WHAT MAKES RNA GENOMES SPECIAL?

SEARCH FOR THE HYDROGEN ATOM OF VIRUSES

WILLIAM M. GELBART
UCLA

115th Statistical Mechanics Conference
RUTGERS UNIVERSITY, BUSCH CAMPUS, HILL CENTER
May 9, 2016
Honoring David Chandler, John Weeks, and Julia Yeomans

(SIMPLEST) VIRUSES ARE JUST:

A COMPOSITE OF

a nucleic acid genome (RNA or DNA)

AND

a protein shell -- **“capsid”**

cylindrical

spherical

helical symmetry

icosahedral symmetry

HOW ARE *RNA VIRUSES* DIFFERENT

(FROM *DNA VIRUSES*)?

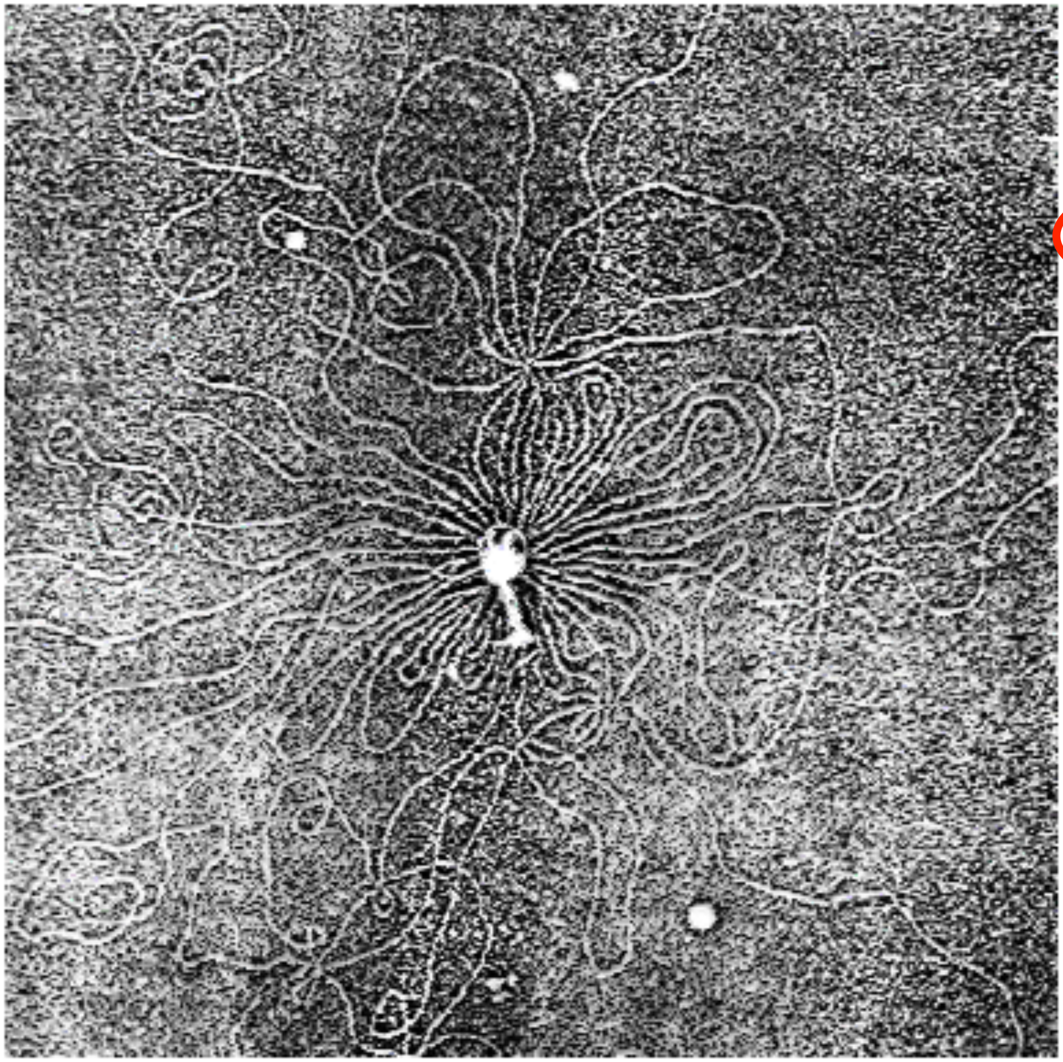
(single-stranded [ss]) **RNA viruses (mostly plant and animal)**

ssRNA is weakly confined; packaged spontaneously

(double-stranded [ds]) **DNA viruses (mostly bacterial)**

dsDNA is strongly confined; packaged by force

A gene of DNA is a very different *physical object* than a gene of RNA


Kleinschmidt et al. (1962)

Osmotically-shocked bacteriophage T2

$$R_{capsid} \approx 25nm \ll R_{DNA} \approx 1\mu$$

DNA "contour length" $L, \approx 20\mu$

DNA "persistence length" $\xi, \approx 50nm$

DNA "size" $(L\xi)^{1/2}, \approx 1\mu$

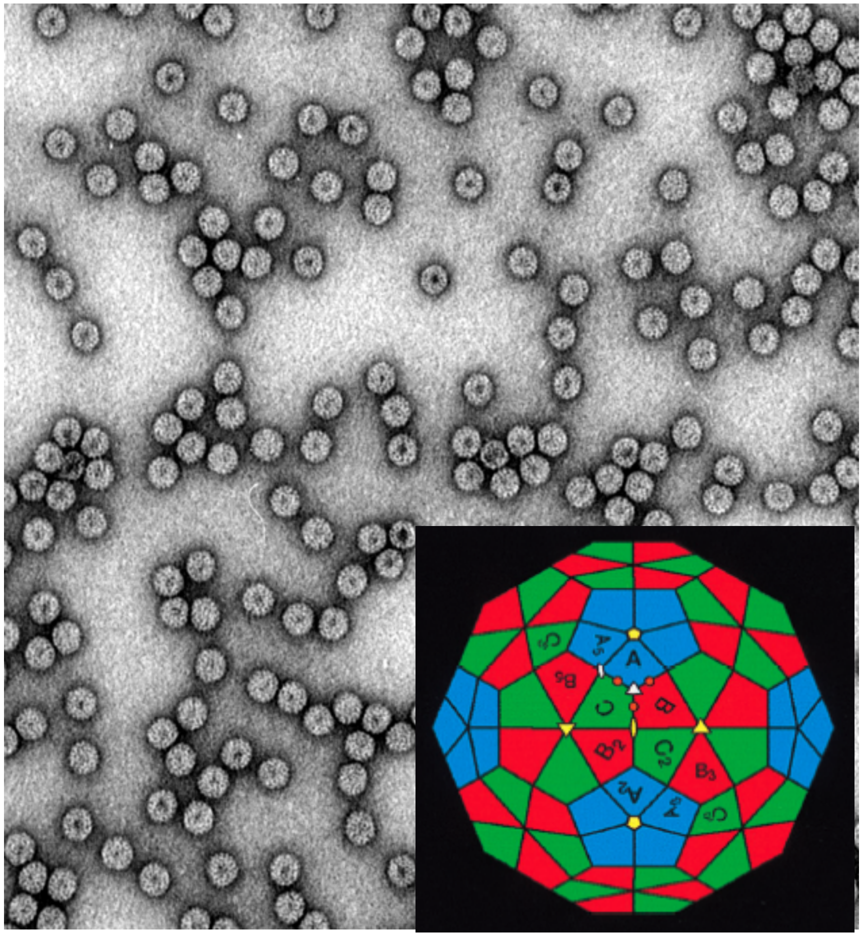
$$L >> \xi$$

$$R_{DNA} \approx (L\xi)^{1/2} \sim M^{1/2}$$

LARGE DNAs ARE LINEAR STATISTICAL OBJECTS WITH WELL-KNOWN CHARACTERISTICS

A LOT OF WORK HAS TO BE DONE, TO PACKAGE THE DNA GENOME *INTO A PRE-FORMED CAPSID* – IT IS … *PRESSURIZED!*

What about viruses with *single-stranded (ss) RNA genomes?*



E.g., Cowpea Chlorotic Mottle Virus (CCMV)

Each identical 28nm-capsid consists of *exactly* 180 copies of one protein, and contains a *different* molecule of the viral RNA genome – RNA1, RNA2, *or* RNA3 (+RNA4) – *each about 3000nt long*

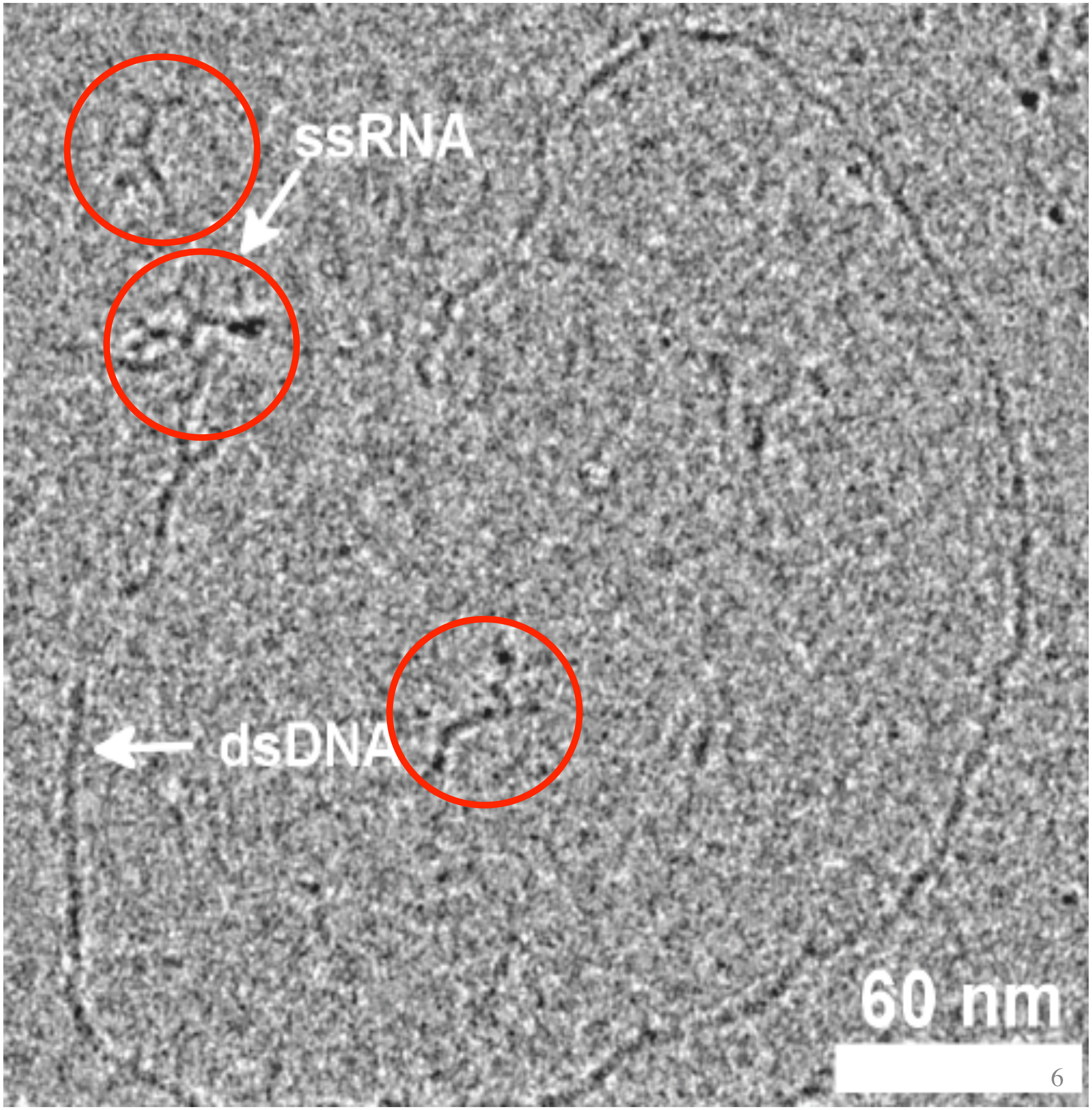
Packaging of genome occurs spontaneously,

***via self-assembly* – no work, no pressure – WHY? HOW?**

BECAUSE ssRNA MOLECULES – *GENES* –

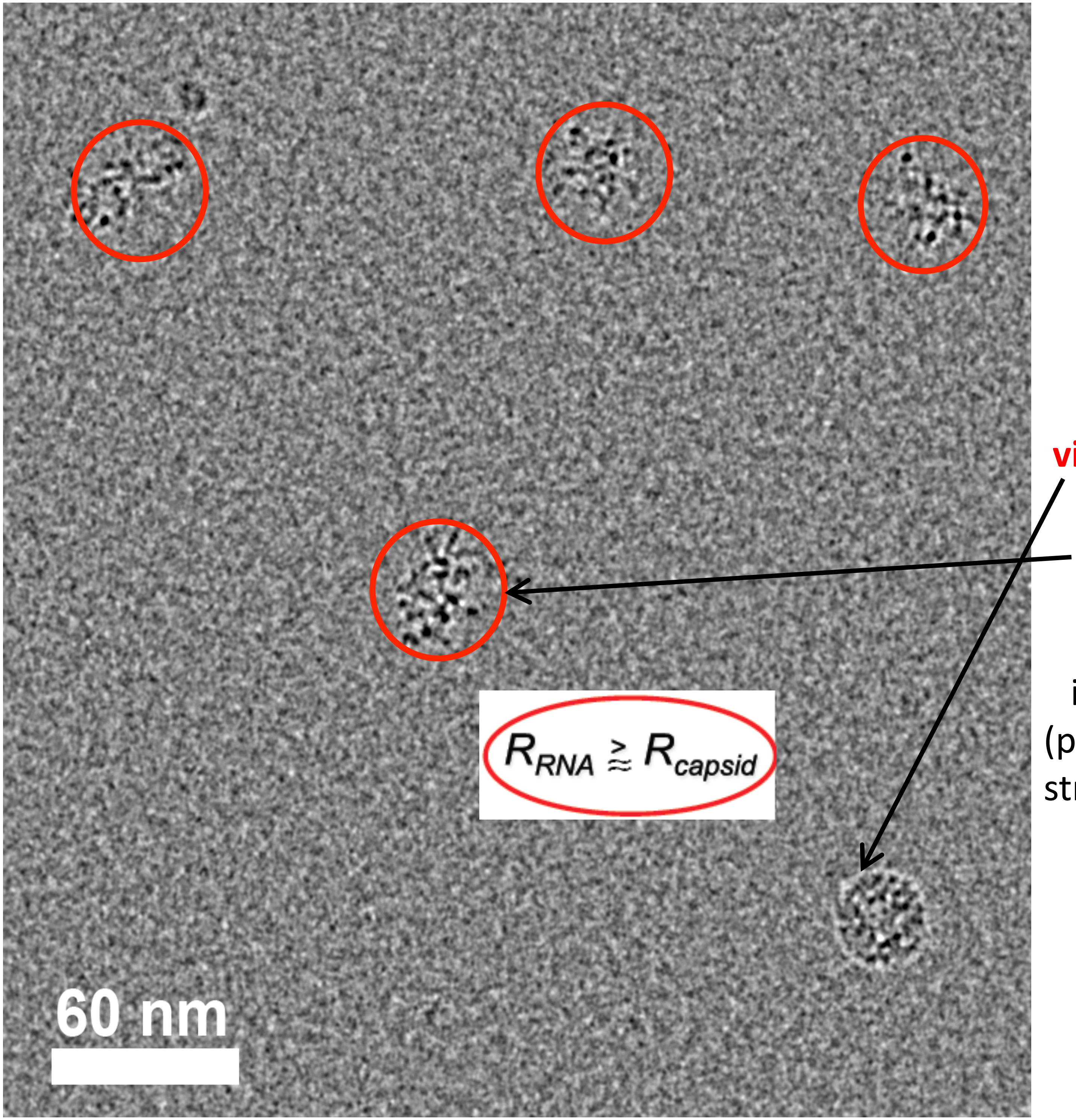
ARE TOTALLY DIFFERENT FROM THEIR DNA COUNTERPARTS…

COMPARISON BETWEEN
2117 nt ssRNA
AND
2117 bp dsDNA



Gopal, Zhou, Knobler, and Gelbart
RNA ***18***, 284 (2012)

TE buffer pH 7.4



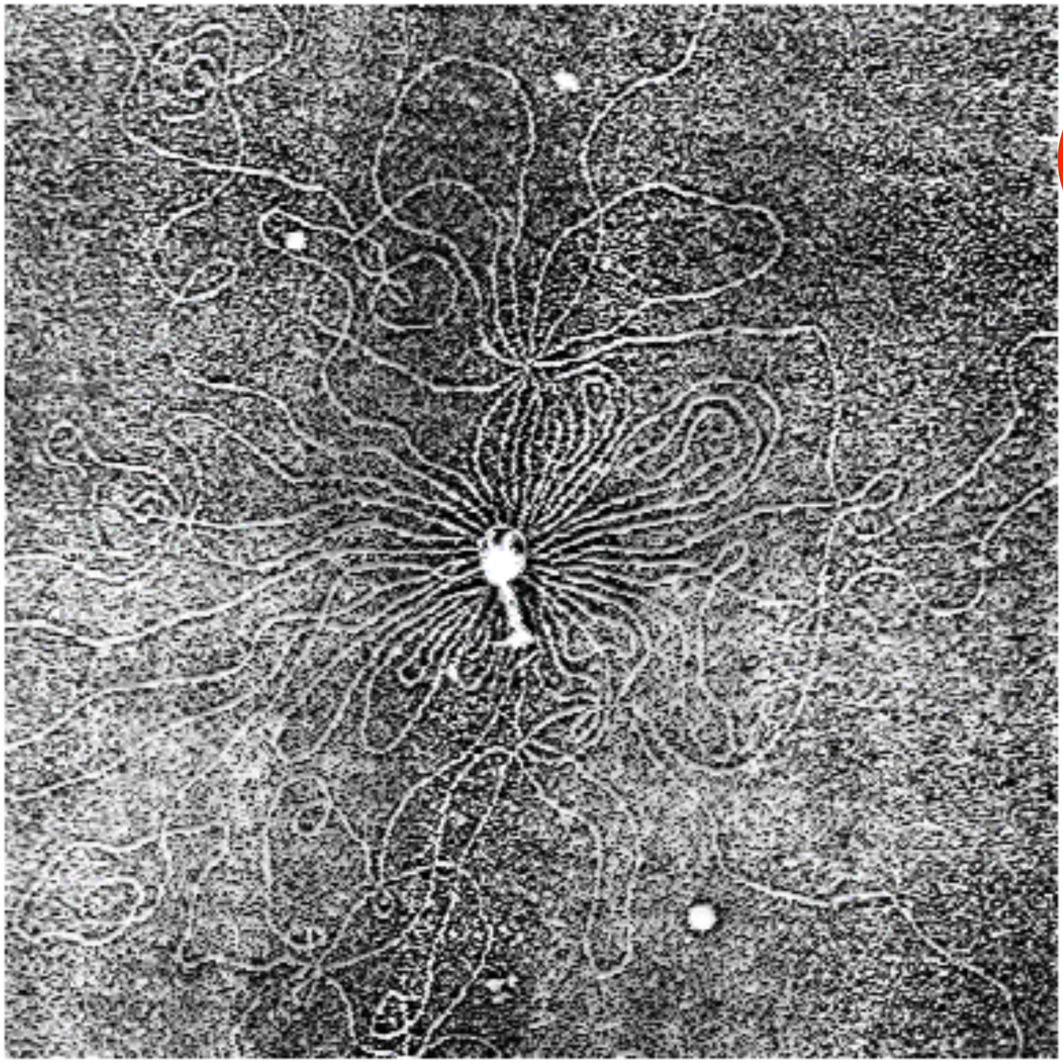
(cryo-EM micrograph)

virion of CCMV, and

its gene content –
ss RNA2 (2774 nt)

in assembly buffer
(physiological pH/ionic
strength, with Mg^{2+})

Gopal, Zhou, Knobler,
Gelbart, *RNA* (2012)



Kleinschmidt et al. (1962)   Osmotically-shocked phage T2

In contrast…

Large DNA is a stiff and linear, statistical object, taking up a lot of space, i.e., it is highly ramified

We know all its configurational properties, independent of sequence, if we know its contour length L and its stiffness ξ

Further,….

$$R_{DNA} \approx (L\,\xi)^{1/2} >> R_{capsid}$$

NOT SO…..

for long RNA molecules

Compactness of ssRNA accounts for smaller size of virus

What determines the size, shape, and flexibility of ssRNA?
It is *not* a linear polymer, *nor* even a simple *branched* one…

Don’t know how its size depends on molecular weight (length),
or on sequence (and secondary structure)….

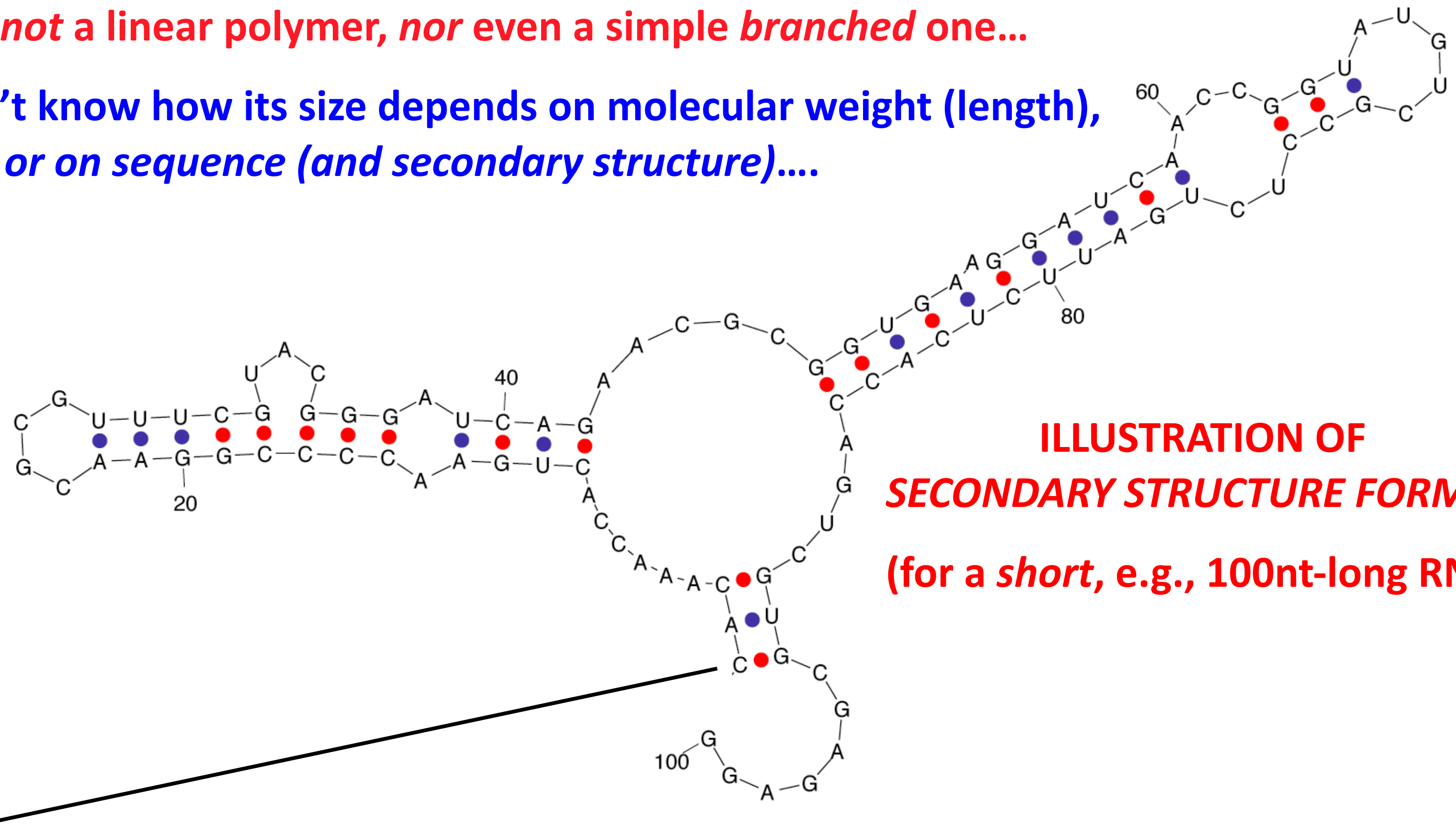


ILLUSTRATION OF
SECONDARY STRUCTURE FORMATION

(for a *short*, e.g., 100nt-long RNA)

(5’)CACAAACCACUGAACCCCGGAACGCGUUUCGUACGGGAUCAGAACGCGGUGAAGGA
UCAACCGGUAUGUCGCCUCUGAUUCUCACCAGUCGUGCGAGAGG(3’)

But what about many *1000s*-of-nt-long RNA?

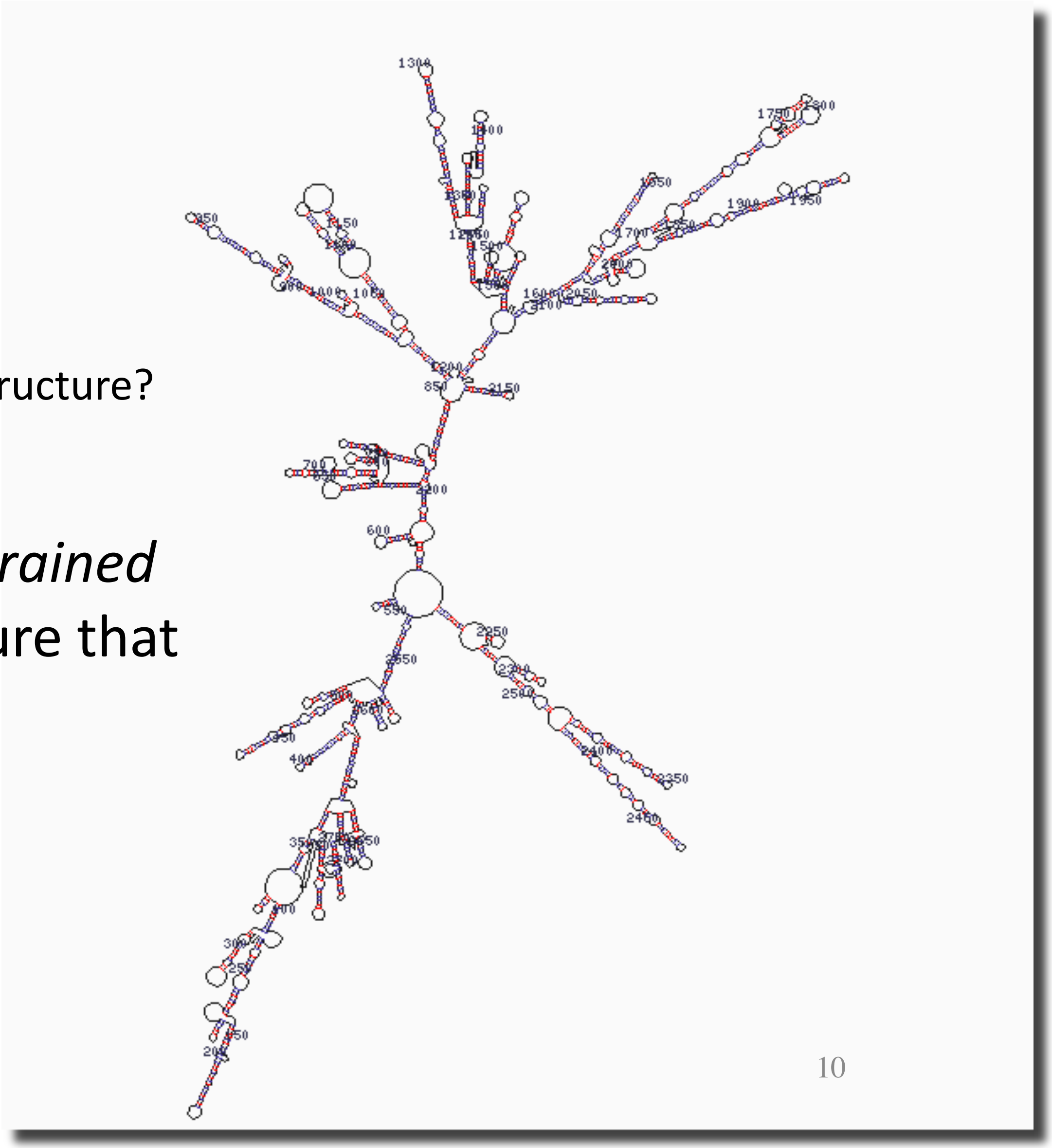
What *coarse-grained* features of its <u>branching</u> and <u>shape</u> and <u>overall size</u> can we infer from its sequence?

How does the *3D size and shape* of an arbitrarily long RNA correspond to its *secondary structure*?

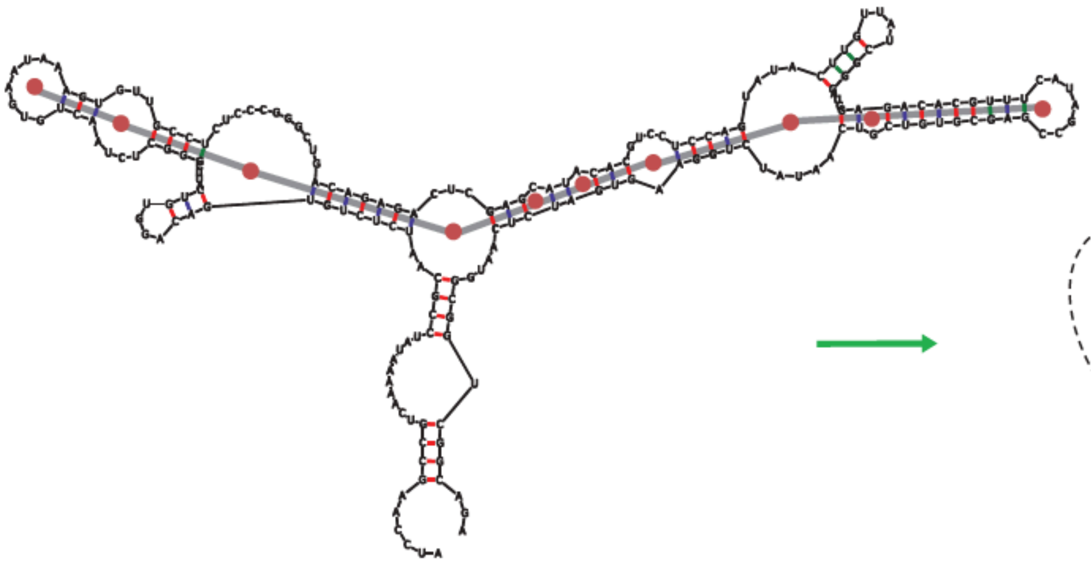
1st example:
RNA2, the molecule comprising the second gene of CCMV, 2774 nt long

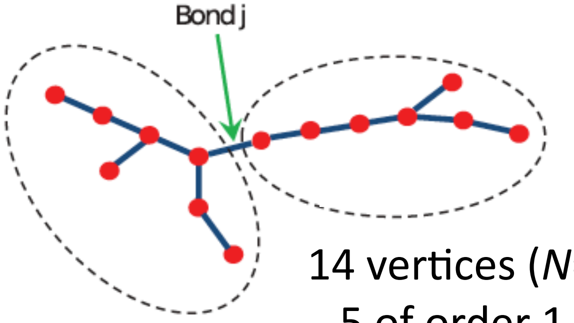
Is anything special about its secondary structure?

Can we identify some *coarse-grained* feature of its secondary structure that determines its overall 3D size?



Each RNA secondary structure can be mapped onto a "tree graph" (and a *primary sequence* is mapped onto an *ensemble* of tree graphs…)





14 vertices (*N*=14)
5 of order 1 (N_1=5)
6 of order 2 (N_2=6)
3 of order 3 (N_3=3)

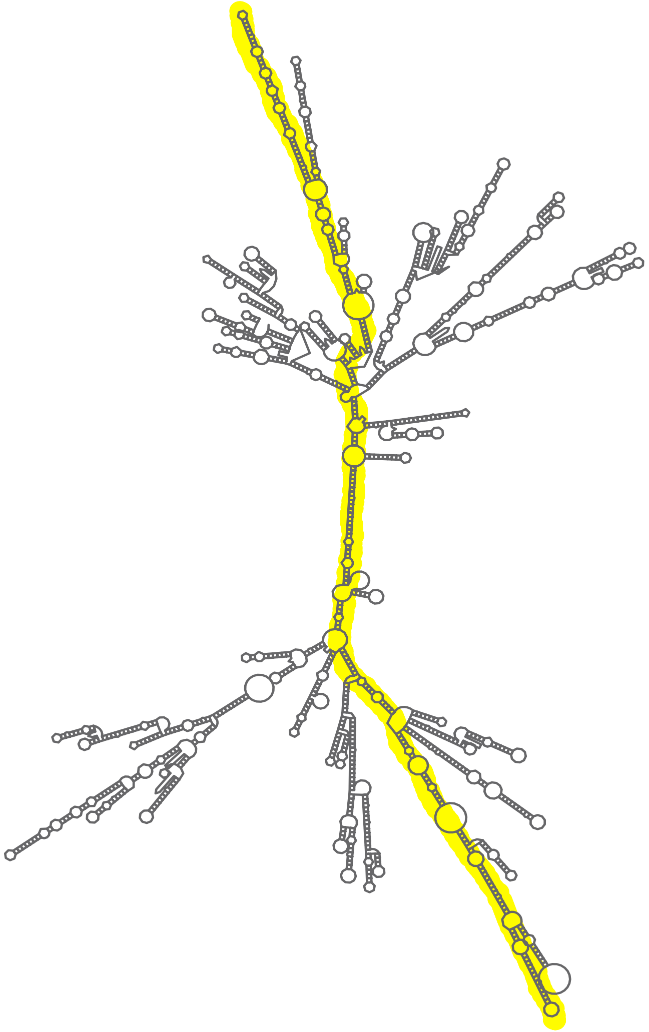
$$\{N_d\} = (N_1, N_2, N_3, \ldots.) = (5,6,3)$$

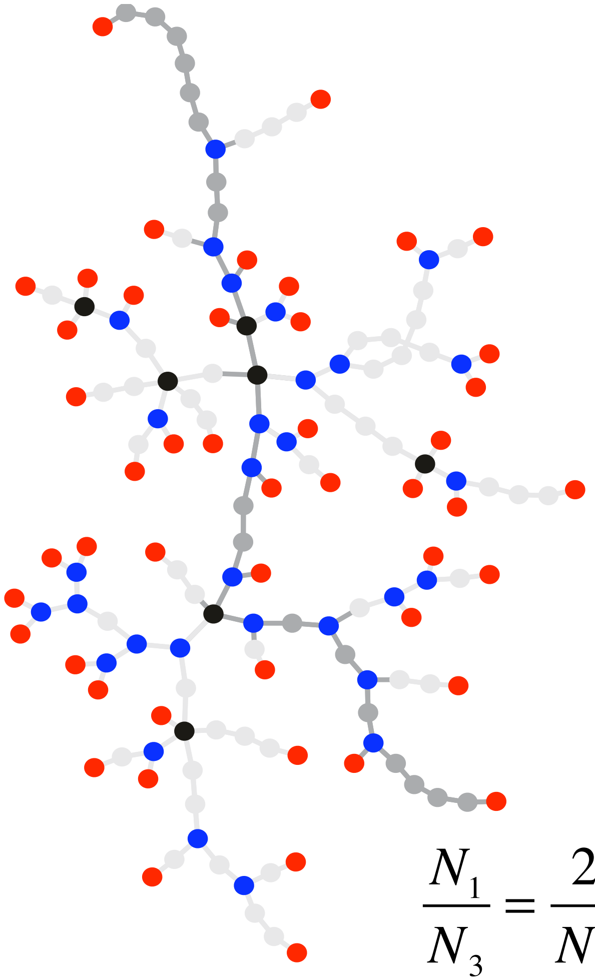
$$N = \sum_{d=1,2,3,\ldots} N_d = 14$$

$$< R_g^2 > = \frac{b^2}{N^2} \sum_{j=1}^{N-1} N_{left}(j)\, N_{right}(j), \quad b = edge\ (Kuhn)\ length$$

KRAMERS, 1940

Back to 2774 nt RNA2 of CCMV:





RNA Length		2774nt
%GC		42
Avg. Duplex Length including single-base bubbles.		5
TREE GRAPHS	Total Vertices in Tree Graph	N=149
	Terminal Vertices (i.e. stem loops)	$N_1=48$
	2-fold Junctions	$N_2=64$
	3-fold Junctions	$N_3=30$
	4-fold Junctions	$\mathbf{N_4= 5}$
	5-fold Junctions	$N_5= 2$
	6-fold Junctions *	$V_6= 0$

$$\frac{N_1}{N_3} = \frac{2}{N_3} + 1 \boxed{+ 2\frac{N_4}{N_3} + 3\frac{N_5}{N_3} +} \ldots$$

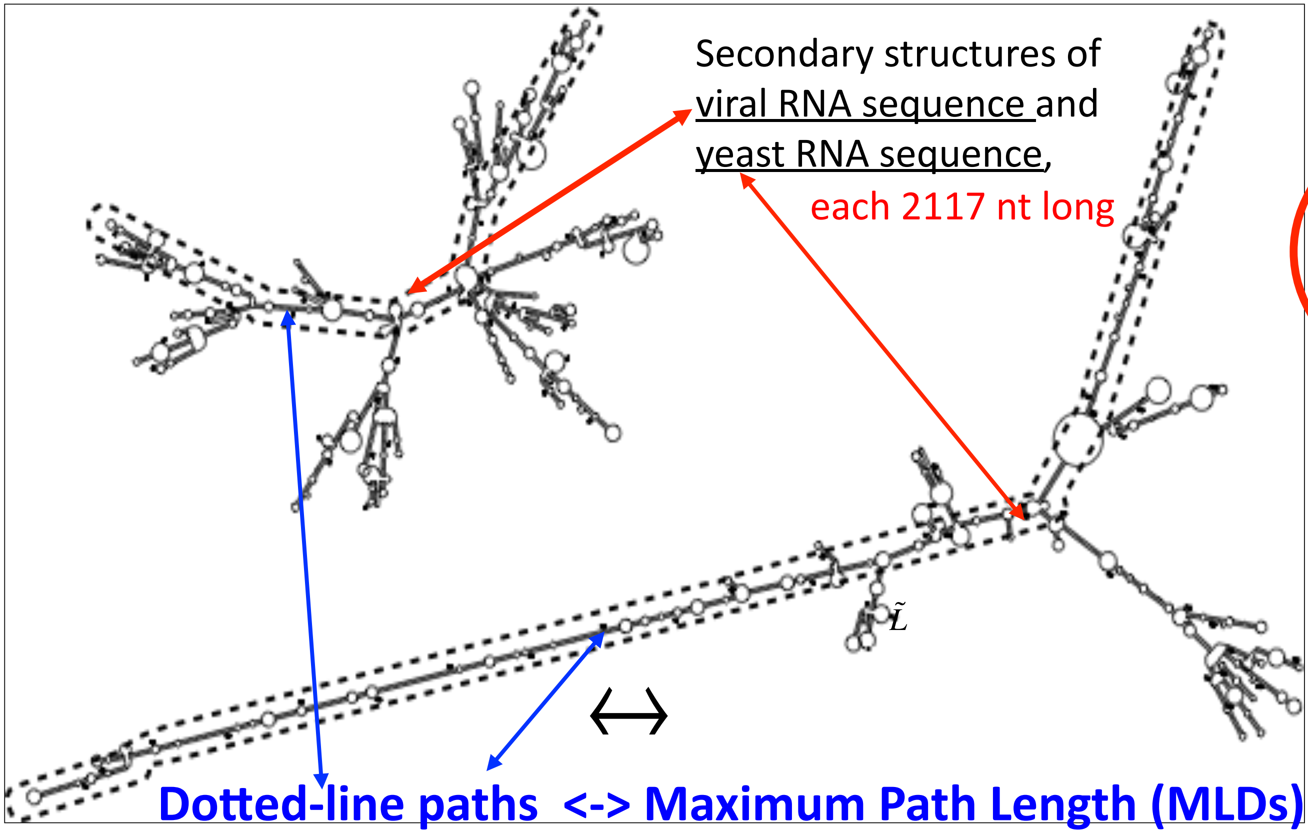
<u>EULER, 1750</u>

<u>LONG RNA IS A FLEXIBLE, *BRANCHED,* STATISTICAL OBJECT WITH *UNKNOWN* CHARACTERISTICS</u>

(E.g., we don't know the significance of the *distribution of vertex orders*, $\{N_d\}$)

For long RNAs: $N_3 >> 1 \Rightarrow (\frac{N_1}{N_3} - 1)$ is a measure of *higher-order branching … compactness*

Another example: RNA 3, the third molecule (2117 nt) of the CCMV genome



Secondary structures of viral RNA sequence and yeast RNA sequence, each 2117 nt long

Dotted-line paths <-> Maximum Path Length (MLDs)

Viral sequences have smaller MLDs, implying smaller Rgs, and hence are *more compact*

Yoffe, Prinsen, Gopal, Knobler, Gelbart, Ben-Shaul, *PNAS (USA)* **105**, 16153 (2008)

WHAT ACCOUNTS FOR THE DIFFERENCES IN THEIR *3D* SIZES?

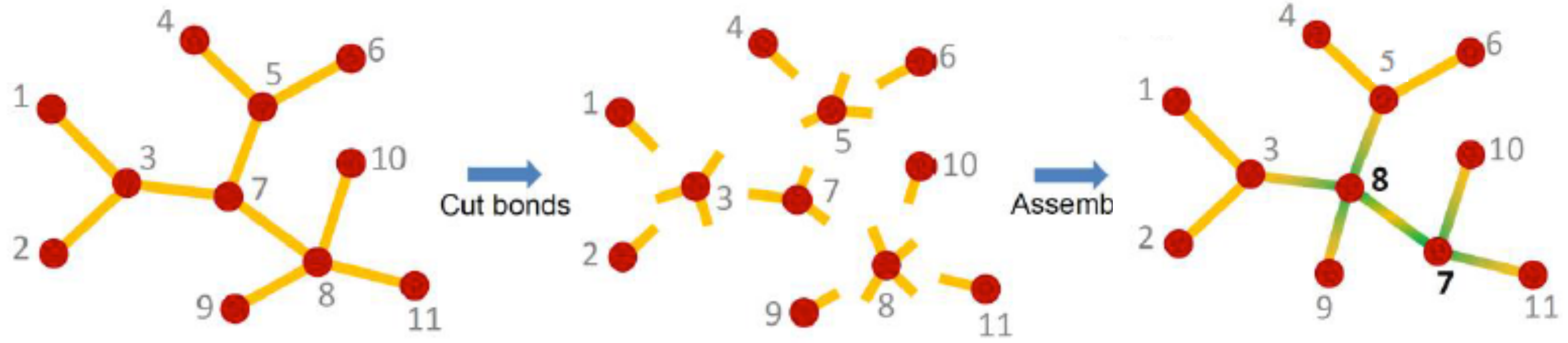
E.g., treat RNA as a *linear homo*polymer with $\tilde{L} \leftrightarrow MLD$:

$$R_g \approx (\tilde{L}\tilde{\xi})^{1/2} \to (MLD\,\tilde{\xi})^{1/2} \propto MLD^{1/2}$$

Here $\tilde{L}$ is the (*effective*) contour length, and $\tilde{\xi}$ is the (*effective*) persistence length

BOTH STRUCTURES HAVE COMPARABLE BRANCHING, E.G., SIMILAR $\{N_d\}$s, BUT ONE HAS A *MORE COMPACT CLUSTERING OF BRANCHING POINTS…*

PRÜFER, 1918 : AN ARBITRARY TREE GRAPH CAN BE REPRESENTED BY **A UNIQUE ORDERED SEQUENCE OF INTEGERS** – “THE PRUFER CODE”



$$\{N_d\} = (N_1, N_2, N_3, ...) = (7,0,3,1)$$

$$N = \sum_{d=1,2,3,...} N_d = 11$$

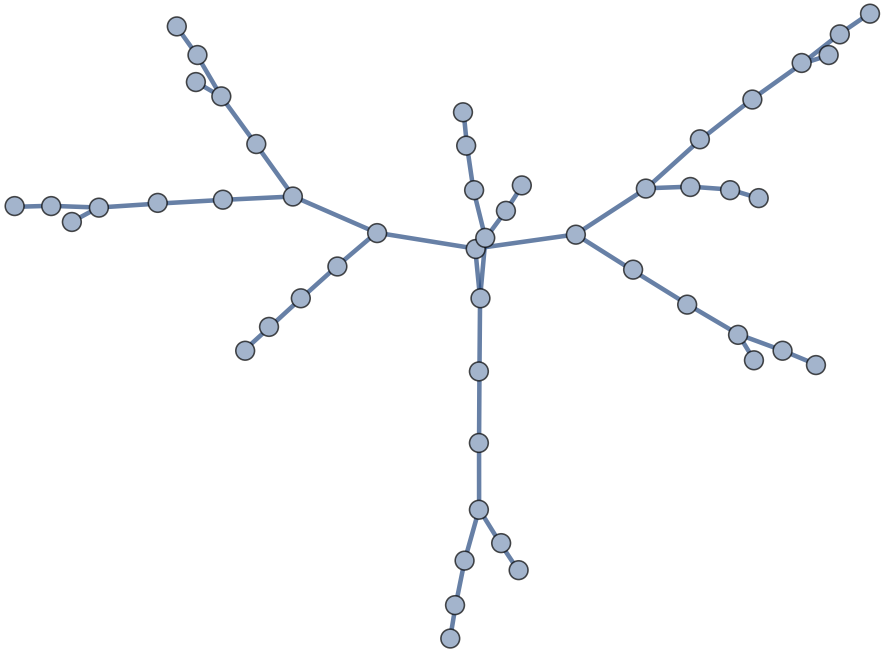
Prufer code: ↓ ↓
3, 3, 7, 5, 5, 7, 8, 8, 8
↑ ↑

Prufer code: ↓ ↓
3, 3, 8, 5, 5, 8, 8, 7, 7
↑ ↑

Here we have “shuffled” the Prufer sequence by permuting two pairs of integers

Prufer shuffling: Leaves invariant *the number of vertices (N)*,
and *the distribution of vertex orders* $\{N_d\}$...but
changes *the connectivity* of the graph, and hence its “size” (compactness)

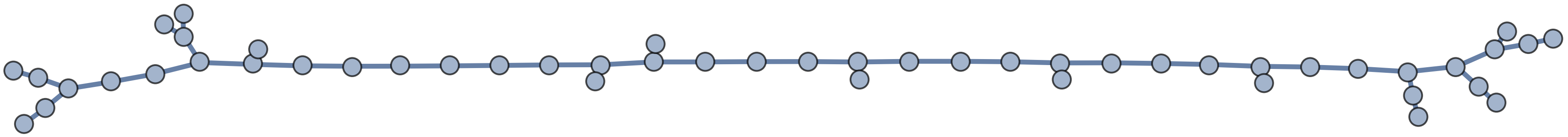
Singaram, Garmann, Knobler, Gelbart, Ben-Shaul, *J. Phys. Chem. B* **119**, 13991 (2015)



These two tree graphs have the *same number of vertices*, and the *same distribution of vertex orders*,

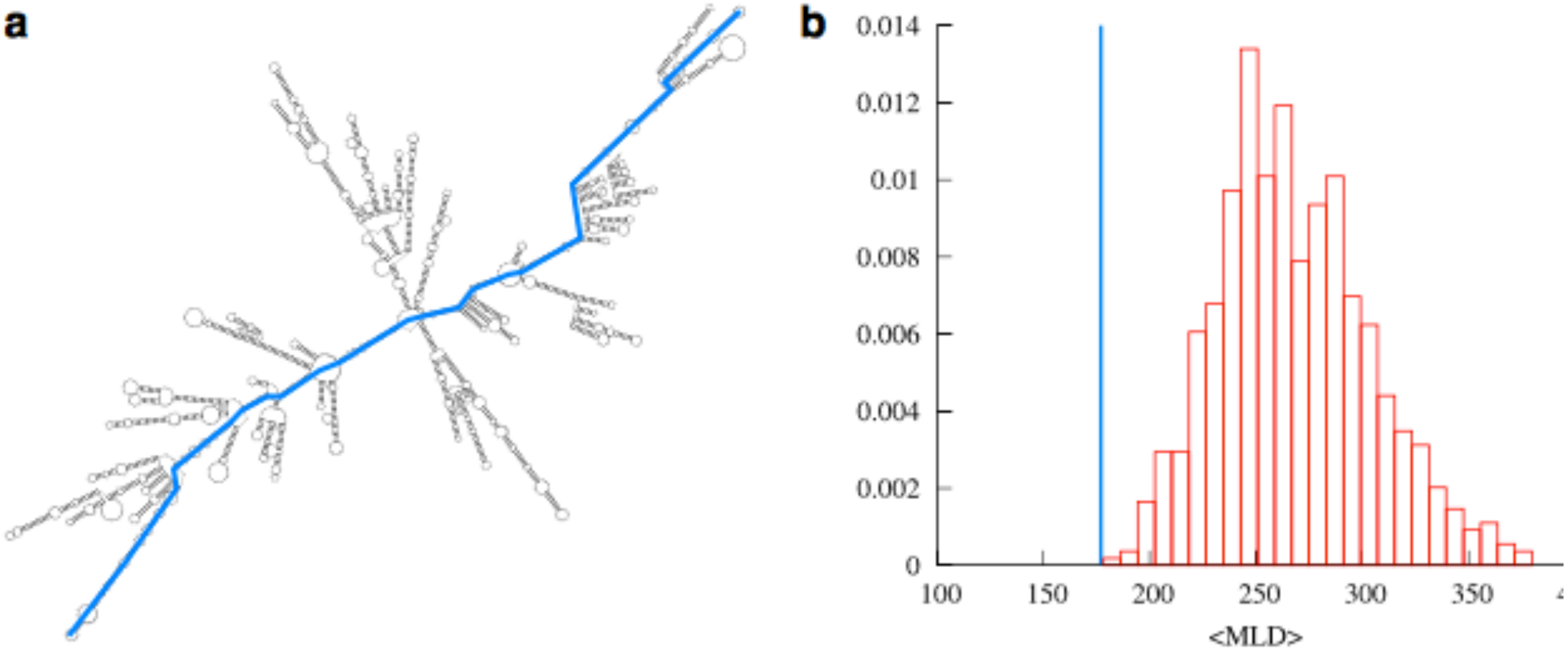
$$\{N_d\} = (N_1, N_2, N_3, \ldots.) = (14, 24, 12)$$

$$N = \sum_{d=1,2,3,\ldots} N_d = 50$$



The 1st has been obtained from the 2nd by successive permutations of its Prufer sequence, *chosen to decrease the maximum path length* (MLD), and hence its radius of gyration, or – equivalently – *the compactness of its branch points*

VIRAL SEQUENCES ARE MORE COMPACT…



a Typical secondary structure of BMV RNA2, with the maximum path length (MLD) shown in **blue**.

b Thermally-averaged maximum path length (**blue line**) of the viral (RNA2) sequence, and the distribution (see **red histogram**) of averaged maximum path lengths for each of many random sequences with the same length and nt composition.

VIRAL RNA GENOMES HAVE EVOLVED – NOT JUST TO CODE FOR CERTAIN PROTEINS – BUT ALSO TO BE *MORE COMPACT*

Viral sequences are more compact….

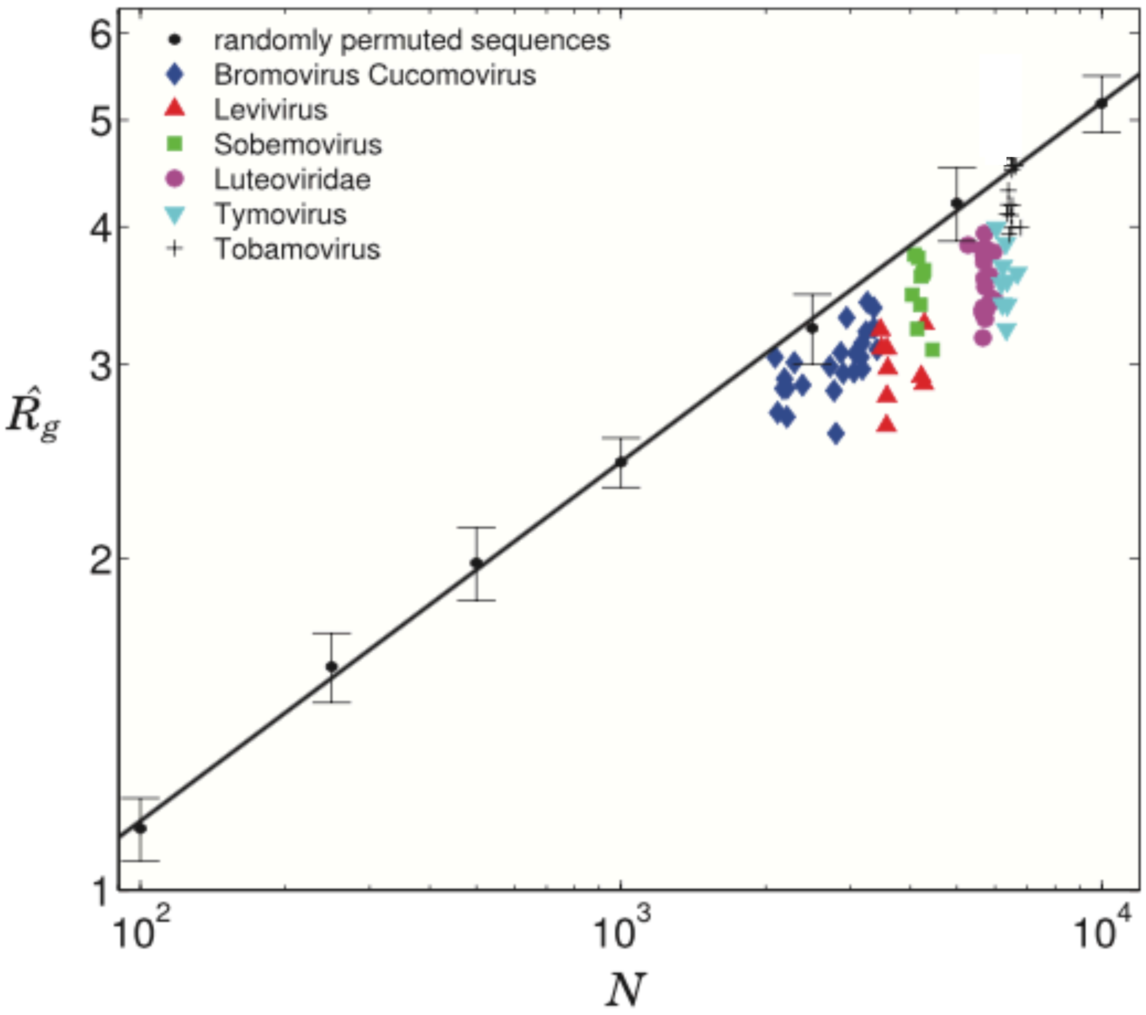
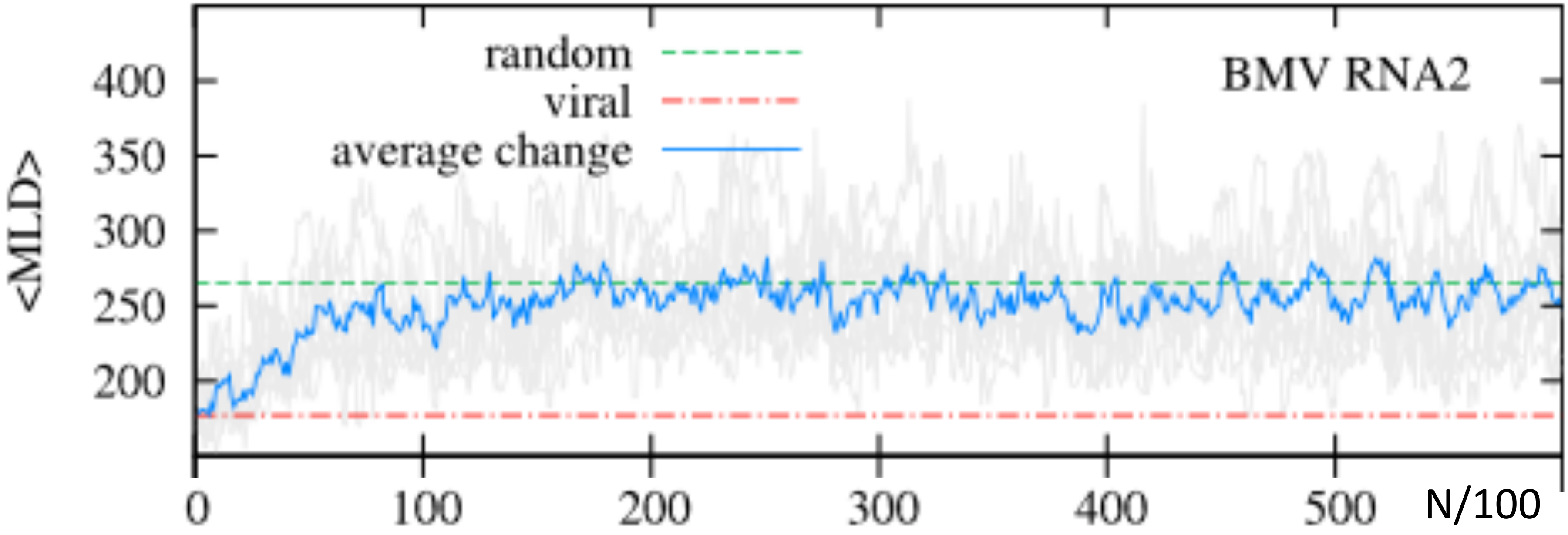


FIG. 3. A log-log plot of the radius of gyration, $\hat{R}_g$ (in units of segment length, b), as a function of RNA sequence length, N. Each black dot represents the average result obtained for 20 randomly shuffled sequences of equal base composition.

SYNONYMOUS MUTATIONS HAVE BEEN RECRUITED TO COMPACTIFY VIRAL SEQUENCES



Average maximum path length (<MLD>) after N synonymous mutations have been introduced into the viral (CCMV RNA2) sequence, for 10 trajectories (grey) and their average (**blue**)
Red dot-dash line: average value for viral sequence
Green dashed line: average value for random sequences with same length and nt composition

Tubiana, Bosic, Micheletti, Podgornik, *Biophys. J.* **108**, 194 (2015)

Ben-Shaul, Gelbart, *Biophys. J.* **108**, 14 (2015)

RNA VIRUSES ARE SMALLER THAN DNA VIRUSES, because

(physical reason) **RNA genomes are more compact, per gene**

(biological reason) **RNA viruses have fewer genes (smaller genomes)**

- **Compactness of RNA genomes derives from their being (effectively) highly branched polymers**

  And viral RNA is more *compactly* branched (*smaller!*) than non-viral RNA, thereby competing better for binding (and packaging) by capsid protein

- ***How* few genes can an RNA virus have?**

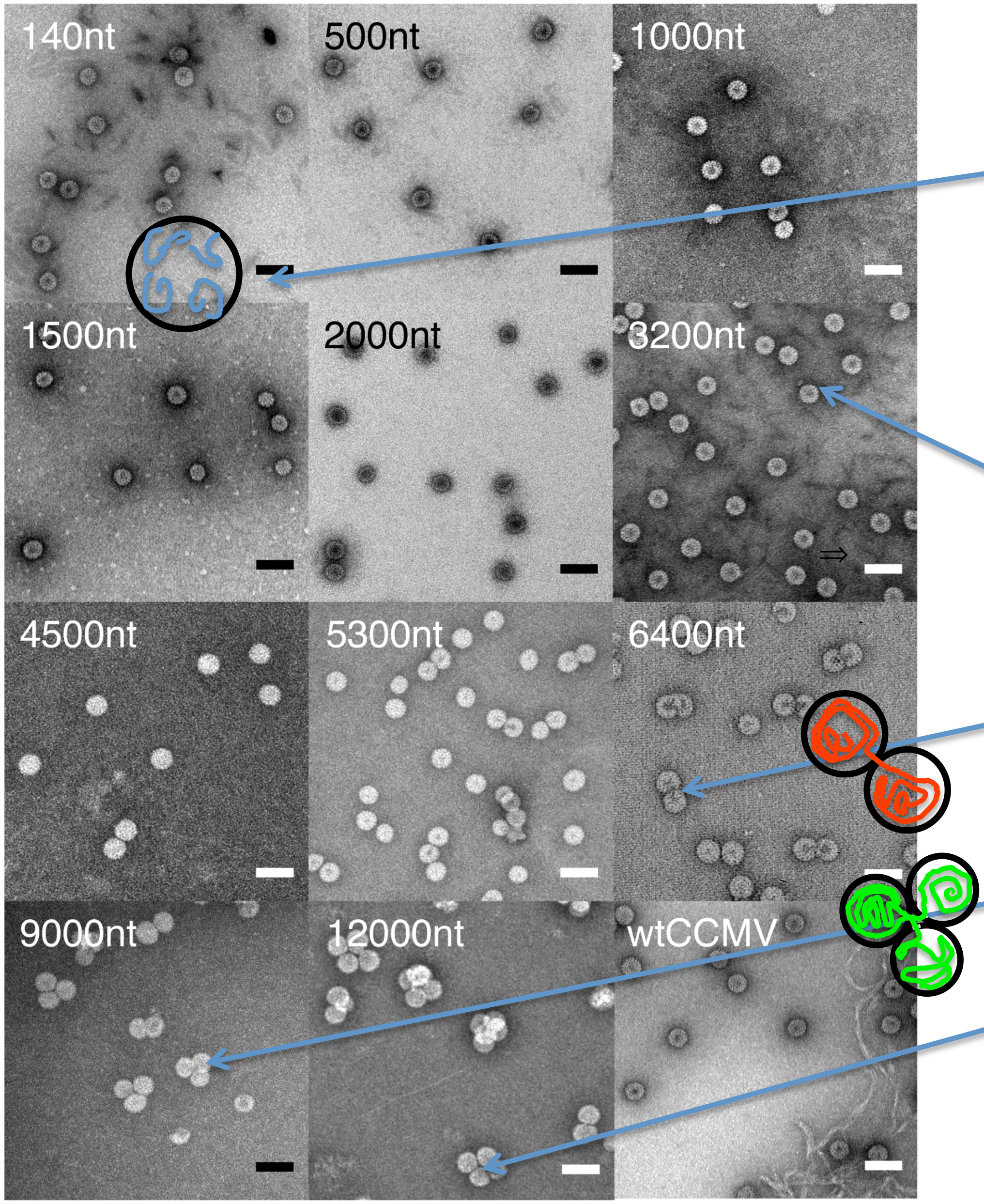
CAN WE MAKE A VIRUS WITH ONLY 2 GENES…?

Work done with

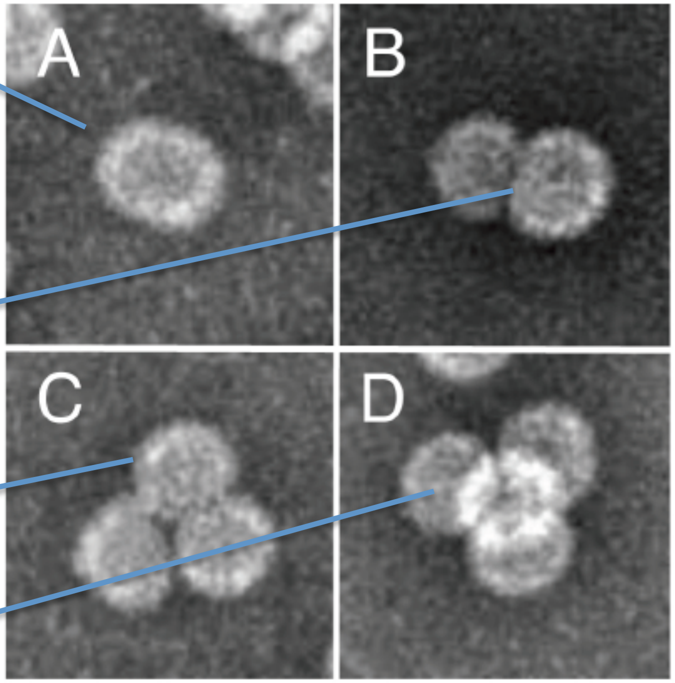
Charles Knobler [UCLA) and Avinoam Ben-Shaul [Jerusalem]

Dr. Ajay Gopal **Walter Singaram**

We (i.e., CCMV protein) can package – *in vitro* – <u>any</u> RNA, of <u>any</u> length:



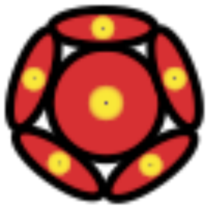
“undersized” RNAs ⟹
many RNAs per capsid
“oversized” RNAs ⟹
many capsids per RNA
(about 3000 bases per capsid)

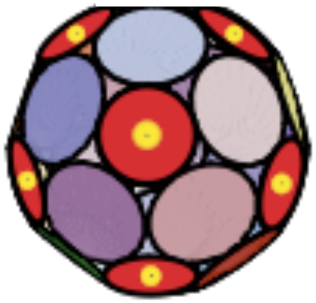


Cadena-Nava, Comas-Garcia, Garmann, Rao, Knobler, Gelbart, *J. Virology* ***86***, 3318 (<u>2012</u>)

Capsid protein subunits form closed, icosahedrally-symmetric, 2D hexagonal lattices: Euler's 12 five-fold defects appear as icosahedrally-positioned pentamers

Minimum-energy **capsid structures correspond to those with the *minimum number (T=1, 3, 4, 7, …) of inequivalent positions* for the 60T protein subunits**


T=1


T=3

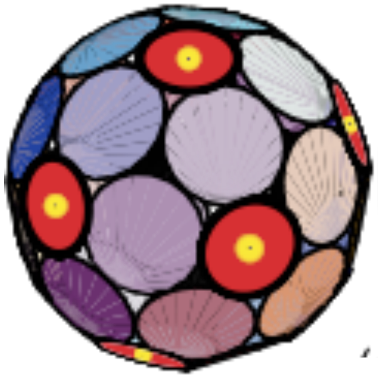
These structures self-assemble, spontaneously, around ssRNA…!

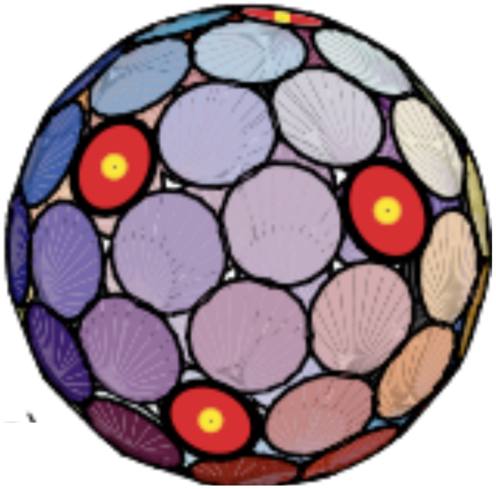
N=12: 12 pentamers,
0=10(T-1) hexamers
60T=60 proteins, *T=1*

N=32: 12 pentamers,
20=10(T-1) hexamers,
60T=180 proteins, *T=3*

N=42: 12 pentamers,
30=10(T-1) hexamers,
60T=240 proteins, *T=4*

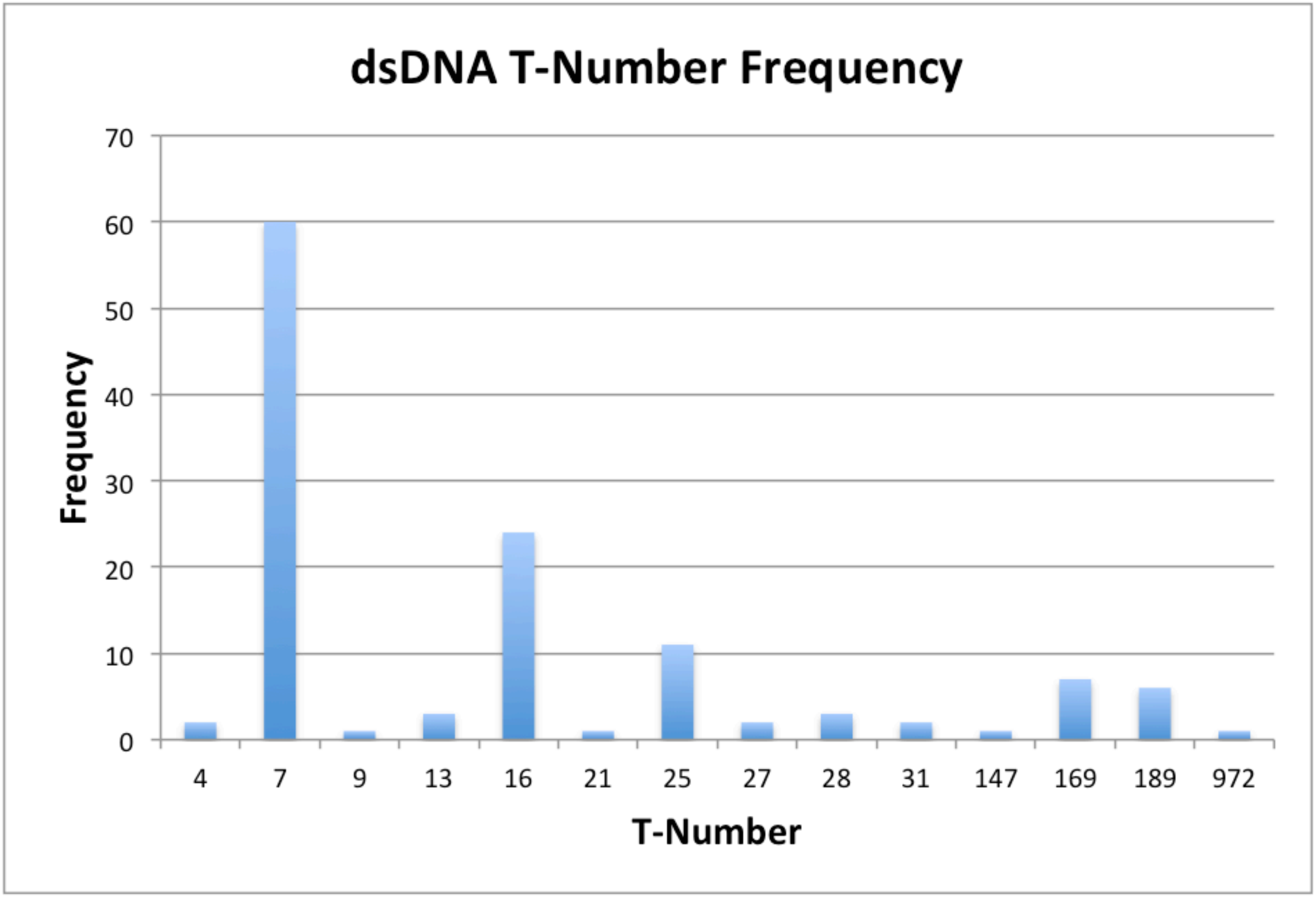
N=72: 12 pentamers,
60=10(T-1) hexamers,
60T=420 proteins, *T=7*


T=4


T=7

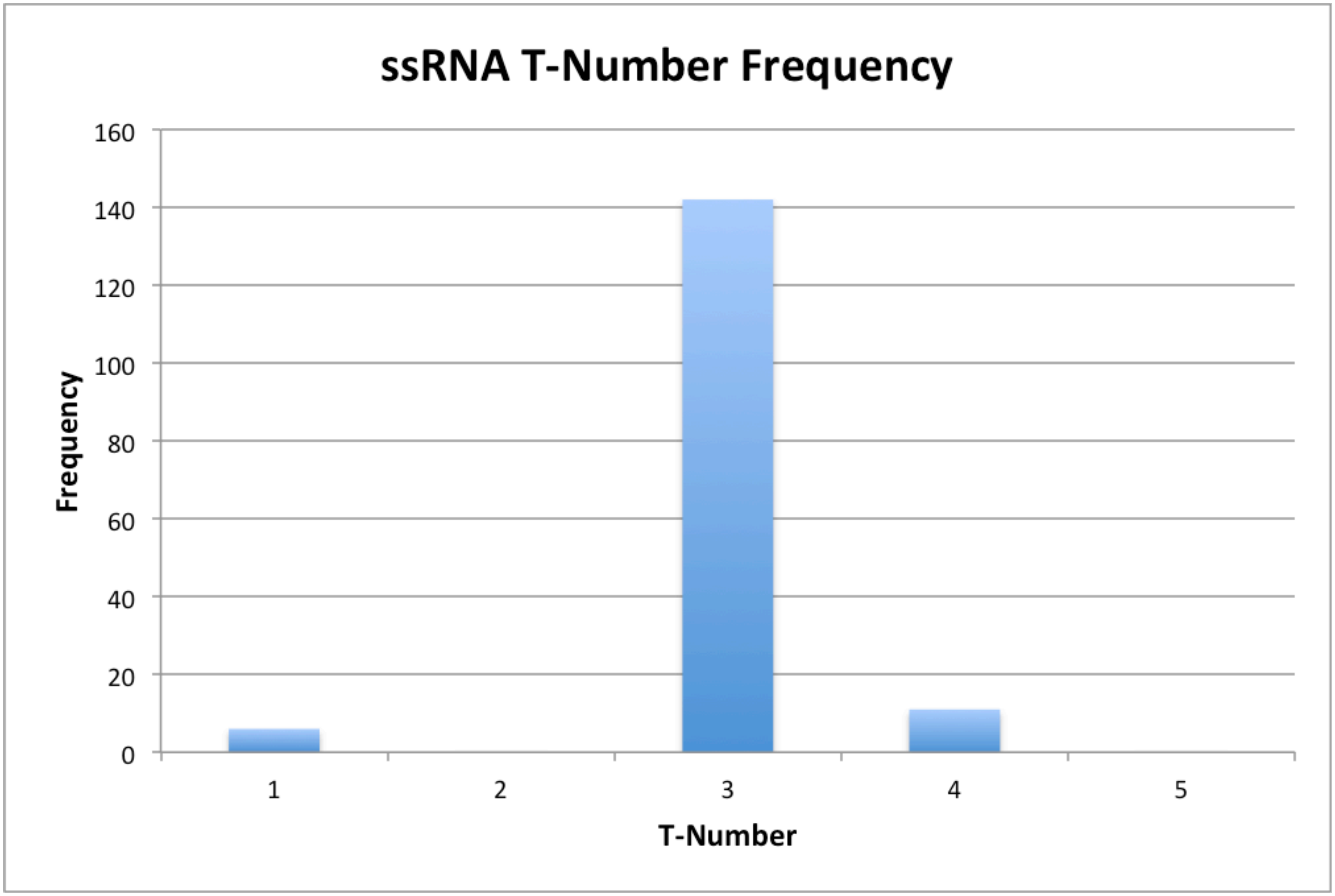
$$T = h^2 + k^2 + hk, \quad h,k = 0,1,2,\ldots$$

Zandi, Reguera, Bruinsma, Gelbart, Rudnick, *PNAS (USA)* 101, 15556 (2004)



$$R_{capsid} \sim T^{1/2}$$

Liya Oster



(Typically, ssRNA viruses have 10-100 times fewer genes than dsDNA viruses!)