Information Theory and Genomes

Vincenzo Manca
University of Verona

Medical Bioinformatics
Natural Computing
Verona, December 2017

Life is ***information represented and processed at molecular level.***

It "has born" when molecules were available to represent and to process information (**polymers and membranes**).

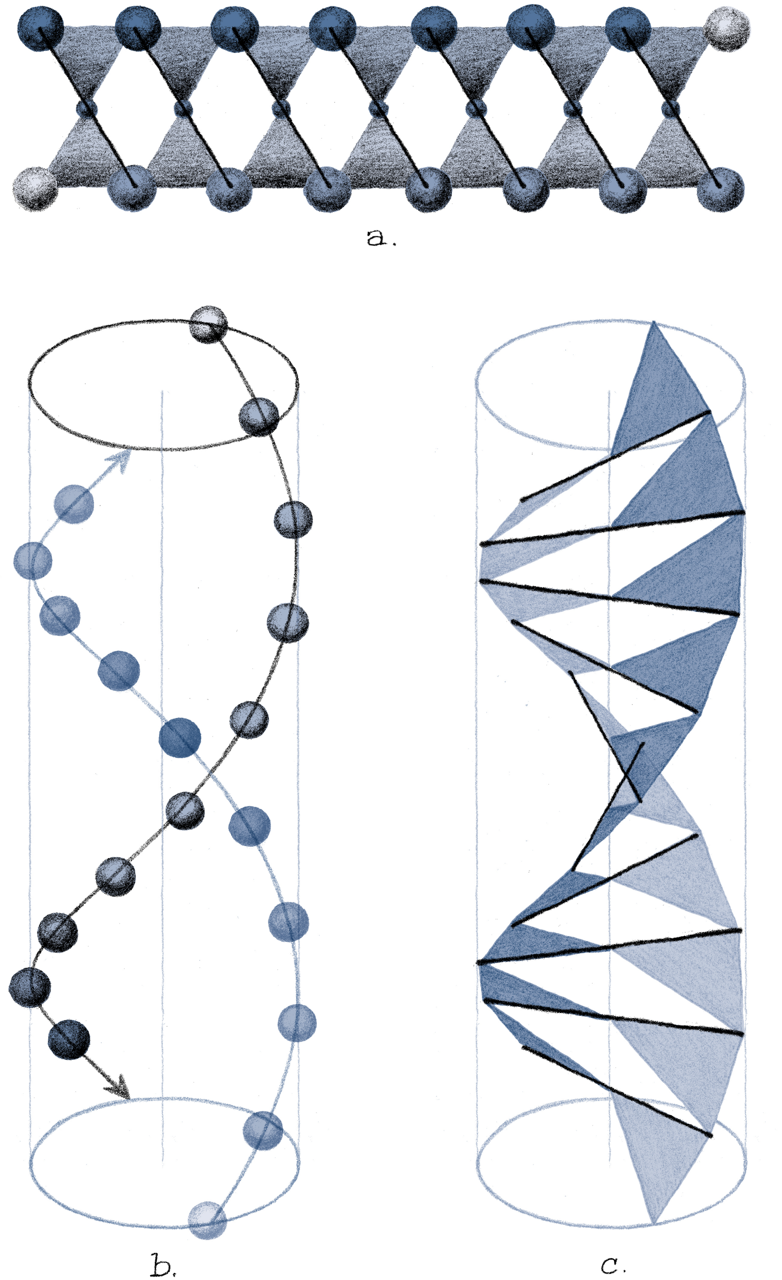
symbolic calculus (*versus numerical/algebraic calculus*) arose in 20° century from Mathematical Logic (formal and automatic information processing):

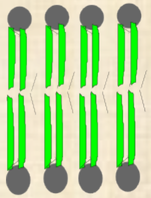
1) There exist processes mathematically definable, but uncomputable. Computation Limits

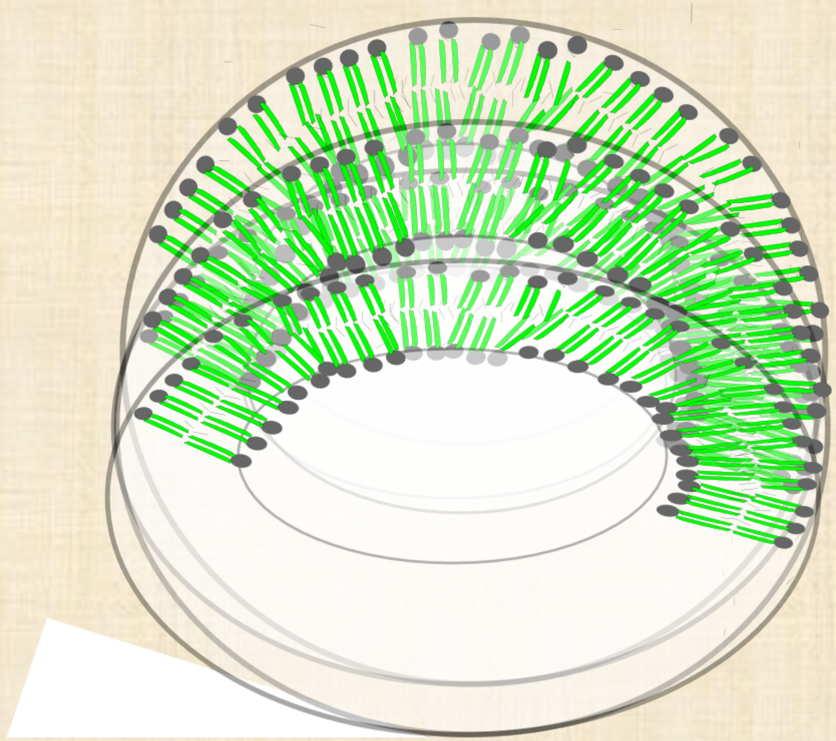
2) There exist universal computation machines able to perform any possible. Computation Power

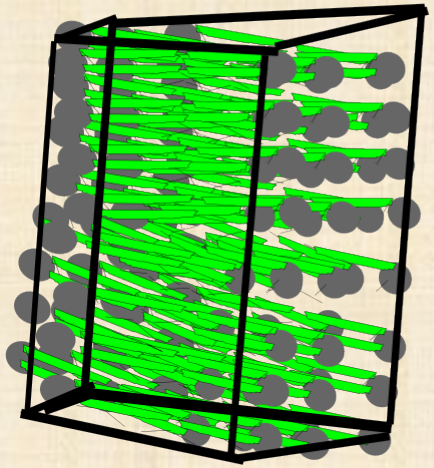
Replication and Universality

- The existence of "universal" computation machines is based on algorithms **of symbolic duplication** (a program is the "mirror" of a machine within another one). Analogously, **biological reproduction** postulates mechanisms of duplication (ds DNA).


a.
b.
c.







Probability and Information

- Shannon 1948: The Information of an event is a function of its probability.

- Probability is distribution (space of events).

Probability

crucial in the scienze since 20° century

- **Cardano e Galileo** : **De ludo aleae**
- **Pascal e Fermat** : **Chevalier de Merè**
- **Jacob Bernoulli** : urn and Bernoulli process : **Ars Conjectandi**
- **De Moivre, Laplace e Bayes** : **gaussian curve and conditional probability**
- **Gauss** : The chance law**: errors follow the gaussian curve**
- **French, Russian, Italian Schools** (Poisson, Cauchy, Borel, Chebicev, Kolmogorov, Cantelli, De Finetti): **distributions, measure theory, laws of large numbers**
- **Boltzmann (Statistical Mechanics**)
- **English Statisticians** (Galton, Pearson, Student, Fisher) **Mathematical Statistics**

Probability Pitfalls

- A pilot has 2% probability of dying in each mission, what is the **probability of dying** in 50 missions?
- 2% x 50 = 100% **ERROR !**
- The same error of the game Chevalier de Meré pointed to Pascal
- The pilot has dead at n-th mission iff he survives in all the previous ones. Summing up from 1 a 49, with p = 0,02 we get:

$$p + (1-p)p + (1-p)^2p + \dots + (1-p)^{49}p = 1 - (1-p)^{50} = 0,64$$

$(1-p)^{50}$ is the probability of surviving up to 50° mission

Modus essendi / Modus conjectandi

- Which way things are?
- Which is the probability that things are in a given way?

Information Theory

- Communication (Hartley, Nyquist, Shannon)
- Coding Theory (Fano, Hamming, Reed, Solomon)
- Cryptography (Hellman, Rivest, Shamir, Adleman)
- Complexity (Kolmogovov, Chaitin) Computation, Chaos
- Cybernetics (Wiener, von Neumann, Langton)
- Foundations (Brillouin, Bennet, Landauer)
- Canonical Quantum Gravity (Wheeler, De-Witt)
- Metabiology (Conrad, Chaitin)

Unification via Information (Carlo Rovelli's books)

Universe's ultimate mechanism for existence might be Information: "it from bit" (Wheeler's last speculation)

Distribution - Information

- X variable assuming values with some multiplicities:

  x_1 , x_2 , x_3 , ….. n_1 , n_2 , n_3 , …..

- If $n = n_1 + n_2 + n_3$, …..
- n_1/n , n_2/n , n_3/n , ….. are frequencies
- p_1 , p_2 , p_3 , ….. . are probabilities (measures of possibility of occurring)
- Shannon calls (X, p) **Source of Information**
- **$-\lg p_e$** is the measure of the information of event **e** with probability **p_e**

Information Paradoxes

Choice, Uncertainty, Information ???

Section 6 of Shannon's booklet

(compare to: Learning/Ignorance/Knowledge)

The paradox is intrinsic to the notion of Event (someting that happens).

The **uncertainty** of E, before it happens, corresponds to the loss of uncertainty, that is, its **information**, when it happened. Both of them correspond to the number of events among which it was **chosen** to happen.

Shannon’s Approach
(Al Kindi’s intuition)

The meaning of a letter in a text is given by its frequency (Caesar Encoding breakdown)

Shannon – The Mathematical Theory of Communication (shannon48.pdf)

Cover & Thomas - Information Theory , Wiley, 1991

Boltzmann’s Tomb

The epochal formula

Thermodynamic Entropy

Carnot's Theorem

A thermodynamic machine between two heat sources:

M (boiler) at temperature T and M_o (condenser) at temperature T_o , with $T > T_o$, taking heat Q from M and giving heat Q_o to M_o and transforms $Q-Q_o$ into mechanical work. In the best efficient machines:

$$Q_o / T_o = Q/T$$

When $T_o = 1$ Q_o is called **entropy (Clausius) denoted by S** therefore **S** = **Q/T** is **minimum heat that M can release** to a condenser at unitary temperature

(Proof: via reversible machines, automata theory style)

Limit to the efficiency of thermodynamical machines

The Second Principle of Thermodynamics

In any isolated system (with no energy exchange with the external world):

$$\Delta S \geq 0$$

$$S_{t+1} - S_t \geq 0$$

$$S_{t+1} \geq S_t$$

Where does “>” come from?

How this relates to Newton Mechanics where laws are equations?

Time irreversibility as probabilistic consequence of complexity

Boltzmann: **S** of Carnot **is** proportional to the **logarithm of** the number **W** of microstates of the thermodynamical macrostate.

GAS

Microstate = position and speeds of all molecules

Let n be the number of particles and k their classes of velocities:

$$n = n_1 + n_2 + ...n_k$$

$S = k \lg W$ (***)

- $W = n! / n_1!\, n_2! \ldots\ldots n_k!$

n_i = number of particles with velocity in the interval i (the whole range of velocity is split in k intervals)

- From Stirling $\lg n! \approx n \lg n$

- $\lg W \approx n \lg n - (n_1 \lg n_1 + n_2 \lg n_2 \;\ldots. + n_k \lg n_k)$

- $S = -k(n_1 \lg n_1 + n_2 \lg n_2 \;\ldots. + n_k \lg n_k) + C$

The impossible theorem

H-Theorem (Boltzmann)

$$\mathbf{H = \Sigma_i \ n_i \ lg \ n_i}$$

H is the discrete microscopic version of thermodynamical entropy (apart: the sign and additive, multiplicative costants).

H-Theorem (1872) In a isolated system:

$$H(t) \geq H(t+1)$$

From Boltzmann to Shannon

$$H_s(X, p) = - \Sigma_x \; p(x) \lg p(x)$$

Shannon 1948

Entropy Th. H_S is completely characterized by 3 conditions:

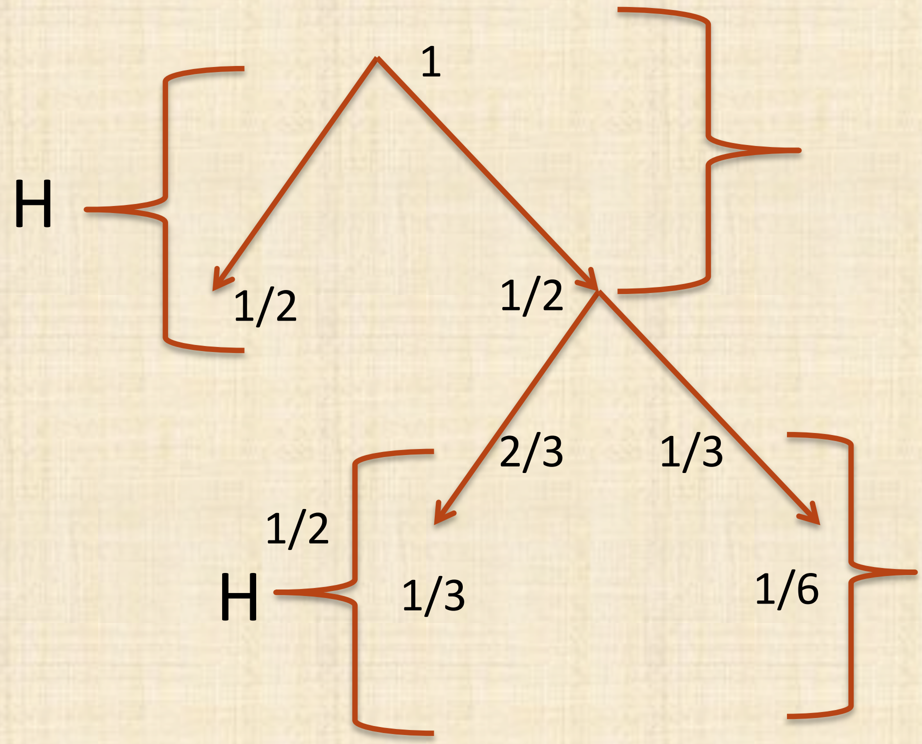
Continuity in p ,

Maximum in p= 1/n ,

Additivity of choices:

$H(1/2 , 1/3 , 1/6) = H(1/2 , 1/2) + 1/2 \, H(2/3, 1/3)$

The additivity of Choices



Shannon Game (abridged version)

Let **X** be a discrete random variable and ask someone, knowing the distribution of **X**, to guess a xalue **x** of **X**, by using the optimal dichotomy strategy (choosing equi-probable intervals): **$x \leq x_0$ or not?** with $P(x \leq x_0) = P(x > x_0$, (Yes/Not answers).

The minimum number of questions that, in average, is sufficient for guessing correctly **x** coincides with the entropy H(X) of the variable X.

- H and H_s are essentially the same thing (von Neumann: suggested the name), En-tropos (internal verse)
- From $inf_i = - \lg p_i$ follows that:

H_s is average information of the source $S = (X, P)$

Double Entropies
(X, p), (Y,q)

- $H(X \times Y) = - \Sigma_{x,y}\ p(x)q(y)\ \lg p(x)q(y)$ product / independent joint
- $H(X \wedge Y) = - \Sigma_{x,y}\ p_{\wedge}q(x,y)\ \lg p_{\wedge}q(x,y)$ joint
- **Very often** $H(X \wedge Y)$ is denoted simply by $H(X, Y)$

$p_{\wedge}q$ requires joint variables (each marginal of an $s(x, y)$, i. e. :

$p(x) = \Sigma_y\, s(x, y)$
$q(y) = \Sigma_x\, s(x, y)$

that is, x and y have the same dependence set,
for ex., height/weight over a population of individuals

Conditional Entropy

- $H(X \mid Y) = - \Sigma_{x,y}\, p_{\wedge}q(x,y)\ \lg\ p|q(x,y)$ conditional

  $p|q(x,y) = p_{\wedge}q(x, y) / q(y)$ conditional probability

- $H(X \mid y) = - \Sigma_{x}\, p|q(x,y)\ \lg\ p|q(x,y)$
- $H(X \mid Y) = \Sigma_{y}\, q(y)\, H(X|y)$

Joint and Conditional Entropies

- $H(X \wedge Y) = H(Y) + H(X|Y)$

- $H(Y \wedge X) = H(X) + H(Y|X)$

Mutual Information

$$I(X, Y) = H(X) - H(X|Y)$$

Is the information that a source gains with respect to another source, that is, the difference of its average information minus the mean of its information conditioned to the values emitted by Y.

Mutual Information and Entropies

$$
\begin{aligned}
I(X, Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X \wedge Y) \\
&= H(X \wedge Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

Mutual information is symmetric, Zero-diagonal (that is, $I(X,X) = 0$), it is not triangular.

Last two equations follow from the joint/ conditional entropies relationships.

Mutual information

$$I(X, Y) = DIV(X_{\wedge}Y\ ,\ X_{x}\ Y)$$

Sender X === Channel ===> Receiver Y

Noise alters data along the channel

What is the information amount that can pass correctly?

Entropic Divergence

$$DIV_{KL}(X , Y) = \Sigma_{x\varepsilon X,\ y\varepsilon Y}\ p(x) \lg [p(x) / q(y)]$$

Mean information difference between distributions (Kullback , Leibler 1951).

How applying this definition to the case of genomic distributions?

We need joint variables!

H theorem is an information theory theorem

1) Maxwell already proved that velocities reach normal Distribution (as a consequence of cause normalization).

2) Elastic collisions guarantee that variance of speed distribution remains constant
(Pytagorean game keeps variance distribution constant).

3) The Gaussian curve is the distribution having maximum Entropy within the class of distributions with a given variance.

****

Information Theory and Genomes

Vincenzo Manca

SECOND PART

(Information Sources and Codes)

From Information Sources to Encoding and Transmission

- Variable values become “Digital Data” via Codes
- Encoded Data are transmitted with a second encoding, **channel encoding** for reducing error transmission

Codes

c : C → D surjective

C strings over A (encodings or codewords, C and **c** will be often identified). D set of data

Only 1 **datum corresponds to** a **codeword** corresponds

Two **code-words** can encode the same datum (genetic code). A code is *redundant* if **C** is injective
(non redundant otherwise)

Types of codes (recovering codewords from a stream)

- Univocal: any string is factorizable in only one way by means of codewords of C
- Instantaneous or prefix-free: no encoding is prefix of an other encoding
- Auto-delimitative: any code-word **w** includes the specification of its length (a prefix of **w** tells the length of **w**)
- Fixed length

Kraft Norm

Let k be the alphabet size

C code over the alphabet

$$|C| = \Sigma_{x \varepsilon C} k^{-|x|}$$

McMillan – Kraft Theorems

- **Th.** McMillan : C univocal iff $|C| \leq 1$

- **Th.** C univocal ➔

exists C' istantaneous t. c. $|C|=|C'|$

(Proof by construction)

Let C be code of a source (X, P)

$$L_C = \Sigma_{w \varepsilon C} |w| p(w)$$

L_C is the average length of C

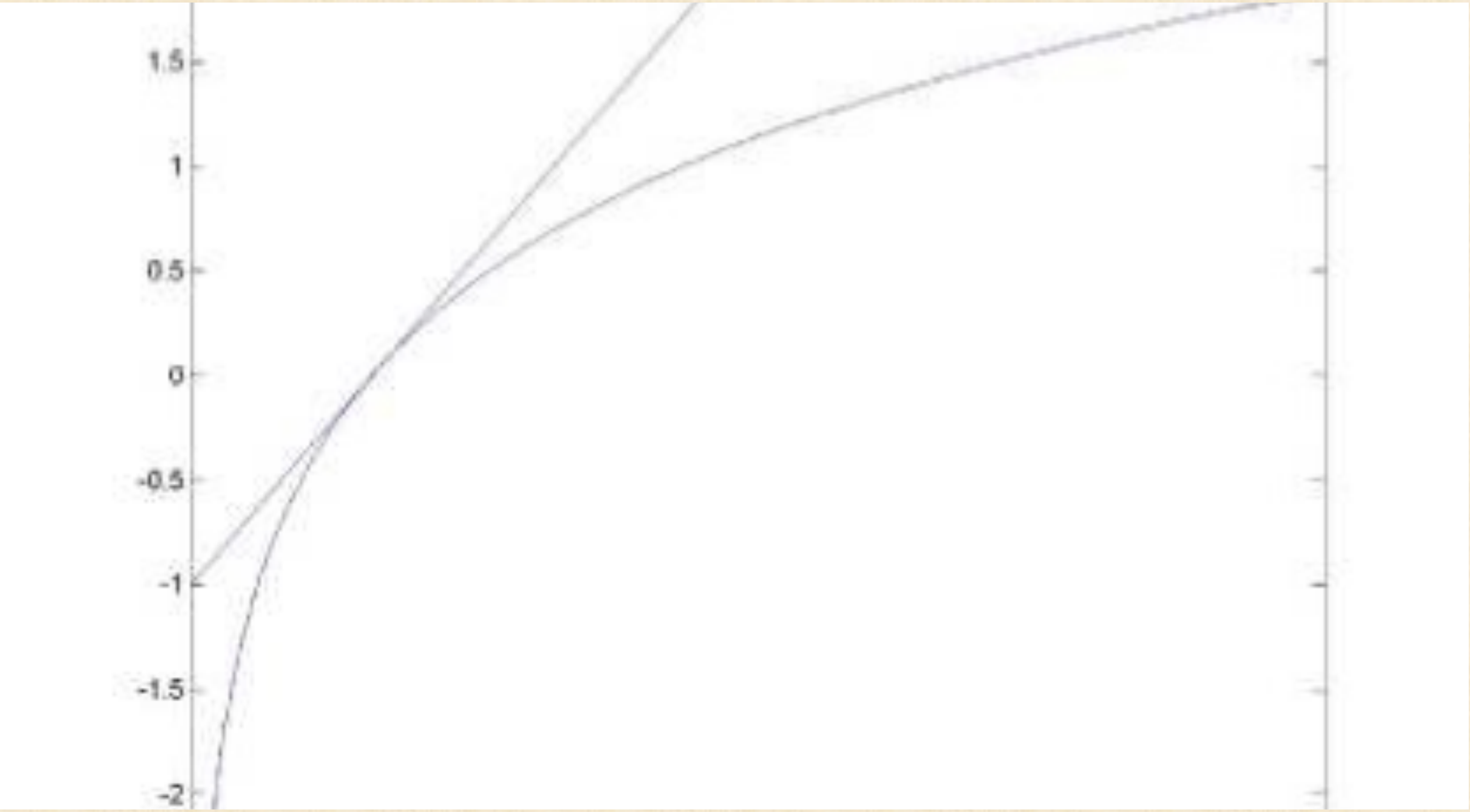
C is optimal if no C' exists with $L_{C'} < L_C$

First Shannon Theorem

$$H(X, p) \leq L_C$$

No code of a source can reach an average encoding length smaller than the entropy of the source

Logarithm Lemma



- $X = q_i / p_i$
- $\ln q_i / p_i \le q_i / p_i - 1$
- and multiplying both members by p_i and summing we have:
- $\Sigma\, p_i \ln q_i / p_i \le \Sigma\, p_i (q_i / p_i - 1) \le 0$
- whence:
- $\Sigma\, p_i \ln q_i / p_i \le 0$
- $\Sigma\, p_i \ln q_i - \Sigma\, p_i \ln p_i \le 0$
- $\Sigma\, p_i \ln q_i - \Sigma\, p_i \ln p_i \le 0$
- $\Sigma\, p_i \ln q_i \le \Sigma\, q_i \ln q_i$
- $- \Sigma\, p_i \ln q_i \ge - \Sigma\, q_i \ln q_i = H$

First Th. Proof

- $H(A, p) = -\Sigma_{a \in A}\ p(a) \log_k p(a)$
- $\leq -\Sigma_{a \in A} -p(a) \log_k q(a)$
- $= -\Sigma_{a \in A}\, p(a) \log_k k^{-|c(a)|} /\ /\!/\ C\ /\!/$
- $= -\Sigma_{a \in A}\, p(a)(\log_k k - |c(a)| - \log_k\ /\!/\ C\ /\!/\ )$
- $= -\Sigma_{a \in A}\, p(a) \log_k k^{-|c(a)|} - \Sigma_{a \in A}\, p(a) \log_k\ /\!/\ C\ /\!/$
- $= \Sigma_{a \in A}\, p(a)\ |c(a)| - \log_k /\!/\ C\ /\!/$

Therefore if C univocal $/\!/\ C\ /\!/\ \leq 1$ whence

- $\log_k /\!/\ C\ /\!/\ = K > 0$, that is:

$$H(A,p) \leq L_C + K$$

whence:

$$H(A,p) \leq L_C$$

Typical sequences

- A sequence is typical for a source (X, p) if the the frequency of any symbol in the sequence coincides with its probabiliy p in the source

- **Th:** The number of typical sequences of length n of (X, p) is $2^{nH(X)}$ and the probability that a sequence of length n is typical for (X, p) is
- $2^{-nH(X)}$

The number of Typical Sequences

- $\log p(\alpha) = \log(p_1^{Np1} \cdot p_2^{Np2} \cdots\cdots p_m^{Npm})$ (N is the length) whence
- $\log p(\alpha) = Np_1 \log p_1 + Np_2 \log p_1 + \cdots + Np_m \log_{pm}$

$= -NH$

Therefore

- $p(\alpha) = 2^{-NH}$
- Typical $= 2^{NH}$

REMARK

We are speaking of simple information sources

Shannon's 2° Th.

The theorem provides conditions to transmit with error probability tending to zero, avoiding transmission errors (autocorrecting codes).

Transmission Rate

Given a Transmission fixed length code where are transmitted M different messages with codewords of length n, the transmission rate R is given by $(M < n)$

$$R = \lg M/n$$

whence, for a binary alphabet $2^{nR} = M$

Capacity

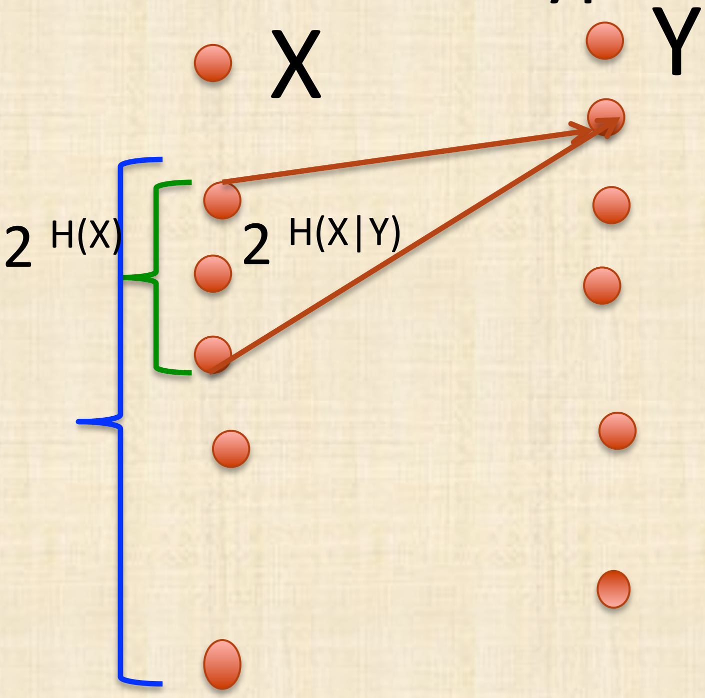
The capacity C of a channel X → Y conecting a Sender Source to a Receiver source:

$$C = \max_S I(X, Y)$$

Where S is the set of possible sources definable over X (or the possible probability distributions definable over variable X)

2° Th: If R < C E → 0 *for* n → ∞

Consider X-typical and Y-Typical (binary alphabet)



$$W = 2^{H(X|Y)} / 2^{H(X)} = 2^{-\mathbf{I}(X,Y)}$$

I = mutual information

W is the average probability that a X-typical transmits a Y-typical. The probability of error is: the number of wrong messages M -1 (all possible messages minus the correct one) multiplied by the average probability W:

$$(2^{nR}-1)W < 2^{nR} W \leq 2^{nR} 2^{-n\mathbf{I}} \leq 2^{nR} 2^{-n\mathbf{C}} = 2^{-n(C-R)} \rightarrow 0$$

Informational Genomics

- An information source (X, p) is a discrete probability distribution (Shannon 1948)
- Let X_G be a variable varying along genome components (positions, segments, strings, …)

How many times $X_G = a$?
p(a) = the frequency of the event $X_G = a$

(X_G , p) is a Genomic Information Source GIS extracted from G

$I(x) = 1/\log_2 p(a) = - \log_2 p(a)$ is the information quantity of a

$E(X_G) = \Sigma_x\, p(x)\, I(x)$ = mean information of (X_G , p)

Infogenomics
An Informational Approach analogous to Genomes
(ENCODE)

- Distributions
- Dictionaries
- Indexes
- Elongation
- Segmentation
- Representation
- Entropies and related notions
- Recurrence
- Randomness
- Para/Meta/Iper-Genomes

- Important genomic distribution are based on genomic dictionaries on genomes, in particular, $D_k(G)$.

- Using the distribution of k-mers we define $E_k(G)$ (w in D) by:

$$\mathbf{E_k(G) = \Sigma_W\ p(w)\ lg\ p(w)}$$

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

$$E_k(G) = -\Sigma_{w \varepsilon D(G), |k|=k}\, prob(w) Inf(w)$$

k-Entropy is the mean information of a genome as information source of k-mers.

- We computed Empirical Entropy for any word length, and for all Human chr.
(k= 18 , $E_k \approx 24$; k=200 $E_k \approx 25$ **!!!**)

Algorithmic basis of k-mer frequency computation

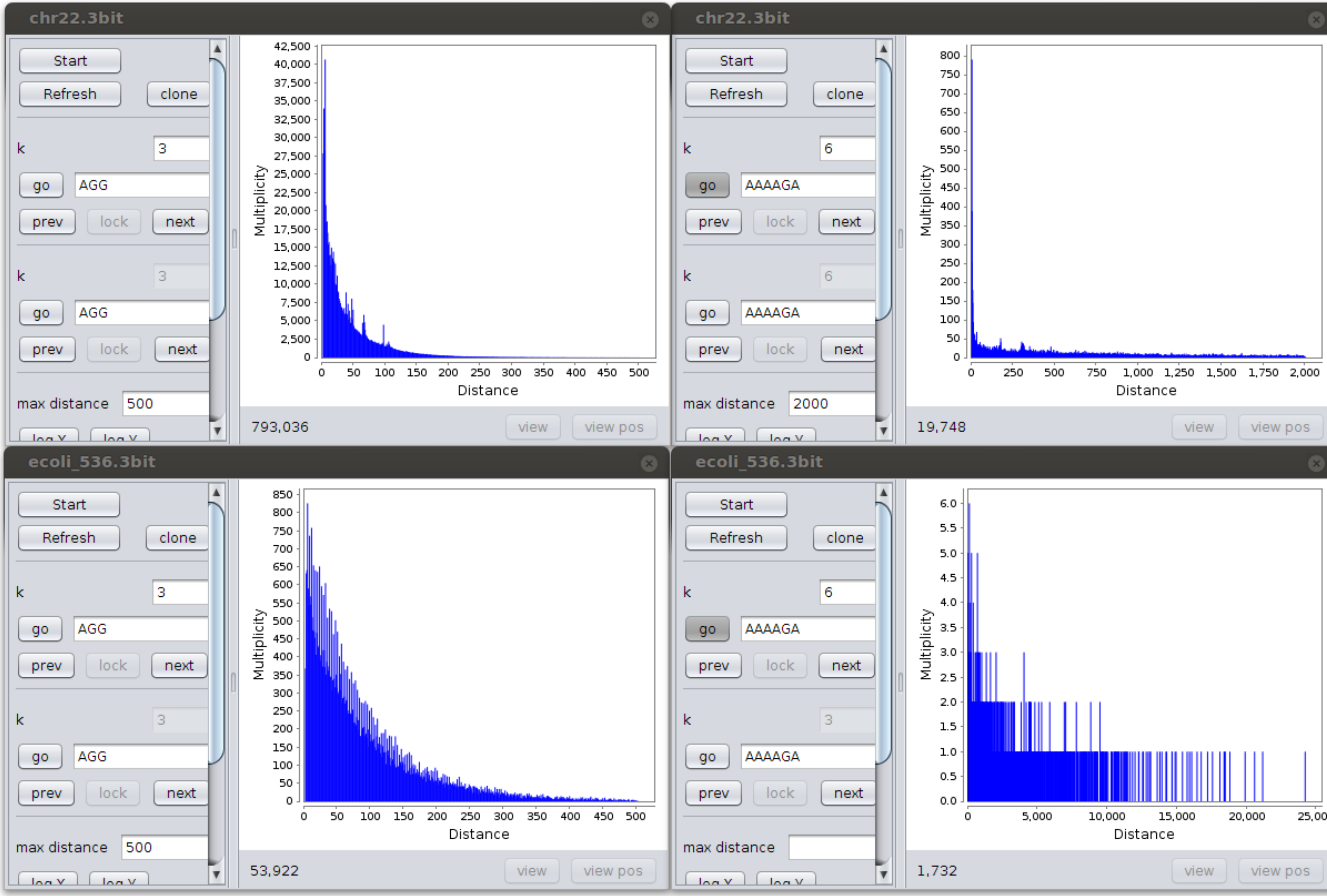
Bonnici V, Manca V – IGTools, J. of Bioinformatics and Proteomics, 2015

- 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

Genomic Distributions

- **Multiplicity** (how many times words of D occur)
- **coMultiplicity** (how many words have a given m-plicity)
- **Segment-Multiplicity** (w.r.t. D and a segment length)
- **Segment-coMultiplicity** (w.r.t. D and a segment length)
- **Segment-Lexicality** (w.r.t. D and a segment length)
- **RDD** (how many times a mer recurrs at a given distance)
- **Repeat-Length** (how many repeats of a given length)
- **Duplex-dist** (how many duplexesat a given distance)


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

Information Correlation and RDD in Genomes

- Trifonof et al. : DNA correlation periodicities, 1980
- Shepherd : DNA periodicities in coding regions, 1981
- Eigen et al. : periodicity in Transfer-RNA, 1981
- Fickett :1982 non min. RDD periodicity in coding regions, 1982
- Li : Mutual information in DNA Strings, 1990
- Herzel et al. : Measuring DNA correlations, 1990
- Li internal correlation in DNA, 1997
- Herzel-Weiss-Trifonof : 10-11 Periodicity, 1999
- Afreixo : 1-RDD min. 2009
- Bastos : 2-RDD min. 2011
- Carpena et al. RDD in keywords finding (non DNA), 2009-2013
- Computational Chemistry 2014

Important classes of k-mers

- **Repeats** $\exists$ i,j,i',j' G[i, j] = G[i', j'] with i≠i' , j≠j'
- **Duplexes** $\exists!$ i,j,i',j' G[i, j] = G[i', j'] i≠i' , j≠j'
  often correctly parenthesized: no ([ )]
- **Hapaxes** $\neg\exists$ i,j,i',j' G[i, j] = G[i', j'] i≠i' , j≠j'
- **Creodes** $\exists$ k > 0 G[i, j]=G[i', j'] → G[i, j+k]=G[i',j'+k]
  G[j+1, j+k] with the maximum k is called Creode-tail

Sequencing = Dictionary (of Reads) → G
Repeats give ambiguity in reconstructing G

- G[i, j] and G[j+1, k] are contiguous strings in G
- G[i, j] and G[i+k, m] k-overlap if

  i+k ≤ j and G[i+k, j] is their overlapping string
- A repeat G[i, j] = G[i', j'] longer than k gives k-overlappings determining positions (i+k) and (j'+k) as k-crossing pairs

$$\alpha\omega\boldsymbol{\beta}\omega\boldsymbol{\gamma} \rightarrow \alpha\omega\boldsymbol{\gamma}\omega\boldsymbol{\beta}$$

Distances between Duplexes or Hapaxes can remove ambiguity

Other classes of k-mers

$G[i_1, j_1]$ is **Memer** if $\exists\ i_2, i_3, i_4, j_2, j_3, j_4$:

- $G[i_1, j_1] = G[i_2, j_2] = G[i_3, j_3] = G[i_4, j_4]$
- $\{G[j_1+1], G[j_2+1], G[j_3+1], G[j_4+1]\} = \{A, C, G, T\}$
- $\{G[j_1+2], G[j_2+2], G[j_3+2], G[j_4+2]\} \neq \{A, C, G, T\}$

moreover it is not proper suffix or prefix of a k-mer with this property and any of its substring has this property.

A memer is a maximal maximally elongable k-mer.

If **w** is a memer, then **w** is a repeat and for all **x** = A, C, G, T, **wx** occurs in G, and also the same property holds for all its prefixes.

- **Minimal Nullomers** (shortest non-occurring k-mers)
- **Tandems** **w---w'** (and poly-tandems)

  (with length and/or structure constraints for ---)
- **Anti-Creodes** (creodes w.r.t. right-left elongation)
- **Twin-creodes** (creodes+anticreodes)
- **Double creodes** (duplexes that are also creodes)
- **Free creode tails** (occurring without creodes)
- **Proper creode tails** (occurring only after creodes)

Other classes of k-mers

$G[i_1 , j_1]$ is **Memer** if $\exists\ i_2 , i_3 , i_4 , j_2 , j_3 , j_4$:

- $G[i_1 , j_1] = G[i_2 , j_2] = G[i_3 , j_3] = G[i_4 , j_4]$
- $\{G[j_1+1] , G[j_2+1] , G[j_3+1] , G[j_4+1]\} = \{A, C, G, T\}$
- $\{G[j_1+2] , G[j_2+2] , G[j_3+2] , G[j_4+2]\} \neq \{A, C, G, T\}$

moreover it is not proper suffix or prefix of a k-mer with this property and any of its substring has this property.

A memer is a maximal maximally elongable k-mer.

If **w** is a memer, then **w** is a repeat and for all **x** = A, C, G, T, **wx** occurs in G, and also the same property holds for all its prefixes.

- **Minimal Nullomers** (shortest non-occurring k-mers)
- **Tandems** **w---w'** (and poly-tandems)

  (with length and/or structure constraints for ---)
- **Anti-Creodes** (creodes w.r.t. right-left elongation)
- **Twin-creodes** (creodes+anticreodes)
- **Double creodes** (duplexes that are also creodes)
- **Free creode tails** (occurring without creodes)
- **Proper creode tails** (occurring only after creodes)

Coverage

The **coverage** of a dictionary D can be considered w.r.t. to single positions or to the whole genome

- How many elements of D pass for a given position?

  (at most k if k is the max length of k-mers in D)

- Which is the fraction of positions k of G where are placed words of D ($i \leq k \leq j$ s.t. $G[i,j]$ is in D)?

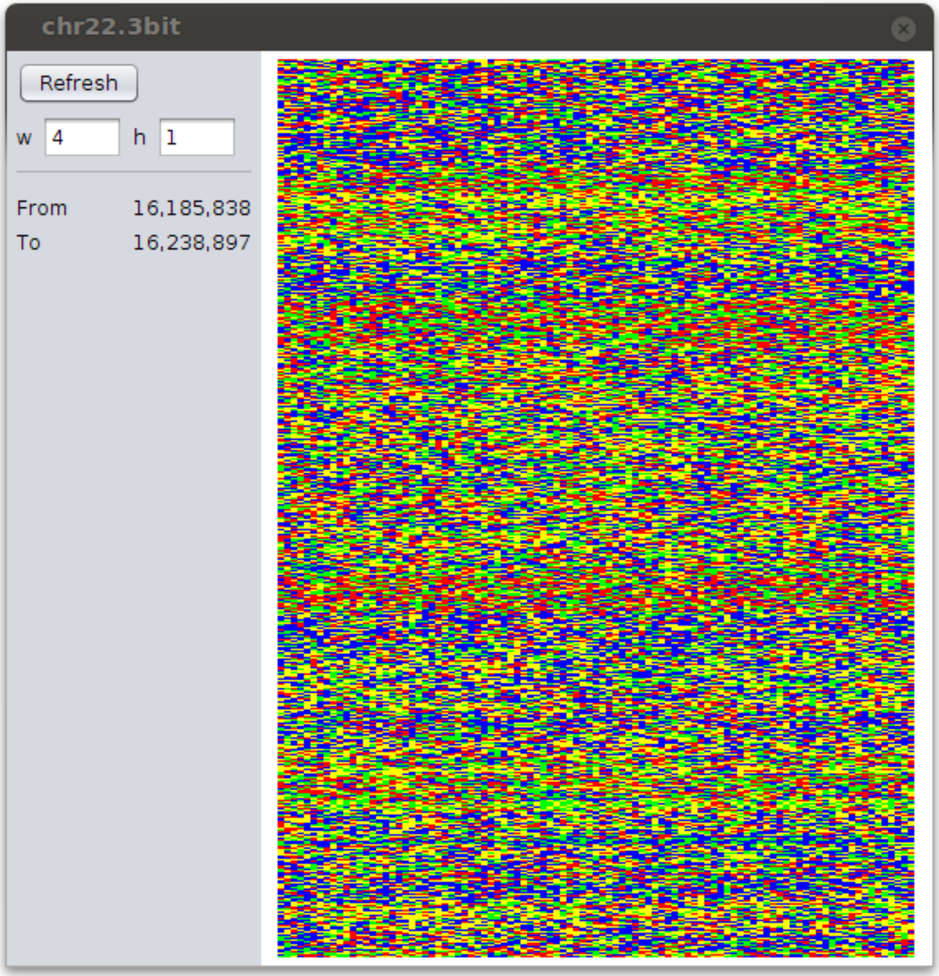
Basic Genomic Indexes

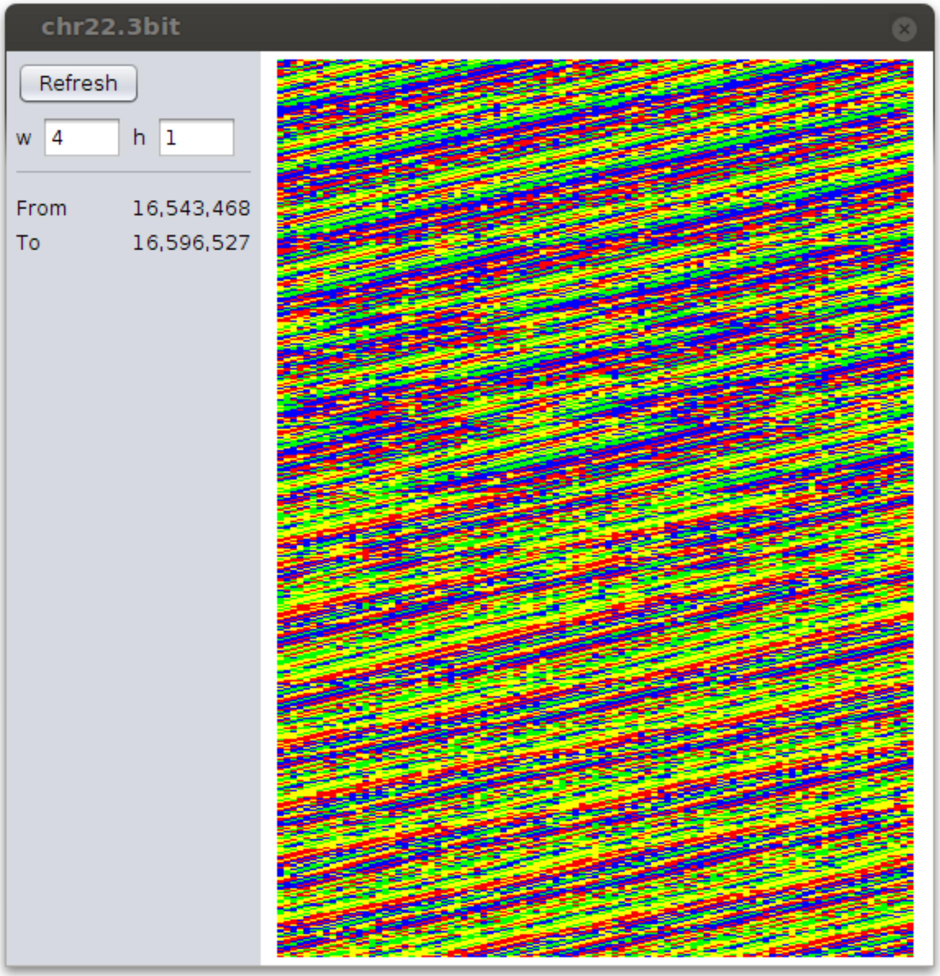
- *LG* — Logarithmic Length (base 4)
- *LX_k* — k-Lexical Multiplicity (how many times k-mers occur in average)
- *MFL* — Minimal Forbidden Length (**MCL** = MFL -1)
- ***MRL*** — **Maximum Repeat length**:

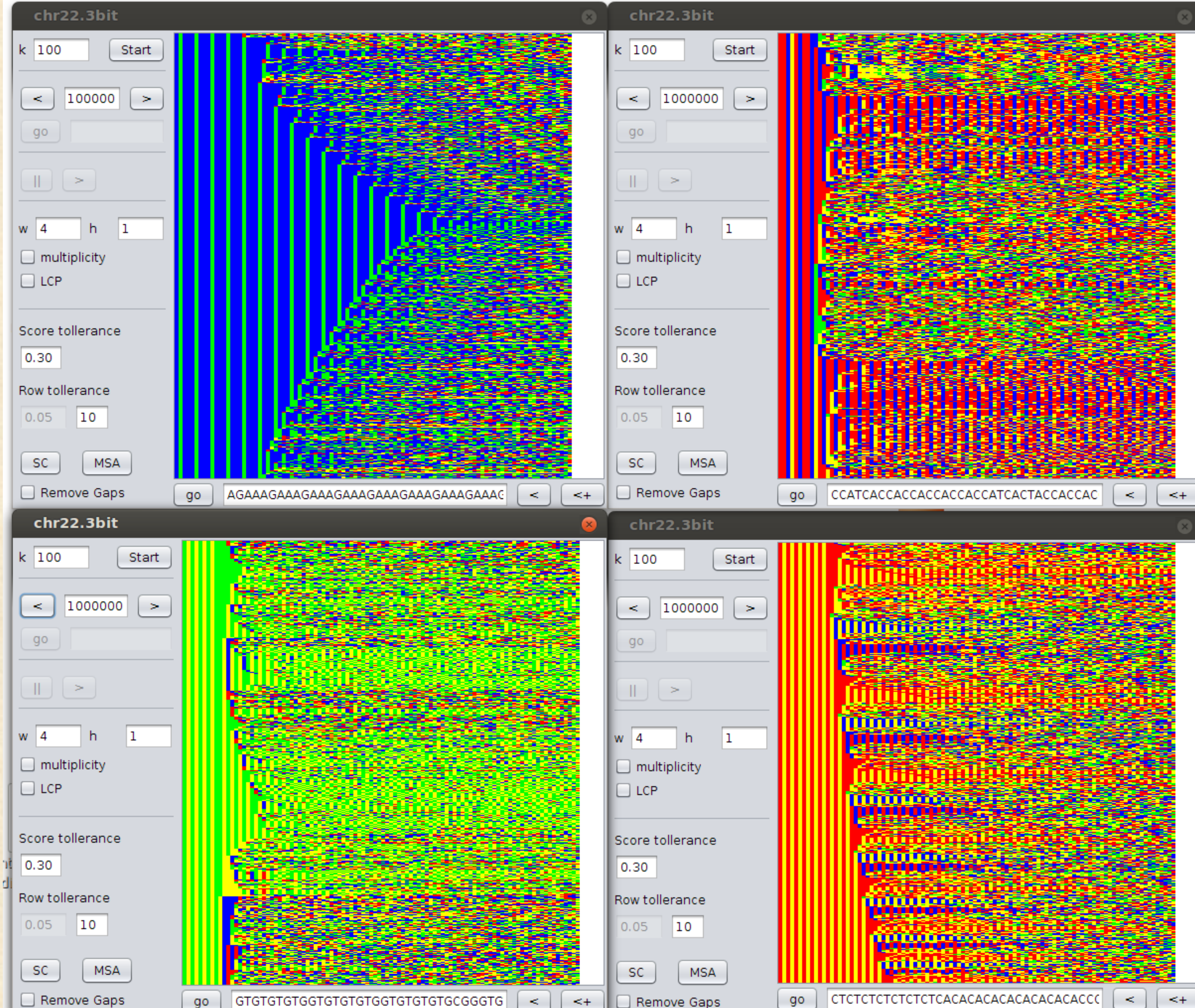
  all the strings of length **MRL+1** are hapaxes of G
- ***MHL*** — **Minimum Hapax length**:

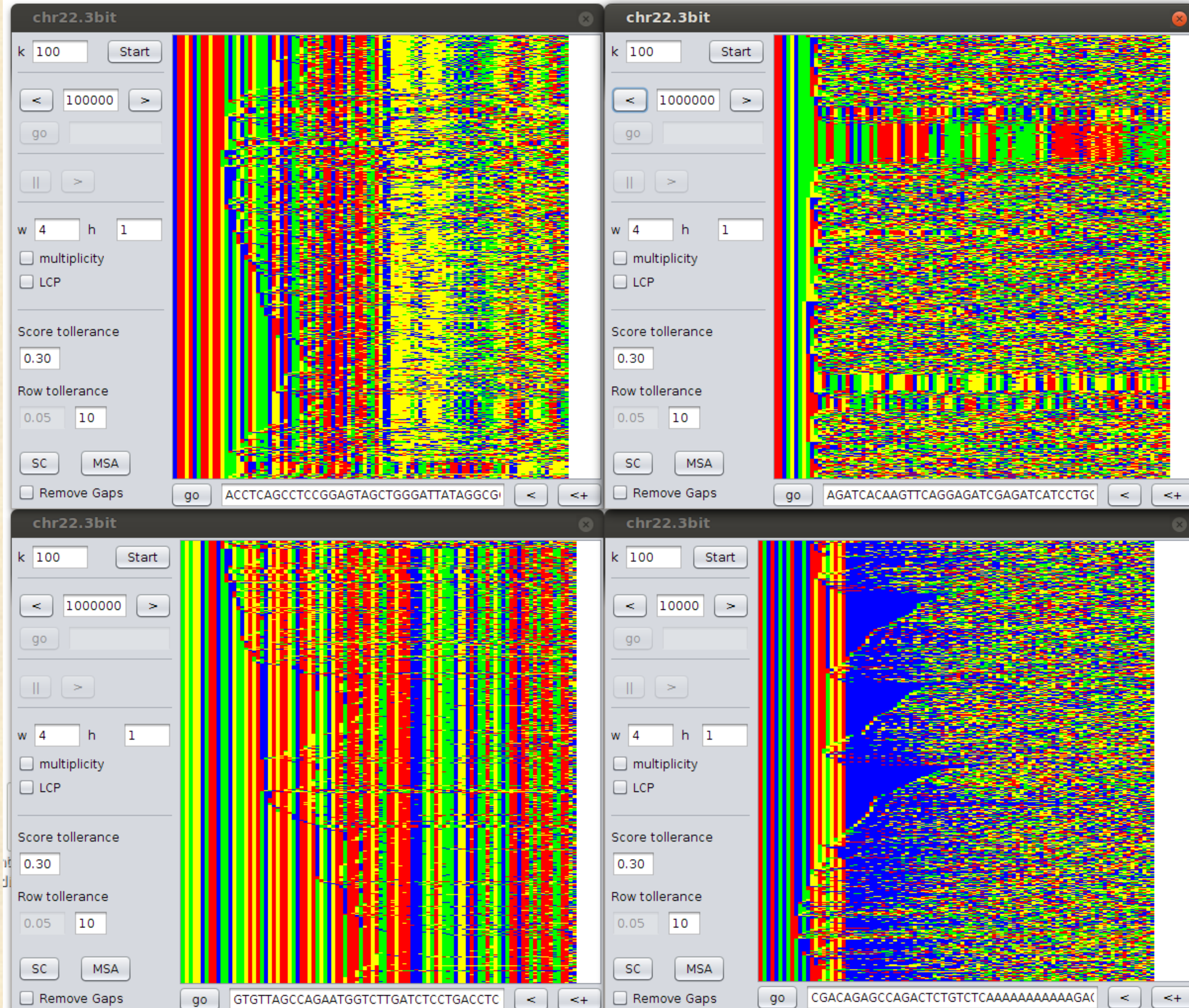
  all the strings of length **MRL-1** are repeat of G
- **COV** — Coverage percentage (w.r.t. a dictionary)
- **PCV** — Positional coverage (w.r.t. a dictionary)
- **E_k(G)** — **E**mpirical k-**E**ntropy →
- **ED_k(G_1, G_2)** — k-**E**ntropic Divergence →
- **Max/min/average length and cardinality of any k-mer class**

Genomic 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 AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA
< <+
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 CTCTCTCTCTCTCTCACACACACACACACACACC
< <+


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 AGATCACAAGTTCAGGAGATCGAGATCATCCTG
< <+
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 CGACAGAGCCAGACTCTGTCTCAAAAAAAAAAAGA
< <+

Bio-bit: a measure of biological information

Bio-bit(G)

provides a comparison between G and $Rand_{|G|}$ by revealing the degree of anti-chaos present in G.

Biobit

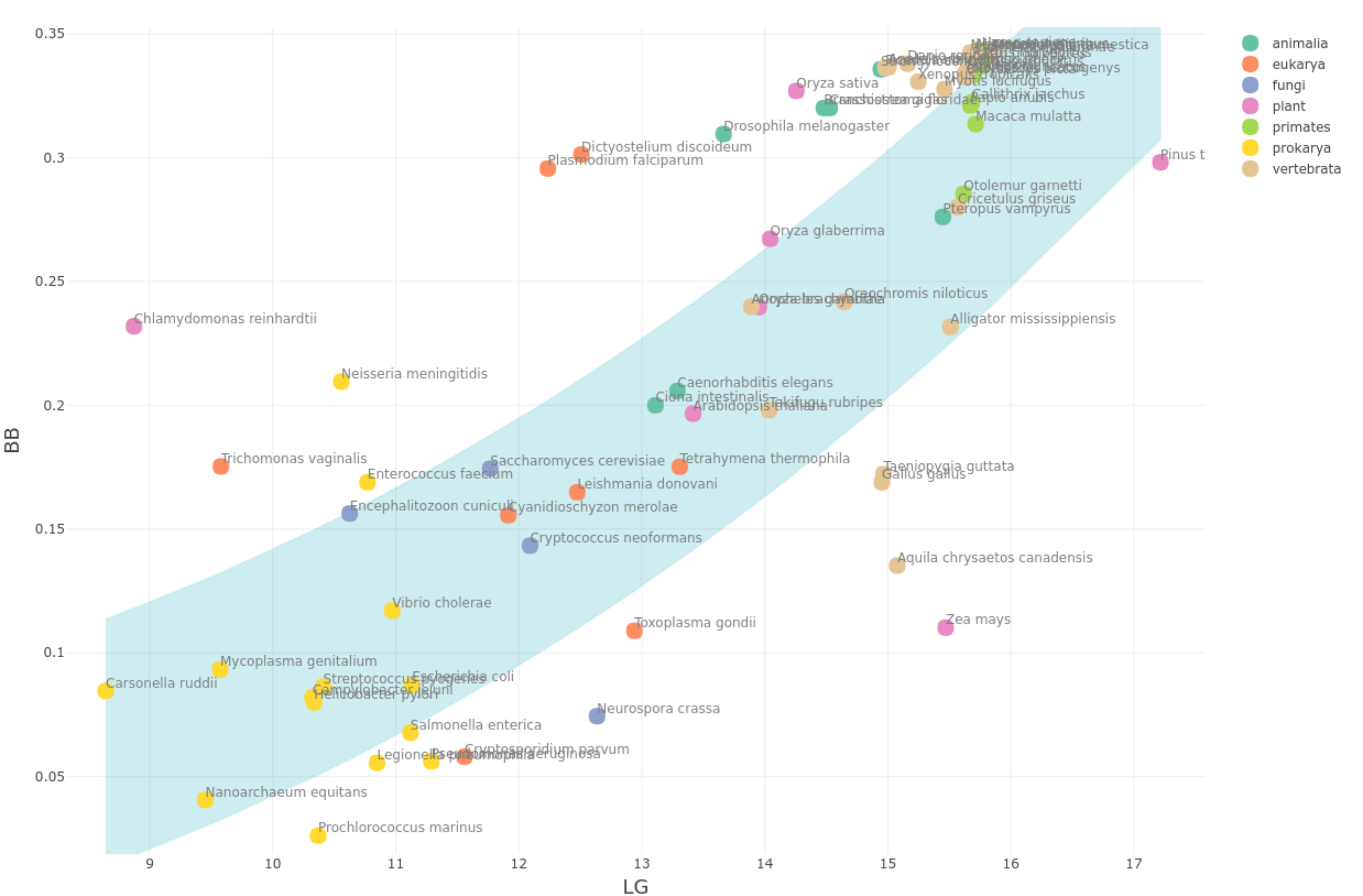
The information that, in the average m-words of G (for suitable m) gain in diverging from random genomes of the same length.

Boltzmann&Shroedinger&Wiener's Neghentropy.

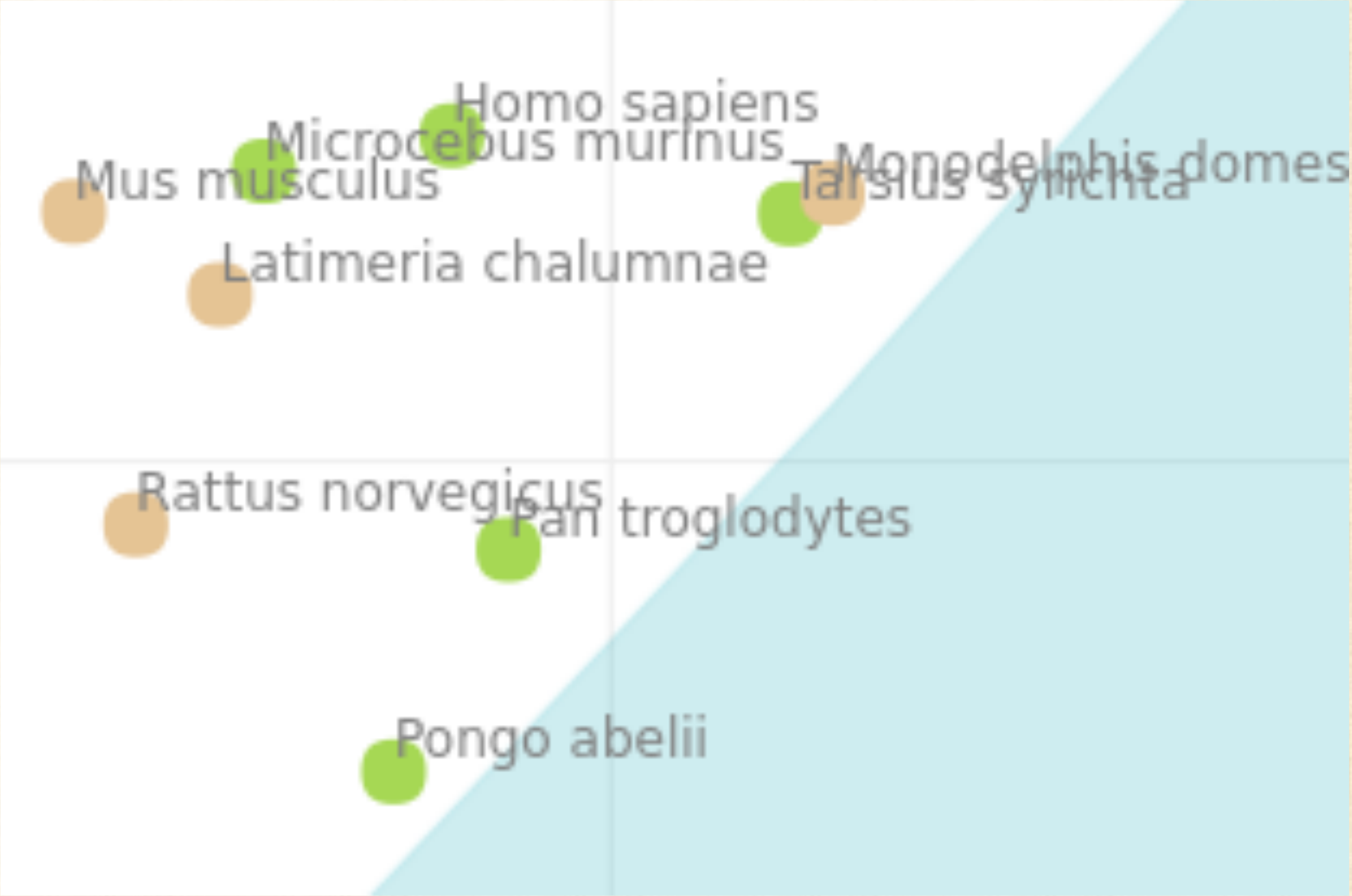
Biobit

The information that, in the average m-words of G (for suitable m) gain in diverging from random genomes of the same length.

Boltzmann&Shroedinger&Wiener's Neghentropy.



animalia
eukarya
fungi
plant
primates
prokarya
vertebrata
BB
LG
0.35
0.3
0.25
0.2
0.15
0.1
0.05
9
10
11
12
13
14
15
16
17
Oryza sativa
Drosophila melanogaster
Dictyostelium discoideum
Plasmodium falciparum
Xenopus tropicalis
Myotis lucifugus
Callithrix jacchus
Papio anubis
Macaca mulatta
Pinus t
Otolemur garnetti
Cricetulus griseus
Pteropus vampyrus
Oryza glaberrima
Oreochromis niloticus
Alligator mississippiensis
Chlamydomonas reinhardtii
Neisseria meningitidis
Caenorhabditis elegans
Ciona intestinalis
Arabidopsis thaliana
Takifugu rubripes
Trichomonas vaginalis
Saccharomyces cerevisiae
Tetrahymena thermophila
Enterococcus faecium
Leishmania donovani
Taeniopygia guttata
Gallus gallus
Encephalitozoon cuniculi
Cyanidioschyzon merolae
Cryptococcus neoformans
Aquila chrysaetos canadensis
Vibrio cholerae
Toxoplasma gondii
Zea mays
Mycoplasma genitalium
Carsonella ruddii
Streptococcus pyogenes
Escherichia coli
Campylobacter jejuni
Helicobacter pylori
Neurospora crassa
Salmonella enterica
Cryptosporidium parvum
Legionella pneumophila
Pseudomonas aeruginosa
Nanoarchaeum equitans
Prochlorococcus marinus



Homo sapiens
Microcebus murinus
Mus musculus
Monodelphis domes
Tarsius syrichta
Latimeria chalumnae
Rattus norvegicus
Pan troglodytes
Pongo abelii

biobit(G)

The formula is not simple to explain,
a mixing of: empirical entropy, logisti map,
 RND, KL divergence, …

biobit is **anti-entropic**,

rather than neghentropic

genome complexity relates to a balance between order and disorder in systems genomes.

Order is related

with functions (for maintaining life),

Disorder with their evolving capacity

References

- Shannon C. The Mathematical Theory of Communication, 1948 (shannon48.pdf)
- Bonnici V, Manca V: Informational Laws of Genome Structure, Scientific Reports (Nature), 2016
- Manca V: The Principles of Informational Genomics, TCS, 2017.
- Manca V: Infobiotics, Springer, 2017.

Open Problems

- Quantum Physics
  - State = vector in Hilbert Space (over Complex field)
  - Measurement = Hermitian Operator
  - Preparation
  - Superposition
  - Entanglement

- Quantum Information
  - Interactive Information Source
  - Mutual information as primitive notion?
  - Informational Reconstruction of Quantum Physics
  - Informational state

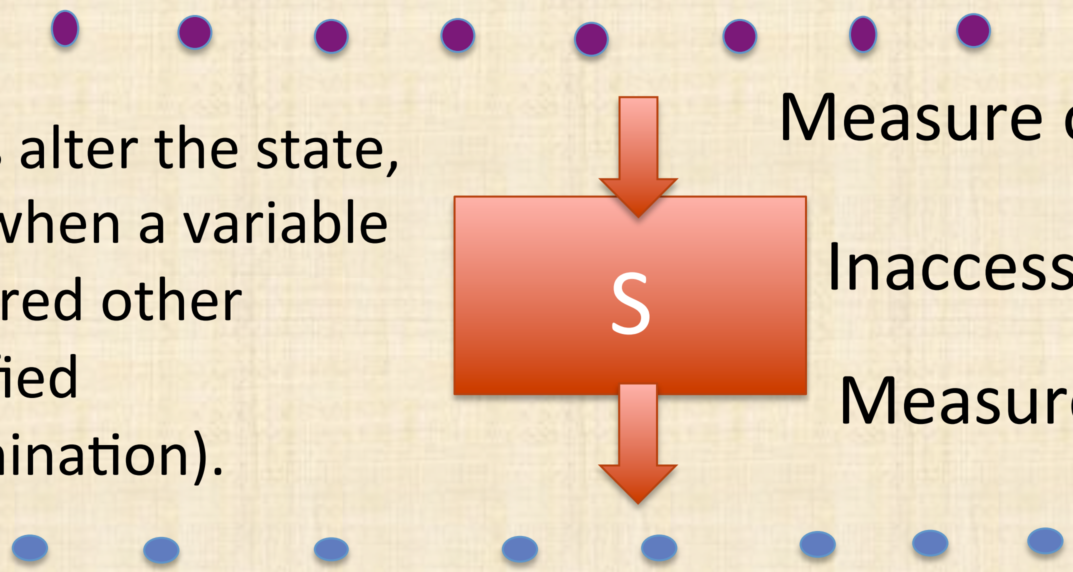
Information Dynamics

- You cannot know as things are when observation changes the dynamics you are observing
- The only information you can get comes from an interaction with a source
- New informational concepts could provide coherent principles for quantum cases

Double Quantum Sources

Input Information Source

Measures alter the state, Because when a variable is discovered other are modified (indetermination).



Measure operations on S

Inaccessible states of S

Measurements of S

Output Information Source

Operations change the probability of Output source. The accessible state of S is only the informational state given by a double intertwined source.

Analogy with living states

In many cases you can know what is inside a cell only by destroying it by missing a part of its complete state.

A general approach to the informational recontruction of inaccessible states could define coherent methodology for describing complex natural systems at different levels.