VIV (Virtual Virus)
A computational simulation of natural virus evolution

Donald S. Burke, M.D.
Professor, Departments of International Health
and Epidemiology
Johns Hopkins Bloomberg School of Public Health

10 June 2002
DIMACS

"Indeed, the microbial biosphere can be thought of as a World Wide Web of informational exchange … What makes microbial evolution particularly intriguing, and worrisome, is a combination of vast populations and intense fluctuations in those populations."

Joshua Lederberg
Nobel Laureate
Science, 14 Apr 00

OUTLINE

I. PRIMER ON BIOLOGICAL VIRUSES

II. MOLECULAR EPIDEMIOLOGY AND EVOLUTION OF KEY BIOLOGICAL VIRUSES (HIV, INFLUENZA)

III. SIMULATION OF HIV EVOLUTION

IV. SPECULATION "METAGENETICS" AND EVOLVABLE CODE ENTITIES

V. CONCLUSIONS

Viruses are the smallest "living" entities:

Code string = $10^{3.5}$ to $10^{5.5}$ nucleotides (bits)

Obligate intracellular pathogens: only grow inside cell

Thousands of distinct, named viruses

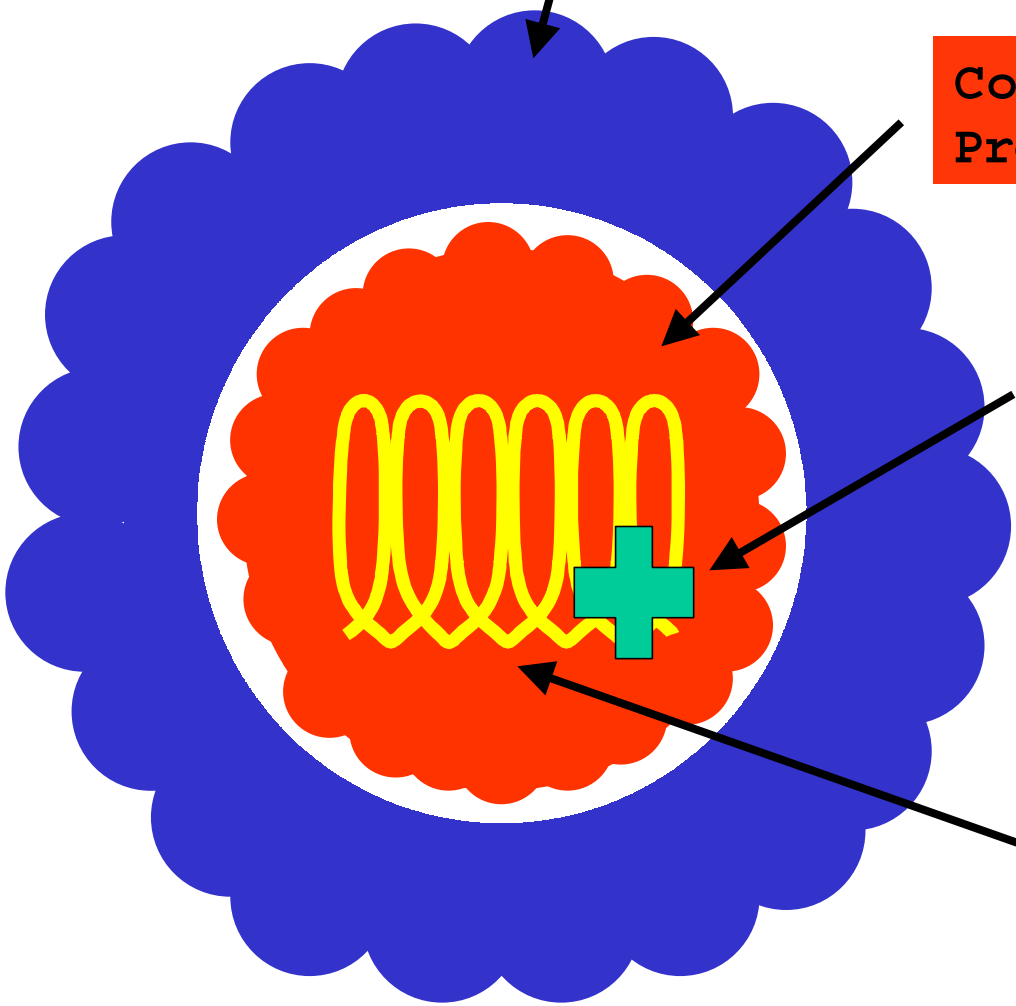
Dozens of "families"

Evolved for billions of years in parallel with cells and multicelluar organisms

Optimized for replication

BASIC VIRUS STRUCTURE

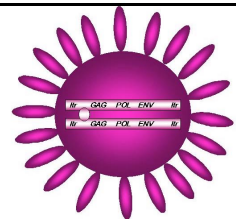
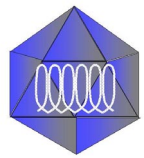
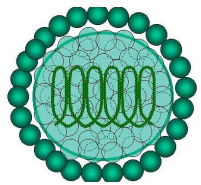
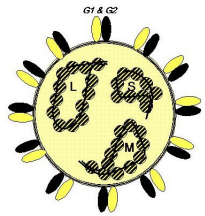
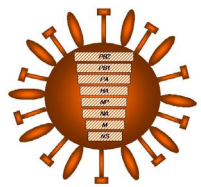
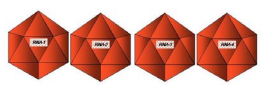
Surface / Envelope
Targets & attaches to cells = address for genome delivery



Core
Protects & masks genome

Polymerase =
Self replicating function.
Replicates virus genome

Virus Genome =
Code String =
DNA or RNA.
Encodes Polymerase, Envelope, and Core

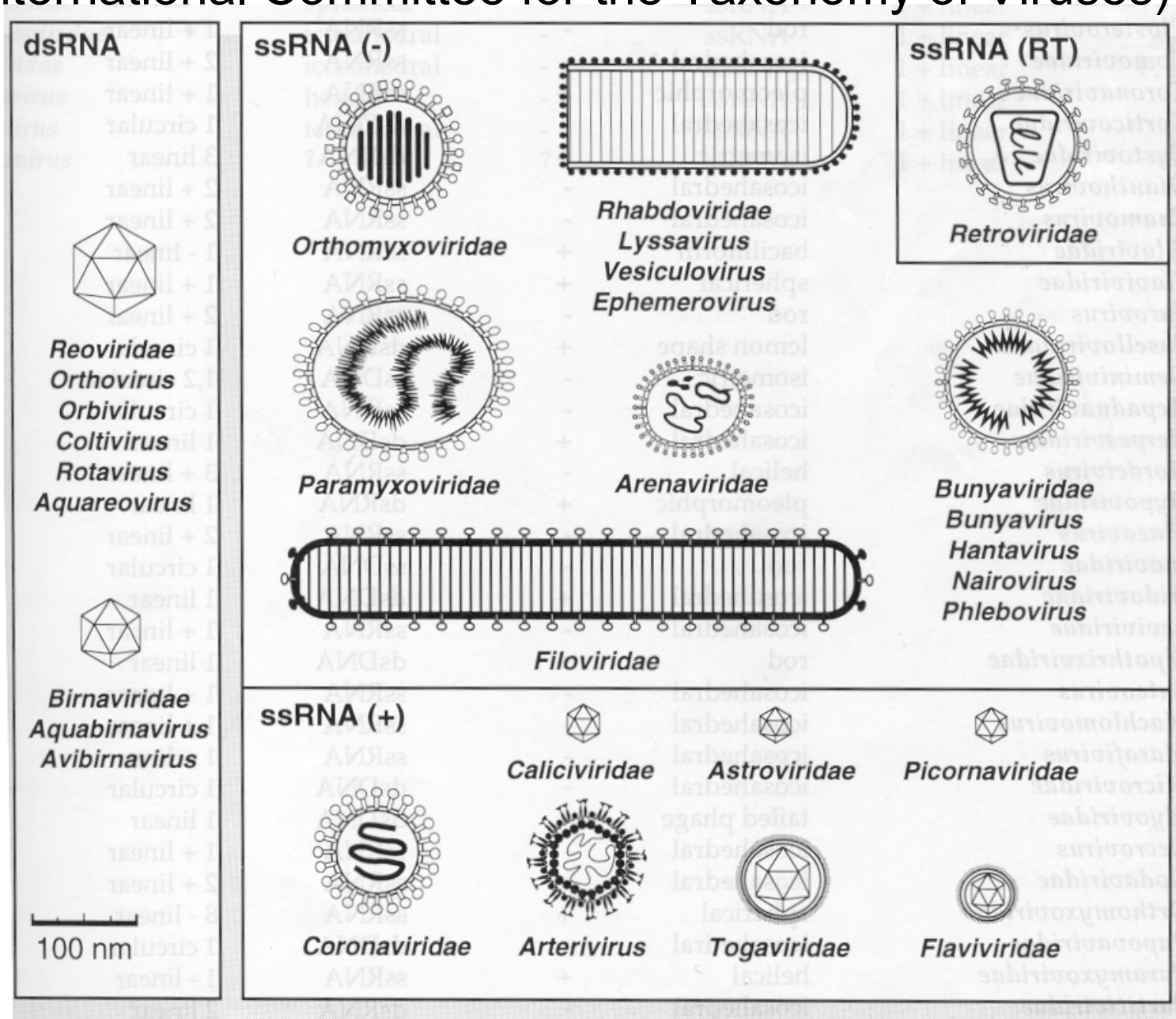
FAMILY	VIRION MORPHOLOGY
RETRO (HIV)	
PICORNA (Polio)	
CORONA (Corona)	
BUNYA (Hanta)	
ORTHOMYXO (Influenza)	
BROMO (plant viruses)	

Code string variations used by viruses

Leu -Arg -Val -Leu -Gly -> Etc

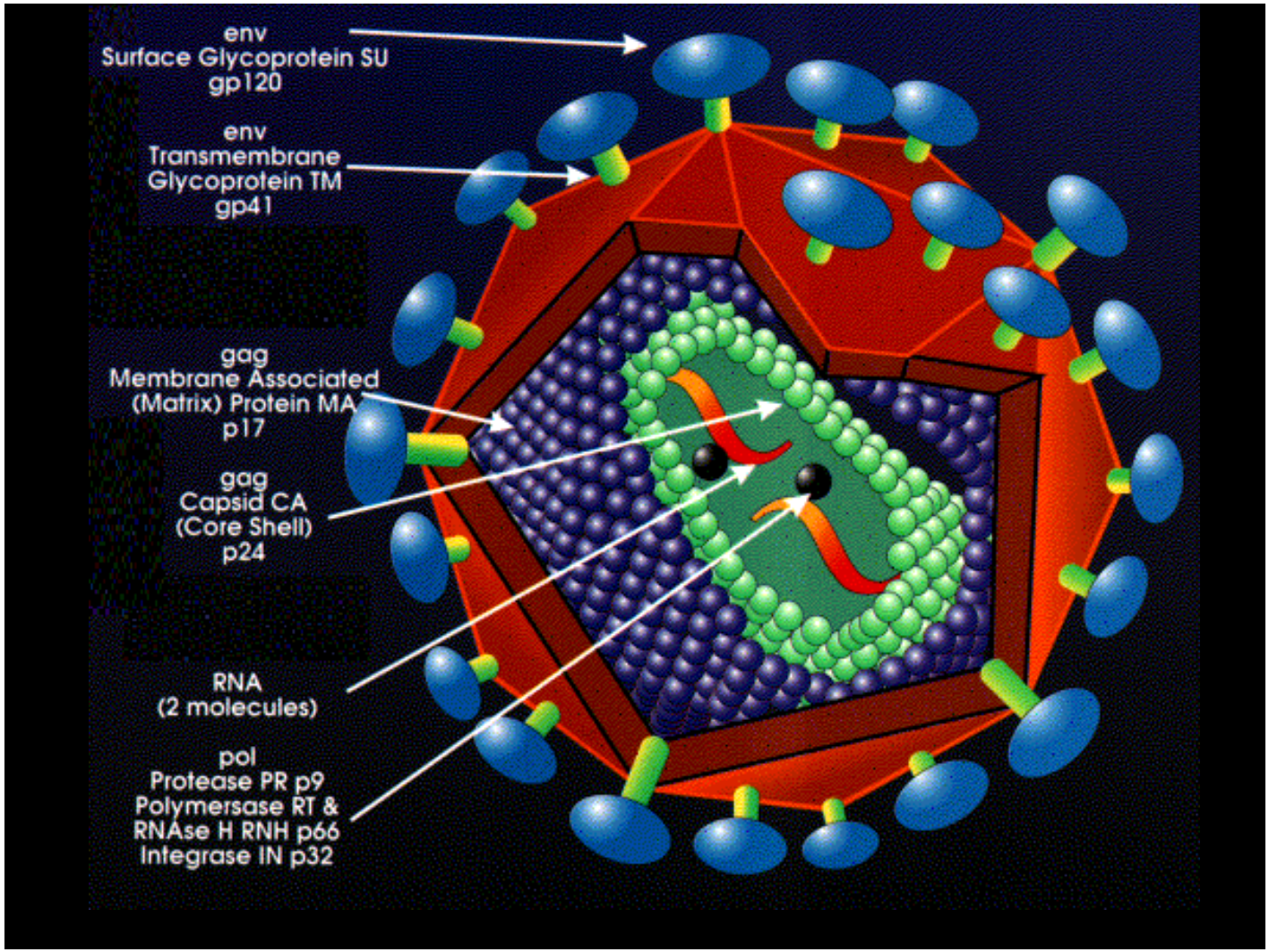
DNA SS	CTA CGA GTC CTC GGT…
DNA DS	CTA CGA GTC CTC GGT…
+ and -	GAT GCT CAG GAG CCA …
RNA+	GAU GCU CAG GAG CCA…
RNA DS	GAU GCU CAG GAG CCA…
+ and -	CUA CGA GUC CUC GGU…
RNA -	CUA CGA GUC CUC GGU…
RNA+	GAU GCU CAG GAG CCA…
(2 copies)	GAU GCU CAG GAG CCA…

RNA Virus Families
(International Committee for the Taxonomy of Viruses)

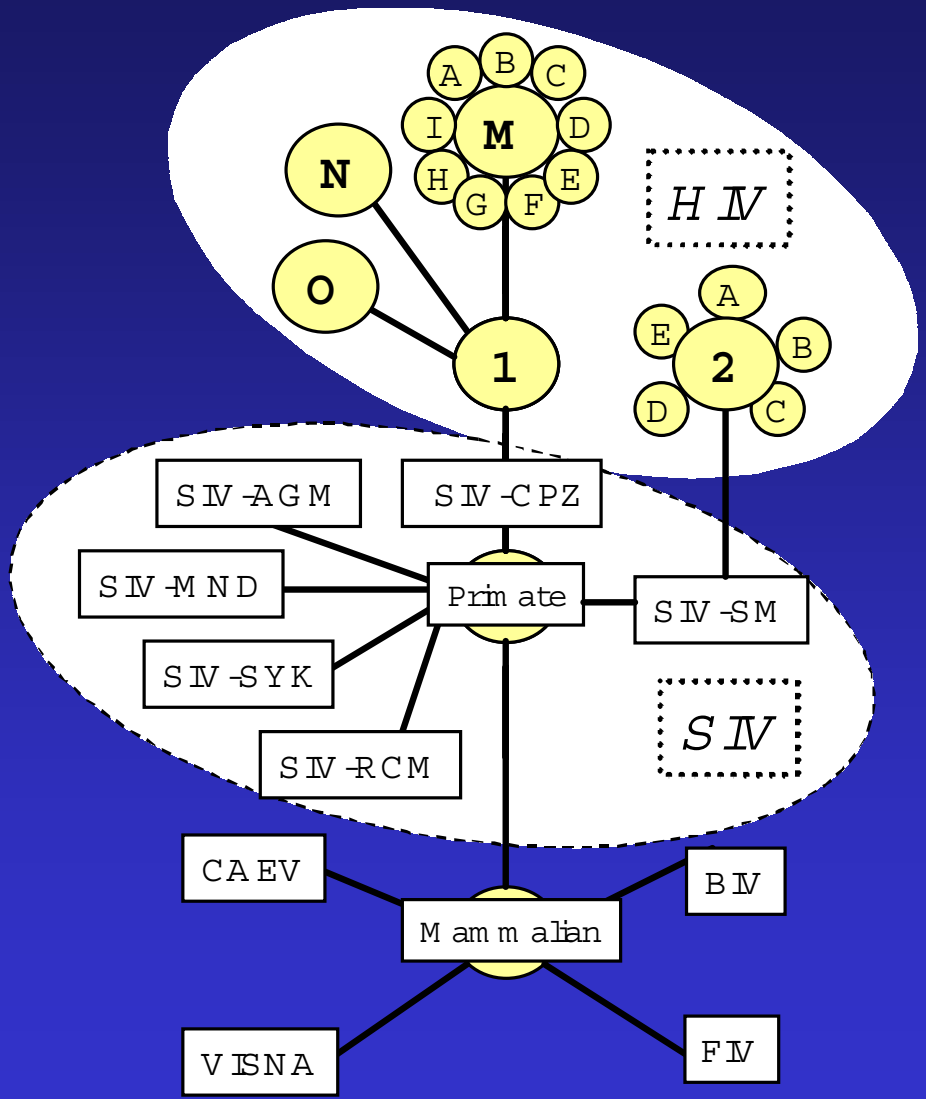


MOLECULAR EPIDEMIOLOGY OF HIV

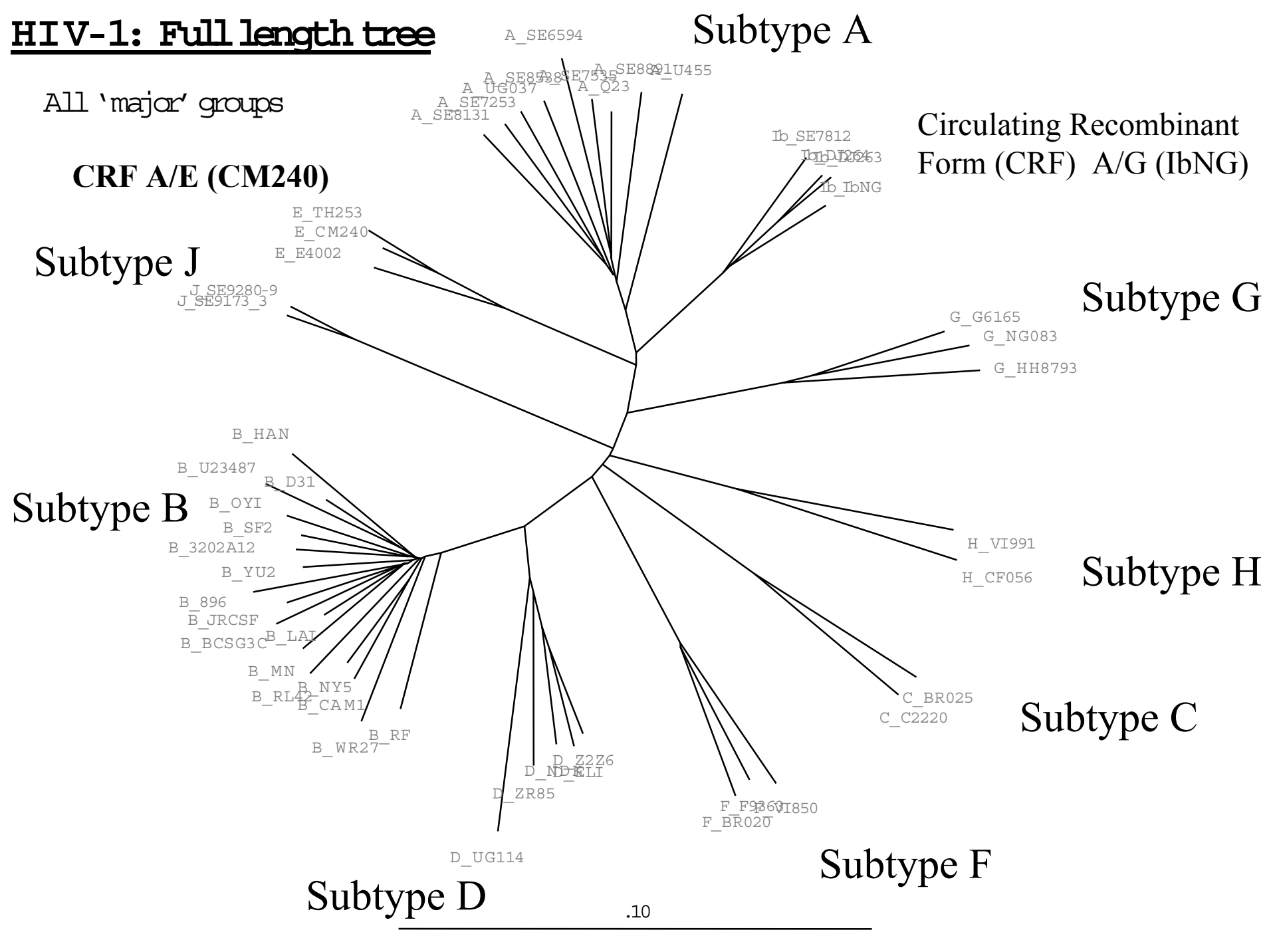
HIV

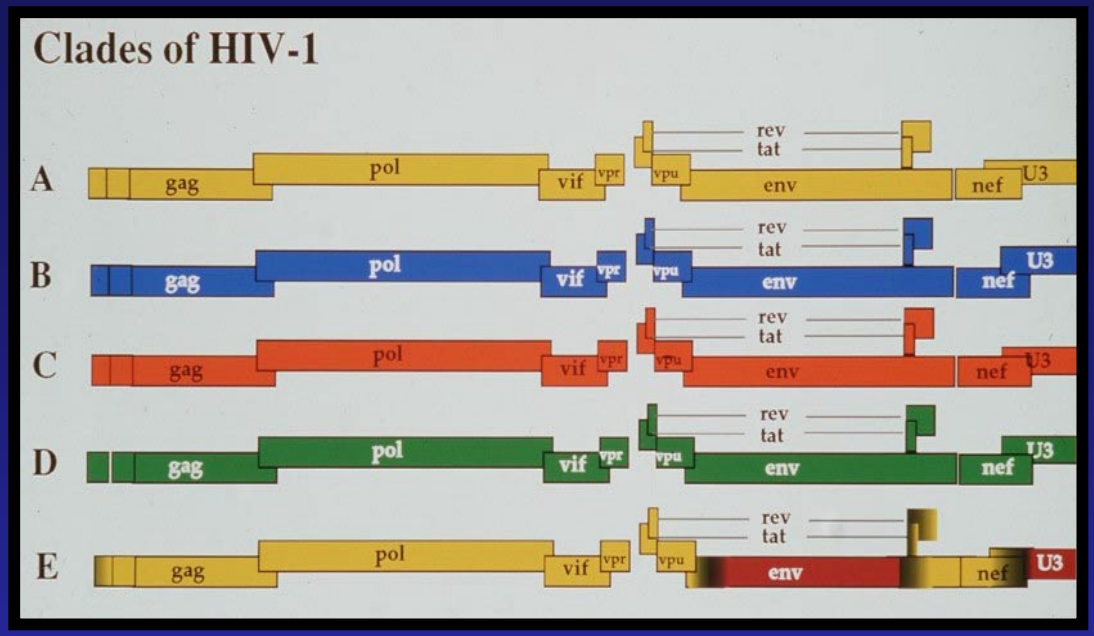

env
Surface Glycoprotein SU
gp120
env
Transmembrane
Glycoprotein TM
gp41
gag
Membrane Associated
(Matrix) Protein MA
p17
gag
Capsid CA
(Core Shell)
p24
RNA
(2 molecules)
pol
Protease PR p9
Polymersase RT &
RNAse H RNH p66
Integrase IN p32

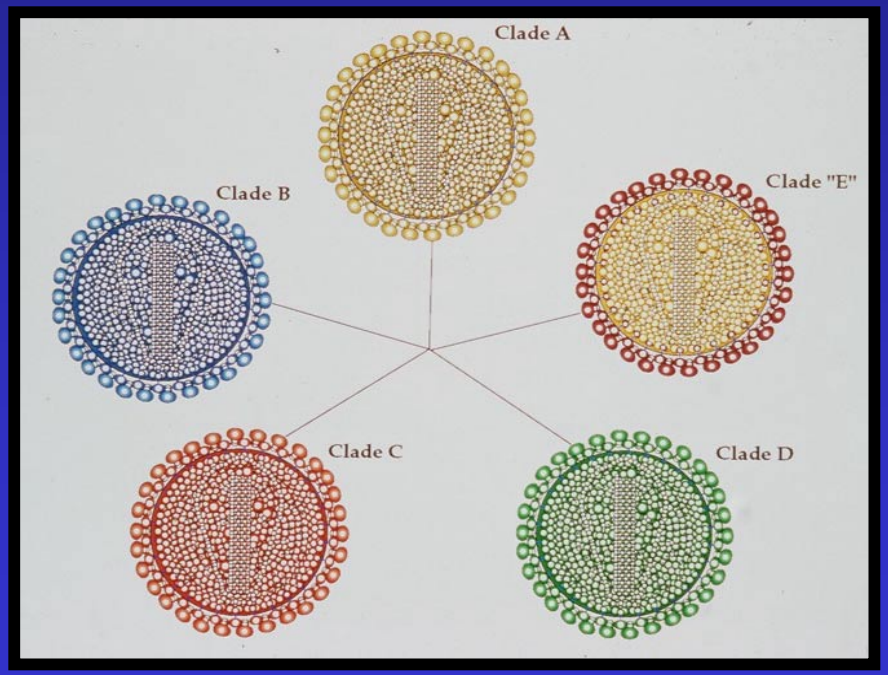
PHYLOGENETIC RELATIONSHIPS OF "HIV", "SIV", AND OTHER LENTIVIRUSE



HIV-1: Full length tree

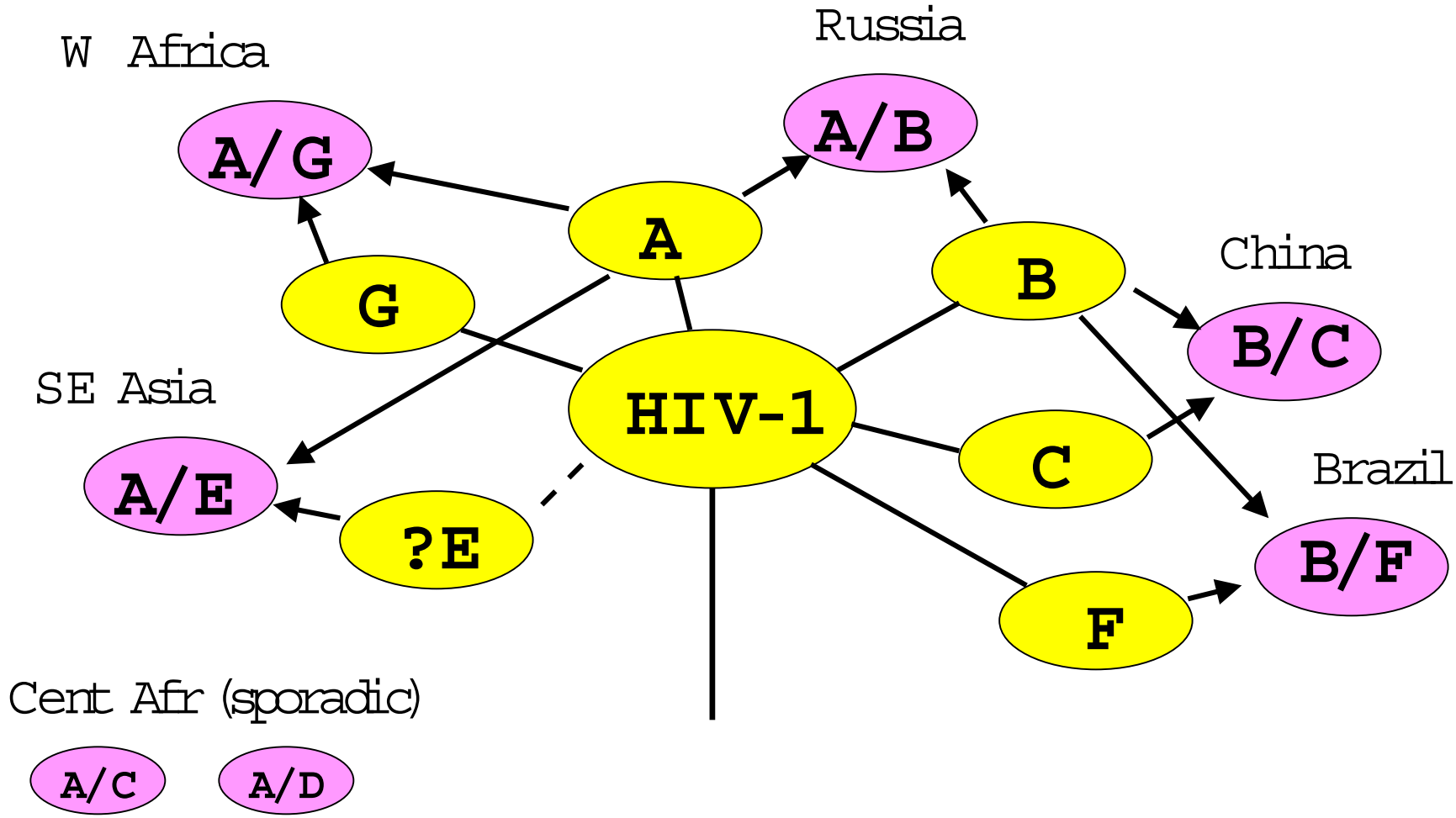


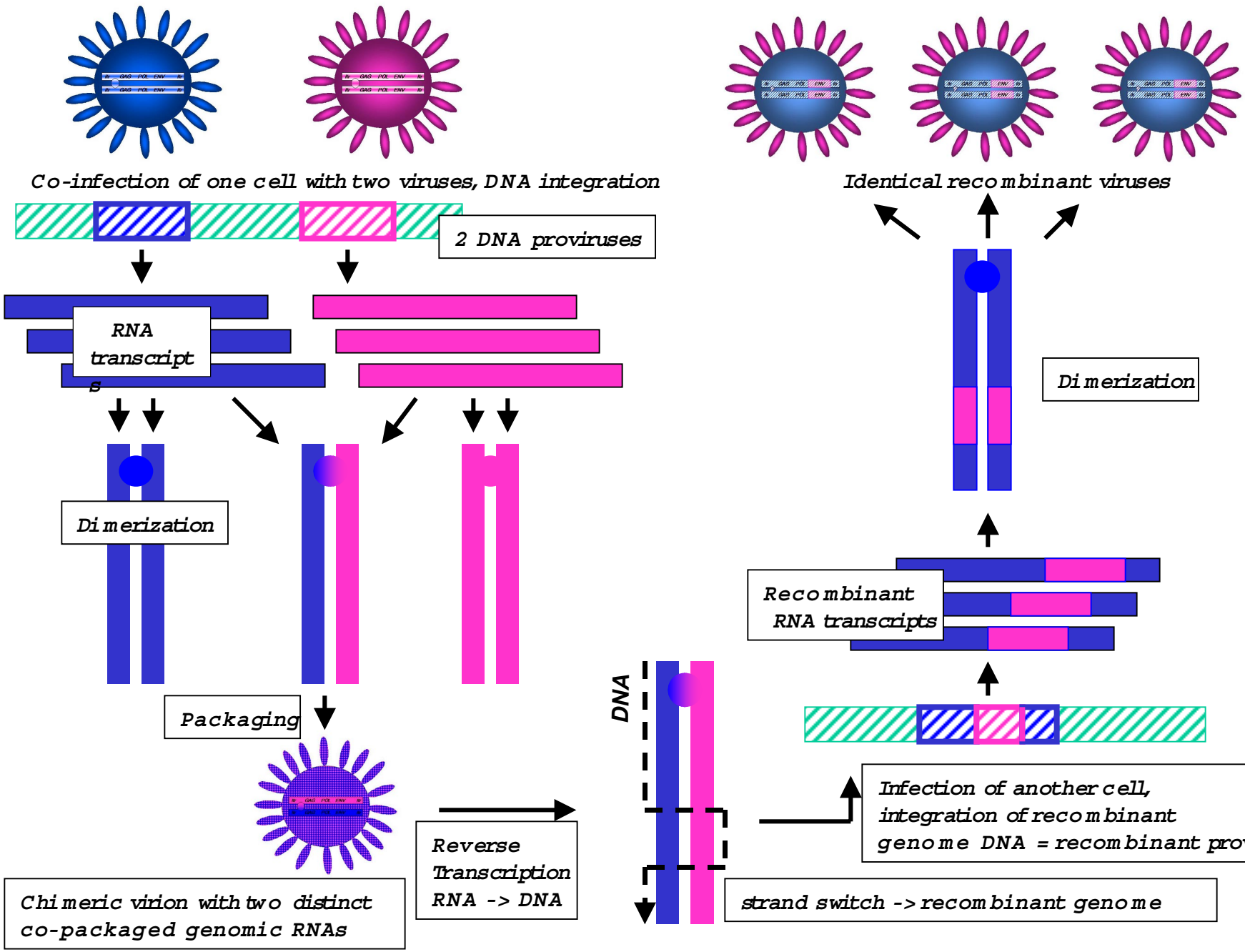


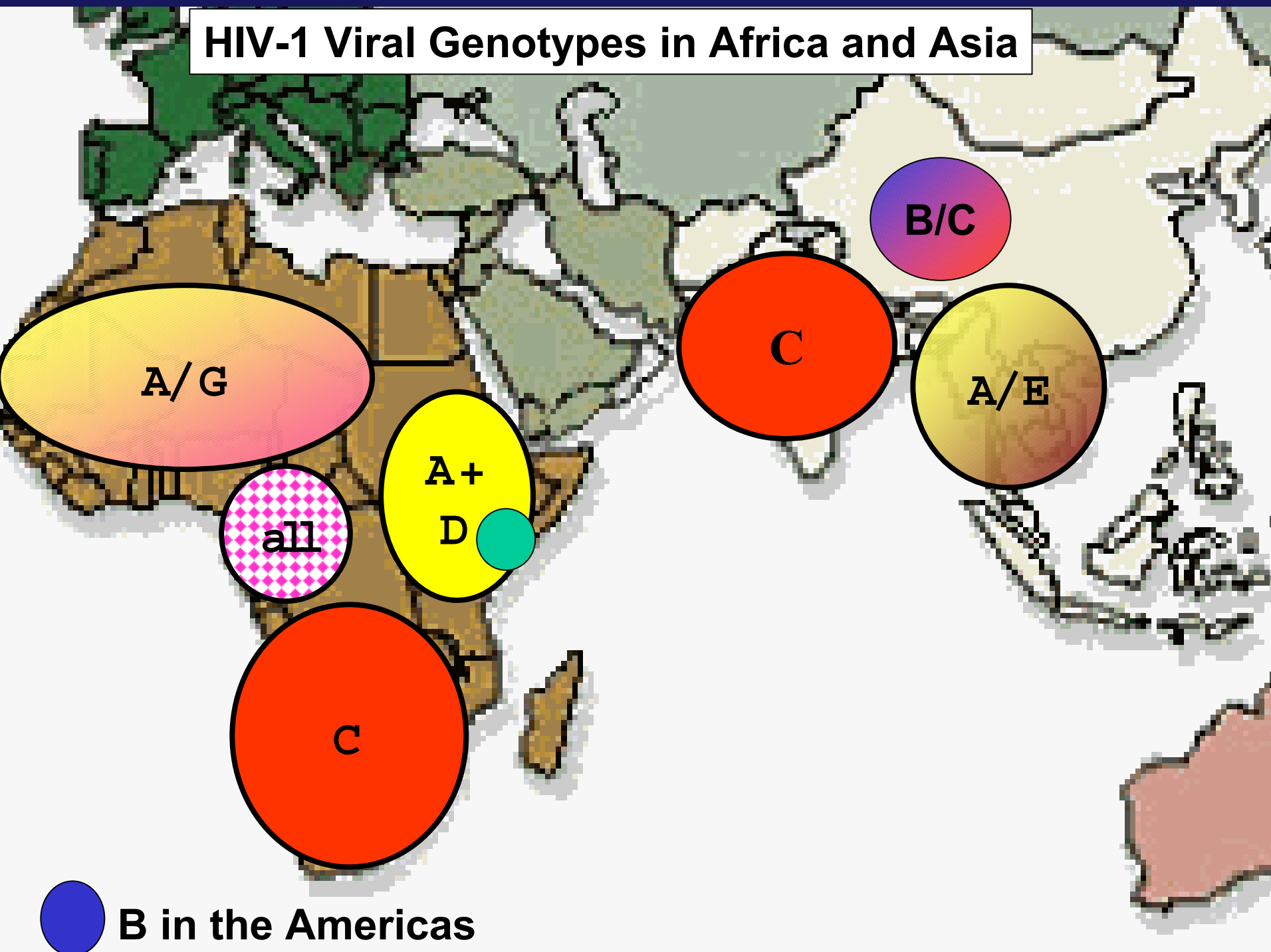


Jean Carr, Francine McCutcha

Recombinant HIV-1 Strains of Major Epidemiological Importance




Co-infection of one cell with two viruses, DNA integration
2 DNA proviruses
RNA transcripts
Dimerization
Packaging
Chimeric virion with two distinct co-packaged genomic RNAs
Reverse Transcription RNA -> DNA
DNA
strand switch -> recombinant genome
Infection of another cell, integration of recombinant genome DNA = recombinant provirus
Recombinant RNA transcripts
Dimerization
Identical recombinant viruses

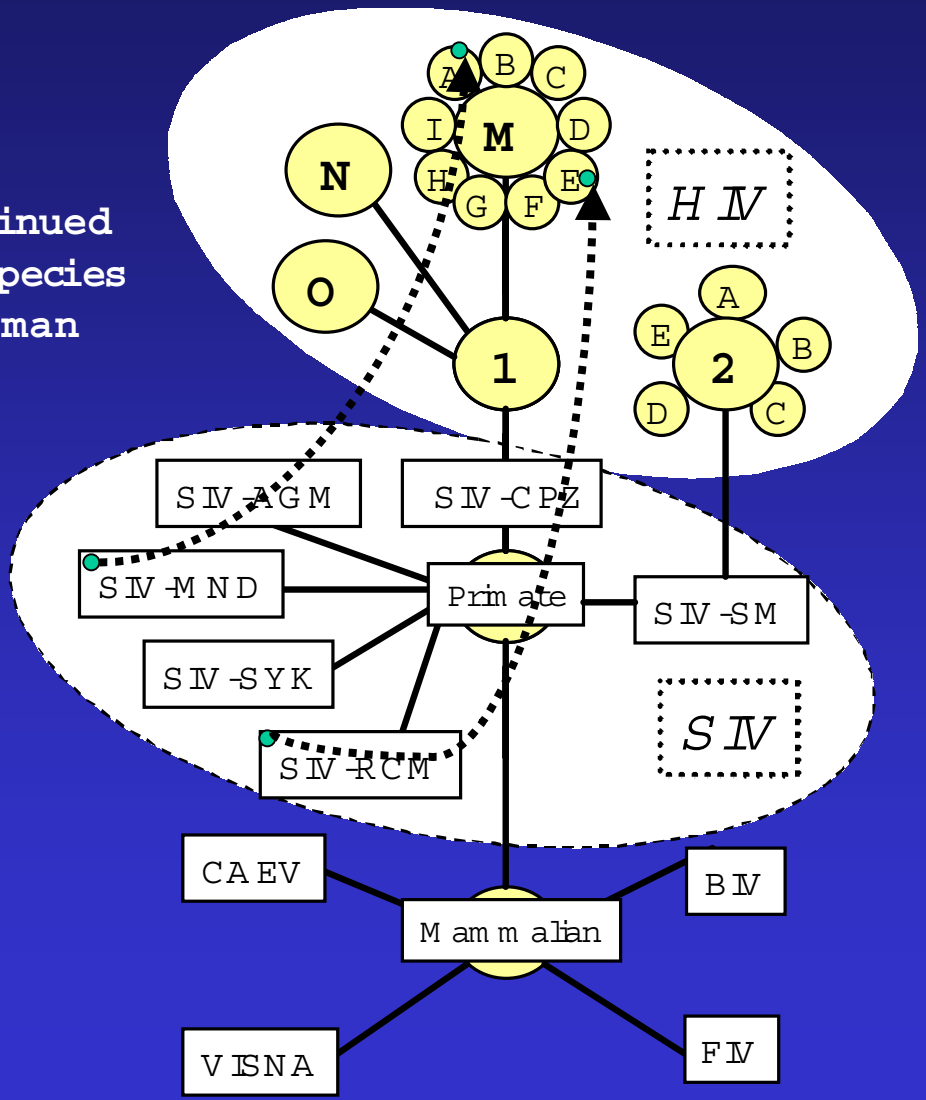

HIV-1 Viral Genotypes in Africa and Asia
B/C
C
A/E
A/G
A+
D
all
C
B in the Americas


NIGERIA
CAMEROON
CENTRAL
AFRICAN REPUBLIC
EQUATORIAL
GUINEA
AFRICAN RAIN FOREST
CONGO
GABON
ZAIRE
UGANDA
ATLANTIC
OCEAN

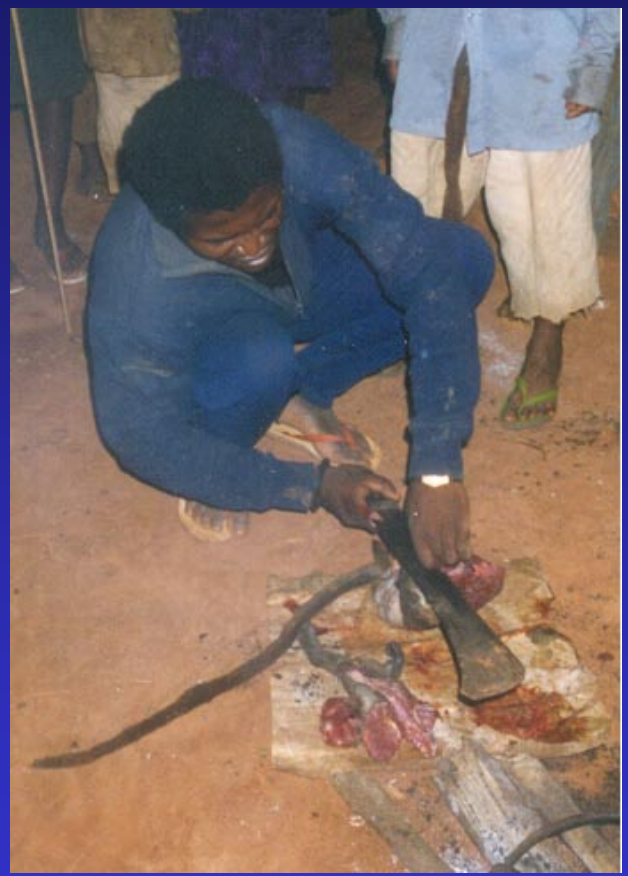
With research team on road to Ngoila, Cameroon

Do all the human HIV sequences derive only from Chimpanzees SIVs, or do SIVs from other species contribute to HIV-1 genetic diversity? Is the mixing ongoing?

Hypothetical continued lentivirus cross-species transmissions to man

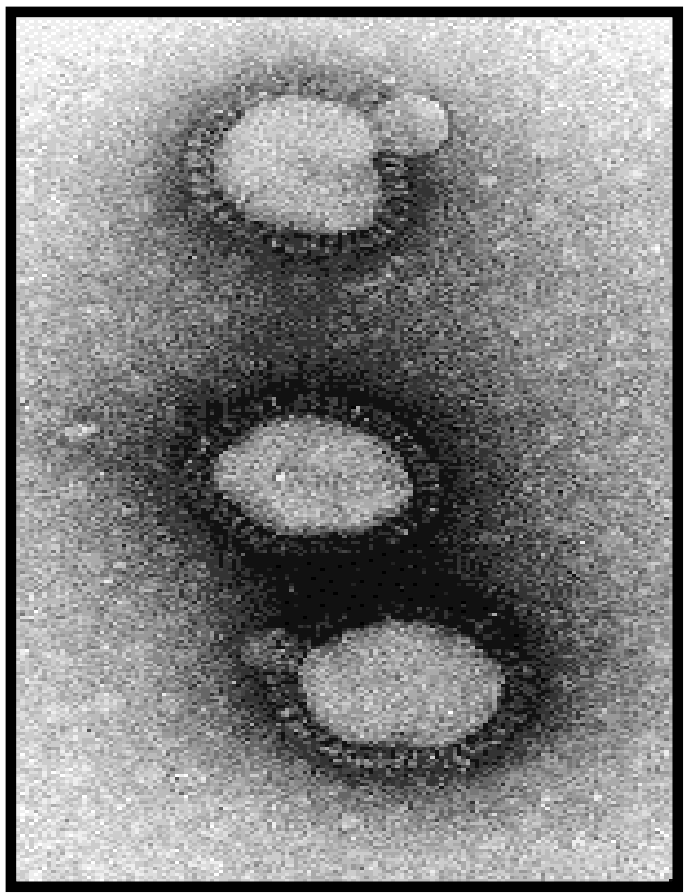


Non-human primates as "bushmeat"



MOLECULAR EPIDEMIOLOGY OF INFLUENZA

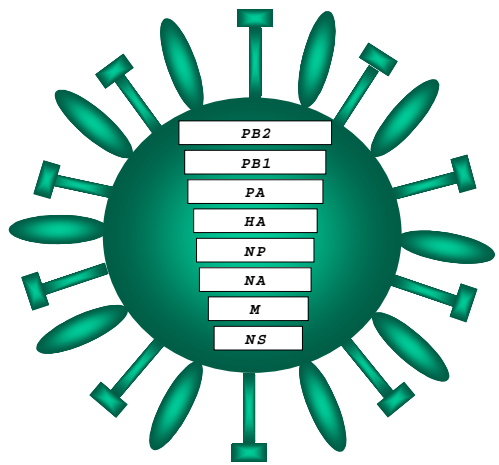
INFLUENZA “A” VIRUS

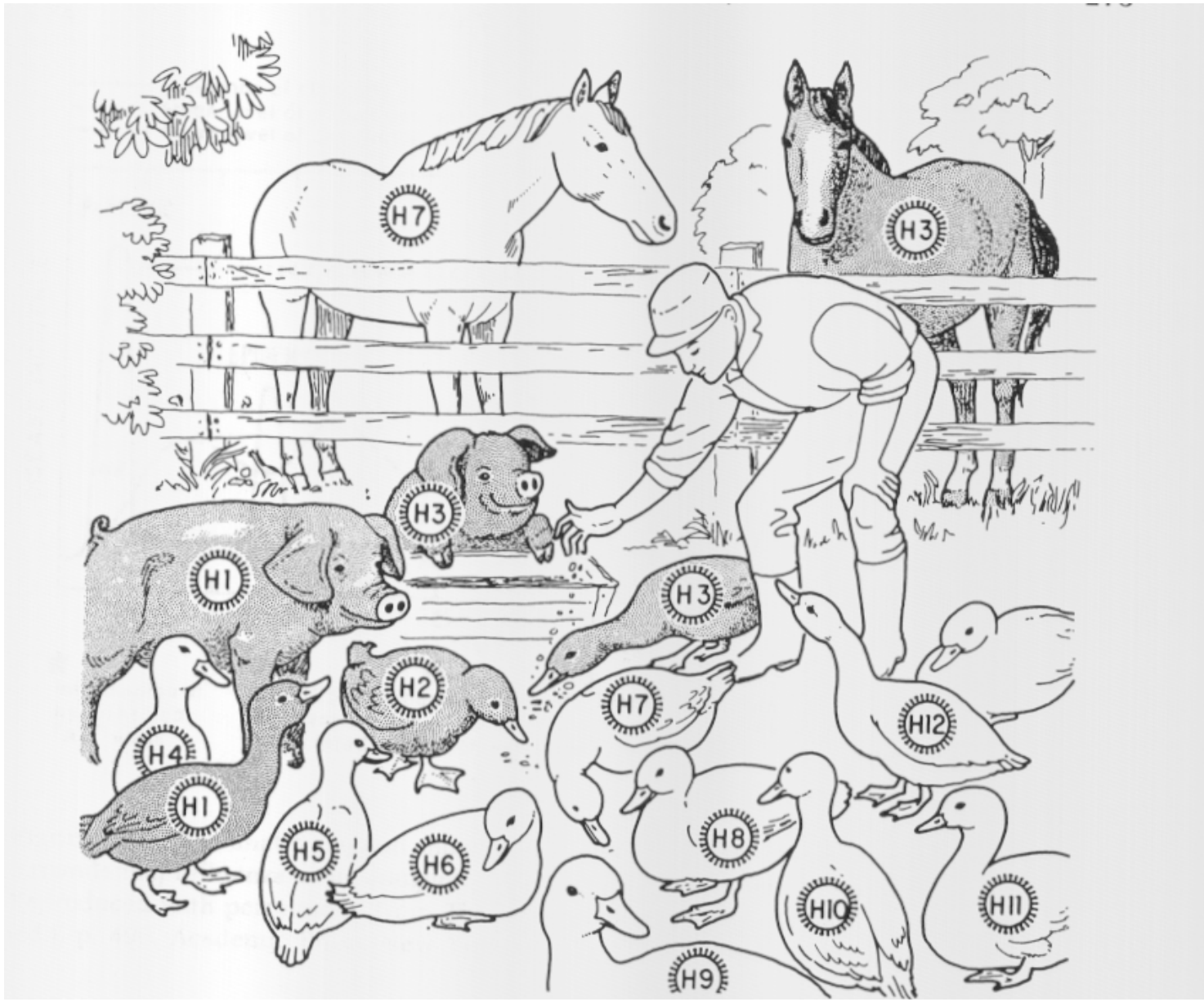


INFLUENZA A VIRUS GENOME RNA SEGMENTS AND CODING ASSIGNMENTS (FOR PROTOTYPE VIRUS A/PR/8/34)

SEG #	LENGTH	PROTEIN	FUNCTION
1	2,341	PB2	POLYMERASE (BASIC)
2	2,341	PB1	POLYMERASE (BASIC)
3	2,233	PA	POLYMERASE (ACIDIC)
4	1,778	HA	HEMAGLUTININ (SURFACE)
5	1,565	NP	NUCLEOCAPSID PROTEIN
6	1,413	NA	NEURAMINIDASE (SURFACE)
7	1,027	M1 & M2	MATRIX
8	890	NS1 & NS2	NON-STRUCTURAL

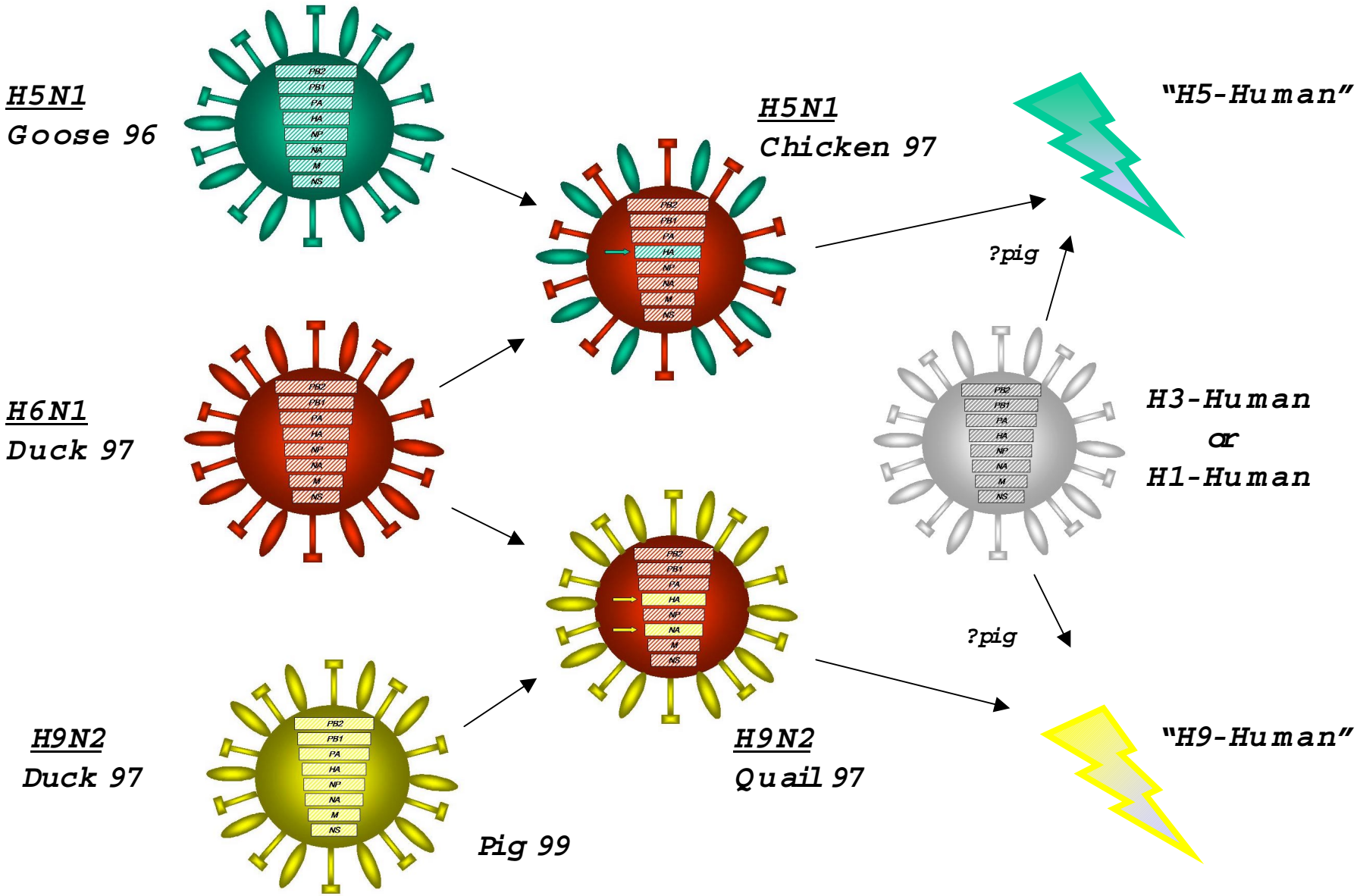
13,588 = TOTAL LENGTH (NUCLEOTIDES)





Variants of "human" influenza A virus subtypes (H1, H2, are found in domestic animals (Kilbourne)

Concern: Possible generation of H5 or H9 Human Influenza A strains

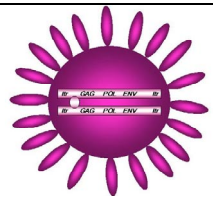
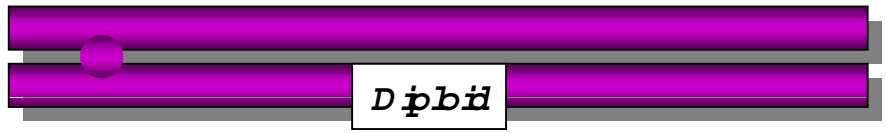
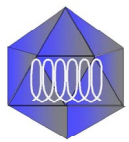
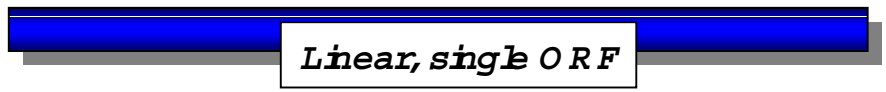
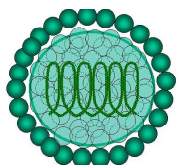
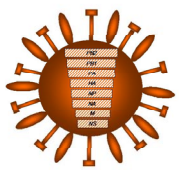
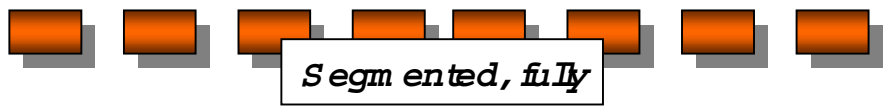
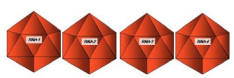
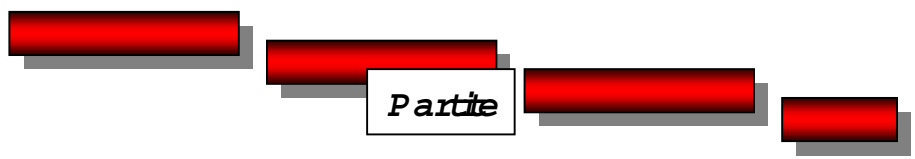


Central hypothesis: RNA virus genomes are structured for "evolvability"

Value of study of RNA virus "evolvability":

- Important pathogens
- Rapidly evolving
- Large gene pool in animals = source for recombination
- Full length nucleotide sequencing commonplace
- Computational simulation feasible
- Insights into the evolution of evolvability

Proposed new classification of RNA viruses based on genome configurations (Donald S. Burke)

FAMILY	VIRION MORPHOLOGY	GENOME CONFIGURATION
RETRO (HIV)		
PICORNA (Polio)		
CORONA (Corona)		
BUNYA (Hanta)		
ORTHOMYXO (Influenza)		
BROMO (plant viruses)		

Natural selection of genomes for "evolvability"?

Mutation rate?

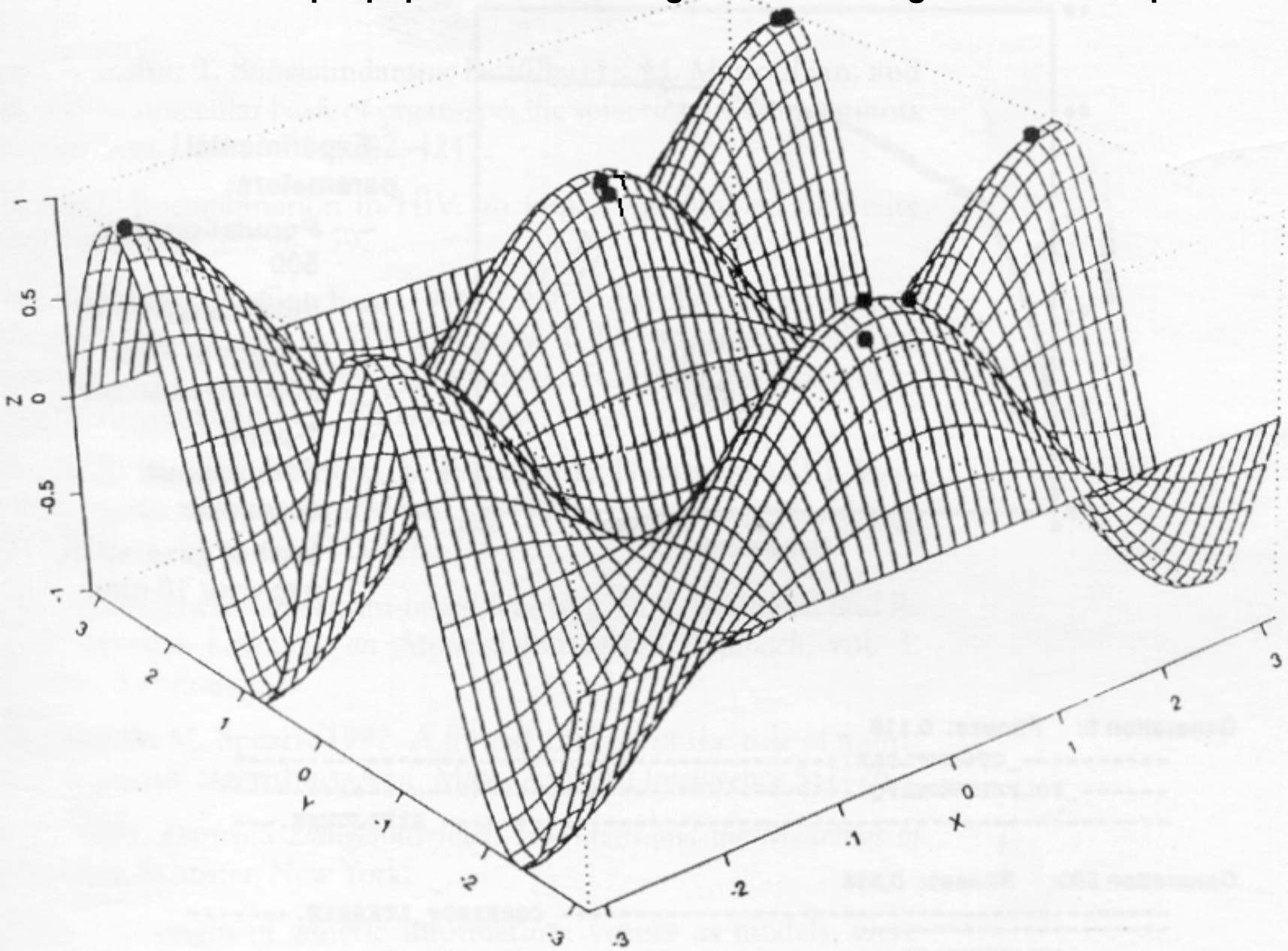
Length ?

Recombination ("cross-over", modularity) ?

Approach: Test the concept using Genetic Algorithms to simulate viral evolution


HAVE CENTER FOR APPLIED RESEARCH
IN ARTIFICIAL INTELLIGENCE

Representation of evolving bit strings in a fitness landscape.
In this example populations of strings are colonizing local fitness optima.



The VIV (Virtual Virus) Model

Approach: Build computational simulation of viral evolution.

Computational problem: Correctly spell the following combination of words, in any order

_ENVELOPE.

_COREPROTEIN.

_POLYMERASE.

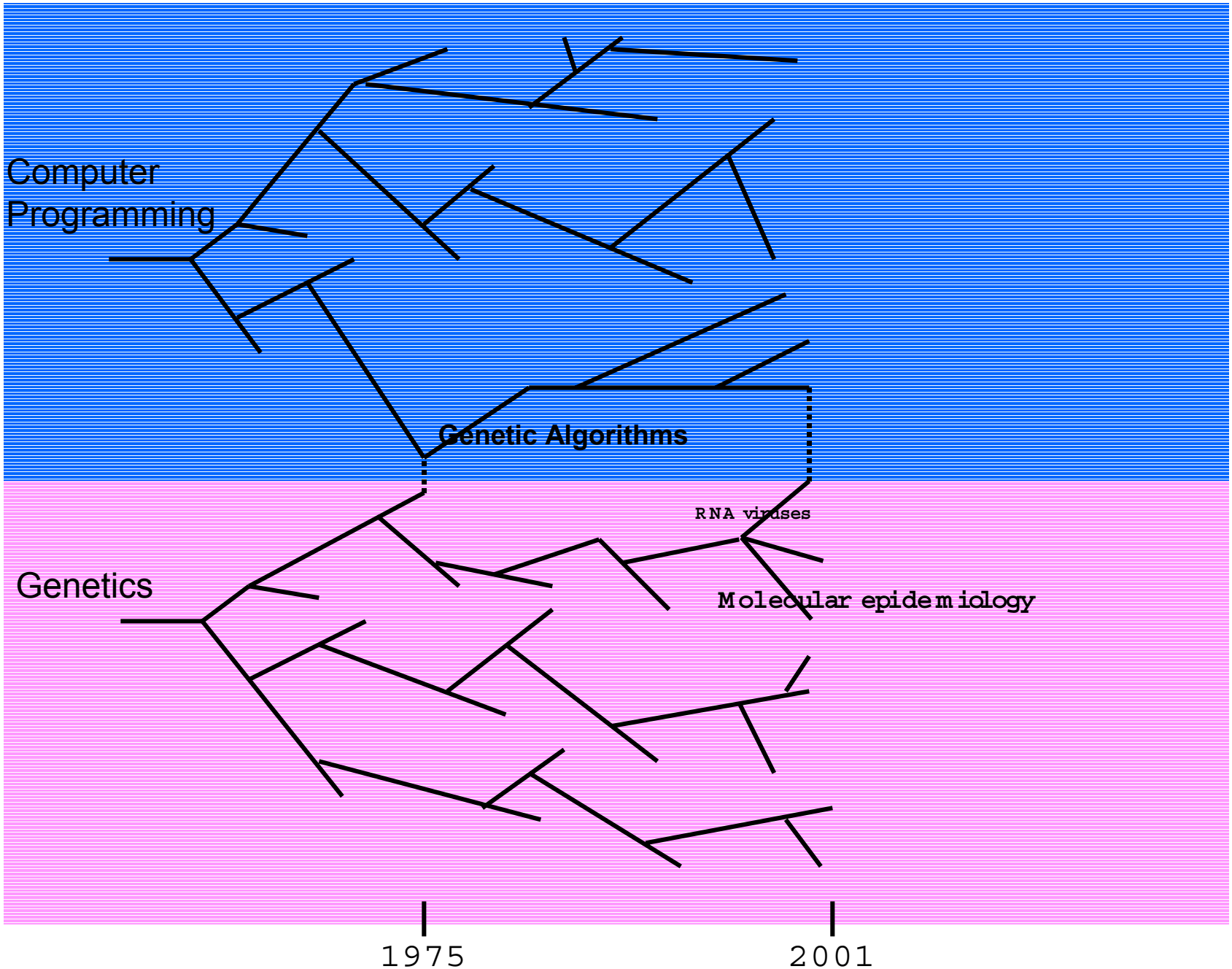
Problem calls for selection of 8+2+11+2+10+2 = 35 codons = 105 bits

---

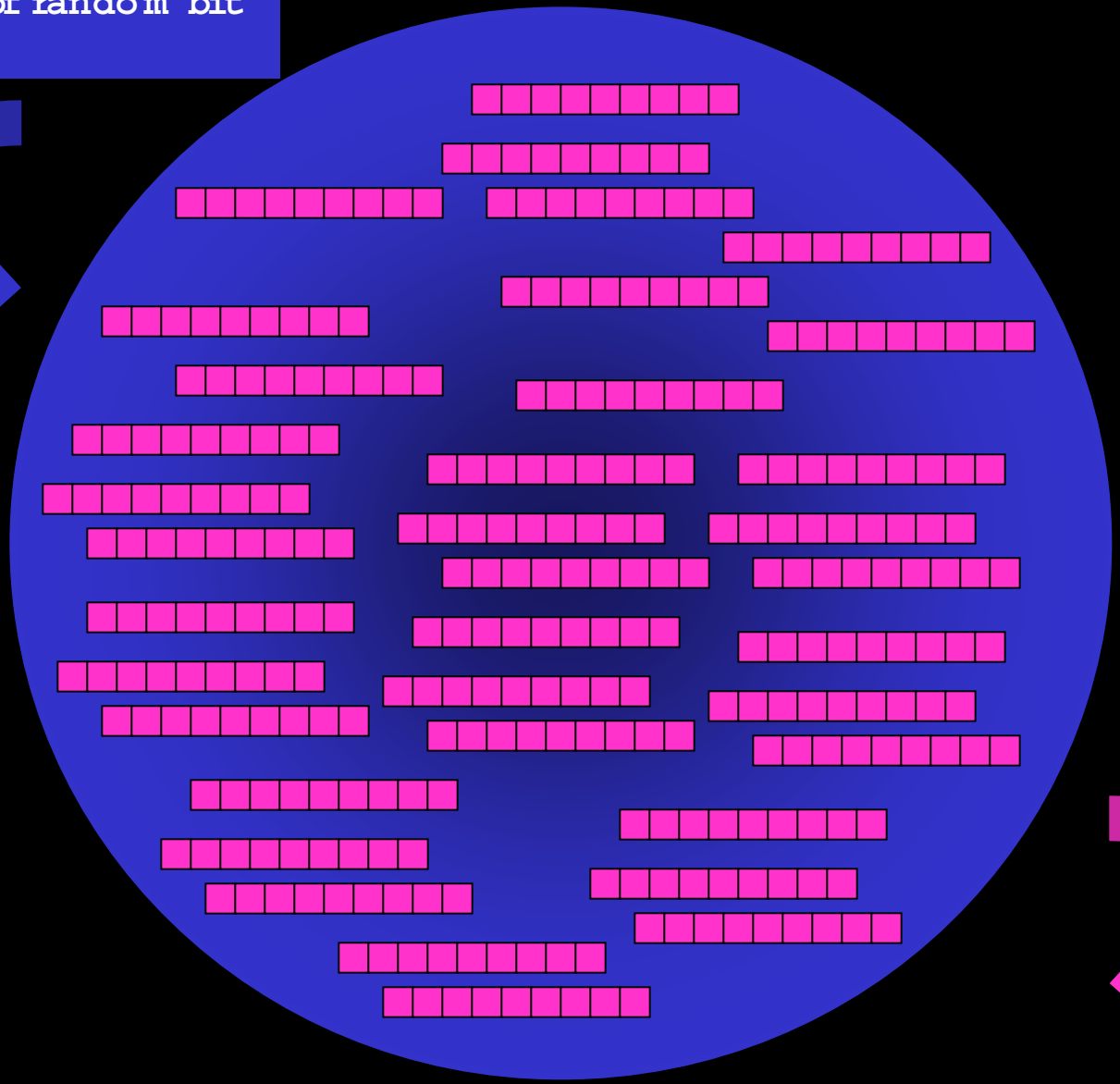
Putting More Genetics into Genetic Algorithms
Burke DS, DeJong KA, Grefenstette JJ, Loggia Ramsey C, Wu AS
Evolutionary Computation 1998; 6: 387 - 410

Genome Length as an Evolutionary Self-adaptation
Ramsey CL, DeJong KA, Grefenstette JJ, Wu AS, Burke DS
Parallel Problem Solving from Nature 1998; 5: 345-353

Visual Analysis of Evolutionary Algorithms
Wu AS, DeJong KA, Burke DS, Grefenstette JJ, Ramsey CL
Proceedings of the Congress on Evolutionary Computation 1999; 2: 1419-25


Computer
Programming
Genetic Algorithms
RNA viruses
Genetics
Molecular epidemiology
1975
2001

Population of random bit strings



Example

A	T	G	T	G	A	T	C	C

The biological genetic code: codons to amino acids

First base	Second base: T		C		A		G	
T	TTT	*Phe*	TCT	*Ser*	TAT	*Tyr*	TGT	*Cys*
	TTC	*Phe*	TCC	*Ser*	TAC	*Tyr*	TGC	*Cys*
	TTA	*Leu*	TCA	*Ser*	TAA	*STOP*	TGA	*STOP*
	TTG	*Leu*	TCG	*Ser*	TAG	*STOP*	TGG	*Trp*
C	CTT	*Leu*	CCT	*Pro*	CAT	*His*	CGT	*Arg*
	CTC	*Leu*	CCC	*Pro*	CAC	*His*	CGC	*Arg*
	CTA	*Leu*	CCA	*Pro*	CAA	*Gln*	CGA	*Arg*
	CTG	*Leu*	CCG	*Pro*	CAG	*Gln*	CGG	*Arg*
A	ATT	*Ile*	ACT	*Thr*	AAT	*Asn*	AGT	*Ser*
	ATC	*Ile*	ACC	*Thr*	AAC	*Asn*	AGC	*Ser*
	ATA	*Met*	ACA	*Thr*	AAA	*Lys*	AGA	*Arg*
	ATG	*Met*	ACG	*Thr*	AAG	*Lys*	AGG	*Arg*
G	GTT	*Val*	GCT	*Ala*	GAT	*Asp*	GGT	*Gly*
	GTC	*Val*	GCC	*Ala*	GAC	*Asp*	GGC	*Gly*
	GTA	*Val*	GCA	*Ala*	GAA	*Glu*	GGA	*Gly*
	GTG	*Val*	GCG	*Ala*	GAG	*Glu*	GGG	*Gly*

The VIV genetic code: codons to alphabet letters

Second base →

First base ↓

	T		C		A		G	
T	TTT	A	TCT	C	TAT	E	TGT	G
	TTC	A	TCC	C	TAC	E	TGC	G
	TTA	B	TCA	D	TAA	F	TGA	H
	TTG	B	TCG	D	TAG	F	TGG	H
C	CTT	I	CCT	K	CAT	M	CGT	O
	CTC	I	CCC	K	CAC	M	CGC	O
	CTA	J	CCA	L	CAA	N	CGA	P
	CTG	J	CCG	L	CAG	N	CGG	P
A	ATT	Q	ACT	S	AAT	U	AGT	W
	ATC	Q	ACC	S	AAC	U	AGC	W
	ATA	R	ACA	T	AAA	V	AGA	X
	ATG	R	ACG	T	AAG	V	AGG	X
G	GTT	Y	GCT	A	GAT	U	GGT	_ start
	GTC	Y	GCC	E	GAC	Y	GGC	_ start
	GTA	Z	GCA	I	GAA	+	GGA	. stop
	GTG	Z	GCG	0	GAG	+	GGG	. stop

COTEWPSOTEEM | POLAMERADE | ENEVELOPE

S1 = 0.591 S2 = 0.800 S3 = 0.938

Raw fitness = [(0.591) ^{2} + (0.800) ^{2} + (0.938) ^{2}] / 3 = 0.623

Note square of fitness is used to accentuate fitness differences between highly fit individuals

Length penalty:

Fitness = [(1 – (L / 7500)] x [Raw fitness]

If length = 200, fitness = 0.97 x raw fitness

length = 500, fitness = 0.93 x " "

length = 2000, fitness = 0.73 x " "

The standard VIV model run:

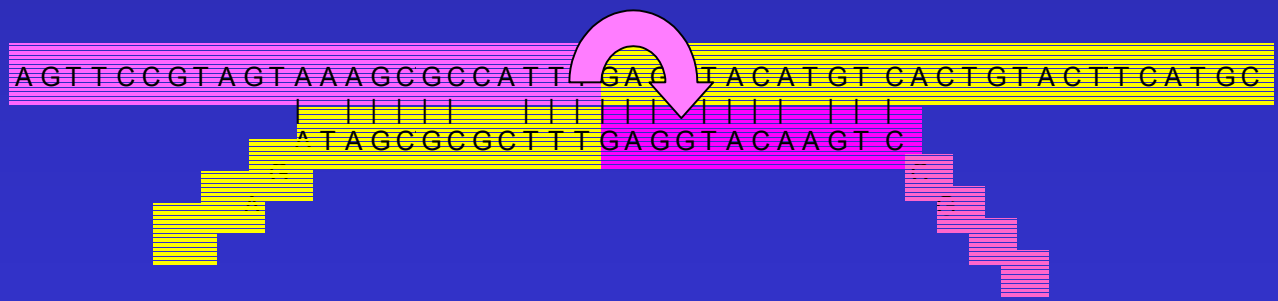
Fixed population size = 500 strings
Init genome lengths = 100 to 500 bits (random)
Max genome length = 7500 bits
Selection proportional to individual string fitness
Experimental run = 2000 generations
Entire population replaced each generation

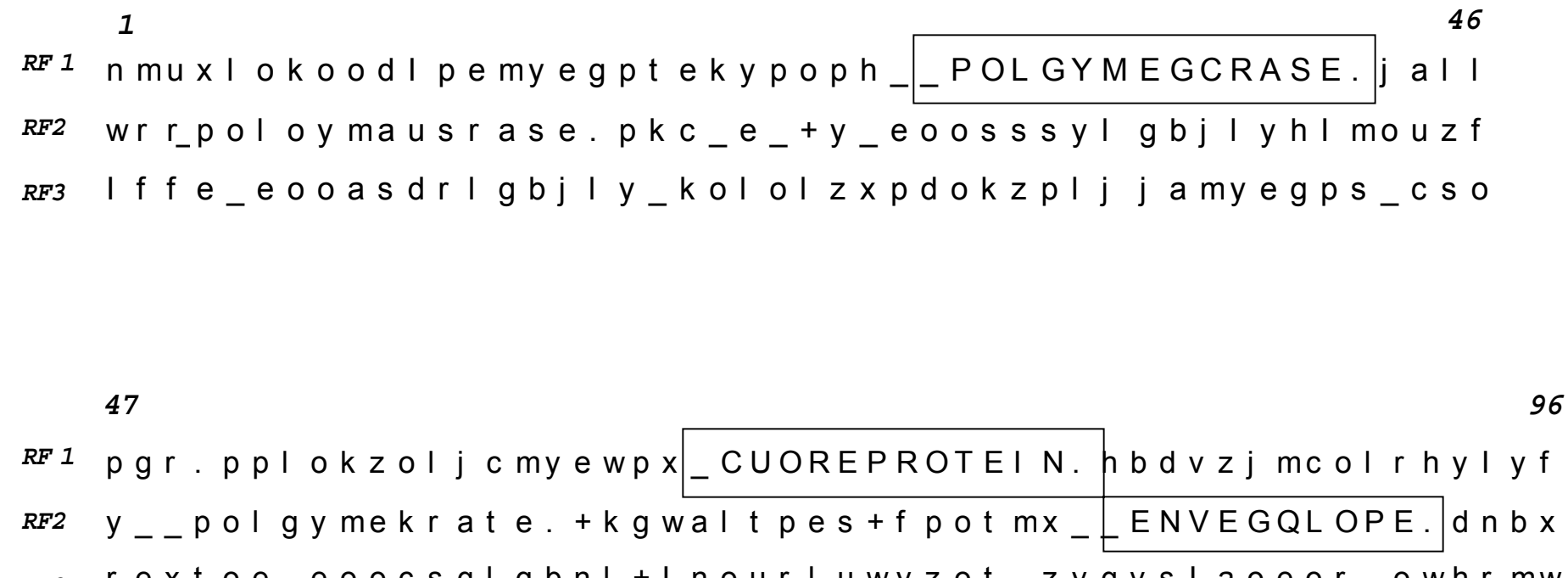
Mechanism of Homologous Recombination

1. Select random 24 bit window on string 1

2. Slide string 2 over window, to identify homologous sequences

3. Break and cross-over (recombine) strings at random point within the window of homology



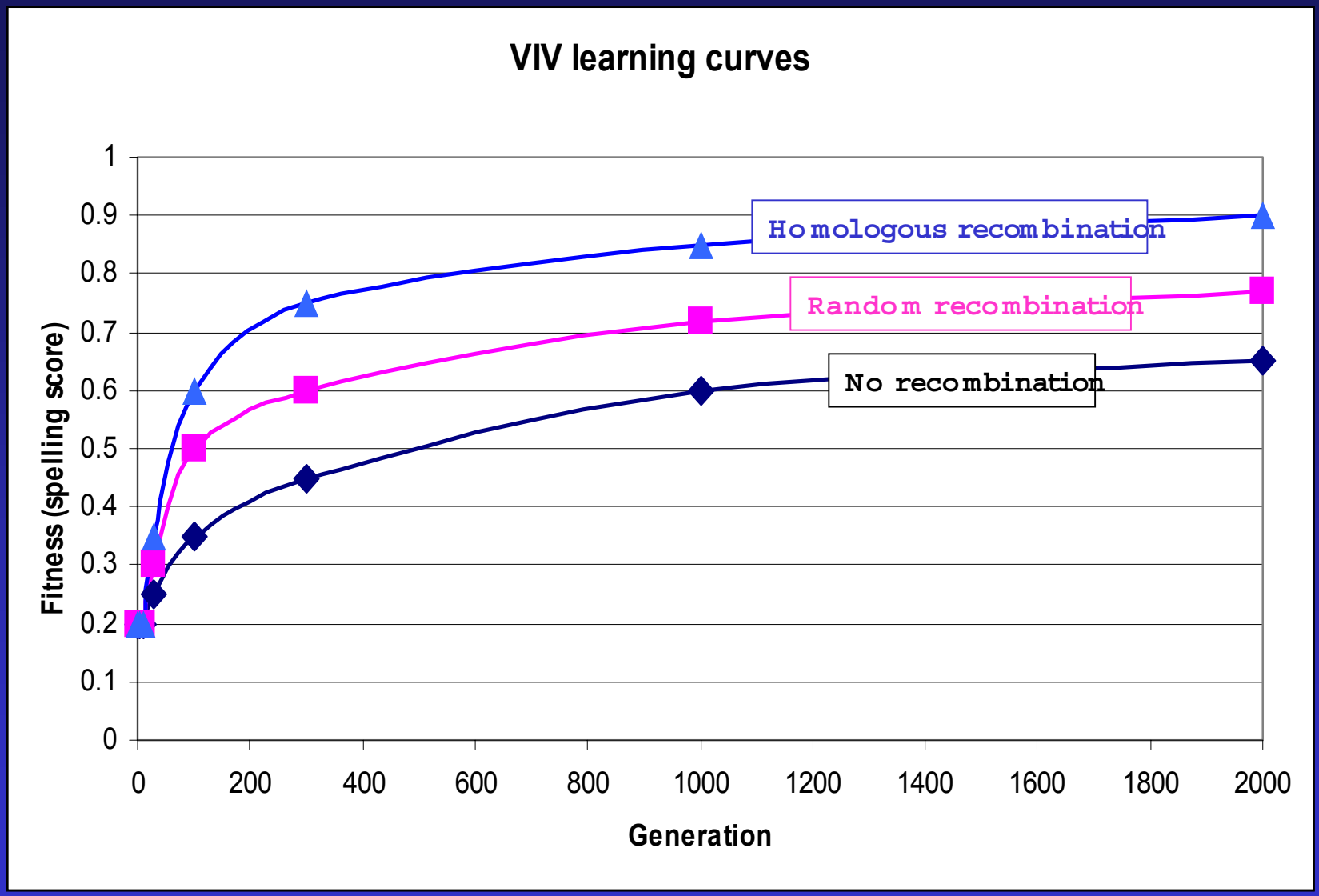


Example of an evolved VIV code string: 1000th generation, raw fitness = 0.8702, length corrected fitness = 0.8365, genome string length = 290 nucleotides, phenotype string length = 96 letters. In this example POLYMERASE and COREPROTEIN are encoded on ORF1, while ENVELOPE is encoded on ORF2

```
      1                                                                  46
RF 1  n m u x l o k o o d l p e m y e g p t e k y p o p h _ _ P O L G Y M E G C R A S E . j a l l
RF2   w r r_p o l o y m a u s r a s e . p k c _ e _ + y _ e o o s s s y l g b j l y h l m o u z f
RF3   l f f e _ e o o a s d r l g b j l y _ k o l o l z x p d o k z p l j j a m y e g p s _ c s o

      47                                                                 96
RF 1  p g r . p p l o k z o l j c m y e w p x _ C U O R E P R O T E I N . h b d v z j m c o l r h y l y f
RF2   y _ _ p o l g y m e k r a t e . + k g w a l t p e s + f p o t m x _ _ E N V E G Q L O P E . d n b x
RF3   r e x t o e _ e o o c s q l g b n l + l n o u r l u w y z o t . z y q v s l a o e o r _ o w h r m w
```

Also note in this example that there are two inactive "pseudogenes" that do not contribute to the fitness

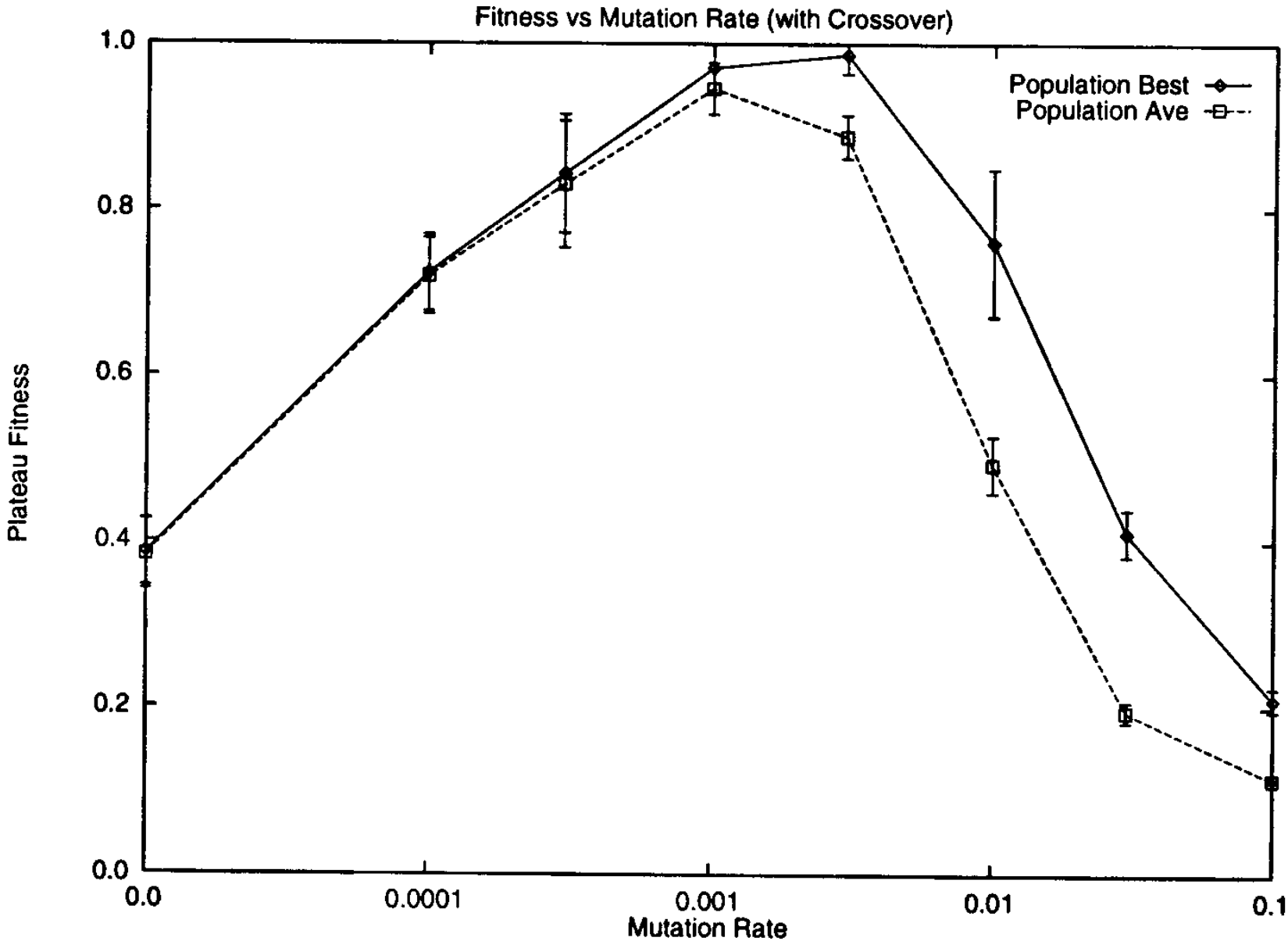


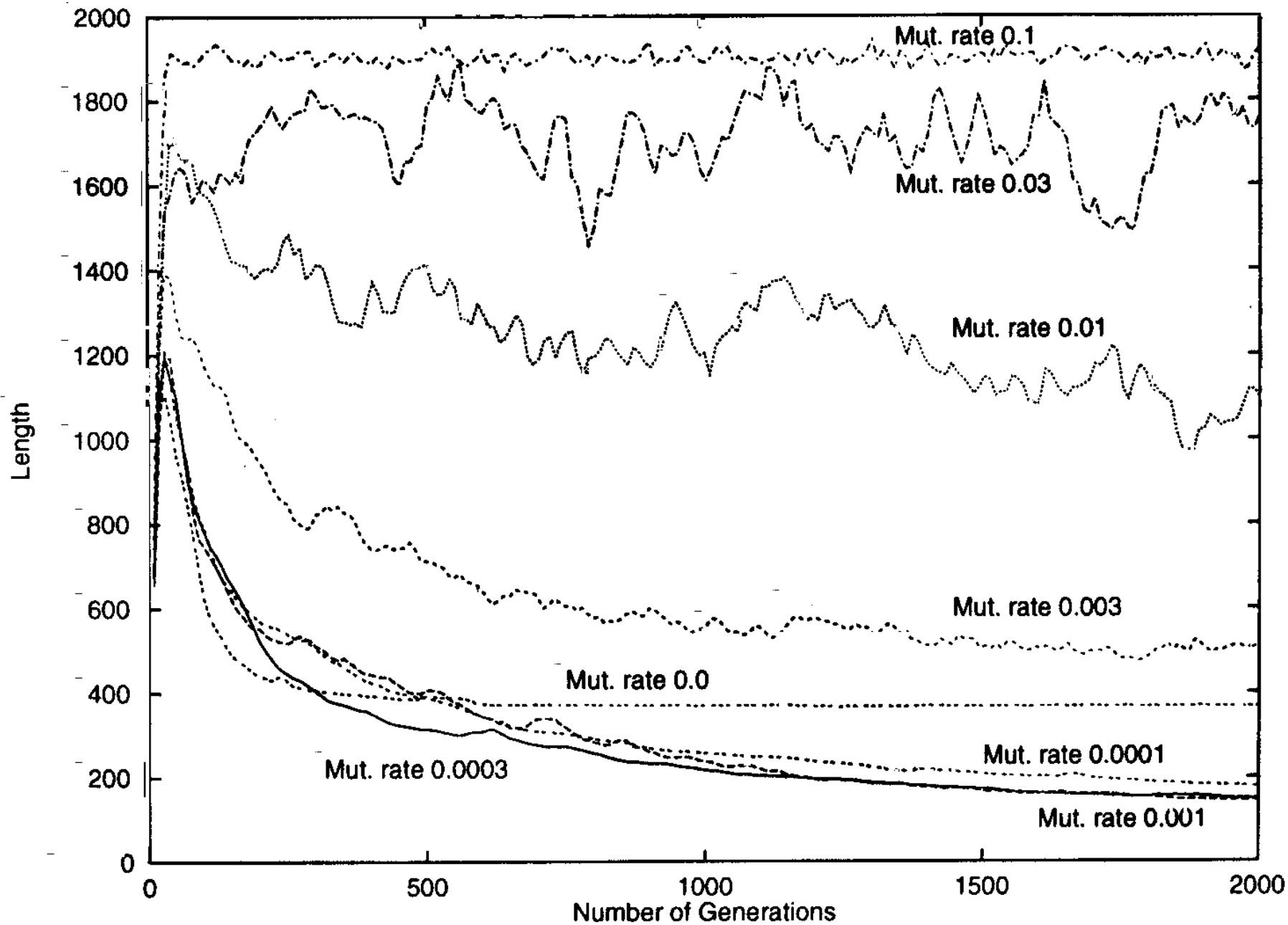
Homologous recombination > Random recombination > No recombination

For a given problem what is the optimum

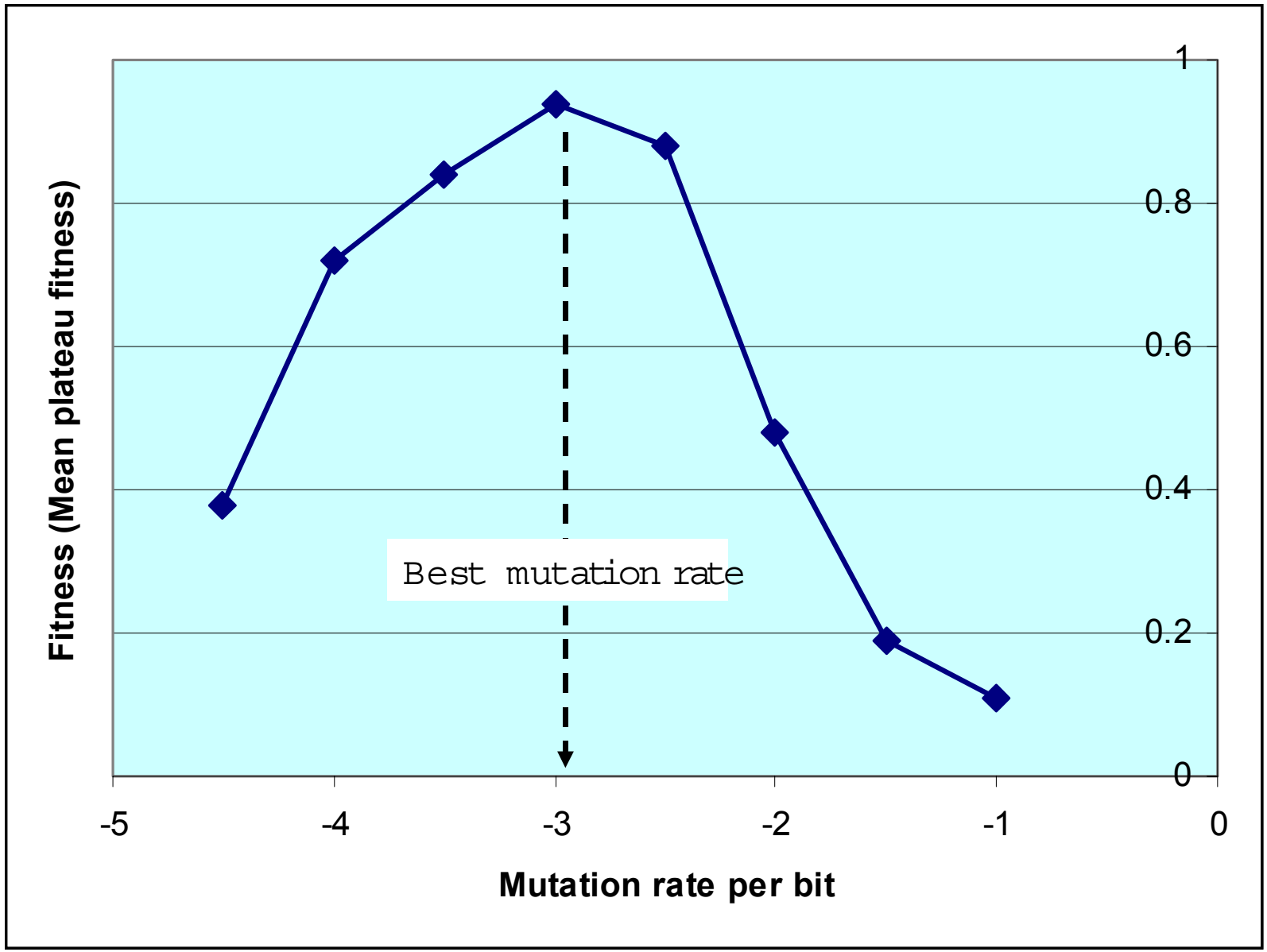
Mutation rate?

Genome length?

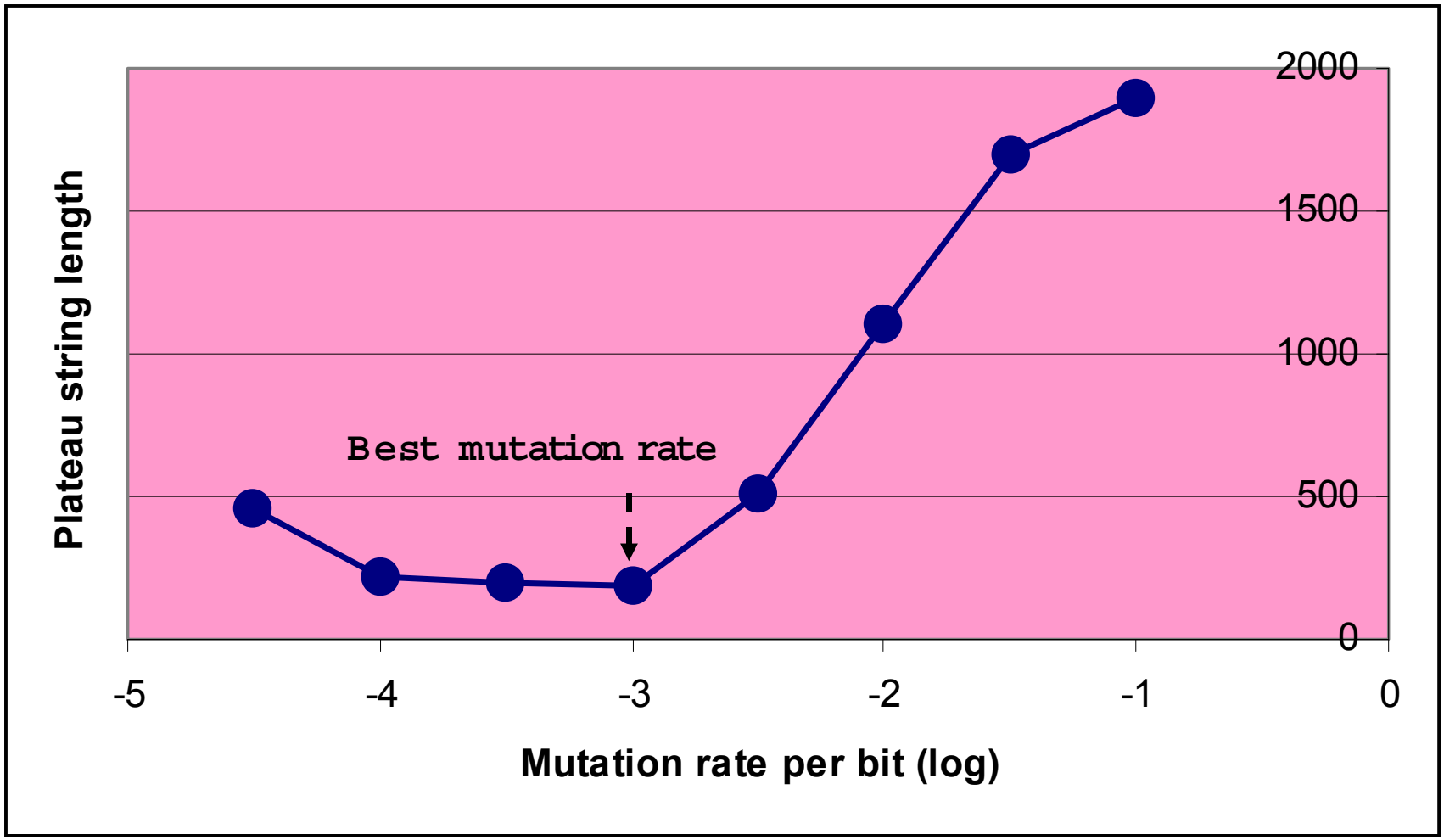

Fitness vs Mutation Rate (with Crossover)
Population Best
Population Ave
Plateau Fitness
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.0001
0.001
0.01
0.1
Mutation Rate


2000
1800
1600
1400
1200
1000
800
600
400
200
0
Length
0
500
1000
1500
2000
Number of Generations
Mut. rate 0.1
Mut. rate 0.03
Mut. rate 0.01
Mut. rate 0.003
Mut. rate 0.0
Mut. rate 0.0001
Mut. rate 0.0003
Mut. rate 0.001

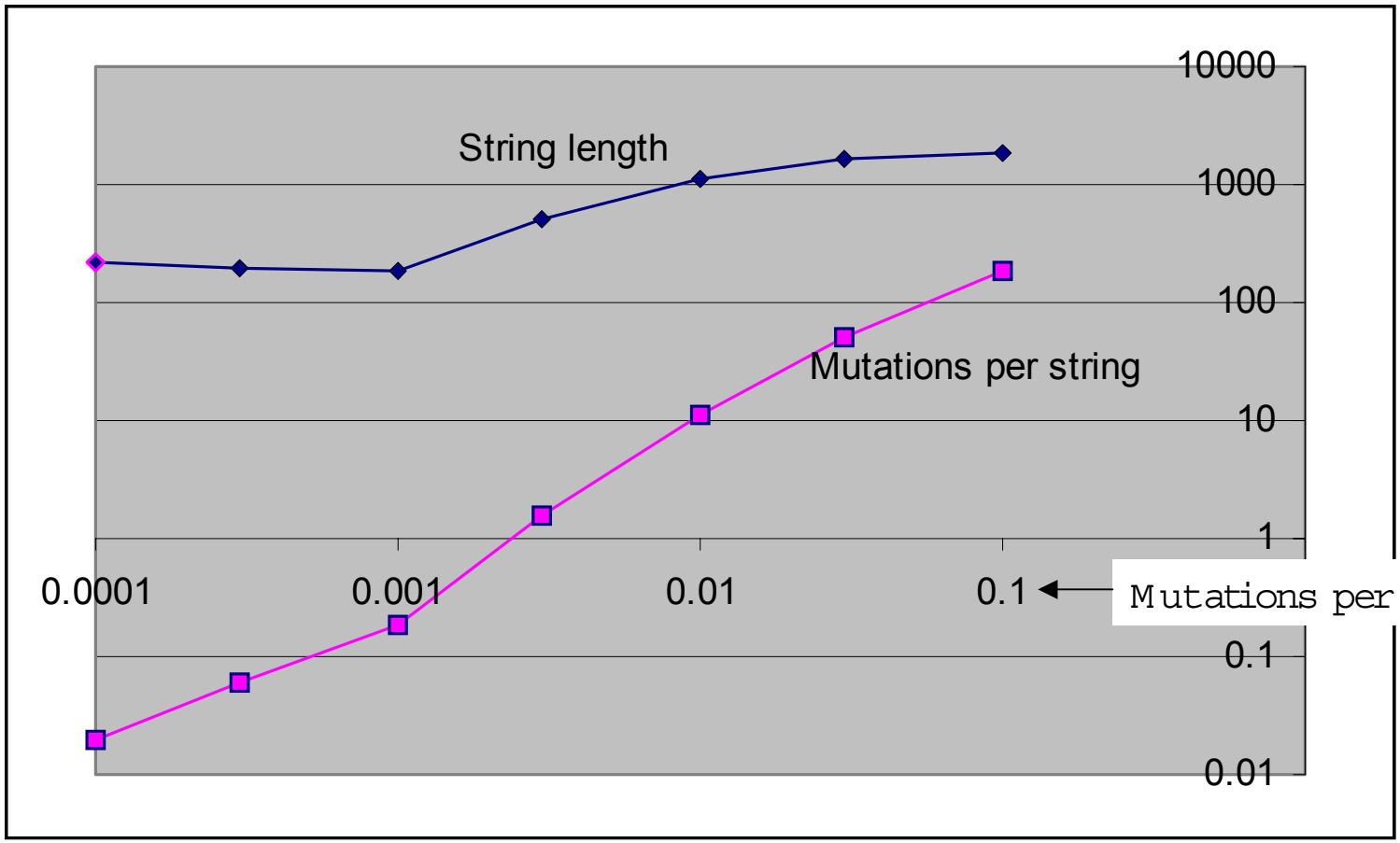
		At Plateau		
per bit (exper)		Fitness	String Length	Mutations per String
0		0.38	460	0
0.0001		0.72	220	0.022
0.0003		0.84	200	0.06
0.001		0.94	190	0.19
0.003		0.88	510	1.53
0.01		0.48	1100	11
0.03		0.19	1700	51
0.1		0.11	1900	190


Fitness (Mean plateau fitness)
1
0.8
0.6
0.4
0.2
0
Best mutation rate
-5
-4
-3
-2
-1
0
Mutation rate per bit

		Plateau		
per bit (exper)		Fitness	String Length	Mutations per String
0		0.38	460	0
0.0001		0.72	220	0.022
0.0003		0.84	200	0.06
0.001		0.94	190	0.19
0.003		0.88	510	1.53
0.01		0.48	1100	11
0.03		0.19	1700	51
0.1		0.11	1900	190


Plateau string length
2000
1500
1000
500
0
Best mutation rate
-5
-4
-3
-2
-1
0
Mutation rate per bit (log)

		Plateau		
per bit (exper)		Fitness	String Length	Mutations per String
0		0.38	460	0
0.0001		0.72	220	0.022
0.0003		0.84	200	0.06
0.001		0.94	190	0.19
0.003		0.88	510	1.53
0.01		0.48	1100	11
0.03		0.19	1700	51
0.1		0.11	1900	190

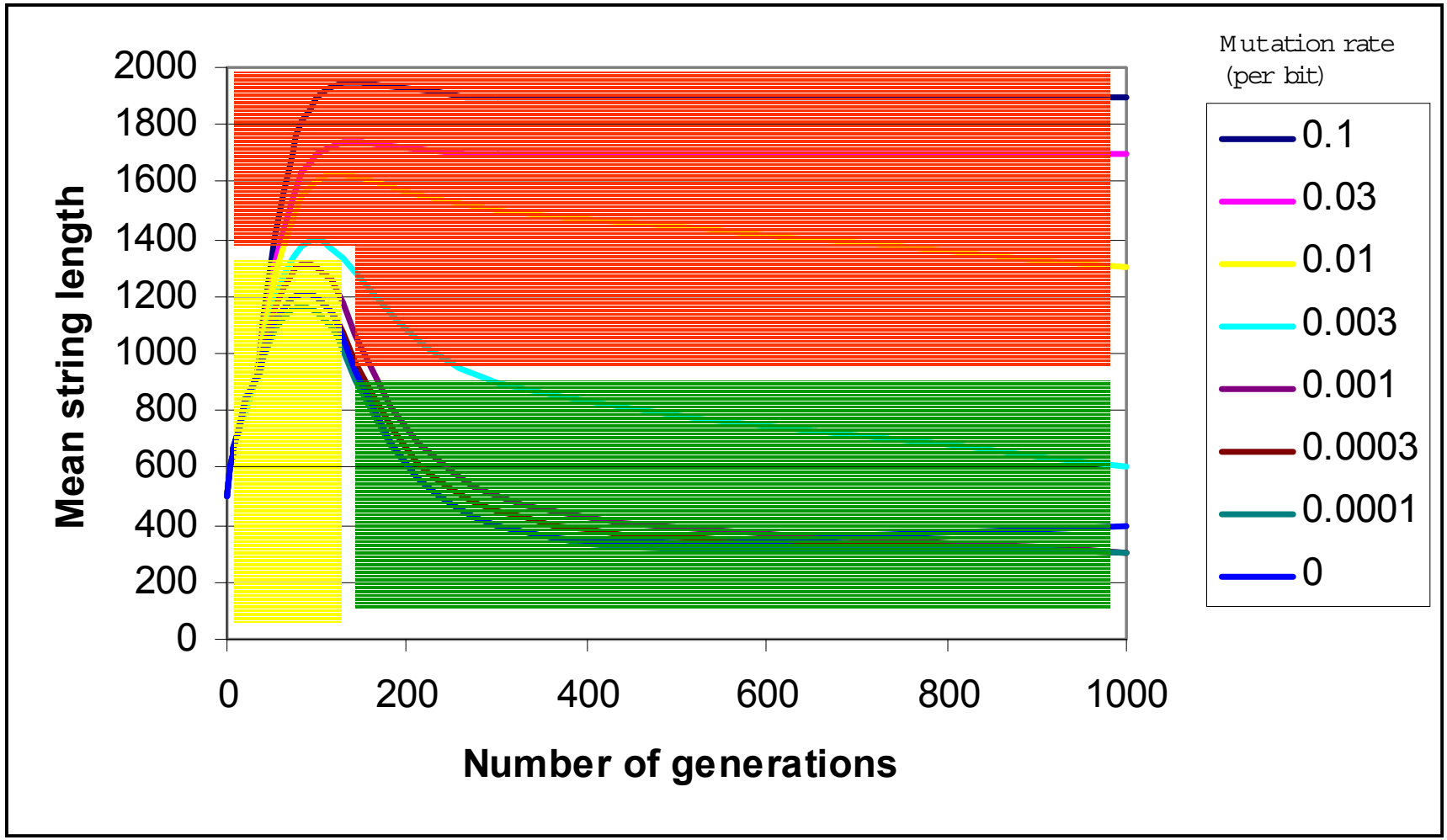


Mut/Bit	Length	Mut/string
0	460	0
0.0001	220	0.02
0.0003	200	0.06
0.001	190	0.19
0.003	510	1.53
0.01	1100	11
0.03	1700	51
0.1	1900	190

```
      1                                                                  46
RF 1  n m u x l o k o o d l p e m y e g p t e k y p o p h _ _ P O L G Y M E G C R A S E . j a l l
RF2   w r r_p o l o y m a u s r a s e . p k c _ e _ + y _ e o o s s s y l g b j l y h l m o u z f
RF3   l f f e _ e o o a s d r l g b j l y _ k o l o l z x p d o k z p l j j a m y e g p s _ c s o

      47                                                                 96
RF 1  p g r . p p l o k z o l j c m y e w p x _ C U O R E P R O T E I N . h b d v z j m c o l r h y l y f
RF2   y _ _ p o l g y m e k r a t e . + k g w a l t p e s + f p o t m x _ _ E N V E G Q L O P E . d n b x
RF3   r e x t o e _ e o o c s q l g b n l + l n o u r l u w y z o t . z y q v s l a o e o r _ o w h r m w
```

Also note in this example that there are two inactive "pseudogenes" that do not contribute to the fitness


Mean string length
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
Number of generations
Mutation rate
(per bit)
0.1
0.03
0.01
0.003
0.001
0.0003
0.0001
0

Conclusions from VIV model regarding string lengths:

If the mutation rate is $>> 1$ mutation per target string:

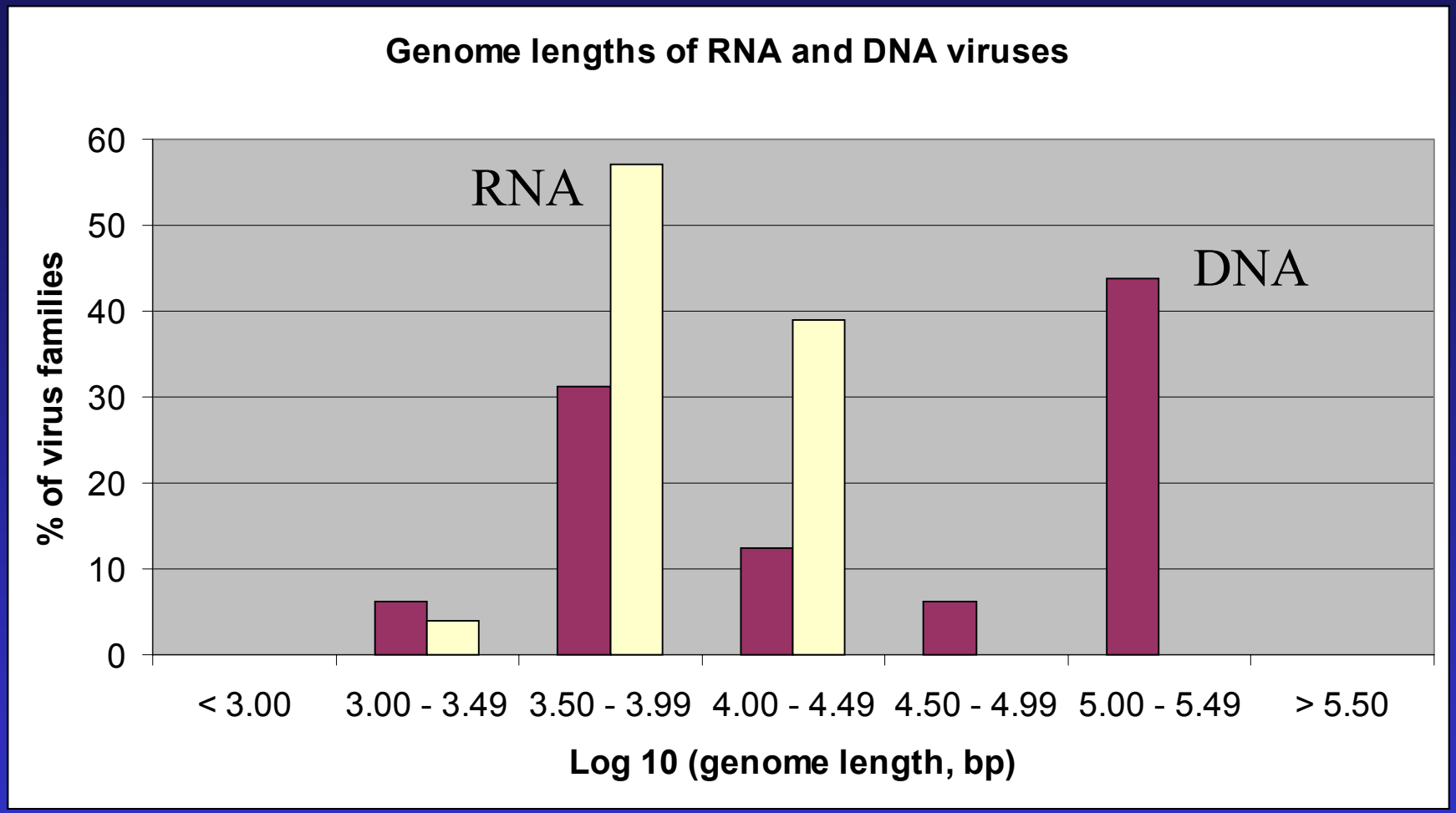
- Unstable state
- String length increases, duplicate genes accumulate
- But mutations per string also increases
  - "Mutational meltdown" $\longrightarrow$ low fitness

If the mutation rate is $<< 1$ mutation per target string

- Initial burst of string lengthening, then shortening
- Slow improvement in fitness

If the mutation rate is ≈ 1 mutation target string

- Initial burst of string lengthening, then shortening
- Rapid improvement in fitness


Genome lengths of RNA and DNA viruses
% of virus families
60
50
40
30
20
10
0
RNA
DNA
< 3.00
3.00 - 3.49
3.50 - 3.99
4.00 - 4.49
4.50 - 4.99
5.00 - 5.49
> 5.50
Log 10 (genome length, bp)

Mutation rates per nucleotide and per genome in real-world biological RNA viral replication

From Eigen, M. The origin of genetic information: viruses as models. Gene. 1993; 135: 37-47.

Life form	Genome size: (number of nt or bp)	Error Rate: (per replication per nt)	Error Rate: (per replication per genome)
RNA			
Bacteriophage Qβ	4200	3×10^{-4}	1.3
Polio-1 virus	7400	3×10^{-4}	0.2
Vesicular stomatitis virus	11000	1×10^{-4}	1.1
Influenza-A virus	14000	6×10^{-5}	0.8
HIV-1 (AIDS virus)	10000	1×10^{-4}	1.0
DNA			
Bacteriophage M13	6400	7×10^{-7}	4.6×10^{-3}
E. coli	4.7 mill	7×10^{-10}	3.3×10^{-3}
Yeast (Saccharomyces cerevisine)	13.8 mill	3×10^{-10}	3.8×10^{-3}
Human	3 billion	$\approx 10^{-12}$	$\approx 3 \times 10^{-3}$

Observation:

For bit strings evolving in silicon

and for RNA strings evolving in nature

there is selection of strings of length L such that

the mutation rate is approximately 1 / L

Hypothesis: For all RNA viruses, it is the fixed RNA polymerase error rate of ≈ 1 per 10,000 nucleotides that dictates the typical observed string length of ≈ 10,000 nucleotides

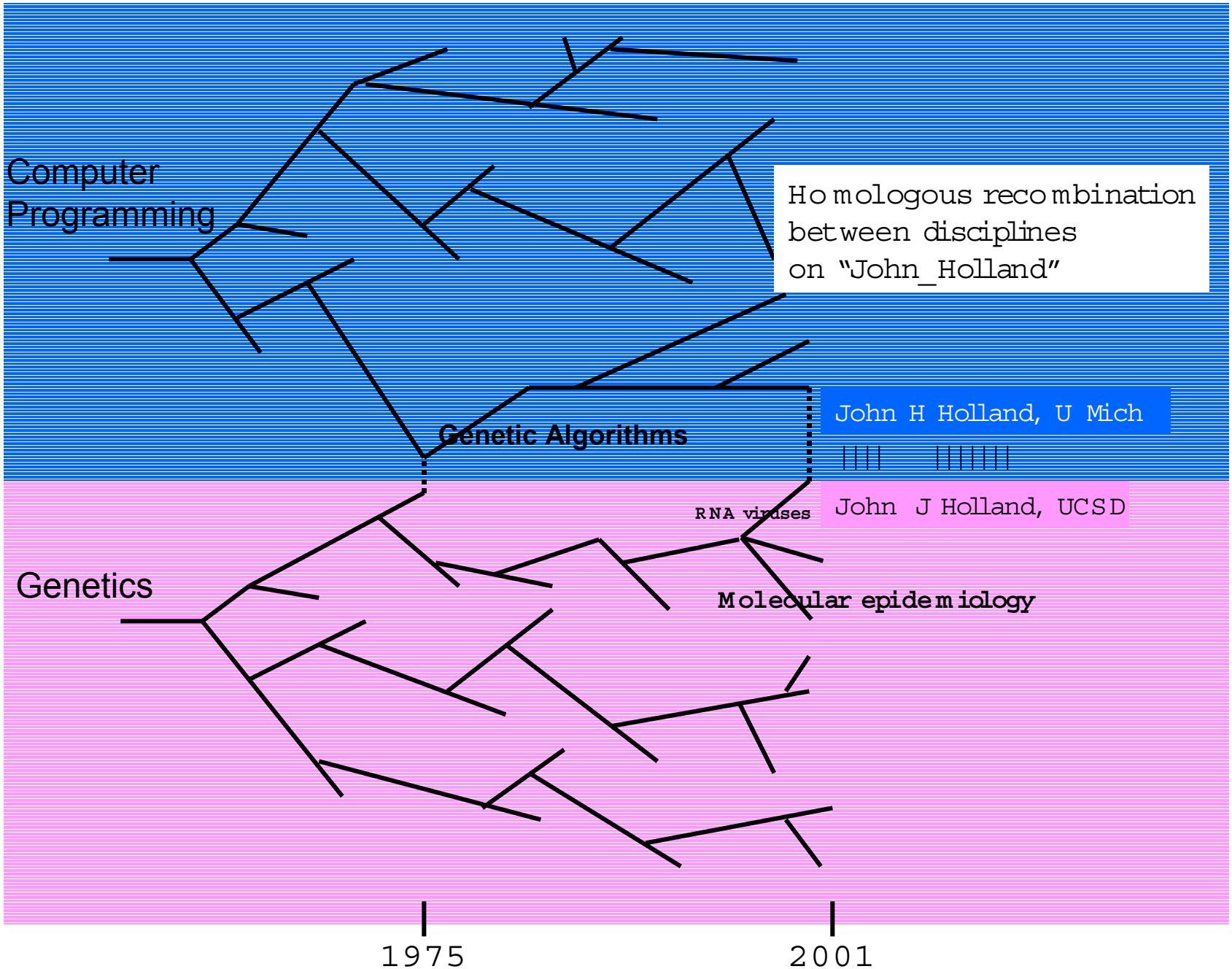
Philosophical speculation: The laws governing evolution of code strings are universal, independent of the code-stuff (ie, there is a science of "metagenetics").

Metagenetics: the universe of possible genotype / phenotype coding systems:

Computer	2 states / bit
	32-64 bits / byte (codon)
	1:1 genotype (code) to phenotype mapping
Biology	4 states / bit
	3 bits / codon
	64:23 (22aa+STOP)
	= 2.8 genotype to phenotype mapping
VIV	4 states / bit
	3 bits / codon
	64:29 (26 letters and + _ .)
	= 2.2 genotype to phenotype mapping
Endless Possibilities	6, 8, 10 states per bit
	3, 4, 5, etc bits per codon
	tight (1:1) or loose (>3:1) genotype to phenotype mapping

Why was 4 / 3 / 2.8 system selected for the biological code?

end


Computer Programming
Homologous recombination between disciplines on "John_Holland"
John H Holland, U Mich
Genetic Algorithms
John J Holland, UCSD
RNA viruses
Genetics
Molecular epidemiology
1975
2001