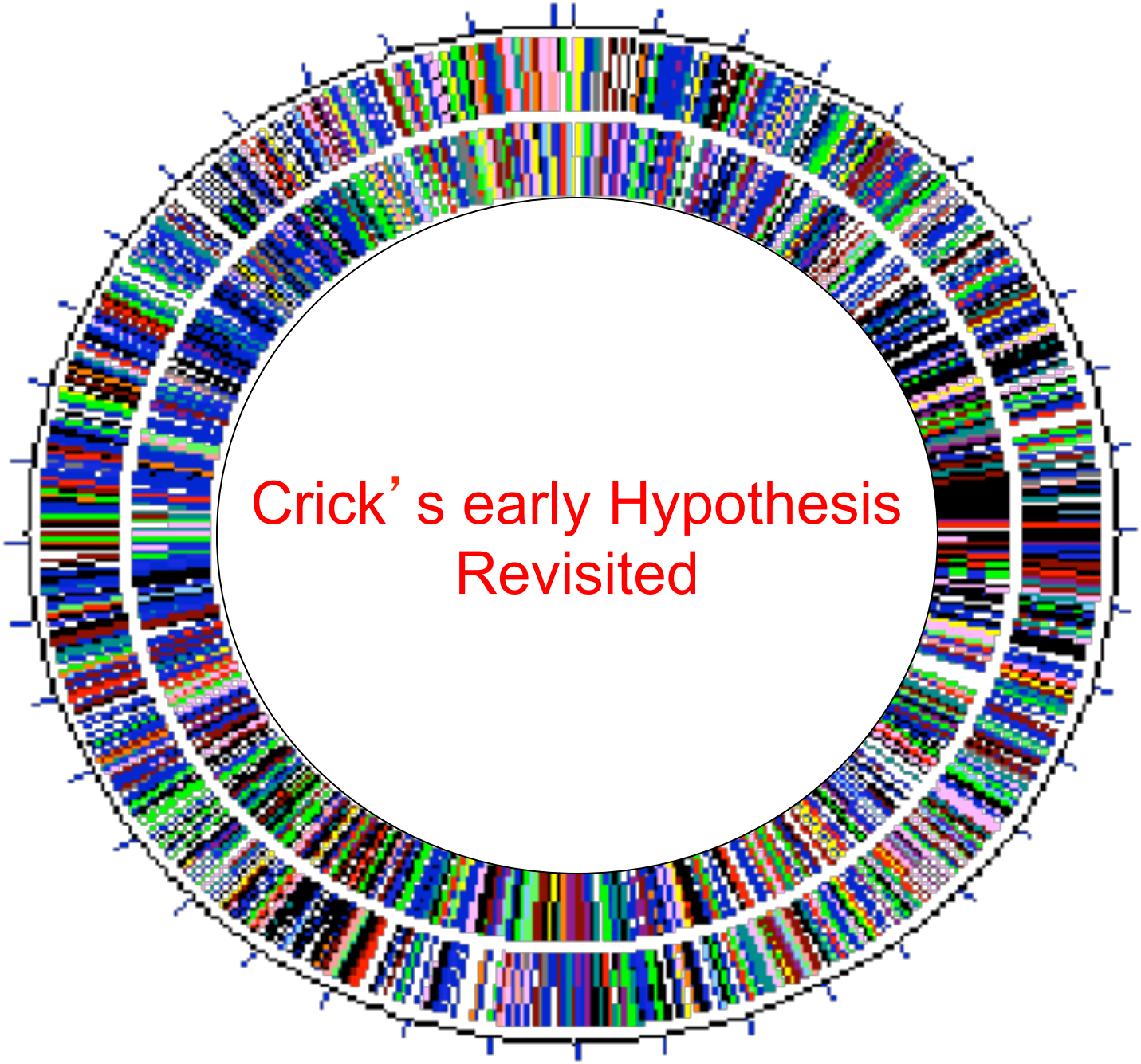

Crick’s early Hypothesis
Revisited

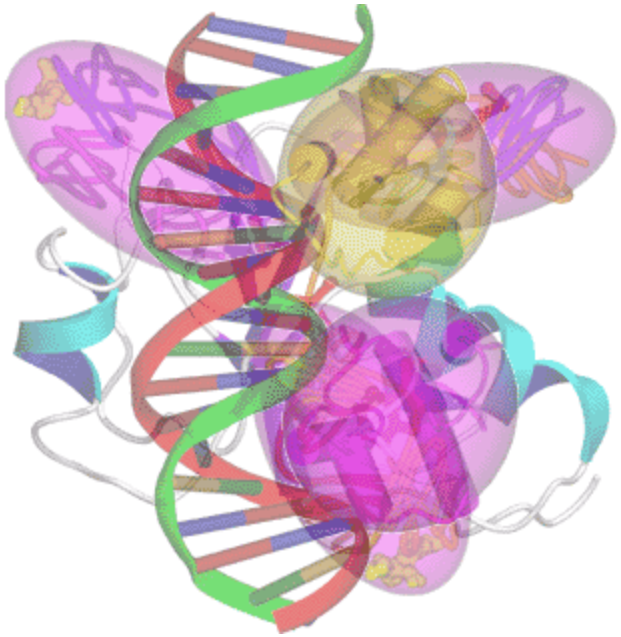
Or The Existence of a Universal Coding Frame

Ryan Rossi, Jean-Louis Lassez

and

Axel Bernal
UPenn Center for Bioinformatics

BIOINFORMATICS



The application of computer technology to the management and analysis of biological data

COMPUTATIONAL BIOLOGY

Biology: the study of living organisms

Why should computer scientists be interested in biology?

replay

Genomes and Genes

The language of life

…..catgcctagactgcatcggtaccatgacatgcatttatagaaca
ctacgcgtaatagccatgatcccatagatacatacagagataca
ctgatagactcgacctcatccgattatatagacctgaaatggctag
ctggacatgcgatcgaatcgagattagcaccatagagtggcata
gccatgcgctgatagcaaaatgccatagctagtgtctaacgtgca
ttgccctggatgacatggctccgatatggcggctgatcgtcgctga
aatgctcgctgcaatggctaggatacagtaatagacgtaatgcc
aatggctgctcgctggatagtcgctgacatcgatcgcctgatatga
tgcgctagctccgcataagatcgctgatcgcta……..

Genetic Code

		2nd base					
		U	C	A	G		
1st base	U	UUU *Phe* UUC *Phe* UUA *Leu* UUG *Leu*	UCU *Ser* UCC *Ser* UCA *Ser* UCG *Ser*	UAU *Tyr* UAC *Tyr* UAA Stop UAG Stop	UGU *Cys* UGC *Cys* UGA Stop UGG *Trp*	U C A G	3rd base
	C	CUU *Leu* CUC *Leu* CUA *Leu* CUG *Leu*	CCU *Pro* CCC *Pro* CCA *Pro* CCG *Pro*	CAU *His* CAC *His* CAA *Gln* CAG *Gln*	CGU *Arg* CGC *Arg* CGA *Arg* CGG *Arg*	U C A G	
	A	AUU *Ile* AUC *Ile* AUA *Ile* AUG *Met*	ACU *Thr* ACC *Thr* ACA *Thr* ACG *Thr*	AAU *Asn* AAC *Asn* AAA *Lys* AAG *Lys*	AGU *Ser* AGC *Ser* AGA *Arg* AGG *Arg*	U C A G	
	G	GUU *Val* GUC *Val* GUA *Val* GUG *Val*	GCU *Ala* GCC *Ala* GCA *Ala* GCG *Ala*	GAU *Asp* GAC *Asp* GAA *Glu* GAG *Glu*	GGU *Gly* GGC *Gly* GGA *Gly* GGG *Gly*	U C A G	

Crick's 1957 Hypothesis

The genetic code has excellent information theoretic properties, it is

comma free

It does not admit ANY form of parasitism.

Dismissed for the past 35 years

Replaced by “Frozen Accident”

- Renewed interest in comma free and circular codes (DNA computing, Arques/ Michel)

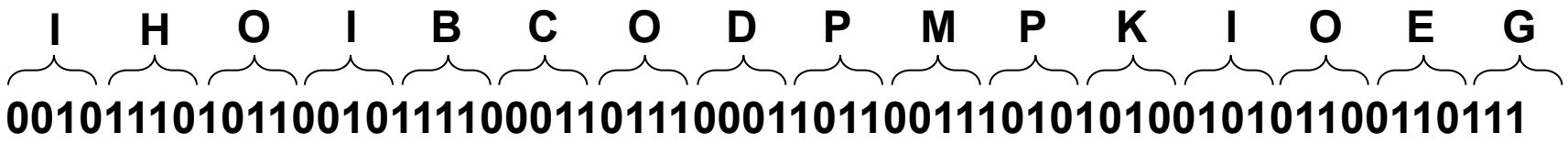
- Time to revisit

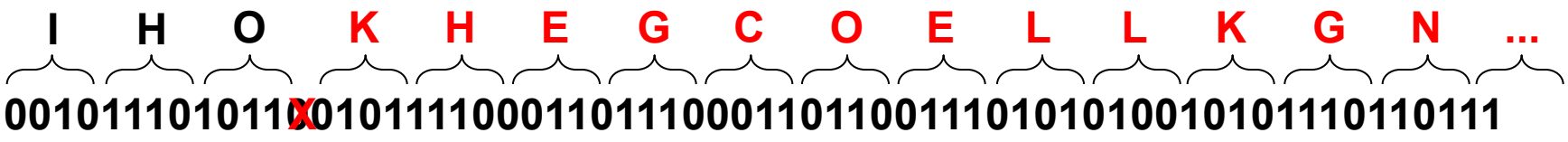
Coding

0000 = A
1111 = B
0001 = C
1000 = D
0011 = E
1100 = F
0111 = G
1110 = H

0010 = I
0100 = J
0101 = K
1010 = L
1001 = M
0110 = N
1011 = O
1101 = P

Communication Error

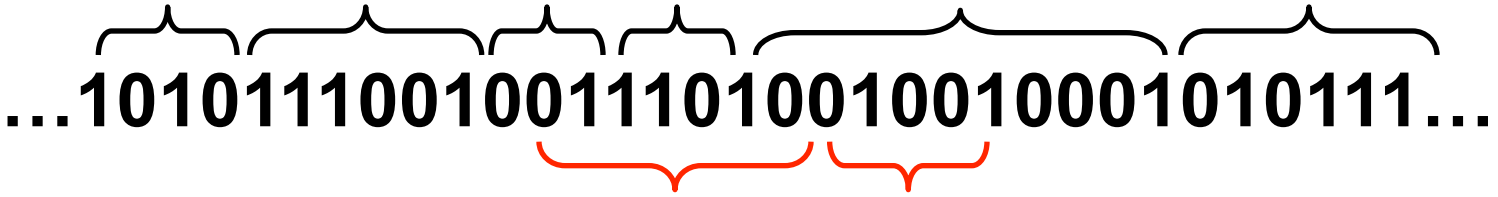




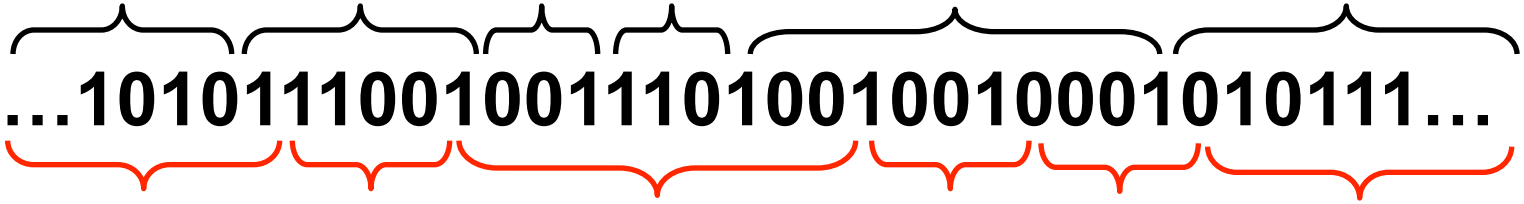
Translation Error
Frameshift

Parasite sub Messages

Bounded Parasitism:



Spread Parasitism:



Biological Implications of comma free

A frameshift will immediately abort the translation

ANY fragment of length 5 in the coding region of ANY gene in ANY organism determines the frame

Universal Frame property

Crick's Hypothesis Revisited

What is the length of the shortest segment of a coding region that defines the frame independently of the organism it comes from?

IF IT EXISTS

Mathematical Concepts

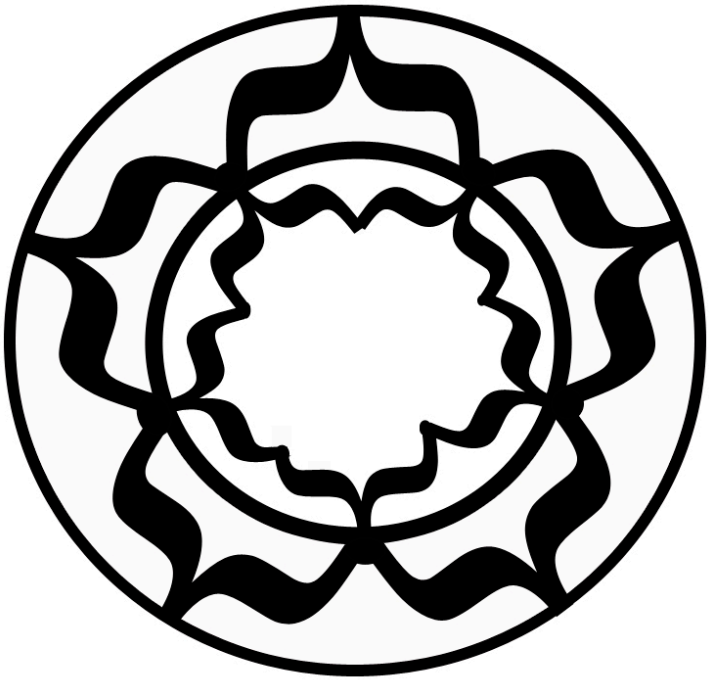
Comma Free Codes

Codes with Bounded and Spread Parasitism

Circular codes

Locally Testable Languages

Similarity Measures



A Circular Code

1

01

001

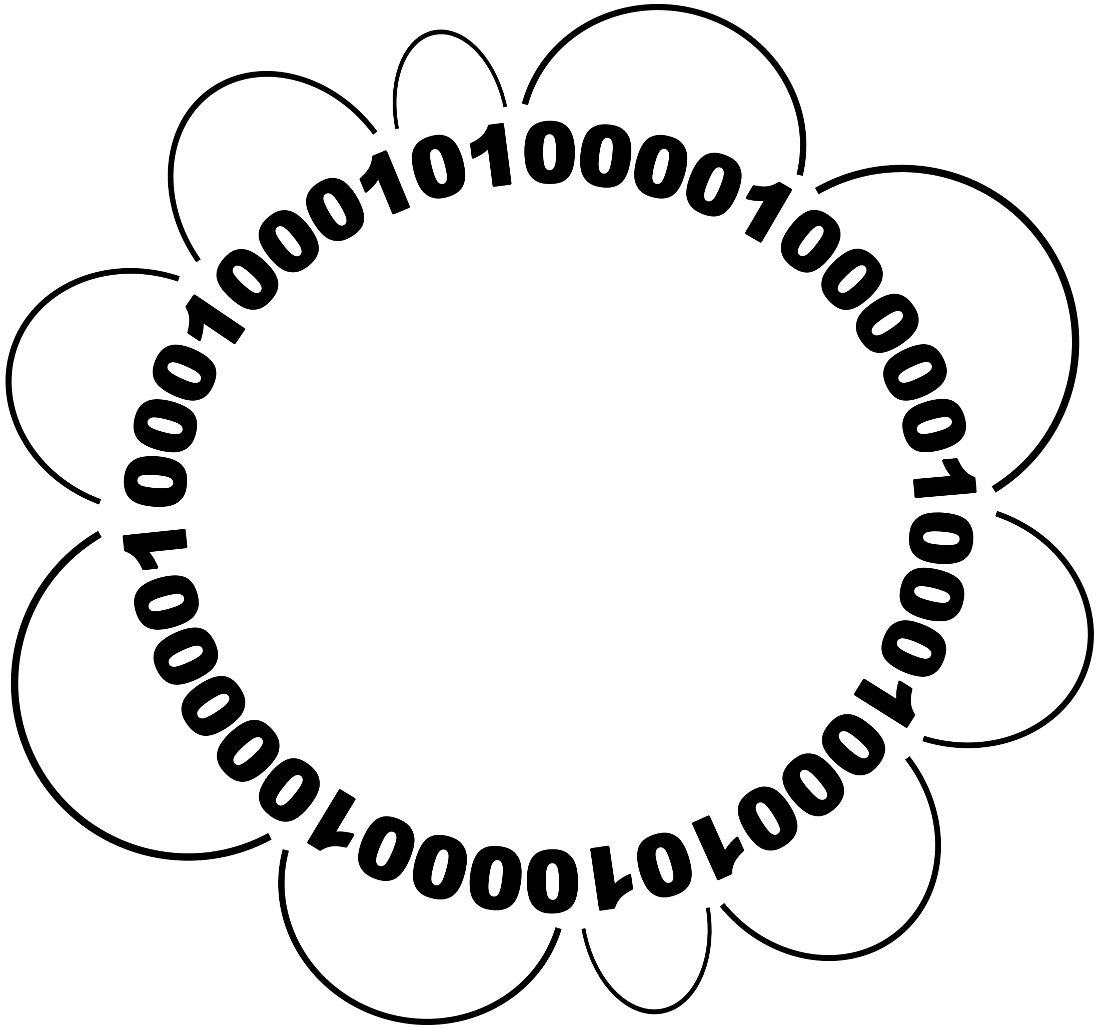
0001

00001

000001

0000001

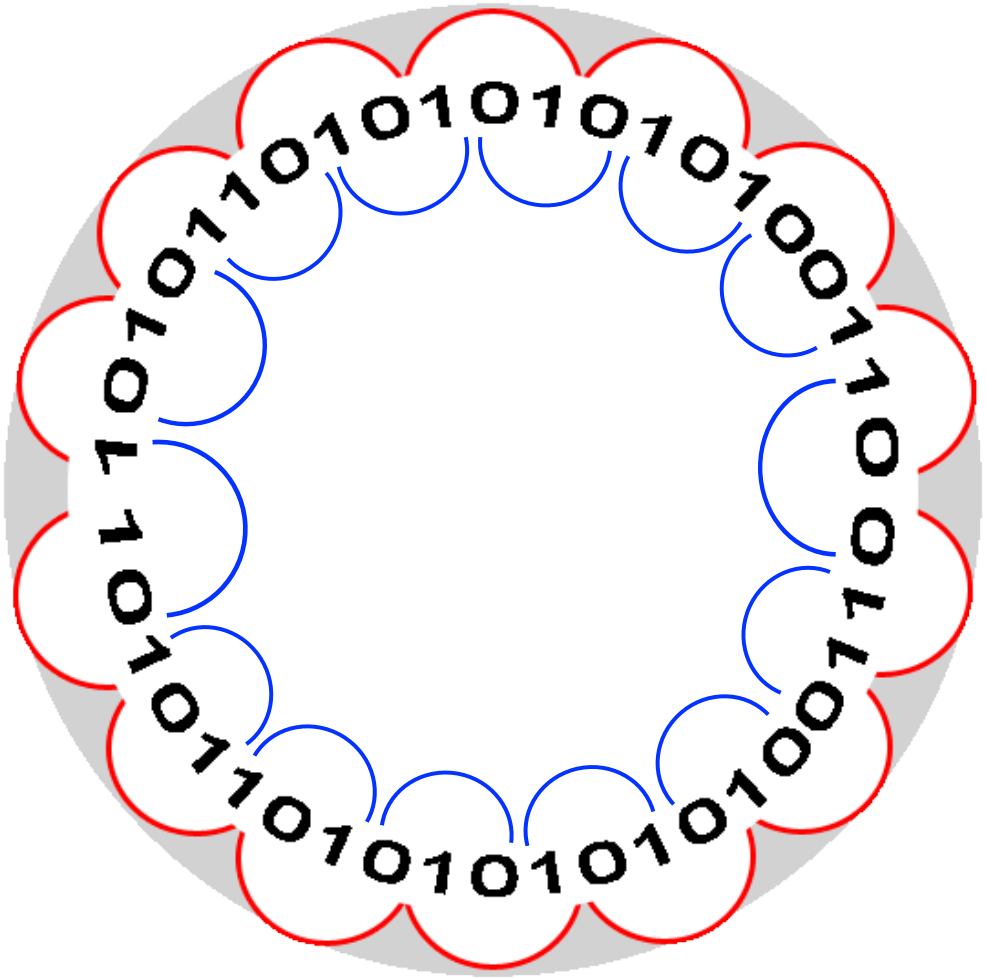
Unique Decomposition



A Non Circular Code

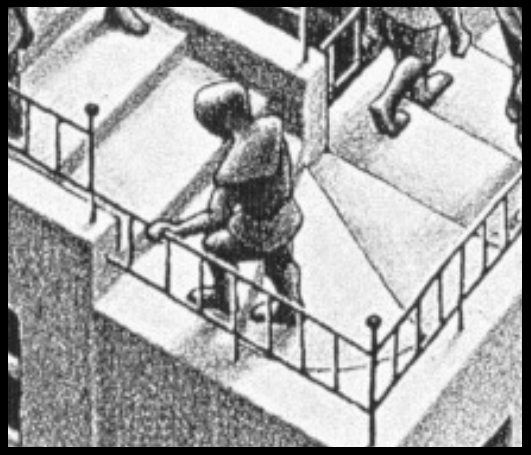
000
111
001
100
011
110
101
010

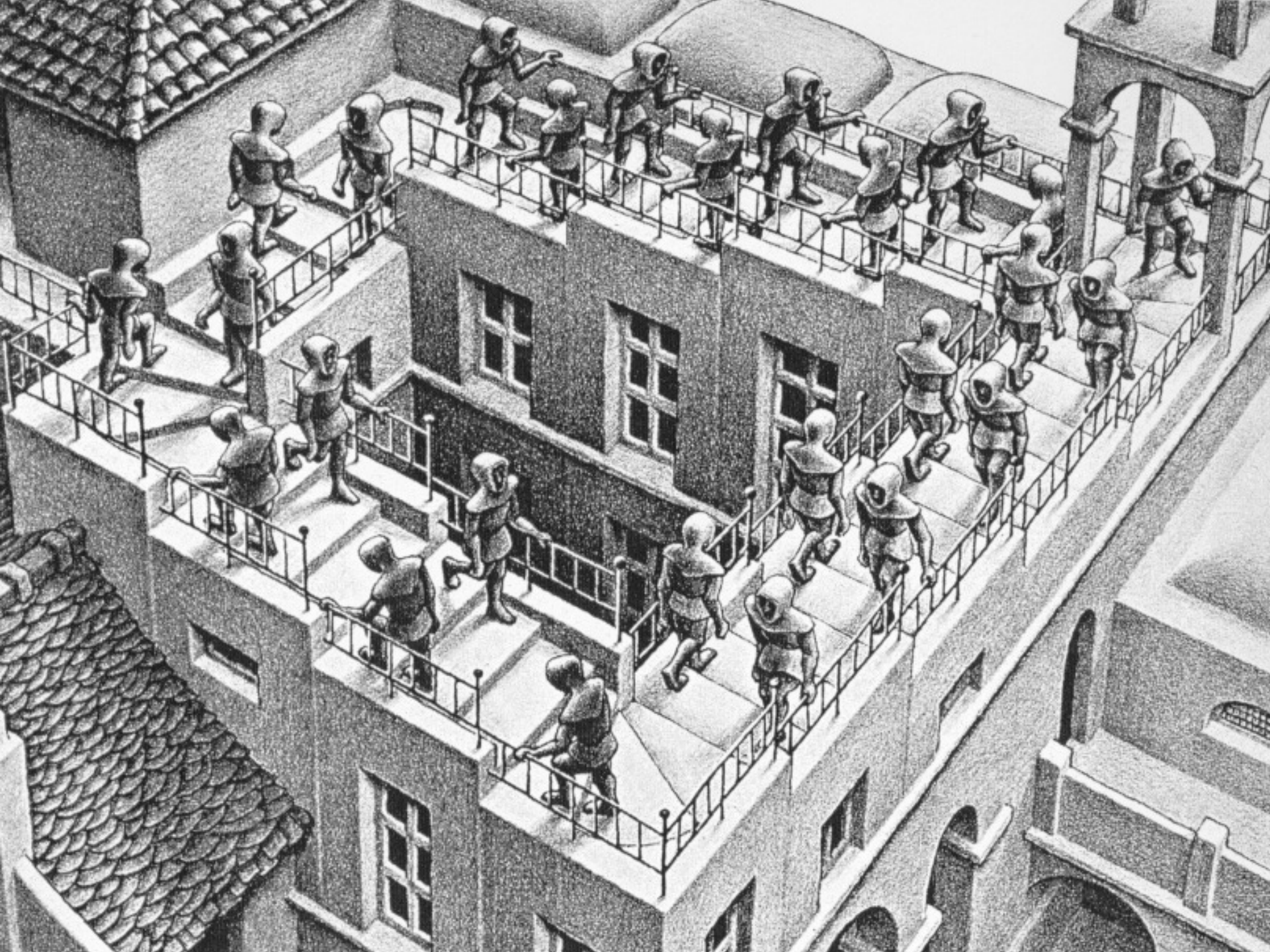
Multiple Possible Decompositions

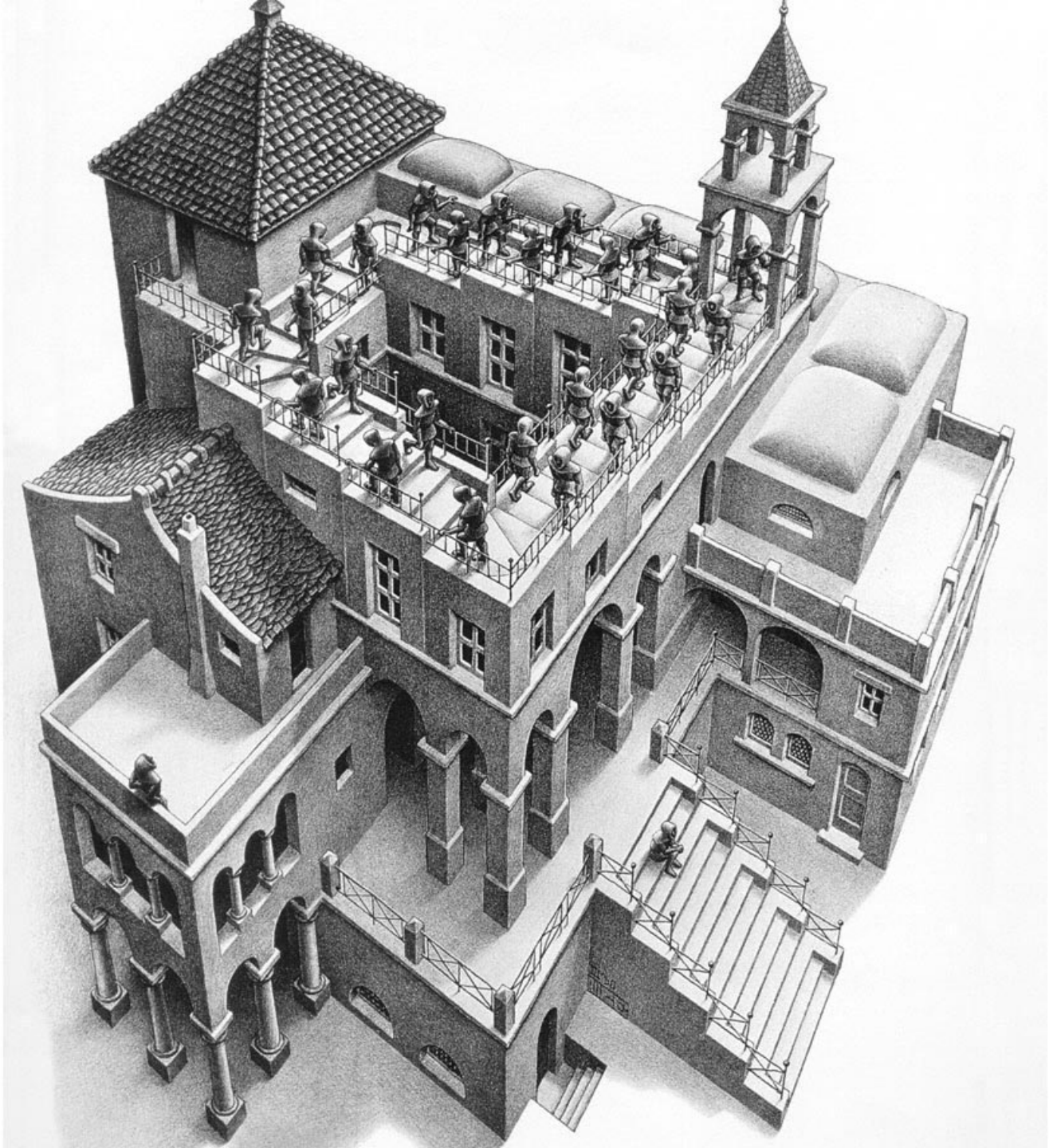


Locally Testable Events

$$\Sigma^* / \Sigma^* 0101 \Sigma^*$$







Theorem

Assumption: code X consists of a finite set of words all of the same length

The following are equivalent:

X has bounded parasitism of degree d

X^{d+1} is comma free

X is circular

X^* is strictly locally testable

Crick's Hypothesis Revisited Again

Genetic code C

Language of Genes $G \neq C^*$

C has good properties then G has good properties
BUT
G may have good properties while C does not.

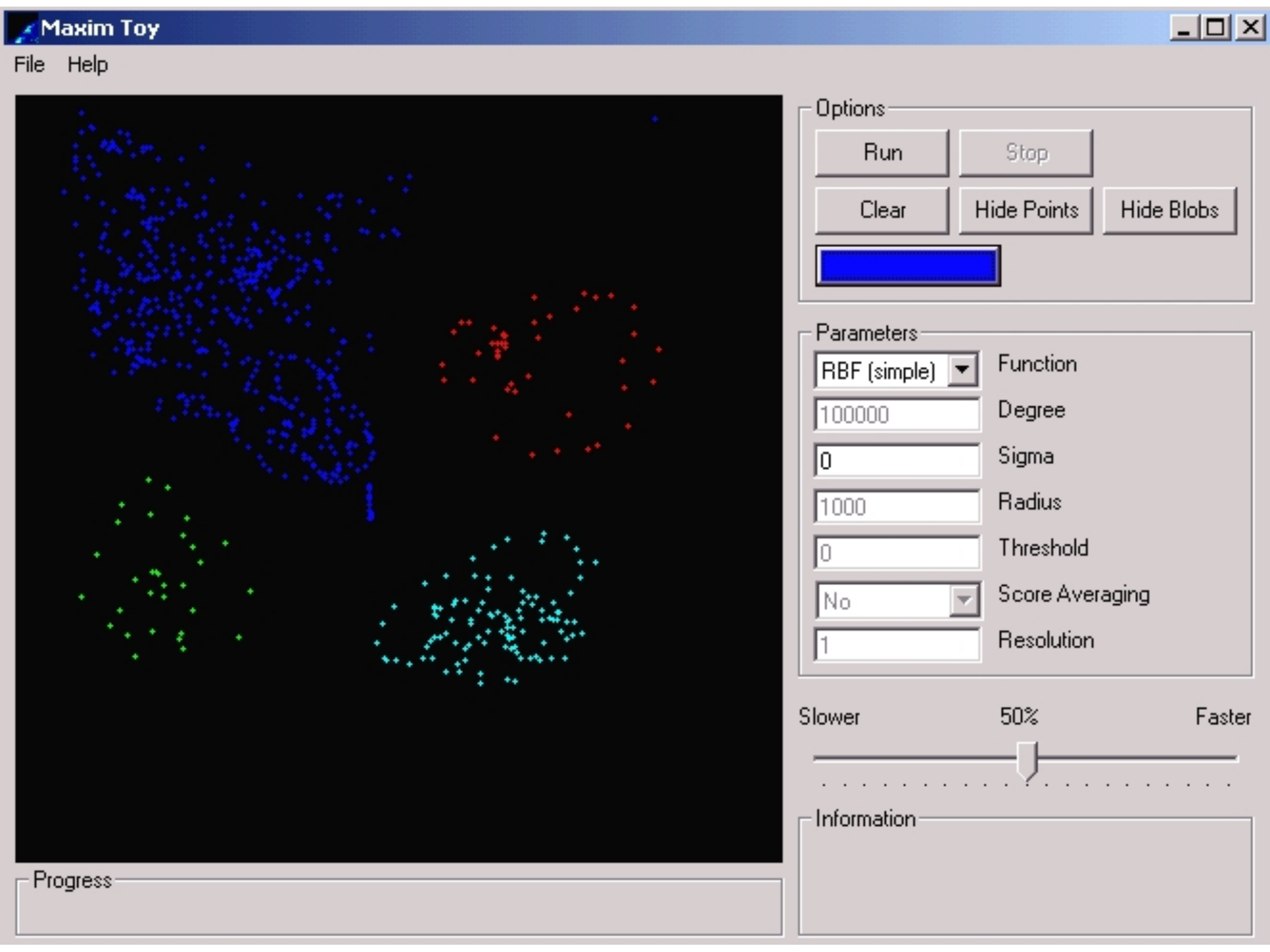
Shift from comma free to Testable by fragments

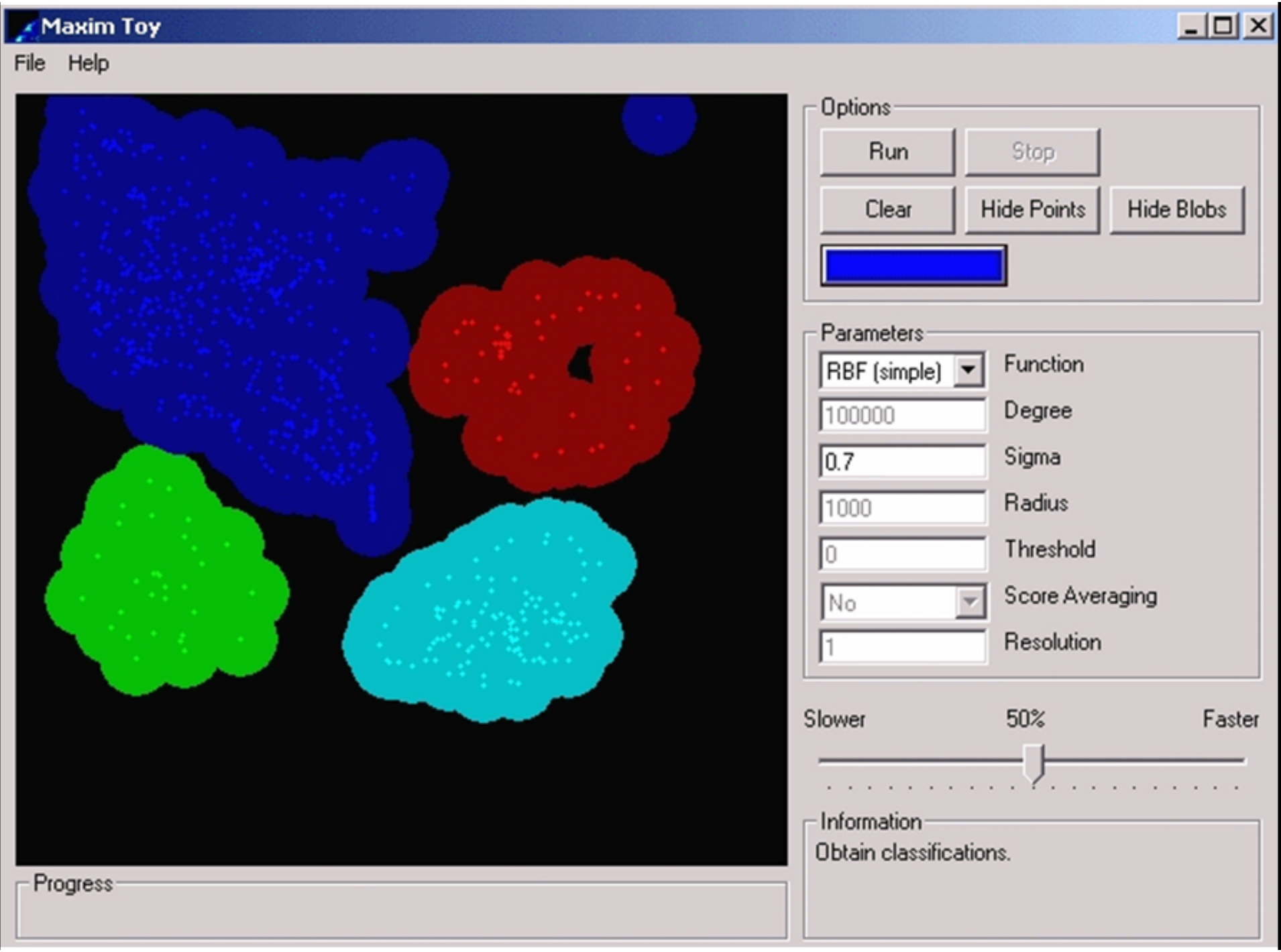
Similarity

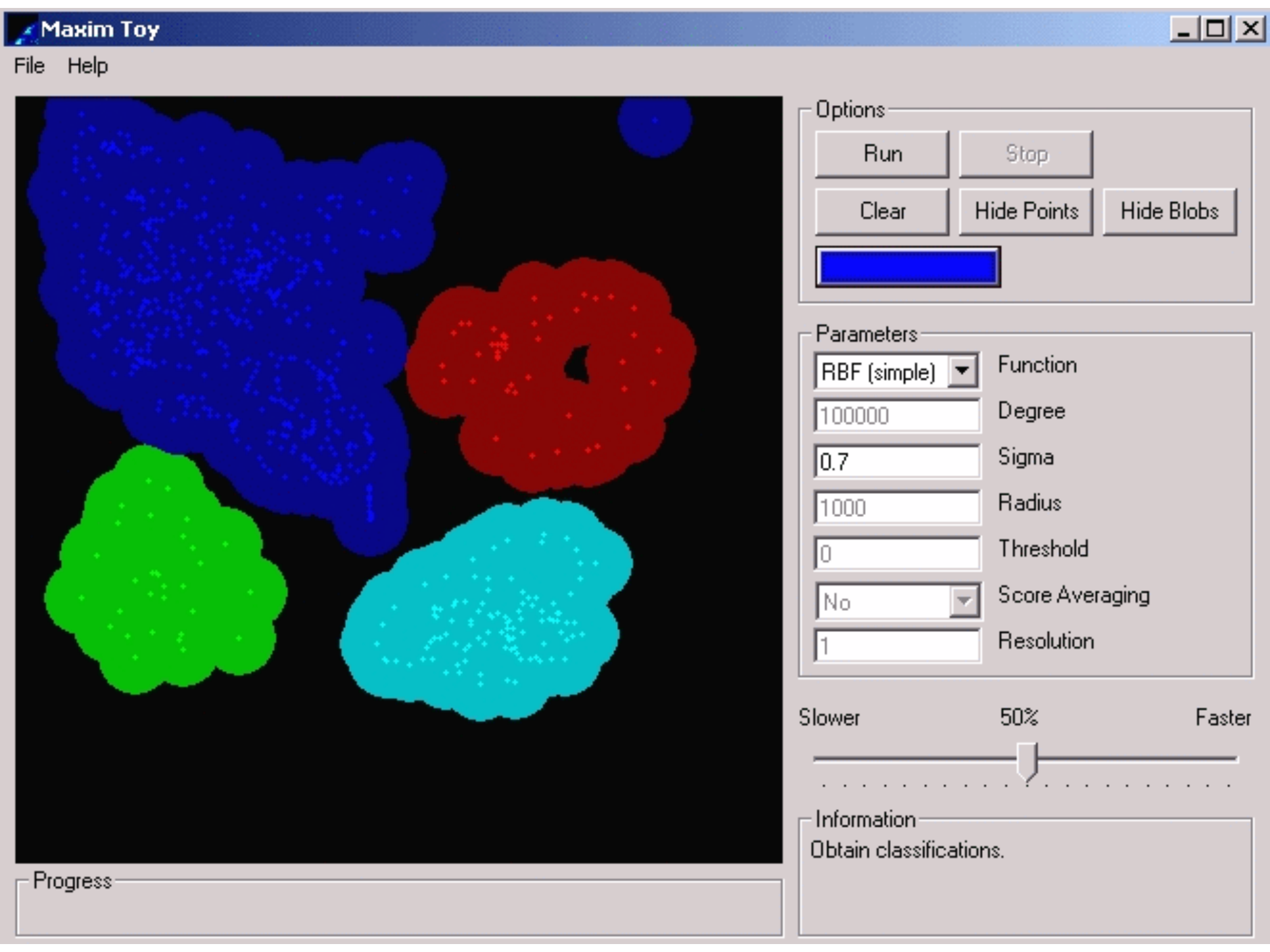
$$S(X,Y) = e^{-\frac{\|X-Y\|^2}{2\sigma^2}}$$

$$S_C(X) = \sum_{X_u \in C} S(X_u, X)$$

$$\xi(X) = \arg c\{\max(S_C(X))\}$$


Maxim Toy
File
Help
Options
Run
Stop
Clear
Hide Points
Hide Blobs
Parameters
RBF (simple)
Function
100000
Degree
0
Sigma
1000
Radius
0
Threshold
No
Score Averaging
1
Resolution
Slower
50%
Faster
Information
Progress


Maxim Toy
File
Help
Options
Run
Stop
Clear
Hide Points
Hide Blobs
Parameters
RBF (simple)
Function
100000
Degree
0.7
Sigma
1000
Radius
0
Threshold
No
Score Averaging
1
Resolution
Slower
50%
Faster
Information
Obtain classifications.
Progress


Maxim Toy
File
Help
Options
Run
Stop
Clear
Hide Points
Hide Blobs
Parameters
RBF (simple)
Function
100000
Degree
0.7
Sigma
1000
Radius
0
Threshold
No
Score Averaging
1
Resolution
Slower
50%
Faster
Information
Obtain classifications.
Progress

Arques/Michel Codes 1998

$$T_0 = \{AAA, TTT\} \cup X_0$$
$$T_1 = \{CCC\} \cup X_1$$
$$T_2 = \{GGG\} \cup X_2$$

X_0 = {AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC, GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC}

X_1 = {ACA, ATA, CCA, TCA, TTA, AGC, TCC, TGC, AAG, ACG, AGG, ATG, CCG, GCG, GTG, TAG, TCG, TTG, ACT, TCT}

X_2 = {CAA, TAA, CAC, CAT, TAT, GCA, CCT, GCT, AGA, CGA, GGA, TGA, CGC, CGG, TGG, AGT, CGT, TGT, CTA, CTT}

T Representations

ATGGGCAAGTAA

Frame0: 1 0 1 2
Frame1: 2 2 2
Frame2: 2 2 0

Training set

- DKEYP-117 zebra fish gene.
- KEGG
- 10620 Nucleotides
- Length of windows 200 in T representation
- C is 1671 Windows (Coding frame)
- C++ 1670 Windows

First Experiment

Testing

Pyrococcus C_000861.1

Window Size: 200

Sigma	F Score	F+ Score	F++ Score
0.4	100%	0%	0%
0.2	100%	0%	0%
0.1	100%	0%	0%
0.01	100%	0%	0%
0.006	100%	0%	0%
0.0058	100%	0%	0%
0.005	100%	0%	0%
0.003	100%	0%	0%
0.002	100%	0%	0%
0.001	100%	0%	0%
0.0006	100%	0%	0%
0.0001	0%	0%	0%

- Consistent with Crick's hypothesis but for the size of the code.
- Comma-free code (words of length 600)

OR

- G is locally testable

- Robustness with respect to overfitting.

General Experiment
Data sets

- We selected 14 different organisms in all three families and extracted 50 genes from each (Ecoli, Pyrococcus, Anopheles gambiae….).
- 100 genes which were selected from KEGG, NCBI, Weizmann Institute (TP53, Atm, HIV, Breast cancer…).
- 1000 genes with various ranges of GC Contents (Center for Bioinformatics, UPenn).

Testing

Saccharomyces_cerevisiae CHR_II NC_001134

Window Size: 200

Sigma	F Score	F+ Score	F++ Score
0.4	100%	55.342%	39.53%
0.2	100%	55.342%	39.53%
0.1	100%	55.342%	39.316%
0.01	100%	54.701%	36.966%
0.006	100%	52.991%	34.188%
0.0058	100%	52.991%	33.761%
0.005	100%	52.778%	32.692%
0.003	100%	50%	30.983%
0.002	100%	45.94%	34.829%
0.001	100%	45.299%	34.402%
0.0006	100%	43.803%	33.974%
0.0001	0%	0%	0%

- Not Comma-free
- Maybe Bounded Parasitism/Circular
- It is testable by fragments

Testing

Streptococcus mutans
Window Size: 200

Sigma	F Score	F+ Score	F++ Score
0.4	100%	100%	0%
0.2	100%	100%	0%
0.1	100%	100%	0%
0.01	100%	100%	0%
0.006	100%	100%	0%
0.0058	100%	100%	0%
0.005	100%	100%	0%
0.003	100%	100%	0%
0.002	100%	100%	0%
0.001	100%	100%	1.02%
0.0006	100%	100%	1.02%
0.0001	0%	0%	0%

- Not Comma-free
- Not Bounded Parasitism/Circular
- Not Locally testable
- But it IS testable by fragments

Interpretation with respect to Crick's Hypothesis

- Existence of a universal coding frame
- Some families fit the local testability/ comma free /BP/circular
- Some families are more susceptible to alternative splicing still they are Testable by Fragments (within the coding sequence)

Strict Algorithm

$$\forall w \in F \quad w \in C / C^{++}$$

$$\exists w \in F^{++} \quad w \in C^{++} / C$$

Relaxed Algorithm

$$F_S^{++} < F_S > 50$$

$$\&$$

$$F_S^{+} - F_S \leq F_S - 50$$

General Results

- 95.4% success with Strict algorithm
- 94.8% success with Relaxed algorithm
- Distribution of failures (concentrated on some organisms)

- Support the Universal Frame Hypothesis
- Existence of underlying mathematical structures

Smallest fragment size

Relaxed Algorithm

fragment of size 10, window size 2

74% success

fragment of size 60, window size 25

90% success

- Keep testable by fragment
- Most probable

Universal Property

Using this gene we are able to find the frame of any other gene.

Human - TP53 Gene

ATGGAGGAGCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCA
GGAAACATTTTCAGACCTATGGAAACTACTTCCTGAAAACAACGTTCTGT
CCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACG
ATATTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAG
AATGCCAGAGGCTGCTCCCCGCGTGGCCCCTGCACCAGCAGCTCCTA
CACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCT
GTCCCTTCCCAGAAAACCTACCAGGGCAGCTACGGTTTCCGTCTGGGC
TTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTG
CCCTCAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGC
TGTGGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGCCATG
AGA.....

Universal Property

Ecoli – dgkA Gene

……..TCGAATAATACCACTGGATTCACCCGAATTATCAAAGCTTCC…..

Pseudomonas fluorescens – ahcY Gene

….TACGGCTGCCGTCACAGCCTGAACGACGCCATCAAGCGCGGC……..

Bos taurus – APOE Gene

……..GCTGGGGCCAGCGAGGGTGCCGAGCGCAGCTTGAGCGCCATC…

Sus scrofa - JAK2 Gene

……ATTGTAACTATTCATAAGCAAGATGGCAAAAGTCTGGAAAGC……

Pyrococcus – OT3 Gene

……CATAGCGTTAACCACTACACCAACAGCGTCGGCAAAATCCTC……

Methanococcus maripaludis – comE Gene

….TTTAACAATTACGCACCTATAACTACAGAACAACAACGTGAT…….

CONCLUSION

- Provided we extend the notion of Comma-Free to the related notion of Testable By Fragment

Crick's 1957 Hypothesis is vindicated:

- There exists a universal frame based on a mathematical model

Coding vs. Non Coding

Algorithm tells us the most likely coding frame under the assumption that we are in the coding region
Not suitable as such to analyze the non coding region.
Need to adapt and refine.

Non coding region contains pseudo genes, gene complements, hypothetical genes, other functional regions in %' UTR and 3' UTR…
Repeats, and apparently random sequences.
Nevertheless we ran an experiment (Augustus) …. 60 pb of transcription vs. translation