Introduzione ai metodi informazionali in biologia (Infogenomics)

Vincenzo Manca
Università di Verona

Infolife – CINI

Roma 9 Giugno 2015

Ringraziamenti

- Infolife/Infogenomics (Mondello meeting)
- My group (2009➔coll., post-doc, Phd, ~ 10/15)
  - *Vincenzo Bonnic*i (Post-doc VR)
  - *Giuditta Franco* (Ric. VR)
  - *Alberto Castellini* (Post-doc Berlin (Potsdam), Max Plank)

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**).

Il ***calcolo simbolico*** (*versus* calcolo *numerico/algebrico*) è emerso nel 20° secolo per l'elaborazione formale (e poi automatica) dell'informazione (Logica matematica e Calcolabilità). Due le scoperte fondamentali:

1) L'esistenza di processi matematicamente definibili, ma non calcolabili. LIMITI DEL CALCOLO

2) L'esistenza di macchine di calcolo universali, ovvero capaci di realizzare qualsiasi processo di calcolo.

POTENZA DEL CALCOLO

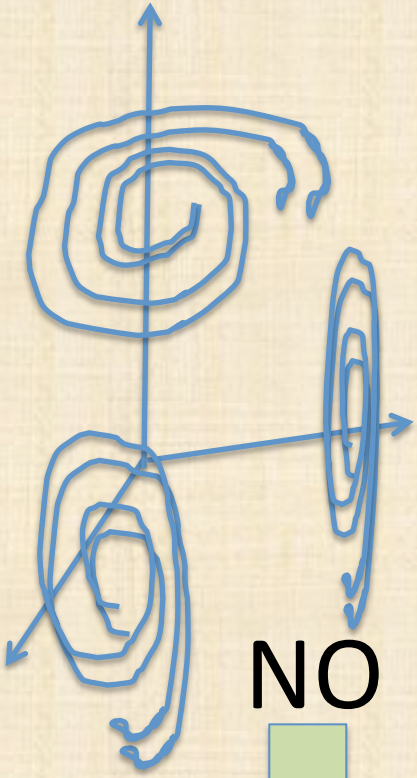
Riproduzione e Universalità

- 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 biologica** postula meccanismi di duplicazione (ds DNA).

DNA “more geometrico demonstratus”

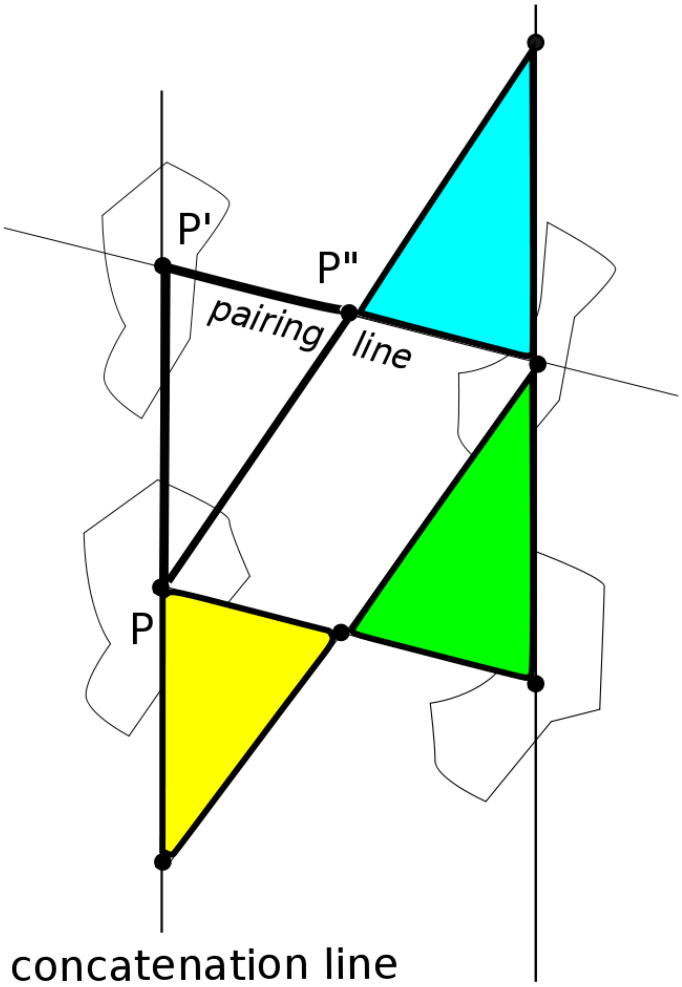
- La duplicazione template-driven è realizzabile in tempo lineare.
- Una struttura bilineare rende la duplicazione template-driven più affidabile e veloce.

Un polimero bilineare è astrattamente una sequenza di monomeri triangolari

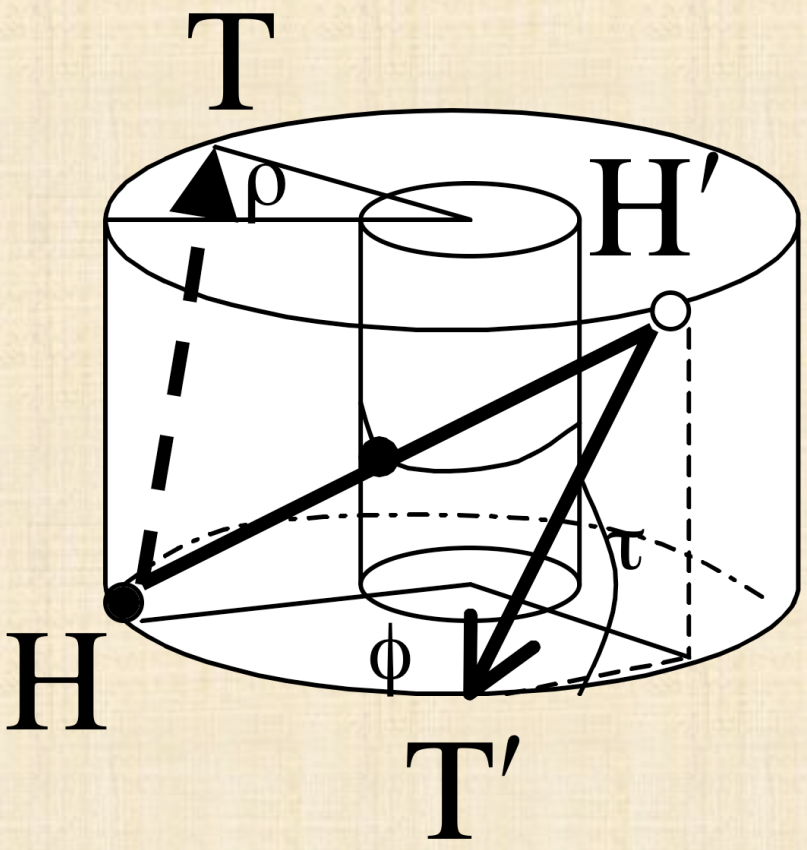


NO

Si avrebbero angoli concatenativi diversi a meno di non usare spirali logaritmiche

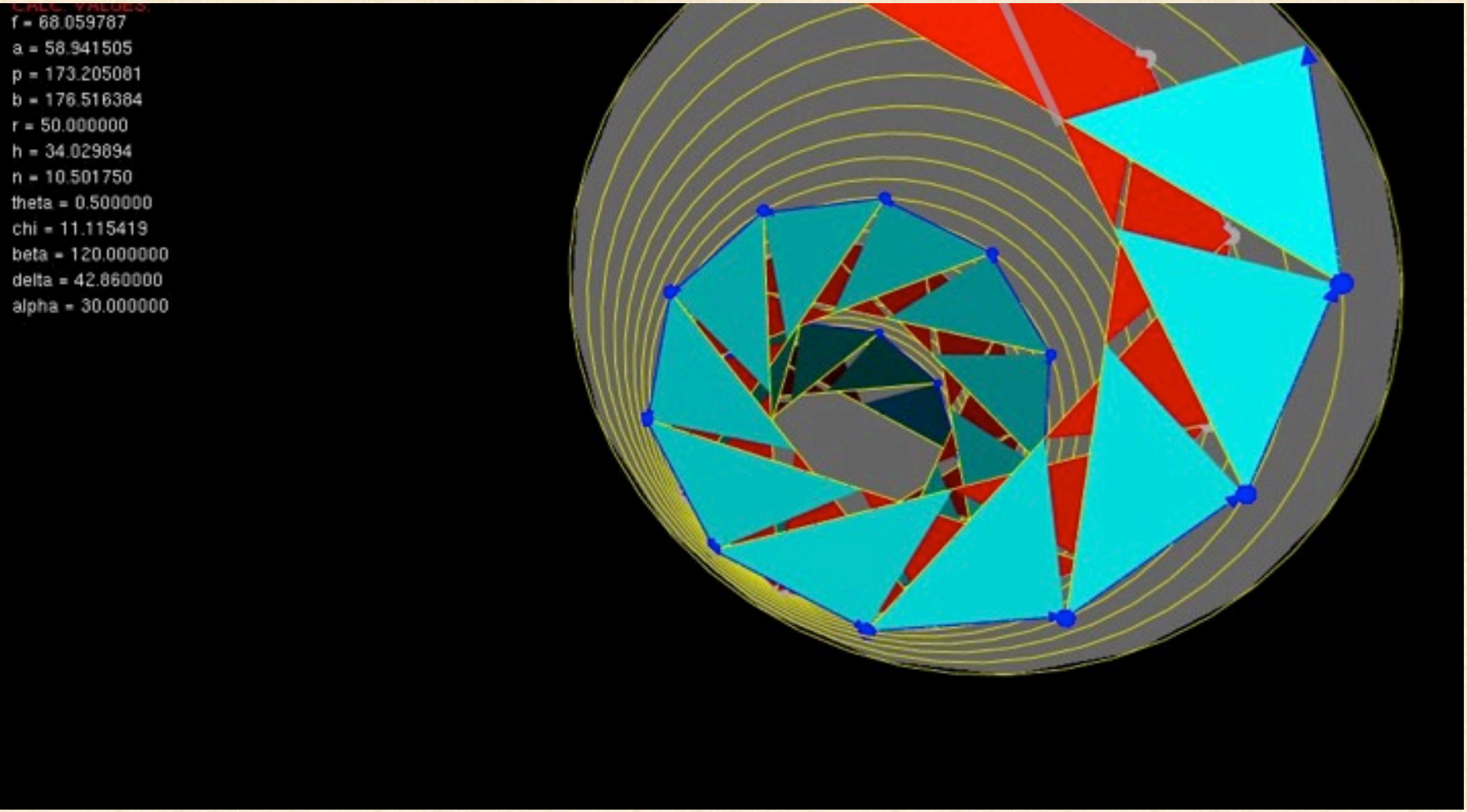


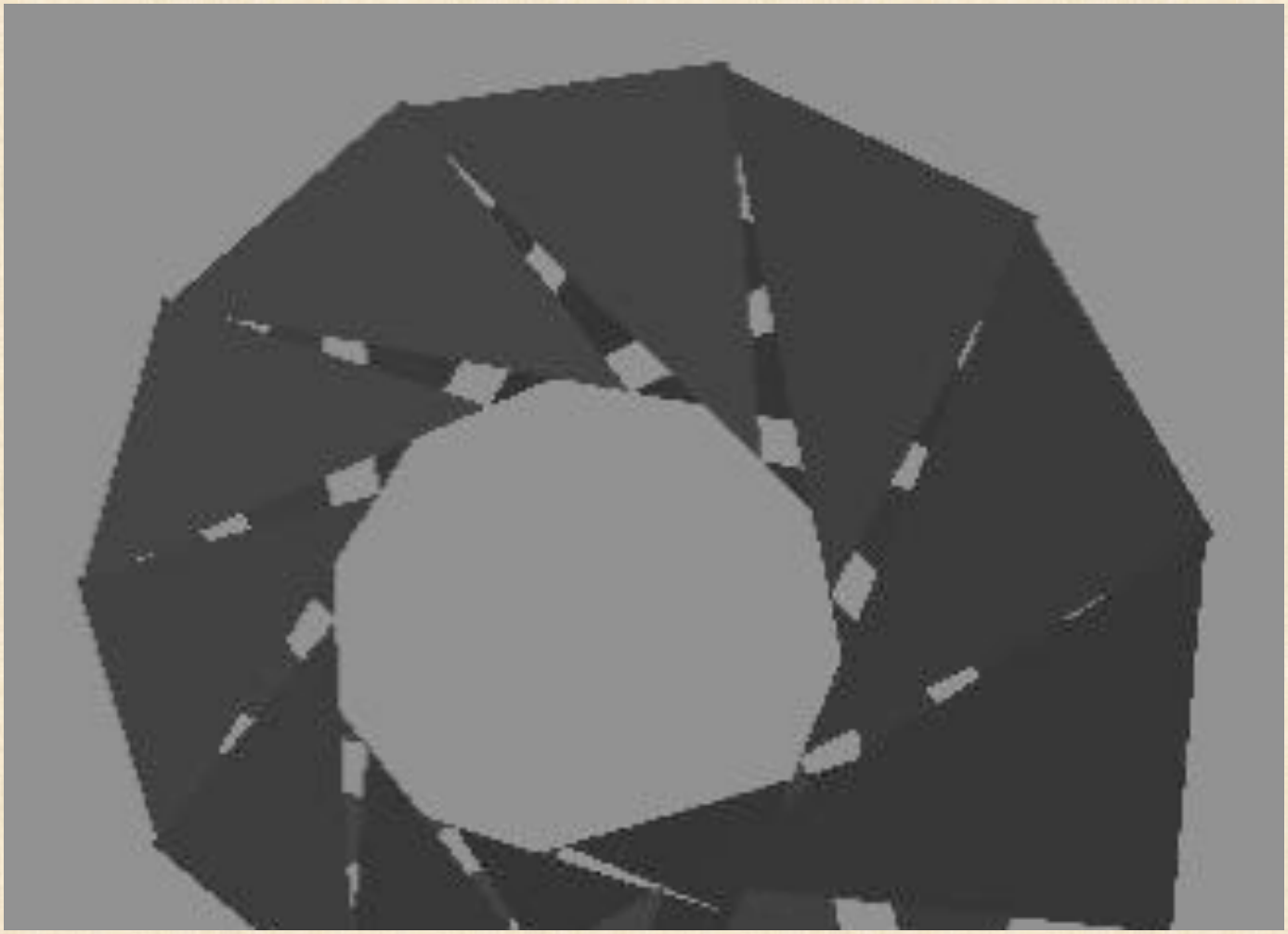
Non può rimanere su un piano



Il parallelismo sarebbe possibile solo avendo molecole di **chiralità diversa**, ma questo complicherebbe notevolmente il processo di copia.

Chiocciola di triangoli monomerici

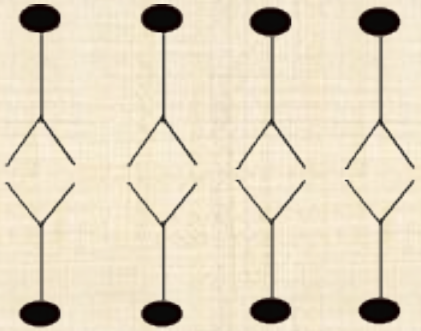


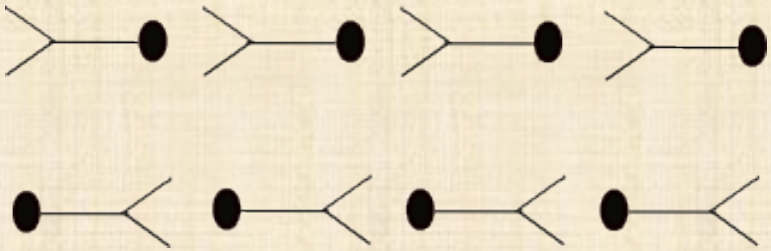


Phospho-lipidic and Phospho-diesteric

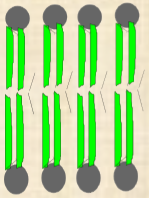
Two forms of Bilinearity (based on chirality)

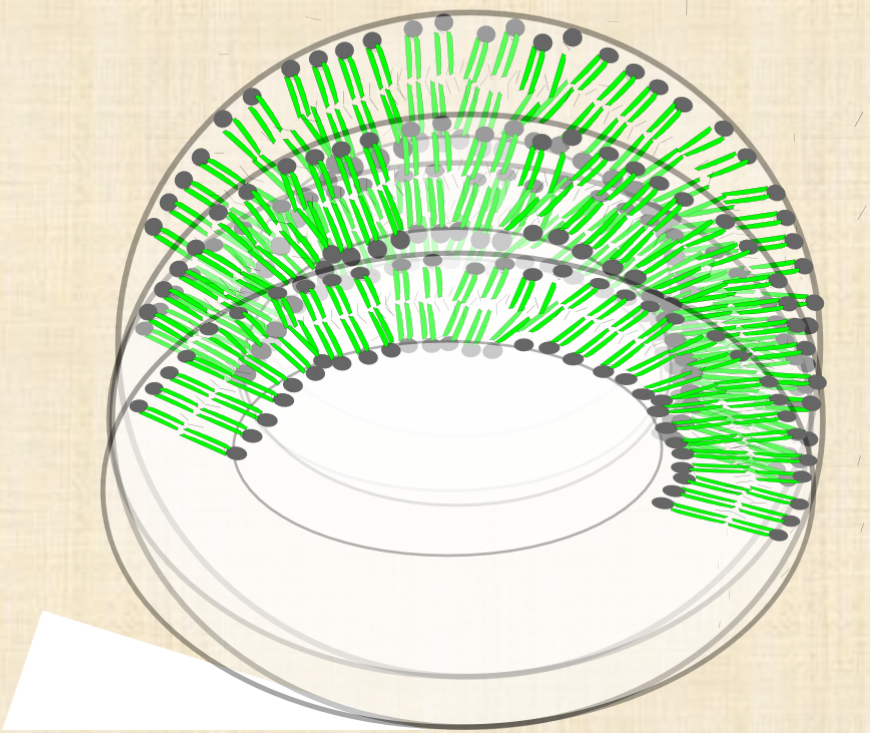
lines of pairs and pairs of lines

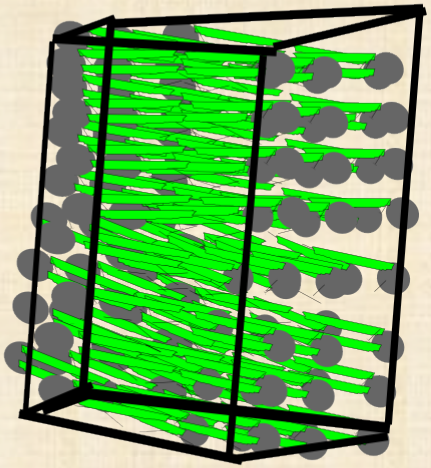




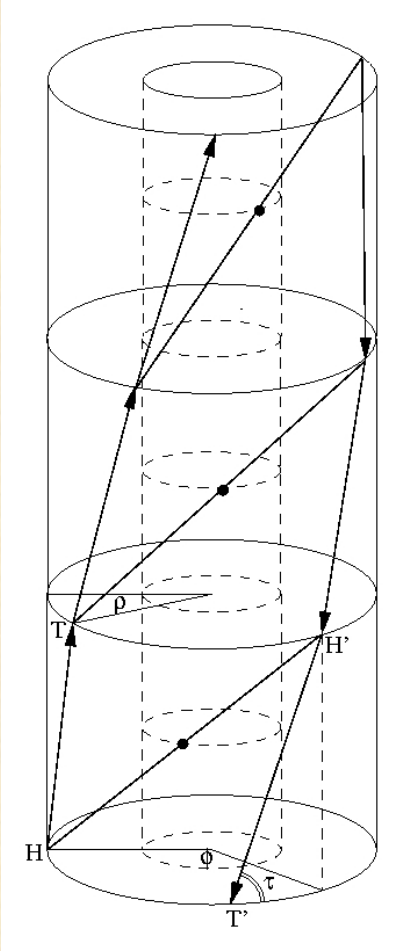
La sfera







Il cilindro



Manca – The logic of bilinear forms, Fundamenta Informaticae, 2005

****

Probabilità e informazione

- Shannon 1948: l'informazione di un evento è funzione della sua probabilità.
- La probabilità è distribuzione (spazio di eventi).
- L'informazione si determina analizzando (opportune) distribuzioni di grandezze in una popolazione di fenomeni.

Probabilità

cruciale in tutte le scienze a partire dal 20° secolo

- **Cardano e Galileo** : De ludo aleae
- **Pascal e Fermat** : Chevalier de Merè
- **Jacob Bernoulli** : urna e processo bernoulliano (Ars Conjectandi)
- **De Moivre, Laplace e Bayes** : gaussiana e inversione del condizionamento
- **Gauss** : leggi del caso
- **Francesi, Russi, Italiani** (Poisson, Cauchy, Borel, Chebicev, Kolmogorov, Cantelli, De Finetti): distribuzioni, leggi dei grandi numeri, misura
- **Boltzmann** (fisica statistica)
- **Inglesi** (Galton, Pearson, Student, Fisher): leggi della statistica

Insidie probabilistiche

- Se un pilota ha il 2% di essere abbattuto in ciascuna missione, qual è **la probabilità di abbattimento** entro 50 missioni?

- Totale perché 2% x 50 = 100% **ERRATO !**

- Stesso errore del gioco sottoposto a Pascal da Chevalier de Meré

- E' abbattuto alla n-esima sse sopravvive a tutte le precedenti e viene abbattuto durate quella. Sommando questi casi esclusivi da 1 a 50 si ha, con p = 0,02

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

$(1-p)^{50}$ è la probabilità di sopravvivere alla 50° missione

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)

Distribuzione - Informazione

- X variabile che assume valori con molteplicità:

  $x_1 , x_2 , x_3 , \ldots\ldots \quad n_1 , n_2 , n_3 , \ldots\ldots$
- Se $n = n_1 + n_2 + n_3 , \ldots\ldots$
- $= n_1/n , \ n_2/n , \ n_3/n , \ldots\ldots$ sono frequenze
- $p_1 , p_2 , p_3 , \ldots\ldots .$ sono probabilità (misure di possibilità di occorrere)
- Shannon chiama (X, p) **Sorgente di Informazione**
- **$-\lg p_e$** è la misura di informazione dell' evento **e** di probabilità **p_e**

Information Paradoxes

Choice, Uncertainty, Information ???

Section 6 of Shannon's booklet

(compare to: Learning/Ignorance/Knowledge)

It 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

Entropia Termodinamica

Teorema di Carnot

Una macchina termica che lavora tra due sorgenti Termiche una M (serbatoio) a temperatura T e una M_o (condensatore) a temperatura T_o con $T > T_o$ prelevando calore da M e restituendolo ad M_o, non può restituire meno calore di $Sx(T/T_o)$, ovvero:

S = Q/T detta **entropia** è il **minimo calore che una macchina (termica) può rilasciare** ad un condensatore a temperatura **T_o quando T_o è assunta come misura di unità termica.**

(dimostrazione: via macchine reversibili, teoria degli automi).

Limite all'efficienza delle macchine termiche

L'irreversibilità temporale come conseguenza probabilistica della complessità

Boltzmann: L'entropia termodinamica di Carnot **S** è proporzionale al **logaritmo** del numero **W** di microstati associati al macrostato termodinamico del sistema.

Sia n il numero di particelle e k le classi delle velocità delle particelle di gas:

$$n = n_1 + n_2 + \ldots\ldots n_k$$

$S = k \lg W$

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

n_i = numero di particelle con velocità compresa nell'intervallo i-esimo

- Per la formula di Stirling $\lg n! \approx n \lg n$

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

- $S = A - k(n_1 \lg n_1 + n_2 \lg n_2 \ldots + n_k \lg n_k)$

Il teorema impossibile

H di Boltzmann

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

H è la versione microscopica dell'entropia termodinamica, cambiata di segno (a meno di costanti additive e moltiplicative).

Teorema H (1872) In un sistema isolato:

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

Da Boltzmann a Shannon

$$H_s = - \Sigma_i \, p_i \, \lg p_i$$

Shannon 1948

Entropy Th. H_S è univocamente individuata dalle 3 condizioni:

Continuità in p_i , **Massimo** in 1/n x n , **Additività** delle scelte:

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

- H e H_s sono la stessa cosa a meno di costanti additive e moltiplicative (von Neumann: "*avrai successo. Pochi sanno veramente cosa sia*"). Entropos (verso interno)
- Da $inf_i = - \lg p_i$ segue che:
- H_s è l' informazione media della sorgente informativa S = (X, P)

Pythagorean Recombination Game confirms Boltzmann's claim

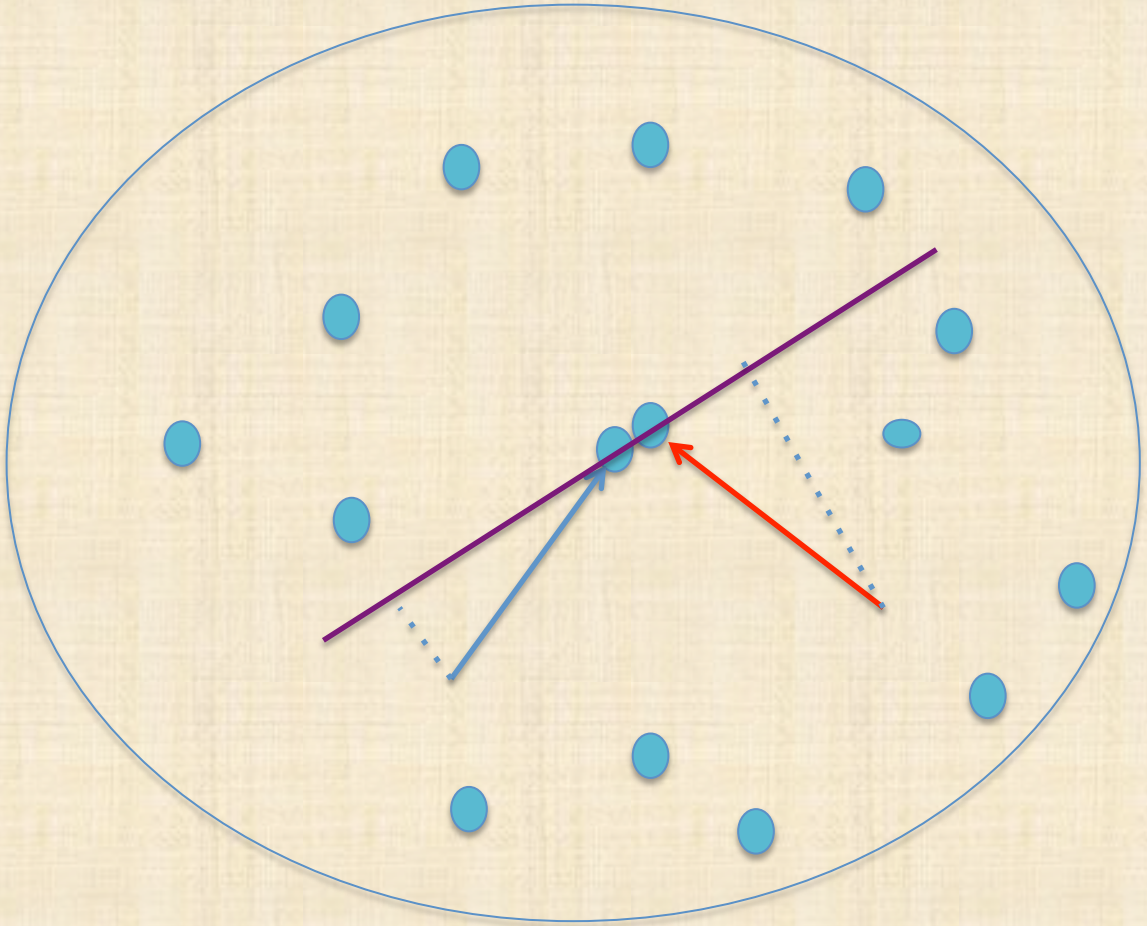
Start with a population P of **random** numbers

- For N steps **do**

- Choose **randomly** a, b in P
- Chose **randomly** a number $a_1 \leq a$ and split a into a_1 and $a_2 = \mathrm{sqr}(a^2 - a_1^2)$, that is, $a = \mathrm{sqr}(a_1^2 + a_2^2)$
- Chose **randomly** a number $b_1 \leq b$ and split a into b_1 and $b_2 = \mathrm{sqr}(b^2 - b_1^2)$, that is, $b = \mathrm{sqr}(b_1^2 + b_2^2)$
- Replace in P numbers a, b with:

$$a' = \mathrm{sqr}(a_1^2 - b_2^2) \quad , \quad b' = \mathrm{sqr}(b_1^2 - a_2^2).$$

2D Gas



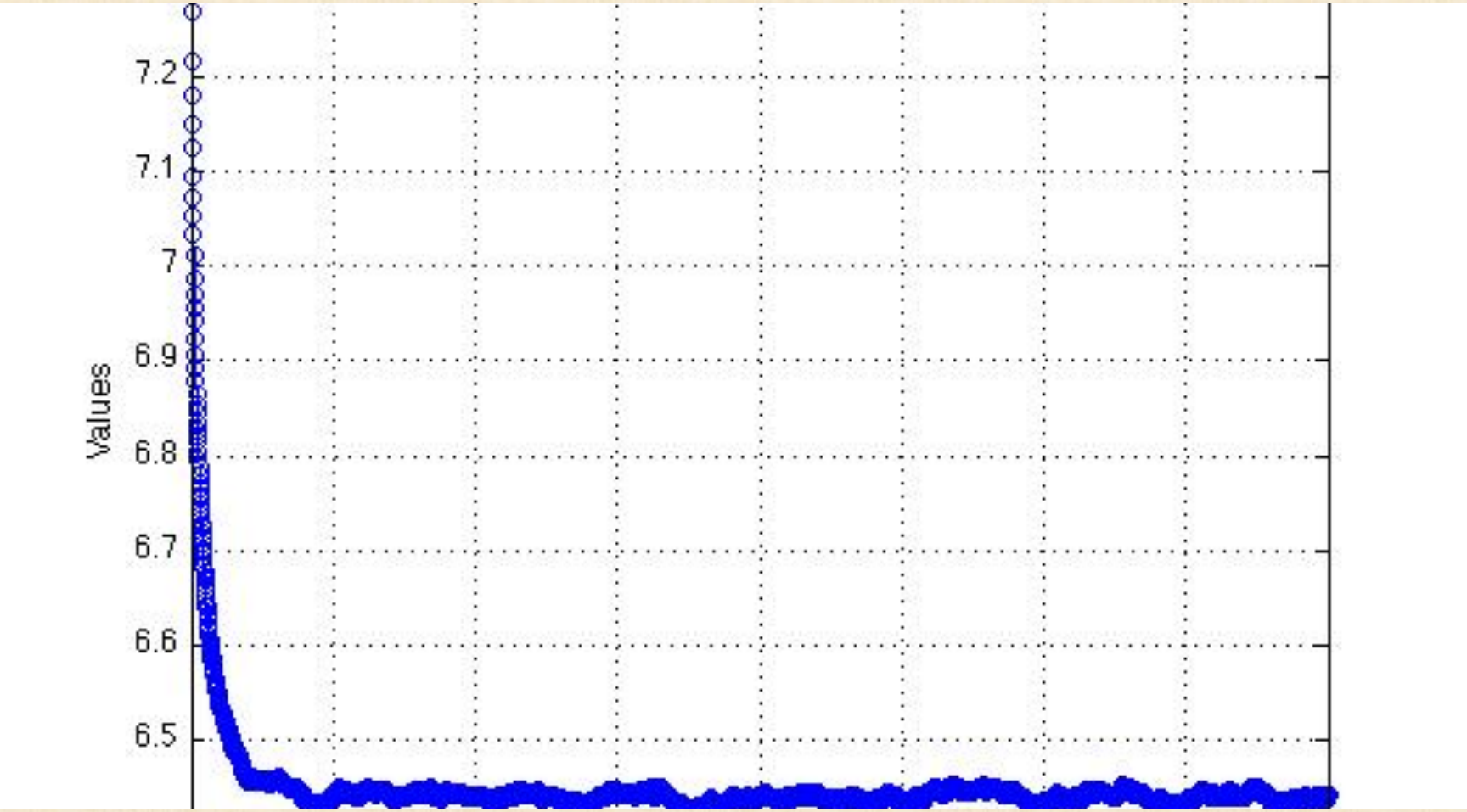
NOTA BENE: l'elasticità degli urti equivale alla conservazione della varianza nella distribuzione delle velocità

A Population of 1000 random numbers

H after 4000 steps 200 collisions/step

Chi = sqrt of a sum of squares normally distributed

H after 4000 steps 200 collisions/step



Cosa è avvenuto?

- Il sistema si “complessifica”: le sue parti diventano interdipendenti;
- le cause si “normalizzano” le distribuzioni di velocità seguono le leggi casuali dei grandi numeri;
- aumenta l’informazione media perché la velocità di ogni particella dipende da tutte quelle che ha incontrato nella storia delle sue collisioni;
- aumenta l’entropia informativa.

V. Manca – Infobiotics: information in biotic systems, Springer, 2013

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.

****

Deterministic Chaos

Algorithmic Generation of randomness

It is the basis of Pseudo-Random Numbers

(on which random genomes are based)

Nothing is more difficult than reproduce a real random process. BUT

π digits, Bernoulli Shifts, Lehmer generators, Logistic Maps, … are algorithmic ways to generate processes that appear as truly random processes (~ 1950 → …)

Poisson, Geometric, Exponential

Il caso è sempre un **urna** (con una dea bendata)

p frazione di palline bianche (successo) e 1-p di nere (insuccesso):

$$\binom{n}{k} p^k (1-p)^{(n-k)}$$

è la probabilità di avere k successi in n estrazioni. Se la media in n estrazioni è λ :

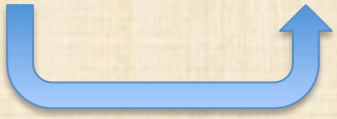
$$\binom{n}{k} (\lambda/n)^k (1-\lambda/n)^{(n-k)}$$

Feller – An introduction to probability theory and its applications, Wiley 1957

per n molto grande (p molto piccola) è
la legge degli eventi (casuali) rari:

$$\binom{n}{k} (\lambda^k/n^k)\, e^{-\lambda}$$

$$n(n-1)(n-2)\ \ldots/k!\ (\lambda^k/n^k)\, e^{-\lambda}$$



$$n(n-1)(n-2)\ \ldots/n^k\ (\lambda^k/k!)\, e^{-\lambda} \approx \mathbf{e^{-\lambda}\lambda^k/k!}$$

success waiting time probability ≈ $\boldsymbol{\lambda\, e^{-\lambda k}}$

$$(1-p)^k\, p = p\,(1-p)^{\mathbf{pk/p}} \approx p e^{-pk}$$

****

Symbolic Sequences (over an alphabet A)

- Alphabet = finite set of simbols (physical objects)
- Sequences, subsequences, strings
- $\alpha, \beta, \gamma, \delta, \eta, \ldots$ **λ** $\alpha(i), \alpha[i], \alpha[i, j], |\alpha|$
- $\alpha\beta$ concatenation (monoid)

  $(\alpha\beta)\gamma=\alpha(\beta\gamma)$, $\alpha\lambda=\lambda\alpha=\alpha$

  $\alpha_{[k]}\beta$ k-overlap concatenation

  $\alpha_{[]}\beta$ maximal-overlap concatenation

A formal Language **L** on A is a set of strings on A

$$L \subseteq A^*$$

- Operations : + , /, · , * , ∩ [ ] $A^* \leftrightarrow \mathbf{N}$

Patterns and Transformations

- G = (A, T, S, R) Chomsky Grammar

Alphabet, Terminals, Start symbol, Rules

- aB → aaB
- Ba→co

- aBaco → aaBaco → aaaBaco → aaacoco

Universal Rewriting Schemata

- Replacement (Chomsky 1957)
- 2-Replacement (Kuroda 1964)
- Prefix-Suffix rotation-replacement (Post 1945)
- Concatenation, Split, Prefix/Suffix Deletion
- Copy-dislocate, Mutation (Transposons) ***

A symbolic example of emergence

The Tri-somatic Grammar

Alphabet = {a,b,c, S,B}

Terminals = {a,b,c}

Start = S

Rules =

- S → aSBc
- S → bc
- cB → Bc
- Bb → bb

$L(G) = \{a^n b^n c^n \mid n > 0\}$

Salomaa – Formal Language theory, Wiley, 1973

****

Infogenomics
An Informational Approach analogous to ENCODE
Representations of long symbolic sequences

- Positions → Simbols
- symbol → set of positions where it occurs

  (symbol spectra)
- Elongation Sequences
- Elongation trees
- (01)-walks (CGR - Chaos Game Representation)
- Auto-similarity distances
- k-RDD

Dizionari genomici

- $D(G) = \{G[i,j] \mid 1 \leq i \leq j \leq |G|\}$ (dim. quadratica w.r.t. |G|)
- $D_k(G) = D(G) \cap \Gamma^k$
- Un dizionario incluso in D(G) dicesi dizionario di G
- $D^{[]}$ è la chiusura di D per maximal overlap concatenation,
  $over(\alpha, D) = \{\beta \text{ in } D \mid \alpha[]\beta \neq \lambda\}$, $|over(\alpha, D)| \leq |\alpha|$, $D[]D' = \ldots$
- Una posizione p di G è m-coperta in D se vi sono m parole di D del tipo G[i,j] con $i \leq p \leq j$
- D copre G se ogni posizione di G è 1-coperta da D (se $D^{[]}$ include D(G))
- D copre minimalmente G se D copre G e non include nessun D' che copre G
- G è (esattamente) D-fattorizzabile se G appartiene a D*

Sequenziamento Genomico

- Dato un dizionario D incluso in D(G) è possibile determinare univocamente G a partire da D?
- In genere no. Ma sotto ipotesi opportune sul coverage e sulla distanza relativa di coppie di parole è possibile farlo con probabilità molto alta.
- Che caratteristiche deve avere un dizionario di D perché G possa essere univocamante determinato da esso?

Basic Genomic Indexes

- ***Ln*** Length
- ***kls*** k-Lexical selectivity $|D_k(G)|/4^k$
- ***mfl*** Maximal Forbidden Length
- ***3-mult*** Codon multiplicity/frequency (k-mer with k ≤ *mfl*)
- ***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)
- **alh , alr** Almost Repeat/hapax Length (95% of length L are repeat/hapax)
- **cov, pcov** The percentage of G covered by D, and the average positional coverage
- **$E_k(G)$, $EE_k(G)$** **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**

Distribuzioni genomiche

- Molteplicità (rispetto a un dizionario D)
- Co-Molteplicità (rispetto a D)
- k-occorrenza (rispetto a D e unit-intervals)
- RDD (di una parola) ***
- Seq-Coverage (di una parola/dizionario rispetto a G)
- Pos-Coverage (di una posizione di G rispetto a D)
- Length-repeats

Minimal k-RDD

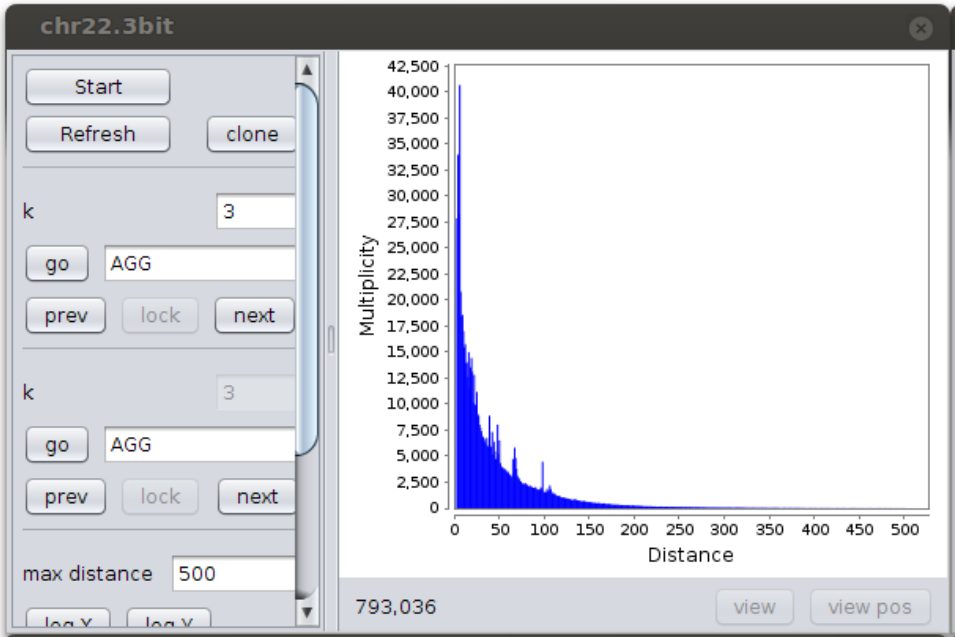
Let α ε D(G) such that:

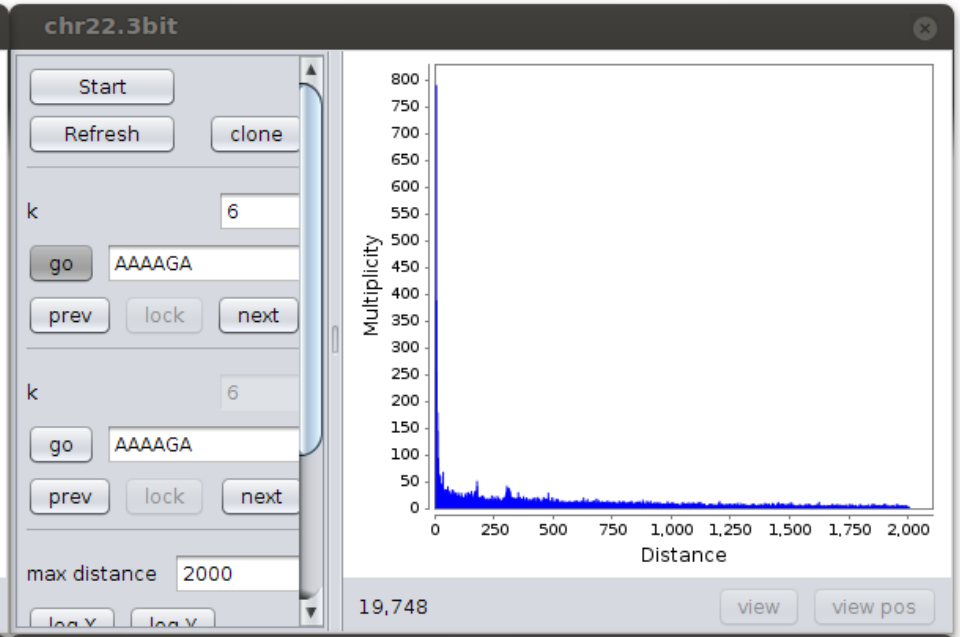
---α---d_1---α---d_2---α---d_1---α---d_3---α---

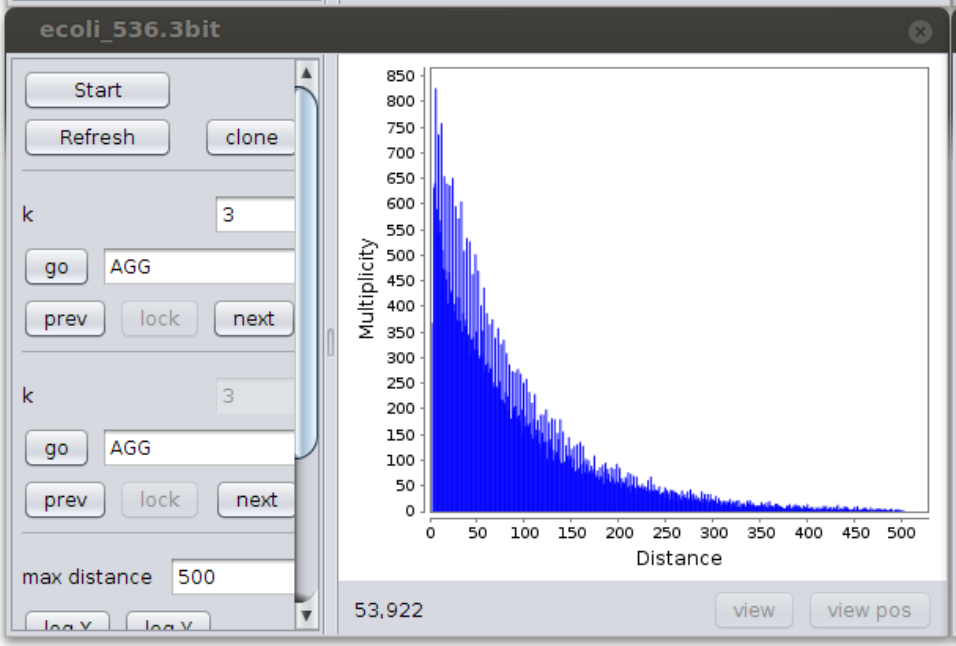
where between two consecutive α no α occurs

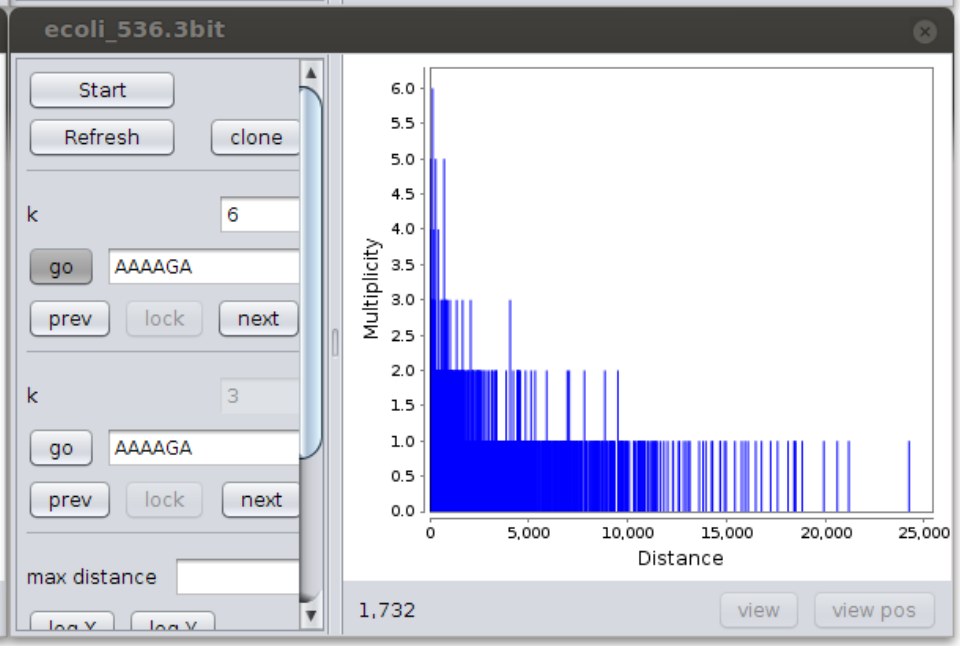
K-RDD(α, G) = d_1 -> n_1 , d_2 -> n_2 , d_3 -> n_3 , - - -

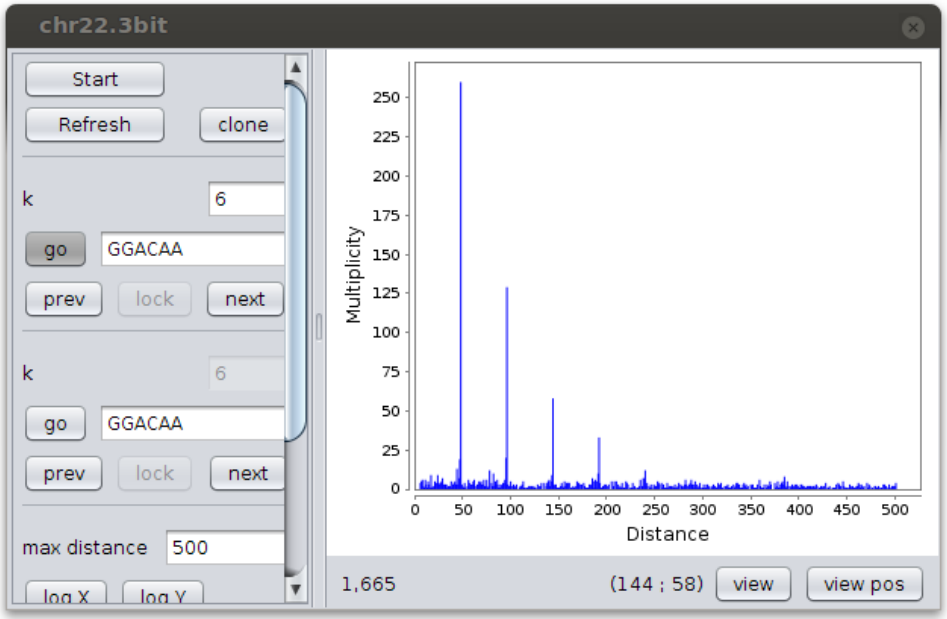
If G is random, for α "not too short, but long enough",
K-RDD(α, G) is geometric/exponential, according to probability theory.

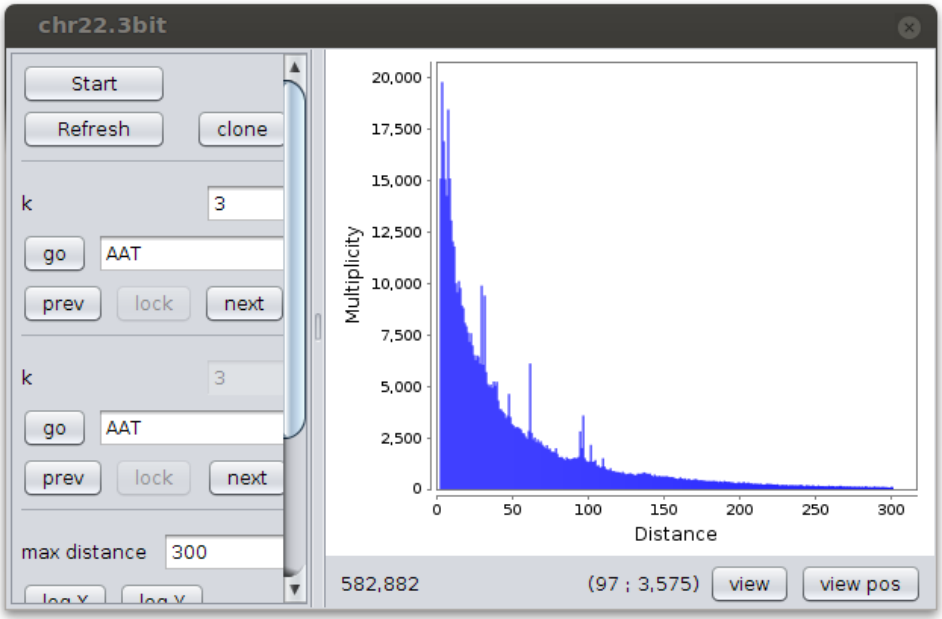

chr22.3bit
Start
Refresh
clone
k
3
go
AGG
prev
lock
next
k
3
go
AGG
prev
lock
next
max distance
500
Multiplicity
Distance
793,036
view
view pos

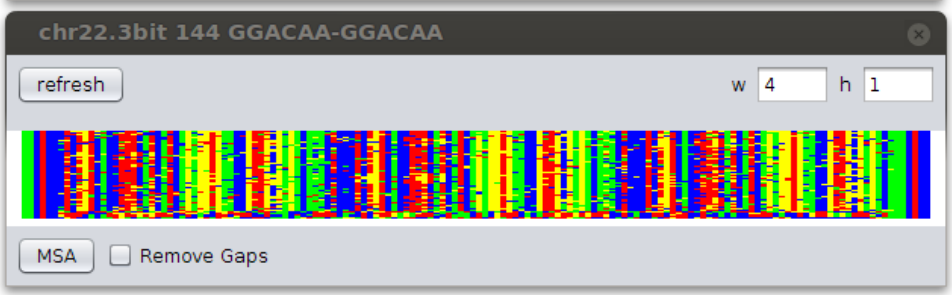

chr22.3bit
Start
Refresh
clone
k
6
go
AAAAGA
prev
lock
next
k
6
go
AAAAGA
prev
lock
next
max distance
2000
Multiplicity
Distance
19,748
view
view pos

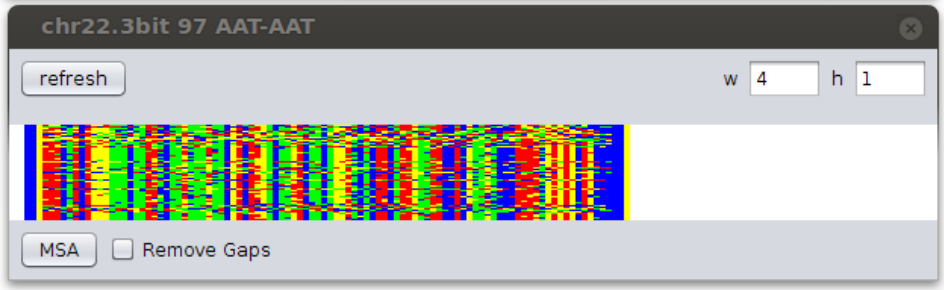

ecoli_536.3bit
Start
Refresh
clone
k
3
go
AGG
prev
lock
next
k
3
go
AGG
prev
lock
next
max distance
500
Multiplicity
Distance
53,922
view
view pos


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

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

****

Recurrence Distance Distribution

- Peak 3-periodicty, for k=1, 2, … in coding regions
- Double exponential distributions
- ncRNA peak 3-periodicity
- Average RDD for k >3
- C_3 coefficients of coding propensity
- Anti 3-periodicity
- Extra peaks and repeatitiveness

Bonnici's IG-Tools investigation about new periodicity phenomena revealing sequence functions

Codici

funzione : C → D surgettiva (copre D)

C stringhe su un alfabeto (codifiche)
D insieme di dati

ad ogni **codifica** corrisponde uno ed **un solo dato**
(una codifica non può corrispondere a due dati distinti)

due codifiche distinte possono codificare uno stesso dato (come nel codice genetico). Il codice è *ridondante* se ciò vale *(non ridondante* altrimenti)

Tipi basilari di codici

- Codici univoci: ogni stringa è fattorizzabile con codifiche in al più un unico modo
- Codici istantanei: nessuna codifica è prefisso di un'altra codifica
- Codici autodelimitanti: la codifica specifica la propria lunghezza
- Codici a lunghezza fissa
- Norma di Kraft nel caso binario $|C| = \Sigma_{x \varepsilon C} 2^{-|x|}$
- **Th.** McMillan : C univoco sse $|C| \leq 1$
- **Th.** C univoco ➔ $\exists$ C' istantaneo t. c. $|C| = |C'|$

Entropic Divergence

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

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

Many other divergences were elaborated, based on different approaches.

Mutual information

$$I(P, Q) = D((P,Q) , P \cdot Q)$$

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

Noise alters data along the channel
What is the information amount that can pass correctly?

Mutual Information Th $I(X, Y) = H(X) - H(X \mid Y)$

Shannon's 2° Th. Provides conditions to transmit with error probability going to zero (autocorrecting codes).

Mutual Information in genomes

$$I_{\alpha,\beta}(G) = \sum_d p_{\alpha,\beta}(G,d) \log [p_{\alpha,\beta}(G,d)/p_\alpha(G)p_\beta(G)]$$

$$I_d(G) = \sum_{\alpha,\beta} p_{\alpha,\beta}(G,d) \log [p_{\alpha,\beta}(G,d)/p_\alpha(G)p_\beta(G)]$$

[Li, 1990, Herzel-Grosse 1995]

****

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, V. Manca - Bioinformatics and Proteomics Review

A New Word Selection Algorithm based on EXP/KL

INGREDIENTS

- Word Recurrence distance distribution RDD(α)
- Waiting time exponential law (in random genomes)
- “Entropic distances” between distributions
- Words extraction by elongation-monotonicity
- Maximal elongations
- Word filtering by different tests

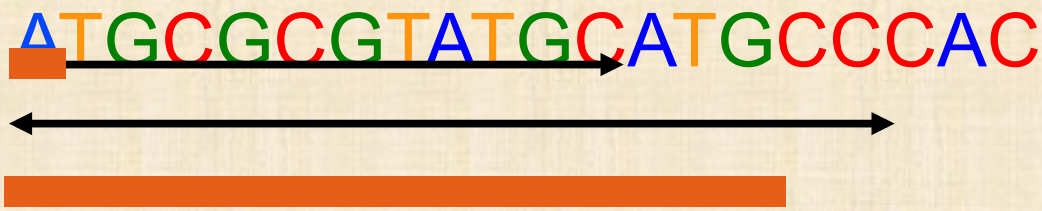
Elongation cases (by V. Bonnici)

Seed extension ➔ Elongation identity

seed

Seed inclusion ➔ Elongation inclusion

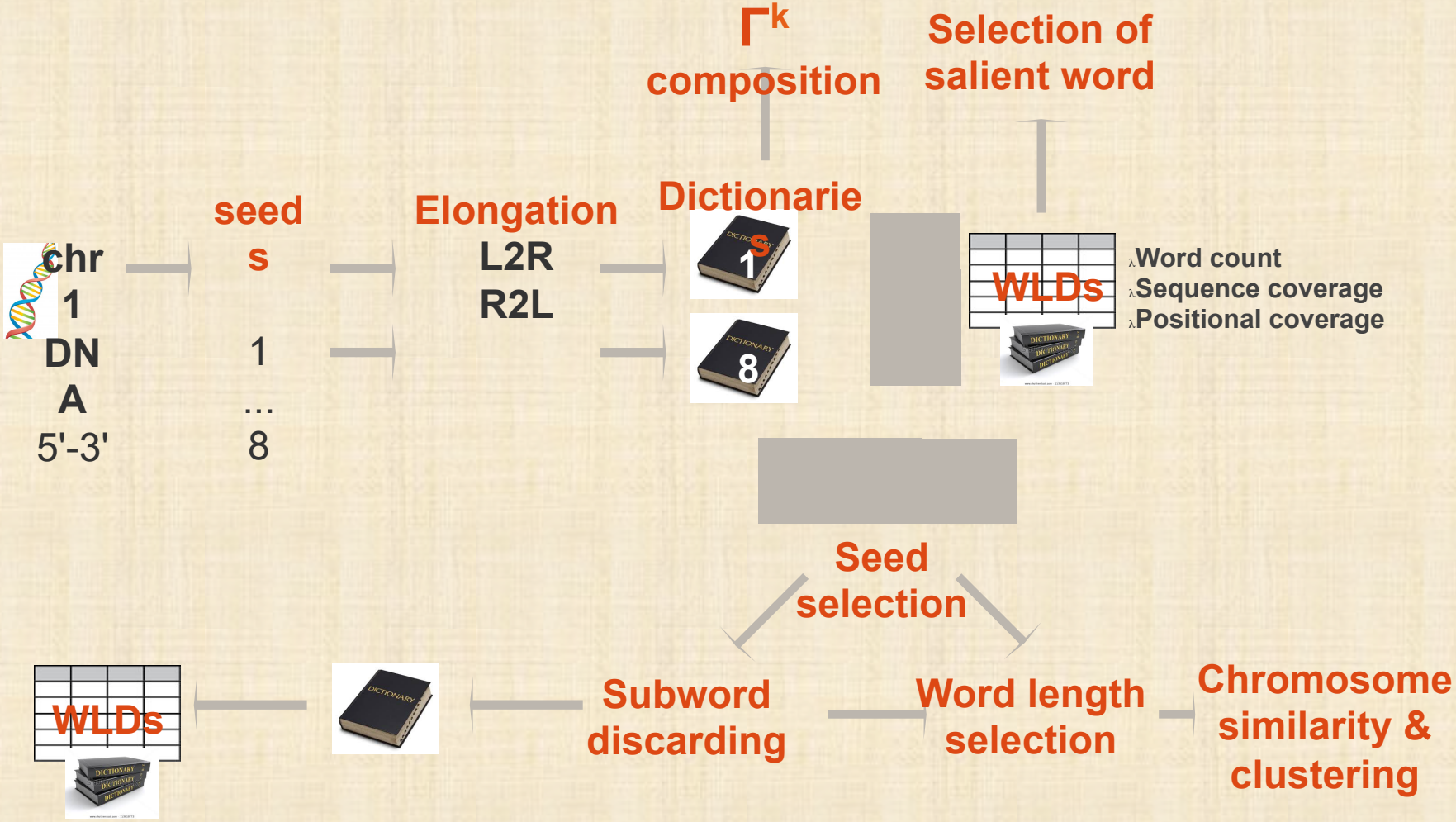
Seed extension ➔ Elongation extension

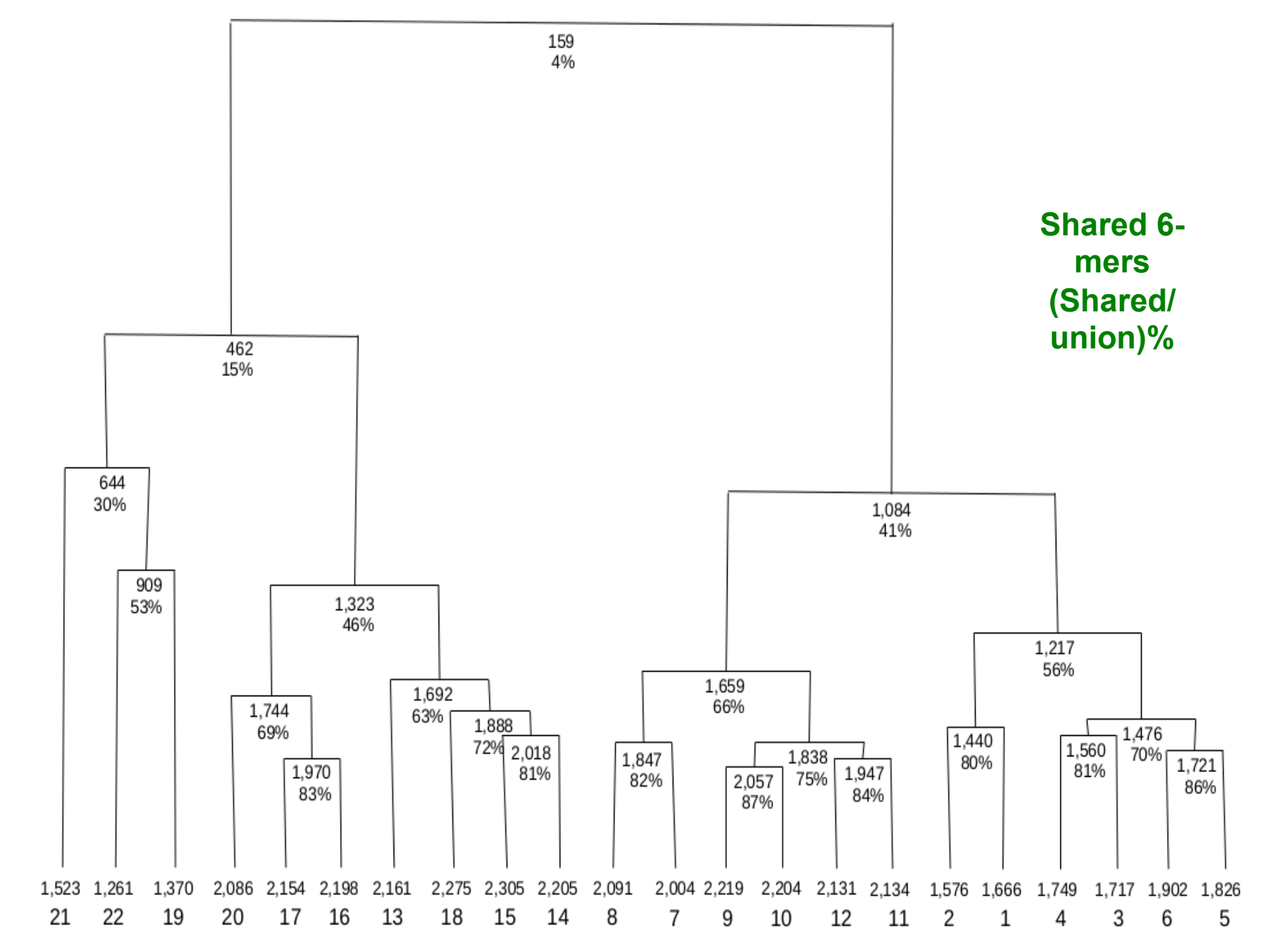


The power of hexamers

	5	5	5
6	1,666	**0.8426**	**2.7715**
7	**2,310**	0.7691	2.5877
8	593	0.1384	1.5184
9	811	0.0615	**1.8791**
10	**2,140**	0.0206	1.1926
11	2,115	**0.0127**	**1.7829**
12	1,363	0.0089	1.1809
13	579	0.0065	1.3769
14	145	0.0065	1.2244
15	33	0.0052	1.1739
16	17	0.0017	1.2539
17	6	0.0013	**1.3957**
18	1	0.0006	1.0000

Informational Analysis Pipeline (by V. Bonnici)




159
4%
462
15%
644
30%
909
53%
1,323
46%
1,744
69%
1,970
83%
1,692
63%
1,888
72%
2,018
81%
1,084
41%
1,659
66%
1,847
82%
2,057
87%
1,838
75%
1,947
84%
1,217
56%
1,440
80%
1,476
70%
1,560
81%
1,721
86%
Shared 6-mers (Shared/ union)%
1,523 1,261 1,370 2,086 2,154 2,198 2,161 2,275 2,305 2,205 2,091 2,004 2,219 2,204 2,131 2,134 1,576 1,666 1,749 1,717 1,902 1,826
21 22 19 20 17 16 13 18 15 14 8 7 9 10 12 11 2 1 4 3 6 5

A jewel in word extraction (chr. 1 hg. 19)

Discarding extracted subwords and partitioning

Radicals and Maximal Morphemes

	5	5	5	5	5	5	5	5	5
6	1,666	**0.843**	**2.772**	1,632	0.841	2.728	34	0.039	1.057
7	**2,310**	0.769	2.588	1,446	0.646	1.922	864	**0.464**	**1.615**
8	593	0.138	1.518	186	0.074	1.297	407	0.086	1.327
9	811	0.061	1.879	71	0.016	1.288	740	0.054	**1.740**
10	2,140	0.021	1.193	63	0.010	1.138	2,077	0.012	1.073
11	**2,115**	0.013	1.783	22	0.007	**1.320**	**2,093**	0.006	**2.269**
12	1,363	0.009	**1.181**	13	0.006	1.228	1,350	0.003	1.093
13	579	0.007	1.377	11	0.005	**1.459**	568	0.001	1.012
14	145	0.006	1.224	8	0.006	1.253	137	0.001	1.055
15	33	0.005	1.174	3	0.003	1.134	30	**0.003**	1.000
16	17	0.002	1.254	2	0.001	1.502	15	0.001	1.000
17	6	0.001	1.396	2	0.001	1.768	4	0.001	1.000
18	1	0.001	1.000				1	0.001	1.000
All	**11,779**	**0.994**	**4.764**	**3,459**	**0.979**	**3.778**	**8,320**	**0.587**	**1.767**

Radicals: ca, sa, le

Maximal morphemes: casale

Dictionary Validation

Words extracted by informational methods are informationally relevant, but what about their biological meaning? (Infogenomics is analogous to ENCODE)

Words are pieces on which genomes were built.

Which categories emerge?

Words are, in this perspective, ***iper-dense information units***

How defining and discovering biological significance?

Can information tell us deep biological mechanisms?

****

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

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

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

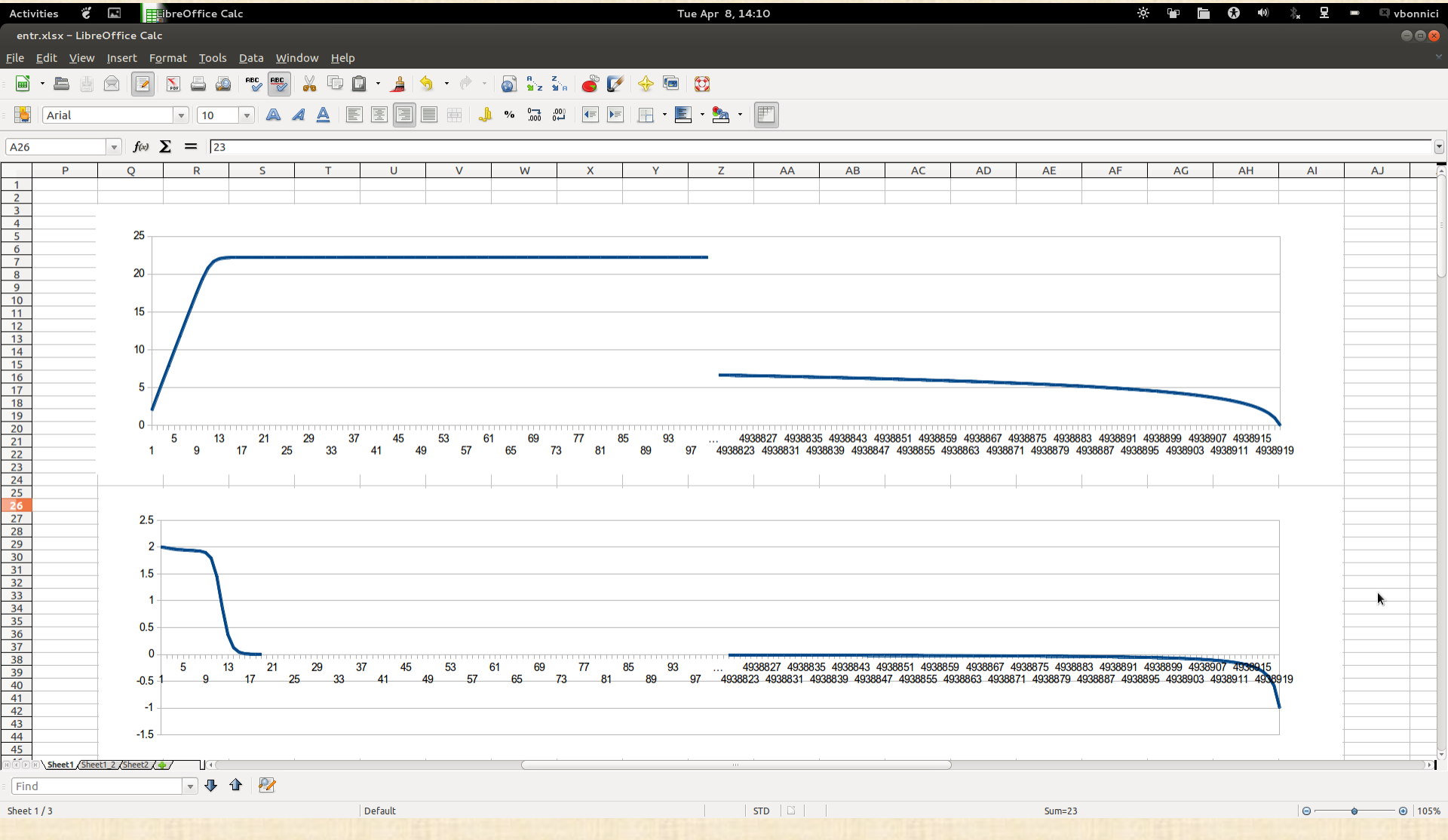
We computed Empirical Entropy
for any word length, and for all Human chr.
(k= 18 , $E_k \approx 24$; k=200 $E_k \approx 25$!!!)

Algorithmic basis of k-mer frequency computation

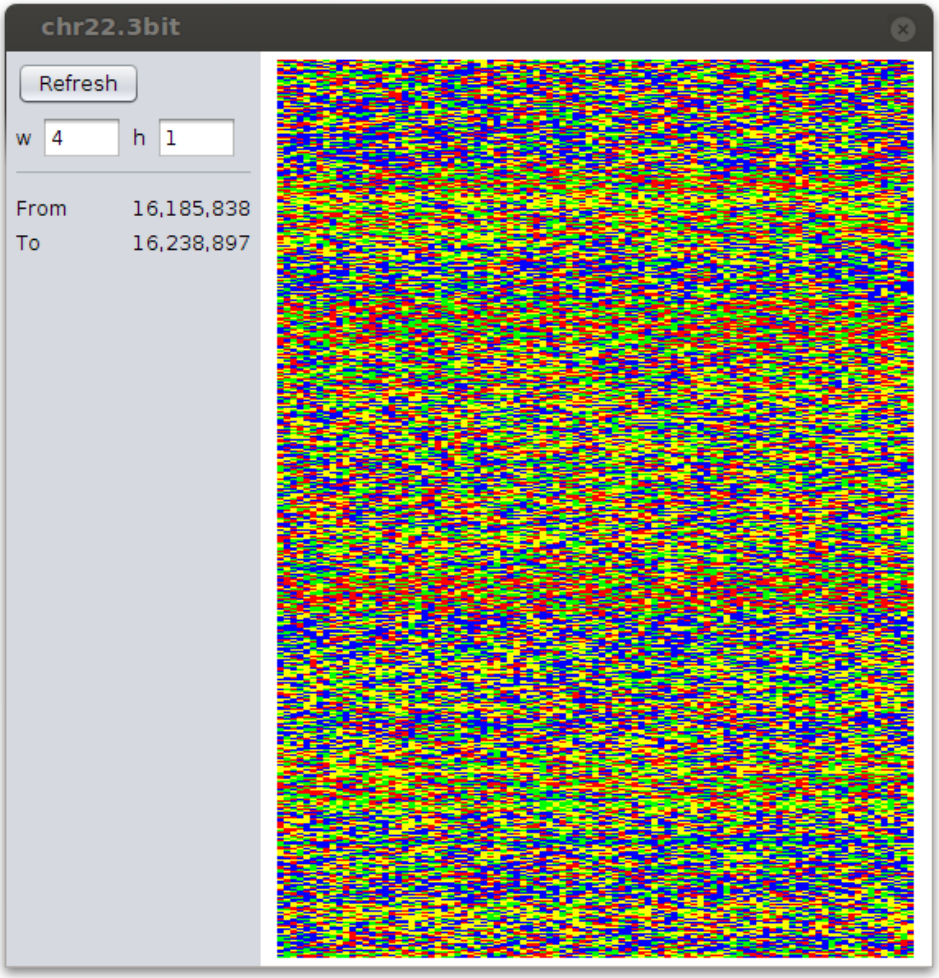
- Suffix trees ST
- Suffix arrays SA
- Enanched SA ESA
- N-extended ESA NESA

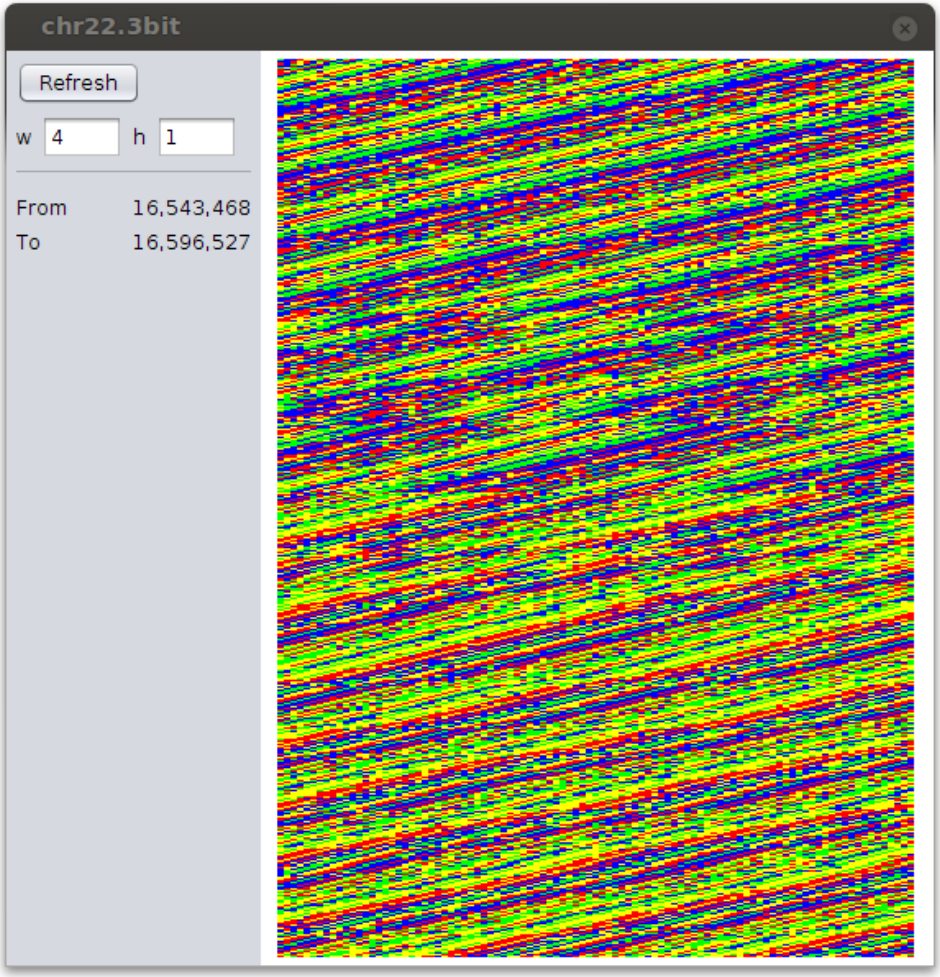
Weiner 73
McCreight 76
Ukkonen 95
Farach 97
Manber & Myers 90
Abouelhoda, Kurtz, Ohlebusch 2004
Kurtz et a. 2008

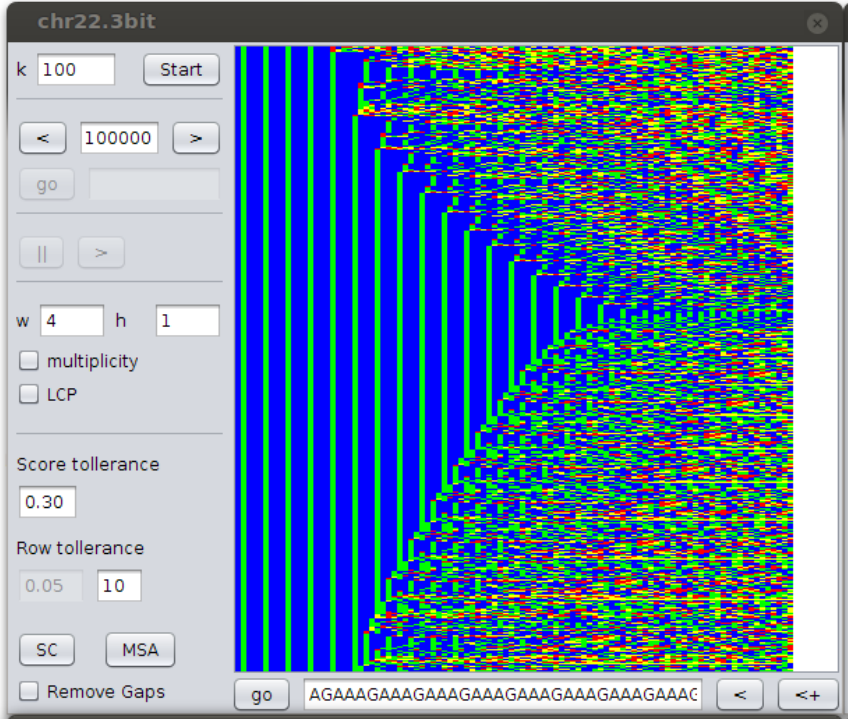
Entropy and Excess Entropy (E. coli)

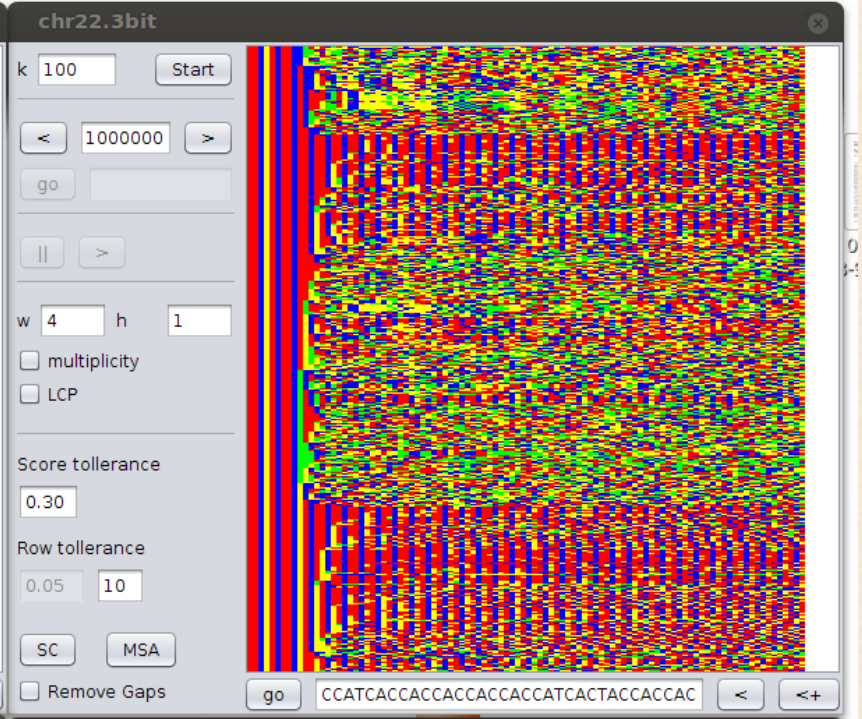


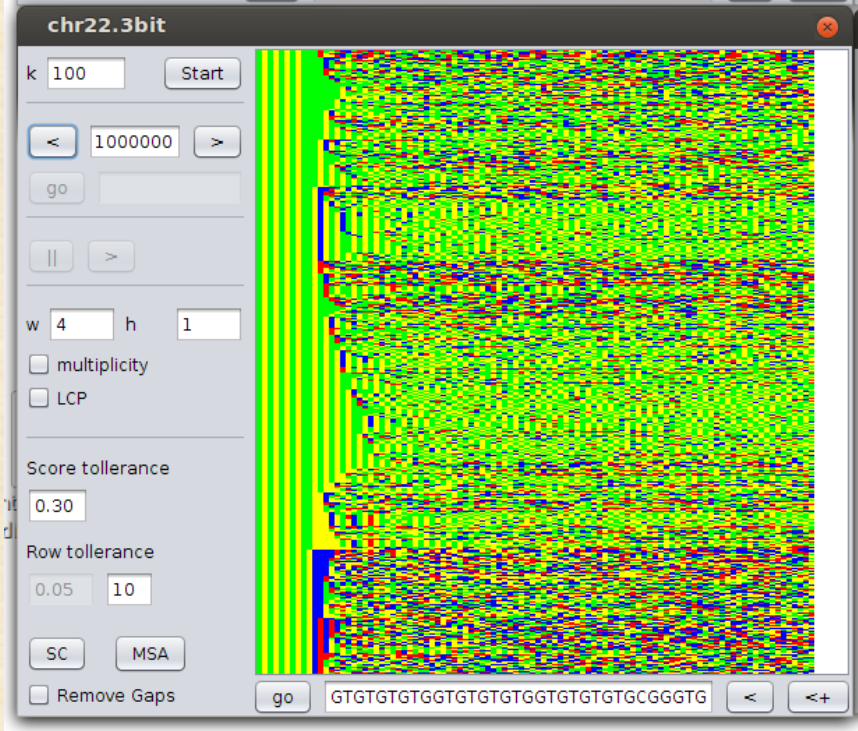
Genomic Chromatic lines

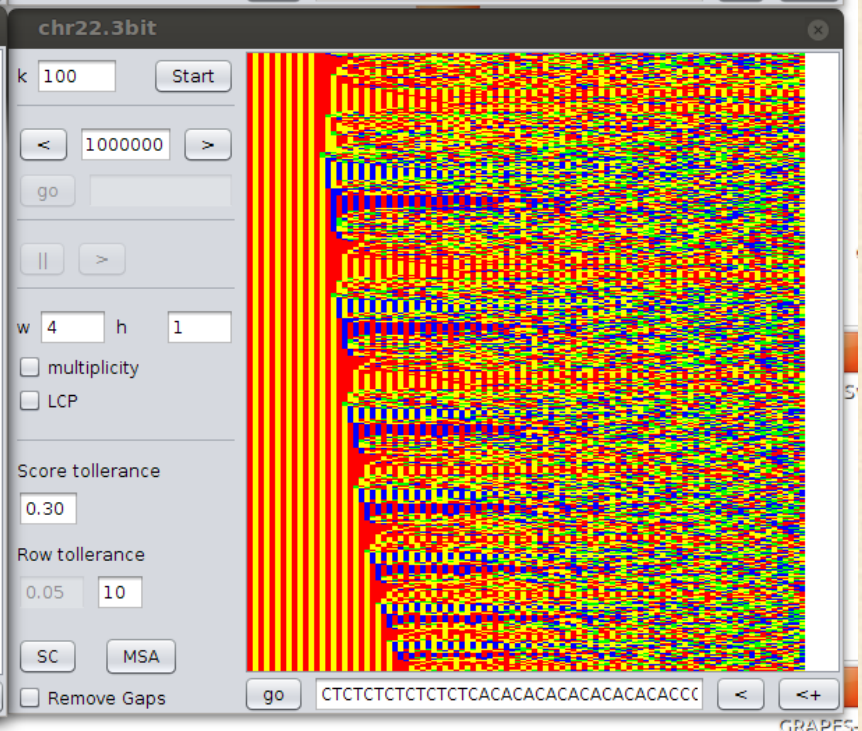


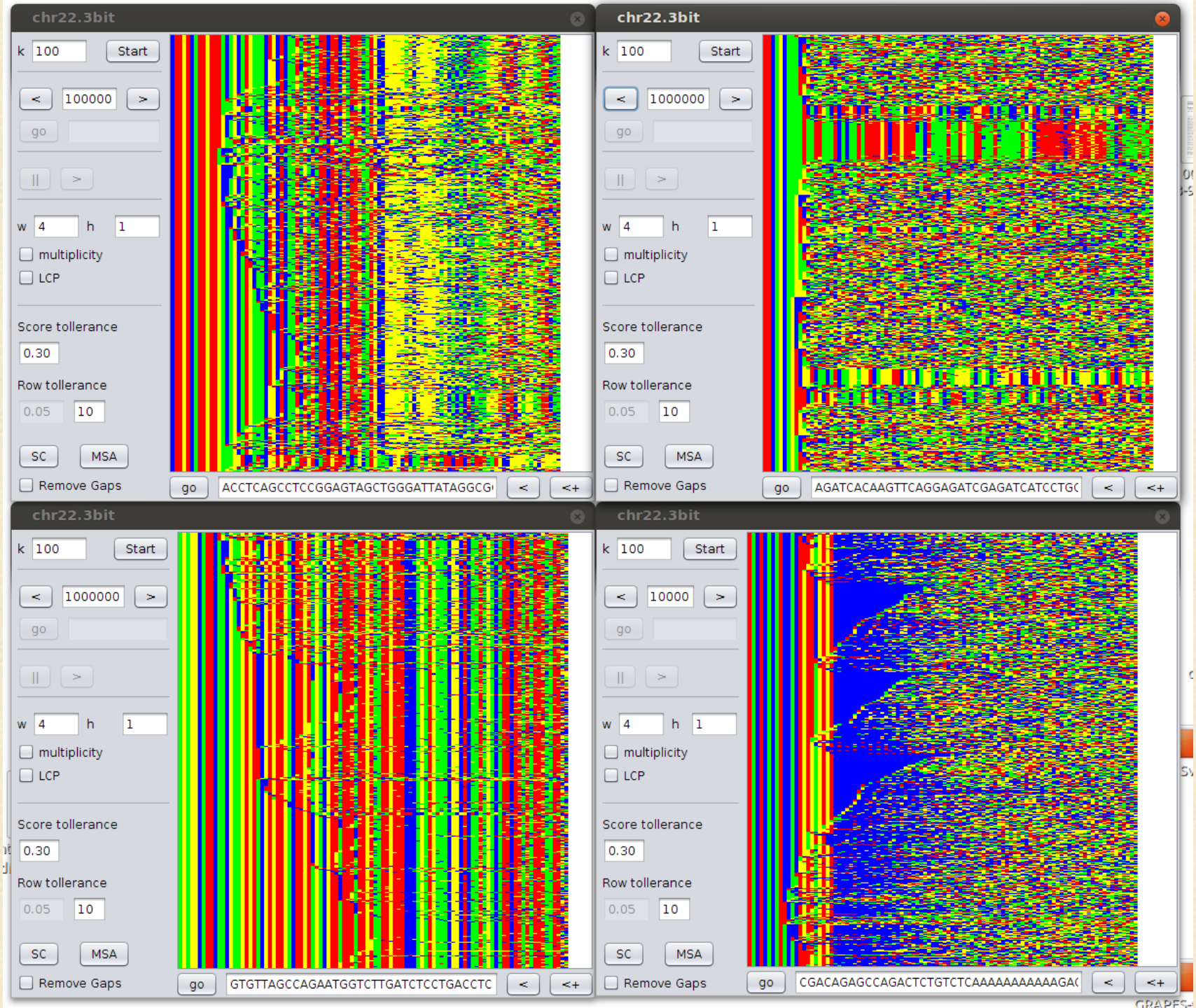


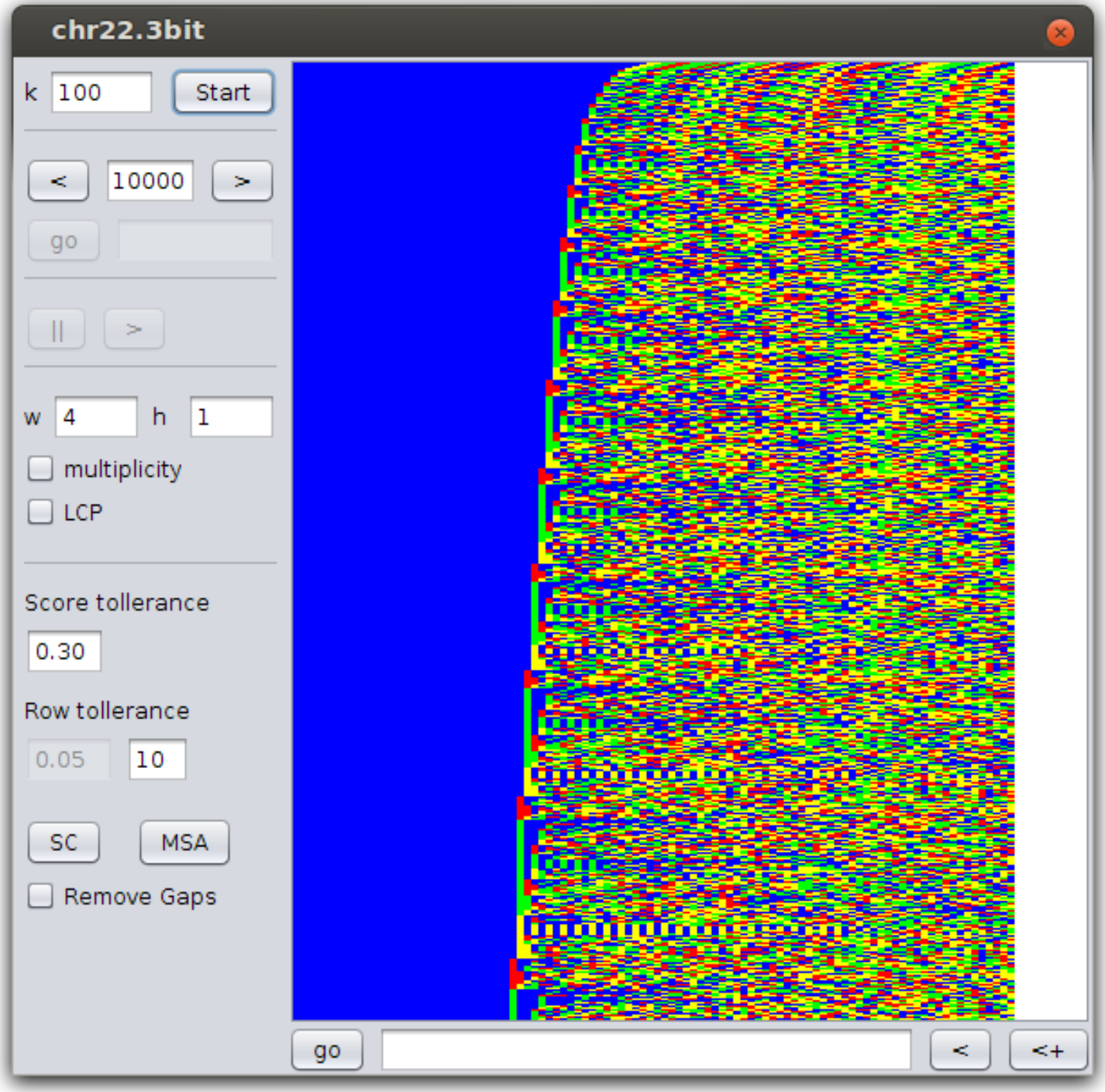

chr22.3bit
k 100
Start
< 100000 >
go
||
>
w 4 h 1
multiplicity
LCP
Score tollerance
0.30
Row tollerance
0.05 10
SC
MSA
Remove Gaps
go
AGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAC
<
<+


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
CCATCACCACCACCACCATCACTACCACCAC
<
<+


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
CTCTCTCTCTCTCACACACACACACACACCC
<
<+


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
<
<+

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(G)

the formula is not simple to explain
(empirical entropy, RND, KL divergence)

It is related to the maximum of KL divergence between G and any R in $RND_{|G|}$

Some biobit computations

Species	$\|G\|$	N%	CG%	Genes	$Bb(G)$
Nanoarchaeum equitans	0.49 Mb	0.00	31	585	0.011
Candidatus Carsonella ruddii	0.15 Mb	0.00	16	182	0.028
Escherichia coli	4.93 Mb	0.00	50	5k	0.038
Mycoplasma genitalium	0.58 Mb	0.00	31	558	0.042
Cyanidioschyzon merolae	14.9 Mb	0.00	55	6k	0.111
Saccharomyces cerevisiae	12.0 Mb	0.00	38	5k	0.145
Trichomonas vaginalis	0.58 Mb	0.36	32	60k	0.147
Arabidopsis thaliana	119 Mb	0.16	36	33k	0.176
Caenorhabditis elegans	100 Mb	0.00	35	46k	0.198
Oryza brachyantha	250 Mb	6.43	41	42k	0.274
Drosophila melanogaster	99 Mb	0.98	42	30k	0.281
Oryza glaberrima	285 Mb	4.00	41	60k	0.375
Vitis vinifera	426 Mb	2.35	35	26k	0.655
Danio rerio	1.34 Gb	0.14	36	40k	1.061
Macaca mulatta	2.88 Gb	11.21	42	30k	1.552
Papio anubis	2.72 Gb	1.54	42	36k	1.603
Callithrix jacchus	2.77 Gb	5.38	41	44k	1.617
Nomascus leucogenys	2.79 Gb	6.57	41	25k	1.721
Sus scrofa	2.59 Gb	10.52	42	34k	1.751
Pongo abelii	3.02 Gb	10.12	41	32k	1.762
Homo sapiens GRCh37	3.09 Gb	7.57	41	39k	1.768
Monodelphis domestica	3.60 Gb	2.88	38	34k	1.795
Pan troglodytes	3.17 Gb	13.33	42	35k	1.804
Rattus norvegicus	2.78 Gb	4.68	42	37k	1.824
Homo sapiens GRCh38	3.08 Gb	4.88	41	56k	1.886
Mus musculus	2.75 Gb	2.86	42	45k	2.151

biobit is an **anti-entropic**, rather than neghentropic, measure of **genome information.**

An upper bound for Bb(G), say it BB(G), is obtained by considering De Bruijn's sequences B(4,k) where $|G| = 4^k + k - 1$ that is, $k \approx \lg_2(|G|)$. In these genomes only k-hapaxes occur.

From Boltzmann to Carnot in cell state analysis

$$PV = nRT$$

Could you recover this law from the dynamical equations of the single particles colliding in the gas?

NO! So analogously, in the cell, we need to abstract from single biochemical molecule dynamics of about 10^4 different types (or macro-types) of molecules (which kind of distribution?)