Calcolo e vita

Relazione Intrinseca

Vincenzo Manca
Università di Verona

Schema della Presentazione

- Prologo (8 slides)
- Parte I (16 slides)
- Parte II (40 slides, diagrammi e figure)
- Conclusioni (10 slides)

Prologo

Prima lezione di *Metodi Informazionali*

C. L. Bioinformatica (triennale – L31 Classe Informatica)

- Polimeri – Sequenze
- Membrane – Multinsiemi
- Interazioni – Grafi
- Evoluzioni – Alberi

19° Secolo: "Information everywhere"

Computer: Digitalizzazione della Matematica

Calcolo Universale 1936-1945

DNA: Digitalizzazione della Vita

Informazione Molecolare 1945-1953

Caos deterministico e processi random

Algoritmi Pseudo-random 1953-1975

Breve excursus storico

- Cibernetica (**Wiener**, **von Neumann**, **Turing**, **Shannon**) **I padri sono … (almeno) … 4**
  - Artificial Neural Networks → EDVAC → Brenner (Shroedinger) → Crick
  - Lindenmayer Systems
  - Cellular Automata → Artificial Life

A logical calculus of the ideas immanent in nervous activity
Bulletin of Mathematical Biology, 1943

- Bioinformatica

Leibniz, …, Babbage, … Post, …., Chomshy, Knuth, ….

  - Databases biologici
  - Algoritmi di allineamento (Smith-Waterman 1981, FASTA 1988, BLAST 1990)
  - Algoritmi di sequenziamento (Haemophilus 1995)
- Natural computing
  - Evolutionary algorithms (Optimization by Evolution)
  - Membrane computing (Multiset Rewriting → Biochemistry)
  - Collective emergence (ACO, BOIDS)
- Infobiotics (V. Manca – Infobiotics, Springer 2013 + Metabolic P Systems – Sholarpedia)
  - **Infosystemics** →
  - **Infogenomics** →
- Metabiology (Chaitin: ”Lo scandalo darwiniano”)

Vincenzo Manca

Infobiotics

The book presents topics in discrete biomathematics. Mathematics has been widely used in modeling biological phenomena. However, the molecular and discrete nature of basic life processes suggests that their logics follow principles that are intrinsically based on discrete and informational mechanisms. The ultimate reason of polymers, as key element of life, is directly based on the computational power of strings, and the intrinsic necessity of metabolism is related to the mathematical notion of multiset.

The switch of the two roots of bioinformatics suggests a change of perspective. In bioinformatics, the biologists ask computer scientists to assist them in processing biological data. Conversely, in infobiotics mathematicians and computer scientists investigate principles and theories yielding new interpretation keys of biological phenomena. Life is too important to be investigated by biologists alone, and though computers are essential to process data from biological laboratories, many fundamental questions about life can be appropriately answered by a perspicacious intervention of mathematicians, computer scientists, and physicists, who will complement the work of chemists, biochemists, biologists, and medical investigators.

The volume is organized in seven chapters. The first part is devoted to research topics (Discrete information and life, Strings and Genomes, Algorithms and Biorhythms, Life Strategies), the second one to mathematical backgrounds (Numbers and Measures, Languages and Grammars, Combinations and Chances).

ISSN 2194-7287
ISBN 978-3-642-36222-4

springer.com

Manca

Infobiotics

Vincenzo Manca

EMERGENCE, COMPLEXITY AND COMPUTATION ECC

Infobiotics

Information in Biotic Systems

Springer

La vita è *informazione rappresentata ed elaborata a livello molecolare. Nasce* quando sono disponibili molecole in grado di rappresentare informazione e processi informativi (polimeri e membrane).

La nozione (generale) di *calcolo simbolico* (*versus* calcolo *numerico/algebrico*) è emersa nel 20° secolo con lo studio dell'elaborazione formale ed automatica dell'informazione (Logica matematica e Calcolabilità).

Un aspetto cruciale evidenzia la relazione profonda calcolo-vita: l'esistenza di macchine di calcolo "universali" è basata su algoritmi di duplicazione simbolica (un programma è il "mirror" di una machina entro un'altra). Analogamente, la riproduzione postula meccanismi di duplicazione (ds DNA).

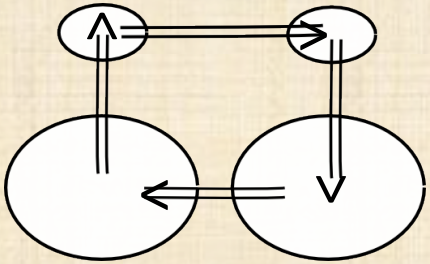
Calcolo-Vita

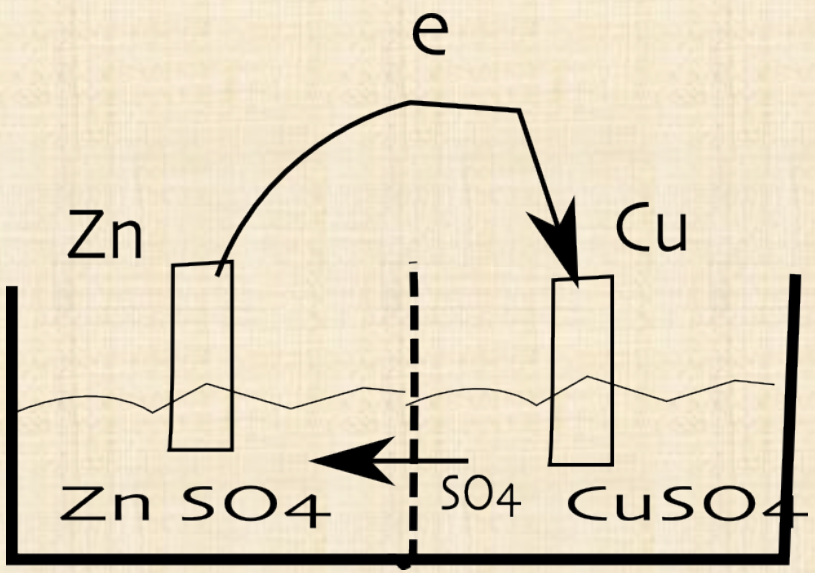
La vita ha suggerito i primi modelli di calcolo simbolico/automatico e la vita non può essere compresa se non si svelano i suoi intimi meccanismi di calcolo (elaborazione dell'informazione molecolare). Ne conosciamo solo alcuni e parzialmente: Codice genetico, Ribosomi trasduttori, …

Il calcolo è utile per elaborare dati (come in migliaia di altri campi), ma il suo rapporto con la vita ha radici e potenzialità di interazione ben più profonde e ancora in gran parte da scoprire.

La pila di Daniell (1996)

- Rimpiazziamo molecole con simboli
- Ricostruiamo il processo elettrochimico come movimento di simboli attraverso membrane
- (essenziale il passaggio da Post a Paun → MP)





Adleman: DNA Computing, 1994

Da molti anni sono convinto che sia uno scandalo non avere una dimostrazione matematica del fatto che l'evoluzione Darwiniana funziona.

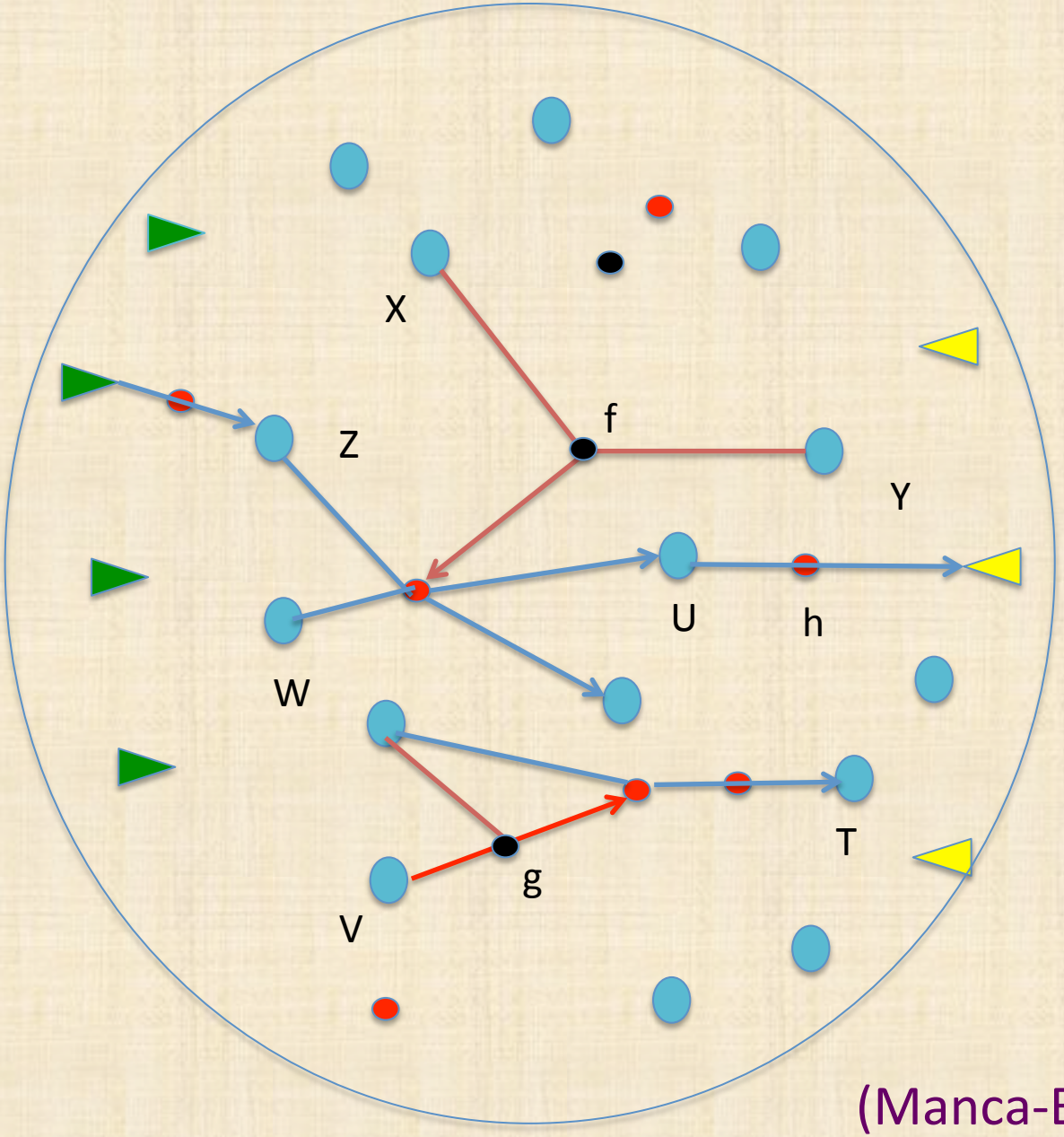
(Chaitin: 10-1-2011, Santa Fe Institute Conference)

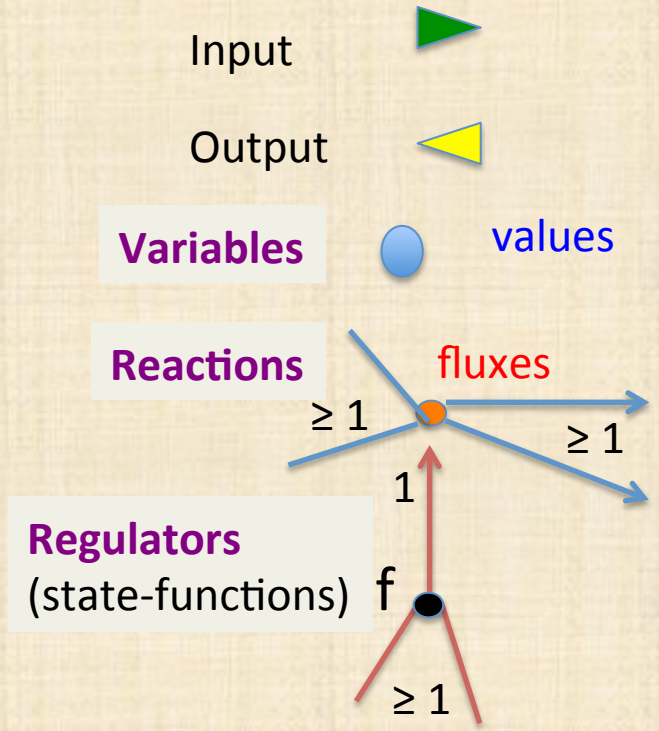
Teorema D ≈ Teorema H (Botzmann)?

PARTE PRIMA

Infosystemics

- **Networks** with two types of nodes and two types of edges (Reactions e Regulators)
- **Discrete** *versus* continuous (ODE) perspective
- Systems of **finite difference** (FDE) recurrent equations
- **Algorithmic** methods of dynamical analysis





MP Graph

(Manca-Bianco, BioSystems, 2007)

Variables, Reactions, Regulators ≈ Variables, Rules

$X_1, X_2, \ldots, X_n$

- The values associated to variables determine the state s of the system

- The dynamics is given by the sequence of states following a given initial state

MP grammars are the textual format of MP graphs

$r : \alpha \rightarrow \beta \; ; \; f$

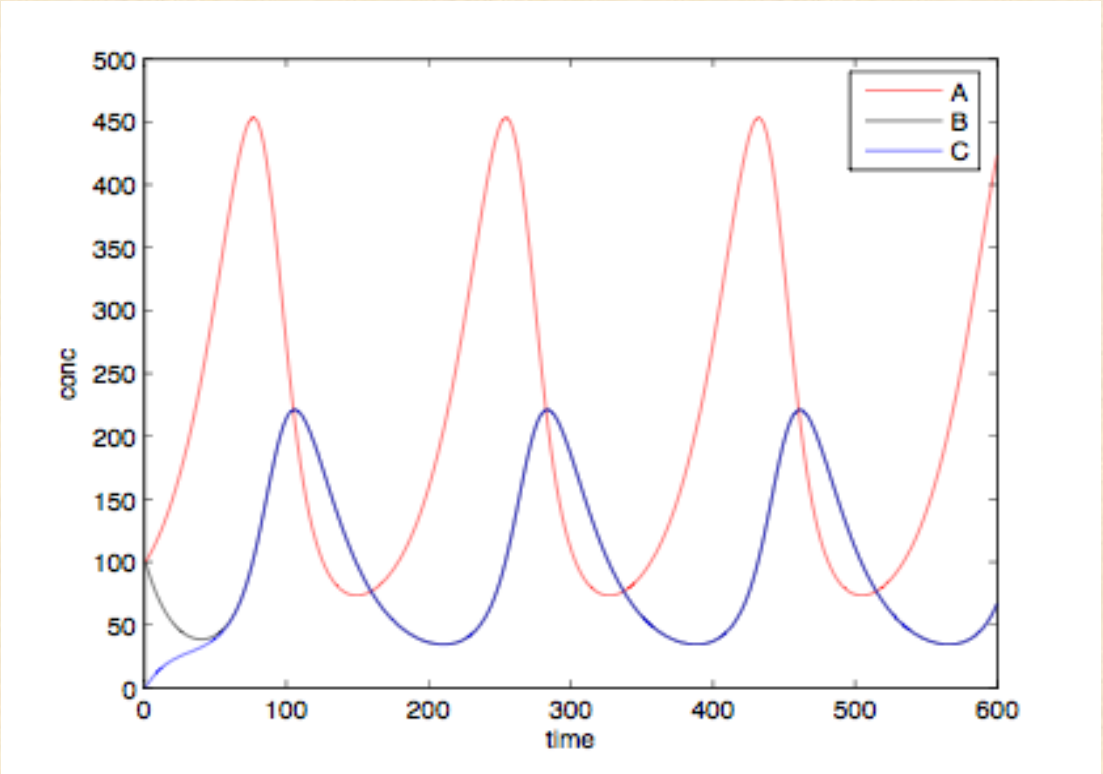
$r : 2X + Y \rightarrow 2Z + W \; ; \; 0.5XY^2$

MP grammars become **MP Systems** when The Initial State and time/mass parameters are fixed

Sirius MP Grammar and Dynamics

$0 \rightarrow A : \phi_1 = 4A/(0.02B + 0.02C + 100)$
$A \rightarrow B : \phi_2 = 0.02AC/(0.02B + 0.02C + 100)$
$A \rightarrow C : \phi_3 = 0.02AB/(0.02B + 0.02C + 100)$
$B \rightarrow 0 : \phi_4 = 4B/(4B + 100)$
$C \rightarrow 0 : \phi_5 = 4C/(4C + 100)$

A [0] = 100
B[0] = 100
C[0] = 0



Finite Difference Equations (MP)

(starting from an initial state)

- $\Delta x_1 = \rho[r_1,x_1]f_1(s) + \rho[r_2,x_1]f_2(s) + \ldots + \rho[r_m,x_1]f_m(s)$
  $- \lambda[r_1,x_1]f_1(s) - \lambda[r_2,x_1]f_2(s) - \ldots - \lambda[r_m,x_1]f_m(s)$

- $\Delta x_2 = \rho[r_1,x_2]f_1(s) + \rho[r_2,x_2]f(s)_2 + \ldots + \rho[r_m,x_2]f_m(s)$
  $- \lambda[r_1,x_1]f_1(s) - \lambda[r_2,x_1]f_2(s) - \ldots - \lambda[r_m,x_1]f_m(s)$

- - - - -

- $\Delta x_n = \rho[r_1,x_n]f_1(s) + \rho[r_2,x_n]f_2(s) + \ldots + \rho[r_m,x_n]f_m(s)$
  $- \lambda[r_1,x_n]f_1(s) - \lambda[r_2,x_n]f_2(s) - \ldots - \lambda[r_m,x_n]f_m(s)$

$$\Delta[s] = A \times U[s]$$

Dynamics Inverse Problem

- A dynamics is an object (or system of objects) changing its state in time: A motion in a n-dimesion space
- Let $\mathcal{S}$ be the sequence of states of an observed dynamics (time series)
- Can you find some internal rules (defined on the states) generating $\mathcal{S}$ when they are applied starting from the initial state of $\mathcal{S}$?
- When this is possible you gain knowledge on the phenomena underlying what you observed.

T =	1	2	3	4	5	6	7	……
x_1 =	0.1	0.2	0.3	0.25	0.15	0.4	0.5	……
x_2 =	0.03	0.02		0.07	0.18	0.32	0.34	0.25 ……
……								
x_n =	0.7	0.13		0.18	0.20	0.15	0.05	0.02 ……

$\mathbf{s}[i] = (x_1[i], x_2[i], \ldots x_n[i]$

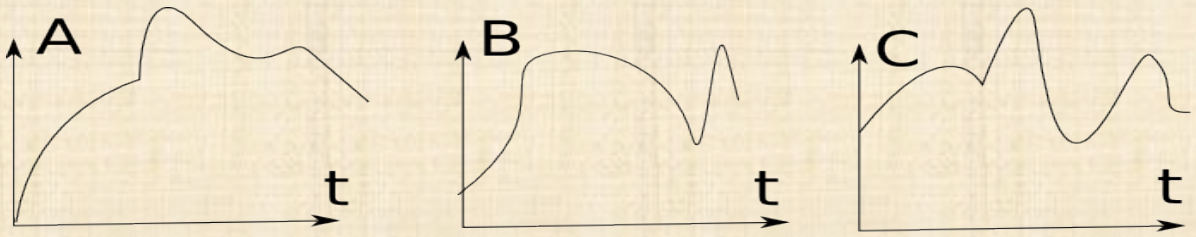
Find δ: $R^n \rightarrow R^n$ such that:

$$\mathbf{s}[i+1] = \delta(\mathbf{s}[i])$$

$$\delta^i(\mathbf{s}[0]) = (\mathbf{s}[i] \mid i = 0, 1, 2, \quad t\text{-}1)$$

MP theory provides algorithms for systematically solving DIPs

Time series

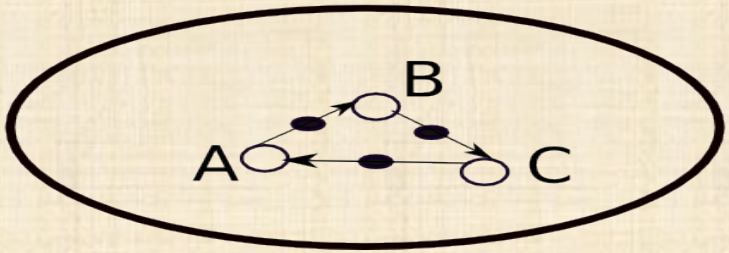


(No knowledge of forces is assumed)

MP- regression algorithms

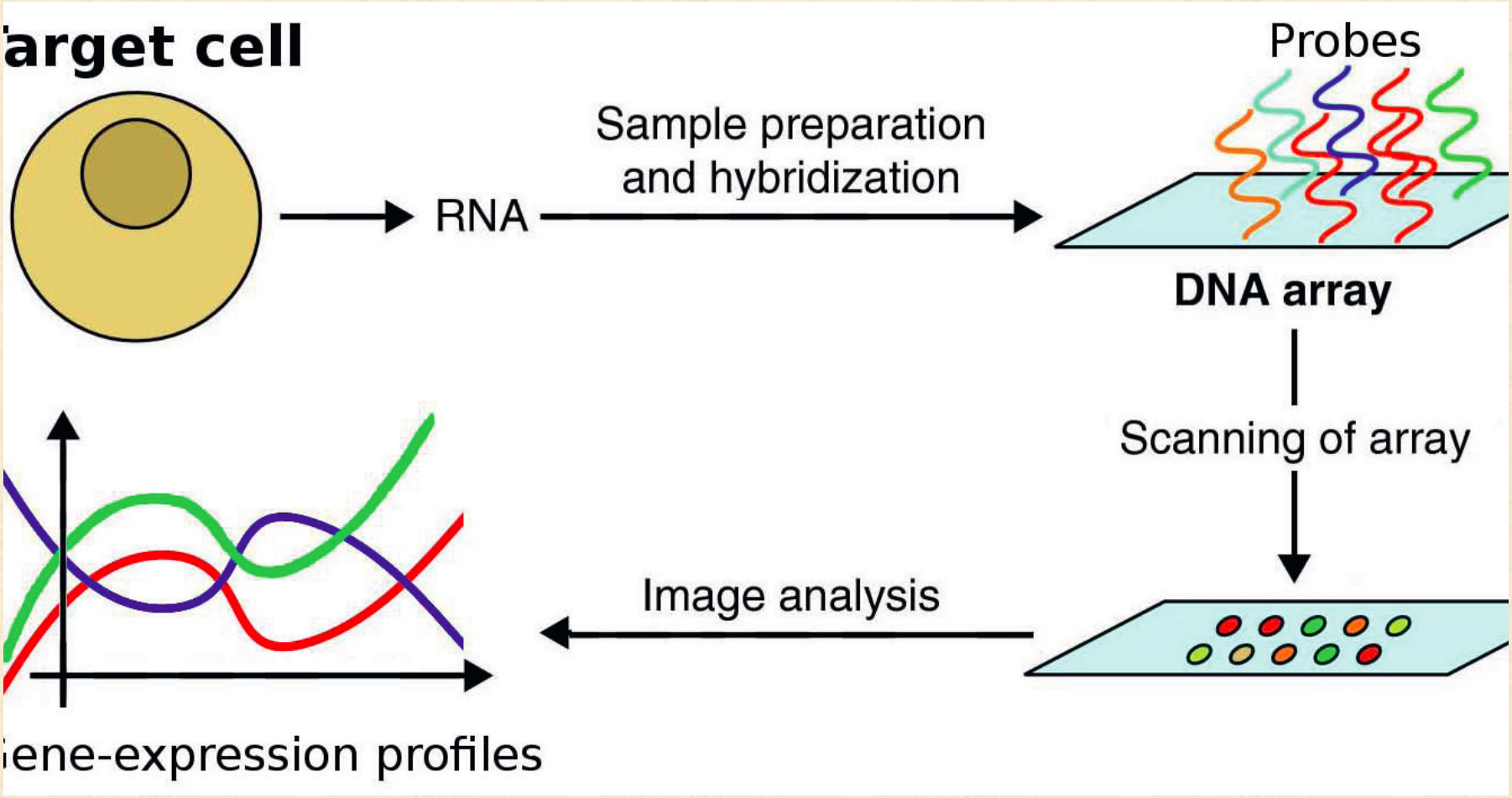
OLGA, MPANN, LGSS, MPSynth, MPTheory

MP Grammar



> 60 papers (2004-2014)

Manca, Bianco, Fontana, Franco, Petterlini, Castellini, Marchetti, Pagliarini, Zorzan, Lombardo
Theoretical C.S., BioSystems, Computer Mathematics, Natural Computing,
Molecular BioSystens, I.J. Nanotechnology and Biomolecular Dynamics, JACM,
Springer and Oxford University Press Volumes



L. Marchetti, V. Manca, R. Pagliarini, A. Bollig-Fischer.
MP Modelling for Systems Biology: Two Case Studies,
In: Applications of Membrane Computing in Systems and Synthetic Biology,
7:223-245, Springer (2014)

The first famous DIP was Halley's question to Newton:

H. – Why planets move on ellypses?

N. – Because I calculated it!

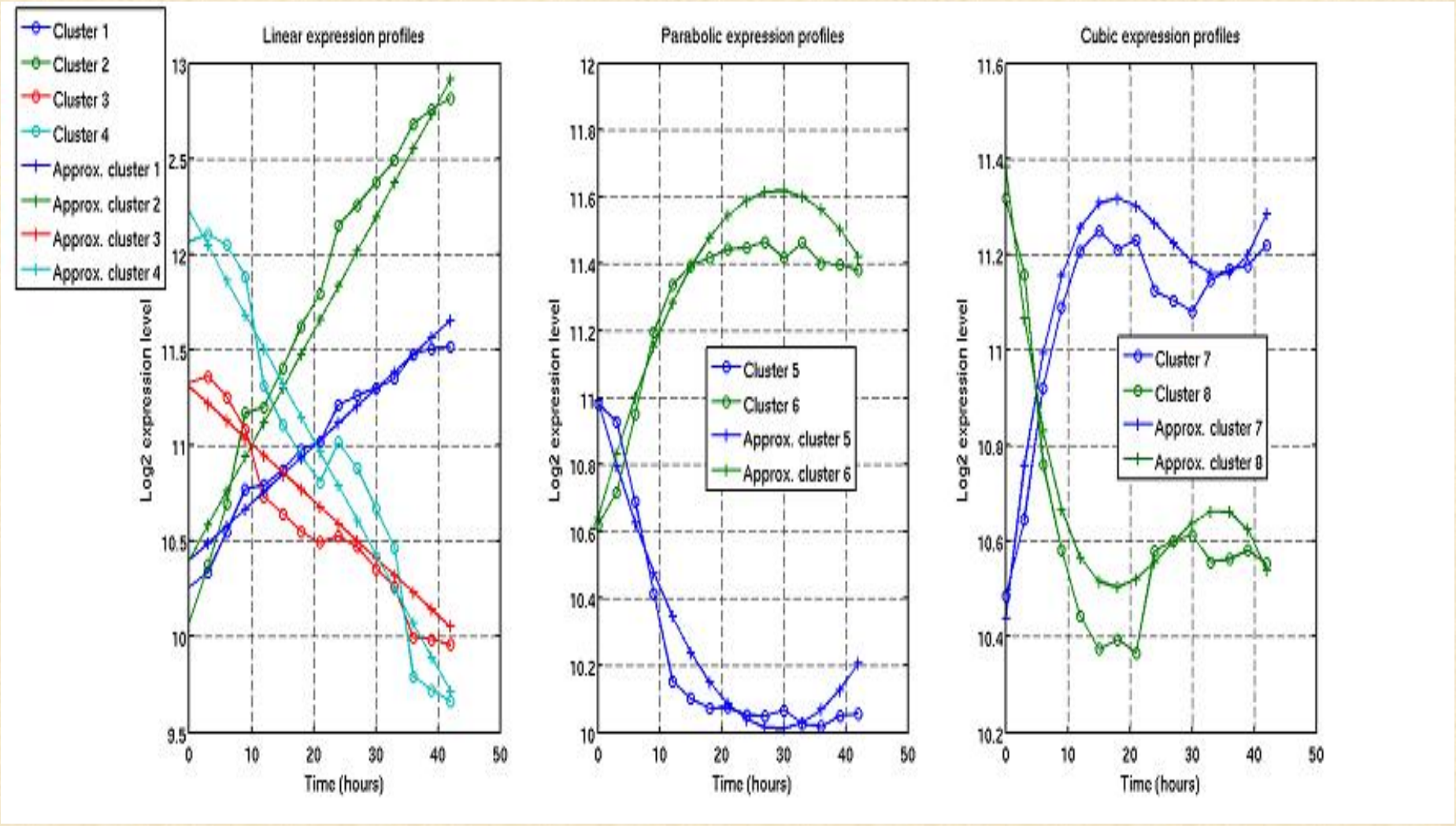
Newton's (with Leibnitz and Euler) solution was found via ODE, by expressing in differential terms dynamics laws under central gravitational forces.

But when you have thousands or millions of causes, force analysis isimpossible. Then, MP regulators replace causes with *dynamical forms* (regulators).

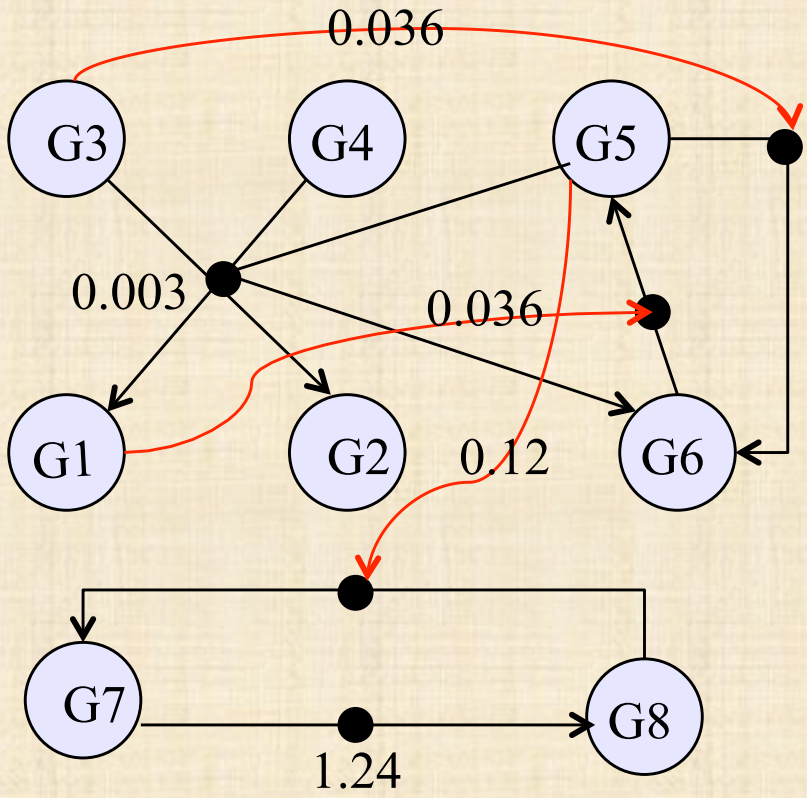
MP Analysis of **Breast Cancer Gene Expression**
Karmanos Cancer Institute (Detroit, USA)

- Consider one breast cancer cell,
- Apply to it an effect E that inhibits the cancer growth factor HER2 **RESISTANCE** !!!
- Measure the expressions of around 22000 genes during the effect of E,
- Collect 22000 time series along 16 time points.
- Filter and normalize data (P values, log scale)

A. Bollig-Fischer, L. Marchetti, C. Mitrea, J. Wu, A. Kruger, V. Manca, S. Draghici. Modeling time-dependent transcription effects of HER2 oncogene and discovery of a role for E2F2 in breast cancer cell-matrix adhesion. BIOINFORMATICS, 2014

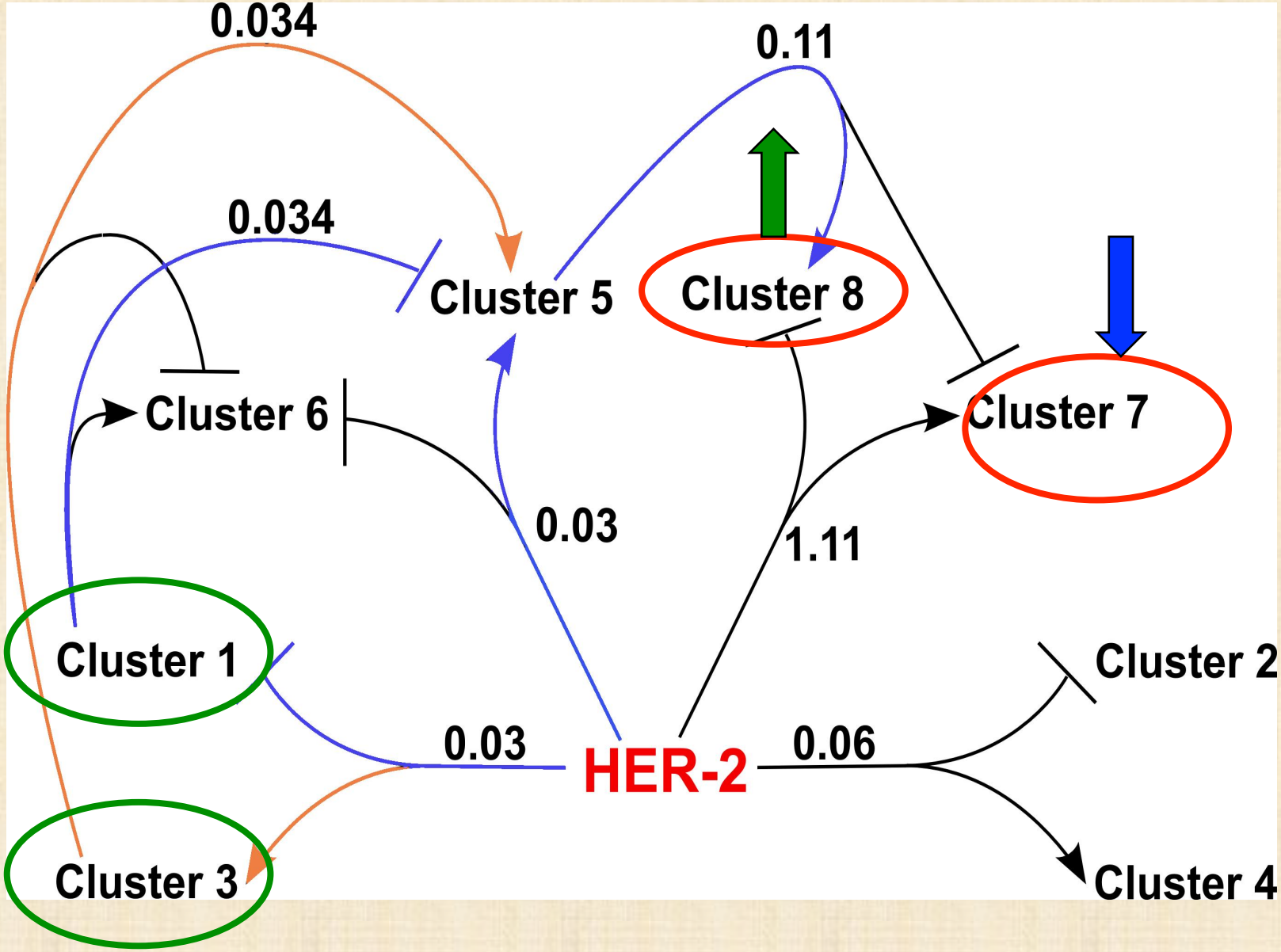

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Approx. cluster 1
Approx. cluster 2
Approx. cluster 3
Approx. cluster 4
Linear expression profiles
Log2 expression level
Time (hours)
Parabolic expression profiles
Cluster 5
Cluster 6
Approx. cluster 5
Approx. cluster 6
Cubic expression profiles
Cluster 7
Cluster 8
Approx. cluster 7
Approx. cluster 8

The MP grammar of Breast-E-HER2 genetic expression dynamics



Le teorie sono reti: solo chi le getta pesca

(Novalis, ripreso da Karl Popper a prologo della "Logica della scoperta scientifica")


0.034
0.11
0.034
Cluster 5
Cluster 8
Cluster 6
Cluster 7
0.03
1.11
Cluster 1
Cluster 2
0.03
HER-2
0.06
Cluster 3
Cluster 4

… and finally

They discovered two genes (??) such that their inhibition eliminates the resistance to HER2, by contrasting the cancerogenesis process.

MP Ingredients

- MP Formulation of **Dynamics Inverse Problems**
- Computational efficiency by **FDE** (deduced by MP Grammars) (FDE = Finite Difference Equations)
- Systematic discovery of regulators**: MP Regression Algorithms**

MP Software

http://mptheory.scienze.univr.it/
http://mplab.sci.univr.it/plugins/mpgs/index.html
http://mplab.sci.univr.it/
http://www.cbmc.it/software/Software.php

PARTE SECONDA

Infogemomics: informational genome annotation

From DNA Computing

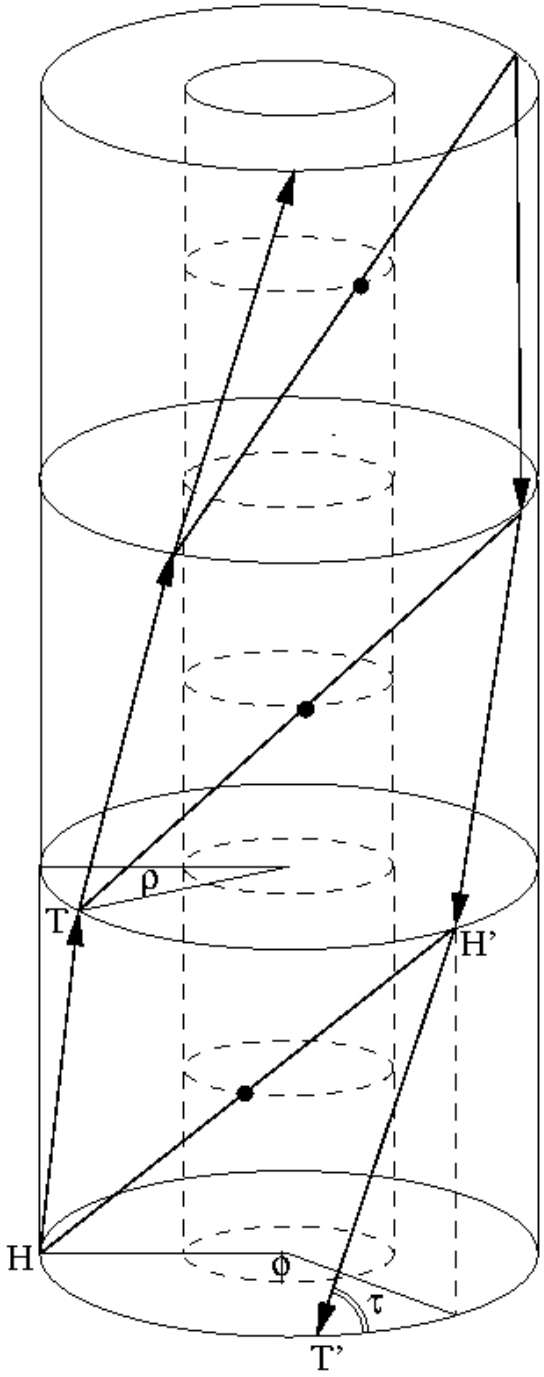
To Computing DNA

Manca, Franco, Castellini, Bonnici, Lombardo, Milanese

Mathematical Biosciences, Soft Computing, NACO, FI, Springer LNCS, BMC Genomics (2 PhD within 10 Phd in Infobiotics)

IG-Tools
V. Bonnici, V. Manca

www.infogenomic-explorer
A. Castellini, V. Manca



DNA Computing and Algorithmic DNA analysis

Efficient sequence duplication implies:

Bilinearity
Complementarity
Antiparalleslism

V. Manca – Fundamenta Informaticae - 2006

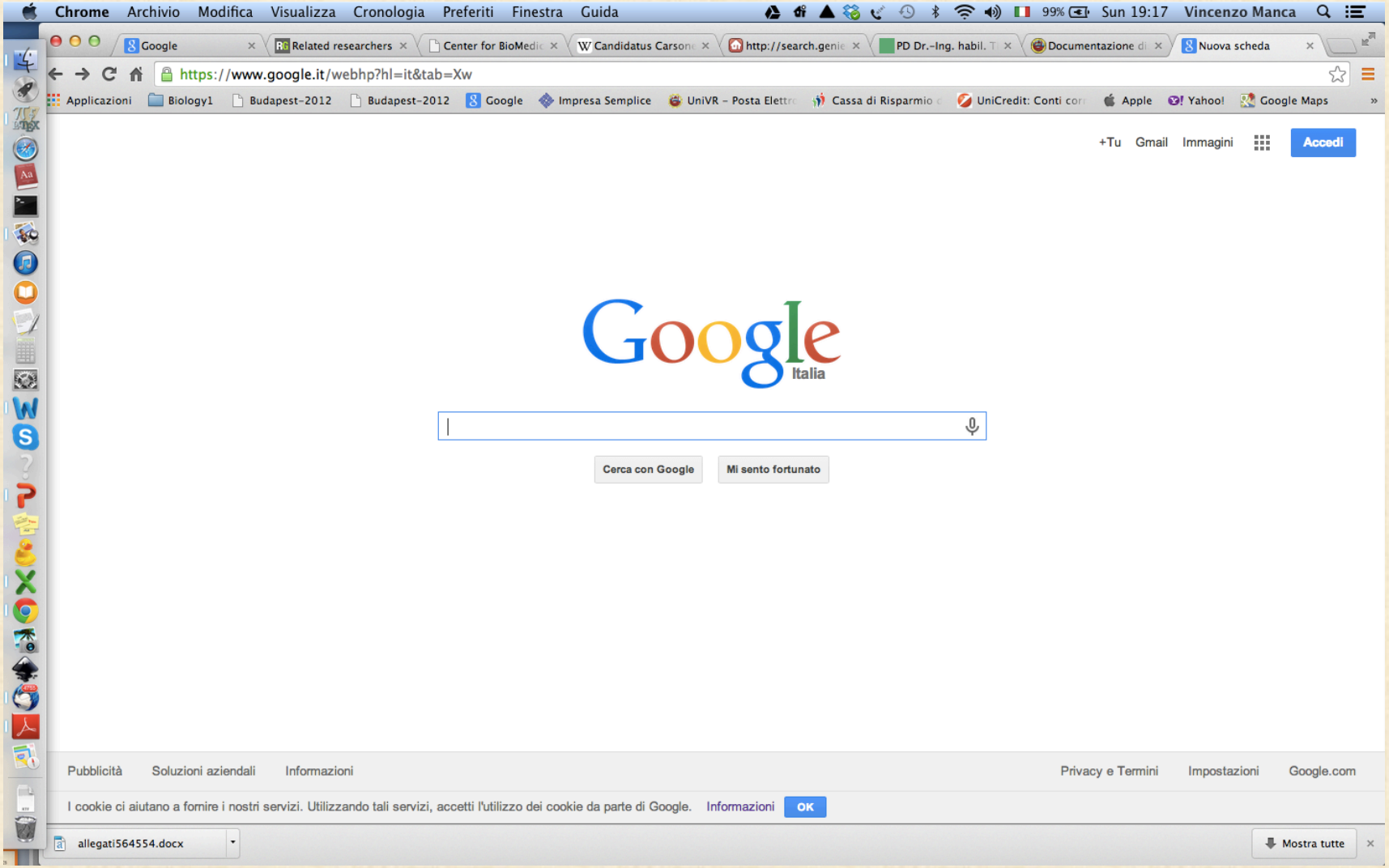
Could you recognize it?

- 3C 21 64 6F 63 74 79 70 65 20 68 74 6D 6C 3E 3C 68 74 6D 6C 20 69 74 65
- 6D 73 63 6F 70 65 3D 22 69 74 65 6D 73 63 6F 70 65 22 20 69 74 65 6D 74
- 79 70 65 3D 22 68 74 74 70 3A 2F 2F 73 63 68 65 6D 61 2E 6F 72 67 2F 57
- 65 62 50 61 20 67 65 22 3E 3C 68 65 61 64 3E 3C 6D 65 74 61 20 69 74 65
- 6D 70 72 6F 70 3D 22 69 6D 61 67 65 22 20 63 6F 6E 74 65 6E 74 3D 22 2F
- 69 6D 61 67 65 73 2F 67 6F 6F 67 6C 65 5F 66 61 76 69 63 6F 6E 5F 31 32
- 38 2E 70 6E 67 22 3E 3C 74 69 20 74 6C 65 3E 47 6F 6F 67 6C 65 3C 2F 74
- 69 74 6C 65 3E 3C 73 63 72 69 70 74 3E 77 69 6E 64 6F 77 2E 67 6F 6F 67
- 6C 65 3D 7B 6B 45 49 3A 22 63 48 42 31 55 4F 37 57 4B 63 54 47 73 77 61
- 37 35 34 48 67 42 41 22 2C 67 65 74 45 49 3A 66 75 20 6E 63 74 69 6F 6E
- 28 61 29 7B 76 61 72 20 62 3B 77 68 69 6C 65 28 61 26 26 21 28 61 2E 67
- 65 74 41 74 74 72 69 62 75 74 65 26 26 28 62 3D 61 2E 67 65 74 41 74 74
- 72 69 62 75 74 65 28 22 65 69 64 22 29 29 29 29 61 3D 61 2E 70 61 72 65
- 20 6E 74 4E 6F 64 65 3B 72 65 74 75 72 6E 20 62 7C 7C 67 6F 6F 67 6C 65
- 2E 6B 45 49 7D 2C 68 74 74 70 73 3A 66 75 6E 63 74 69 6F 6E 28 29 7B 72
- 65 74 75 72 6E 20 77 69 6E 64 6F 77 2E 6C 6F 63 61 74 69 6F 6E 2E 70 72
- 6F 74 6F 63 6F 6C 3D 20 3D 22 68 74 74 70 73 3A 22 7D 2C 6B 45 58 50 49
- 3A 22 33 31 32 31 35 2C 33 35 37 30 32 2C 33 37 31 30 32 2C 33 38 33 37
- 31 2C 33 39 35 32 33 2C 33 39 39 37 37 2C 33 33 30 30 30 32 35 2C 33 33
- 30 30 31 31 39 2C 33 33 30 30 31 32 34 2C 33 20 33 30 30 31 33 33 2C 33
- 33 35 2C 33 33 30 30 31 33 37 2C 33 33 30 30 31 33 30 30 31

Is this better? (the same information)

- <!doctype html><html itemscope="itemscope" itemtype="http://schema.or
- g/WebPage"><head><meta itemprop="image" content="/images/google_favi
- con_128.png"><title>Google</title><script>window.google={kEI:"cHB1UO7
- WKcTGswa754HgBA",getEI:function(a){var b;while(a&&!(a.getAttribute&&(
- b=a.getAttribute("eid"))))a=a.parentNode;return b||google.kEI},https:
- function(){return window.location.protocol=="https:"},kEXPI:"31215,357
- 02,37102,38371,39523,39977,3300025,3300119,3300124,3300133,3300135,33
- 00137,3300152,3310000,4000116,4000354,4000553,4000624,4000648,4000742,
- 4000833,4000879,4000955,4001064,4001131,4001145,4001188,4001192,400126
- 7,4001281,4001437,4001441,4001449,4001459,4001568",kCSI:{e:"31215,3570
- 2,37102,38371,39523,39977,3300025,3300119,3300124,3300133,33001,35,330
- 01373300152,3310000,4000116,4000354,4000553,4000624,4000648,4000742,40
- 00833,4000879,4000955,4001064,4001131,4001145,4001188,4001192,4001267,
- 4001281,400144001441,4001449,4001459,4001568",ei:"cHB1UO7WKcTGswa
  754Hg37,BA"},authuseml:funct r:0,

Genomes are not sequences but texts



```
\documentclass[11pt]{amsart} \usepackage{geometry}  \geometry{letterpaper}  \usepackage{graphicx}
\usepackage{amssymb} \usepackage{epstopdf}
\DeclareGraphicsRule{.tif}{png}{.png}{`convert #1 `dirname #1`/`basename #1 .tif`.png}
\title{On the numbers of life codes}
\author{Vincenzo Manca}  %\date{}    % Activate to display a given date or no date
\begin{document} \maketitle \section{Introduction}
In this paper we discuss and speculate about the following questions:
\begin{enumerate}
\item Why 4 different types of nucleotides? \item Why a 3-mer genetic code? \item Why 20 different amino acids?
\end{enumerate}        \hspace{2mm}        \ldots  \ldots \ldots  \ldots \\
As we will argue in the position of the related problems, these numbers are strongly related and intrinsically connected to basic principles in the informational organisation of organic matter.
\section{Representation levels}  \ldots  \ldots  \ldots  \ldots
\begin{equation} ax^2 + bx + c =0 \end{equation}
\begin{equation} x=-b \pm \frac{\sqrt{(b/2)^2-4ac}}{2a} \end{equation}
\begin{verbatim}(3) \begin{equation} x=-b\pm\frac{\sqrt{(b/2)^2-4ac}}{2a} \end{equation} \end{verbatim}
(4)$$10101101111000000000000111111111010101010001100100010100111001010101100000011$$
$$01100010100010101010101010101010101010000000000000011111010101110111001000001$$
$$10101000010101010101010101010101000000001111111111010010001011011101110101100 1$$
$$00011000010001101010101010101010000010001000001111110101011000110111011101110 1101$$
$$10111011001011100101010010100101010010101111000000000000000000011101000110011 0$$
$$01010101000000111111111110100100010011000010001000001110000011111000001111110$$
\section{Types of codes and redundant codes}  \ldots  \ldots  \ldots  \ldots \end{document}
```

ON THE NUMBERS OF LIFE CODES

VINCENZO MANCA

1. Introduction

In this paper we discuss and speculate about the following questions:

(1) Why 4 different types of nucleotides?
(2) Why a 3-mer genetic code?
(3) Why 20 different amino acids?

.....

As we will argue in the position of the related problems, these numbers are strongly related and intrinsically connected to basic principles in the informational organisation of organic matter.

2. Representation levels

.....

(1) $$ax^2 + bx + c = 0$$

(2) $$x = -b \pm \frac{\sqrt{(b/2)^2 - 4ac}}{2a}$$

(3) `\begin{equation} x=-b\pm\frac{\sqrt{(b/2)^2-4ac}}{2a} \end{equation}`

(4)

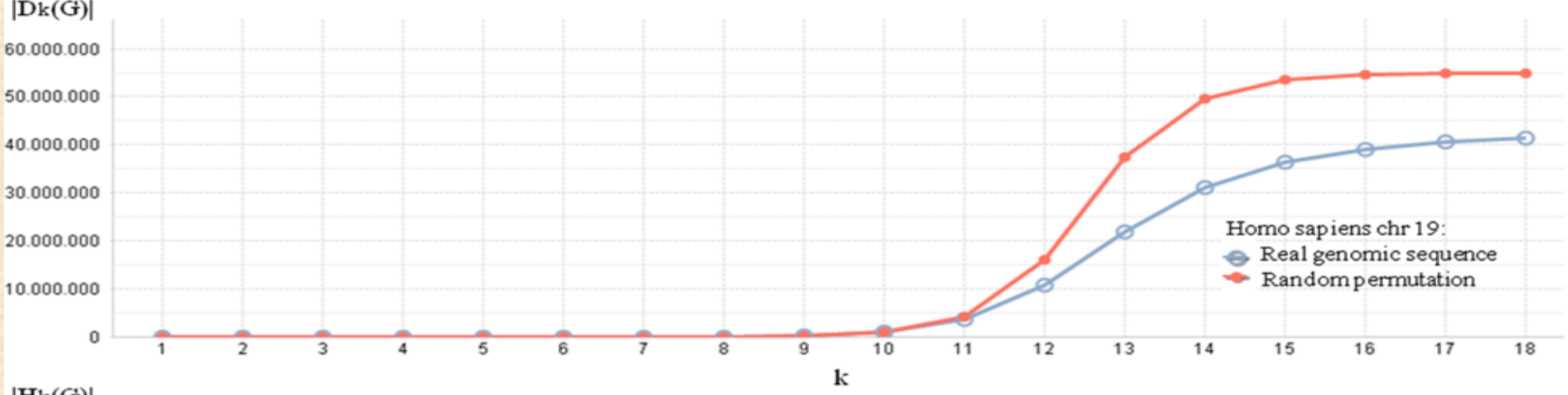
```
10101101111000000000001111111110101010100011001000101001110010101011000001101
01100010100010101010101010101010101010000000000000011111010101110111001000010
10101000010101010101010101010101000000001111111111010010001011011101110101100
10001100001000110101010101010100000100010000011111101010110001101110111011101
10111011001011100101010010100101010010101111000000000000000000011101000110011
01010101000000111111111110100100010011000010001000001110000011111000001111110
```

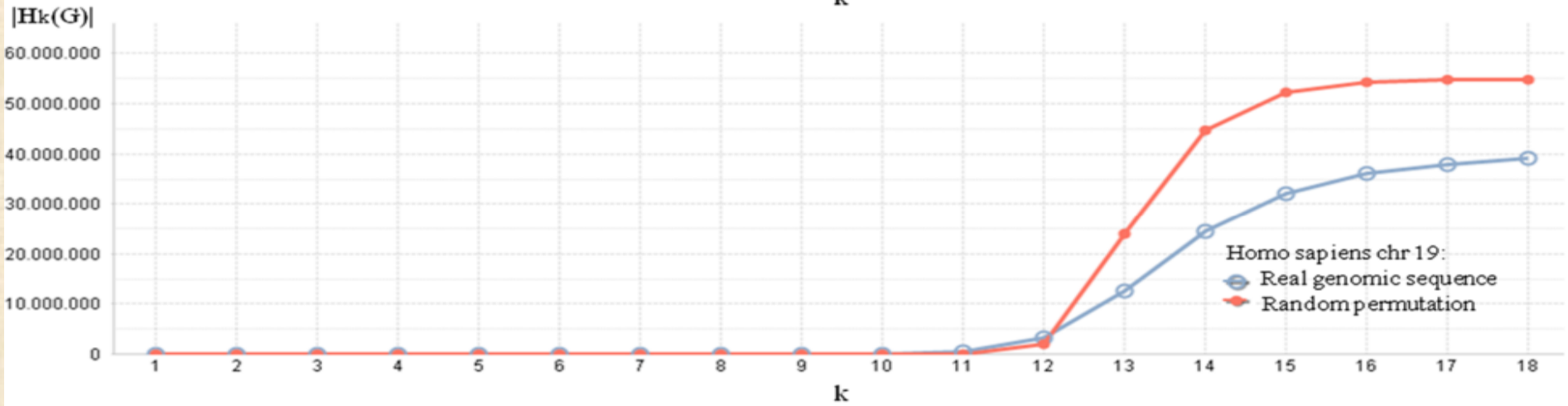
3. Types of codes and redundant codes

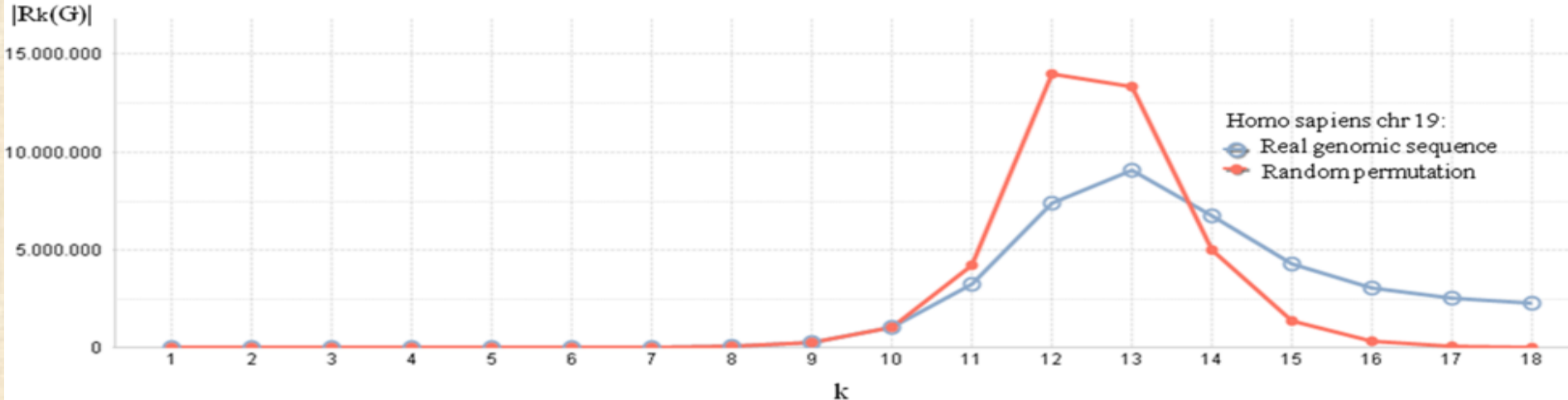
.....

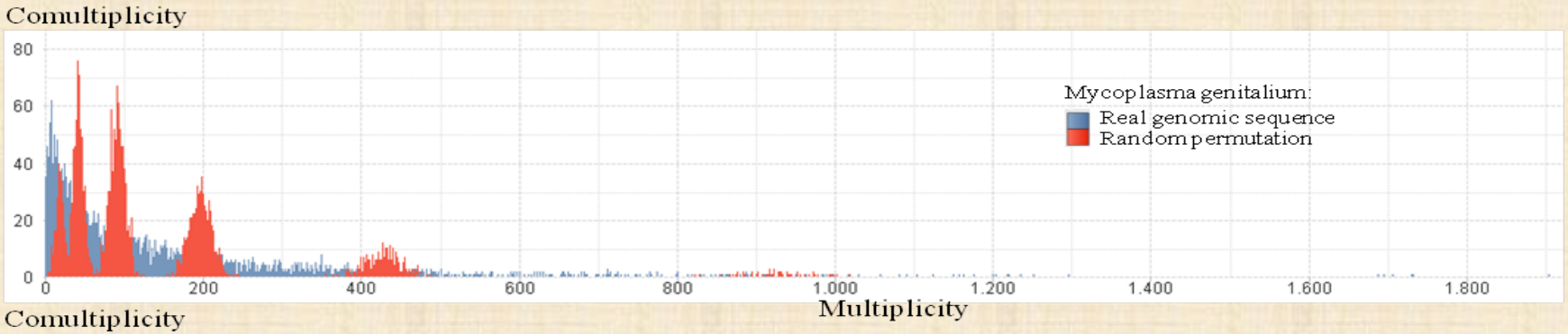
Some Informational Analyses
A systemic approach

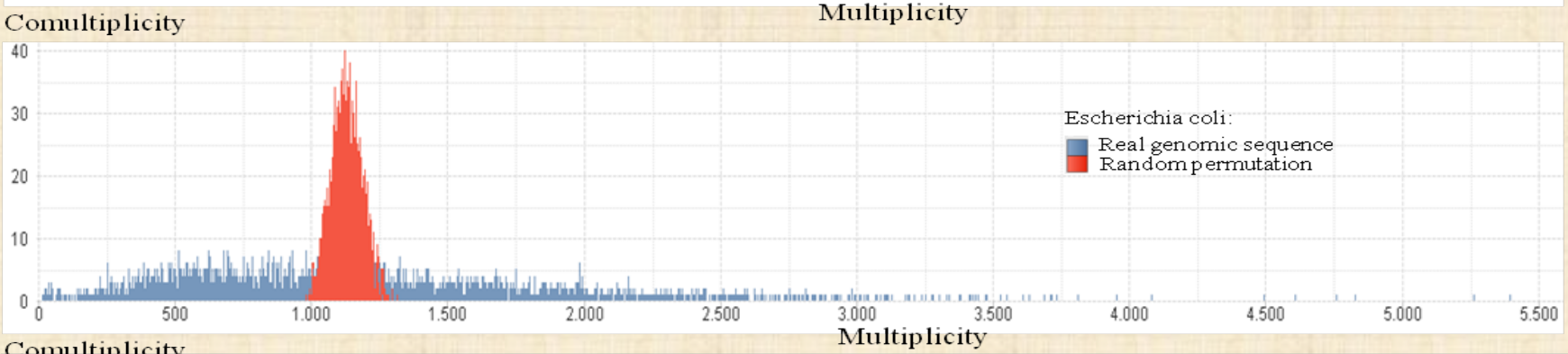
- Genome Distributions and Spectra
- Codon Occurrences and Recurrences
- Elongation Matrices
- Segmentation (3-mer minimal completeness) **
- Genomic Indexes (random-normalized)
- Genomic dictionaries
- Genomic Representations and Visualizations

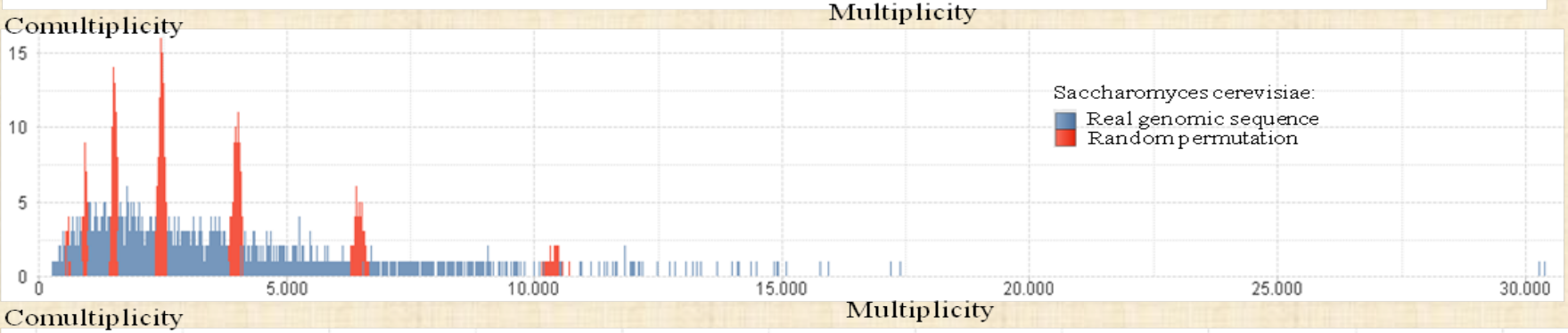

|Dk(G)|
60.000.000
50.000.000
40.000.000
30.000.000
20.000.000
10.000.000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
k
Homo sapiens chr 19:
Real genomic sequence
Random permutation

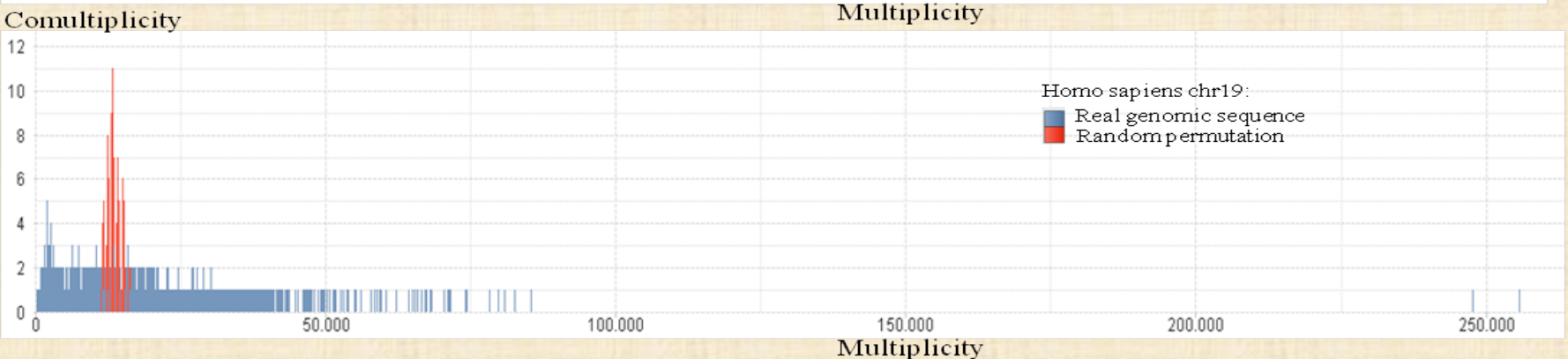

|Hk(G)|
60.000.000
50.000.000
40.000.000
30.000.000
20.000.000
10.000.000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
k
Homo sapiens chr 19:
Real genomic sequence
Random permutation


|Rk(G)|
15.000.000
10.000.000
5.000.000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
k
Homo sapiens chr 19:
Real genomic sequence
Random permutation

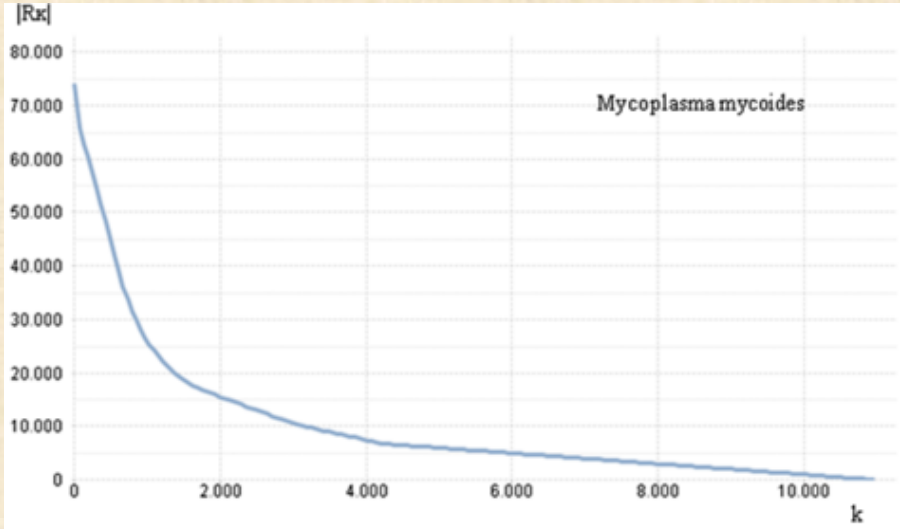

Comultiplicity
80
60
40
20
0
0
200
400
600
800
1.000
1.200
1.400
1.600
1.800
Multiplicity
Mycoplasma genitalium:
Real genomic sequence
Random permutation

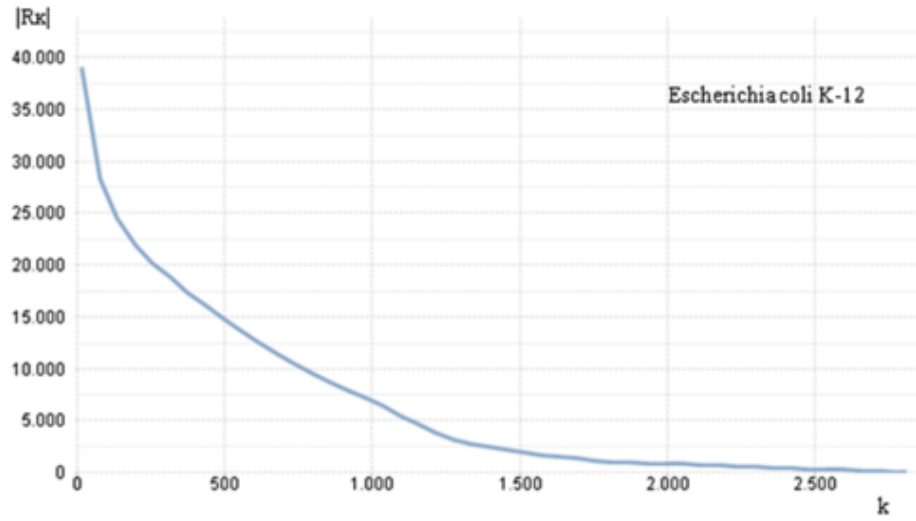

Comultiplicity
40
30
20
10
0
0
500
1.000
1.500
2.000
2.500
3.000
3.500
4.000
4.500
5.000
5.500
Multiplicity
Escherichia coli:
Real genomic sequence
Random permutation

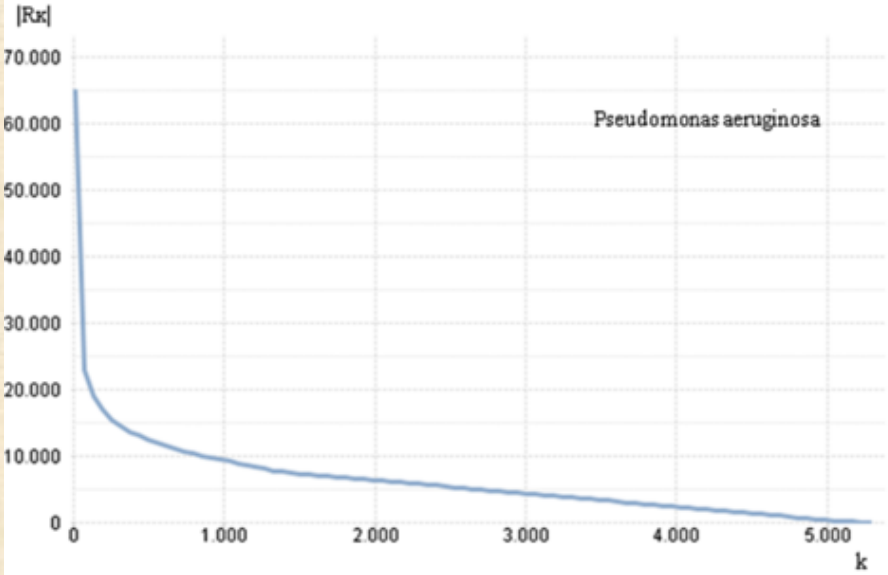

Comultiplicity
15
10
5
0
0
5.000
10.000
15.000
20.000
25.000
30.000
Multiplicity
Saccharomyces cerevisiae:
Real genomic sequence
Random permutation

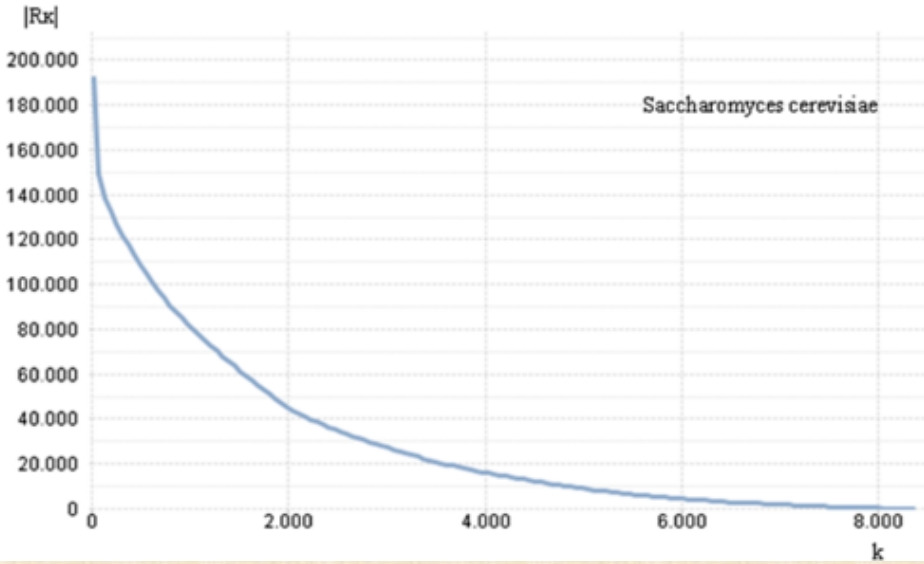

Comultiplicity
12
10
8
6
4
2
0
0
50.000
100.000
150.000
200.000
250.000
Multiplicity
Homo sapiens chr19:
Real genomic sequence
Random permutation

Lenght - repeat multiplicity

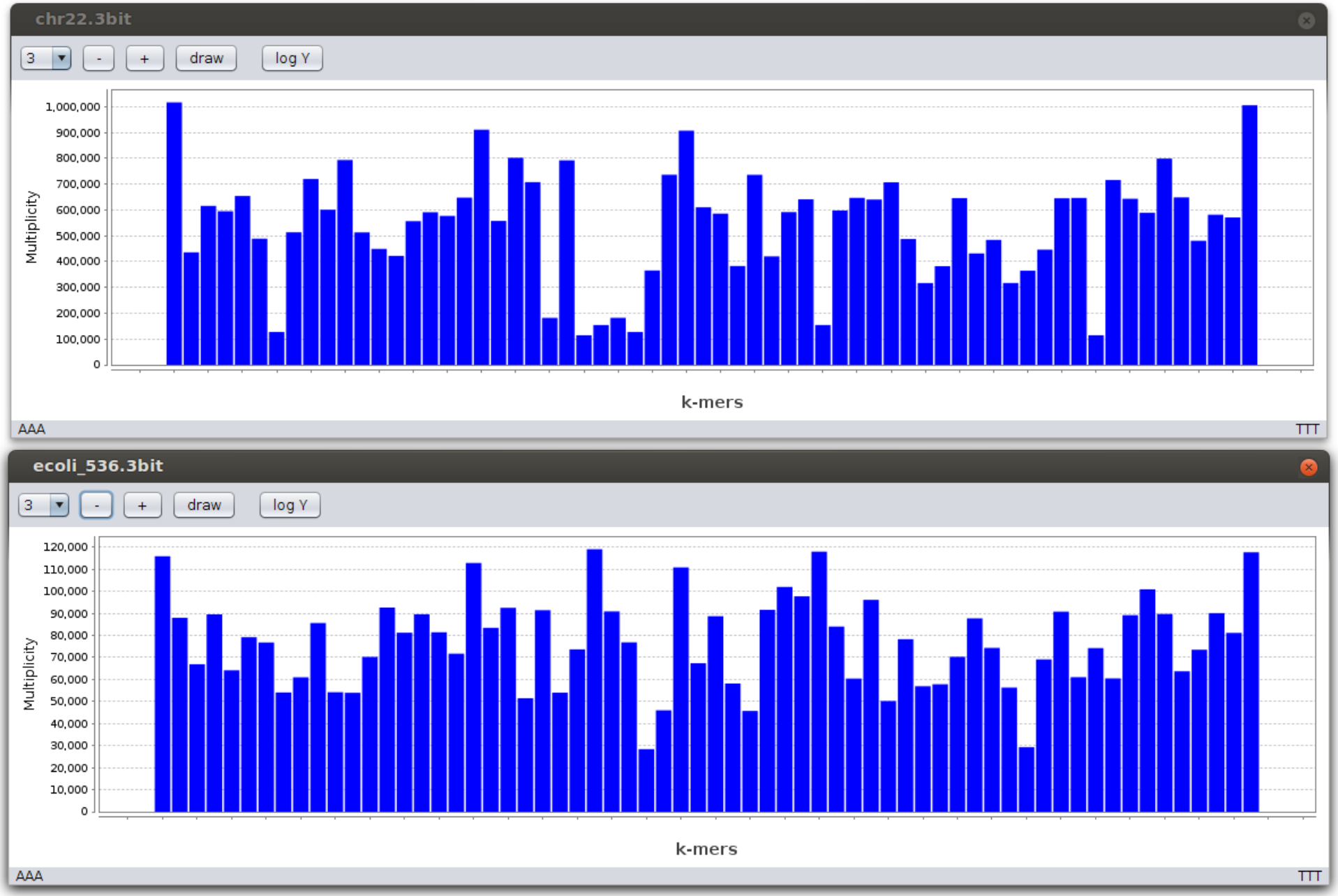


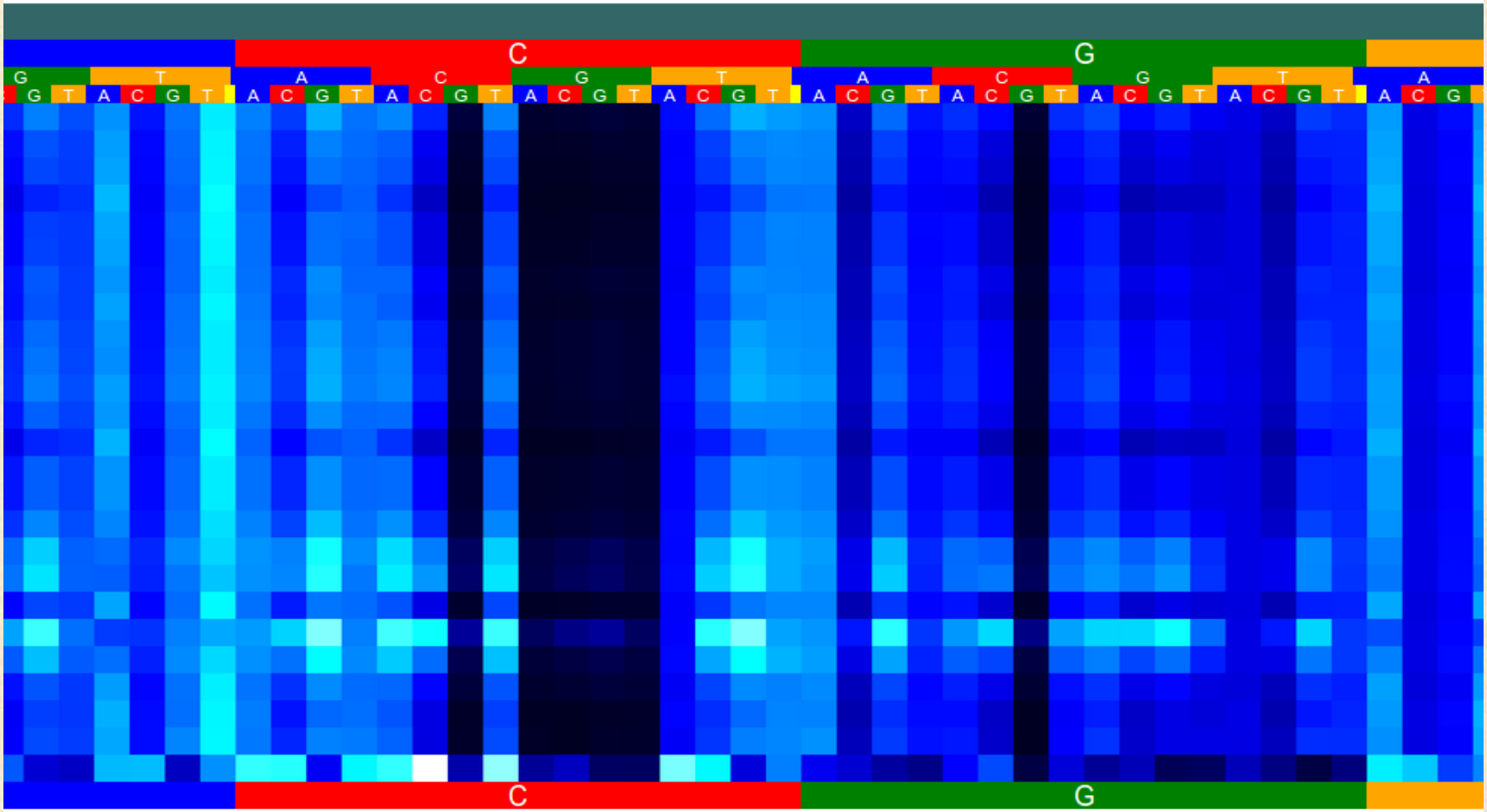




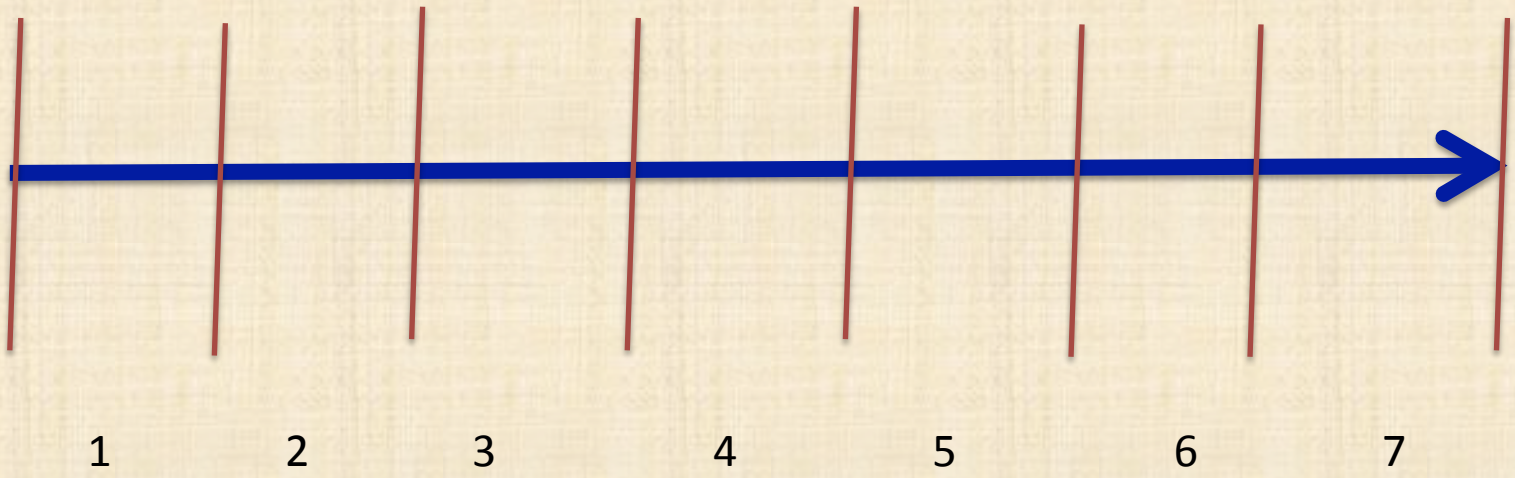


A maximal repeat cannot occurs more than 4 times (usually ≤ 2)
(How many kinds of k-repeat, and how many times they repeat?)


chr22.3bit
3
-
+
draw
log Y
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
Multiplicity
k-mers
AAA
TTT
ecoli_536.3bit
3
-
+
draw
log Y
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Multiplicity
k-mers
AAA
TTT

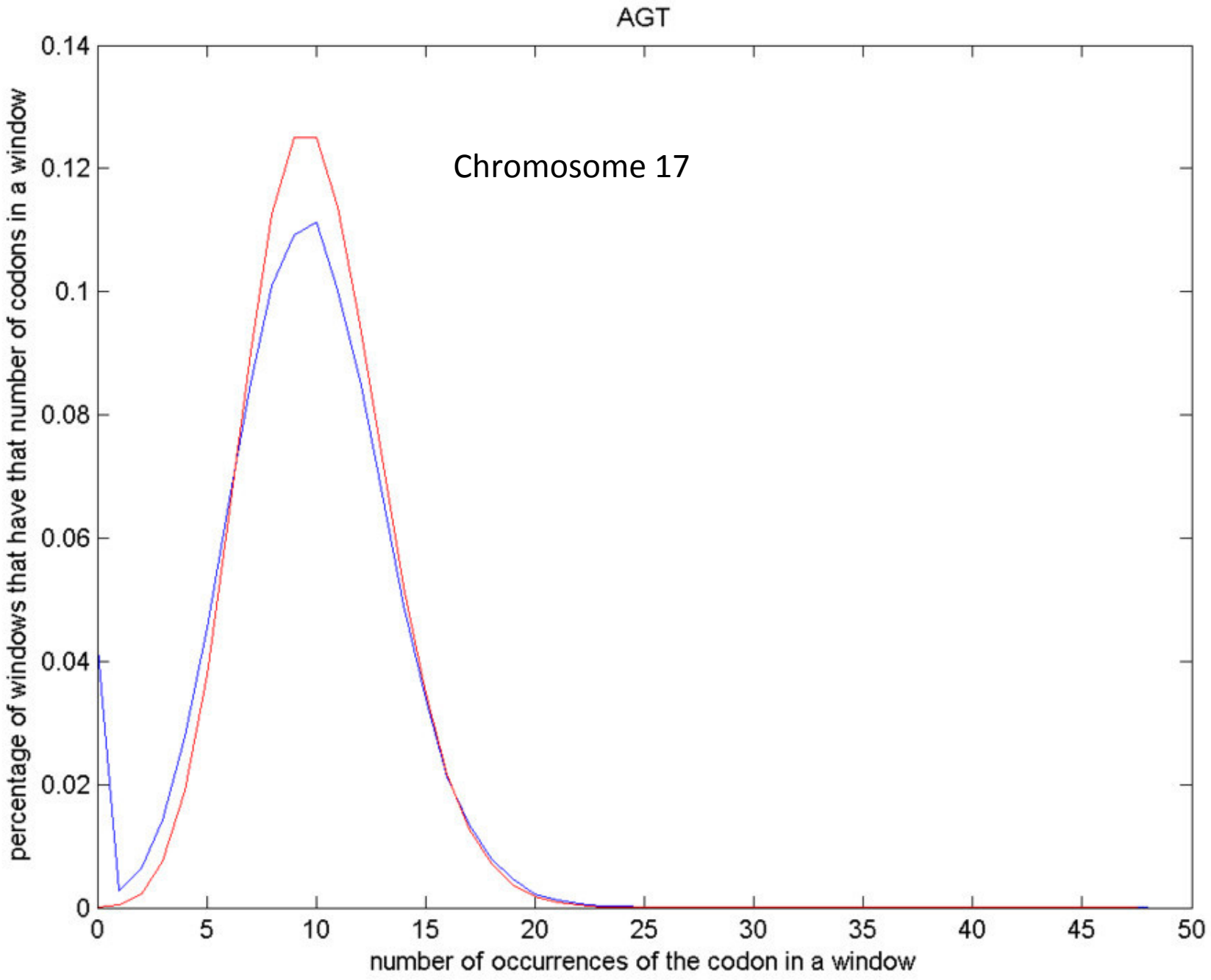

C
G
C
G

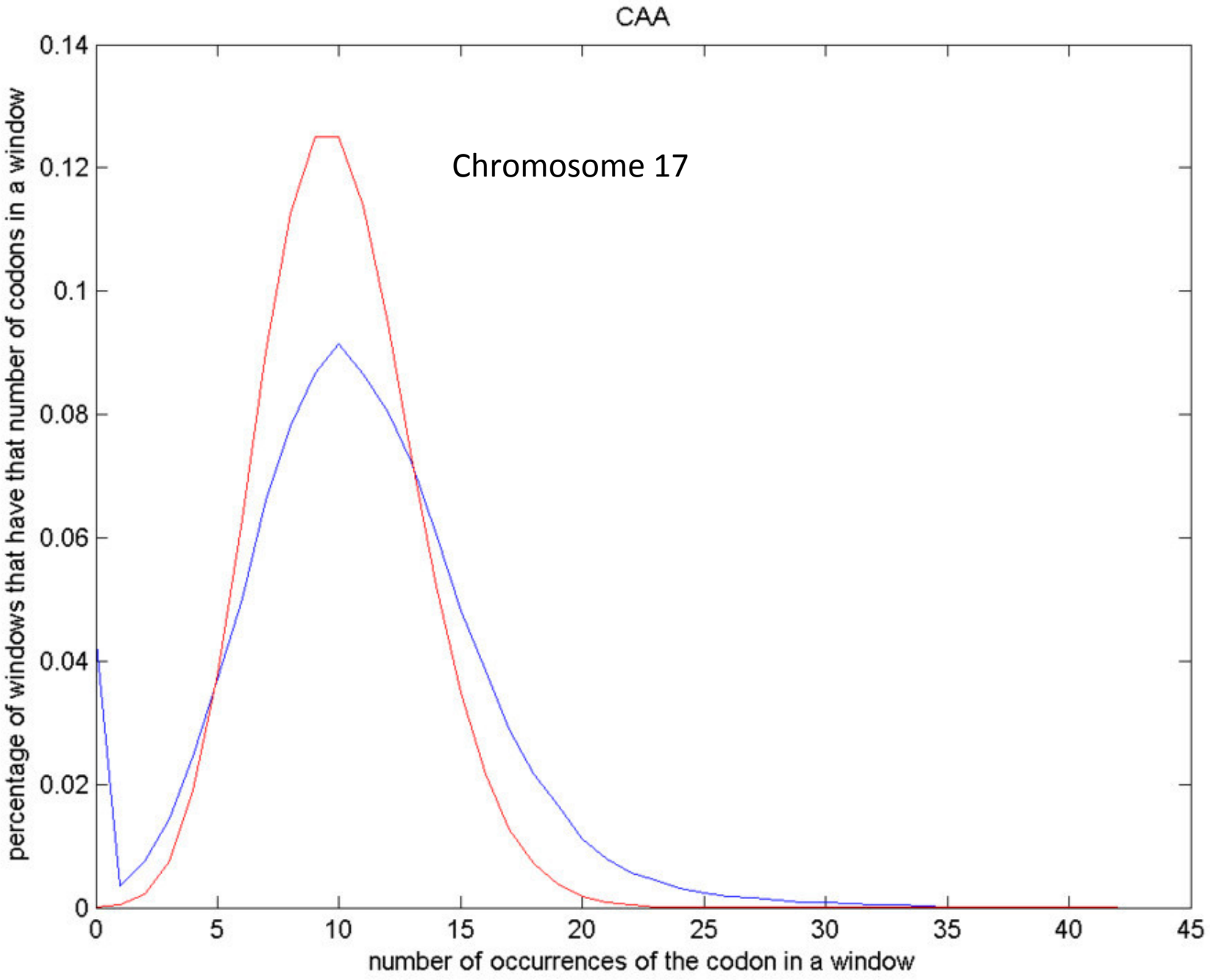
Codon Occurrence Distribution

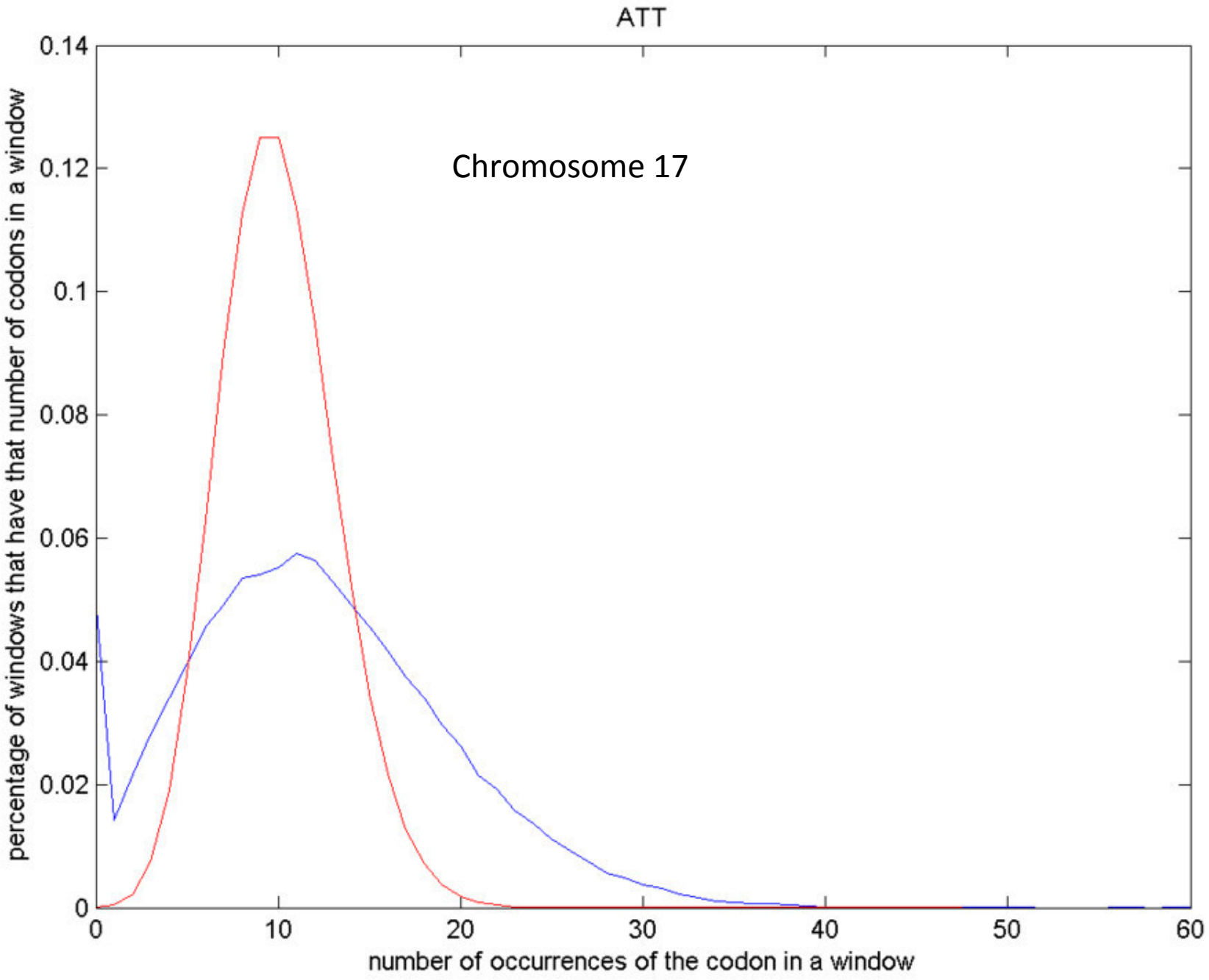


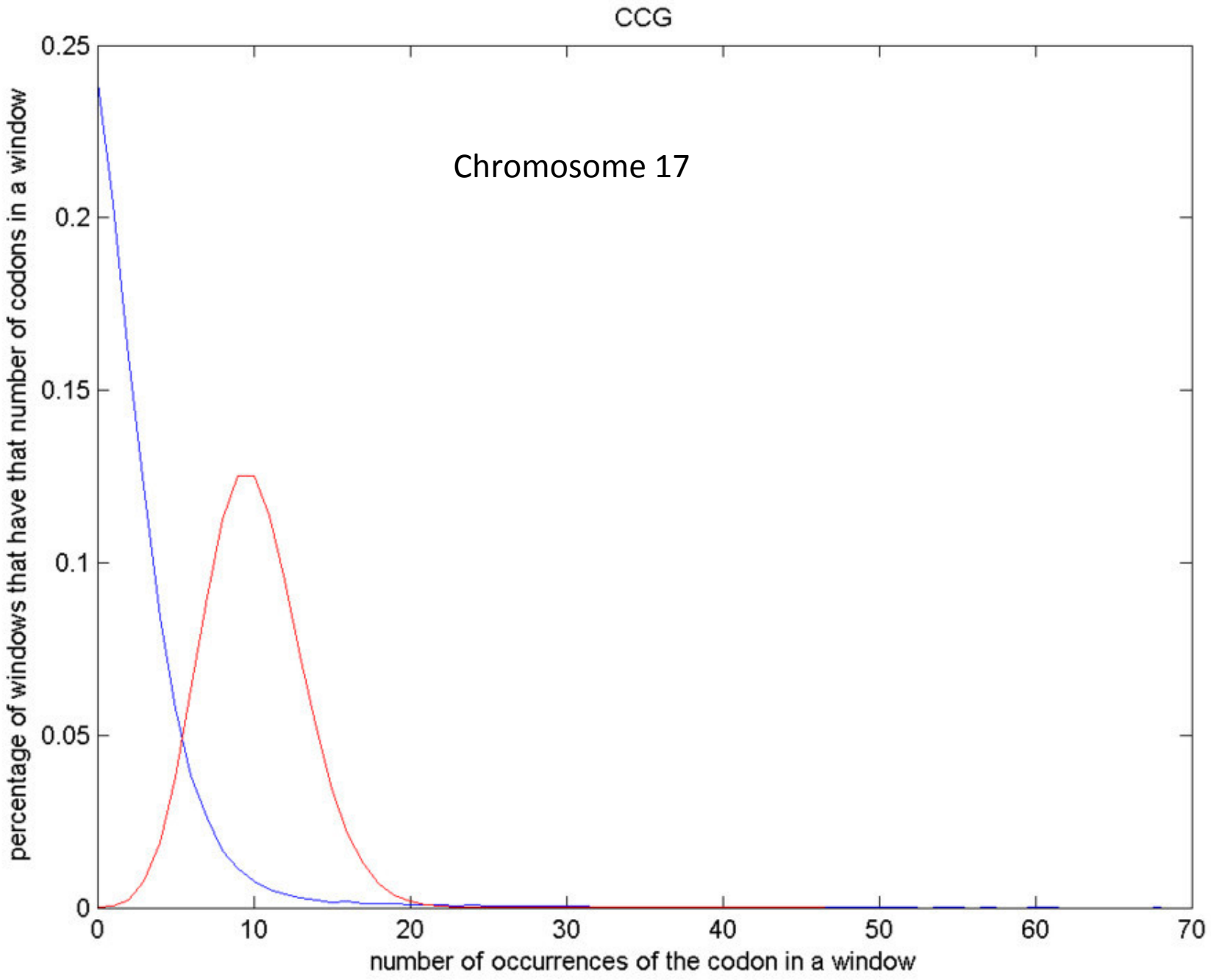
Count the number/percentage of 640-windows where "XYZ" occurs k times, k= 0, 1, 2, ….

J. Percus, Mathematics of Genome Analysis, Cambridge, 2002
(for the analytical determination of the window size)


AGT
Chromosome 17
percentage of windows that have that number of codons in a window
number of occurrences of the codon in a window
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
35
40
45
50


CAA
Chromosome 17
percentage of windows that have that number of codons in a window
number of occurrences of the codon in a window
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30
35
40
45


ATT
Chromosome 17
percentage of windows that have that number of codons in a window
number of occurrences of the codon in a window
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60

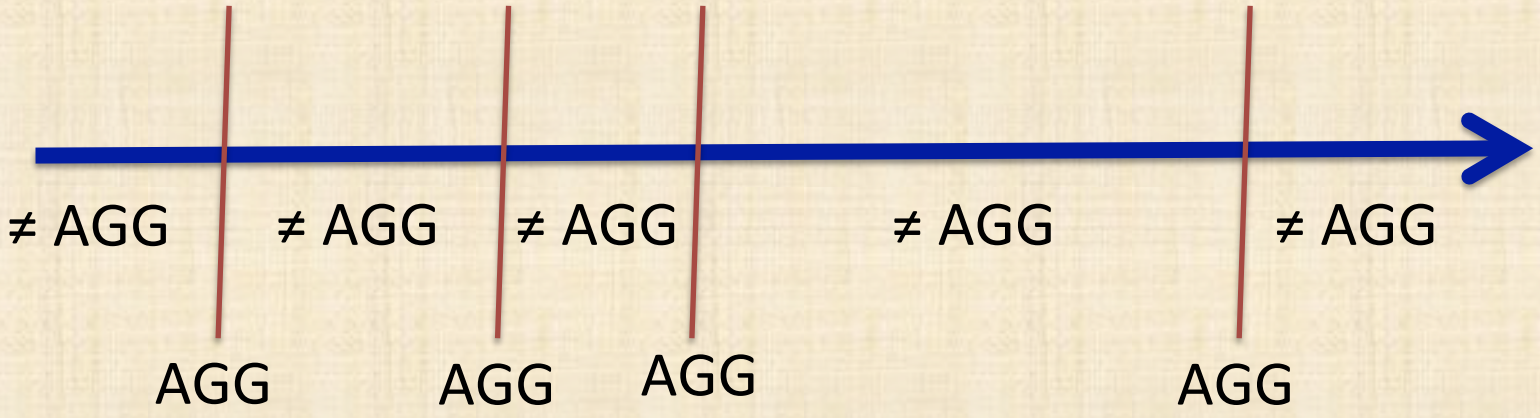

CCG
Chromosome 17
percentage of windows that have that number of codons in a window
number of occurrences of the codon in a window
0.25
0.2
0.15
0.1
0.05
0
0
10
20
30
40
50
60
70

Senza titolo — Modificato

AAA	V	20	21	24	20	21	21	18	V	V	V	19	V	V	V	V	V	V	V	V	V	V	V
AAC	V	9	10	10	9	10	9	10	V	9	9	9	V	V	V	V	8	9	7	9	V	V	9	V
AAG	V	12	13	12	13	13	13	13	V	12	13	13	V	V	V	V	11	12	10	12	V	V	13	V
AAT	V	15	16	18	15	16	15	16	V	14	14	15	V	V	V	V	11	15	V	V	V	V	16	V
ACA	V	12	13	13	12	12	12	13	V	12	12	12	V	V	V	V	12	12	10	12	V	V	13	V
ACC	8	7	7	7	7	7	7	7	V	8	8	8	V	V	V	V	9	7	10	9	V	V	7	V
ACG	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
ACT	11	11	11	11	11	11	11	11	V	11	11	11	V	V	V	V	10	11	10	11	V	V	11	V
AGA	V	13	14	14	13	13	13	13	V	13	14	13	V	V	V	V	13	14	12	14	V	V	13	V
AGC	V	9	9	8	9	9	9	9	V	9	9	9	V	V	V	V	11	9	11	10	V	V	8	V
AGG	V	10	10	9	9	10	10	10	V	10	10	10	V	V	V	V	12	9	12	12	V	V	10	V
AGT	11	11	11	11	11	11	11	11	V	11	11	11	V	V	V	V	11	11	10	10	V	V	11	V
ATA	V	12	13	14	12	12	11	13	V	10	12	12	V	V	V	V	V	12	V	V	V	V	12	V
ATC	9	9	9	9	9	9	9	9	V	9	9	9	V	V	V	V	8	9	8	9	V	V	9	V
ATG	V	12	12	13	13	12	12	12	V	12	12	12	V	V	V	V	10	12	10	12	V	V	13	V
ATT	V	16	16	19	16	16	14	16	V	14	14	15	V	V	V	V	12	16	V	V	V	V	15	V
CAA	V	12	12	12	12	12	12	12	V	12	12	12	V	V	V	V	11	12	11	12	V	V	12	V
CAC	V	9	9	8	9	9	9	9	V	10	9	9	V	V	V	V	11	9	12	10	V	V	9	V
CAG	V	12	12	11	12	12	13	12	V	13	13	13	V	V	V	V	16	12	17	15	V	V	11	V
CAT	V	12	12	13	12	12	12	13	V	12	12	12	V	V	V	V	11	13	9	11	V	V	13	V
CCA	V	11	11	10	11	11	11	11	V	12	12	12	V	V	V	V	14	11	16	14	V	V	11	V
CCC	V	6	6	5	5	6	6	6	V	6	6	6	V	V	V	V	8	6	V	8	V	V	6	V
CCG	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
CCT	V	9	9	9	10	10	10	10	V	11	10	10	V	V	V	V	13	9	13	12	V	V	10	V
CGA	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
CGC	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
CGG	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
CGT	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
CTA	9	9	9	9	9	9	9	9	V	9	9	9	V	V	V	V	8	9	7	8	V	V	9	V
CTC	V	9	9	8	8	9	9	9	V	9	9	9	V	V	V	V	11	9	12	10	V	V	8	V
CTG	V	12	12	11	12	12	13	12	V	13	13	13	V	V	V	V	16	12	18	15	V	V	12	V
CTT	V	12	13	13	13	13	12	13	V	12	13	13	V	V	V	V	11	13	9	12	V	V	13	V
GAA	V	13	13	13	12	13	11	13	V	12	12	13	V	V	V	V	10	12	9	11	V	V	12	V
GAC	6	6	6	6	6	6	6	6	V	6	6	6	V	V	V	V	7	6	7	7	V	V	6	V
GAG	V	9	9	8	8	8	9	8	V	9	9	9	V	V	V	V	10	9	12	10	V	V	8	V
GAT	9	9	9	9	9	9	9	9	V	9	9	9	V	V	V	V	8	9	8	9	V	V	9	V
GCA	9	9	9	8	9	9	9	9	V	9	9	9	V	V	V	V	10	9	11	10	V	V	9	V
GCC	V	5	5	5	5	5	5	5	V	6	5	5	V	V	V	V	9	5	12	8	V	V	5	V
GCG	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
GCT	V	9	9	8	9	9	9	9	V	10	9	9	V	V	V	V	11	9	11	10	V	V	8	V
GGA	V	9	9	8	9	9	9	9	V	10	10	9	V	V	V	V	11	9	11	10	V	V	9	V
GGC	V	5	5	5	5	5	5	5	V	6	5	5	V	V	V	V	9	5	12	8	V	V	5	V
GGG	V	6	6	5	6	6	6	6	V	6	6	6	V	V	V	V	8	6	V	8	V	V	5	V
GGT	8	7	7	7	7	7	7	7	V	8	8	7	V	V	V	V	9	7	10	9	V	V	7	V
GTA	7	8	8	8	8	8	7	8	V	7	7	8	V	V	V	V	7	8	6	7	V	V	8	V
GTC	6	6	6	6	6	6	6	6	V	6	6	6	V	V	V	V	7	6	7	7	V	V	6	V
GTG	V	9	9	8	9	9	9	9	V	9	10	9	V	V	V	V	11	9	11	11	V	V	9	V
GTT	V	9	10	10	10	9	9	9	V	9	9	9	V	V	V	V	8	9	8	9	V	V	9	V
TAA	V	13	14	15	14	14	12	13	V	12	12	12	V	V	V	V	8	13	V	V	V	V	13	V
TAC	7	8	8	8	8	8	7	8	V	7	7	8	V	V	V	V	7	8	6	7	V	V	8	V
TAG	9	9	9	9	9	9	9	9	V	8	9	9	V	V	V	V	8	9	7	8	V	V	9	V
TAT	V	13	12	15	13	12	11	12	V	10	10	11	V	V	V	V	V	12	V	V	V	V	14	V
TCA	V	13	13	13	13	13	13	13	V	13	13	13	V	V	V	V	13	13	12	13	V	V	13	V
TCC	V	9	9	8	9	9	9	9	V	9	10	9	V	V	V	V	11	9	11	10	V	V	9	V
TCG	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
TCT	V	14	13	14	13	13	13	14	V	14	13	14	V	V	V	V	13	13	11	13	V	V	13	V
TGA	V	13	13	13	13	13	13	13	V	13	13	13	V	V	V	V	12	13	12	13	V	V	13	V
TGC	9	9	9	9	9	9	9	9	V	10	9	9	V	V	V	V	10	9	11	10	V	V	9	V
TGG	V	11	11	10	11	11	12	11	V	12	12	11	V	V	V	V	15	11	15	13	V	V	11	V
TGT	V	12	13	13	13	12	12	12	V	12	12	12	V	V	V	V	11	12	11	12	V	V	13	V
TTA	V	14	15	16	14	13	13	13	V	12	11	13	V	V	V	V	9	14	V	V	V	V	13	V
TTC	V	13	13	13	13	13	12	12	V	12	12	12	V	V	V	V	11	13	9	12	V	V	12	V
TTG	V	12	12	13	12	12	12	12	V	12	12	12	V	V	V	V	11	12	10	11	V	V	12	V
TTT	V	20	22	23	22	22	23	19	V	V	V	19	V	V	V	V	V	V	V	V	V	V	V	V

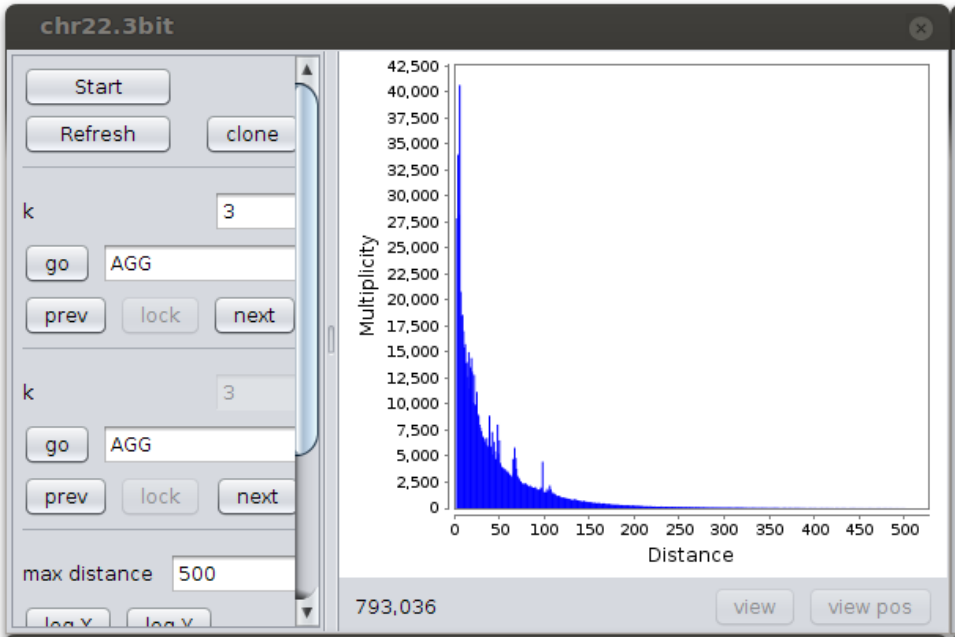
Codon Recurrence

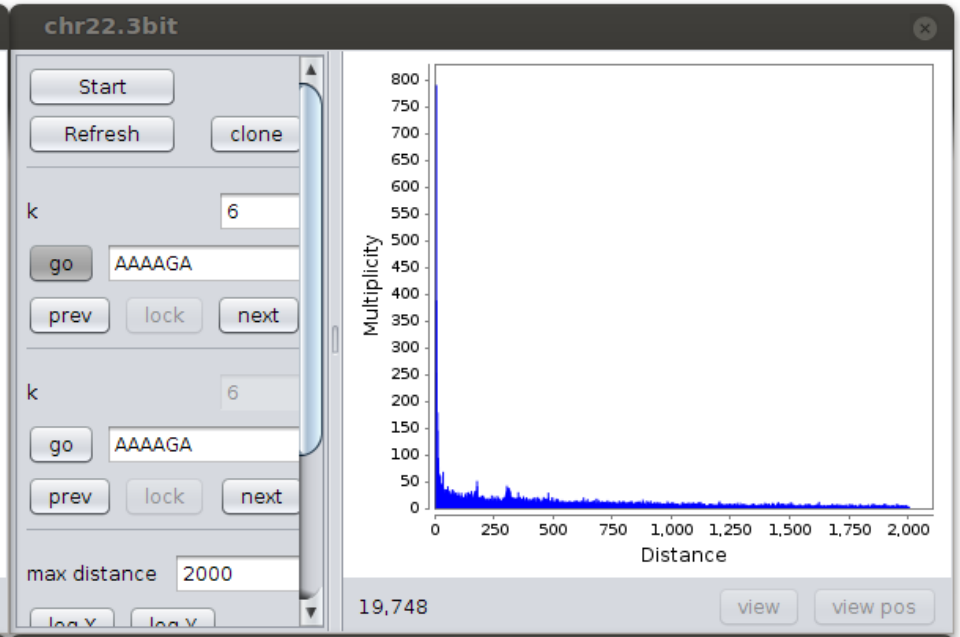
let us fix a codon, say AGG, and slice the genome:

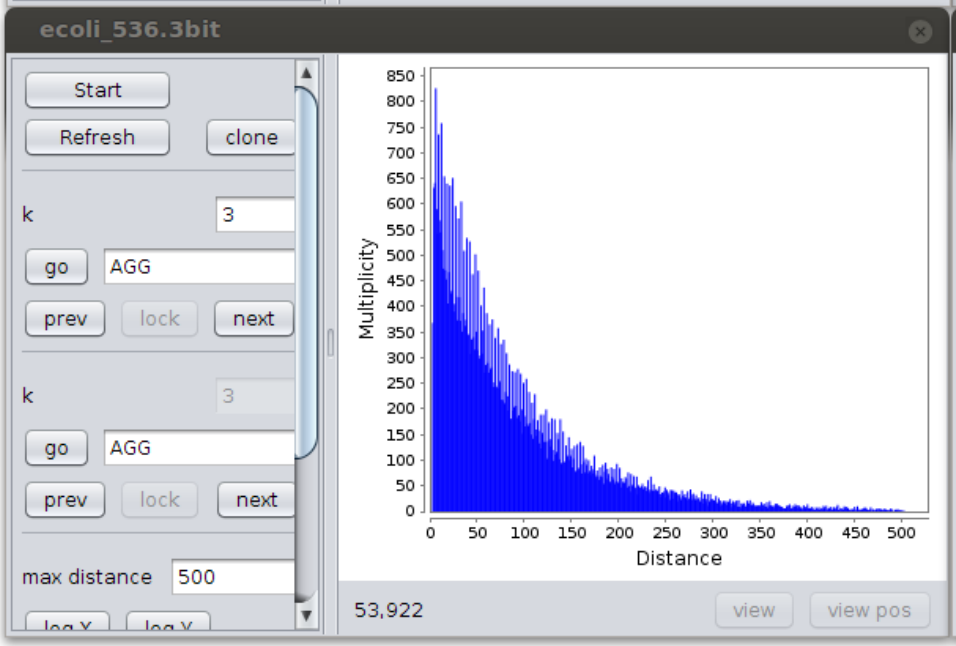


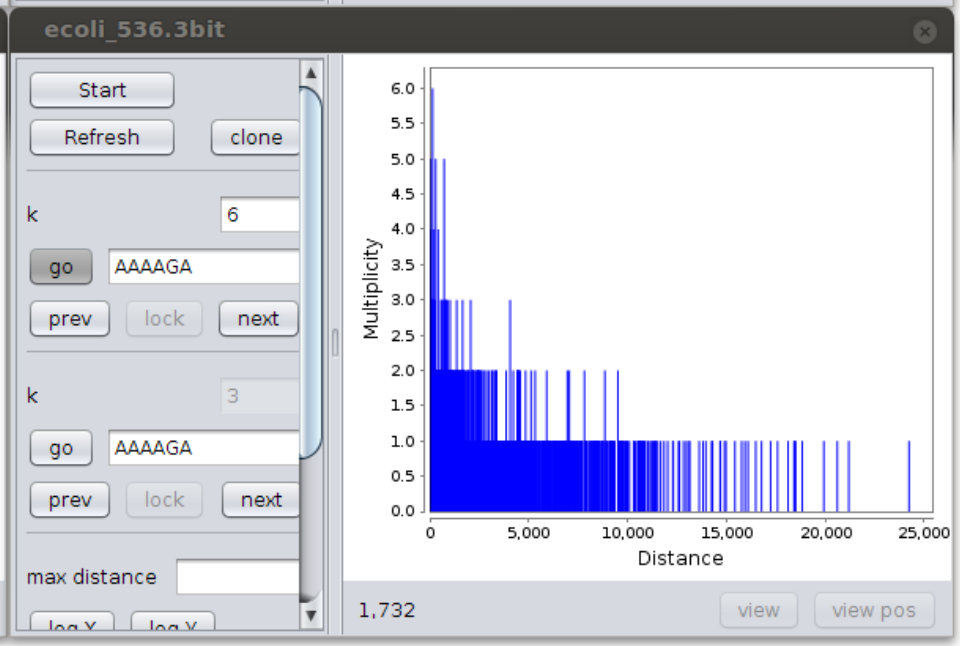
x= segment length, y= number of segments of length x

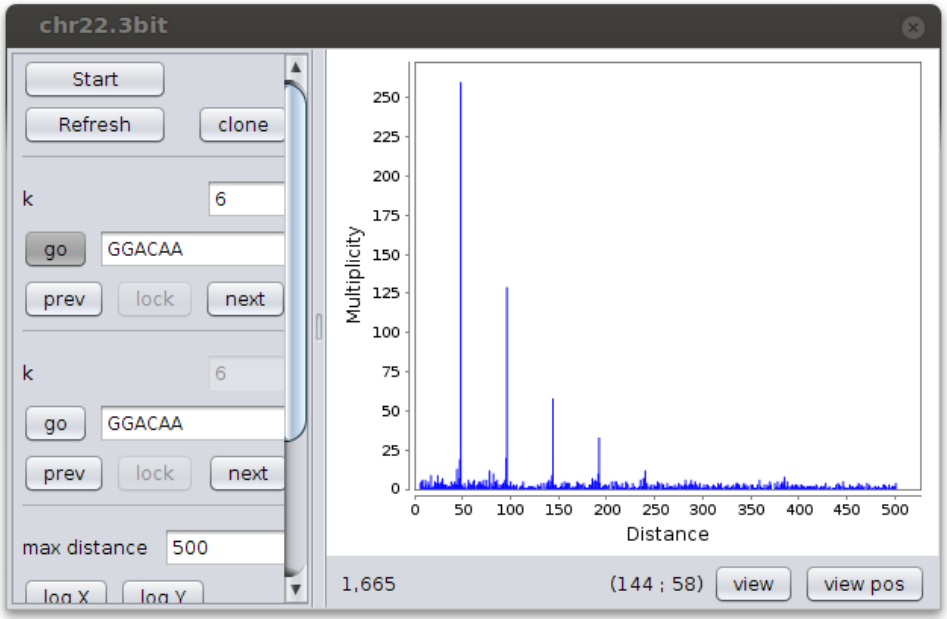
when codon occurrences follow Poisson laws, then their recurrences follow exponential distributions.

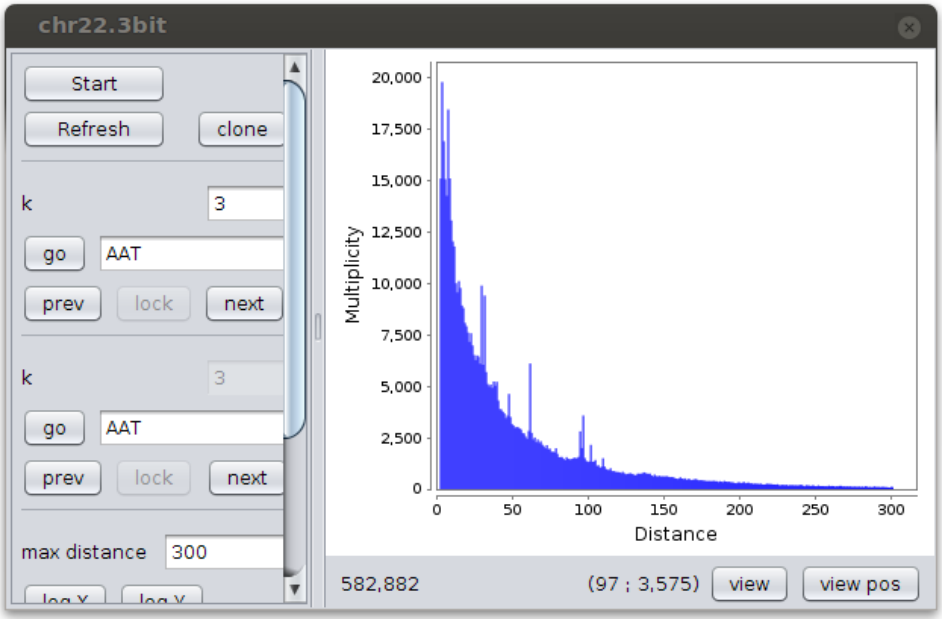

chr22.3bit
Start
Refresh
clone
k
3
go
AGG
prev
lock
next
k
3
go
AGG
prev
lock
next
max distance
500
Multiplicity
Distance
793,036
view
view pos

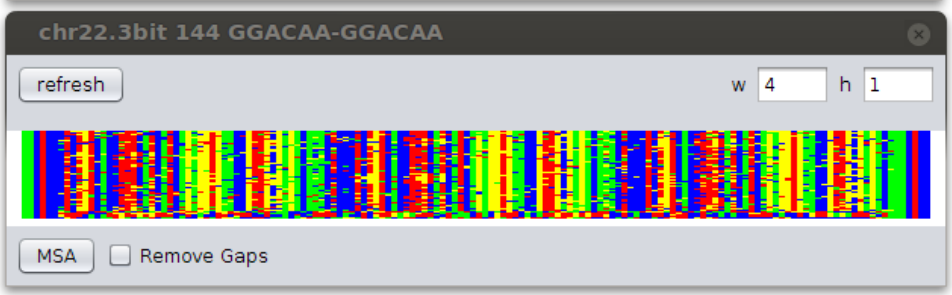

chr22.3bit
Start
Refresh
clone
k
6
go
AAAAGA
prev
lock
next
k
6
go
AAAAGA
prev
lock
next
max distance
2000
Multiplicity
Distance
19,748
view
view pos

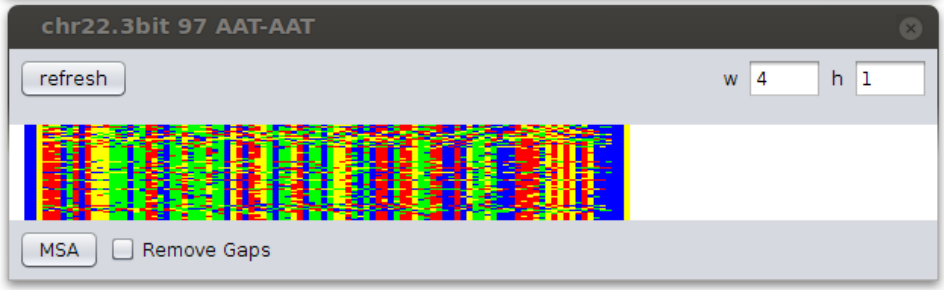

ecoli_536.3bit
Start
Refresh
clone
k
3
go
AGG
prev
lock
next
k
3
go
AGG
prev
lock
next
max distance
500
Multiplicity
Distance
53,922
view
view pos


ecoli_536.3bit
Start
Refresh
clone
k
6
go
AAAAGA
prev
lock
next
k
3
go
AAAAGA
prev
lock
next
max distance
Multiplicity
Distance
1,732
view
view pos


chr22.3bit
Start
Refresh
clone
k
6
go
GGACAA
prev
lock
next
k
6
go
GGACAA
prev
lock
next
max distance
500
log X
log Y
Multiplicity
Distance
1,665
(144 ; 58)
view
view pos


chr22.3bit
Start
Refresh
clone
k
3
go
AAT
prev
lock
next
k
3
go
AAT
prev
lock
next
max distance
300
Multiplicity
Distance
582,882
(97 ; 3,575)
view
view pos


chr22.3bit 144 GGACAA-GGACAA
refresh
w
4
h
1
MSA
Remove Gaps


chr22.3bit 97 AAT-AAT
refresh
w
4
h
1
MSA
Remove Gaps

6-mer Elongation Matrix

Hex(1,1) Hex(1,2) - - - - - - - - - - - - - - - - - - Hex(1,m1) He(1,m1)

Hex(2,1) Hex(2,2) - - - - - - - - - - - - - - - - - - Hex(2,m2) Hex(1,m2)

- -

- -

- -

- -

- -

- -

Hex(4095,1) Hex(4095,2) - - - - - - - - - - - - - - - - - Hex(4095,m4095)

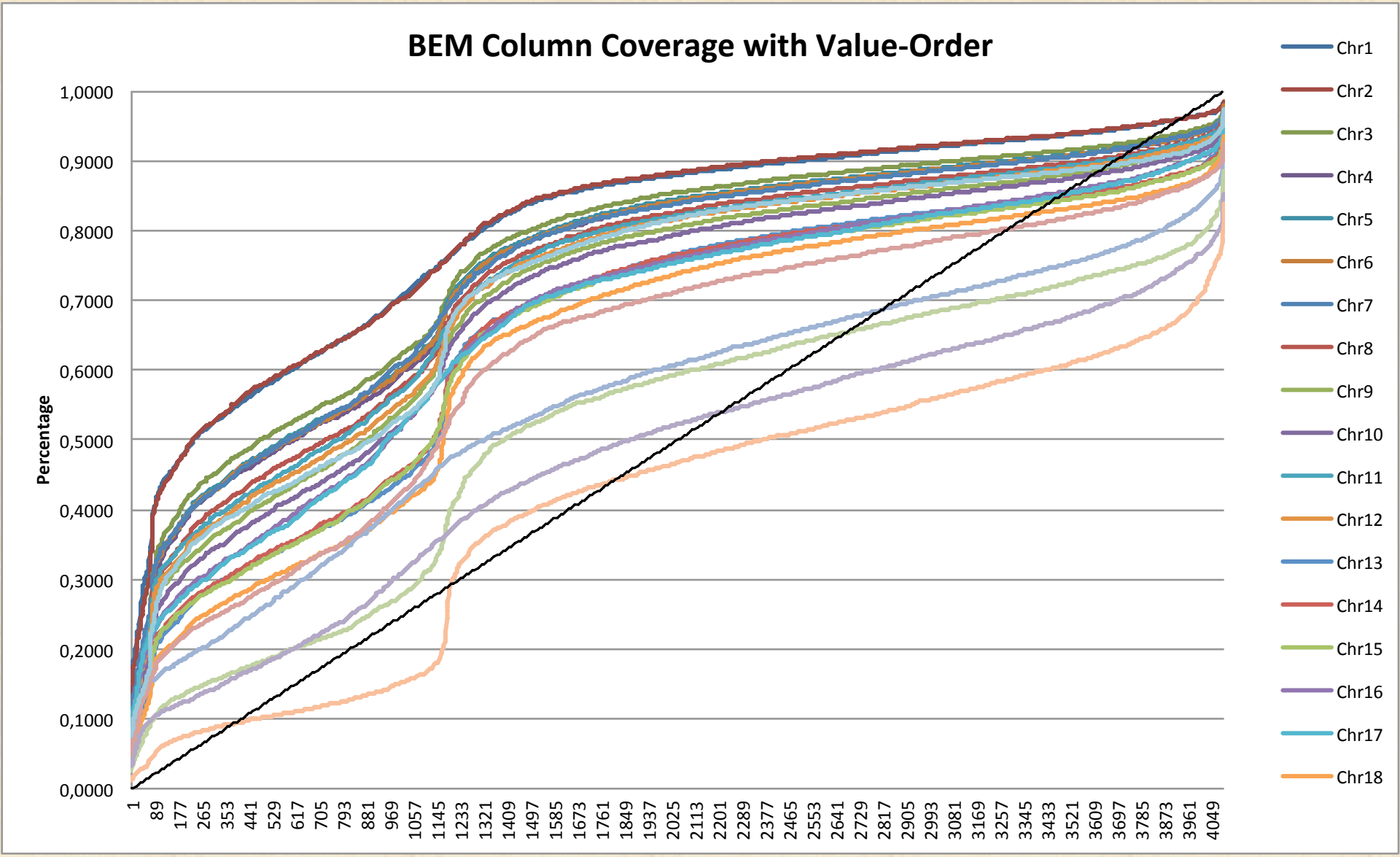
Hex(4096,1) Hex(4096,2) - - - - - - - - - - - - - - - - - Hex(4096,m4096)

Kinds of Elongation Matrices

- SEM_k ($k=0, 1, 2, \ldots, 5$)
- FEM
- DEM
- MEM
- NEM
- BEM
- PEM

More than 60 Gigabyte of Matrices for H. sapiens

Zipf Elongation Distribution



Theorems on Elongation Matrices

Th1: $SEM_0 + G[1,6] \approx G$

Th2: $DEM + NEM \approx G$

Th3: $FEM \approx D_{12}(G)$

Th4: $FEM \approx SEM_0 + \ldots + SEM_5$

Th5: $PEM + G \approx D_k(G)$, $k > 6$

Th6: $SEM_0 + G[1,6] \Longrightarrow PEM$

Th7: $SEM_0 + G[1,6] \approx D_k(G)$, $k > 6$

The proofs are based on the Algoritms constructing the various elongation matrices.

Basic Genomic Indexes

$|G|/(1+4^k)$

The fraction of Γ^k that occurs in a random genome with the same length as G (Fofanov)

$|D_k(G)|/4^k$

The fraction of Γ^k that occurs in G

- ***Ln*** Length
- ***k-ls*** k-Lexical selectivity
- ***mfl*** Maximal Forbidden Length
- ***3-mult*** Codon multiplicity/frequency (in general, k-mer frequency)
- ***mrl*** Maximum Repeat length (+1 = all-hapax ***lub*** least upper bound)
- ***mhl*** Minimum Hapax Length (-1 = all-repeat ***glb*** greatest lower bound)
- ***arl , ahl*** Average (Repeat/hapax) Length (also frequency weighted)
- **$E_k(G)$, $EE_k(G)$** **E**mpirical k-**E**ntropy , **E**xcess **E**mpirical k-**E**ntropy

Castellini, Franco, Manca:
A dictionary based informational genome analysis,
BMC Genomics, Sept. 2012, 13:485

Shannon's Approach
(Al Kindi's intuition)

$$Inf_2(word) = - \log_2(prob(word))$$

$$E = - \Sigma_w \, prob(w) Inf(w)$$

Entropy is the mean information of a text

We computed (empirical) Entropy
for (almost) any word length for all H. chr.
($k= 18$, $E_k \approx 24$; $k=200$ $E_k \approx 25$ **!!!**)

Algorithmic basis of k-mer frequency computation

- Suffix trees ST
- Suffix arrays SA
- Enanched SA ESA
- N-extended ESA NESA

Weiner 73
McCreight 76
Ukkonen 95
Farach 97
Manber & Myers 90
Abouelhoda, Kurtz, Ohlebusch 2004
Kurtz et a. 2008

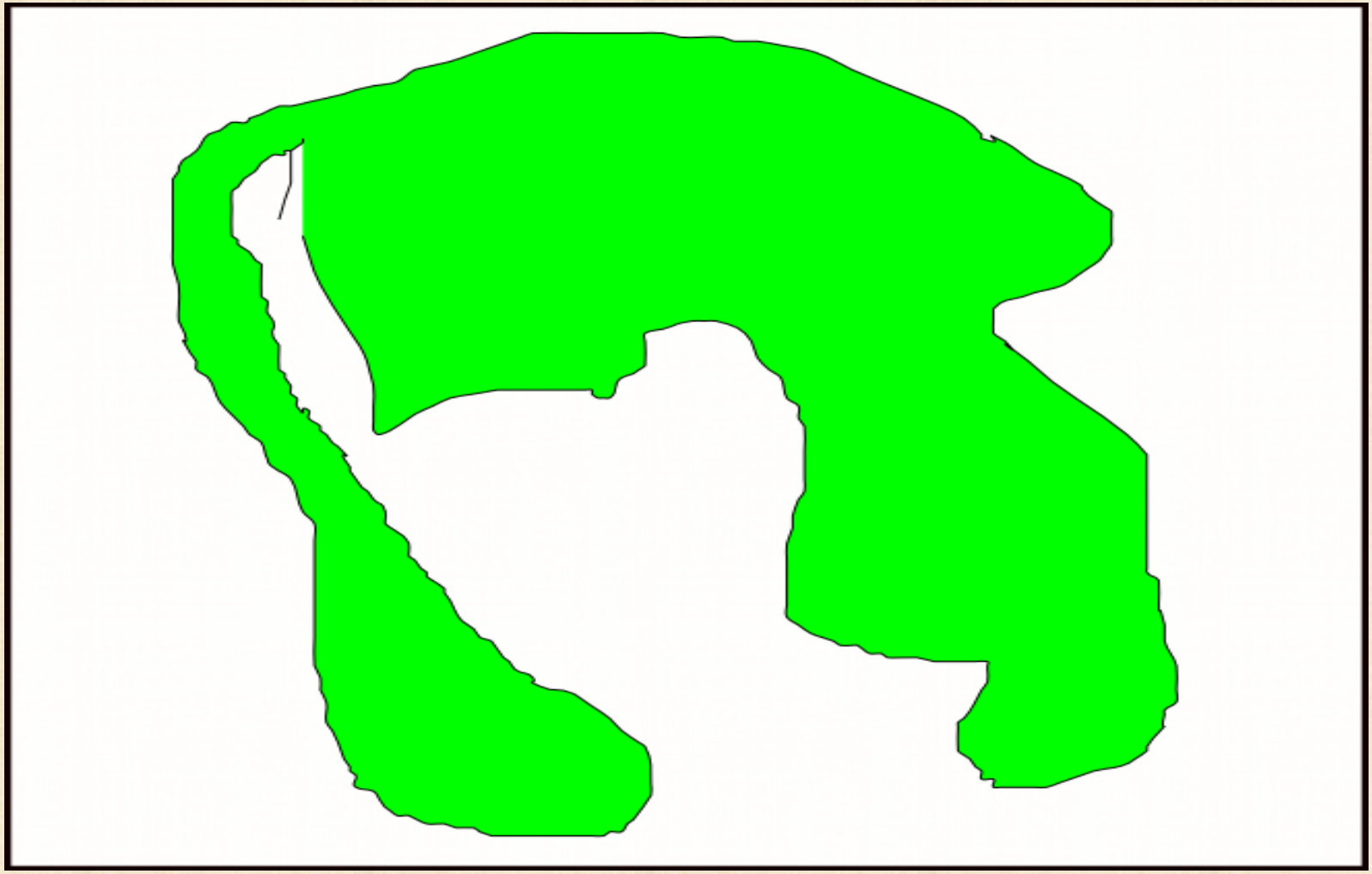
Measures of biological information

- $M = \text{argmax}_k \, E_k(\text{Rand}_{|G|}) = mrl(\text{Rand}_{|G|}) + 1$

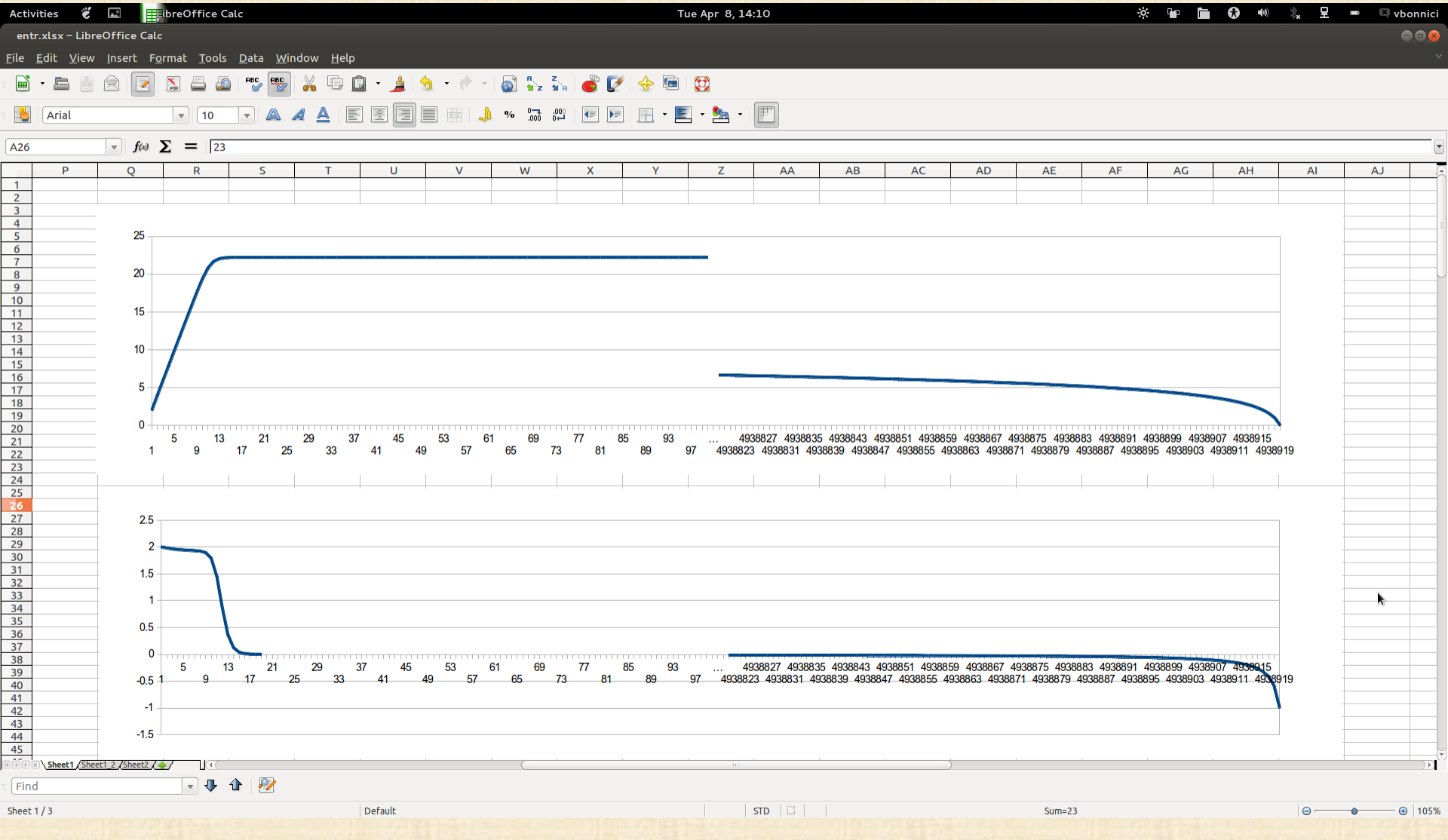
- **Bio-bit(G)** $= E_M(\text{Rand}_{|G|}) - E_M(G)$
  $= \log_2(|G| - M+1) - E_M(G) \approx \boxed{\log_2(|G|) - E_M(G)}$

- **Evo-Info(G)** $= (G/M) \cdot$ Bio-bit(G)

Monte Carlo Integration



Entropy and Excess Entropy (E. coli)



Genomic Dictionaries

Sets of words (of length 10-100) with relevant recurrence properties.

Carpena et al. – C index
Phys. Rev. E - 2009

Carpena et alii's Approach

1. Distance word recurrence distribution d(α)
2. Standard dev. of d(α) and mean normalization σ(α)
3. Geometric distributions $p^{d-1}(1-p)$ with p = n/L
4. Geometric Normalization $\sigma_{nor}(\alpha)$
5. Random Normalization $\sigma_{nor}(\alpha, n)$
6. Index of clusterization C(α)
7. Selection of words by elongation from initial seeds by means of stability w.r.t. the word relevance indexe C.

IG-Tools
V. Bonnici Phd

www.infogenomic-explorer
(Castellini, Manca)

A Genomic Word Selection Algorithm

- Word Recurrence distribution
- Parameter evaluation of the nearest Geom. Distr.
- Average distance recurrence distribution
- Normalization of distributions
- Entropic divergence (Kullback-Leibler) between distributions (symmetric extension)
- Extraction of Recurrence Indexes for 6-mer seeds
- Word Selection by elongation stability

Representations and Visualizations

10 almost unconventional methods of genome representation

Positions → Symbols

Words → Positions

Characteristic vectors

Elongation trees

Elongation matrices

Permutation matrices

(01)-walks (CGR - Chaos Game Representation)

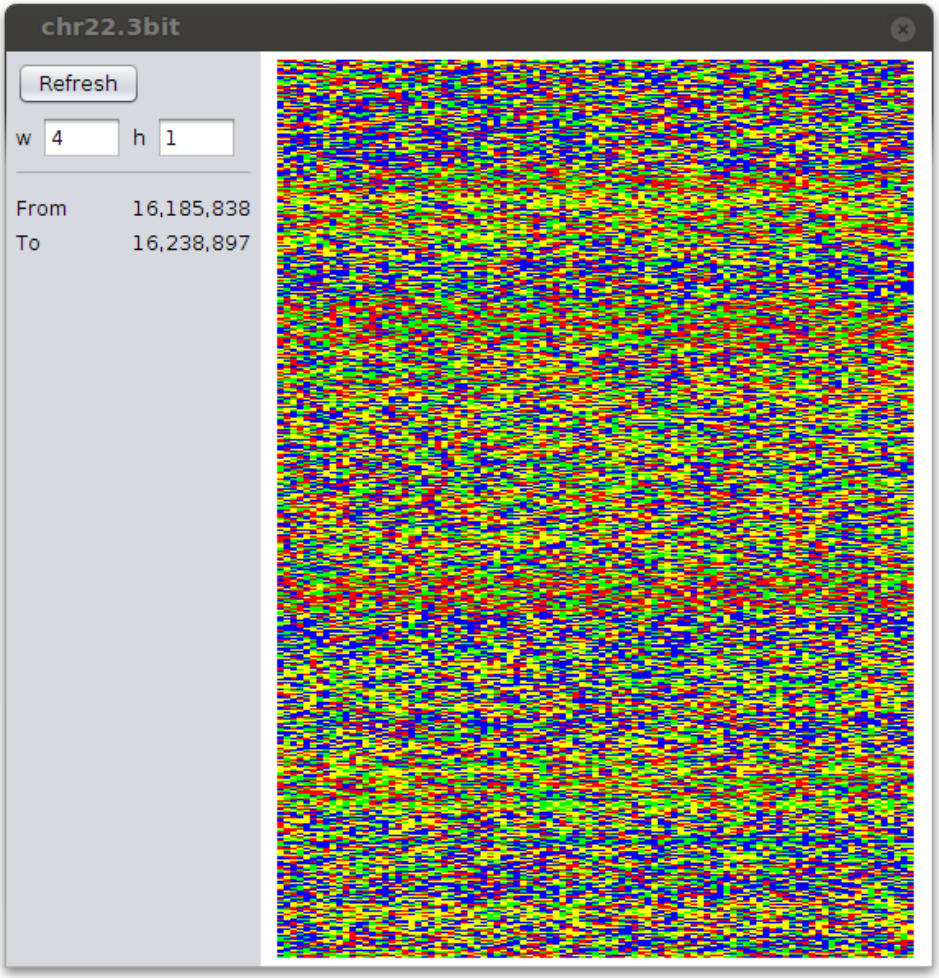
Univocal dictionaries

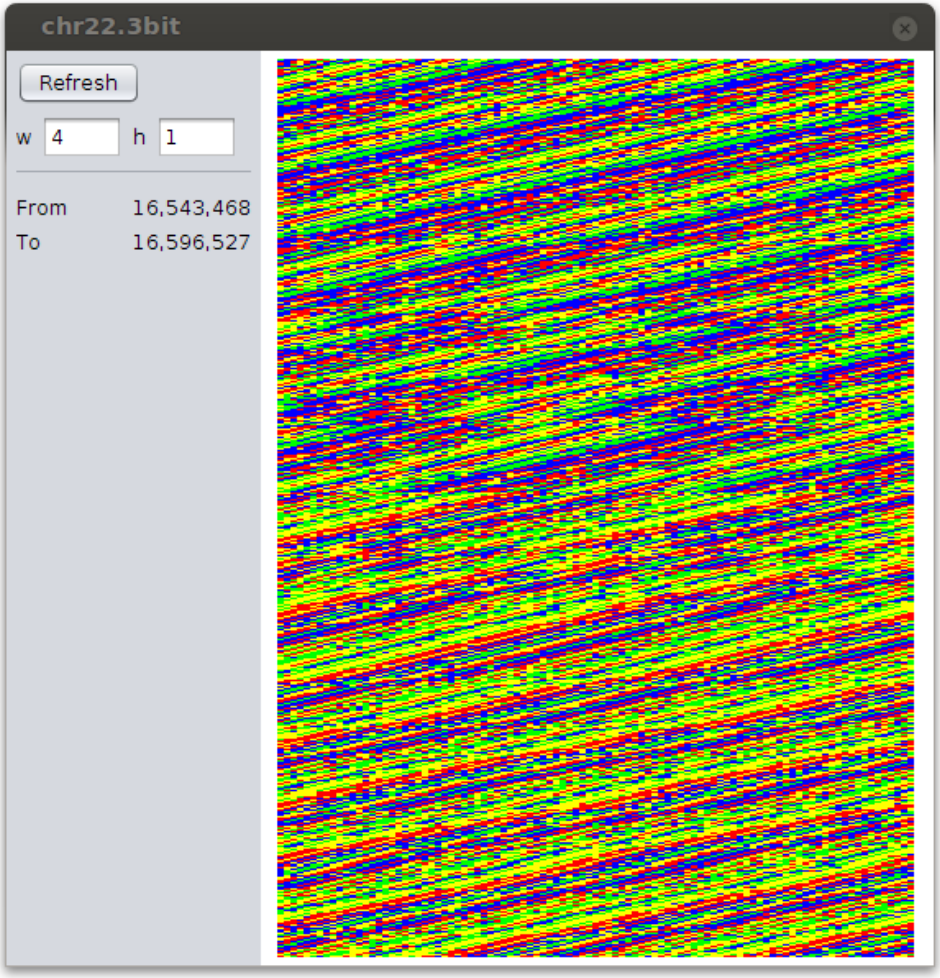
Autosimilarity distances

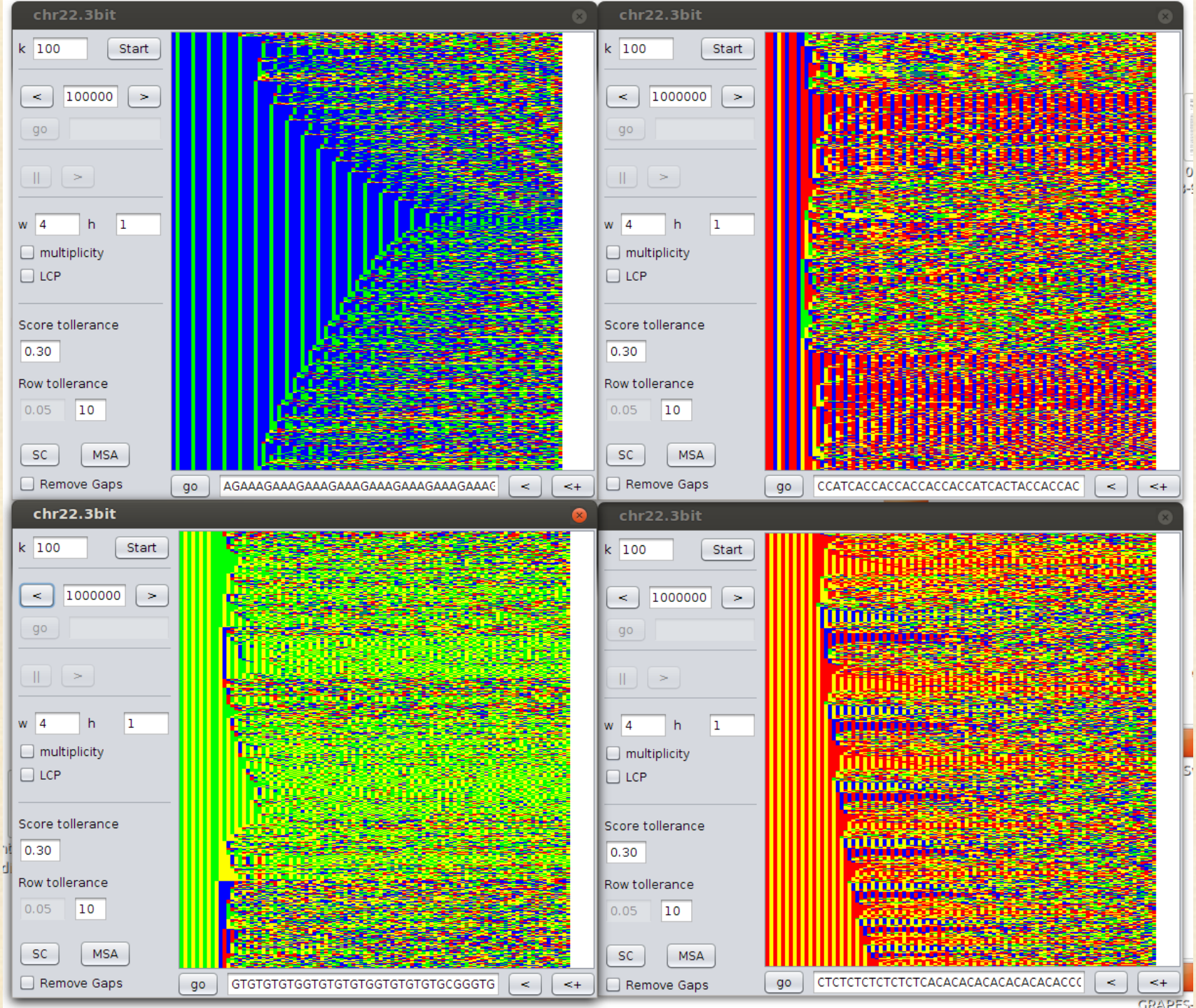
Word distance sequences

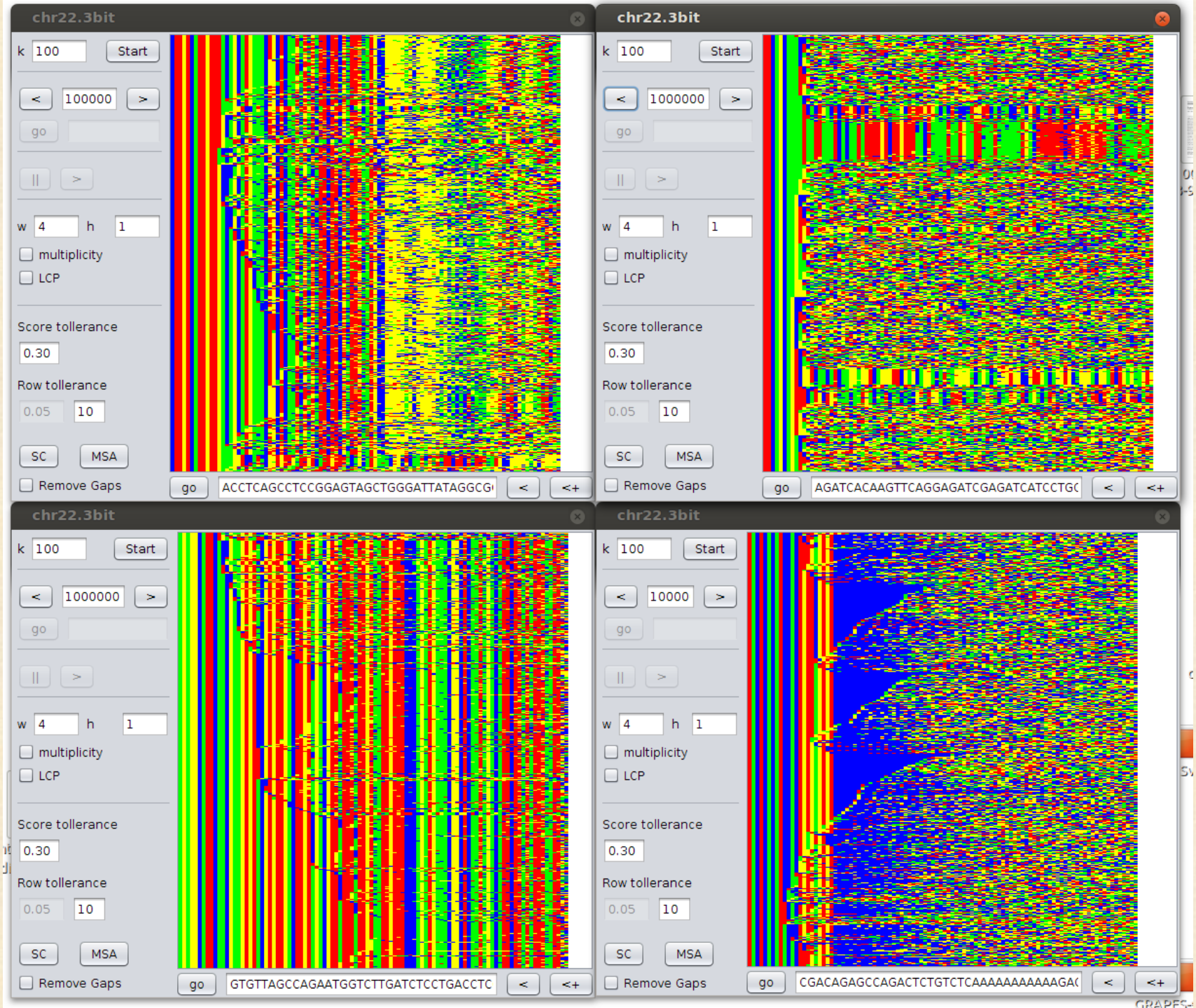
Lexicographic indexes

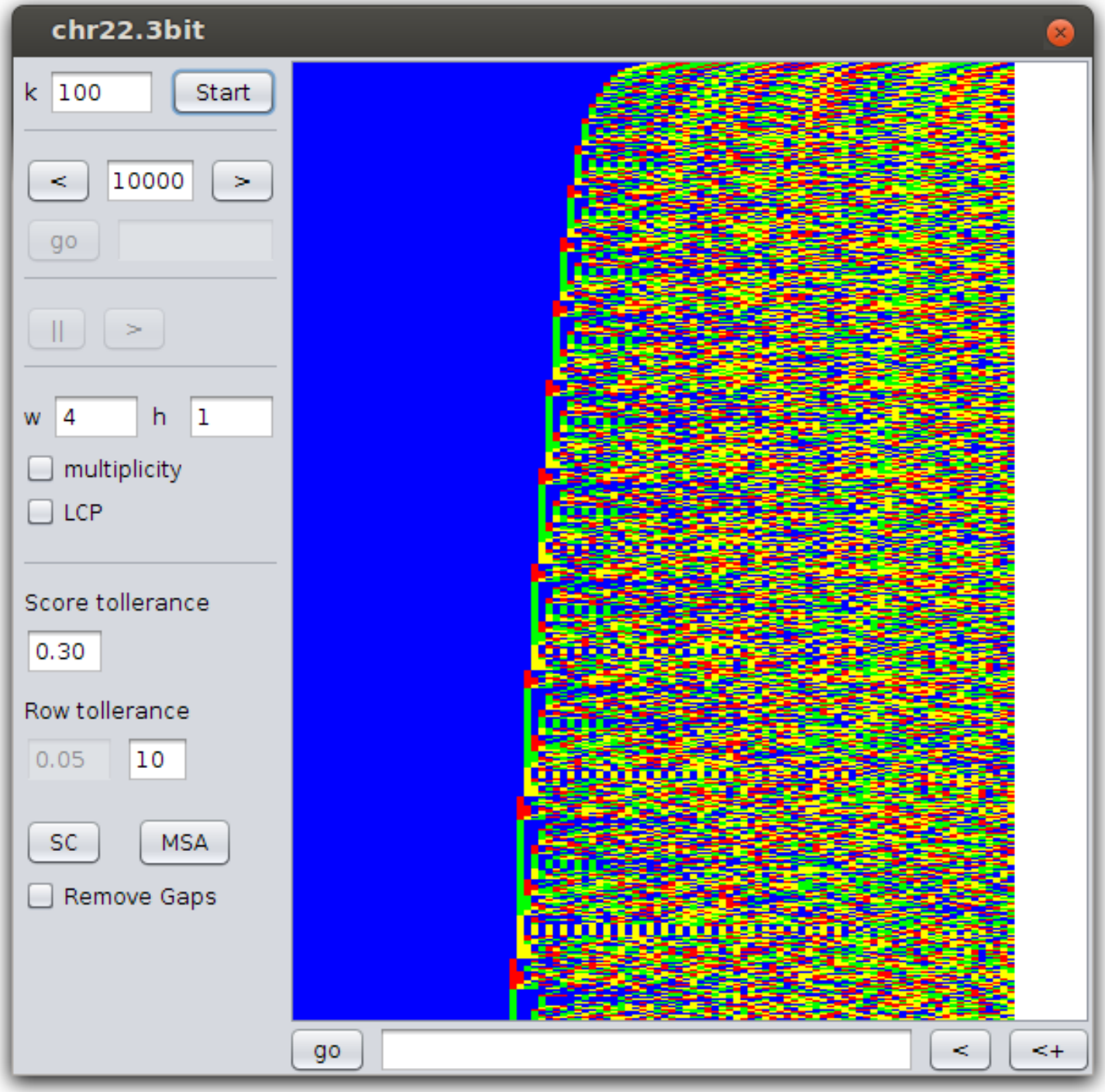
Chromatic lines



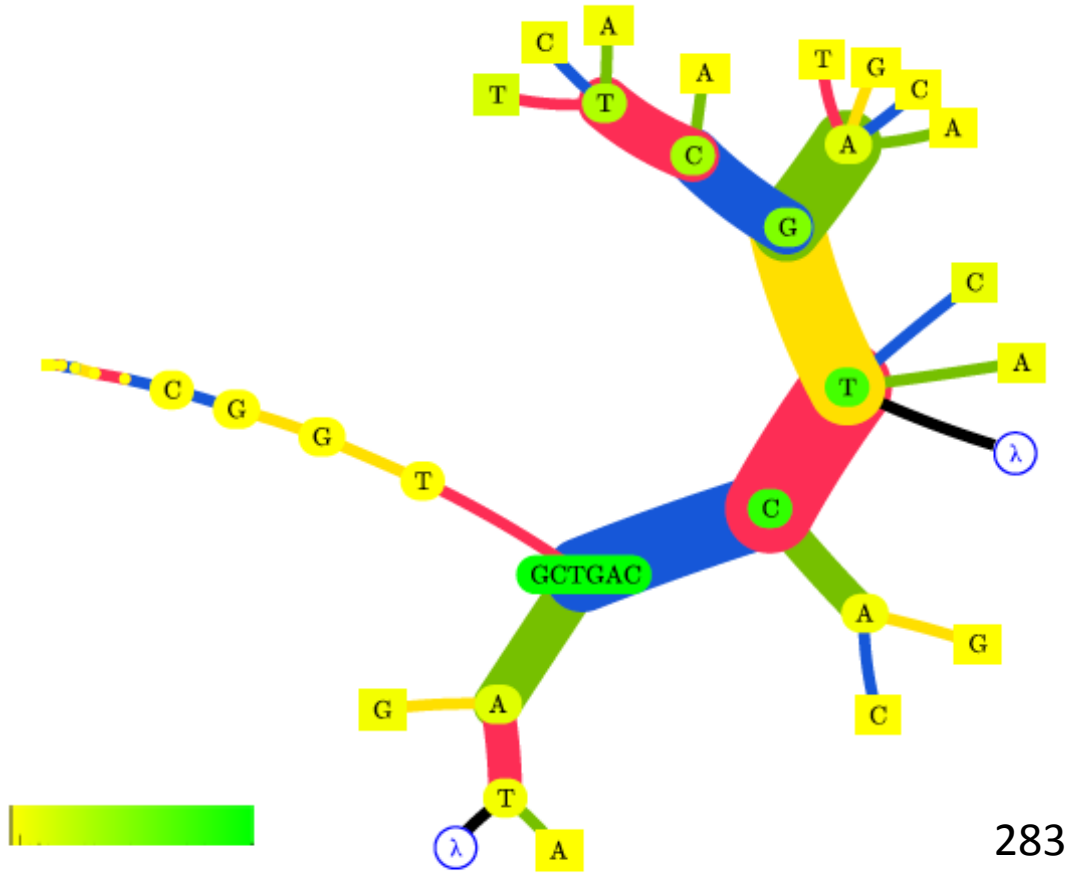



chr22.3bit
k 100
Start
< 100000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC MSA
Remove Gaps
go AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAC < <+
chr22.3bit
k 100
Start
< 1000000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC MSA
Remove Gaps
go CCATCACCACCACCACCACCATCACTACCACCAC < <+
chr22.3bit
k 100
Start
< 1000000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC MSA
Remove Gaps
go GTGTGTGTGGTGTGTGTGGTGTGTGTGCGGGTG < <+
chr22.3bit
k 100
Start
< 1000000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC MSA
Remove Gaps
go CTCTCTCTCTCTCTCACACACACACACACACCC < <+


chr22.3bit
k 100
Start
< 100000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC
MSA
Remove Gaps
go
ACCTCAGCCTCCGGAGTAGCTGGGATTATAGGCG
< <+
chr22.3bit
k 100
Start
< 1000000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC
MSA
Remove Gaps
go
AGATCACAAGTTCAGGAGATCGAGATCATCCTGC
< <+
chr22.3bit
k 100
Start
< 1000000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC
MSA
Remove Gaps
go
GTGTTAGCCAGAATGGTCTTGATCTCCTGACCTC
< <+
chr22.3bit
k 100
Start
< 10000 >
go
|| >
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC
MSA
Remove Gaps
go
CGACAGAGCCAGACTCTGTCTCAAAAAAAAAAAGAC
< <+


chr22.3bit
k
100
Start
<
10000
>
go
||
>
w
4
h
1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05
10
SC
MSA
Remove Gaps
go
<
<+

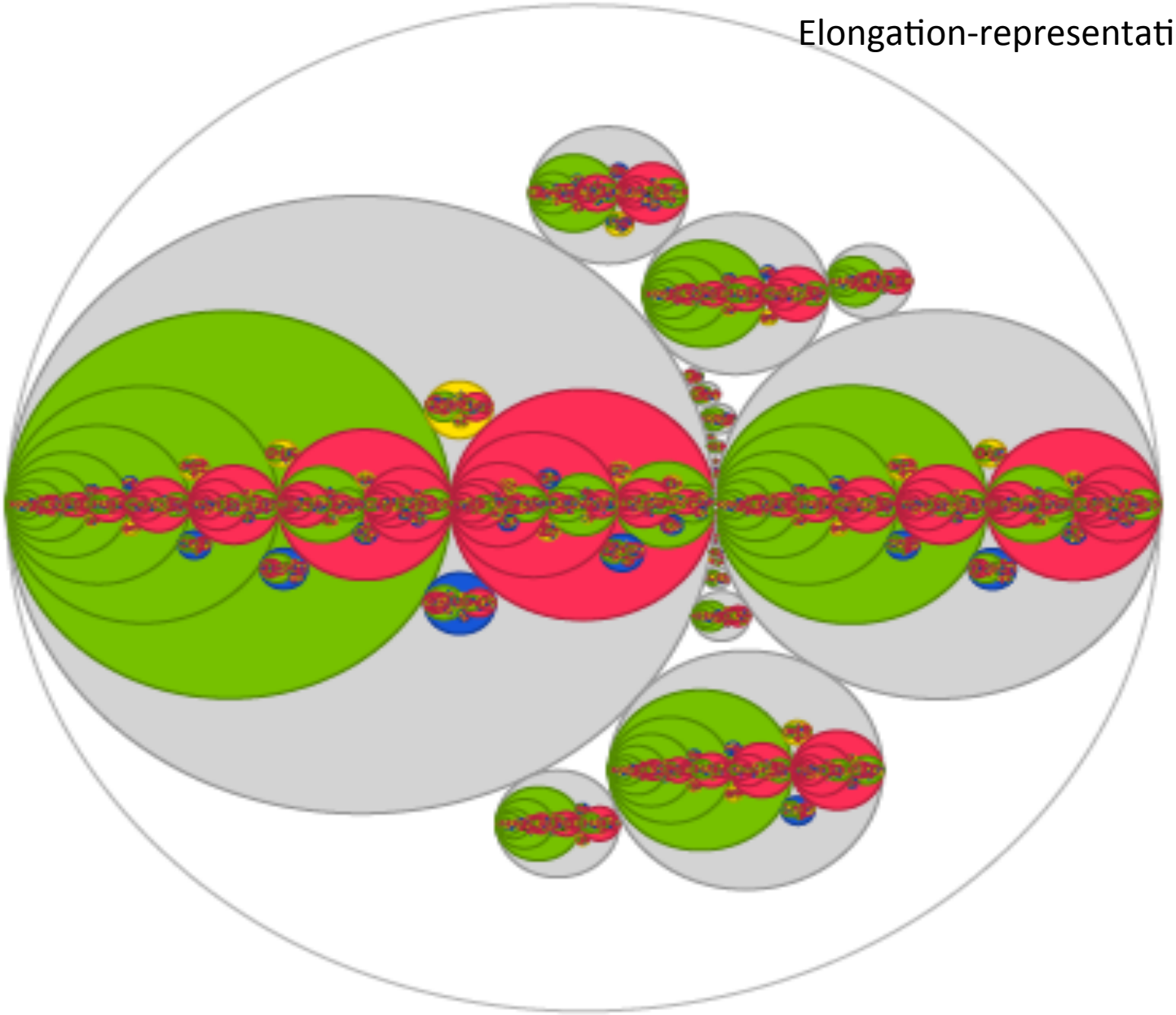
Elongation trees (hyperbolic trees, XML trees)



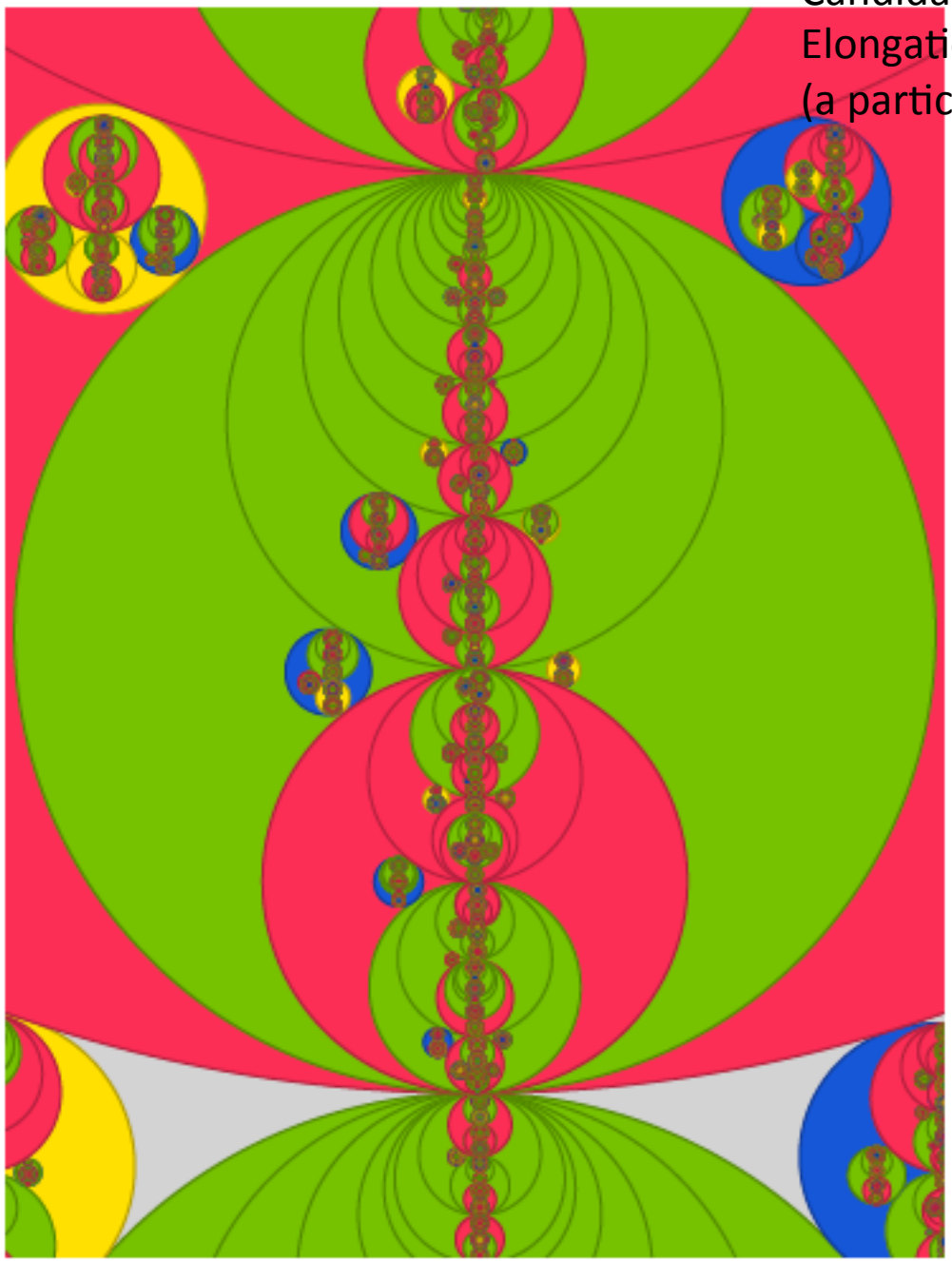
283 bp

⟨GCTGAC⟩ ⟨4, 22, 0, 1⟩ ⟨0, 0, 2, 2⟩ ⟨2, 0, 0, 20⟩ ⟨0, 0, 1, 0⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨1, 0, 0, 0⟩ ⟨0, 1, 1, 0⟩ ⟨EOF⟩ ⟨2, 3, 14, 0⟩ ⟨0, 0, 1, 0⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨4, 10, 0, 0⟩ ⟨0, 1, 0, 0⟩ ⟨1, 1, 1, 1⟩ ⟨1, 0, 0, 9⟩ ⟨0, 1, 0, 0⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨2, 2, 0, 5⟩ ⟨0, 0, 0, 1⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨EOF⟩ ⟨0, 0, 1, 0⟩ ⟨0, 1, 0, 0⟩ ⟨0, 0, 0, 1⟩ ⟨0, 1, 0, 0⟩ ⟨EOF⟩.

Candidatus Carsonellae ruddii
Elongation-representation k=4



Candidatus Carsonellae ruddii
Elongation-representation k=7
(a particular)



A Conclusion Lesson

From
Biomolecular Reductionism
To
Biomolecular Holism

The astronomy analogy:
Tycho Brahe → Kepler → Newton

The Inventors' Paradox (Polya) and Related Principles
Life thinks BIG (The solution of Eigen's Paradox)

First problem

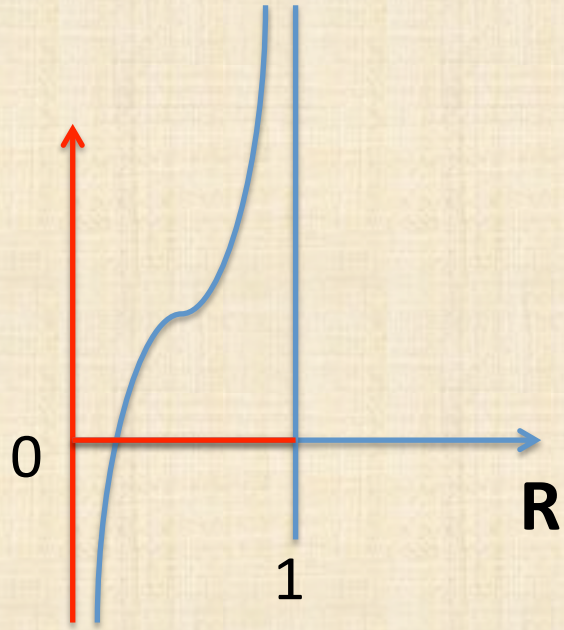
What about comparisons of these indexes-curves between normal and pathological genomes?
What about ENCODE and infogenomics annotations?

Caro Informatico mi aiuti ad elaborare questi dati secondo questi criteri di analisi biologica?

Caro biologo mi aiuti a spiegare queste sorprendenti regolarità matematiche e le altrettanto sorprendenti regolarità delle irregolarità?

The infinity of the egg
(mirror strategy)

A set is **infinite** *if and only if* it is 1-to-1 (bijection) with a proper part of itself (Dedekind).



Galileo's Paradox **N <--> 2N**

0, 1, 2, 3, 4 … **N <--> N' ✚ N''**

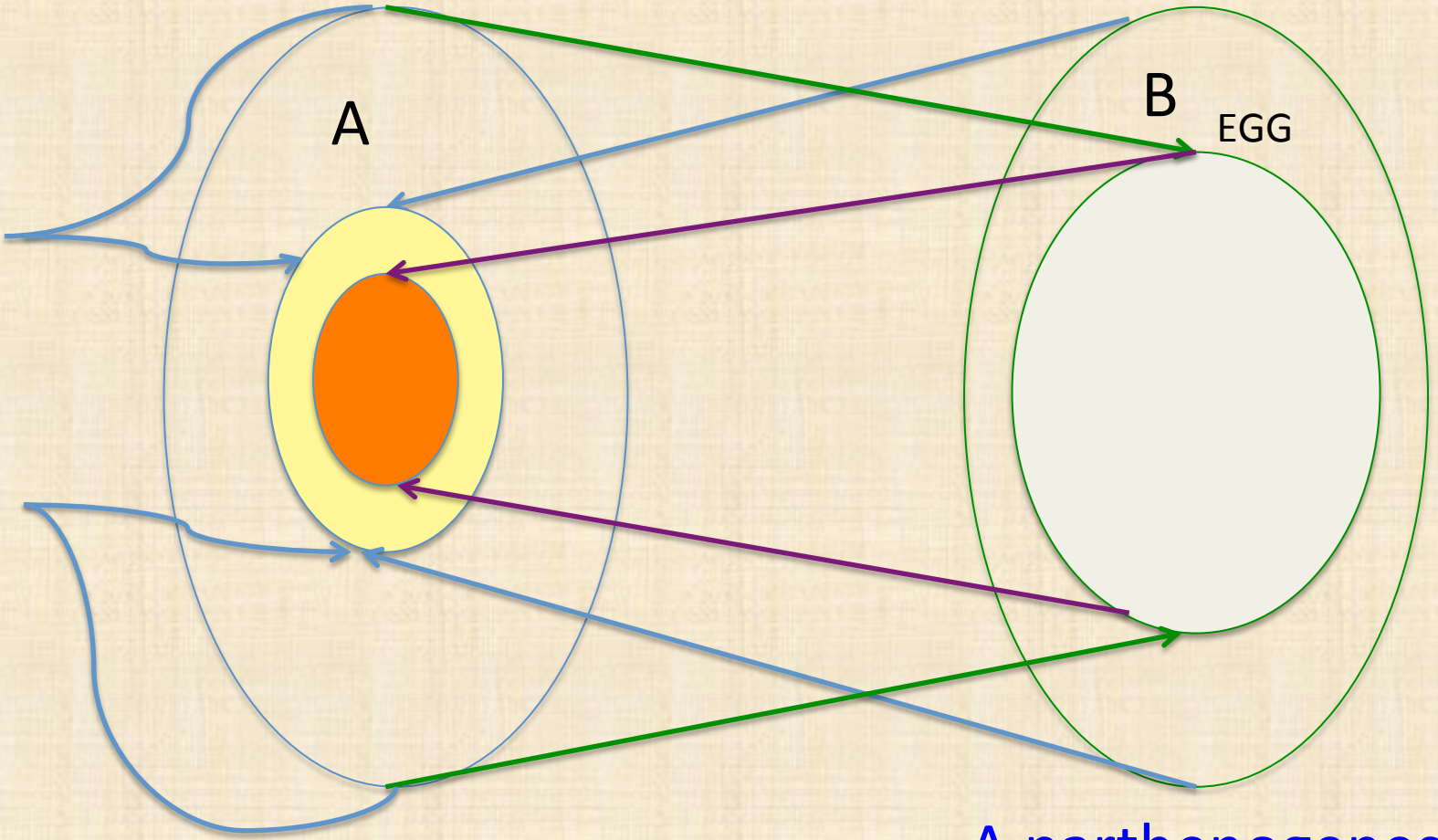
0, 2, 4, 6, 8 … **N <--> N' <--> N''**

R <--> (0,1) <--> $\{0,1\}^N$ <--> P(N) <--> N^N

P(N) <--> N^N

follows from **SB Th.**

A >---> B , B >---> A ➔ A <--> B (*and* A , B infinite)



A parthenogenesis view

Schröder-Bernstein Theorem

An improved version of Halmos'Proof, where yellow part needs to be "duplicated".

Egg's Paradox: Where the trick? Organisms are open!

Egg, *Infinity*, and Computational *Universality*

These concepts are based on mirror-duplication phenomena:

1. An Egg contains the whole information of the organism it is part of

2. An infinite set is 1-to-1 with a proper part of itself

3. A universal computation machine includes the mirror of any machine it is able to simulate, and of its computational universality (its operating system).

4. The famous Turing's 1936 paper discovers the *semidecidability* (creativity) by reformulating Cantor's diagonal argument of infinity.

➔ Metabiology (cell computability)

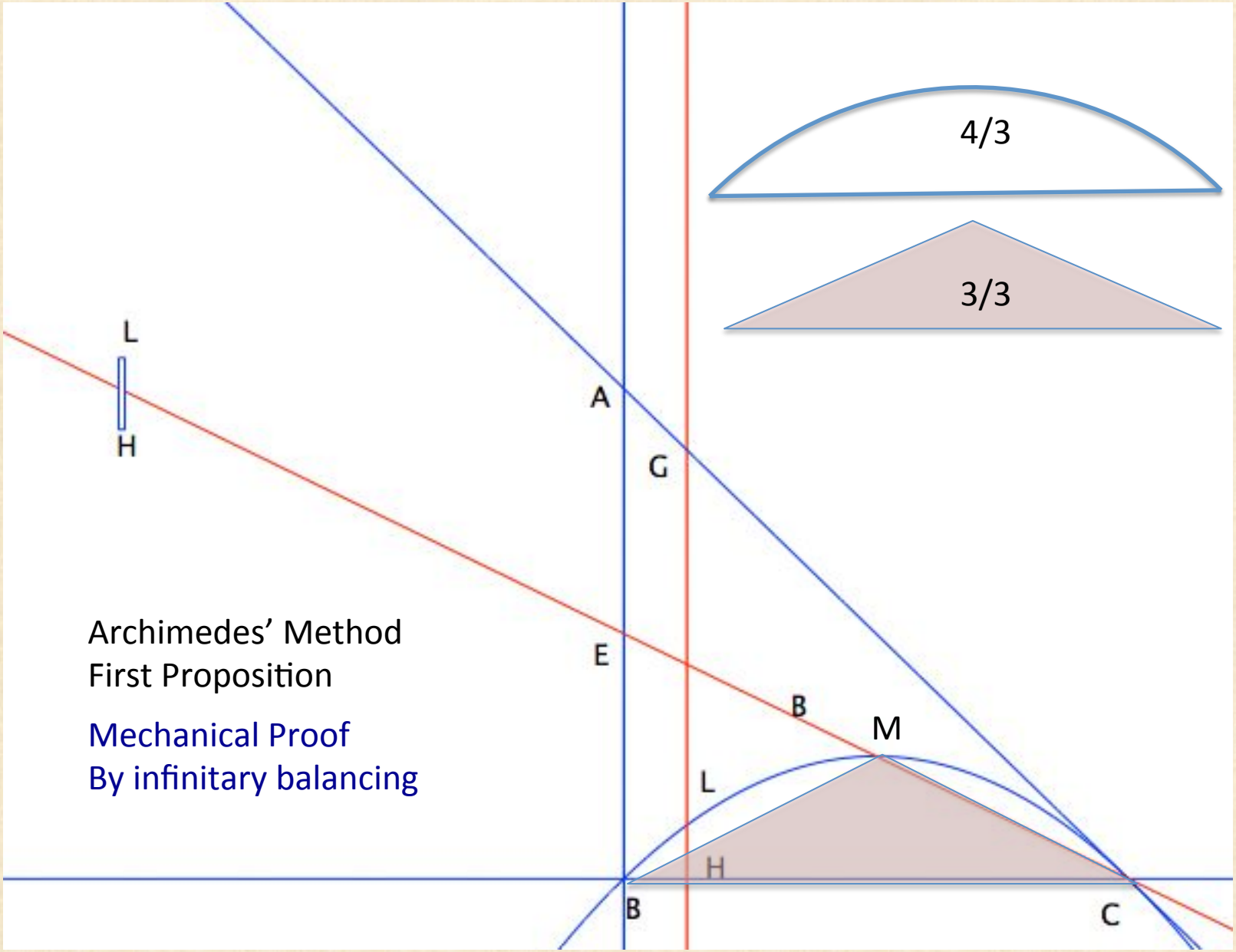
Second Problem:
Where is the genetic kernel?

- Is there, and of what kind is it, a hierarchy among genes?
- Comparing Computer OS with Genomes
  - Epigenetics concerns with I/O procedures
  - Where the OS "kernel" ?
  - Where the structure of sw levels (network levels)?
  - How programs/process duality is realized?
- Genes are regulated, but where regulation is "programmed" or "pre-programmed" (in some way)?
- Programmability - Emergence - Evolvability !!!

  formal models of this interaction could suggest

  new keys to understand genomic functionalities

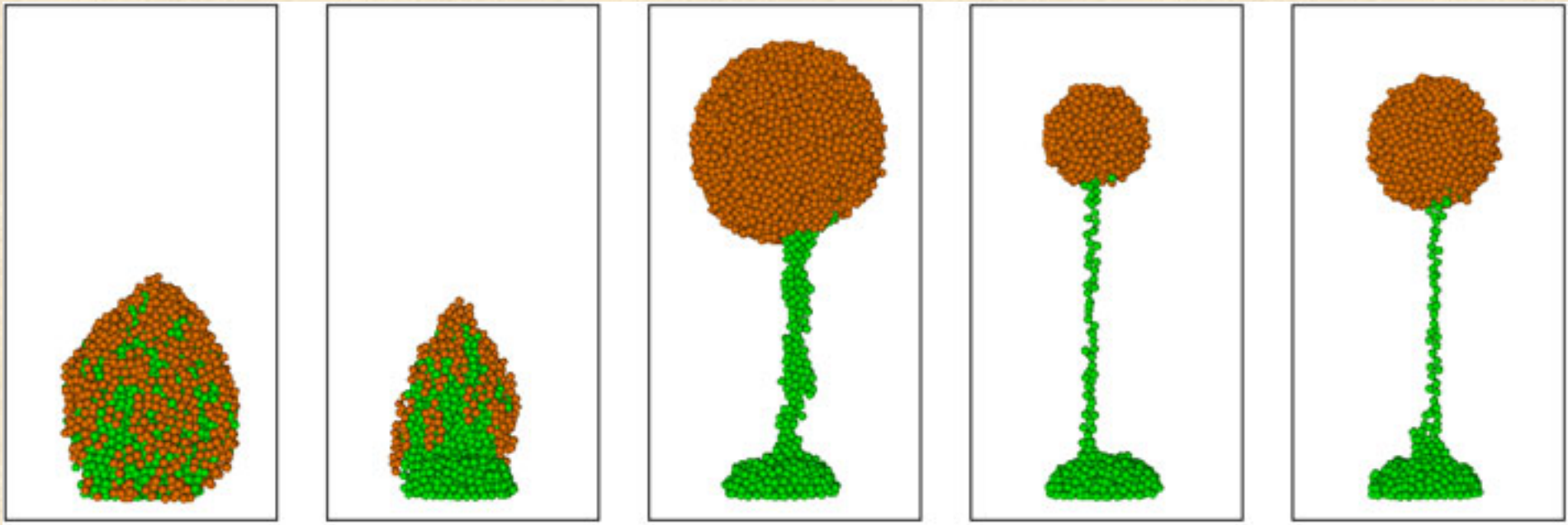
→ Metabiology (Conrad 83-90, Chaitin 2011)



Archimedes' Method
First Proposition

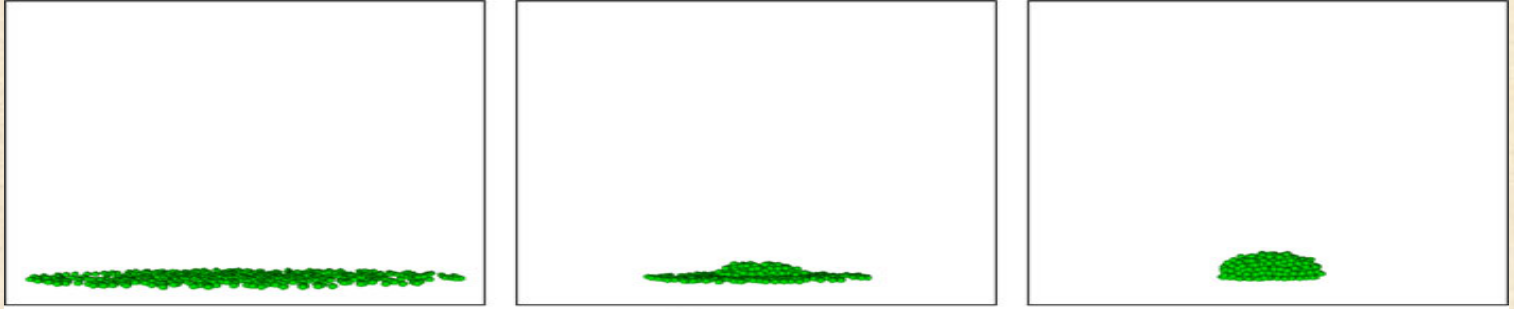
Mechanical Proof
By infinitary balancing

Algorithmic Morphogenesis
computing cell population motions
(25 parameters in an iterative replacement procedure)
Dictyostelium discoideum

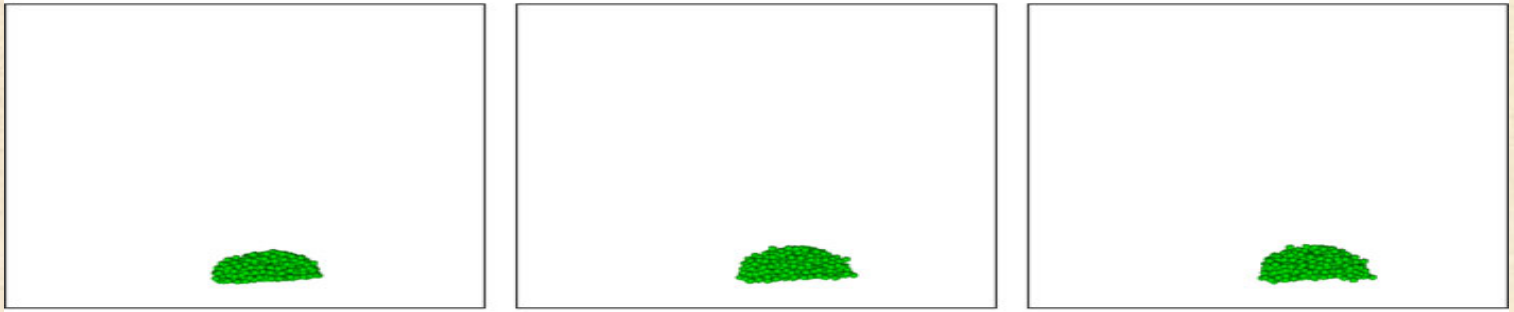


(a) $d_{\text{hid.}} = 1.99$, $t = 940$ **(b)** $d_{\text{hid.}} = 1.995$, $t = 940$ **(c)** $d_{\text{hid.}} = 2$, $t = 2220$ **(d)** $d_{\text{hid.}} = 2.05$, $t = 2220$ **(e)** $d_{\text{hid.}} = 2.1$, $t = 2220$

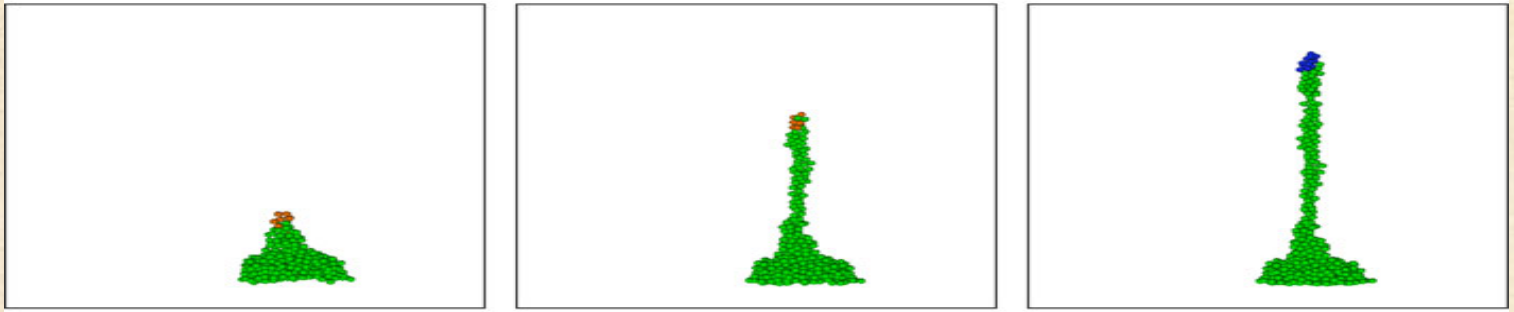
Morphogenesis through moving membranes
Vincenzo Manca • Giovanni Pardini
Natural Computing – Vol. 13, 3, pp. 403-419, 2014



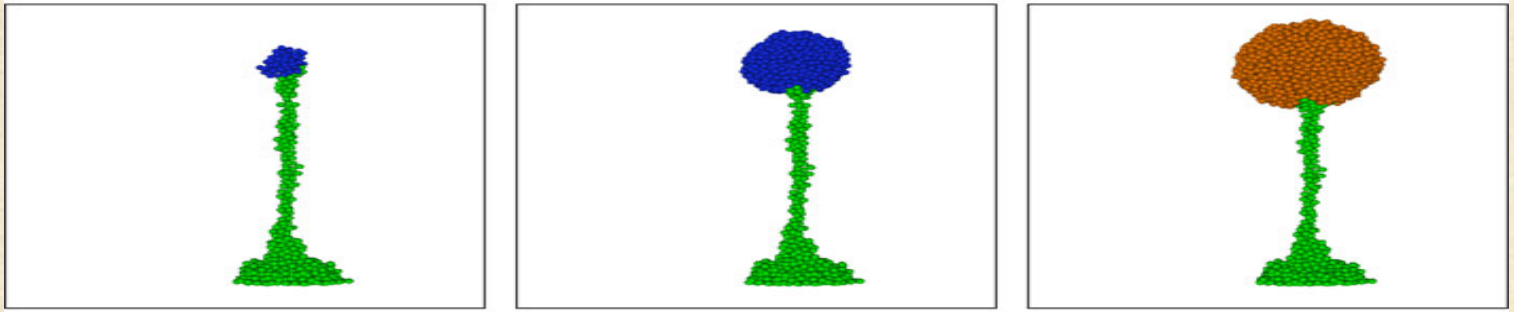
(a) $t = 0$ **(b)** $t = 240$ **(c)** $t = 520$



(d) $t = 600$ **(e)** $t = 760$ **(f)** $t = 1040$



(g) $t = 1280$ **(h)** $t = 1920$ **(i)** $t = 2320$



(j) $t = 2420$ **(k)** $t = 2520$ **(l)** $t = 2800$