11-1-2013

A Bioinformatics Method for Identifying RNA Structures within Human Cells

Stephen Donald Huff

A Bioinformatics Method for Identifying RNA Structures within Human Cells

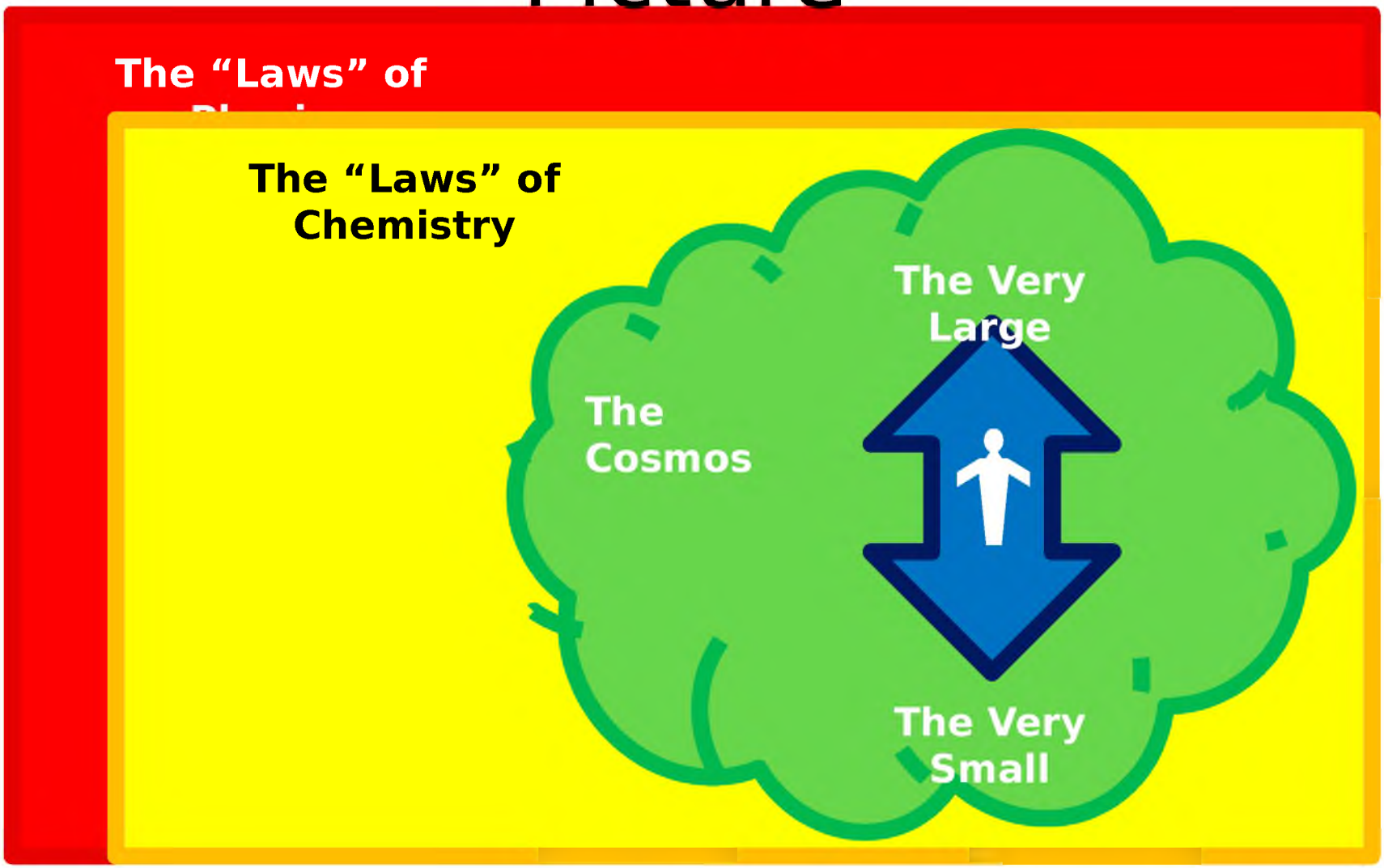
Stephen Huff
Biological Informatics Group
USAFRL – RHDJ
WPAFB

Agenda

- Background
- Overview
- The Problem
- The Solution (?)
- Ideally, Resultant Knowledge

(Bioinformatics = biology + computers)

Background – The BIG Picture



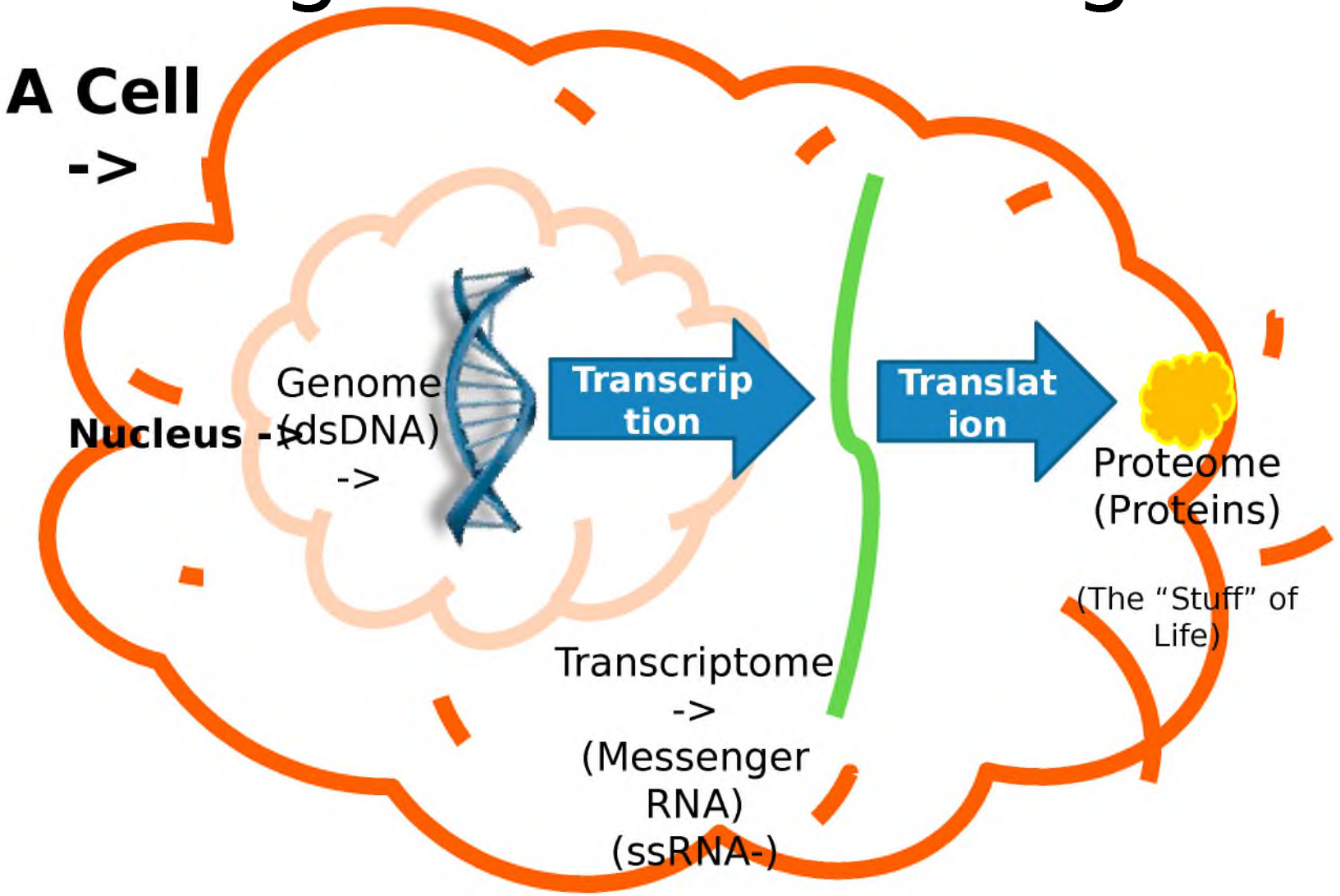
Background – The Human Picture

Ir a es courtesy of MS Office ClipArt, per fair use.

Background – The Dogma

Images courtesy of MS Office ClipArt, per fair use.

Background – The Dogma

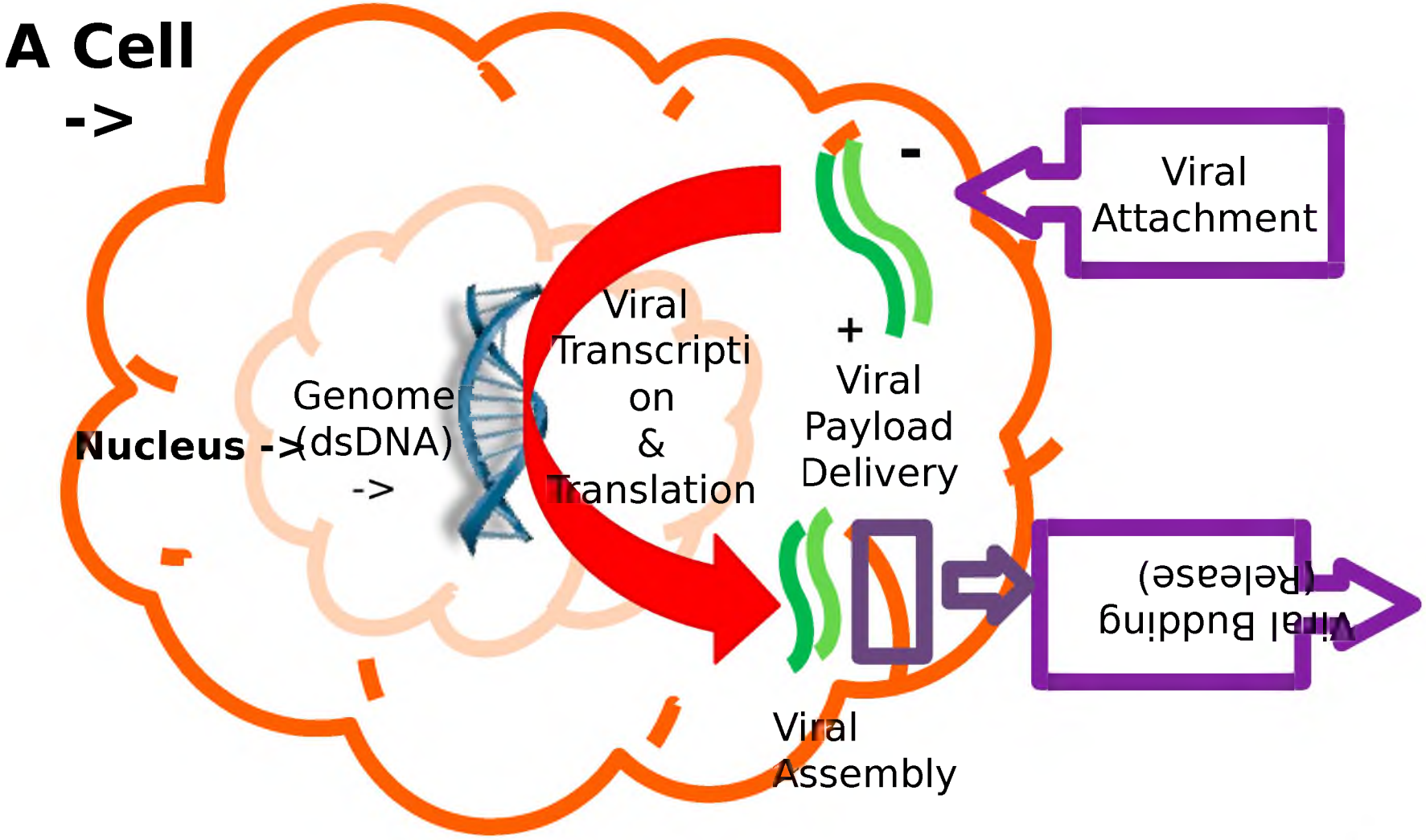


Images courtesy of MS Office ClipArt, per fair use.

Background – Viral Infection

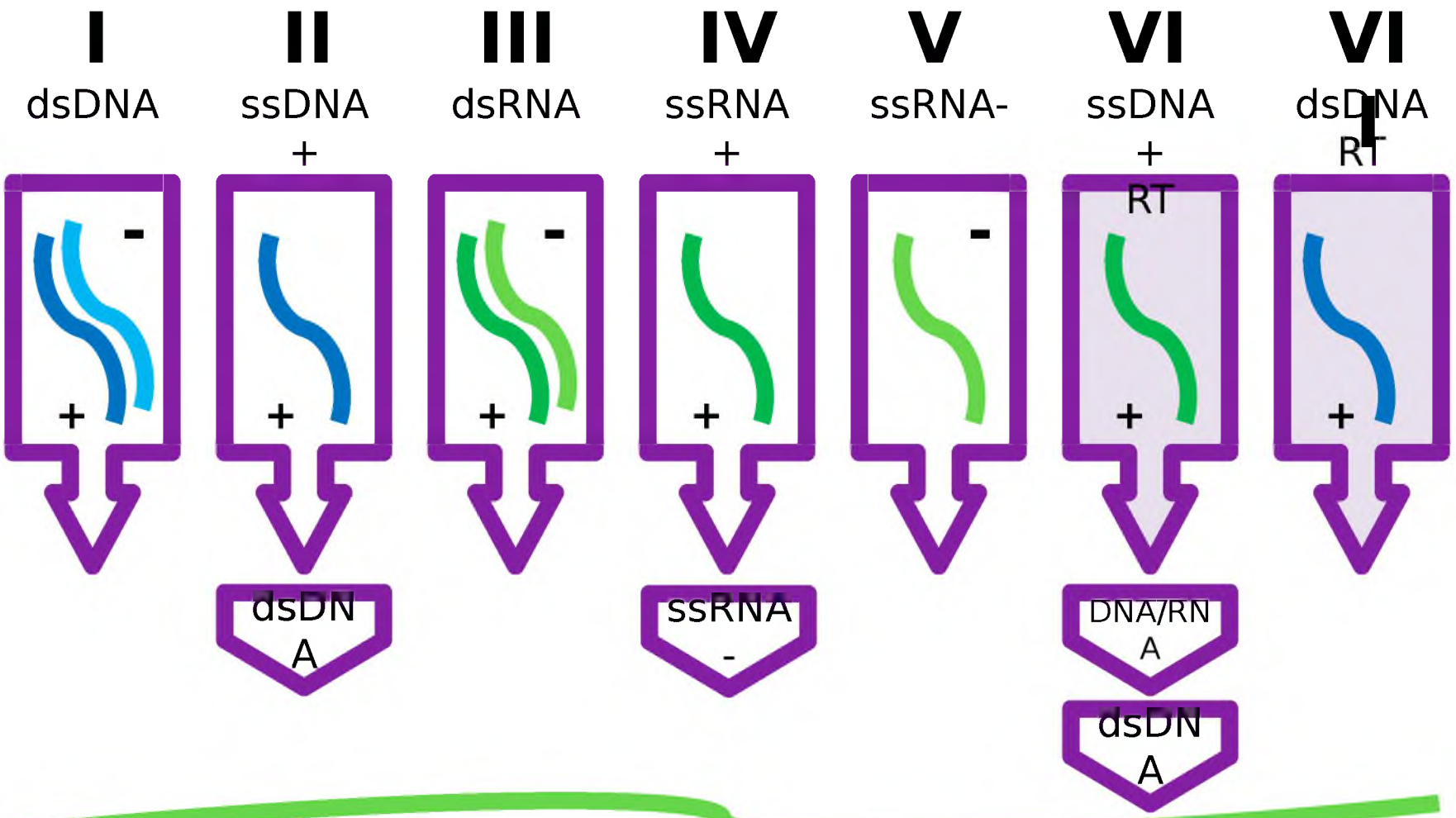
Viral infection can be defined as the process of "hijacking" a host cell to change it into a VIRAL FACTORY (producing mainly viral genomes and proteins NOT host genomes and proteins).

Background – Viral Infection



Remember the Dogma… DNA -> RNA -> Protein (The "Stuff" of Life)

Background - Viruses



ssRNA- Messenger RNA (Translates to Proteins)

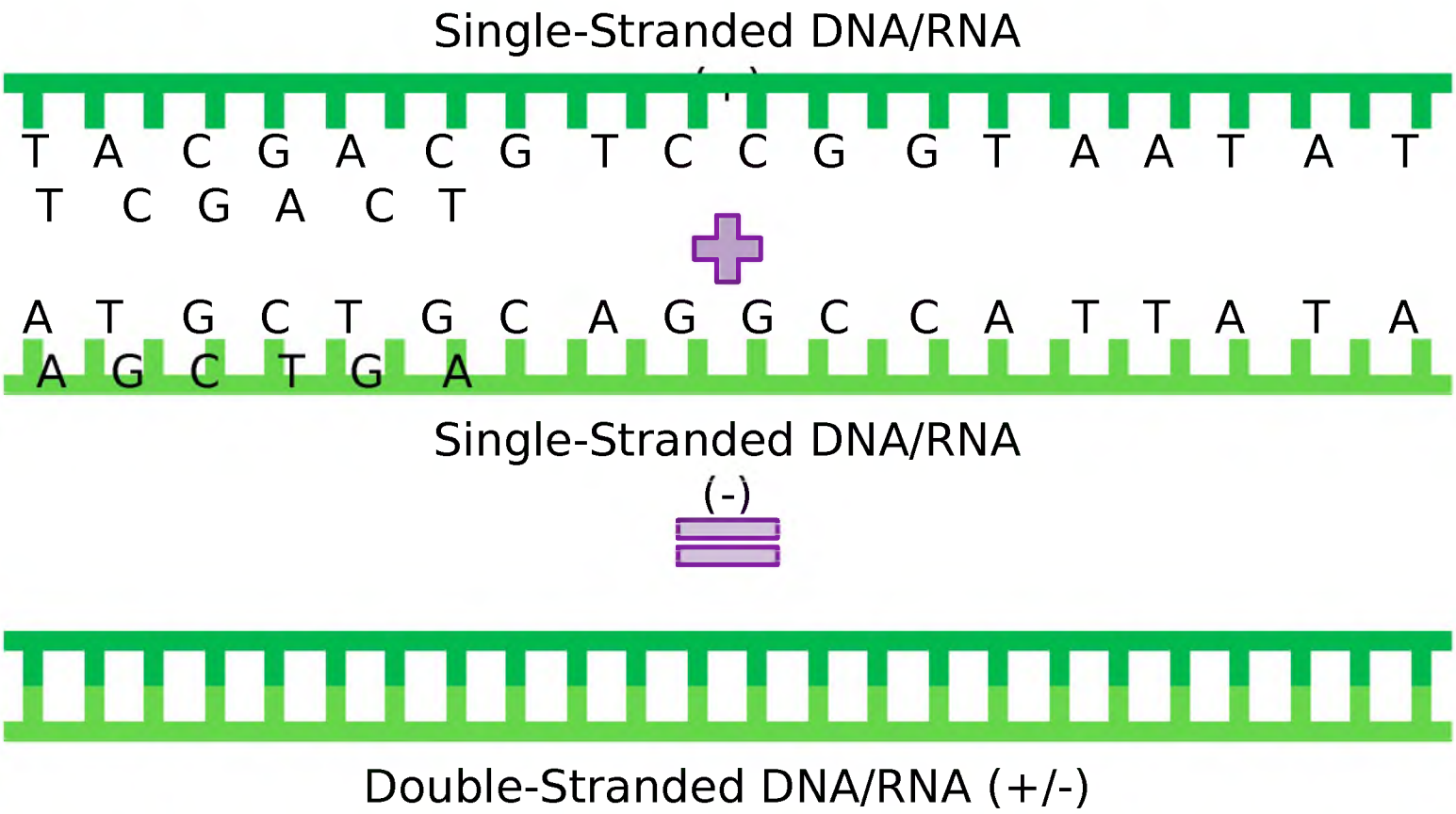
Overview – Virus vs. YOU

- Influenza virus as example vector
  - ssRNA-, replicates in nucleus
  - Genome ~ 14Kb (kilo-bases), segmented (8)
  - Proteome ~ 12
- Human as example host
  - dsDNA, multiple organs, tissues and cell types
  - Genome ~ 3Gb (giga-bases), segmented (22+1)
  - Proteome ~ 21,000 +

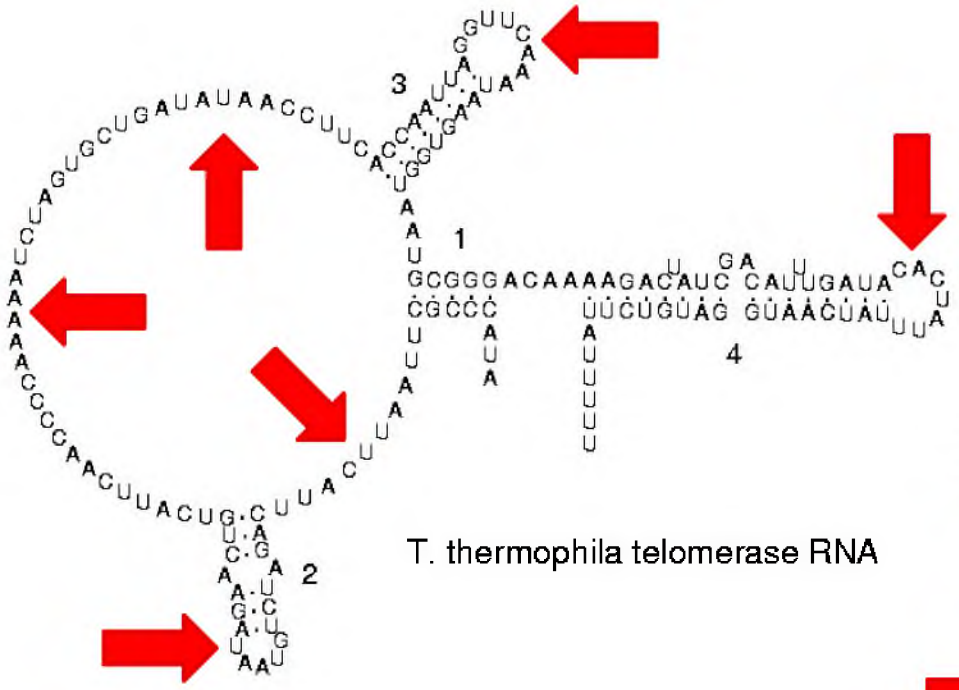
Overview – Curious Question

- SO… how can something so SMALL (virus) so completely overwhelm something so LARGE (host)???
- Traditional view focuses on proteins…
  - 12 proteins overwhelm and undermine 21K+???
- Could other phenomena affect the system?
  - Nucleotide phenomena… specifically, formation of secondary, tertiary (and quartenary?) structures

Overview – Nucleotide Pair Bonding



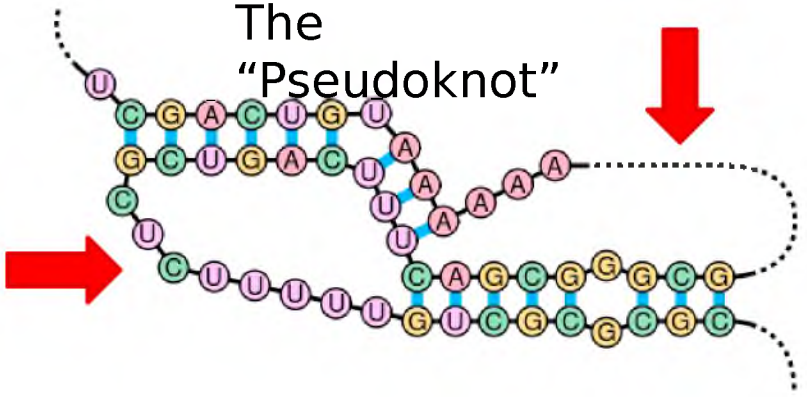
Overview – Nucleotide Self-Bonding



T. thermophila telomerase RNA

Multiple “Hairpins”

The “Pseudoknot”



Single-stranded nucleotides free to bond… with just about ANYthing biological

Images courtesy of WikiPedia, per fair use.

Overview – Interesting Possibilities

- Many examples of nucleotide-structure-based control have recently been discovered
  - Silencing RNA, micro RNA, nucleolar RNA, etc…
  - Exert surprising biological control (start/stop/alter protein synthesis and other functions)
  - Others probably exist (so called “junk” DNA)
- Potential for biological exploitation (therapies) is VAST but unknown

The Problem

- Given the human transcriptome/proteome (RNA -> protein), identify key control structures
  - Compared to what?
  - 21K+ candidates, each with 10K-10M+ nucleotides
  - Permutations may be more numerous than stars in the known COSMOS (given mutational variants)
- Currently, this problem is intractable

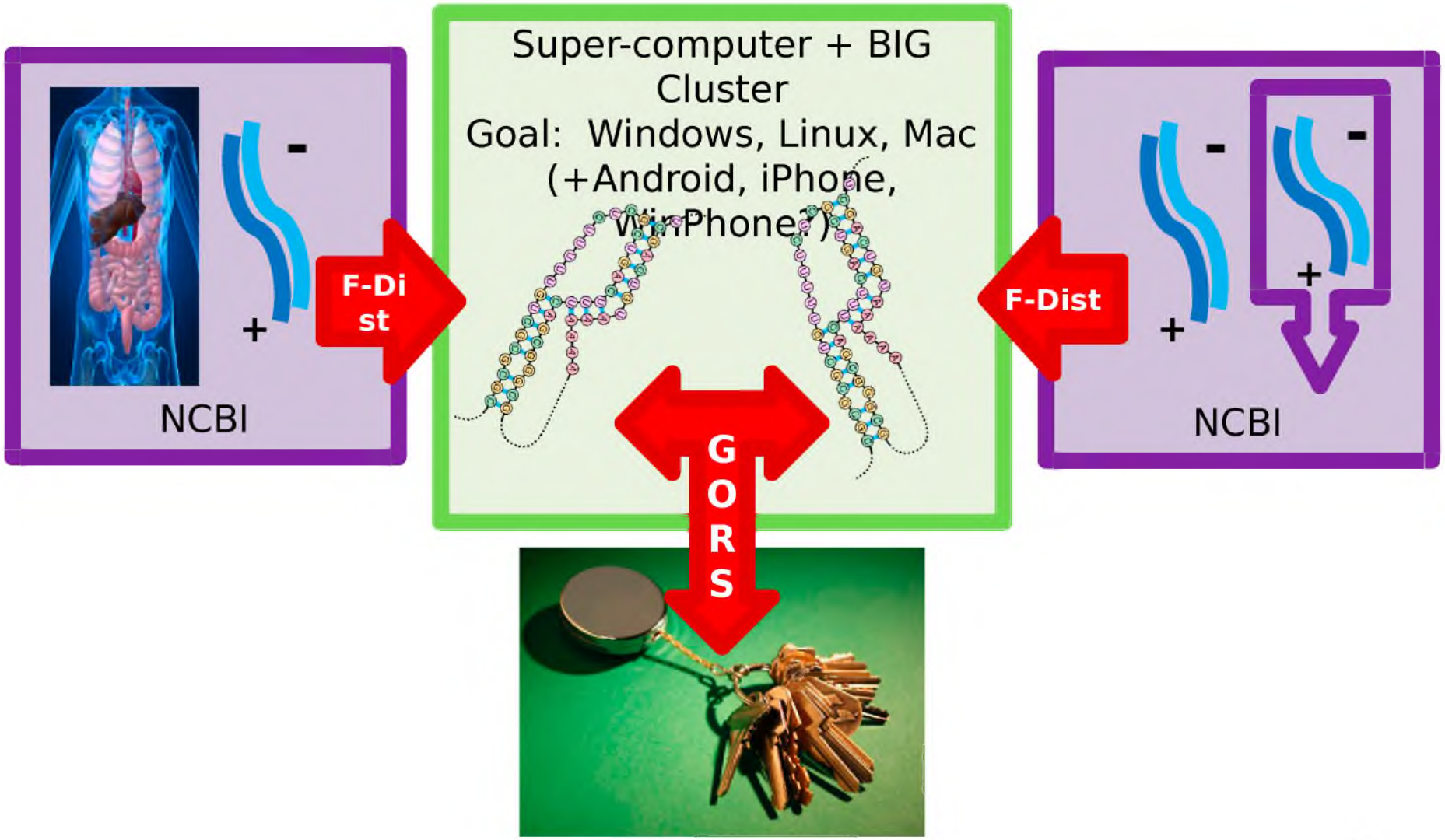
The Solution (?)

- Viruses have been “picking biological locks” for eons
  - More than just proteins – nucleotide structures are probably involved, too
  - They have evolved to become VERY GOOD at this
- Use smaller viral genomes to identify keys
  - Presumably they are RICH in such structures
- Fold viral & host transcripts then

The Solution - Method

1. Use evolutionary distance to choose viral candidates (proprietary software)
2. Fold viral candidates (Rosetta/ViennaRNA)
3. Fold host transcriptome (Same as 2)
4. Use GORS to identify conserved structures within virus (proprietary software)
   - These are likely to be vital to virus-host relation
5. Compare to human
6. Validate in wet-lab via high-throughput screens

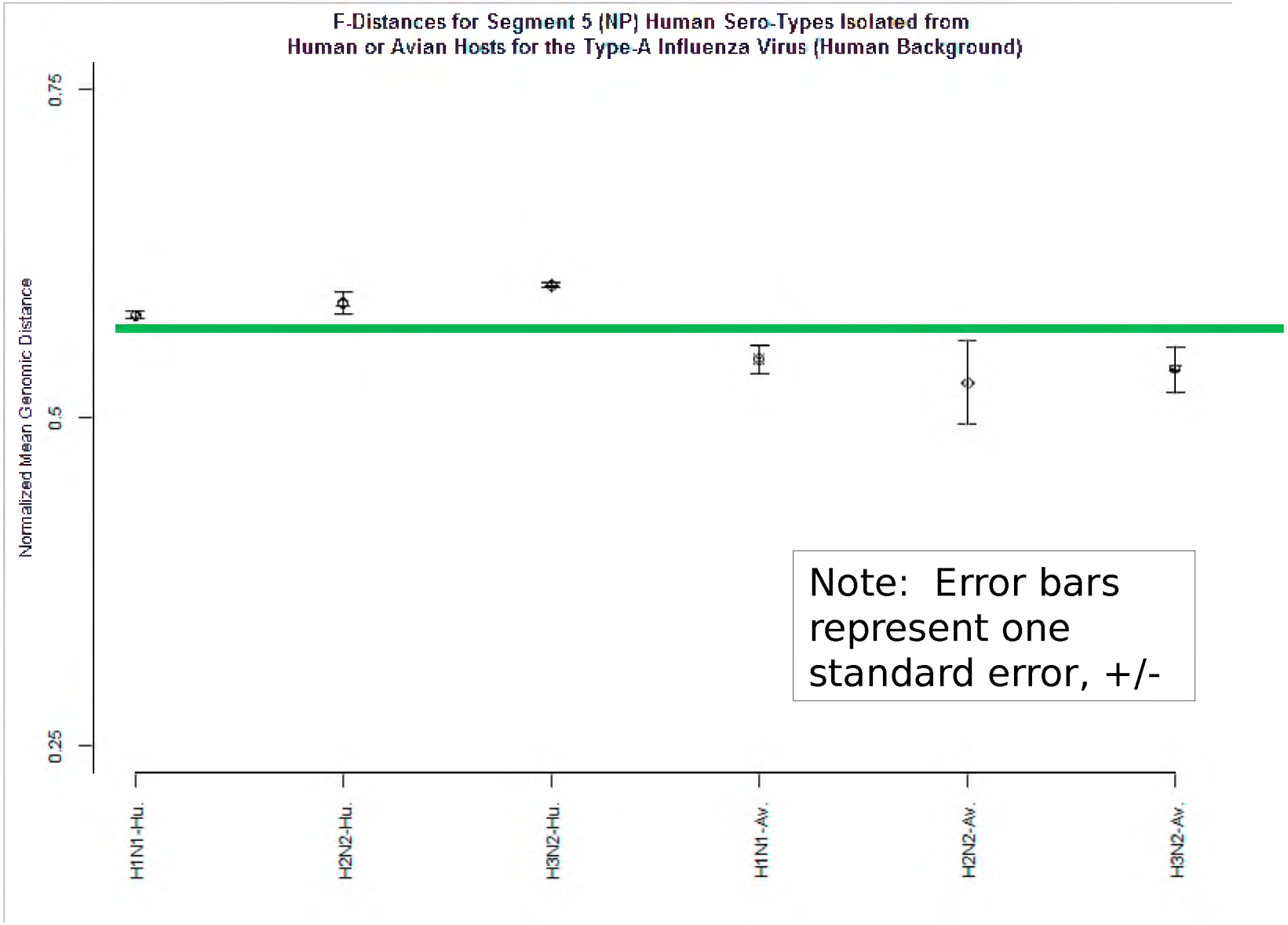
The Solution Graphically

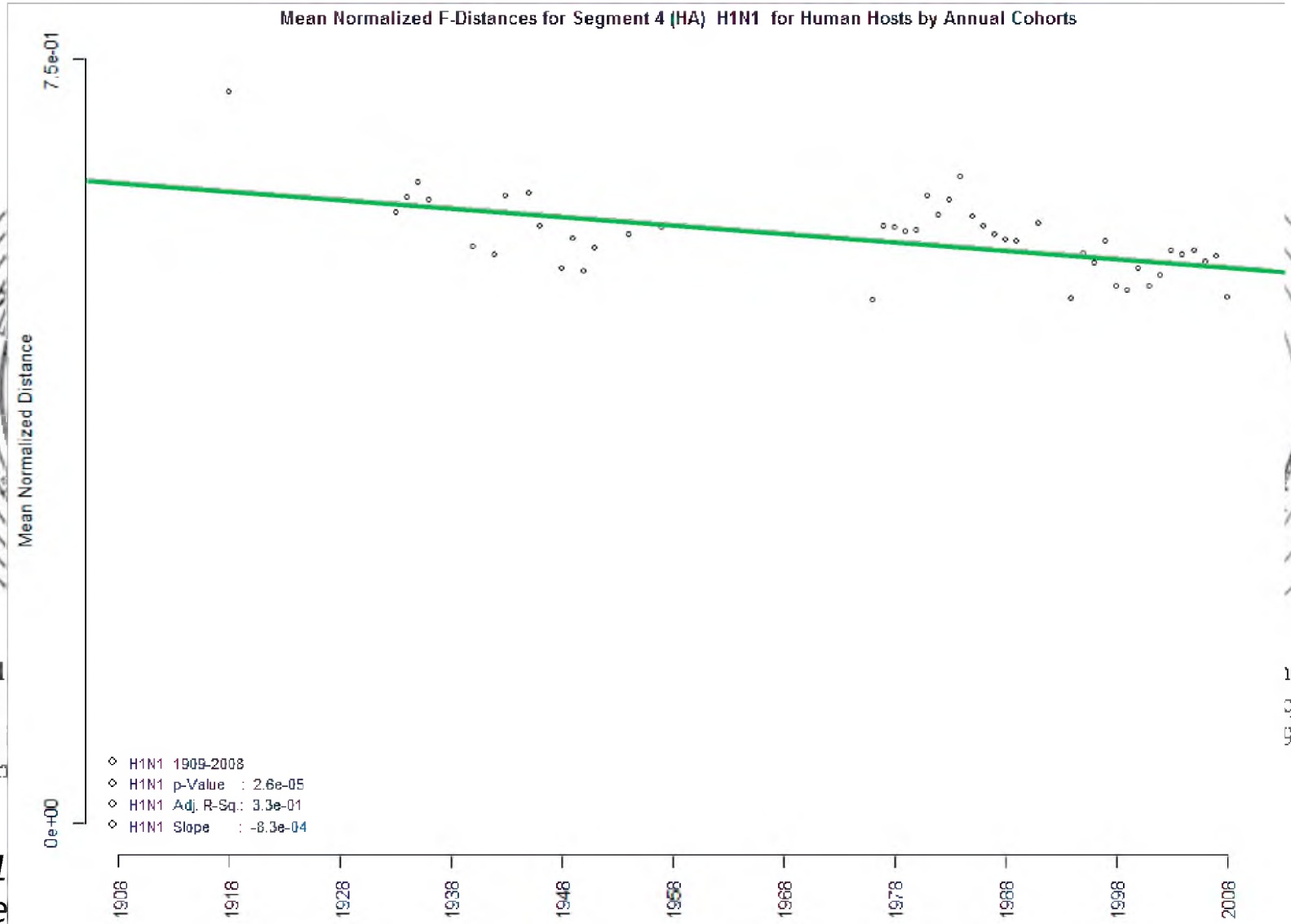


Images courtesy of MS Office ClipArt, per fair use.

Solution – Fofanov-Distance

- Influenza Type-A as example parasite
  - Segmented genome, currently attenuating to human host (SOMEtimes)
    - Some segments attenuated to human, others to avian/swine or other host(s)
    - Temporal changes
  - Use of wrong segments at wrong times = noise
- F-Distance is novel tool for non-heuristic analysis of genomic distance, computationally intense due to exhaustive mutational analysis

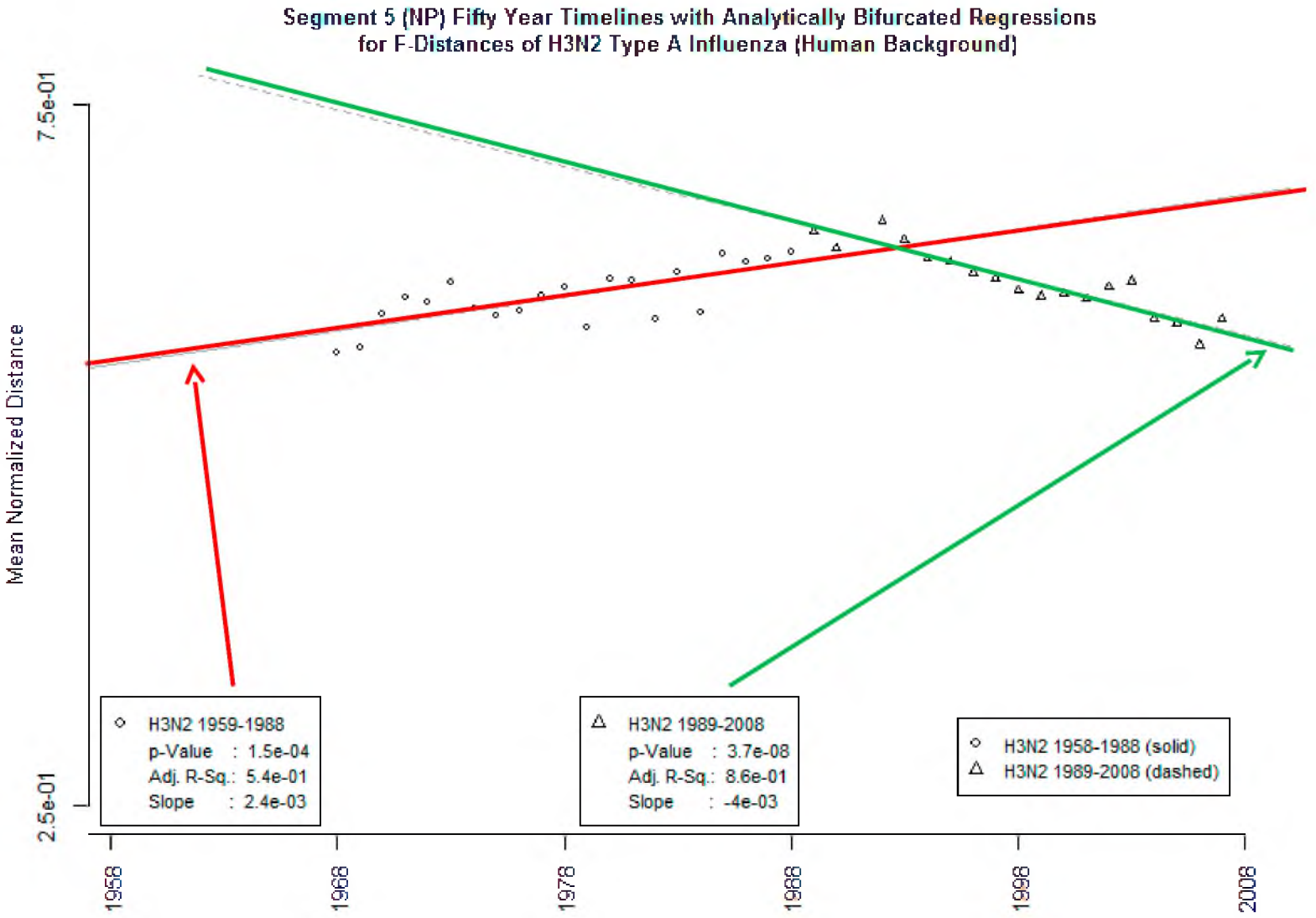

F-Distances for Segment 5 (NP) Human Sero-Types Isolated from
Human or Avian Hosts for the Type-A Influenza Virus (Human Background)
Normalized Mean Genomic Distance
0.25
0.5
0.75
H1N1-Hu.
H2N2-Hu.
H3N2-Hu.
H1N1-Av.
H2N2-Av.
H3N2-Av.
Note: Error bars represent one standard error, +/-



(a)

Figure 1 ... humans, inferred ... quences isolated ... 973 and 2008. In ...

2011 ... dime ... antigenically variable RNA viruses.", Proceedings of the Royal Society B: Biological Sciences. Online May


Segment 5 (NP) Fifty Year Timelines with Analytically Bifurcated Regressions
for F-Distances of H3N2 Type A Influenza (Human Background)
Mean Normalized Distance
7.5e-01
2.5e-01
1958
1968
1978
1988
1998
2008
H3N2 1959-1988
p-Value : 1.5e-04
Adj. R-Sq.: 5.4e-01
Slope : 2.4e-03
H3N2 1989-2008
p-Value : 3.7e-08
Adj. R-Sq.: 8.6e-01
Slope : -4e-03
H3N2 1958-1988 (solid)
H3N2 1989-2008 (dashed)

Solution – RNA Folding

- ViennaRNA and Rosetta are best candidates
  - Open source software for Windows and Linux
- Accepts RNA sequences as input, uses lab-validated thermodynamics and chemistry
- Folds primary structure (sequence) into secondary and tertiary structures

Solution - GORS

- Primarily, need to identify changing patterns of structures over time (within parasite)
  - Conservation indicates need (survival pressure)
  - This indicates something KEY to structures in host
- GORS is statistical method for such analysis
- May have to “invent” new statistical methods or adapt existing methods

Solution - Products

- F-Distance
  - R-extension (dll), overlaid by custom GUI (Windows, Mac, smart-devices)
  - Publication
- ViennaRNA/Rosetta
  - Custom computational “pipeline”
- GORS
  - R-extension (dll), GUI, “pipeline”, publication
- Final results – publication and software suite

Ideally - Resultant Knowledge

- Improved grasp of virus-host relationships
  - Influenza vaccine, new generation?
- Toxin mitigation (RHDJ focus)
  - Structures bond to metallic ions, small molecules, proteins, other nucleotides, etc…
- Exploit discovered structures for therapies
- Investigate biology from a new perspective

Questions

Background Array IO

Sequence	Binary Version	Index	Present
AAAAAAAAAAAAAAAA	00000000000000000000000000000000	0	0
AAAAAAAAAAAAAAAT	00000000000000000000000000000001	1	0
AAAAAAAAAAAAAAAG	00000000000000000000000000000010	2	0
AAAAAAAAAAAAAAAC	00000000000000000000000000000011	3	0
AAAAAAAAAAAAAATA	00000000000000000000000000000100	4	1
AAAAAAAAAAAAAATT	00000000000000000000000000000101	5	0
AAAAAAAAAAAAAATG	00000000000000000000000000000110	6	0
AAAAAAAAAAAAAATC	00000000000000000000000000000111	7	0
AAAAAAAAAAAAAAGA	00000000000000000000000000001000	8	1
AAAAAAAAAAAAAAGT	00000000000000000000000000001001	9	0
AAAAAAAAAAAAAAGG	00000000000000000000000000001010	10	0
AAAAAAAAAAAAAAGC	00000000000000000000000000001011	11	0
.....	. . . . .	…	…
TCCCCCCCCCCCCCCC	01111111111111111111111111111111	4n - 3	0
GCCCCCCCCCCCCCCC	10111111111111111111111111111111	4n - 2	0
CCCCCCCCCCCCCCCC	11111111111111111111111111111111	4n - 1	0

Foreground Array IO