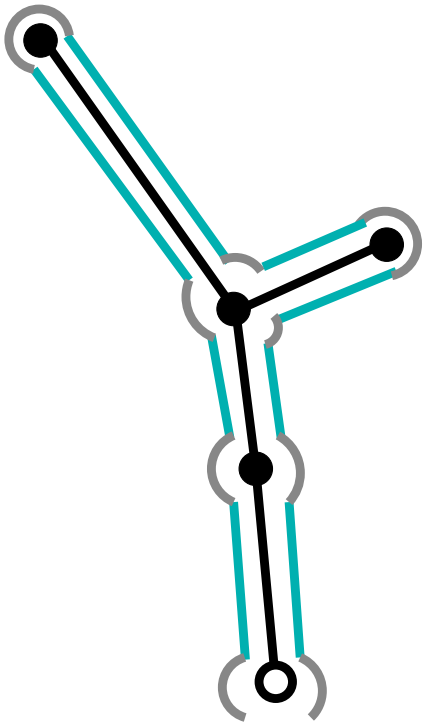
Definition. *A plane tree T is a rooted tree whose subtrees at any vertex are linearly ordered. A vertex with k children has degree k.*

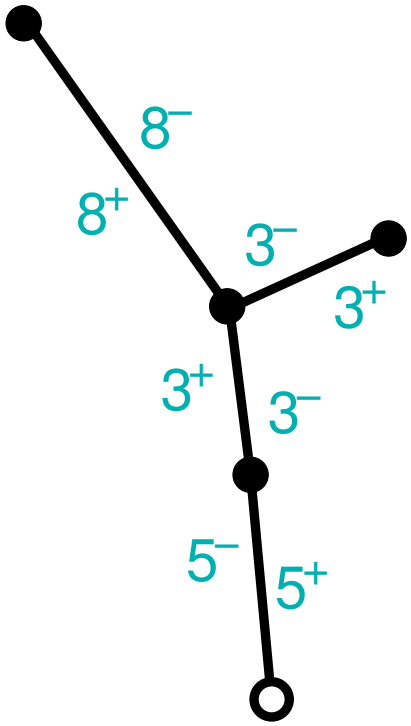
$$\mathcal{T}_n = \{T \mid n \text{ edges (and } n+1 \text{ vertices)}\}$$

$\mathcal{T}_3 =$ 

A Combinatorial Model of RNA Folding





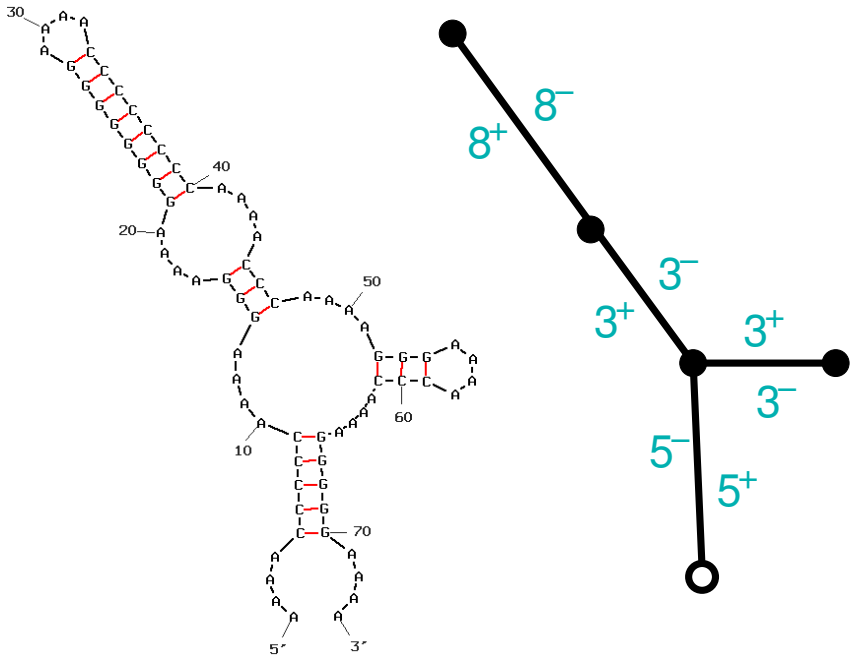


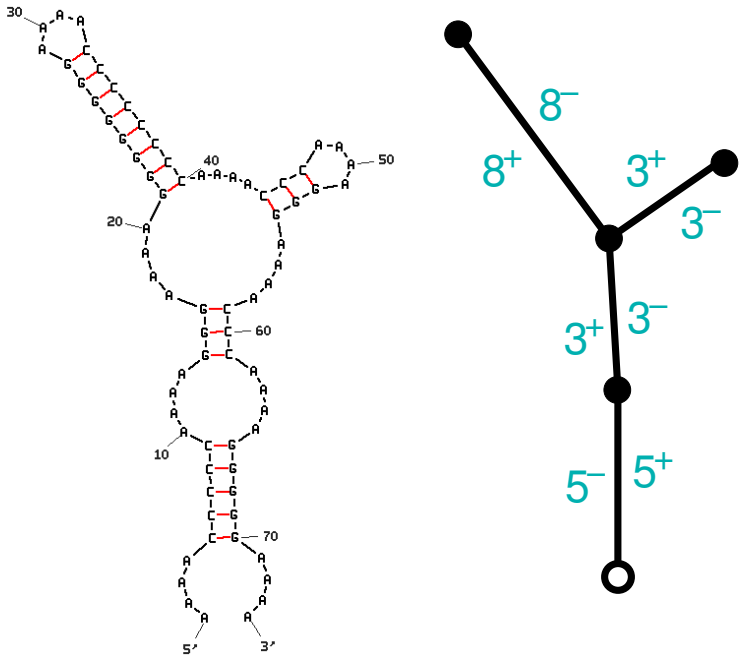
Let $\rho(k^+) = \overbrace{\mathrm{G}\ldots\mathrm{G}}^{k \text{ times}}\,\mathrm{AAAA} = \mathrm{G}^k\mathrm{A}^4$ and $\rho(k^-) = \mathrm{C}^k\mathrm{A}^4$ for $k \in \mathbb{N}$.

Consider $R = \mathrm{A}^4\rho(s)$ for strings like $s = 5^-\,3^+\,8^+\,8^-\,3^-\,3^+\,3^-\,5^+$.

How Do Helices Encode Structure?

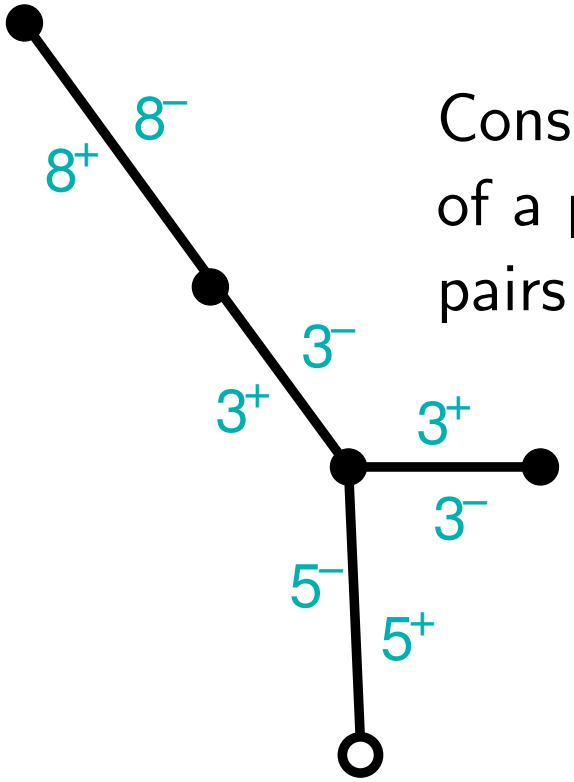
Consider $R = \mathrm{A}^4\rho(s)$ for strings like $s = 5^- 3^+ 8^+ 8^- 3^- 3^+ 3^- 5^+$.





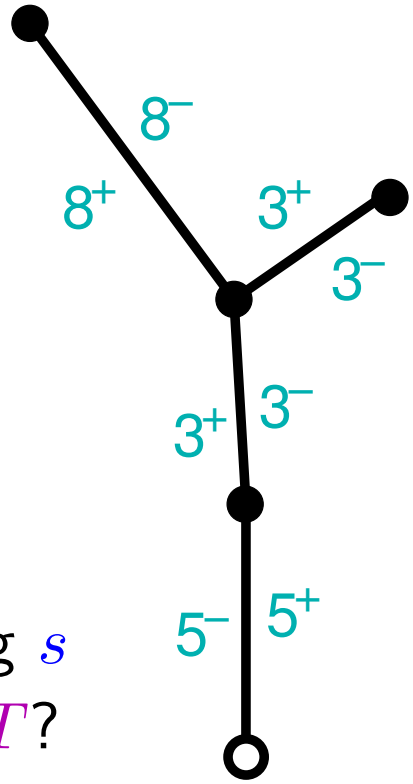
Two distinct T for two different secondary structures of R!

Analyzing Strings Encoding Trees



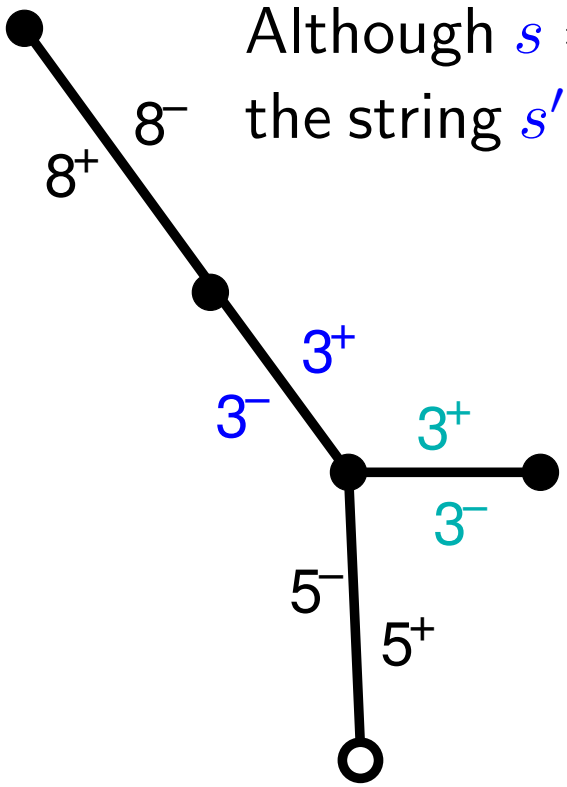
Consider labeling the boundary of a plane tree T with integer pairs k^+, k^- to form a string s.

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



Theorem. [H] *Let m be the maximum number of edges incident on a vertex in T. Then $\lceil \frac{m}{2} \rceil$ distinct k are necessary and sufficient.*

Local Constraints Give Global Structure

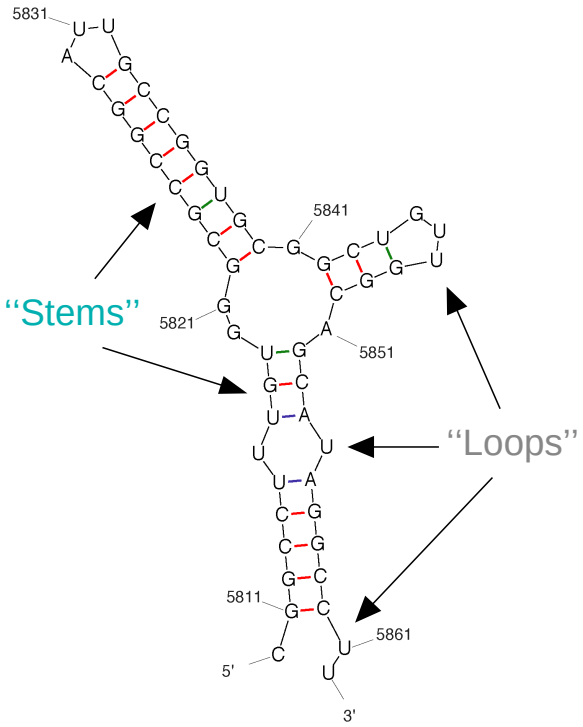


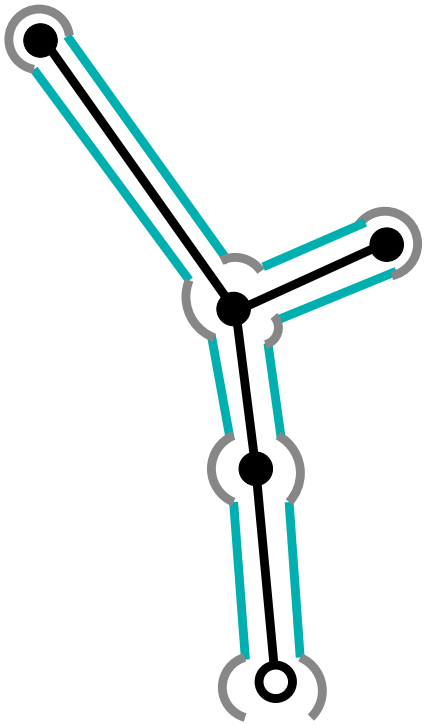
Although $s = 5^- \, 3^+ \, 8^+ \, 8^- \, 3^- \, 3^+ \, 3^- \, 5^+$ has two foldings, the string $s' = 5^- \, 3^- \, 8^+ \, 8^- \, 3^- \, 3^+ \, 3^+ \, 5^+$ has exactly one!

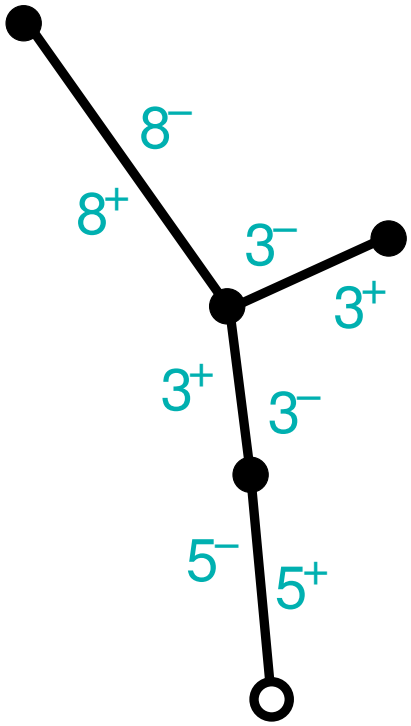
With at least $\lceil \frac{m}{2} \rceil$ distinct integers k, label the boundary of T using each k^-, k^+ symbol at most once in entering (resp. exiting) a vertex.

The resulting string s refolds uniquely to T.

A Combinatorial Model of RNA Folding



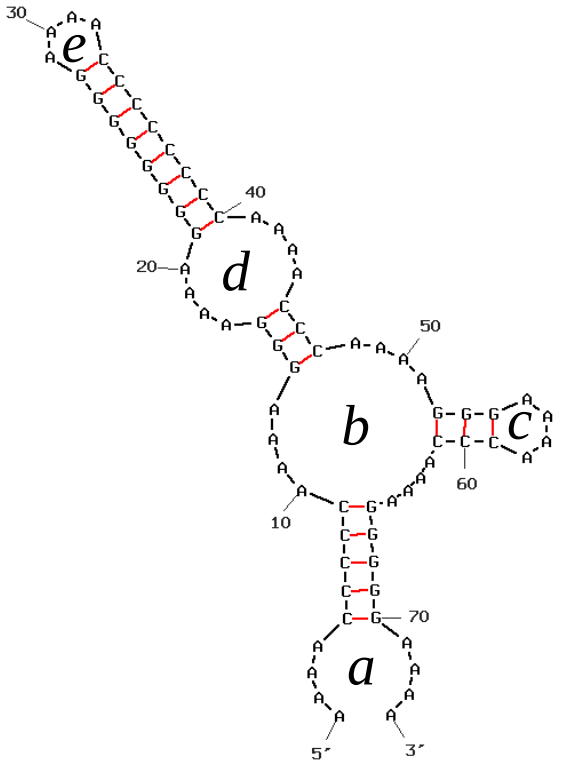




Let $\rho(k^+) = \overbrace{\mathrm{G}\ldots\mathrm{G}}^{k \text{ times}}\,\mathrm{AAAA} = \mathrm{G}^k\mathrm{A}^4$ and $\rho(k^-) = \mathrm{C}^k\mathrm{A}^4$ for $k \in \mathbb{N}$.

Consider $R = \mathrm{A}^4\rho(s)$ for strings like $s = 5^-\,3^+\,8^+\,8^-\,3^-\,3^+\,3^-\,5^+$.

Contradictory Loop Energies?



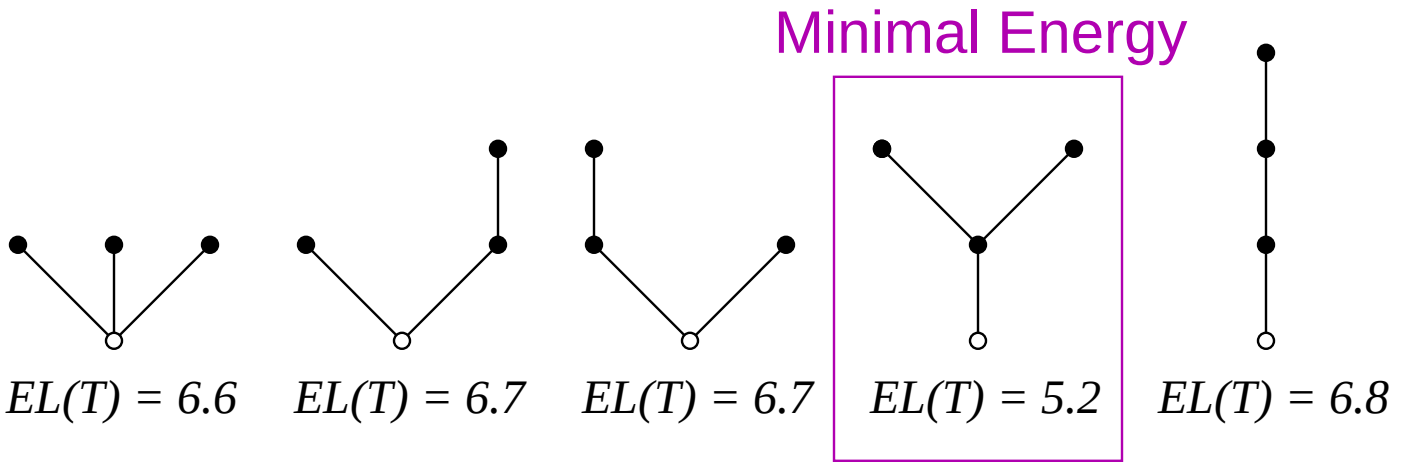
Loop free energy decomposition

Loop	Type	k	ΔG
a	external	n/a	-1.60
b	multibranch	3	-1.10
c	hairpin	1	4.50
d	interior	2	2.30
e	hairpin	1	4.50

Analyzing Vertex Energies

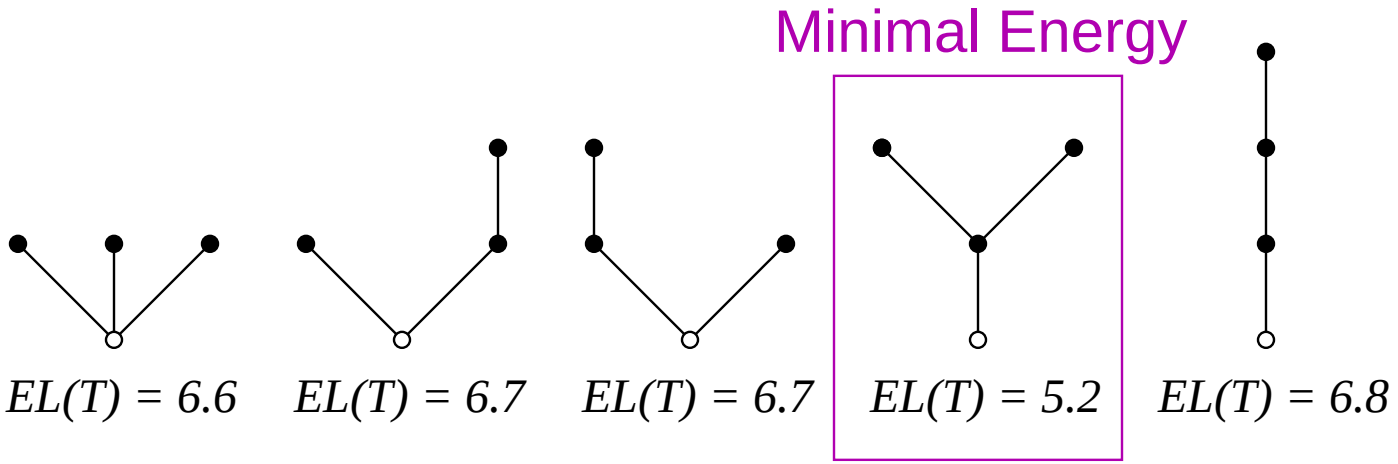
For a tree T, let $EL(T)$ be the sum of the energy for each vertex.

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



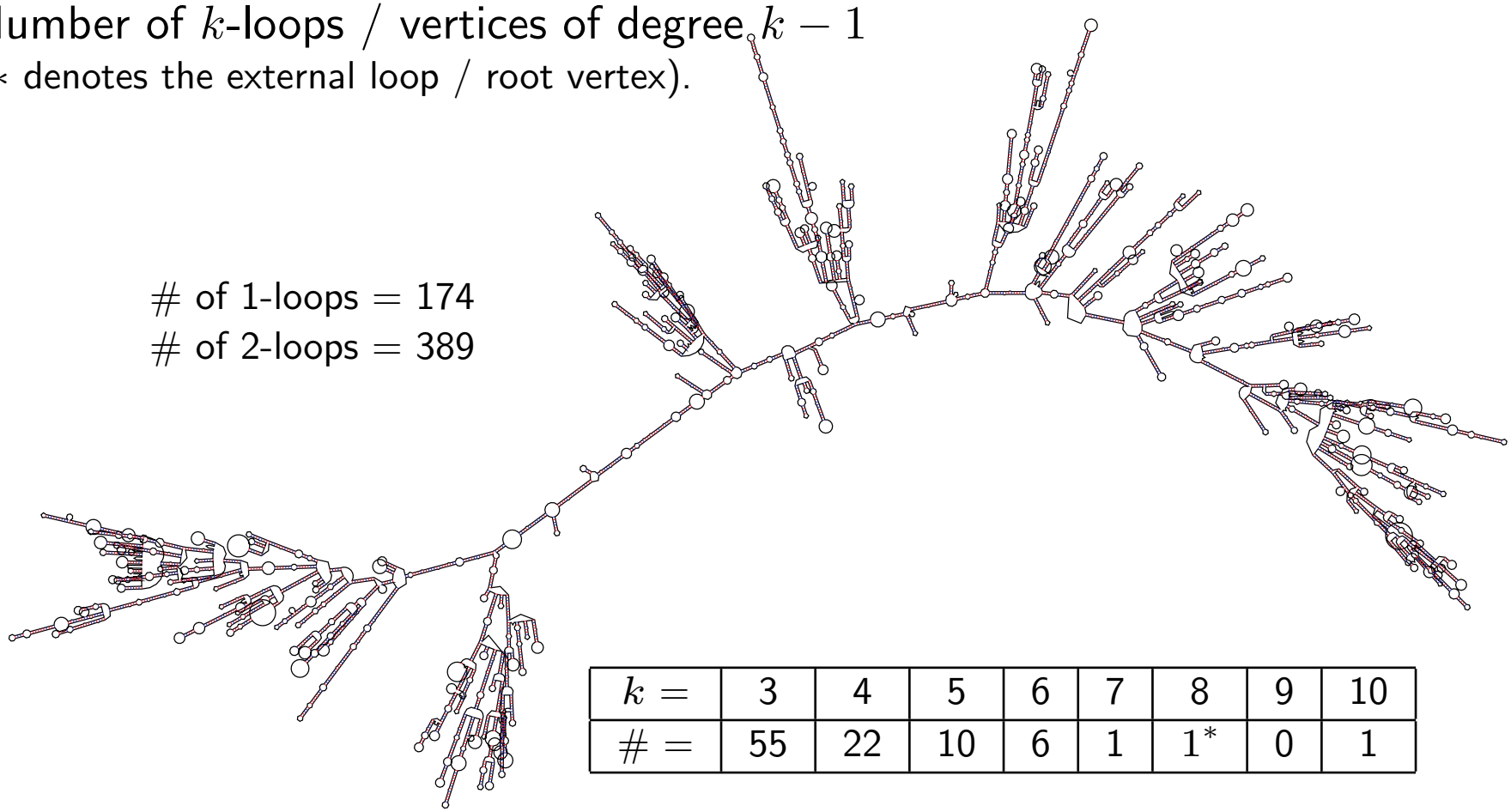
Plane trees T with 3 edges and their total loop energies $EL(T)$.

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.)*

Hepatitis C Secondary Structure Prediction

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).

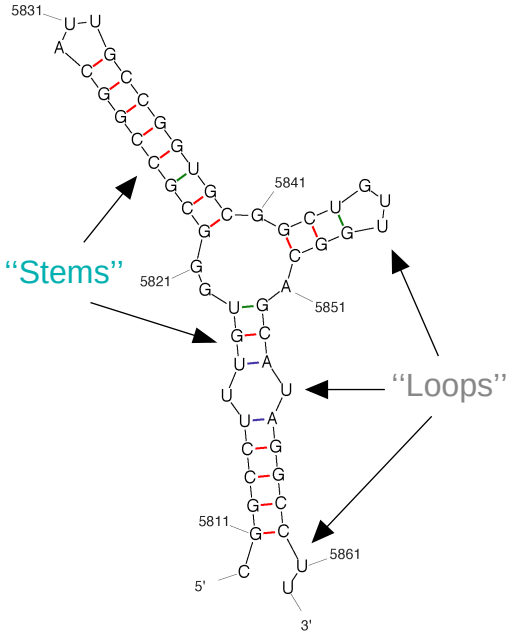


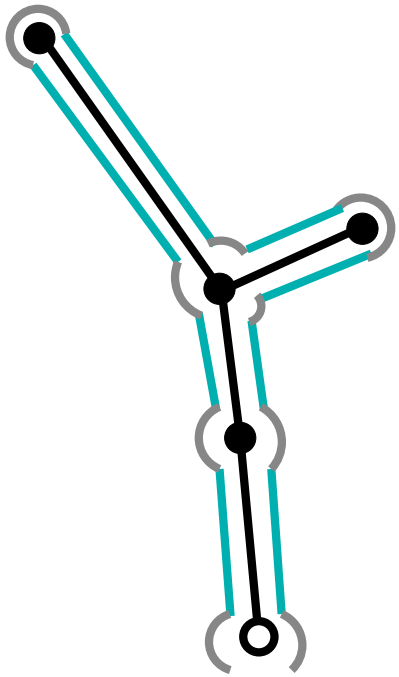
$k=$	3	4	5	6	7	8	9	10
$\#=$	55	22	10	6	1	1^*	0	1

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.



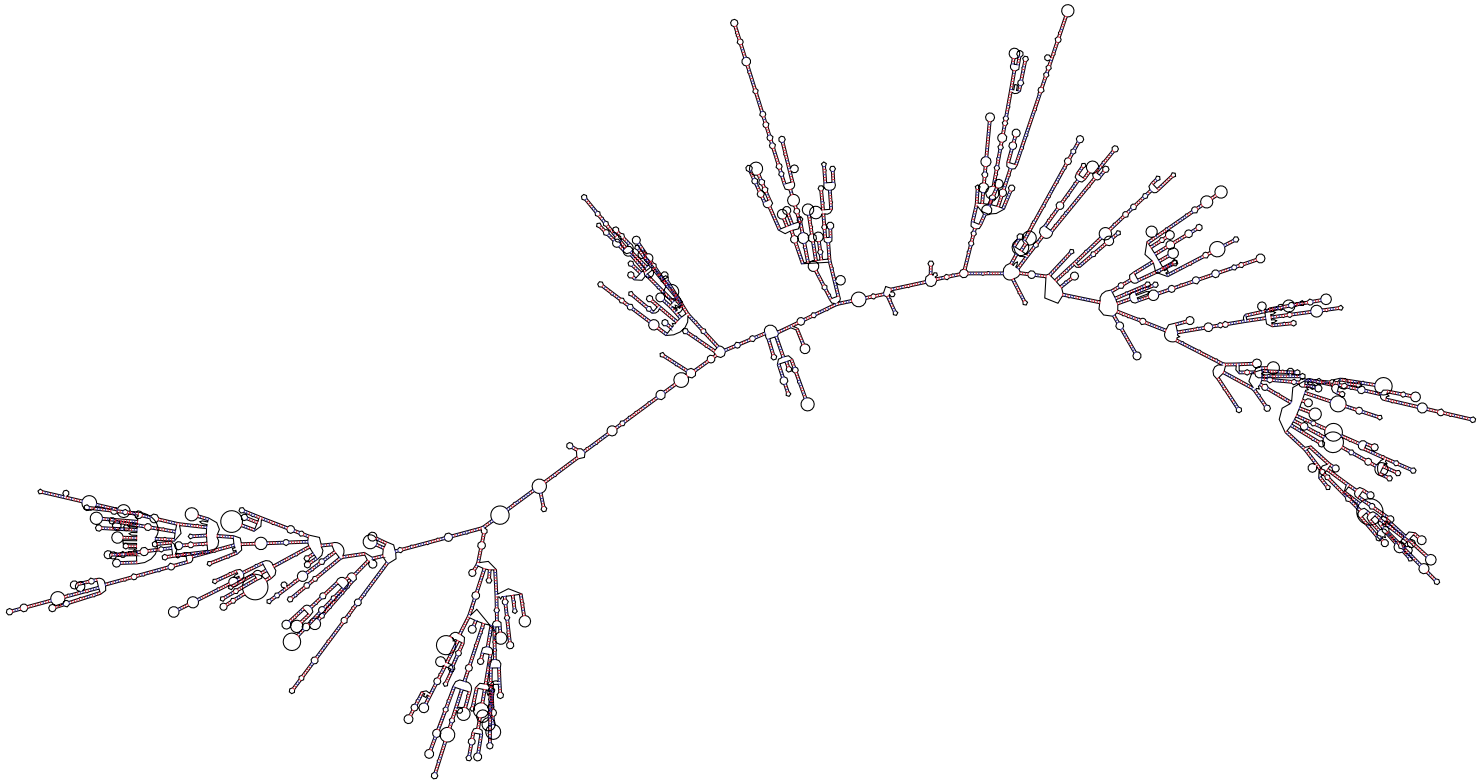


Result 2. Associated loop energies are minimized by maximizing vertices with three edges.

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

In Conclusion

Fundamental questions about the structure and function of RNA viral genomes can be addressed combinatorially.



Acknowledgments

- Dr. Laxmi Parida and DIMACS for organizing this workshop on Detecting and Processing Regularities in High Throughput Biological Data.

- Hepatitis C Secondary Structure Prediction data made publicly available by Prof. Ann C. Palmenberg and Dr. Jean–Yves Sgro.

- Predicted RNA foldings courtesy of Michael Zuker's `mfold` algorithm.

- 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."