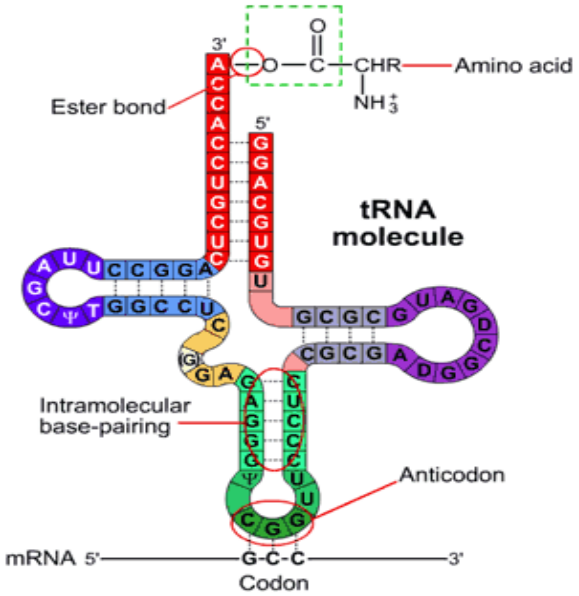
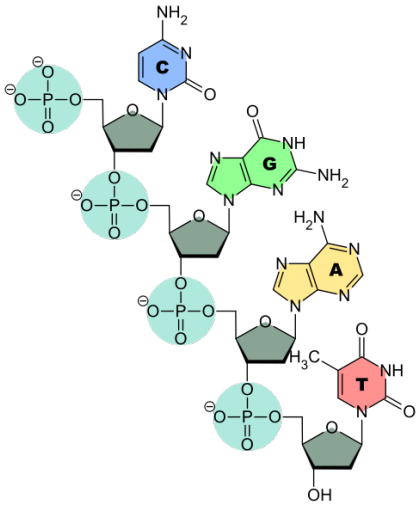
RNA vs. Ideal Randomly Branched Polymers & Prufer Shuffle

Walter Surendra Singaram, Ajay Gopal
Li Tai Fang, Aron Yoffe.
Hebrew University of Jerusalem
Bill Gelbart - UCLA





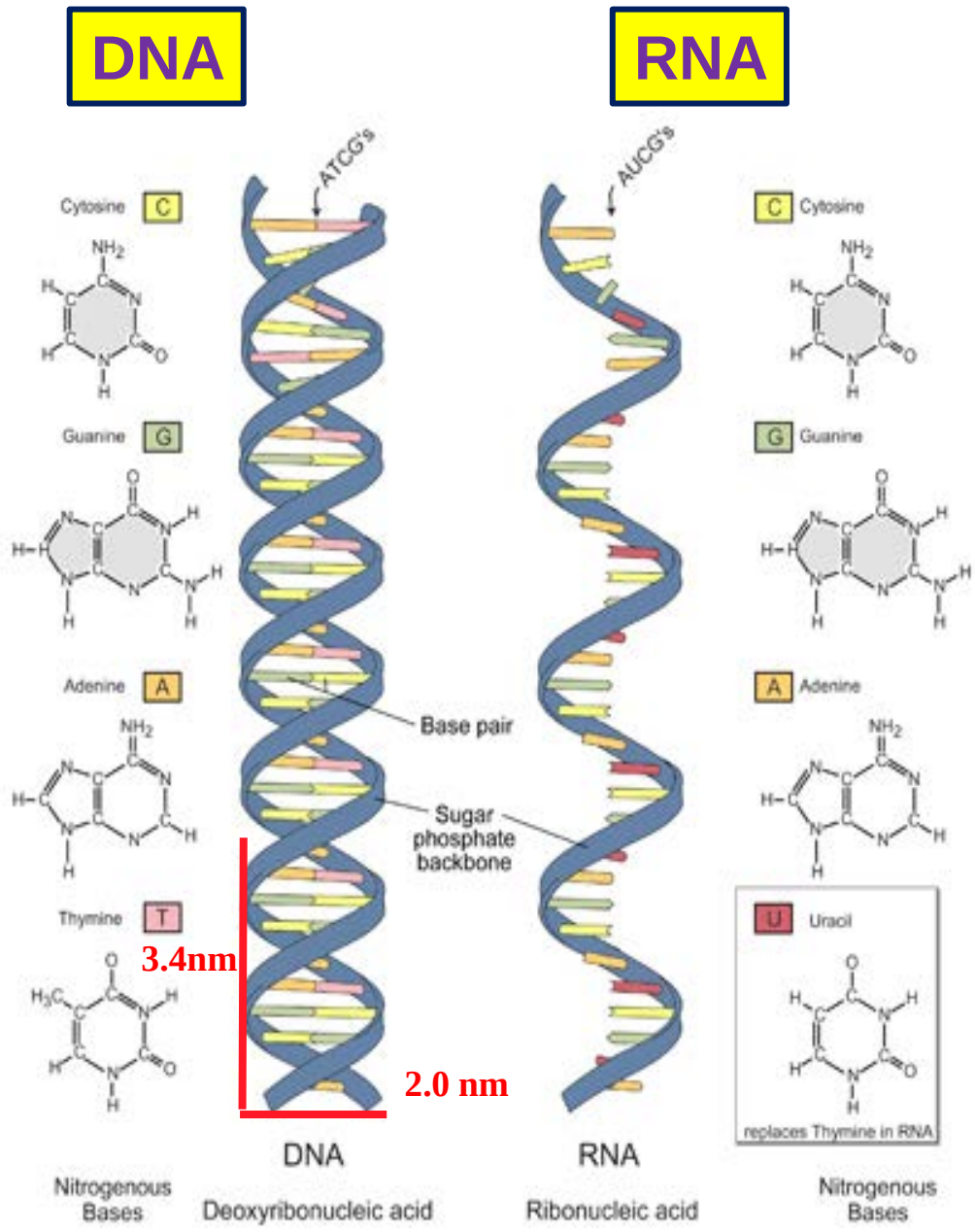


Image adapted from: National Human Genome Research Institute.

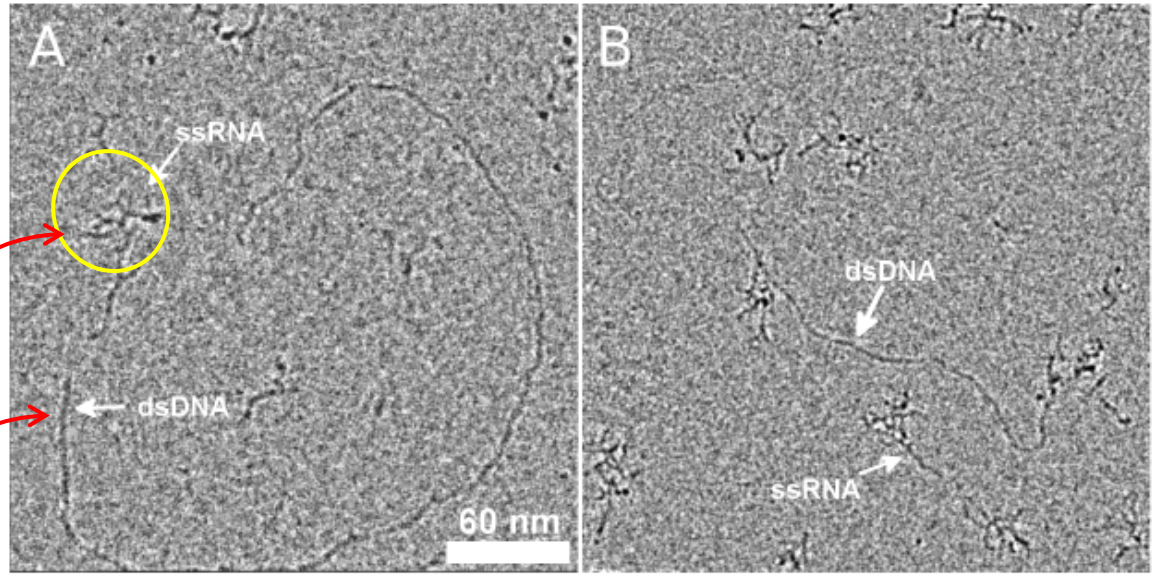
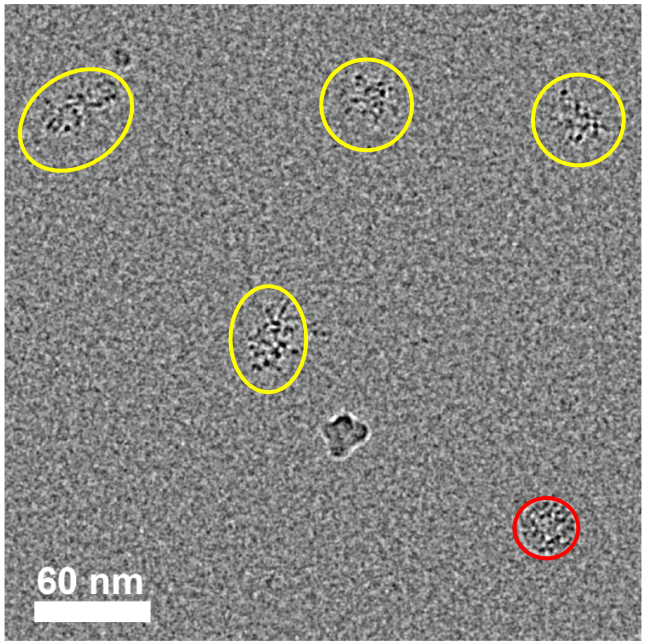
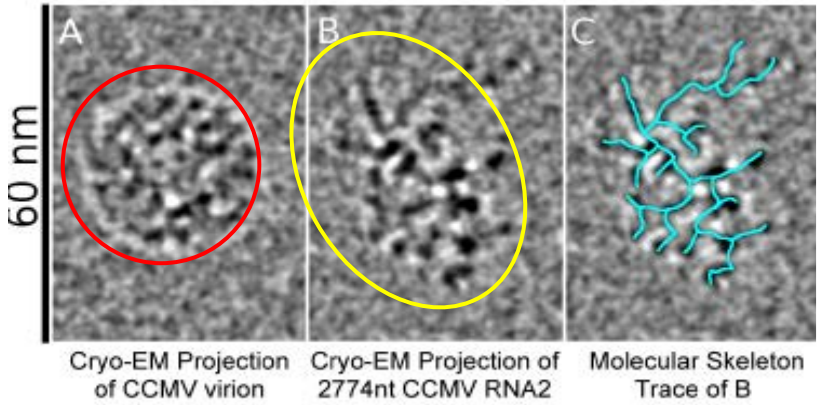


Figure S1: Comparison of ssRNA and dsDNA by Cryo-EM. Molecules of a 2117 nt ssRNA imaged in low-ionic-strength TE (Panel A) and physiological Mg^{2+}-containing buffers (Panel B)(see Methods). A small amount of 2141 base-pair dsDNA (transcription template)



assembly buffer



CCMV RNA2 (2774nt) – middle and upper three, plus virus – lower right

Relationship Between RNA and Capsid Sizes:

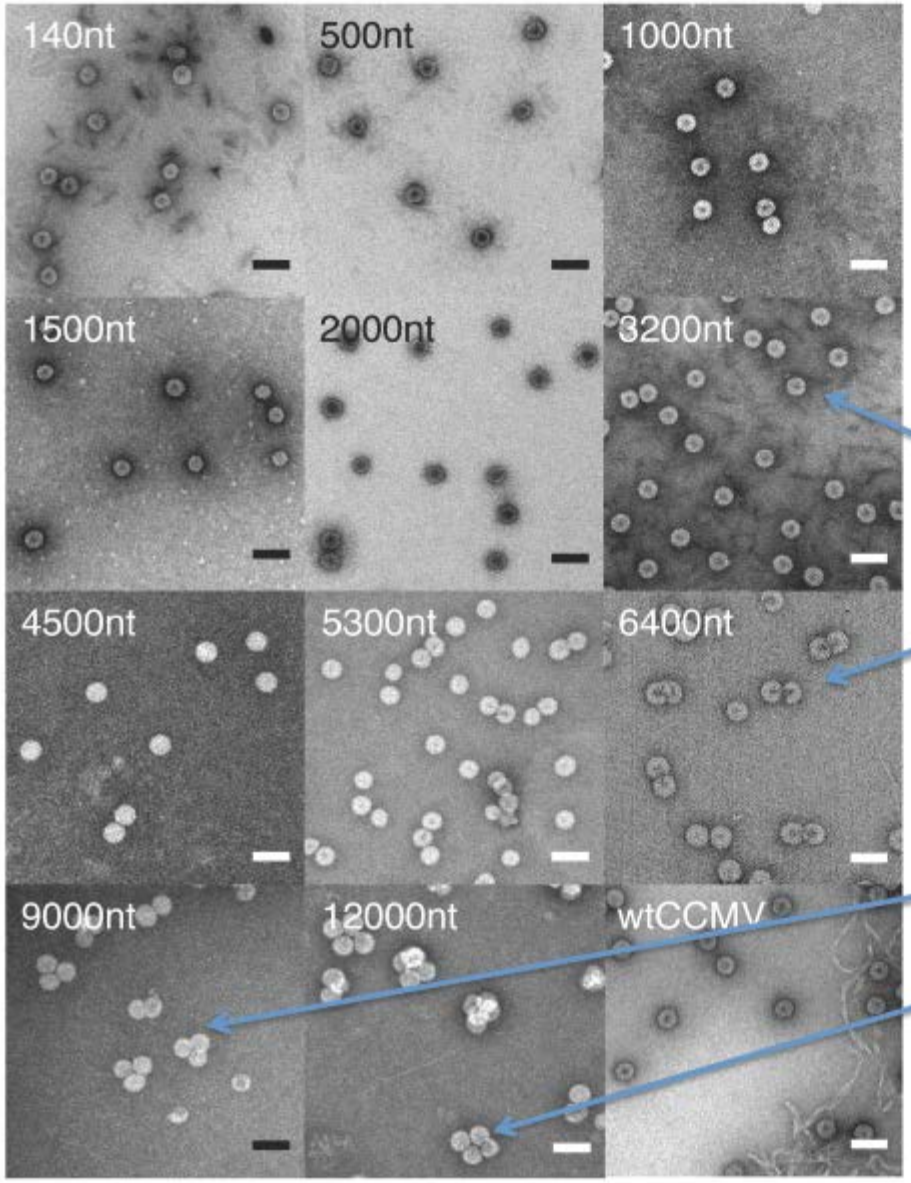
Cowpea Chlorotic Mottle Virus (CCMV) and
Brome Mosaic Virus (BMV)

These Viruses are Tripartite.

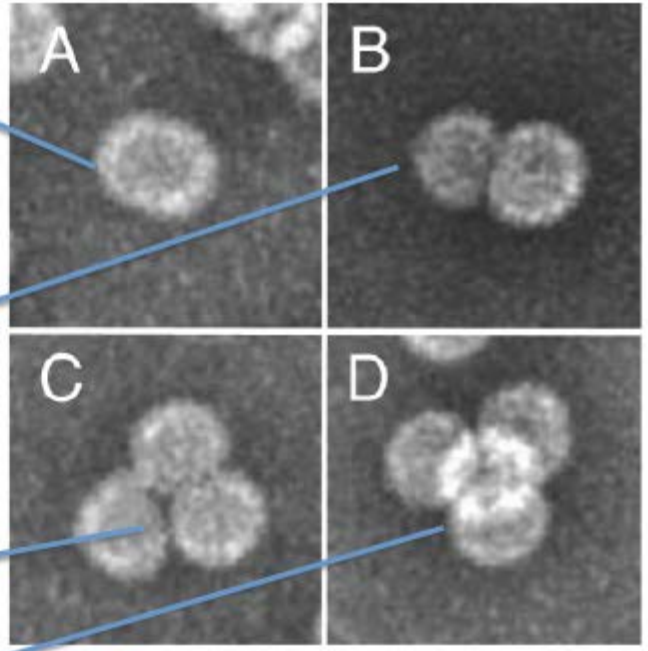
RNA 1 - 3200 nt,
RNA 2 – 2800 nt,
RNA 3 (2200)+ RNA 4 (800)= 3000 nt.
All three in the same (T=3) capsid

The sizes of capsid and genome are correlated.
Charge matching can also be important
Possibly – both size and charge should match

Assembly experiments (RNA + Capsid Protein) Reveal Definite RNA/Capsid size Correlations

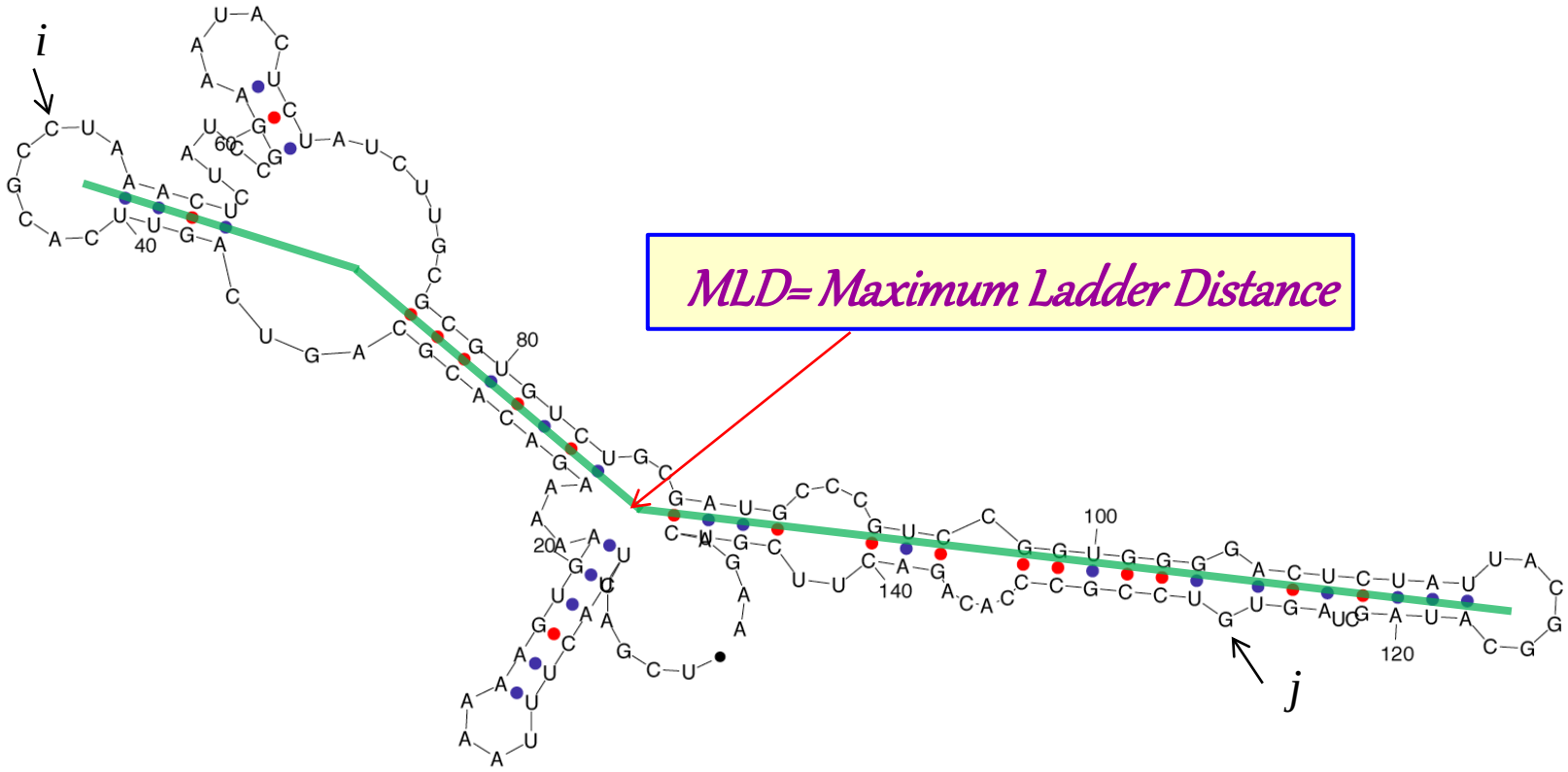


Electrostatics drive the binding of protein to RNA , and **preferred curvature** of protein determines size of capsids...

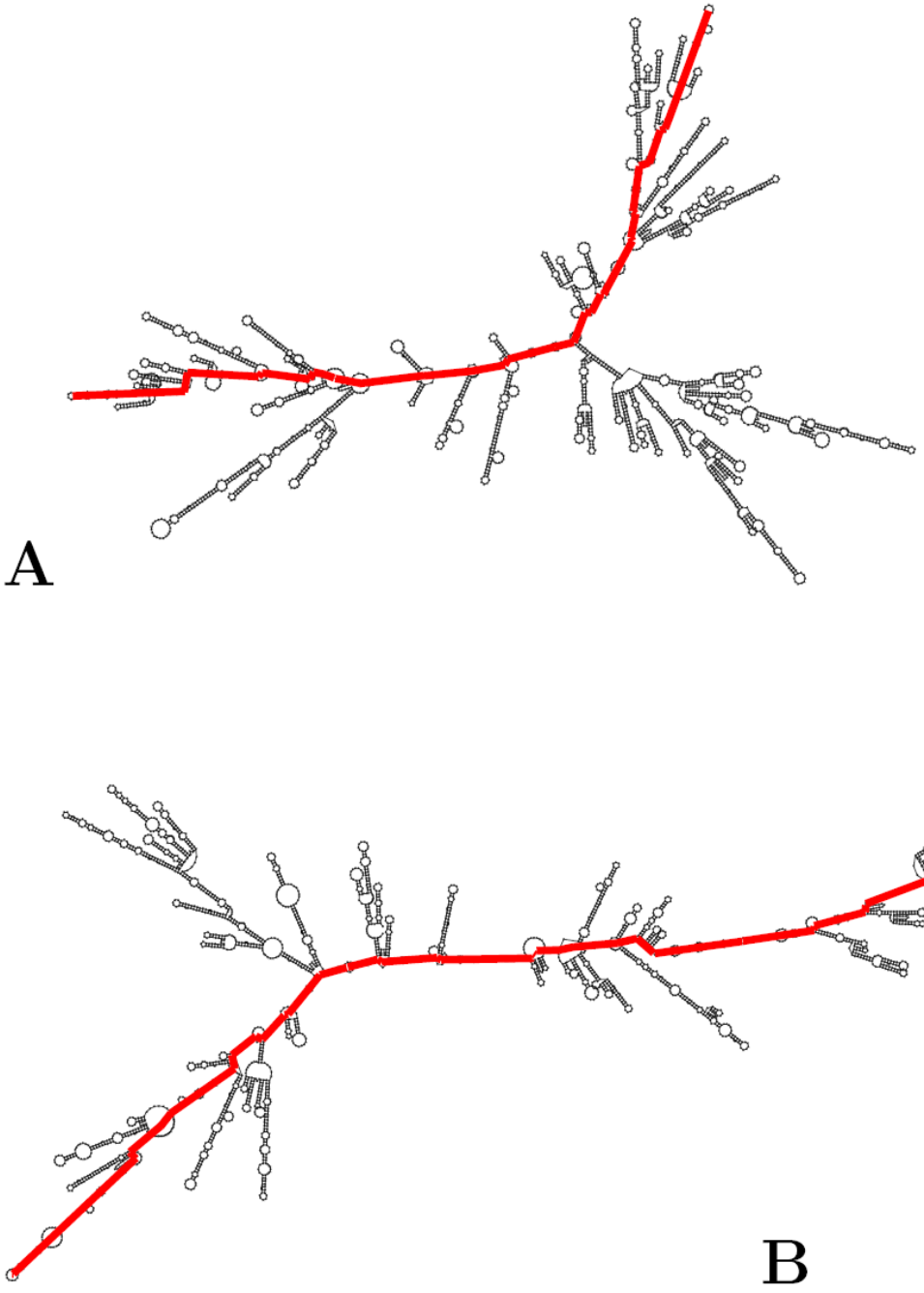


Cadena, Comas, Garmann, Rao, Knobler, Gelbart, *J. Virology* **86**, 3318 (2012)

Secondary RNA structure of ssRNA and the LD*



* Bundschuh and Wha Phys. Rev. B 2002



Viral RNA are more compact than non-viral RNA of similar length

Minimum Free energy (MFE) Structures and MLDs of:

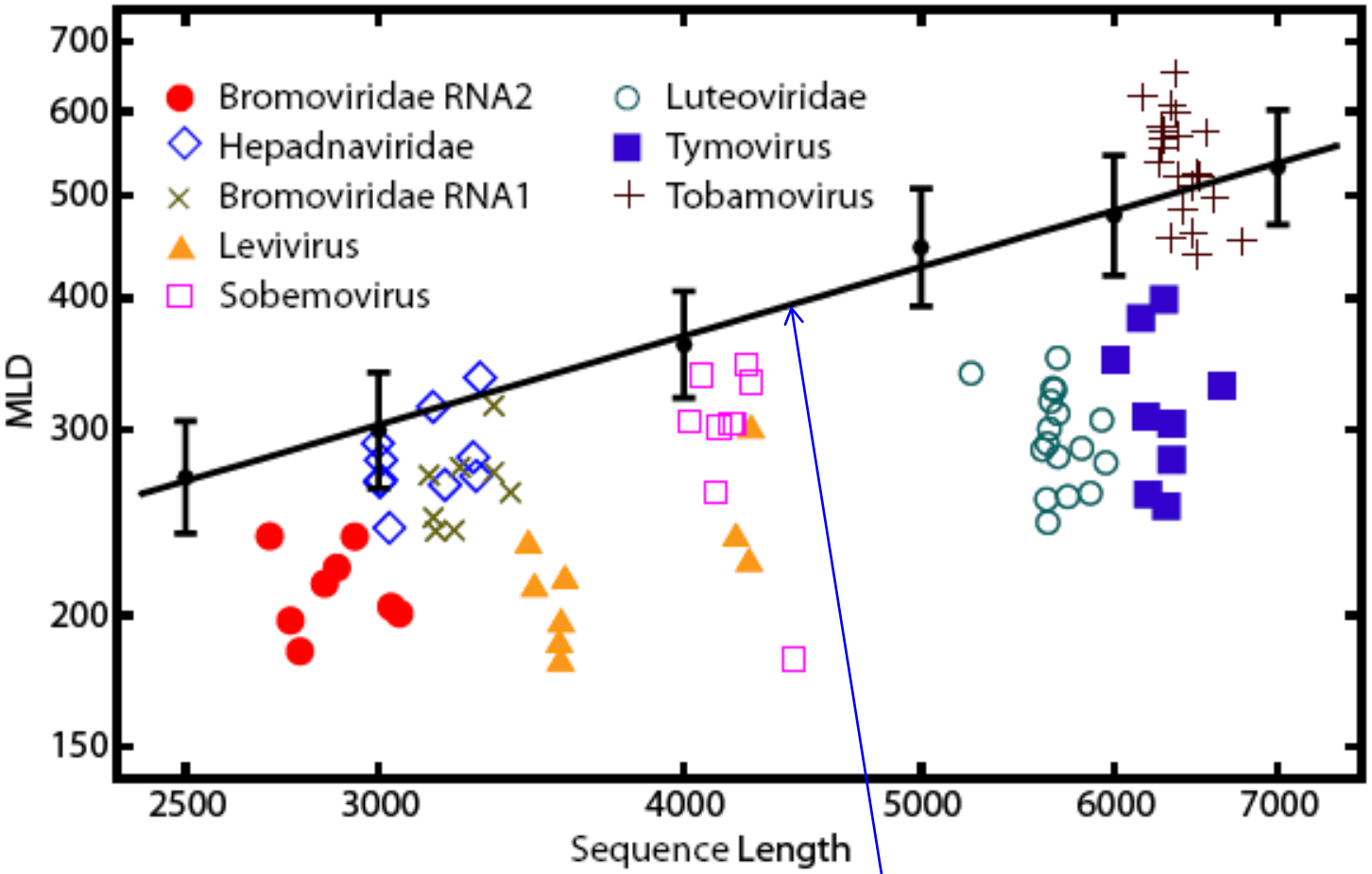
A: *3200-nt viral brome mosaic virus (BMV) RNA*

MLD = 207

B: *MFE of Random 3200-nt*,

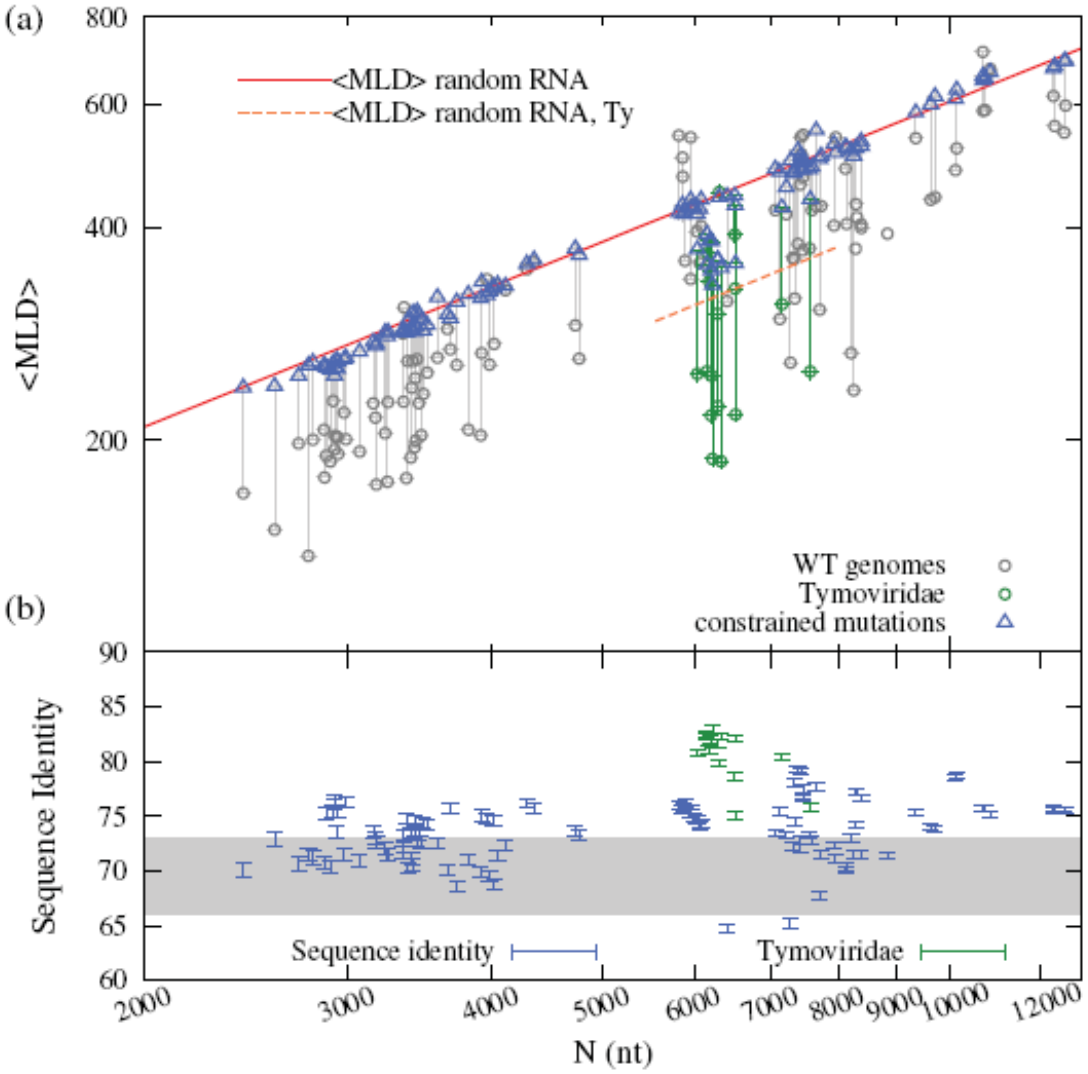
MLD = 354

Calculated by:
Surrendra Walter Singaram
for
ABS & WMG, BJ 2015



$$\langle MLD \rangle \sim N^{2/3}$$ *Random nt Sequences*

From Yoffe et al. PNAS 2008



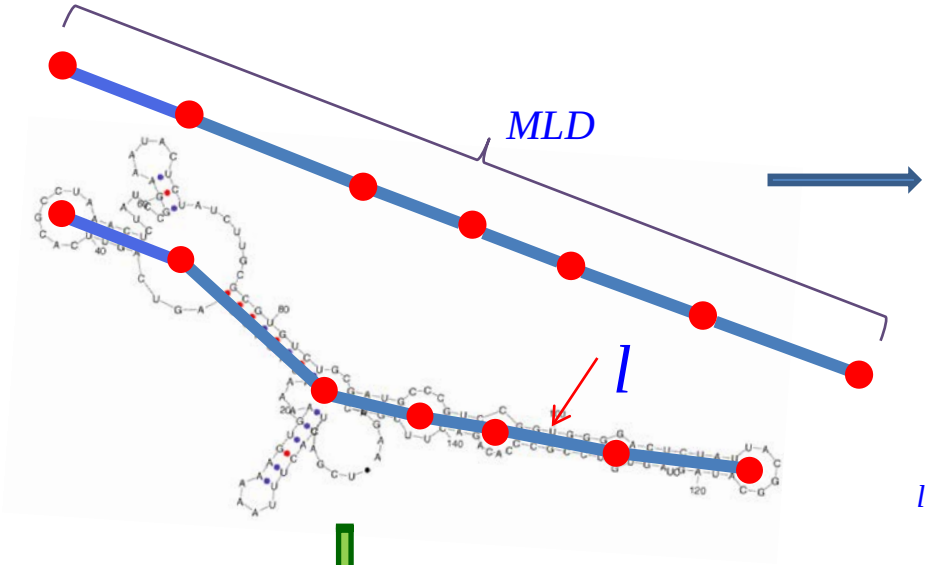
BJ 2015

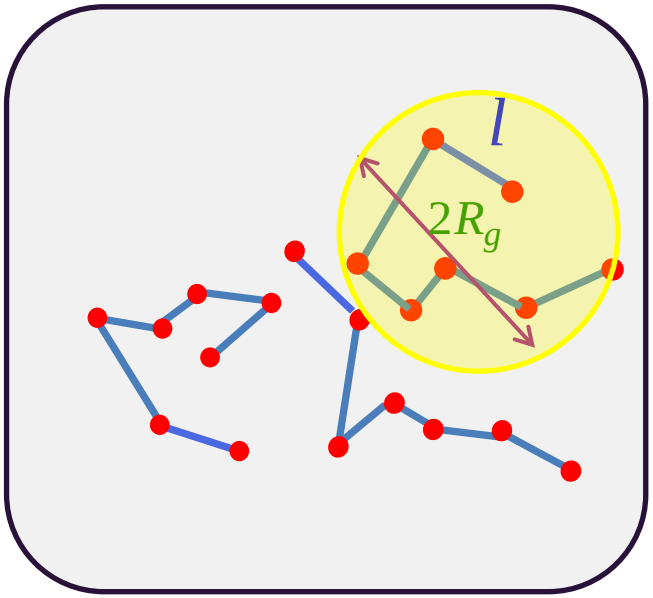
Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses

Luca Tubiana,[1,*] Anže Lošdorfer Božič,[1,2] Cristian Micheletti,[3] and Rudolf Podgornik[1,4,5]

ABS&WMG; New And Notable Commentary, BJ 2015

From MLD to Rg





$l \approx \langle k \rangle \approx 5\text{bp}$

$$R_g \sim l \left(\frac{\langle MLD \rangle}{l} \right)^{1/2} \sim \langle MLD \rangle^{1/2} \sim N^{1/3}$$

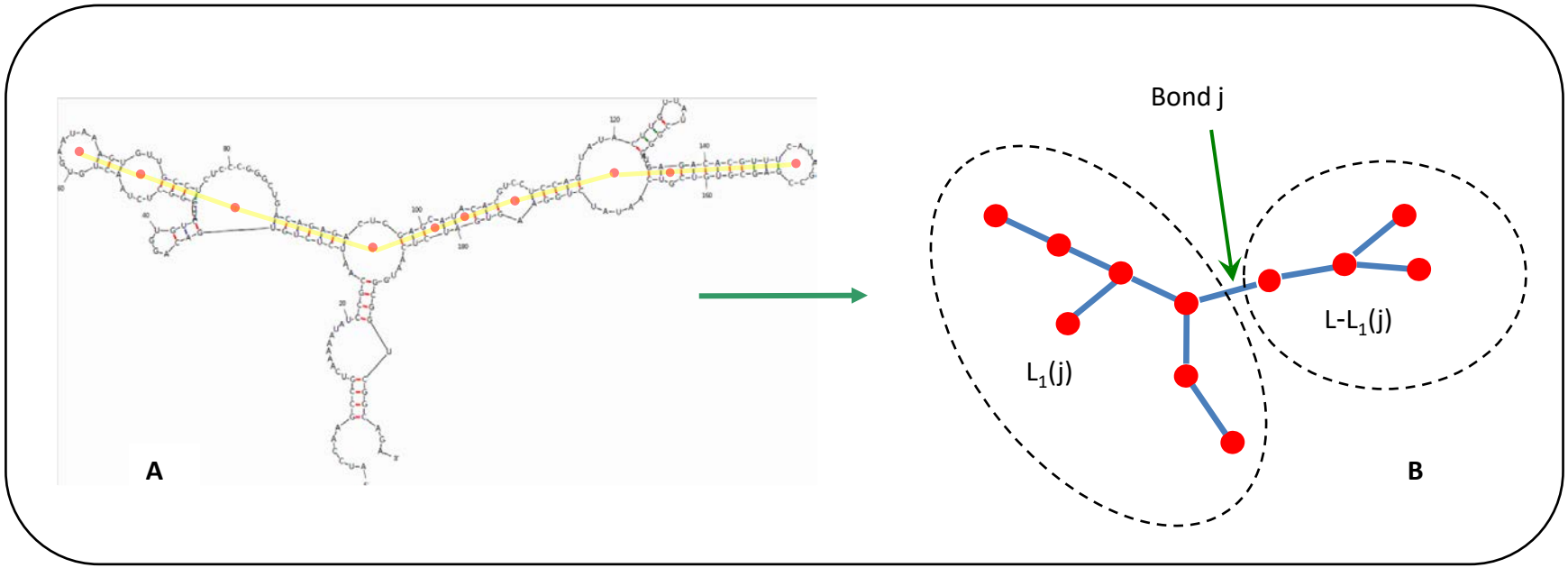
$$R_g \sim N^{\nu}$$

Ideal linear polymer	ν=1/2
Linear w Excluded Volume	ν=3/5
Collapsed polymer	ν=1/3
Ideal Randomly branched	ν=1/4

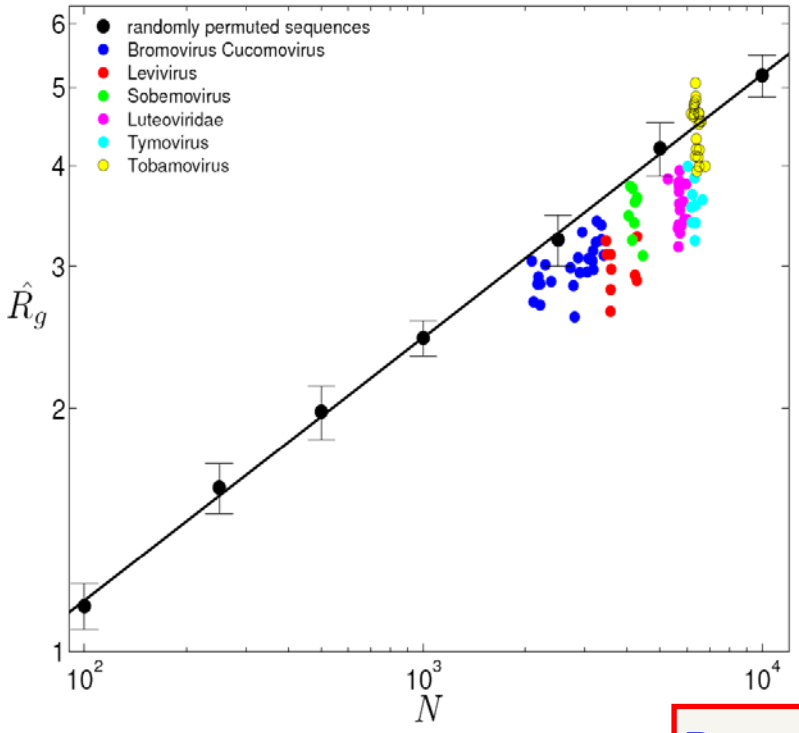
More direct calculation of R_g

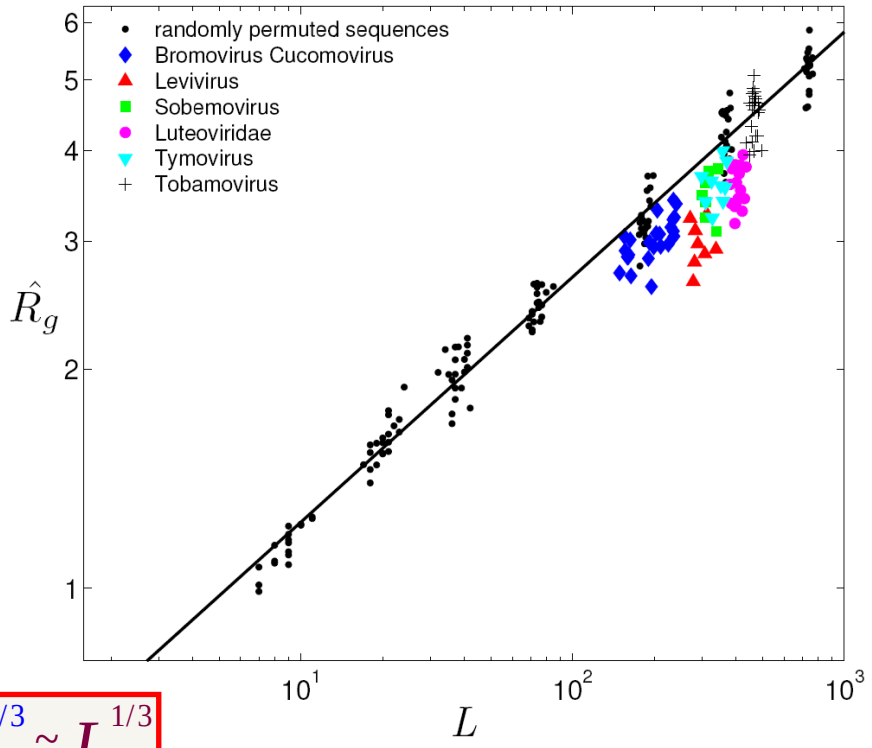
1. *Map Secondary Structure to a Tree Graph*
2. *Use Kramers Formula*

$$\left\langle R_g^2 \right\rangle = \frac{b^2}{L^2} \sum_{j=1}^{L-1} L_1(j)[L - L_1(j)]$$



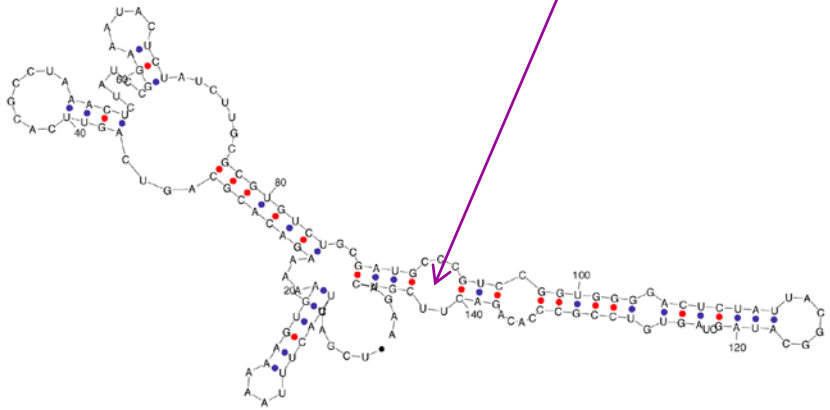
Fang, Gelbart and ABS, JCP 2011

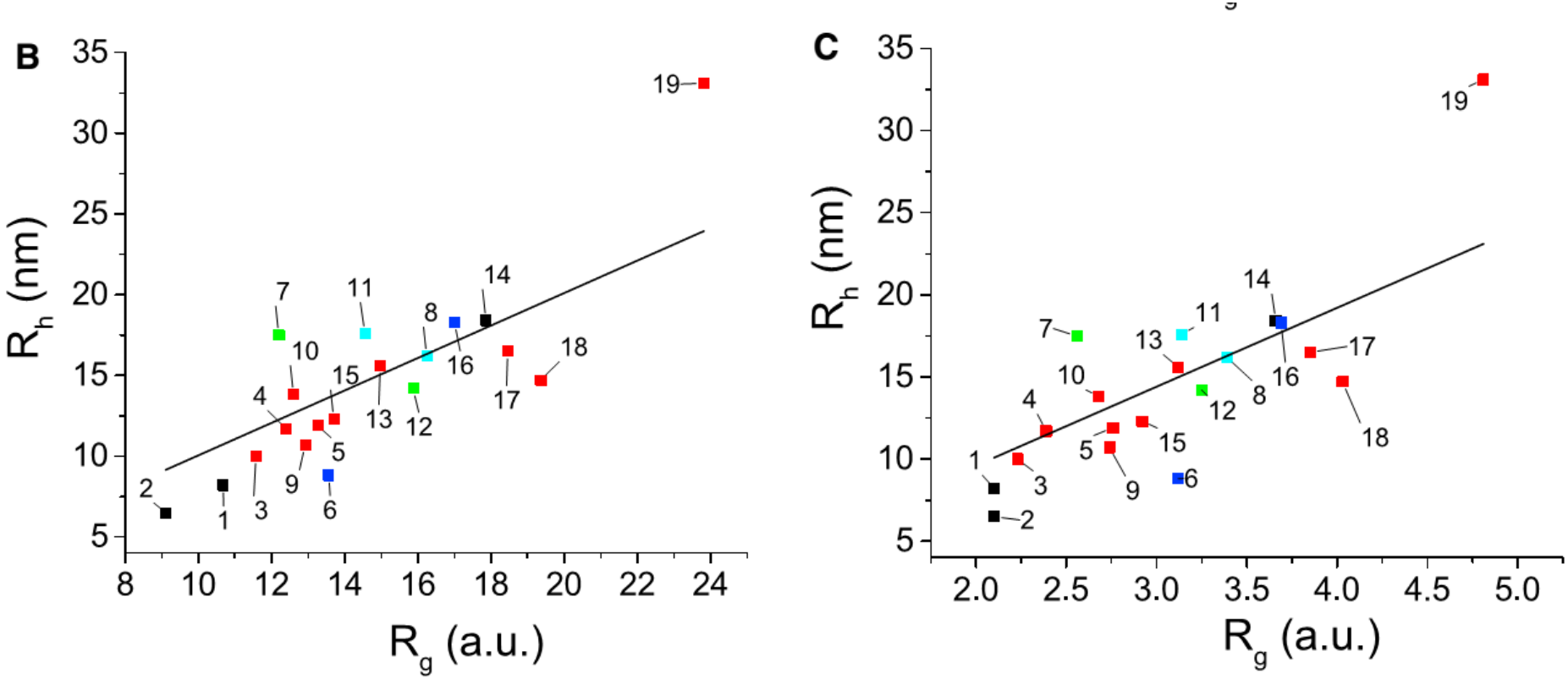




$$R_g \sim N^{1/3} \sim L^{1/3}$$

N=Number of nucleotides, *L*= Number of loops, $L \sim N$



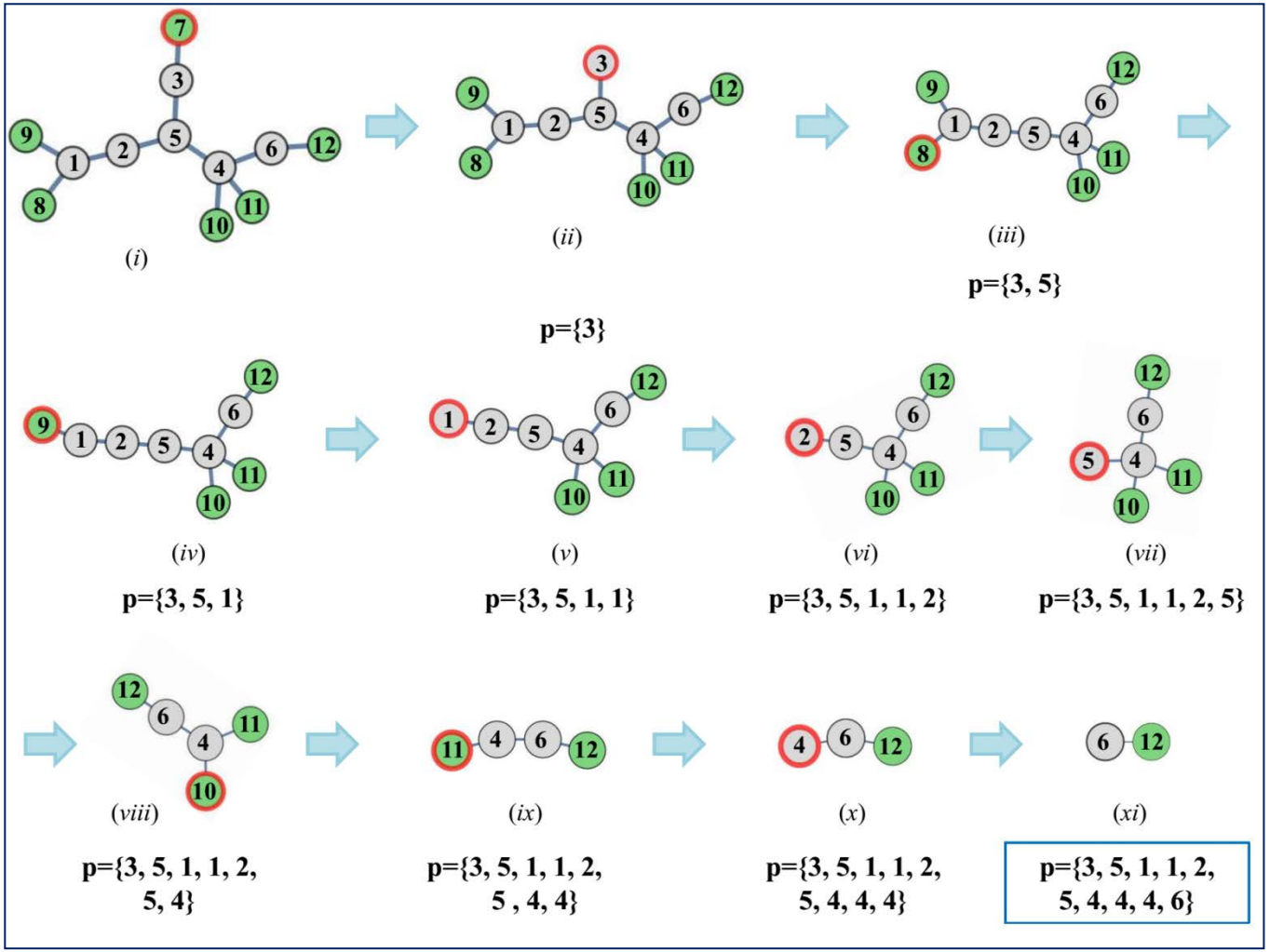


Borodavka, Singaram, Fessl, Stockley, M. Gelbart, Ben-Shaul, Tuma

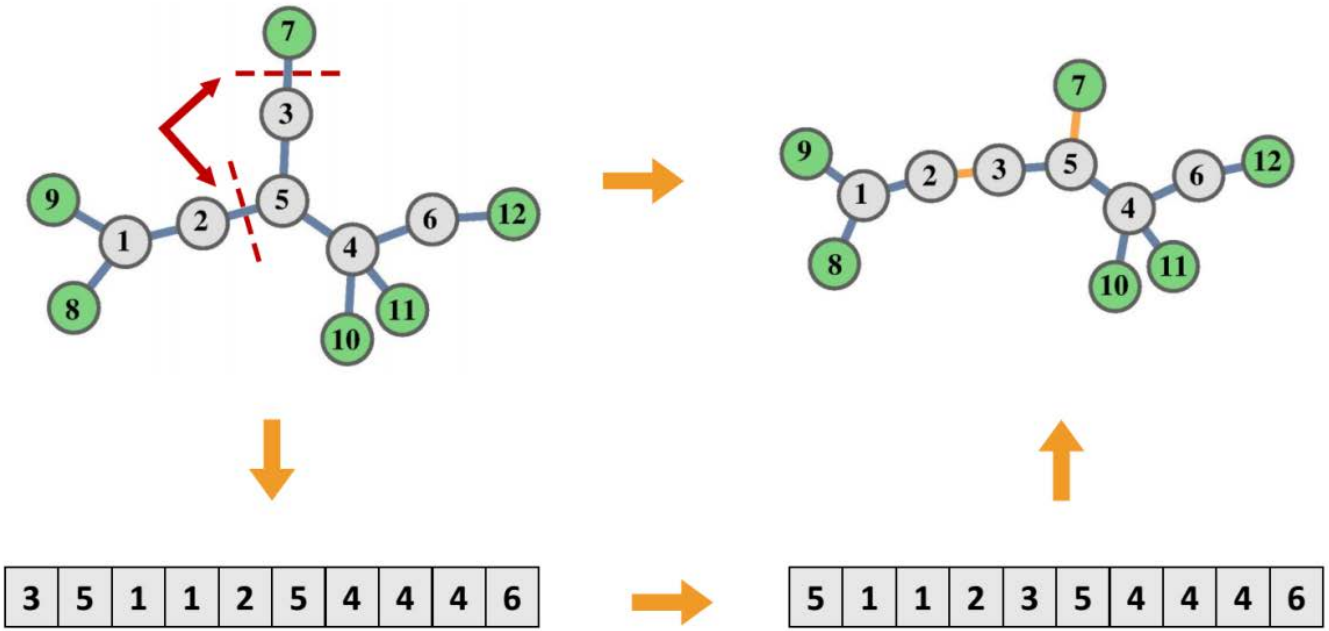
Sizes of large RNA molecules in dilute solution are dominated by the branching patterns of their secondary Structures; BJ (in press)

Generating and analyzing branched polymer configurations using Prüfer sequences

Surendra Walter Singaram, Ajaykumer Gopal, ABS (JPCB)

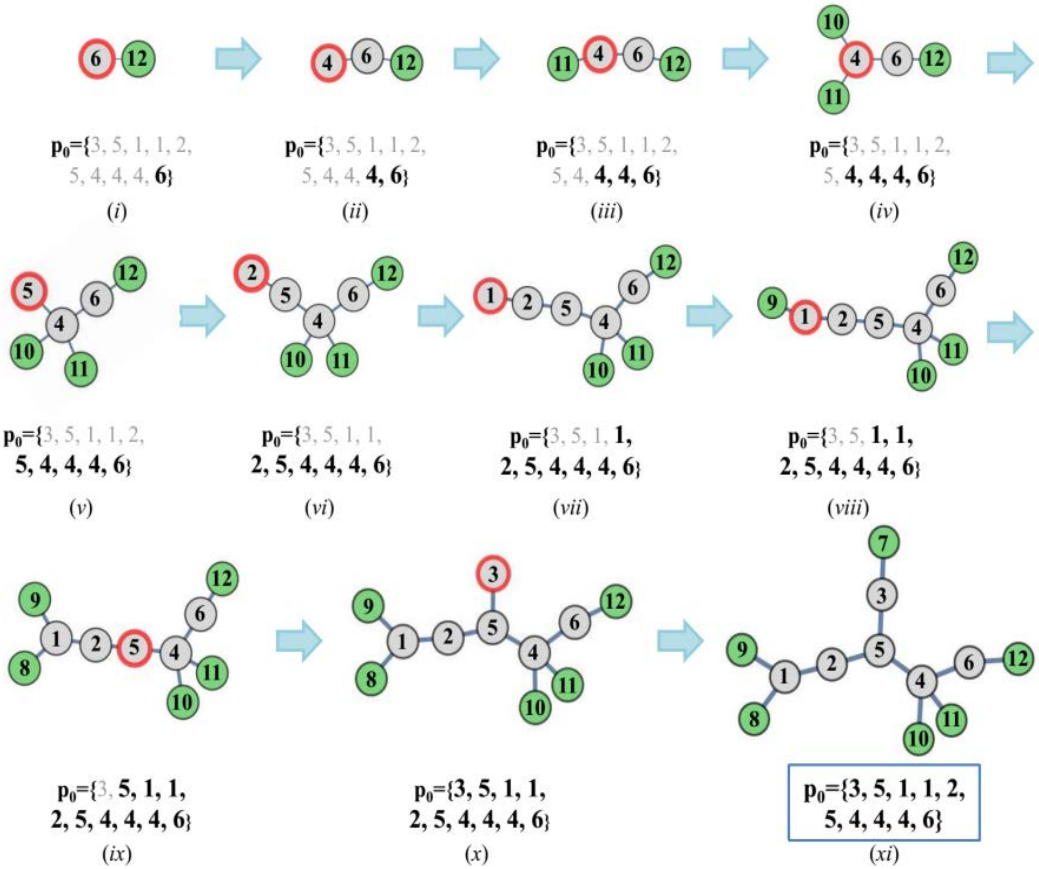


Only skeletal vertices appear in the sequence – no leaves

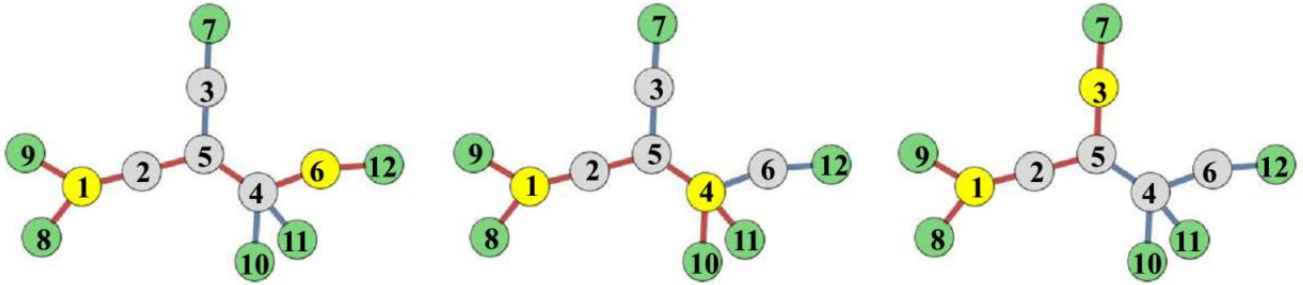


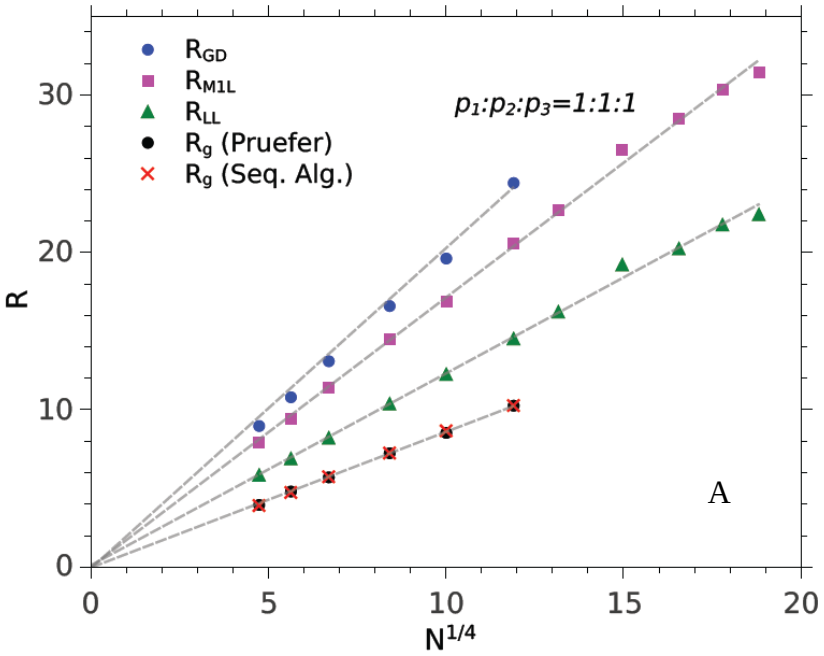
Permutations (shuffling) of the elements in the sequence produce sequences of different structure but of the same vertex degree distribution

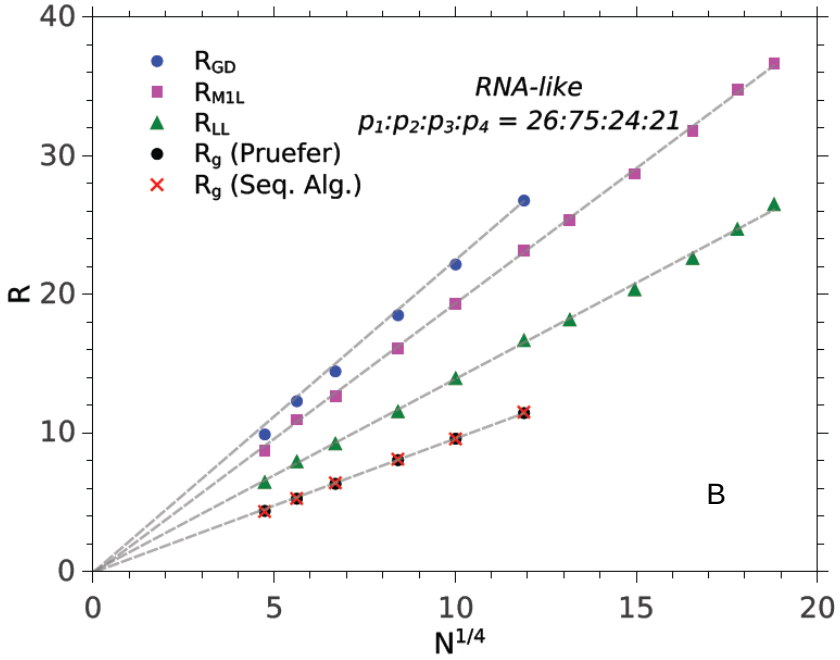
The tree graph can be recovered from its sequence



Leaf to Leaf trajectory lengths can be derived
From the sequence




R_{GD}
R_{M1L}
R_{LL}
R_g (Pruefer)
R_g (Seq. Alg.)
$p_1:p_2:p_3$=1:1:1
A
R
$N^{1/4}$
0
10
20
30
0
5
10
15
20


R_{GD}
R_{M1L}
R_{LL}
R_g (Pruefer)
R_g (Seq. Alg.)
RNA-like
$p_1:p_2:p_3:p_4$ = 26:75:24:21
B
R
$N^{1/4}$
0
10
20
30
40
0
5
10
15
20

$$R \propto N^{1/4}$$

On the secondary and tertiary (2D) structure of long ssRNA and their Relationship

Work appears in:

1) Predicting the Sizes of Large RNA Molecules
   Yoffe et al, PNAS 2008
2) The Ends of a Large RNA Molecule are Necessarily close
   Yoffe et al, Nucleic Acid Research 2010
3) A Sequential Folding Model Predicts Length Independent Secondary Structure Propreties of Large ssRNA Molecules
   Fang et al, JPCB, 2011
4) The Size of RNA as an Ideal Random Polymer
   Fang et al, JCP 2011
5) The Unusual Compactness of Viral RNAs
   Gopal et al, PloS One 2014
6) Viral RNAs are Indeed Compact
   ABS&WMG BJ 2015
7) Role of RNA Branchedness in the Competition for Viral Capsid Protein
   Singaram et al JPCB 2015
8) Sizes of large RNA molecules in dilute solution are dominated by the branching patterns of their secondary structures
   Borodavka et al BJ… 2016
9) A Prüfer-Sequence Based Algorithm for Calculating the Size of Ideal Randomly Branched Polymer
   Singaram et al JPCB (2016)

Questions:

What makes viral RNA more compact

What is the origin of the $R_g \sim N^{1/3}$ of random sequence RNA

Many question regarding RNA-CP interactions
and Viral assembly

Viral assembly and competition experiments
UCLA – Gelbart and Knobler Laboratory

Assembly of CCMV (BMV) viruses take place in two stages:

(i) Capsid protein (CP) and RNA are mixed at neutral pH.
The CP exist as dimers and bind electrostatically to the RNA.
At a "magic ratio" of 1CP dimer/20 nt all RNA charges are neutralized.

(ii) Viral assembly takes place after lowering the pH to ~4.8,
reducing inter-CP repulsion.
T=3 capsids (180 CP) are the dominant structures.

RNA packaging competitions: Experiments done at UCLA.

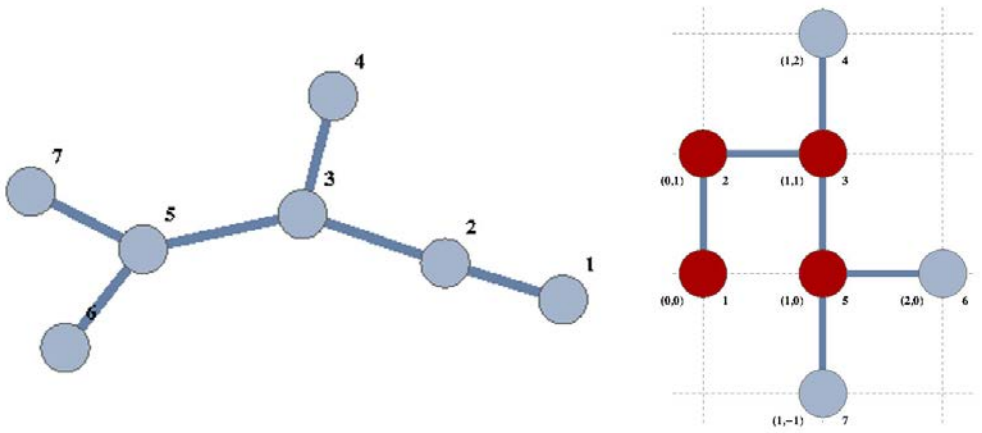
Example: (i) **1500nt RNA is mixed and saturated by Capsid Proteins** (CP) at Neutral pH .

(ii) **RNA 3000nt added**. No assembly at this pH.

(iii) pH is lowered to 5 → Viruses formed. **All containing only 3000nt RNA**

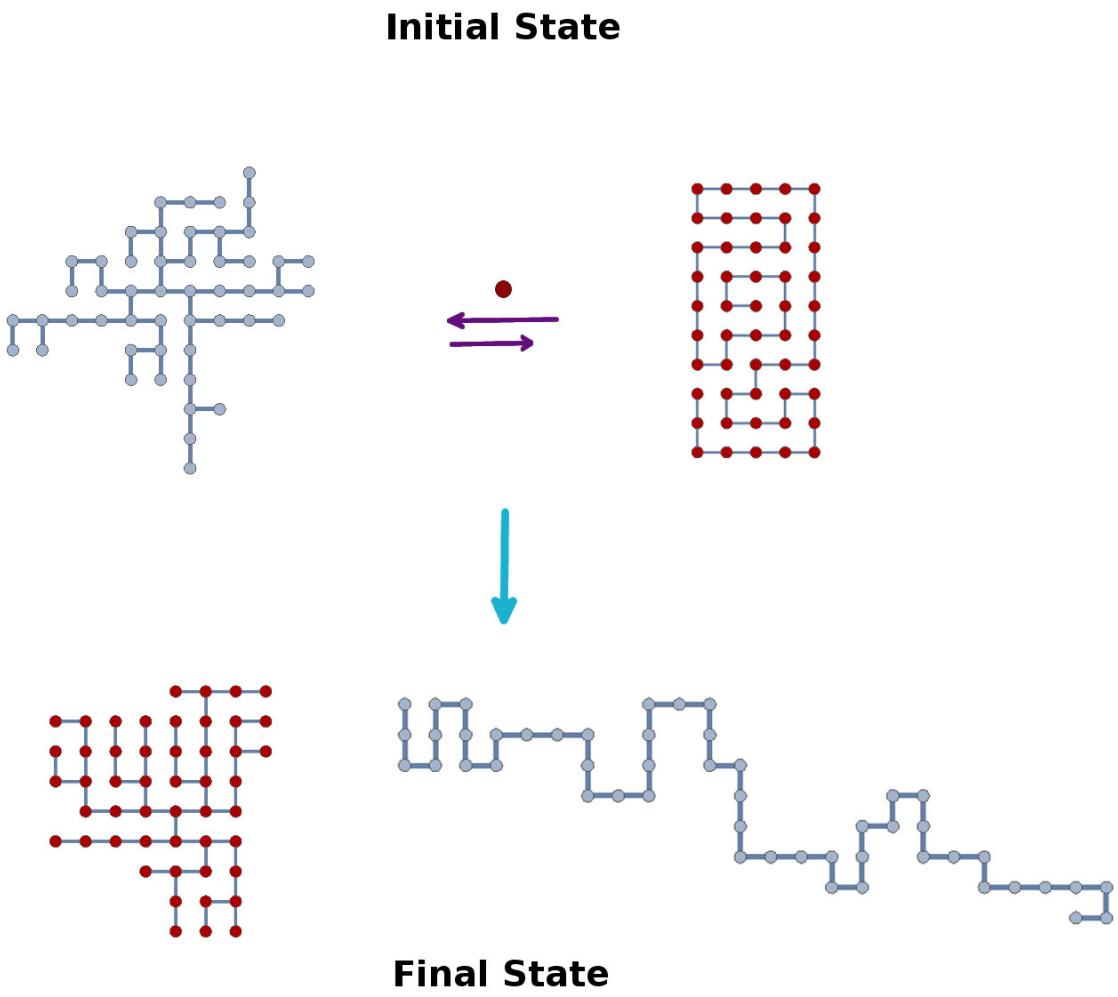
Redistribution of CP takes place at Neutral pH, prior to full assembly

Modeling Virus Competition Experiments
Walter Surendra Singaram



(a) (b)

Branched polymer (Left) vs. Linear Polymer (Right)
Proteins (Red) go from linear to branched.
Process driven by entropy gain of linear polymer

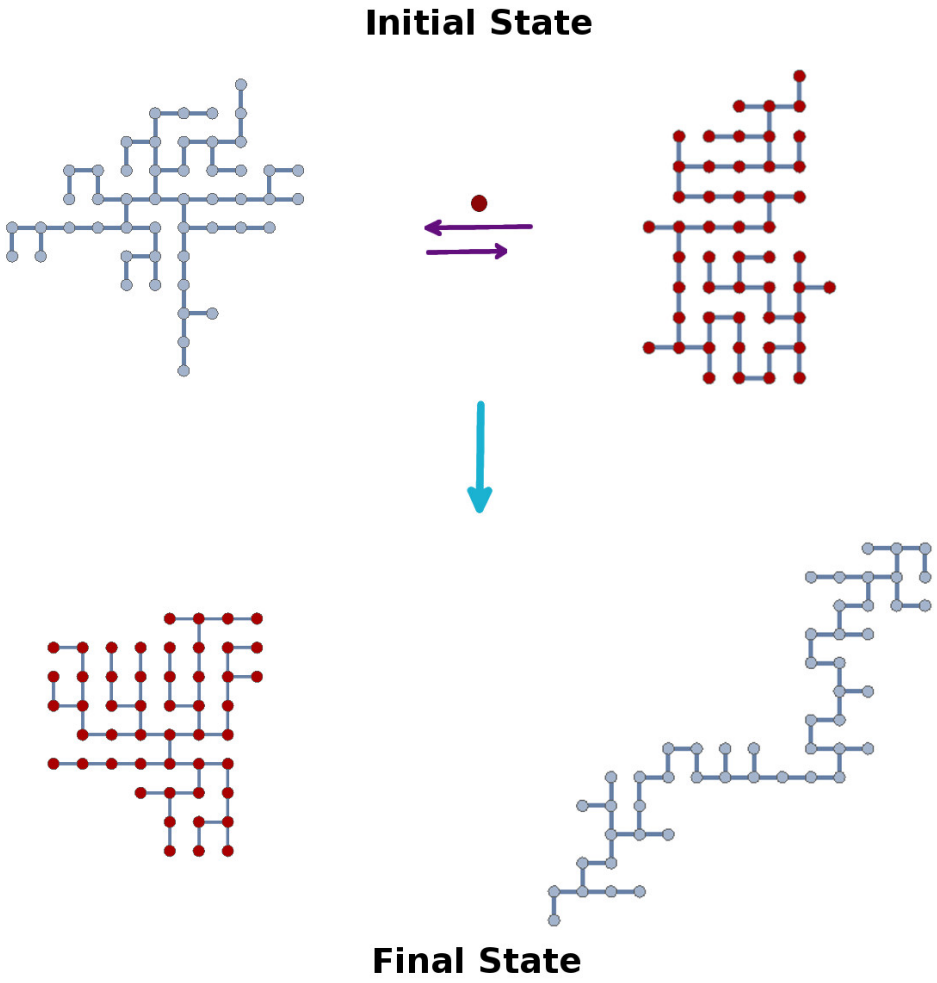


Compact (small Rg) Branched polymer (Left) vs. Extended (Large Rg) Branched Polymer (Right)*

Proteins (Red) go from extended to compact

Process driven by entropy of extended and energy by compact

(* Both polymers have the same vertex distribution)

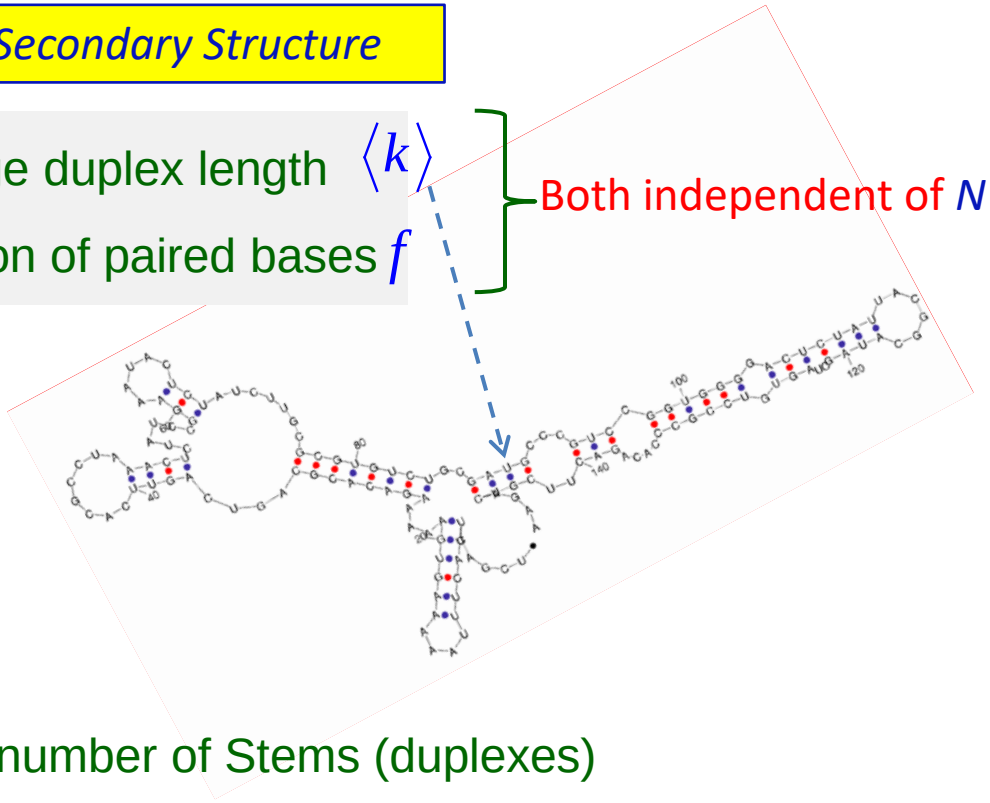


Some general properties of RNA Secondary Structure

L ~ *N* because: 1. The average duplex length $\langle k \rangle$

2. The fraction of paired bases f

Both independent of *N*



$\langle k \rangle \approx 4.5,\ f = 0.65$
for random and most viral sequences

3. *L*=*S* The number of Stems (duplexes)

4. The average degree of a loop is 2.

$$\langle d \rangle = \frac{1}{L}\sum_i d_i = \frac{1}{L}\sum_d dL_d = 2 - \frac{2}{L} \approx 2$$

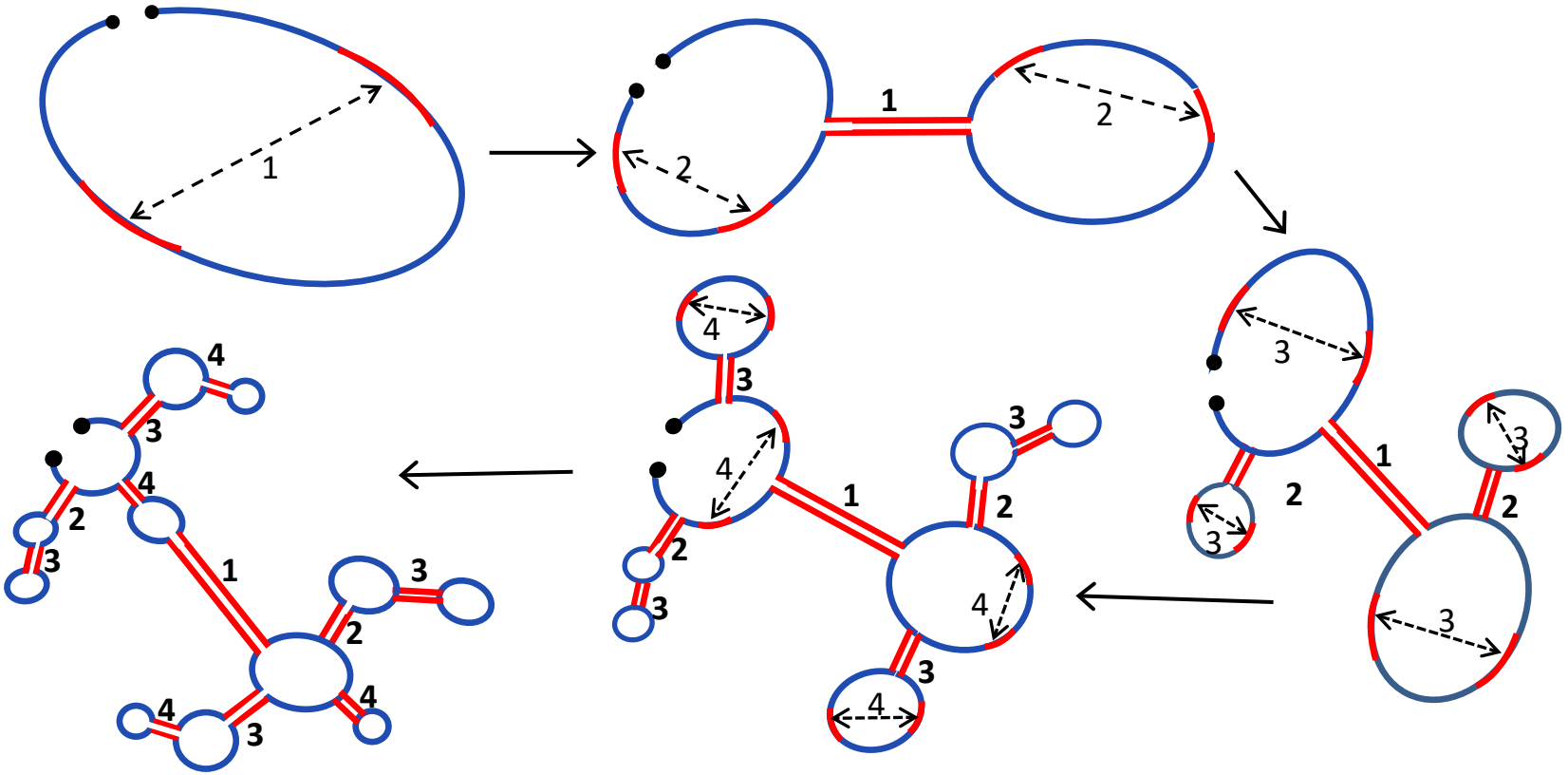
These, and other results can be explained based on a simple RNA folding model:

A simple Sequential Folding Model of RNA (Fang et al – JPCB 2011)

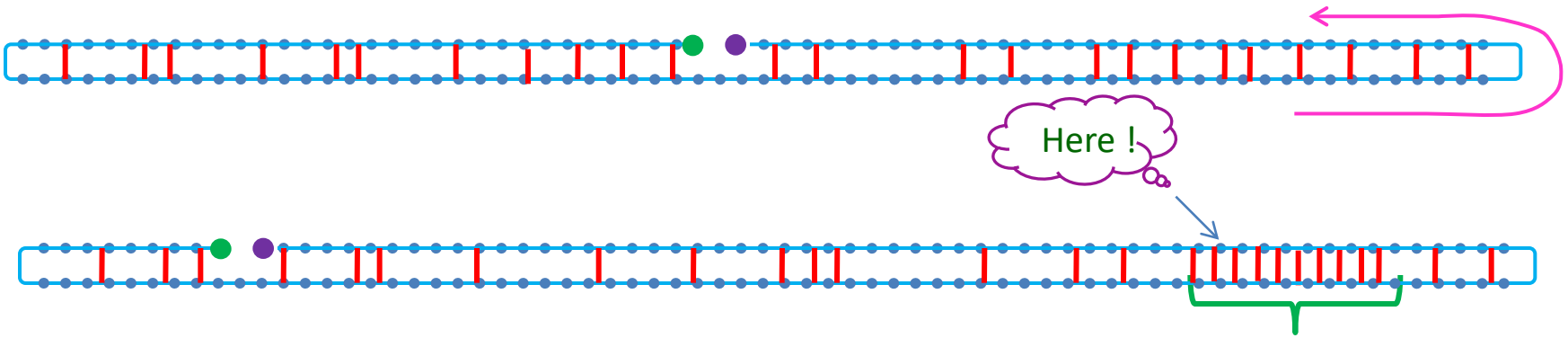
Explains Basic Properties of Secondary Structure of Random RNA:

Independence of <k> and f on sequence length.

SFM: At every generation each loop is divided by the largest possible duplex



Align and slide to find longest duplex



A duplex of exactly k bp is found with probability:

$$P(k) = (1-\alpha)\alpha^{k}(1-\alpha)$$

(α=1/4 for uniform A,C,G,U mixture, etc)

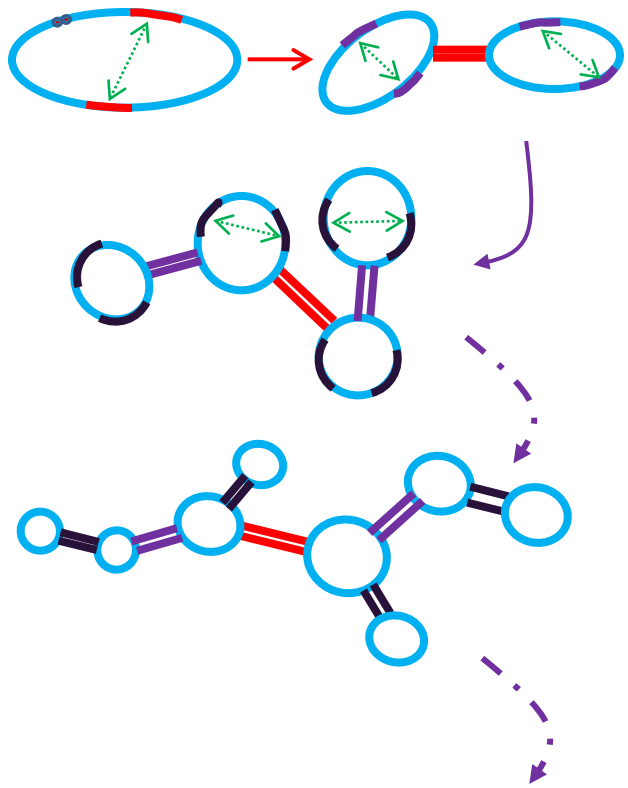
Total Number of possible Alignments of pairs of k sequences:

$$(1/2)(N-2k-l)(N-k) \approx (1/2)N^{2}$$

Assuming one longest duplex of length k_0 $\longrightarrow$ $(1/2)N^{2}(1-\alpha)^{2}\alpha^{k_0} \approx 1$

$$k_0 \simeq \frac{1}{\ln 2}\ln N - 0.9 \qquad (\alpha = 1/4)$$

Generation	duplex length	#of duplexes
0	$k_0 \simeq \frac{1}{\ln 2} \ln N - 0.9$	$1=2^0$
1	$k_1 \simeq \frac{1}{\ln 2} \ln\left(\frac{N}{2}\right) - 0.9$	$1=2^1$
m	$k_m \simeq \frac{1}{\ln 2} \ln\left(\frac{N}{2^m}\right) - 0.9$	$1=2^m$
l=last	$k_l \simeq \frac{1}{\ln 2} \ln\left(\frac{N}{2^l}\right) - 0.9$	$1=2^l$



$$\frac{N}{2^l} \approx N_l = \text{(penultimate) smallest loop} \sim 30$$

$$\langle k \rangle \simeq \frac{\ln N_l}{\ln 2}$$

Average duplex length - $\langle k \rangle$ Independent of *N*;

Average fraction of bases in duplexes *f* - Independent of *N*;

$$f = \frac{4\langle k \rangle}{N_l}$$

Crude Estimate

$N_l \sim 30 \rightarrow \langle k \rangle \approx 4.9,\ f = 0.65$

($\rightarrow$ smallest loop ~10 bases)

<k> ~ 4-5 and f~0.6 are similar for both viral and non-viral RNAs

$$\langle k \rangle = \frac{1}{\sum_0^l 2^m} \sum_0^l k_m 2^m = \frac{1}{\sum_0^l 2^m} \sum_0^l 2^m \left(\frac{\ln(N/2^m)}{\ln 2} - 0.9 \right)$$

$$= k_0 - (l-1) = \frac{\ln N}{\ln 2} - 0.9 - \frac{\ln(N/N_l)}{\ln 2} + 1 = \frac{\ln N_l}{\ln 2}$$

$$f = \frac{2\langle k \rangle \times (\#\, duplexes)}{N} \simeq \frac{2(\ln N_l / \ln 2) \times 2^{l+1}}{N}$$

$$= \frac{2(\ln N_l / \ln 2) \times 2 \times (N/N_l)}{N} = \frac{4 \ln N_l}{N_l \ln 2} = \frac{4\langle k \rangle}{N_l}$$

$$\langle k \rangle \simeq \frac{\ln N_l}{\ln 2}$$

Average duplex length - $\langle k \rangle$ Independent of *N*;

Average fraction of bases in duplexes *f* - Independent of *N*;

$$f = \frac{4\langle k \rangle}{N_l}$$

Crude Estimate

$N_l \sim 30 \rightarrow \langle k \rangle \approx 4.9,\ f = 0.65$

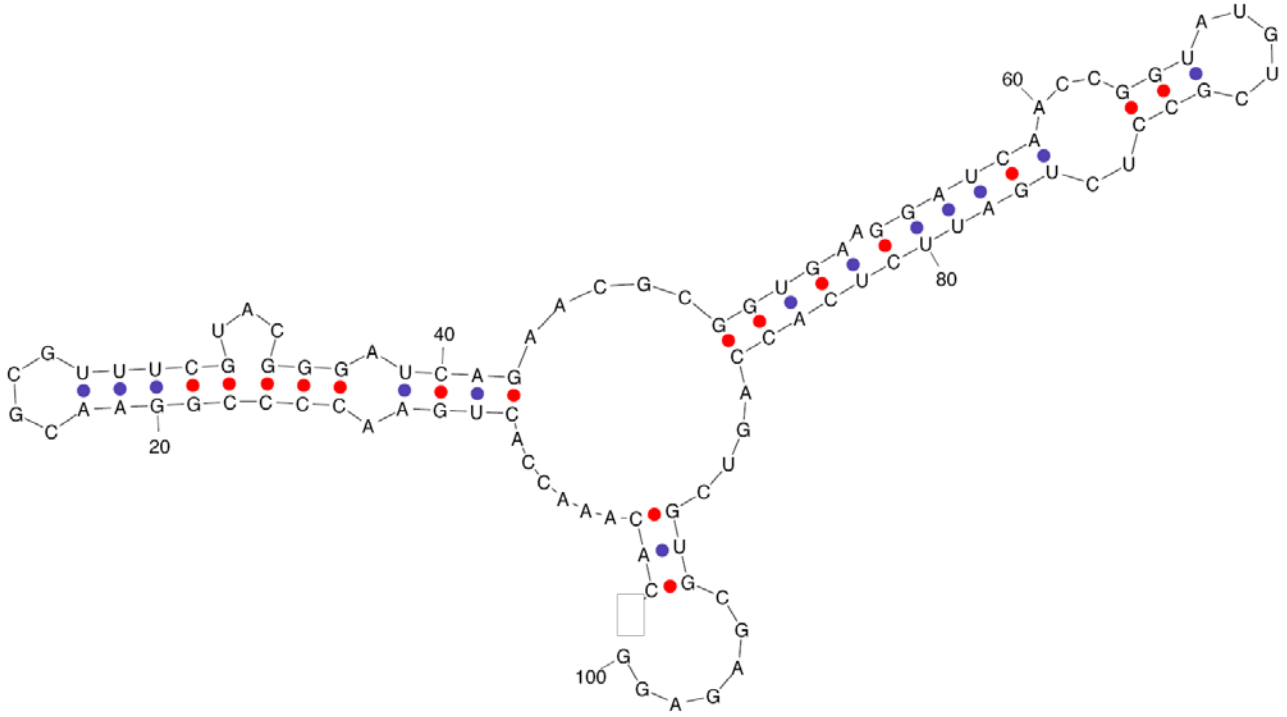
($\rightarrow$ smallest loop ~10 bases)

<k> ≈4.5 and f≈0.6

for both viral and non-viral RNAs

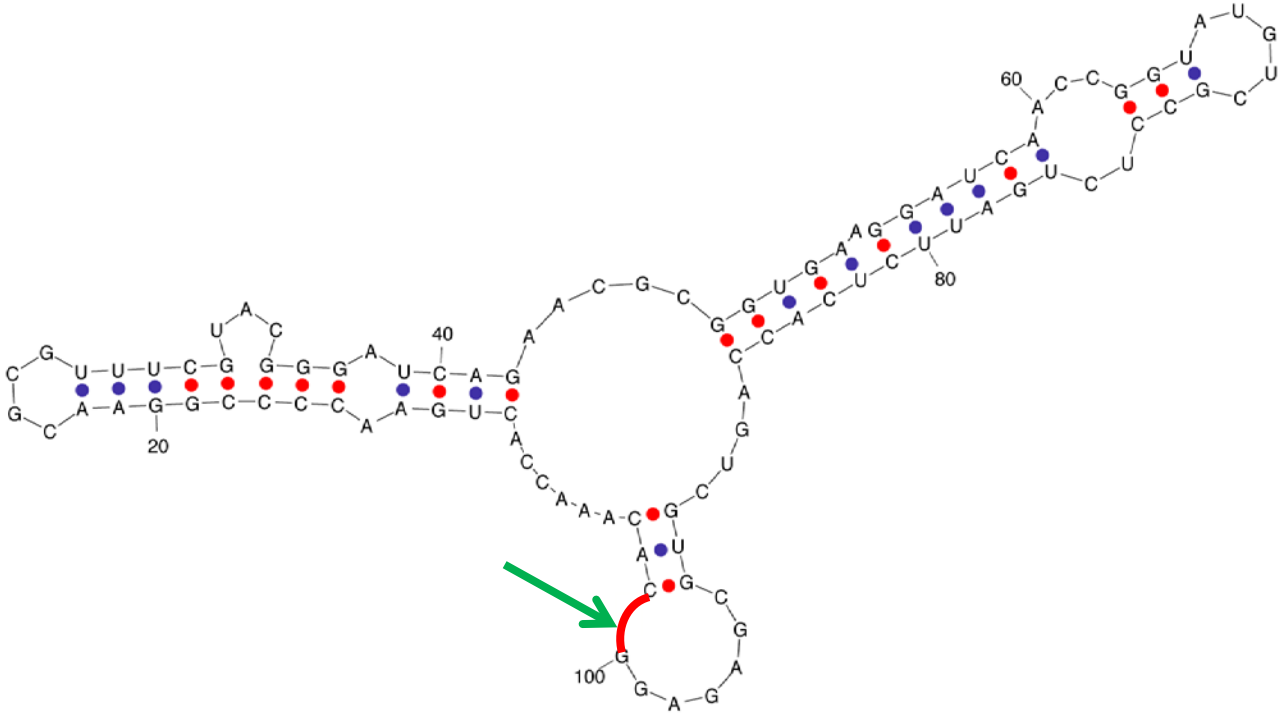
THE ENDS OF LARGE RNA MOLECULES ARE NECESSARILY CLOSE

1. "Closing the ends" of Linear ssRNA - thus forming Circular RNA does not (significantly) alter the secondary structure of the molecule – **because no base pairs are formed or disrupted**



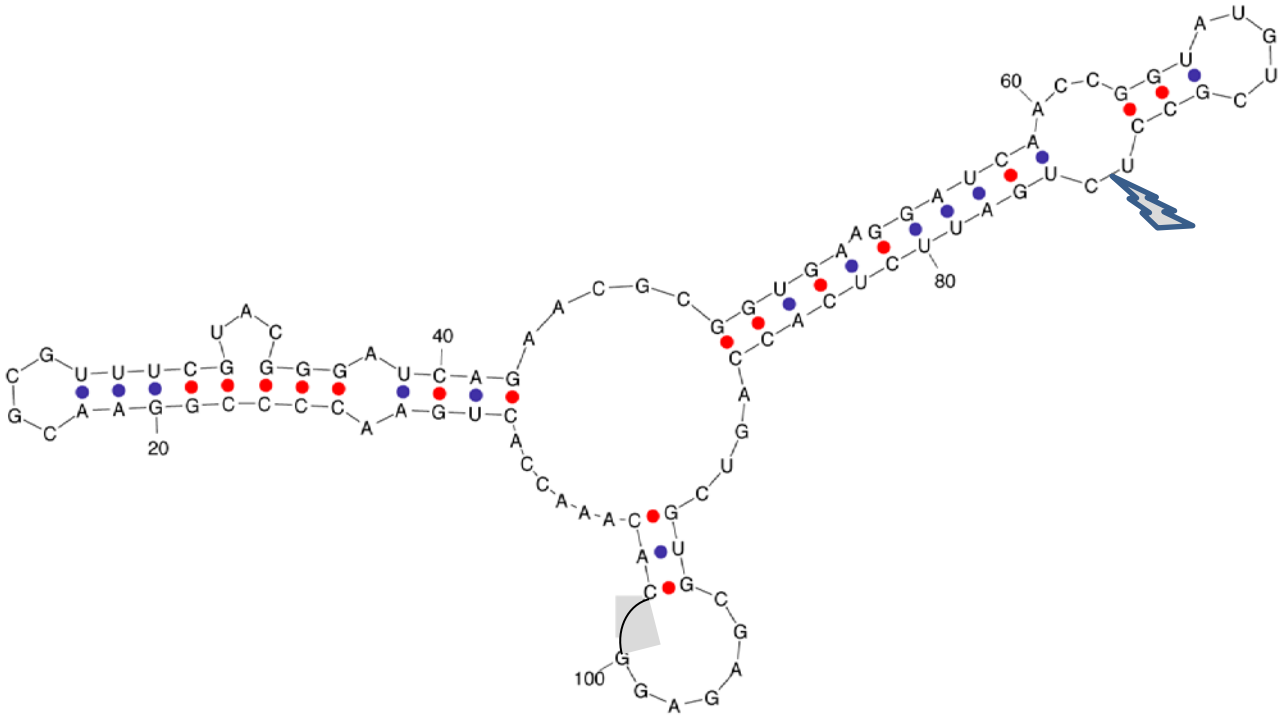
THE ENDS OF LARGE RNA MOLECULES ARE NECESSARILY CLOSE

1. "Closing the ends" of Linear ssRNA - thus forming Circular RNA does not (significantly) alter the secondary structure of the molecule – **because no base pairs are formed or disrupted**

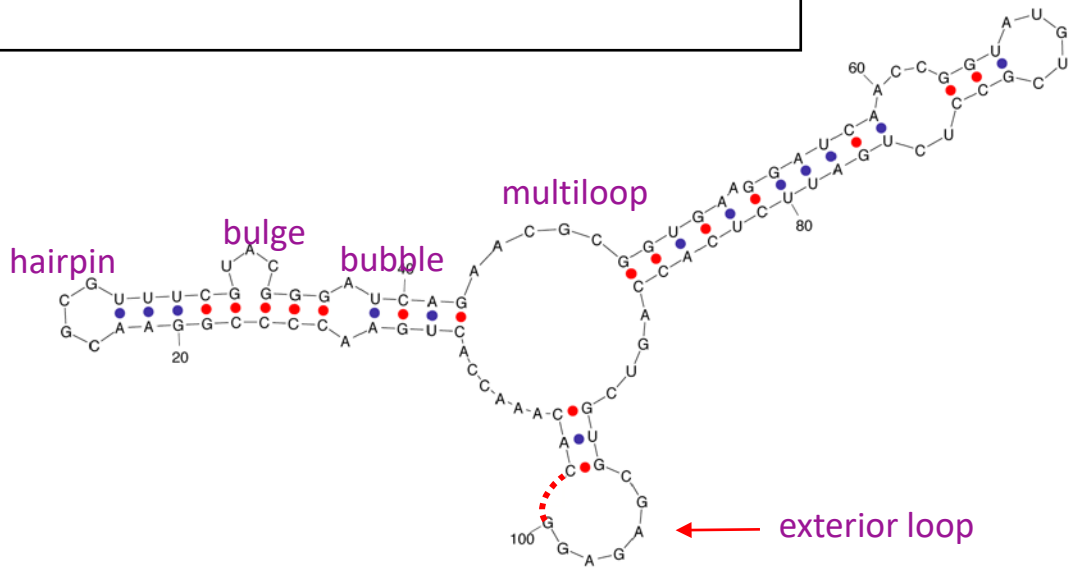


THE ENDS OF LARGE RNA MOLECULES ARE NECESSARILY CLOSE

2. Conversely, cutting a loop of a Circular RNA, thus forming a Linear RNA also does not (significantly) alter the secondary structure of the molecule – **because no base pairs are formed or disrupted.**



We thus assume: Exterior loops of linear RNA are severed internal loops of circular RNA



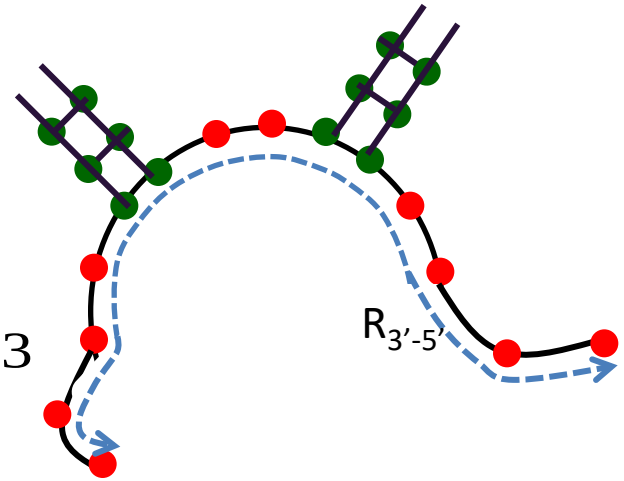
The 3'-5' distance is:

$$R_{3'-5'} \equiv l_{ss} + l_{ds}$$

l_{ss} = number of ss links

l_{ds} = number of ds links

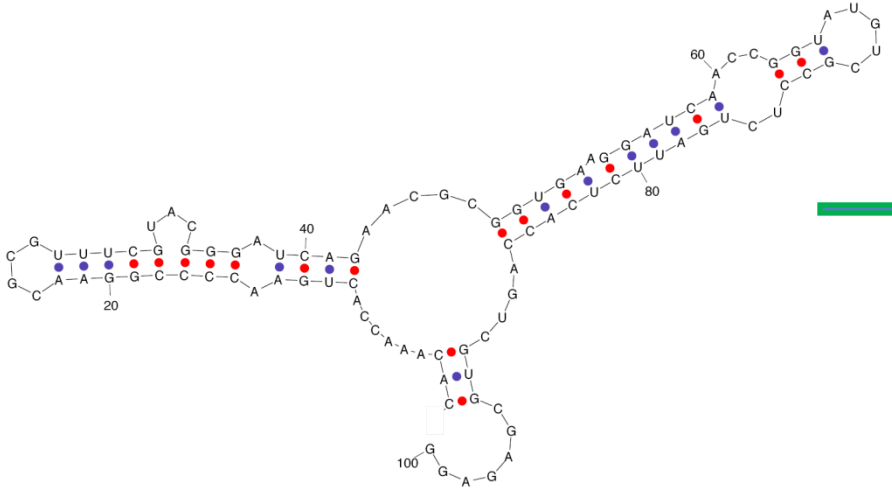
numbers of ss and ds links

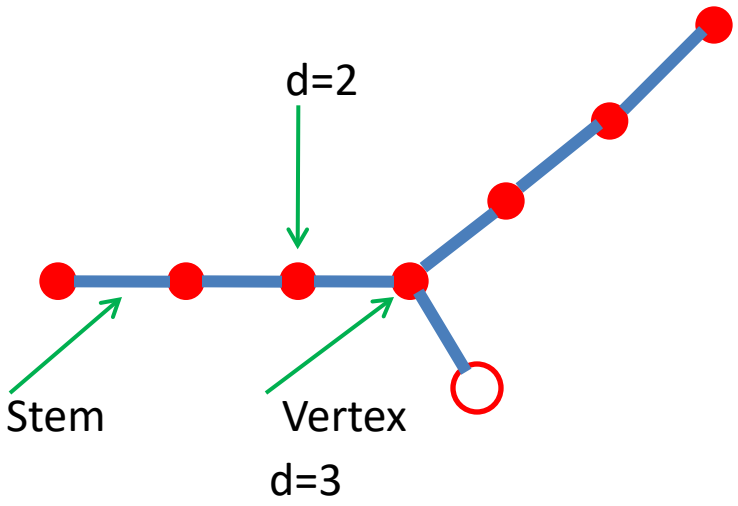


$$R_{3'-5'} \equiv l_{ss} + l_{ds} = 11+2=13$$

To calculate 3'-5' distance we note:

1. The secondary structure of ssRNA can be reduced to a simpler (tree graph) structure





S=number of stems (duplexes), L=number of vertices (loops)

$$S = L - 1$$

$$\langle d \rangle = \frac{1}{L}\sum_i d_i = \frac{1}{L}\sum_d dL_d = 2 - \frac{2}{L} \approx 2$$

L_d = # of vertices of degree d

Comprehensive analysis of RNA secondary structure topology, were carried out by Hofacker et al, and earlier by Waterman

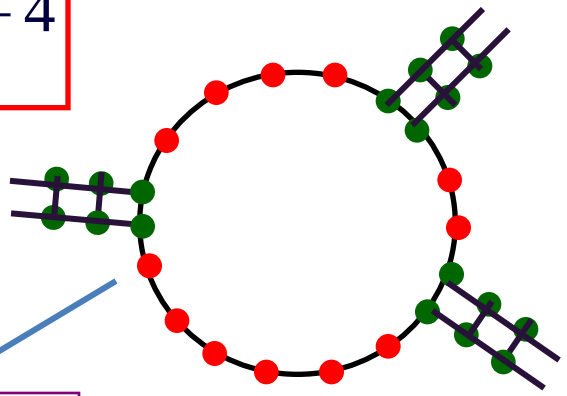
2. The average number of unpaired bases per internal loop is

$$\langle l \rangle \simeq \frac{2(1-f)}{f}\langle k \rangle$$

The total number of bases, per loop, is

$$\langle R \rangle \simeq \langle l \rangle + 2\langle d \rangle$$
$$= \frac{2(1-f)}{f}\langle k \rangle + 2\langle d \rangle = \frac{2(1-f)}{f}\langle k \rangle + 4$$

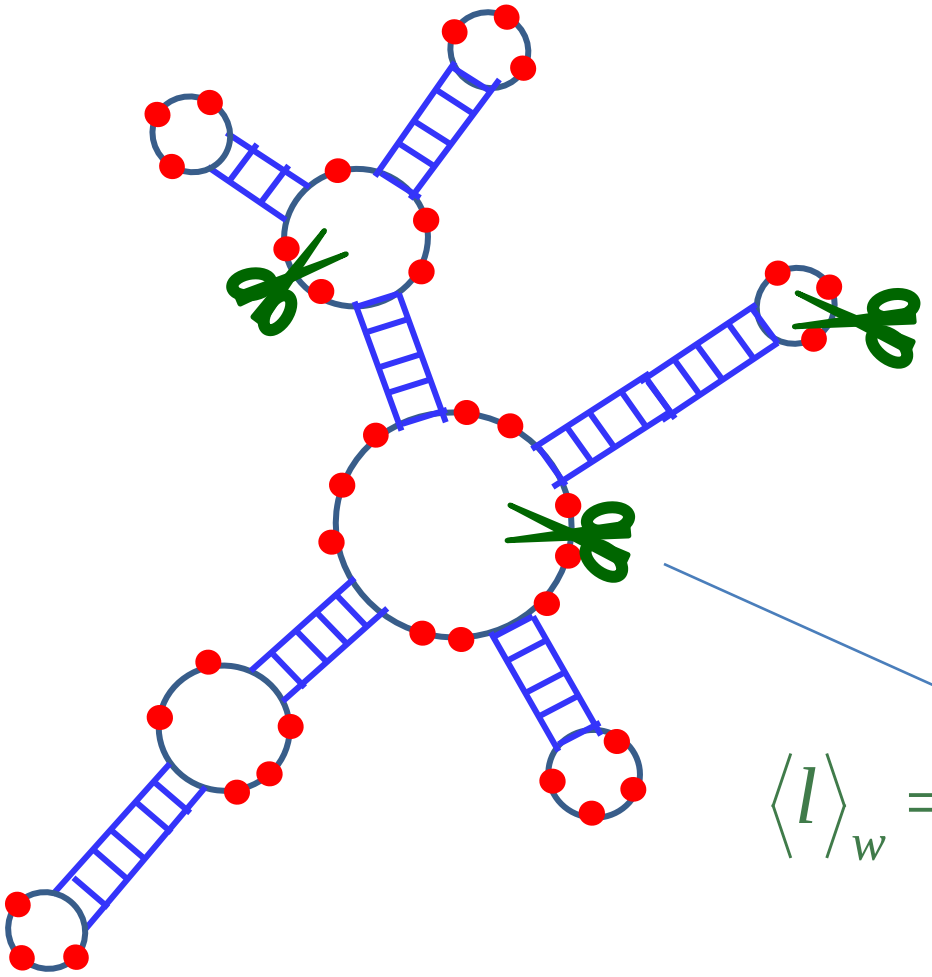
$\langle R \rangle$ *Is Independent of N !*



$$f = 0.6,\ \langle k \rangle = 4.5 \rightarrow \langle R \rangle = 10$$

$$\langle l \rangle = \frac{\#\ \text{unpaired bases}}{\#\ \text{loops}} = \frac{(1-f)N}{L} = \frac{(1-f)N}{S+1} = \frac{(1-f)N}{(fN/2\langle k \rangle)+1} \simeq \frac{2(1-f)}{f}\langle k \rangle$$

3. The average 3’-5’ distance is determined by the weight average length of the ss portions corresponding to randomly severed internal loops.



$$\langle l \rangle_w = \frac{10\times10+5\times5+3\times3+\cdots}{10+5+3+\cdots}$$

Let $n(l)$ = number of internal loops containing

l ss-bases. $\rightarrow$ The total number ("weight") of ss-bases in

l-loops is $l \times n(l) = W(l)$. Randomly severing near *any*

unpaired ss-base of *any* internal loop yields an

external loop containing l ss-bases with probability:

$$W(l) / \sum_{l \geq 1} W(l) = l \times n(l) / \sum_{l \geq 1} l \times n(l) = l \times n(l) / \sum_{l \geq 1} l \times n(l)$$

$\rightarrow$ The average number of ss-bases in an external (open) loop is

$$\langle l \rangle_{ext} = \langle l \rangle_w = \sum_{l \geq 1} l W(l) / \sum_{l \geq 1} W(l)$$

$$= \sum_{l \geq 1} l^2 \times n(l) / \sum_{l \geq 1} l \times n(l) = \langle l^2 \rangle_n / \langle l \rangle_n$$

($\langle \ \rangle_w$ and $\langle \ \rangle_n$ denote Weight and Number Averages, respectively)

Random distribution of the $W = \sum_{l \geq 1} W(l)$ ss-bases among

the L interior loops: $\rightarrow P(l) \equiv n(l) / \sum_{l \geq 1} n(l) = e^{-\lambda l} / q$ with $\langle l \rangle_n = \sum_{l \geq 1} l P(l)$

$\rightarrow \lambda = \ln[\langle l \rangle_n / (\langle l \rangle_n + 1)]$

($q = \sum_{l \geq 1} e^{-\lambda l} = 1/(e^{\lambda} - 1)$). This finally yields the $\langle l \rangle_w = \langle l \rangle_{ext} = 2 \langle l \rangle_{\text{int}} - 1$

On average – the number of ss-bases per external loop is:

$$\langle l \rangle_{ext} = 2\langle l \rangle_{\text{int}} - 1$$

And because, on average, there are two branches (4 ds-bases) per loop:

$$\langle R_{3'-5'} \rangle = 2\langle l \rangle_{\text{int}} + 3 = \frac{4(1-f)}{f}\langle k \rangle + 3$$

For $f = 0.6$ and $\langle k \rangle = 4.5$ we find $\langle R_{3'-5'} \rangle \approx 15$

Conclusions and Summary

Viral RNA is Less Extended, More Branched, than Nonviral RNA

A Simple (Sequential) Folding Predicts Basic Secondary Structure Properties of RNA

The Two Ends of Linear ssRNA are Necessarily Close

Rg of ssRNA Sacles as $Rg \sim N^{1/3}$.

Main Present effort: Viral Self Assembly: Experiment and Theory