An INFOrmational GENOMICS approach

(based on joint work with V. Bonnici, A. Castellini, V. Manca)

Dr Giuditta Franco

Department of Computer Science, University of Verona, Italy

Computational genomics

Study of genomes is referred to as *genomics* - while *genetics* concerning the study of (single of groups of) genes.

Computational/statistical genomics include gwas (genome wide association studies), wgs (whole genome sequencing algorithms), *advanced data structures (suffix trees, arrays, BWT)* and big data approaches (to process massive amount of data).

Local sequence (similarity) analysis is traditionally performed by **alignment-based methods** (FASTA, BLAST). Traditional **alignment-based methods** are applied for multiple sequence analysis and comparison, with phylogenetic and pharmacogenomics applications.

Alignment free methods

Recent (ten years) **alignment-free methods**[1] in computational genomics focus on a systemic view, based on empirical studies of frequencies of DNA k-mers (genome factors of length k), out of whole genomes/proteomes [2], where FL and information theories are applied.

For a fixed length code, empirical frequencies are computed on k-mers as normalized multiplicities.

[1] Vinga et al. '03, Fofanov et al., 04, Chor et al. '09, Searls '10

[2] set of proteins possibly expressed by a cell, tissue, organism.

Dictionary-based genomics

As a *very long* string over an alphabet of four/five symbols, a genome may be seen as a text book, which language has to be completely deciphred. Information is comprised in words, *lineraly* arranged by unknown synthax and semantics.

Sequences become **bags of words**, used in computer vision, for document classification by clustering of strings which keep only their multiplicity/feature (*no order*, no grammar).

Investigation and comparison (e.g., similarity analysis) by **dictionaries** or multisets. Models common to natural language processing, information retrieval (IR), computational linguistics. This abstraction excludes tandem repeats.

Informational genomics: main ingredients

Genomes are information source, whose words are random variable values, data encodings, and a probability distribution is given by frequencies [3].

Infogenomics employs methods from FLT [4] and IT to define genomic indexes, dictionaries, distributions, elongation and segmentation methods, sequence similarity measures, gene networks [5].

[3] Sims et al. PNAS, 2008. Robins et al. Journal of Bacteriology 2005
[4] V. Brendel at al. Nucleic Acids Research 12, 1984.
[5] A. Castellini et al. BMC Genomics 2012, Nat Comp 2015.

Lecture outline

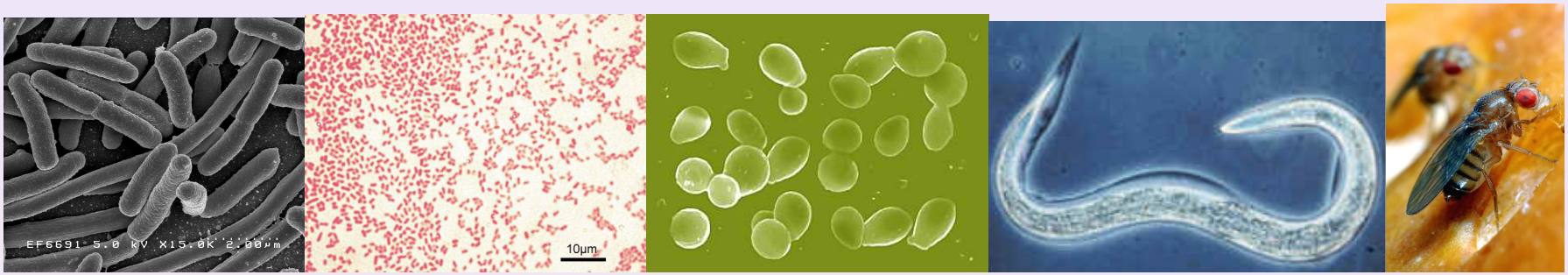
An informational dictionary-based alignment-free approach:

- Genomic dictionaries, distributions, indexes.
- Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) by **dictionary intersections**
- Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
- Genomic Recurrent Distance Distribution (**RDD**), and cluster analysis on repeat sharing **gene networks**.

Software Tools

Our work focused on real genomes

Dictionaries of 60 specific genomes, deeper analysis on 12



Organism Genome	Length	Genes	Type
Nanoarchaeum equitans	490,885	585	Minimal archaeum
Mycoplasma genitalium	580,076	476	Minimal bacterium
Mycoplasma mycoides	1,211,703	1,016	Venter's experiment bacterium
Haemophilus influenzae	1,830,138	1,717	First sequenced bacterium
∗*Escherichia coli*	4,639,675	4,685	Bacterium model (K-12)
∗*Pseudomonas aeruginosa*	6,264,404	5,566	Ubiquitous bacterium
∗*Saccharomyces cerevisiae*	12,070,898	6,275	Unicellular eukaryote (Yeast)
Sorangium cellulosum	13,033,779	9,700	Longest genome bacterium
Homo sapiens chr. 19	63,800,000	2,066	Highest gene density H. chr
∗*Caenorhabditis elegans*	100,267,632	19,000	Worm (around 1000 cells)
∗*Drosophila melanogaster*	129,663,327	14,000	Insect (fruit fly)
Homo sapiens chr. 1	247,000,000	3,511	Longest Human chr

Basic definitions

Alphabet $\Gamma = \{a, t, c, g\}$, genome $G \in \Gamma^\star$.

- $D_k(G)$: **genomic k-dictionary** collecting all k-mers in G, $F_k(G)$ of **forbidden k-words**, which do not occur in G

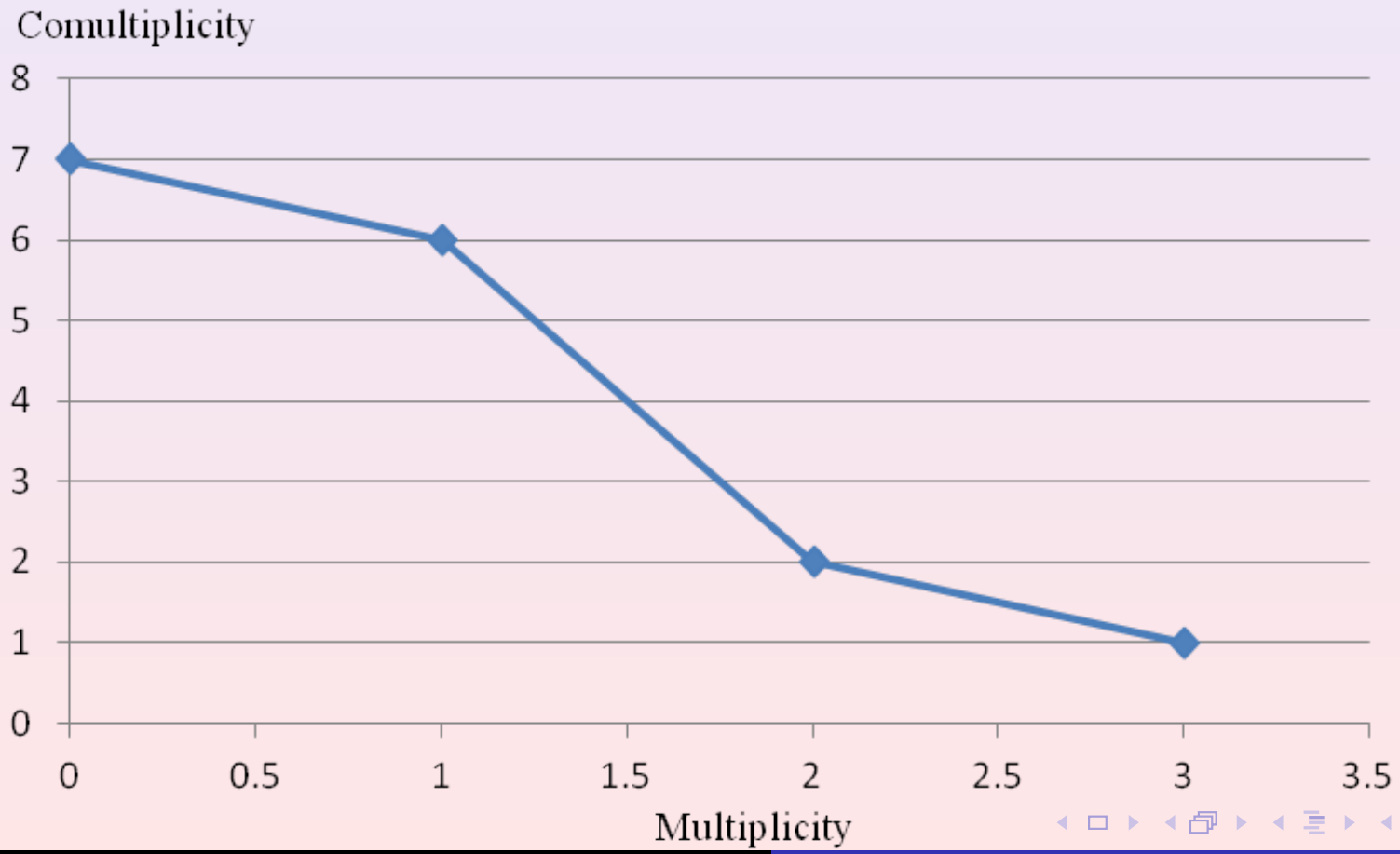
  Ex: *attaggatcttaat* has nine 2-words: six occurring once (aa, ag, tc, ct, ga, gg), two occurring twice (ta, tt), one (at) occurring 3 times, and seven 2-forbidden.

- $H_k(G)$: dictionary of **hapax k-words**, occurring once in G, $R_k(G)$ the set of **k-repeats**, occurring in G at least twice

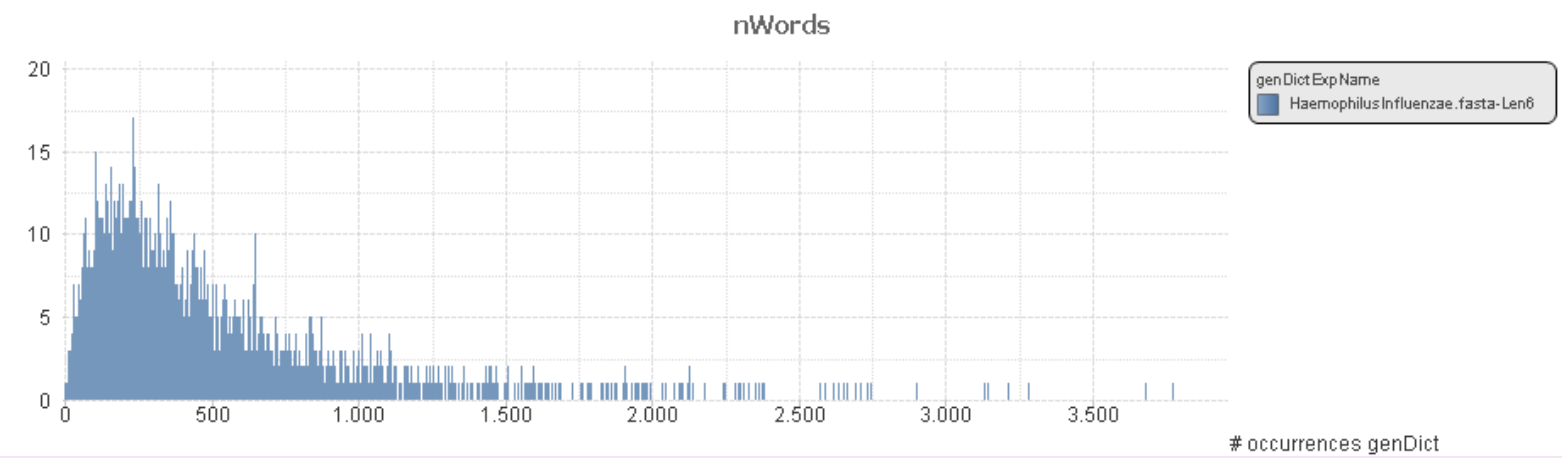
- $T_k(G)$ is the **k-factor multiset** of G: a funcion over $D_k(G)$ associating each string to its number of occurrences in G

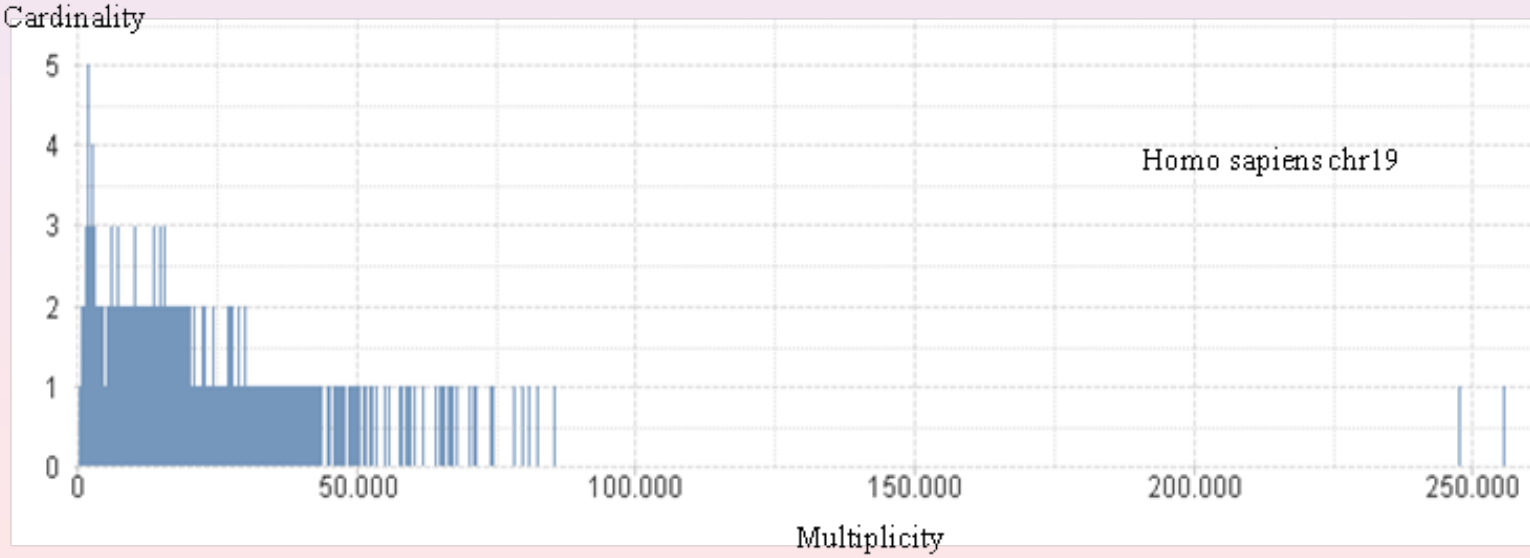
Multiplicity-comultiplicity profiles

Multiplicity-comultiplicity 2-distribution diagram for the sequence *attaggatcttaat*



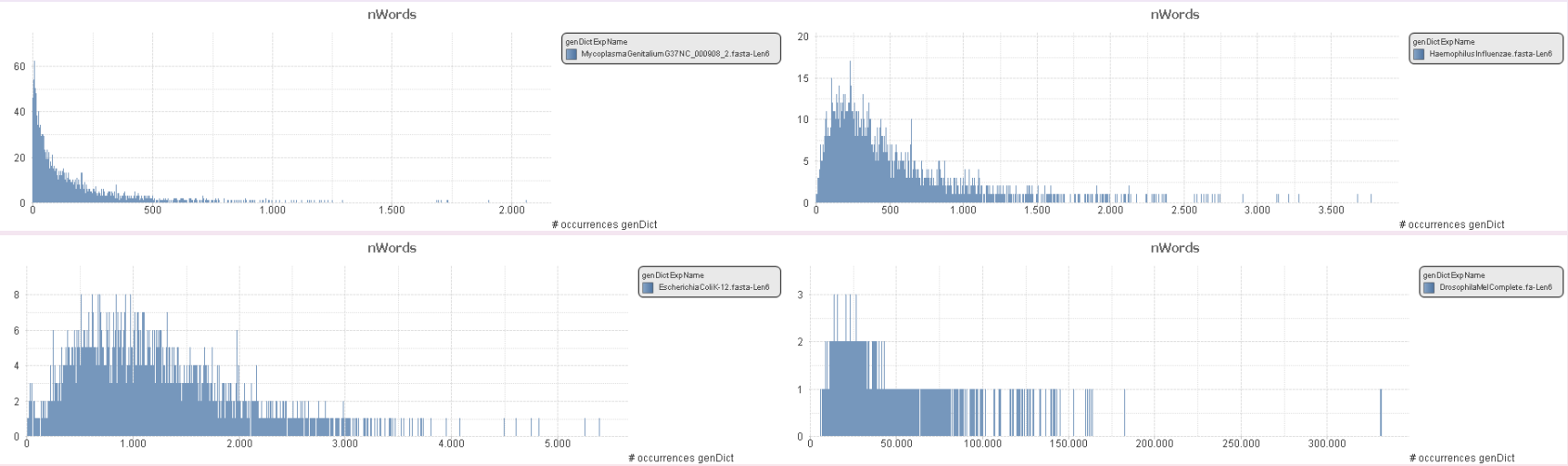
Genomic profiles





A related string problem

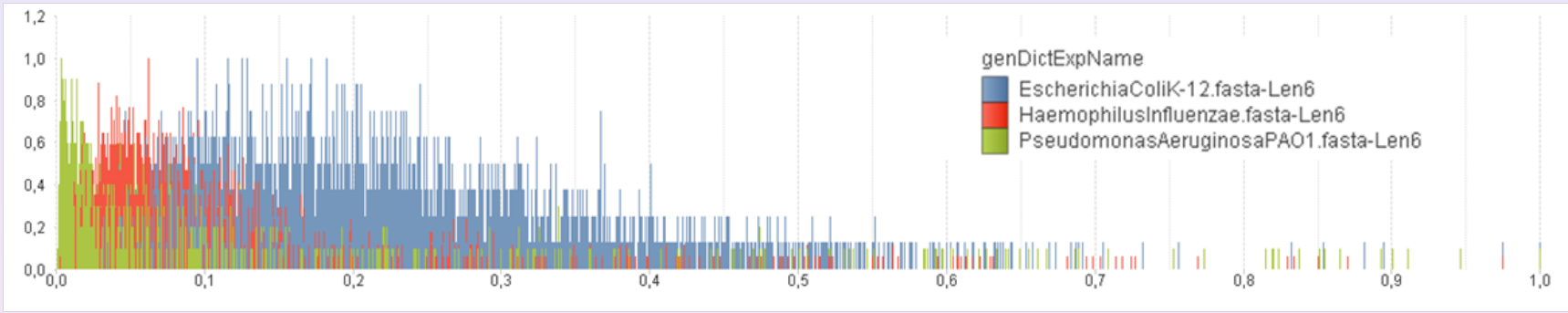
Related to the *word assembly problem*: genome reconstruction method, from a dictionary (examers), or a mult-comult diagram.



M. genitalium, H. influenzae, E. coli, D. melanogaster.

These studies could improve existent genome reconstruction algorithms, by estimations of reads length repeatability.

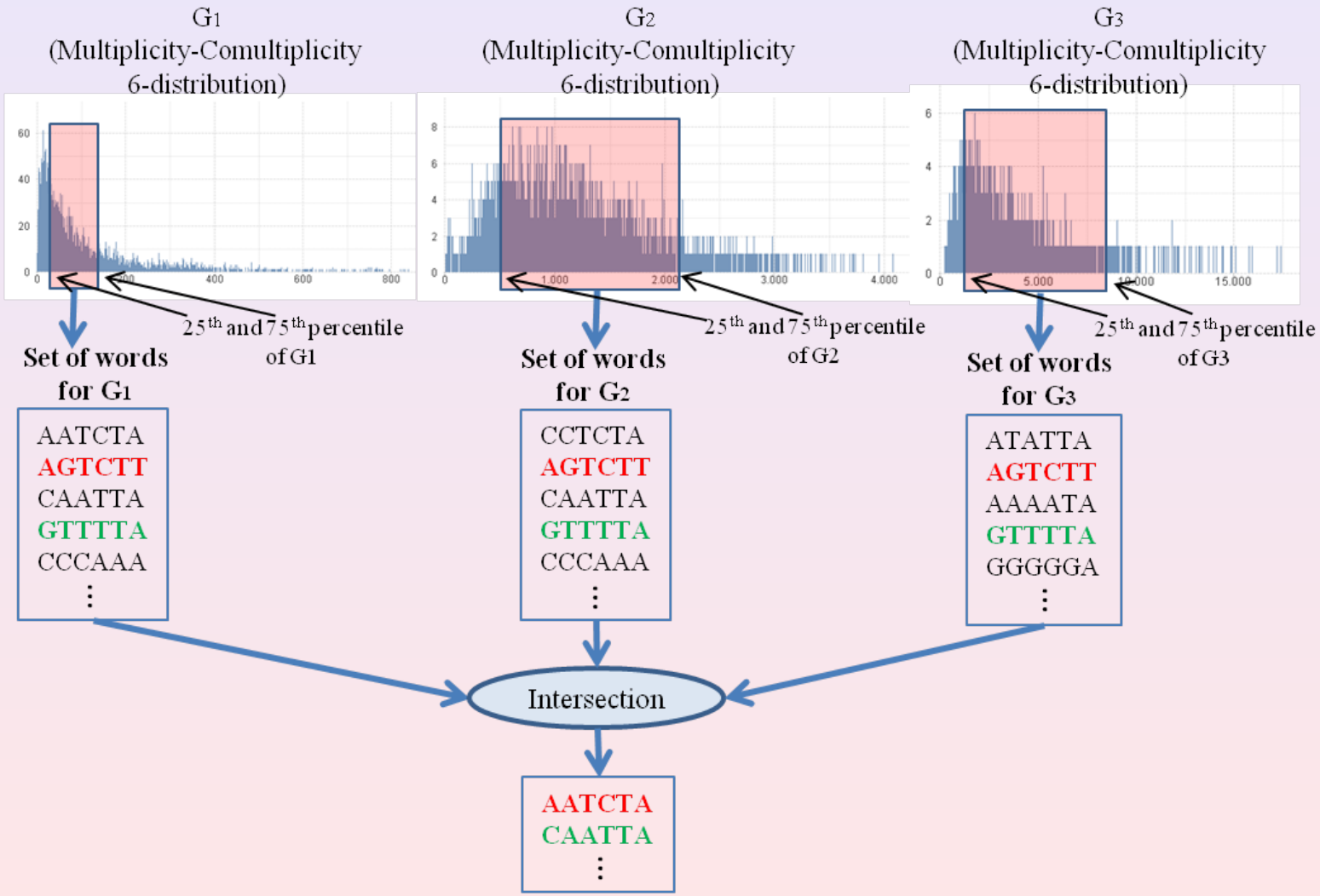
Comparison among genomic profiles - examples



Above, *E. coli*, *H. influenzae*, *P. aeruginosa* are compared by their normalized genomic (6-)profiles.

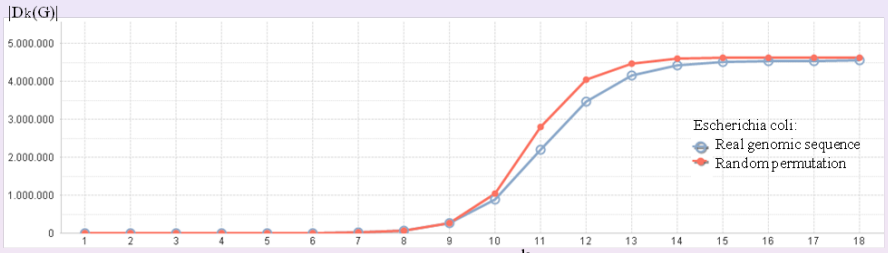
Aside, *M. genitalium*, *E. coli*, *S. cerevisiae*, *H. sapiens chr 19*, are compared with one random permutation (in red) by the multiplicity-comultiplicity profiles.

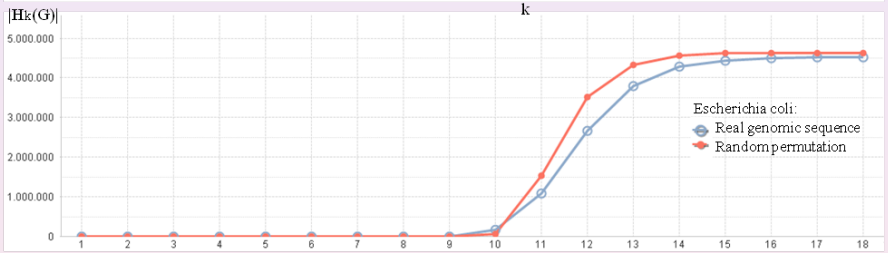
Fairly occurring motifs common to genomes (UCE)

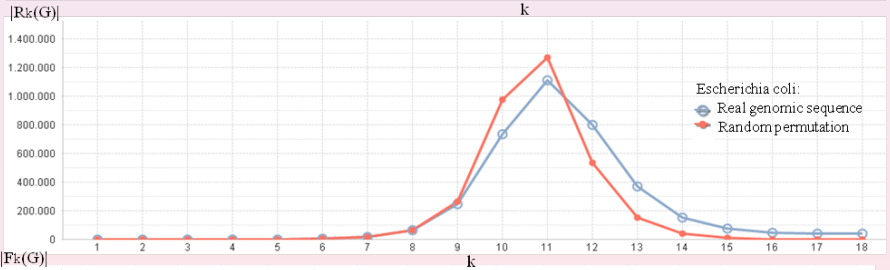


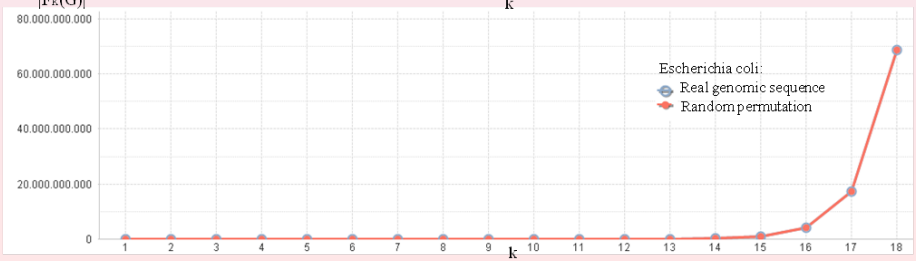
Sizes of *k*-dictionaries: : real vs random genomes

E. coli 's genomic dictionaries

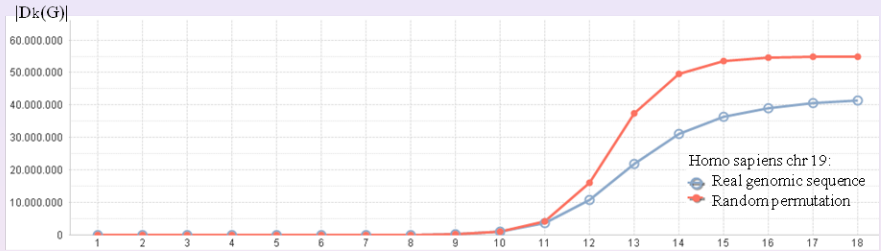


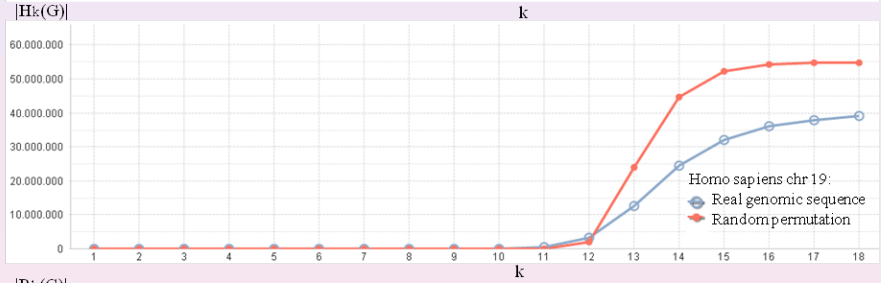


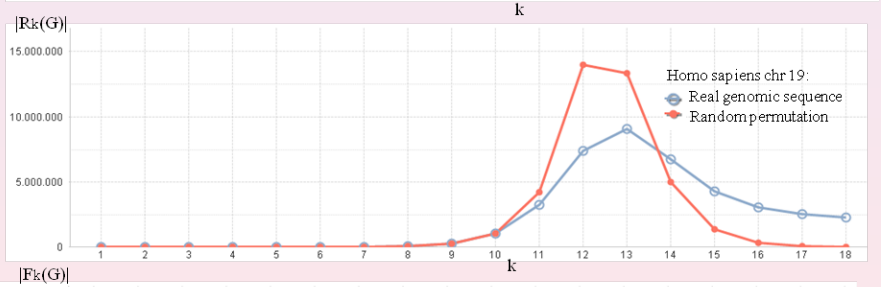


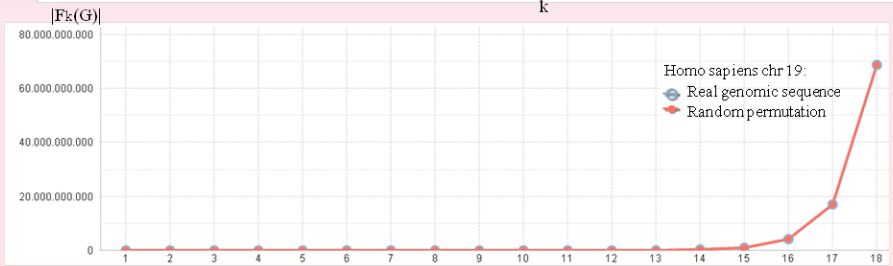


H. sapiens chr19 's dictionaries



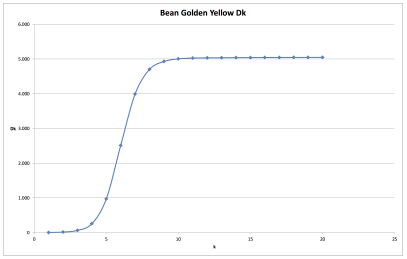


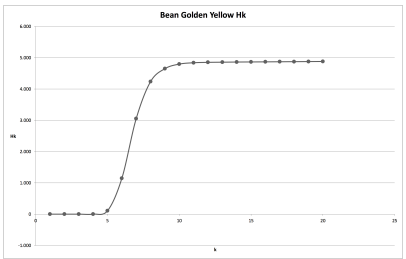


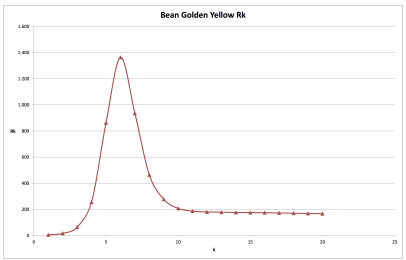


Sizes of *k*-dictionaries: : virus and bacterium

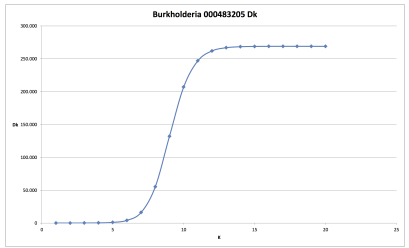
BeanGoldenYello 's dicts

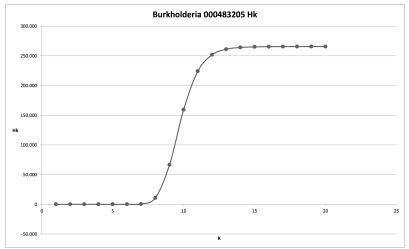


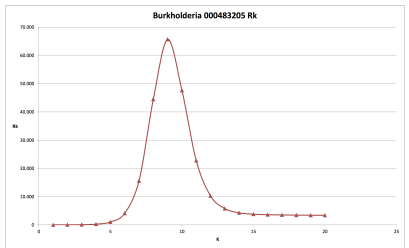




Burkholderia 's dicts







A related open problem

Observed curves of $|D_k|$ (of $|H_k|$, $|R_k|$, and F_k) exhibit a similar shape for some genomes having a sensibly different length.

Open problem: the discovery and comprehension of some rule explaining the empirically evident relationship among genome length n, factor length k, and k-dictionaries cardinality.

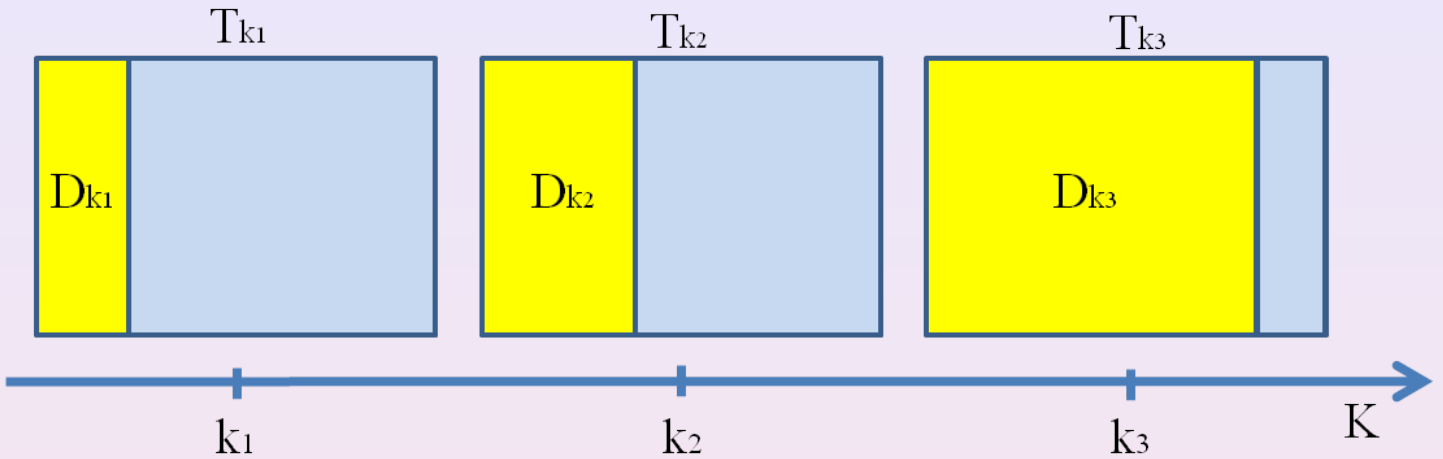
Phase transitions for hapax/repeat cardinality ratio

$DT_k = \frac{|D_k(G)|}{|T_k(G)|}$ (k-lexicality), $HR_k = \frac{|H_k(G)|}{|R_k(G)|}$

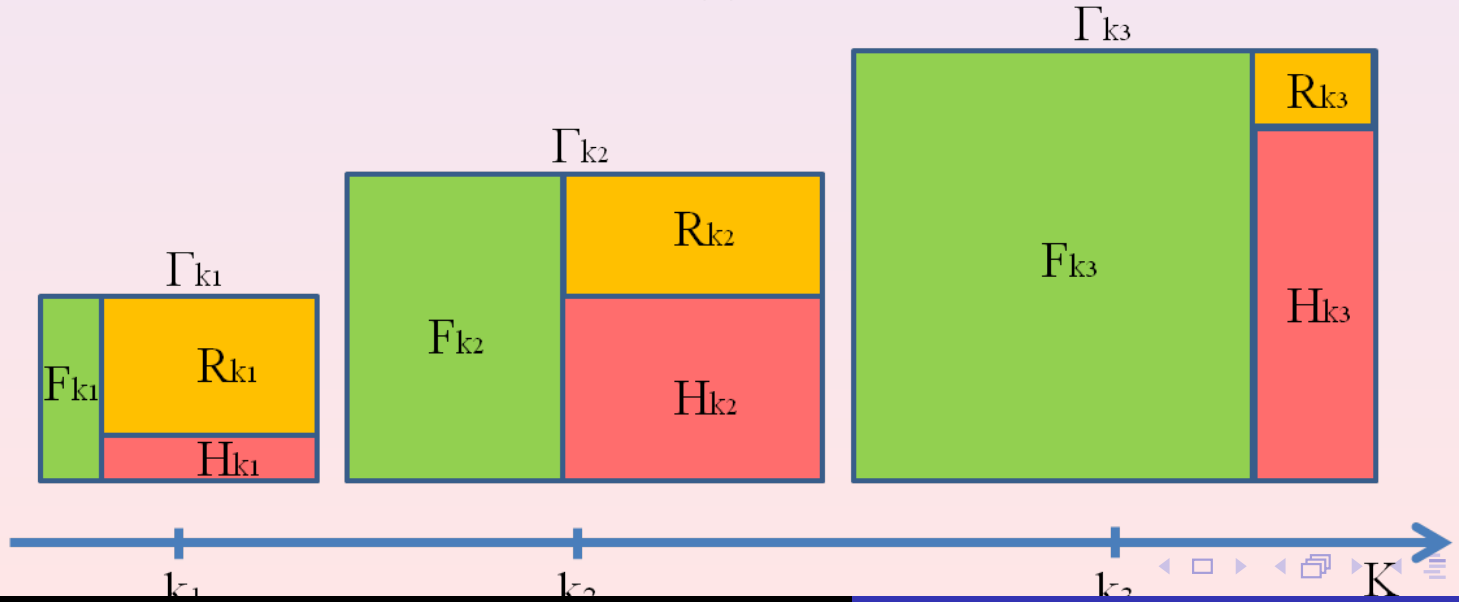
Genomes	DT_6	DT_{12}	DT_{18}	HR_6	HR_{12}	HR_{18}
N. equitans	0.008	0.87	0.99	1.468×10^{-3}	8.39	737.25
M. genitalium	0.007	0.85	0.98	8.65×10^{-3}	7.175	91.44
M. mycoides	0.003	0.53	0.81	9.661×10^{-3}	2.169	12.33
H. influenzae	0.002	0.81	0.98	0	5.240	88.93
E. coli	0.0009	0.74	0.98	⋮	3.331	115.84
P. aeruginosa	⋮	0.47	0.98		1.564	93.76
S. cerevisiae		0.54	0.95		1.518	58.67
S. cellulosum		0.29	0.96		0.993	41.12
H. sapiens chr19		0.19	0.75		0.455	17.27
C. elegans		0.13	0.89		0.286	19.86
D. melanogaster		0.12	0.90		0.114	32.56

$|T_k|$ = number of k-mers counted with their multiplicity (|G|-k+1). HR18 explains why microarrays work well.

Analysis of (relatively) long repeats reduction



(a)



Genomic Dictionaries

- D(G) = {G[i,j] |1 ≤ i ≤ j ≤ |G|} (square dim. w.r.t. |G|)
- $D_k(G)$ = D(G) ∩ Γ^k
- L included in D(G) is a dictionary of G
- A position p of G is m-covered in D if there are m words G[i,j] of D with i ≤ p ≤ j **(positional coverage)**
- D covers G if every position of G is k-covered with k ≥ 1 by D **(lexical coverage)**
- D minimally covers G if D covers G and no D' included in D covers G
- G is D-segmentable if G belongs to D*

Slide courtesy of prof. Vincenzo Manca, Univ. Verona, IT

Basic Genomic Indexes

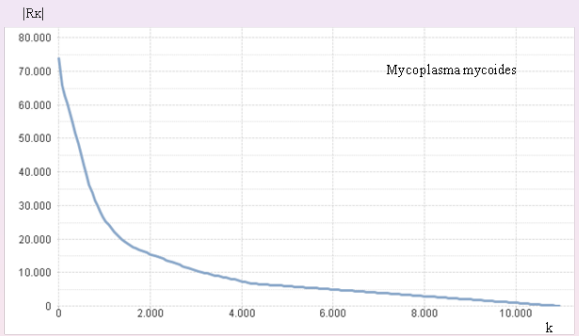
- ***LG*** Logarithmic Length
- ***LX_k*** k-Lexical Multiplicity
- ***MFL*** Minimal Forbidden Length (**MCL** = MFL -1)
- ***MRL*** **Maximum Repeat length** (+1 = all-hapax ***lub*** least upper bound)
- ***MHL*** Minimum Hapax Length (-1 = all-repeat ***glb*** greatest lower bound)
- **COV** Coverage percentage (w.r.t. a dictionary)
- **POV** 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
- **IDG** **Inverse de Brujin graph indexes**

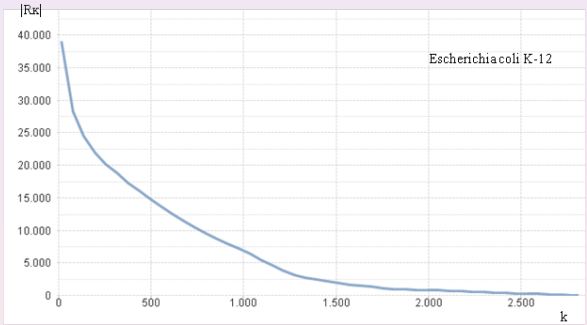
Slide courtesy of prof. Vincenzo Manca, Univ. Verona, IT

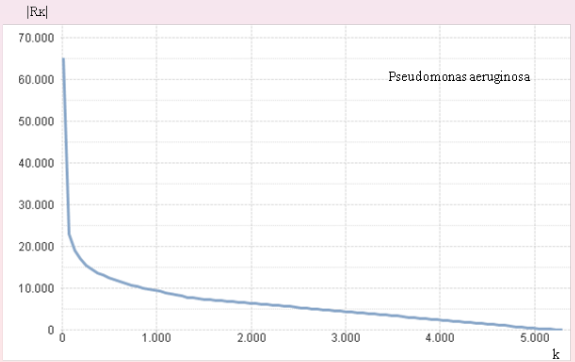
Informational indexes

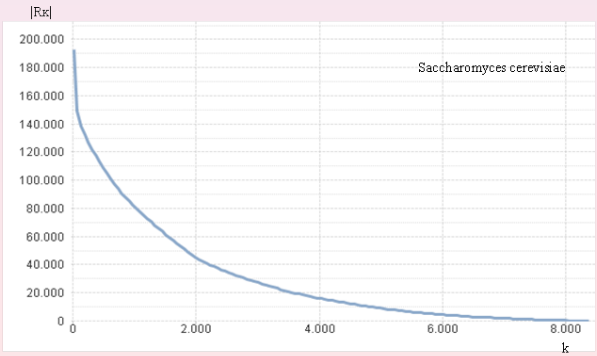
Minimal Hapax Length (genome itself is an hapax, any word including an hapax is an hapax), *Maximal Repeat Length* (any subword of repeat is a repeat), *Minimal Forbidden Length*

Genomes	MF	MR
N. equitans	6	139
M. genitalium	6	243
M. mycoides	6	10,963
H. influenzae	7	5,563
E. coli	7	2,815
P. aeruginosa	8	5,304
S. cerevisiae	9	8,375
S. cellulosum	7	2,720
H. sapiens chr19	9	2,247
C. elegans	10	38,987
D. melanogaster	11	30,892



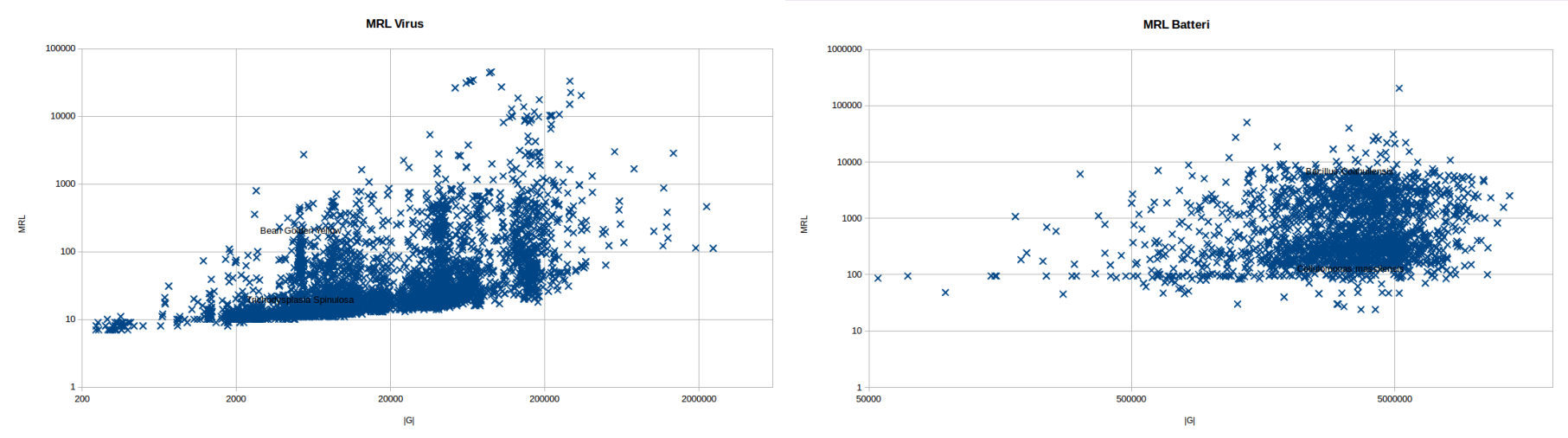






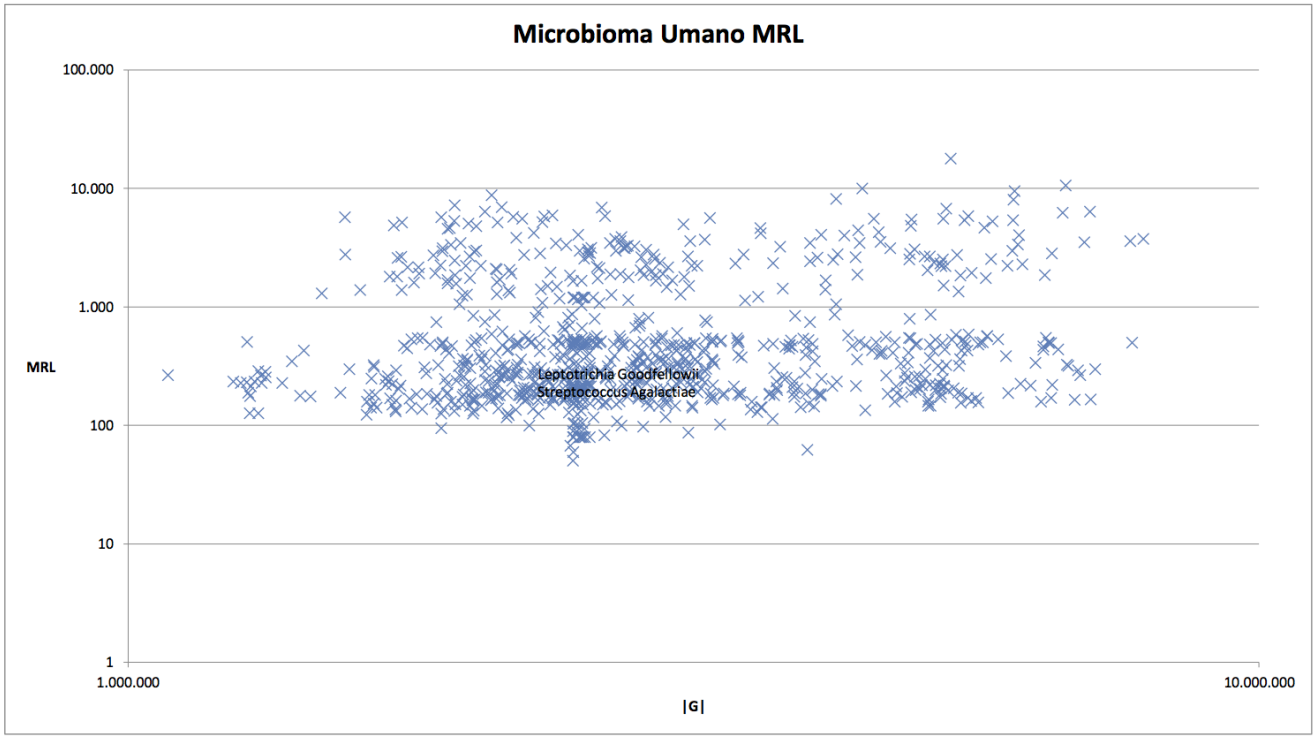
MRL of virus and bacteria

97% viruses has MRL ranging 10-10^3, 93% bacteria 100-10^4



Slides courtesy of Dantoni-Mancini-Sartea, Univ. Verona, IT

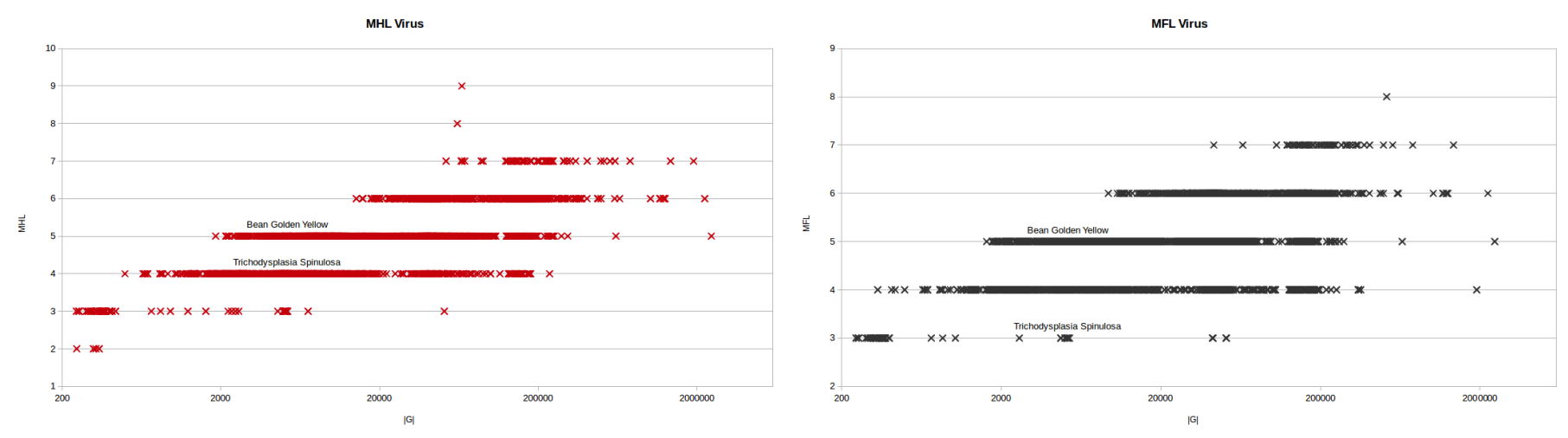
Human Microbiome MRL



Slide courtesy of Dantoni-Mancini-Sartea, Univ. Verona, IT

MHL and MFL for viruses

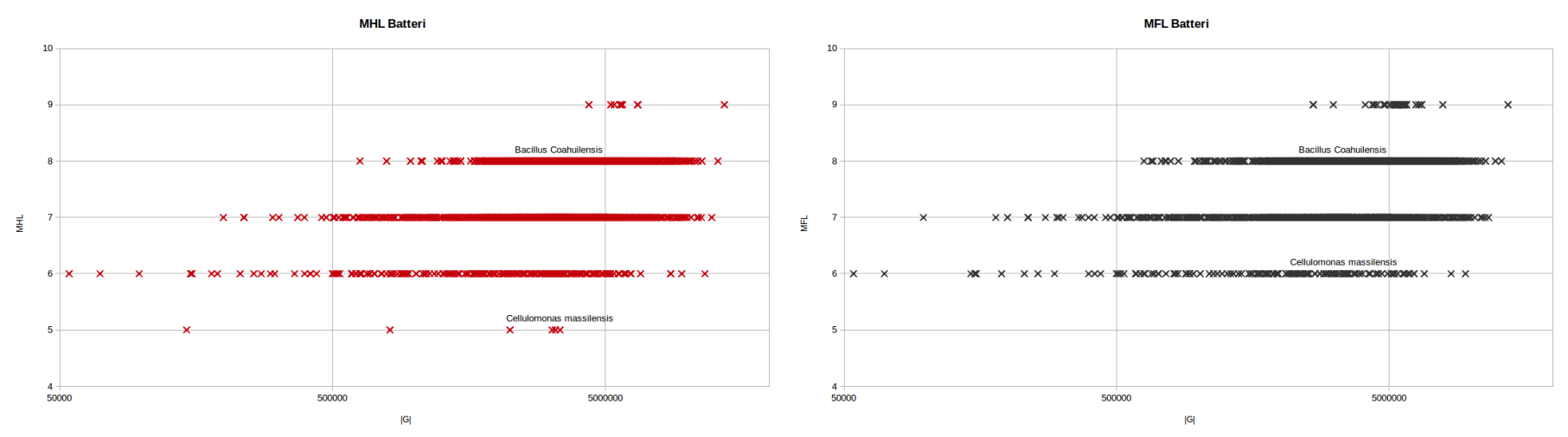
Major part of viruses has MHL and MFL between 4 and 6.



Slides courtesy of Dantoni-Mancini-Sartea, Univ. Verona, IT

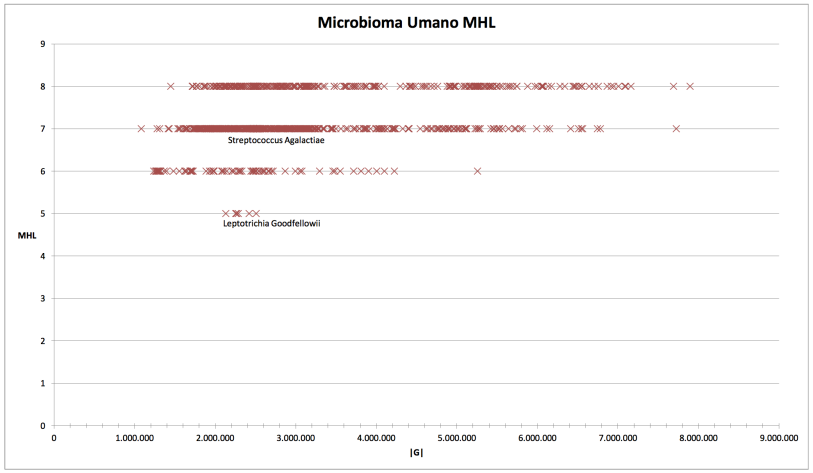
MHL and MFL for bacteria

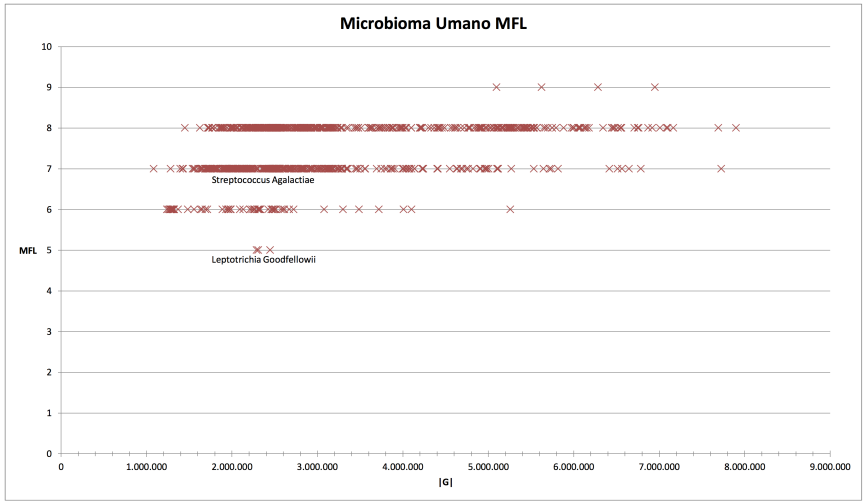
Major part of bacteria has MHL and MFL between 6-8.



Slides courtesy of Dantoni-Mancini-Sartea, Univ. Verona, IT

Human Microbiome: MHL and MFL confirmed





Slides courtesy of Dantoni-Mancini-Sartea, Univ. Verona, IT

Informational indexes: some properties

For any genome G (of length n):

$H_k = \Gamma^k \cap H, R_k = \Gamma^k \cap R \Rightarrow D_k = H_k \uplus R_k, \Gamma^k = D_k \uplus F_k$

$AR_k = \frac{|T_k \backslash H_k|}{|R_k|}$ average k-factors repeatability

$S_n = \lfloor \lg_4 n \rfloor - MF + 1$ factor length selectivity

$MFL \leq MHL + 1$

Importance of genomic repeats

Repeats located in genes (insertion of ALU sequences) have a relevant role for survival and evolution of primates, as they may alter the rate of protein production

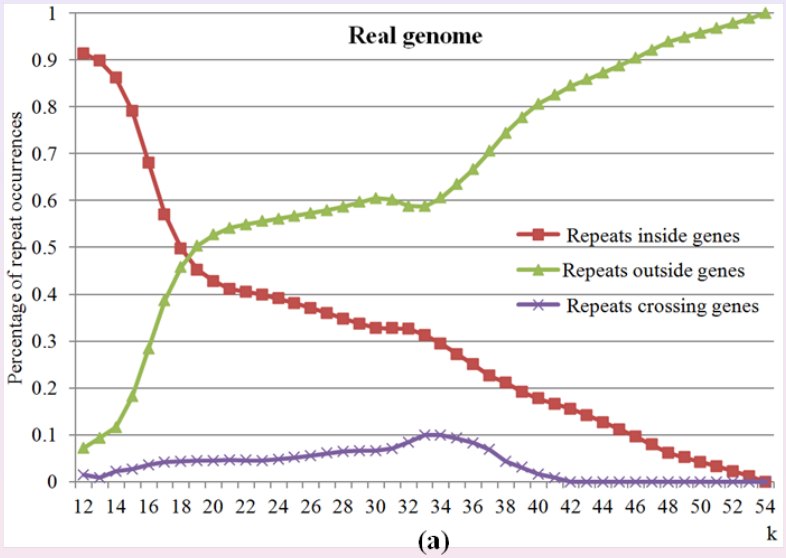
Correlation with number of mutations in (encoding regions of) many cancer-related genes (such as PTEN)

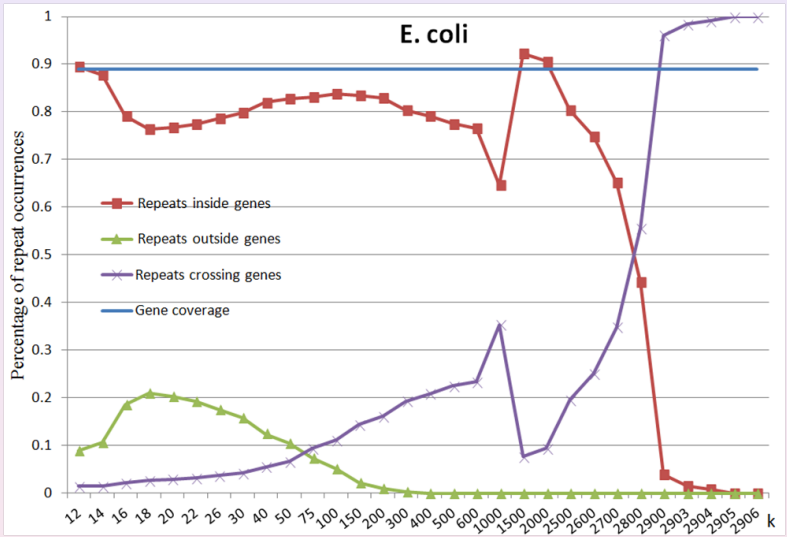
Tumors with microsatellite instability (MSI), characterized by massive instability in repeated sequences, display a strong microsatellite mutator phenotype (MMP)

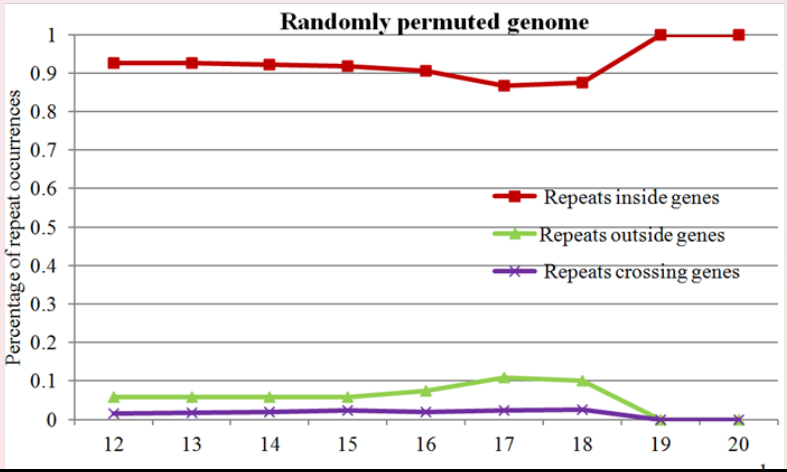
Fragile X syndrome (a genetic disorder inducing mental retardation) is associated to the expansion of CGG affecting the (FMR1) gene on the X chromosome

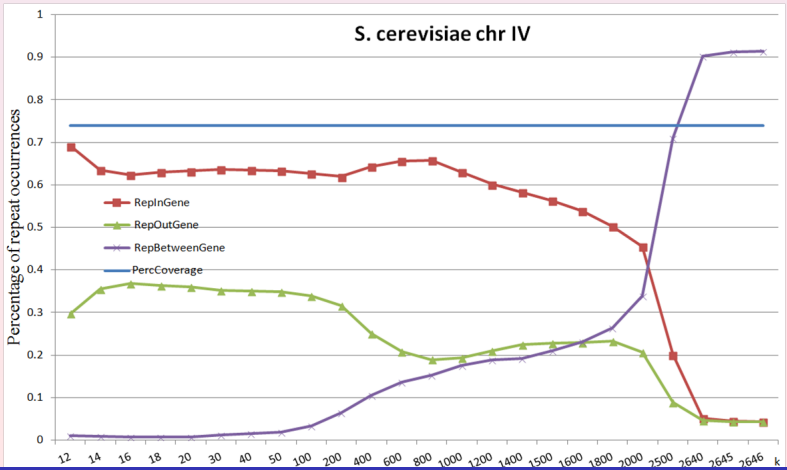
Repeat localization (inside, outside, across genes)

Percentage of *k*-repeat occurrences, with the length *k*









Repeat sharing gene networks

Def. A k-parametrized, labeled graph $G_k = (V_k, E_k)$, where:

- V_k are nodes associated to the genes of the organism,

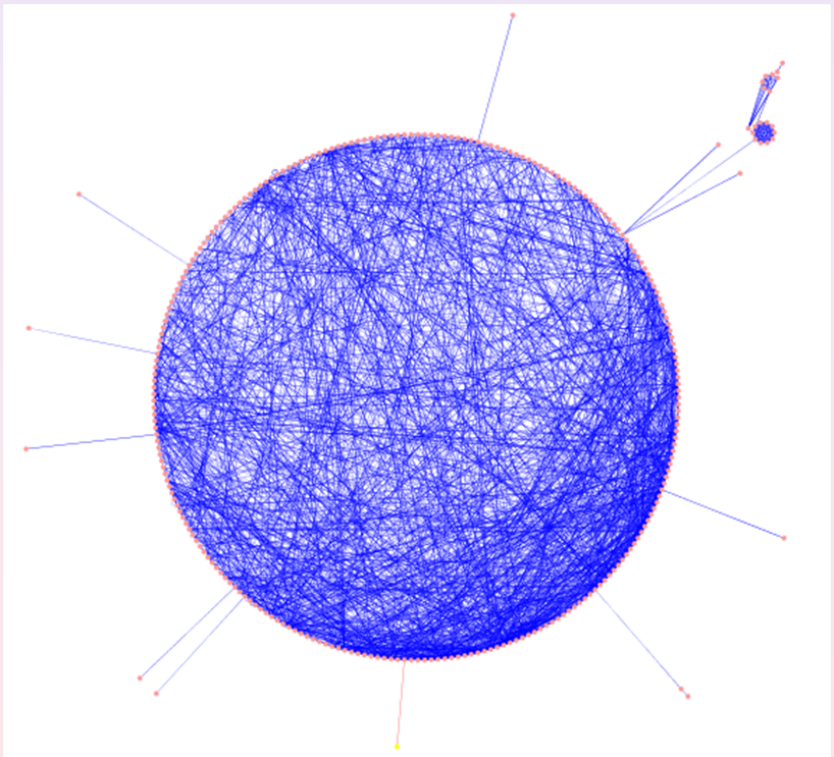
- two genes are connected (in E_k) if they share at least one k-repeat. The set of shared k-repeats is the edge label.

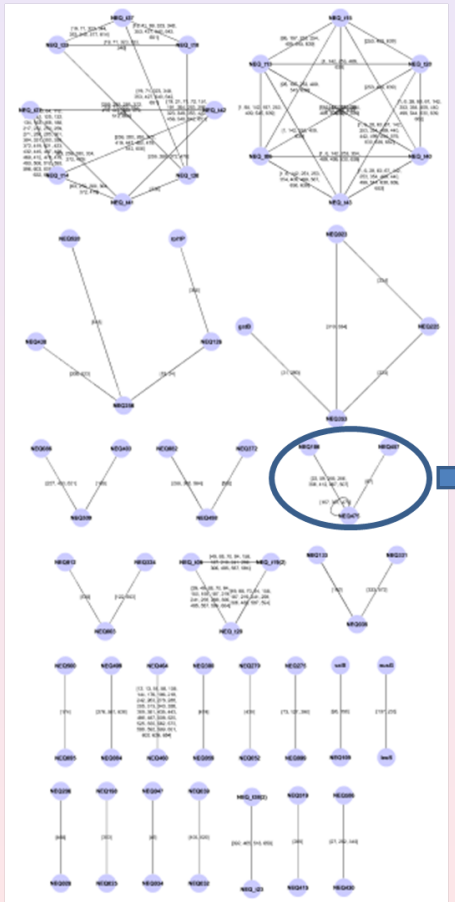
By sketching the k value, nodes and edges in G_k decrease (as isolated nodes are deleted), until the network disappears[6]. A break-point ($k = 18$) where the networks pass from having a big connected component to being a set of "vanishing" clusters.

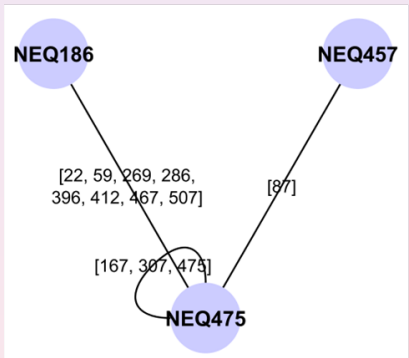
[6] In *N. equitans*, G_{54} is empty, while MR = 139

Genetic repeats within a specific genome

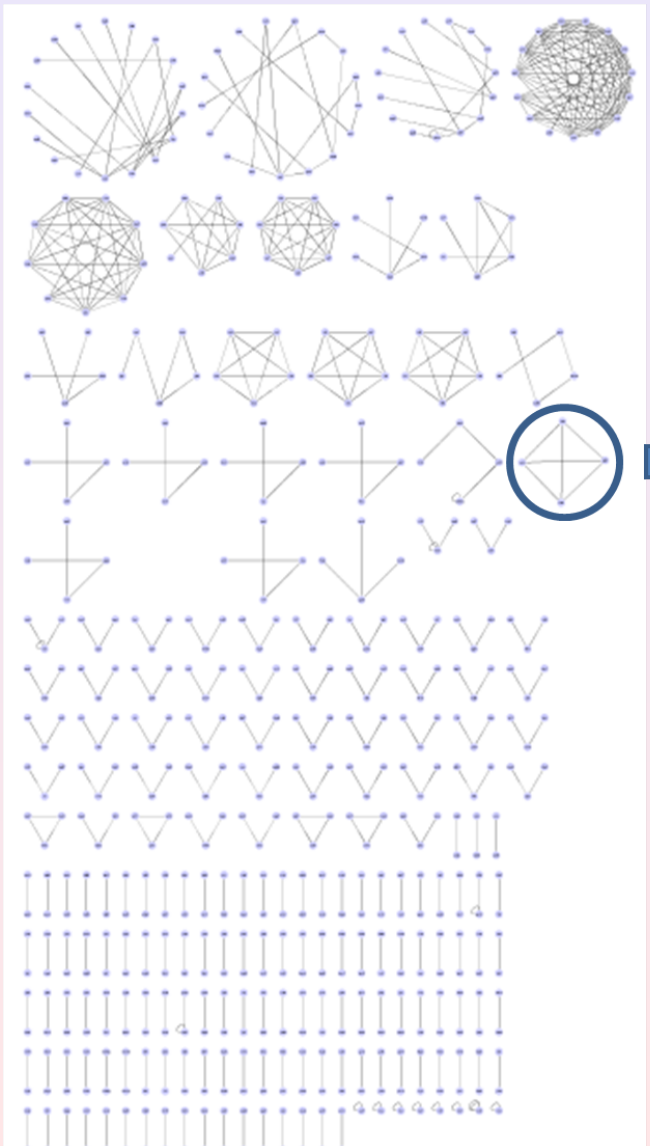
N. equitans 's gene networks G_{14} - G_{18}

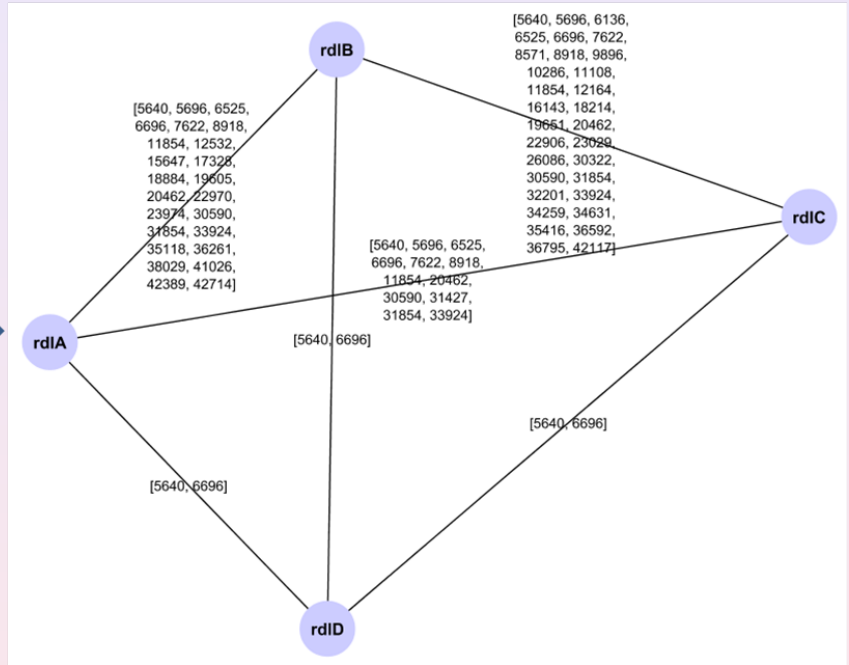






Another example: *E. coli* 's gene network G_{18}



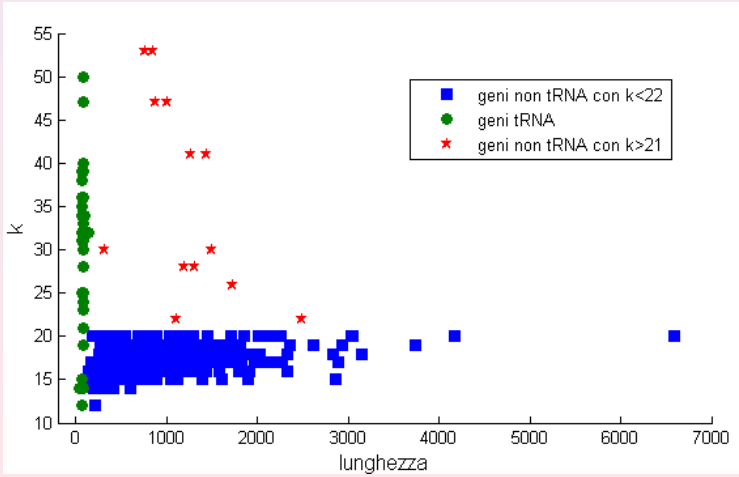


Investigation on repeat sharing gene networks

Max degree, max labels weight, number of cliques (its variation with k) - results on *N. equitans, E. coli, S. cerevisiae*

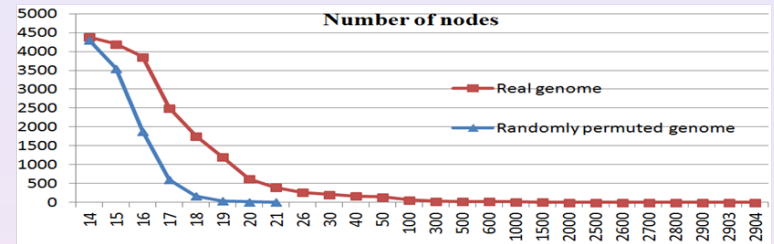
Highly connected genes, for k long enough, have similar functionality and turned out involved in important biological pathways, such as DNA repair and replication.[7]

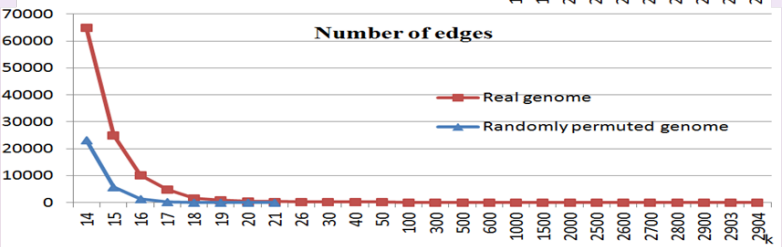
Gene length compared with repeat length k, to measure edge significance (genes shorter than 100 encode for tRNA)

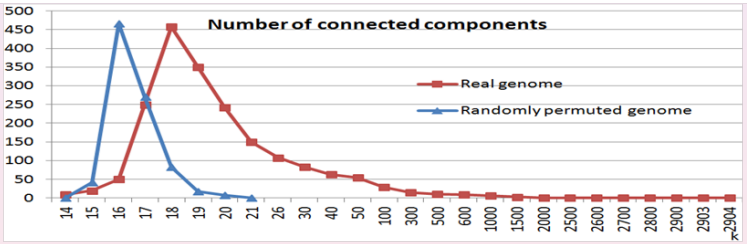


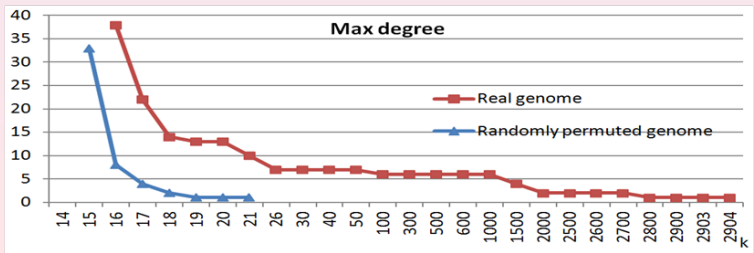
[7] A. Castellini et al., Natural Computing 2015

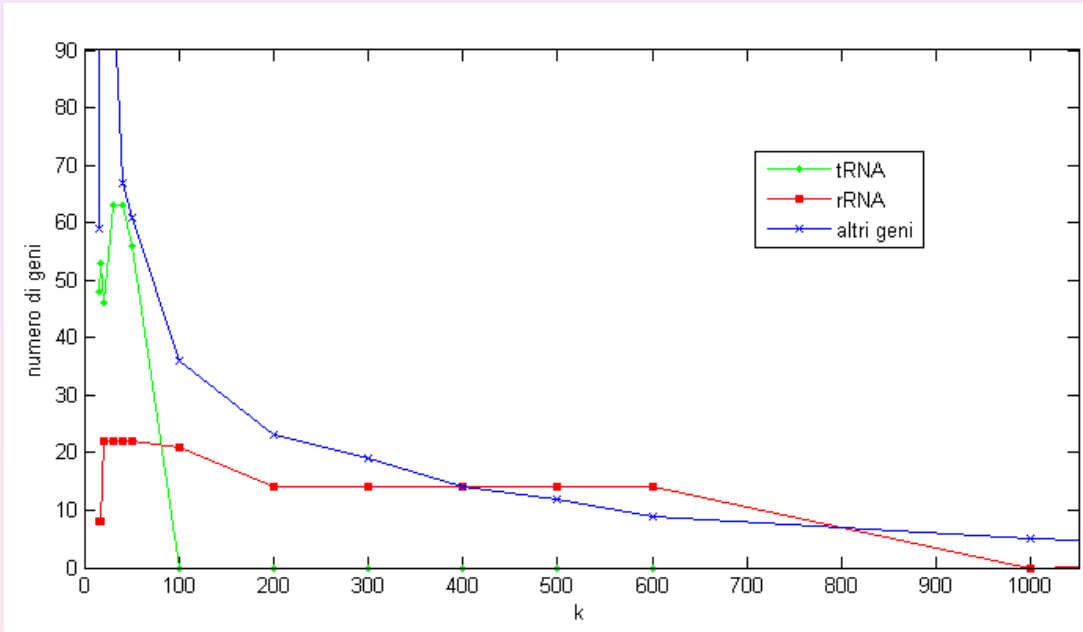
Basic E.coli network analysis





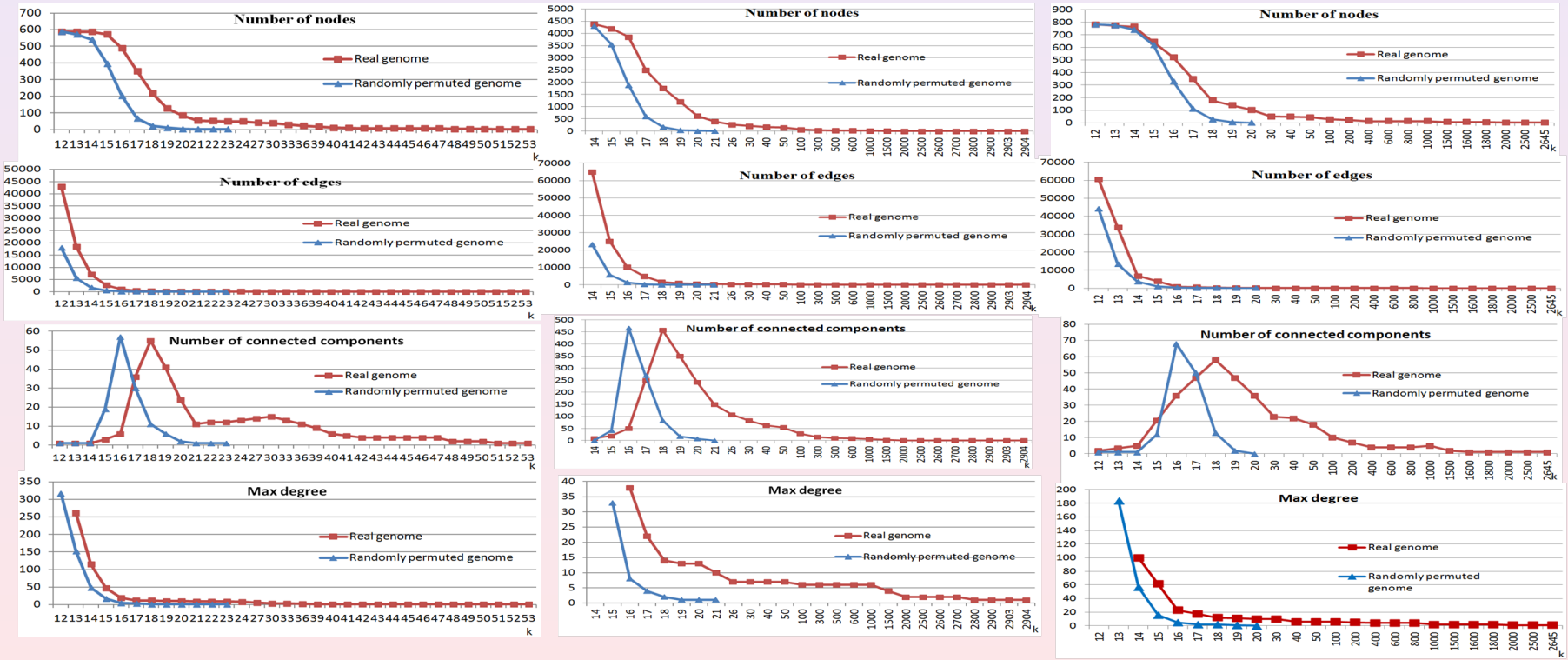






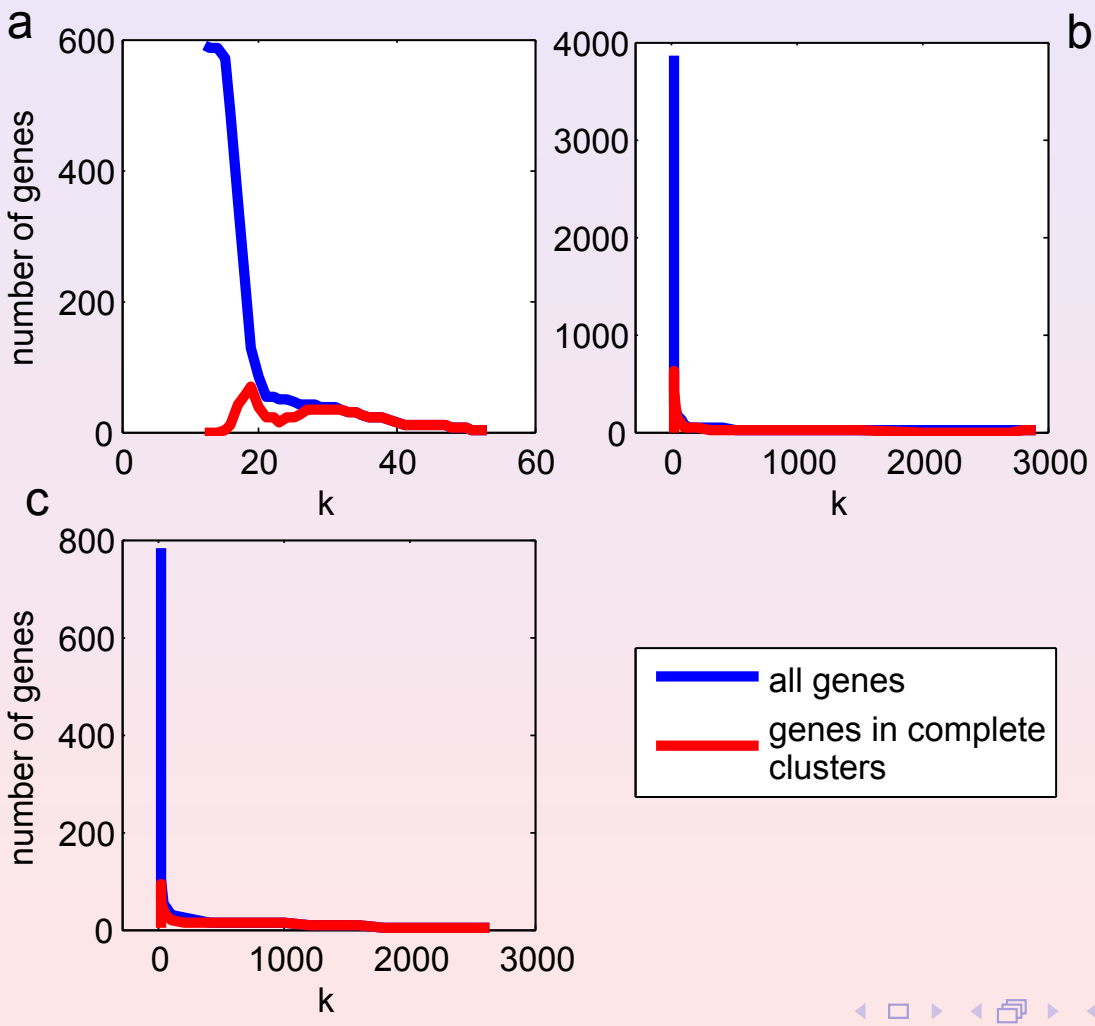
Similar network patterns

Network curves observed for *N. equitans, E. coli, S. cerevisiae* do not depend on genome length or gene coverage.

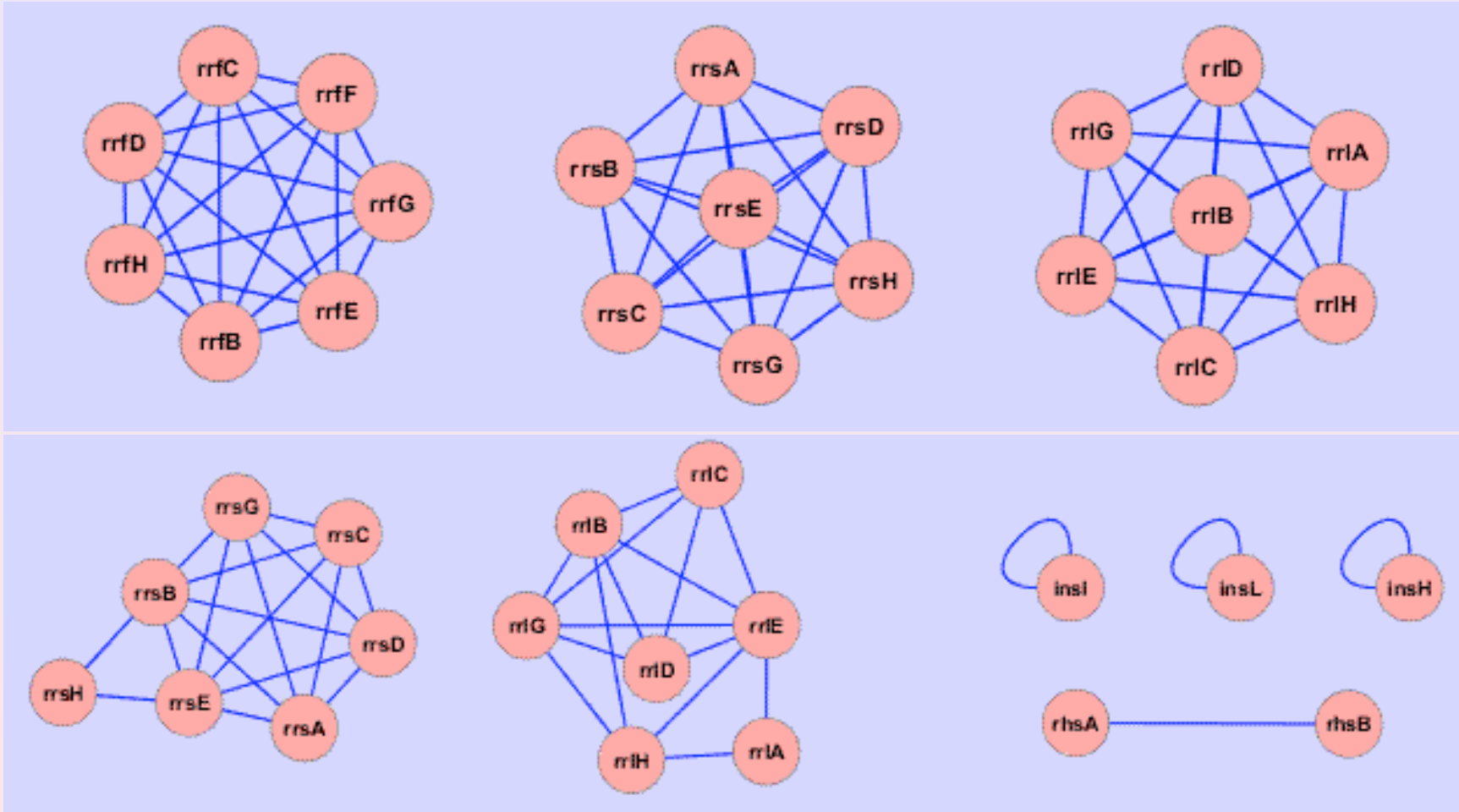


Complete clusters (N. equitans, E. coli, S. cerevisiae)

For longer repeats, gene networks tend to aggregate in cliques.



Examples of cliques (in E. coli)



Repeat clique analysis

According to the range values of k, we may deduce:

$k = 1, \ldots, 12$: repeats are completely random

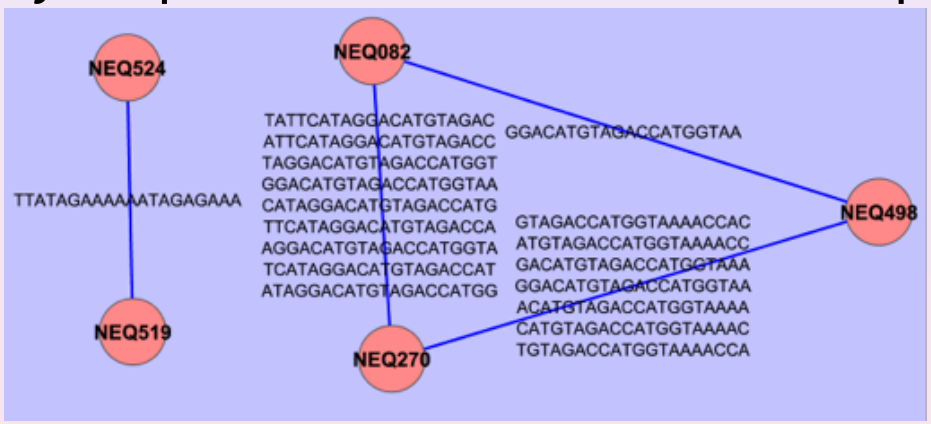
$k = 13, \ldots, 20$: some repeats are present only in couples of genes, only few have a biological role (14-15: first repeats present even in non-tRNA genes)

$k > 21$: Repeats (all, for $k > 40$) have a biological role, they belong to paralogous genes and have a same reading frame for protein translation.

Clique analysis - a few observations

◇ Relatively few genes are involved in cliques, and the number of cliques varies similarly in the three organisms.

◇ In *every* cliques, there is at least one repeat *common to all* genes.



◇ Such a repeat encodes for a protein/enzyme core, and has the same reading frame in *all* the genes where it occurs.

◇ Almost all cliques are composed by paralogous genes.

Minimal k-RDD

Let $\alpha \in D_k(G)$ such that:

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

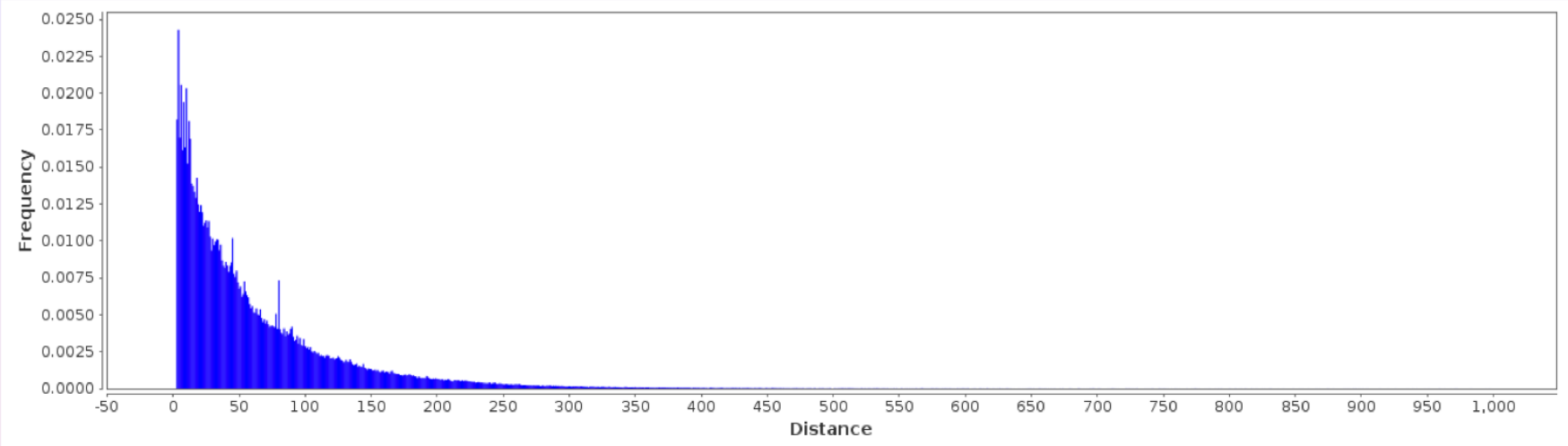
where no α occurs between two consecutive α

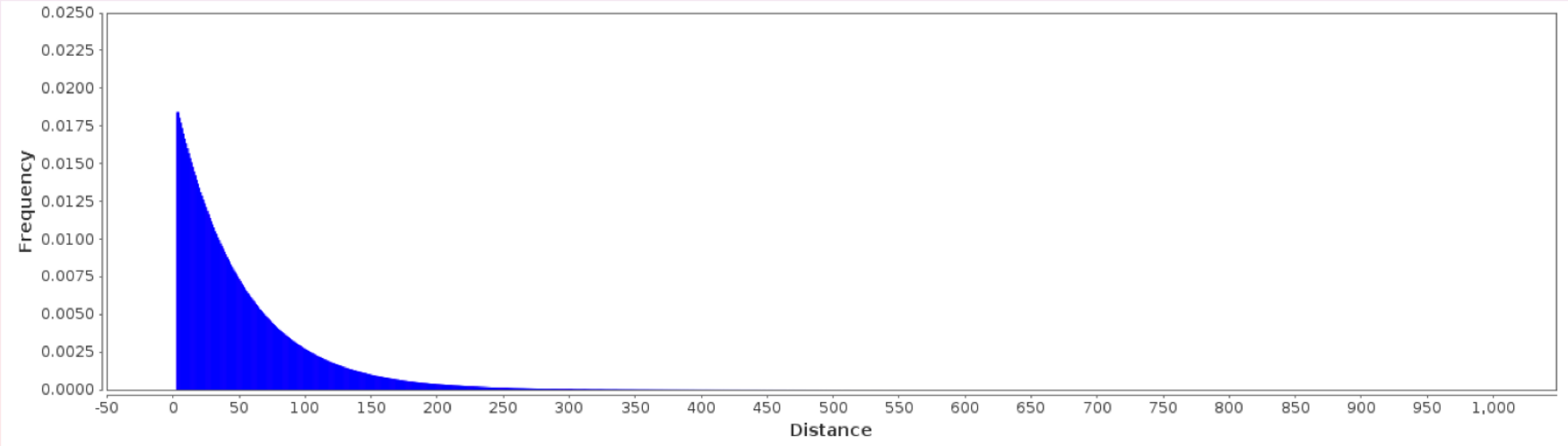
K-RDD(α, G) = d_1 -> n_1 , d_2 -> n_2 , d_3 -> n_3 , - - -

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

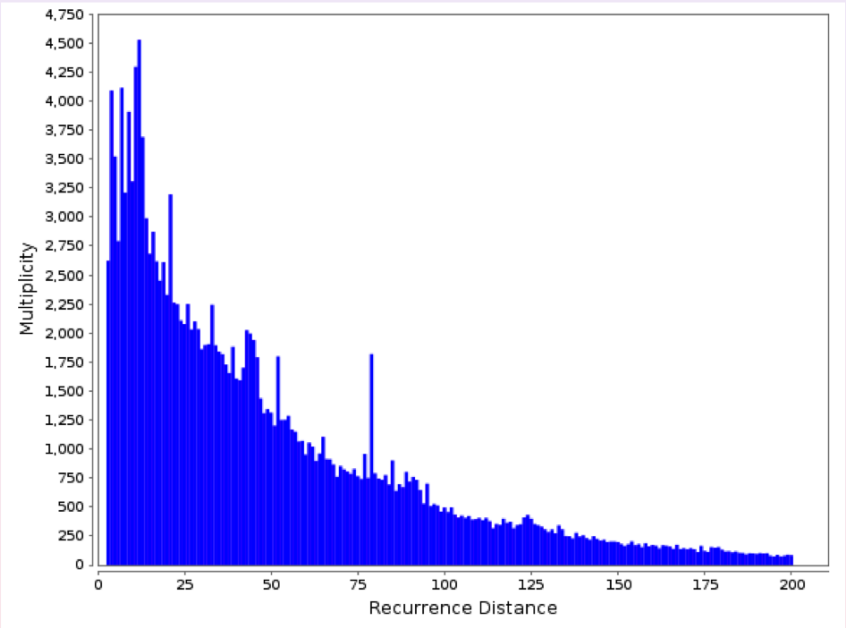
Slide courtesy of prof. Vincenzo Manca, Univ. Verona, IT

ATG in human chr 22 vs extimated Exp Distr

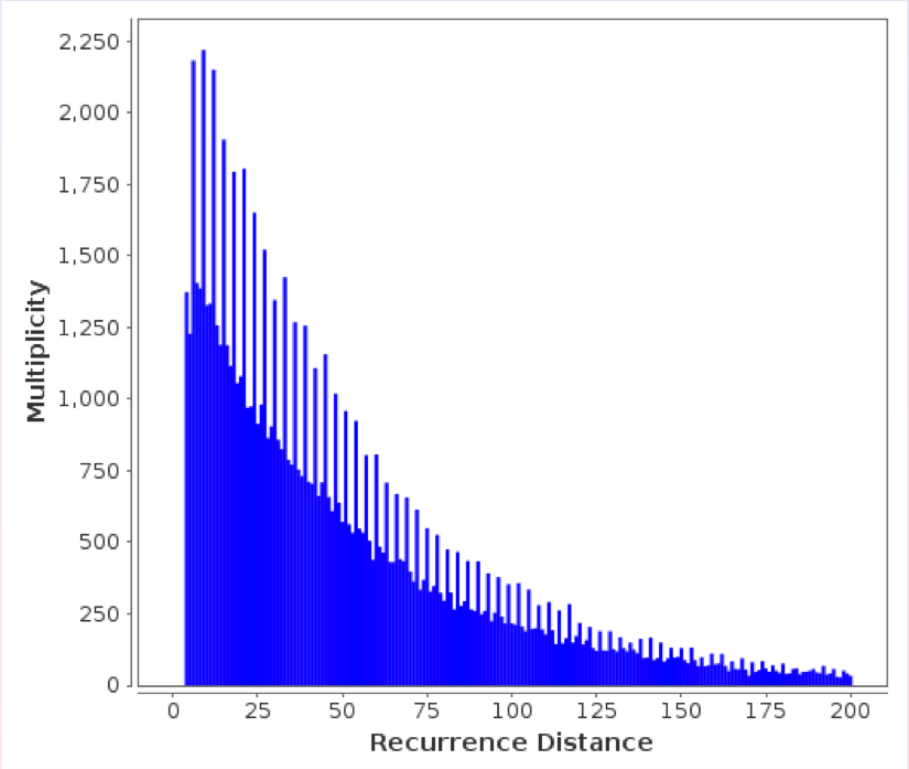


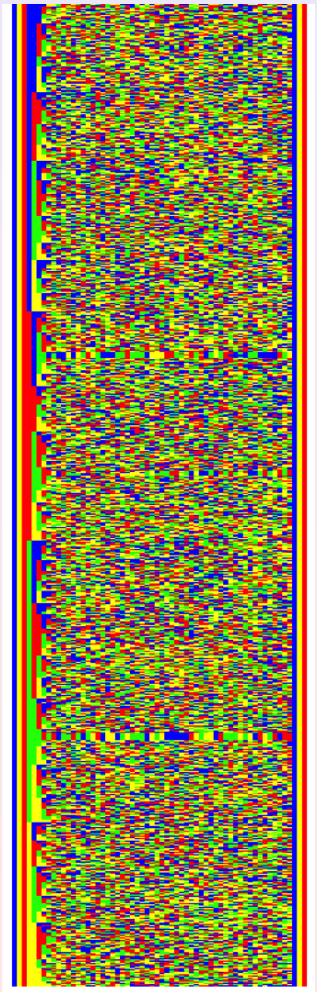


ATG in human chr 22 and sequences at distance 81

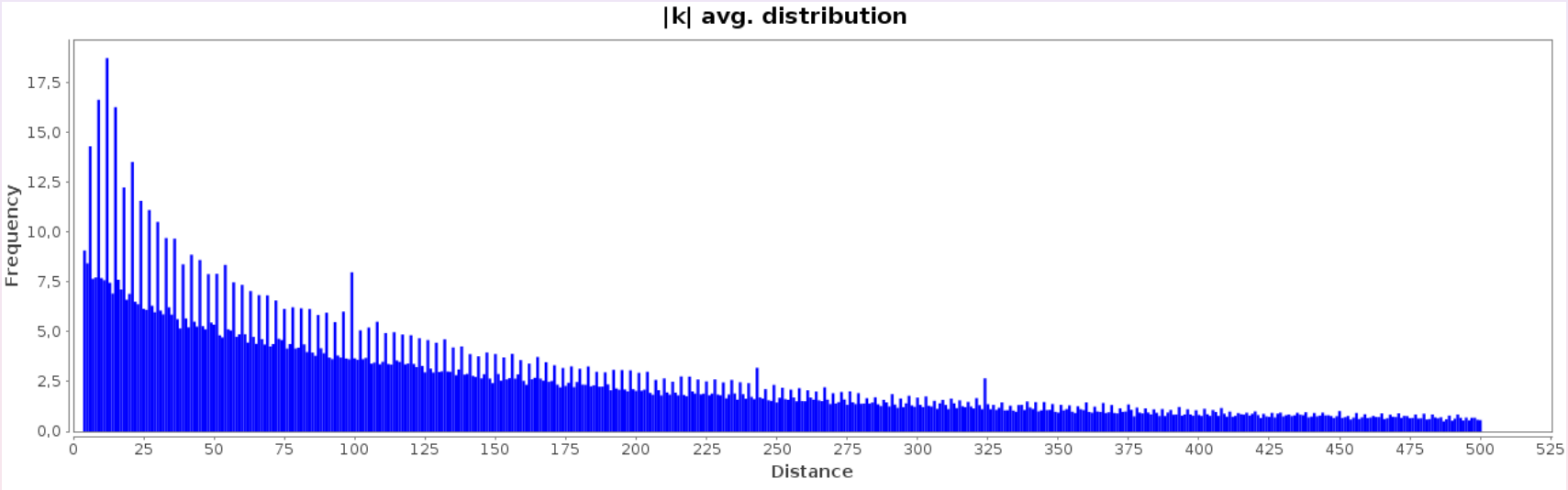


ATC in E. coli: peaks with non-repetitive elements



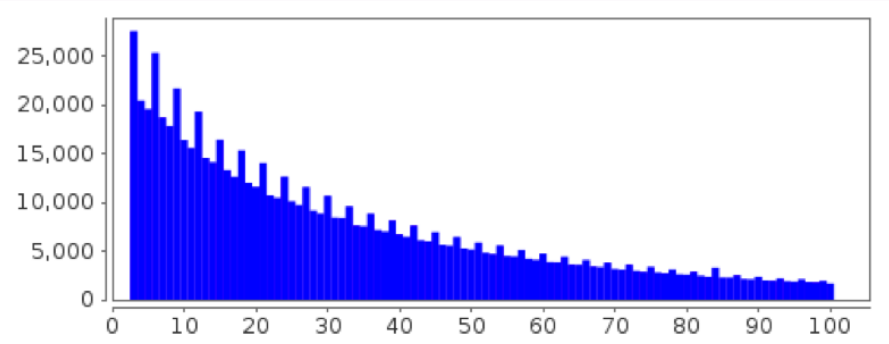


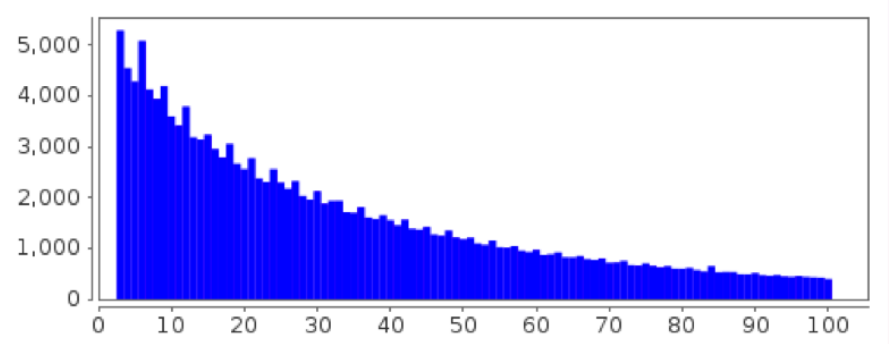
Average RDD for k=4, Emiliana huxley virus86



RDD in exonic regions, k=3

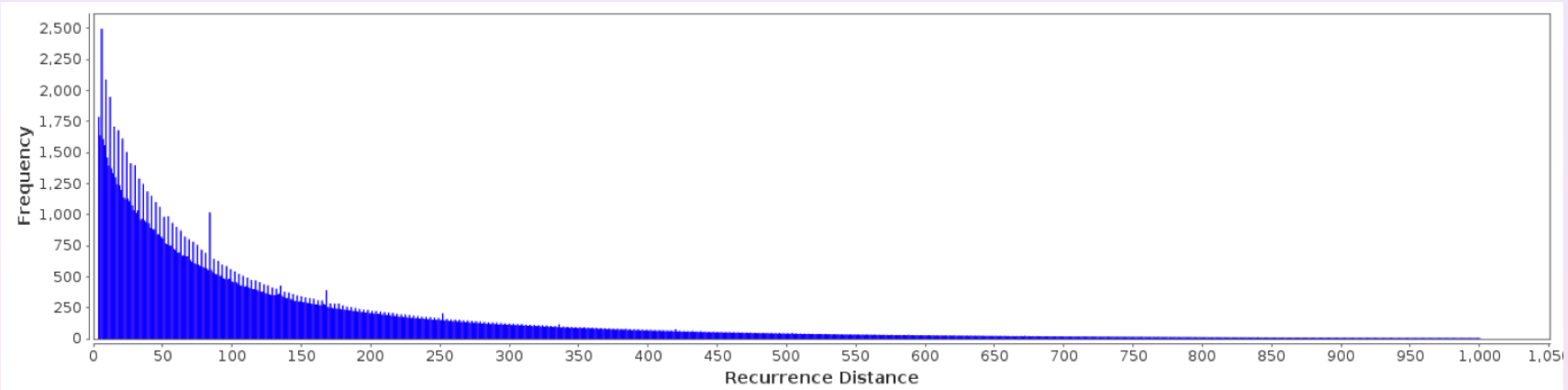
Peaks at distance 3 disappear in transcripts (introns + exons) - they are a *codonic language*



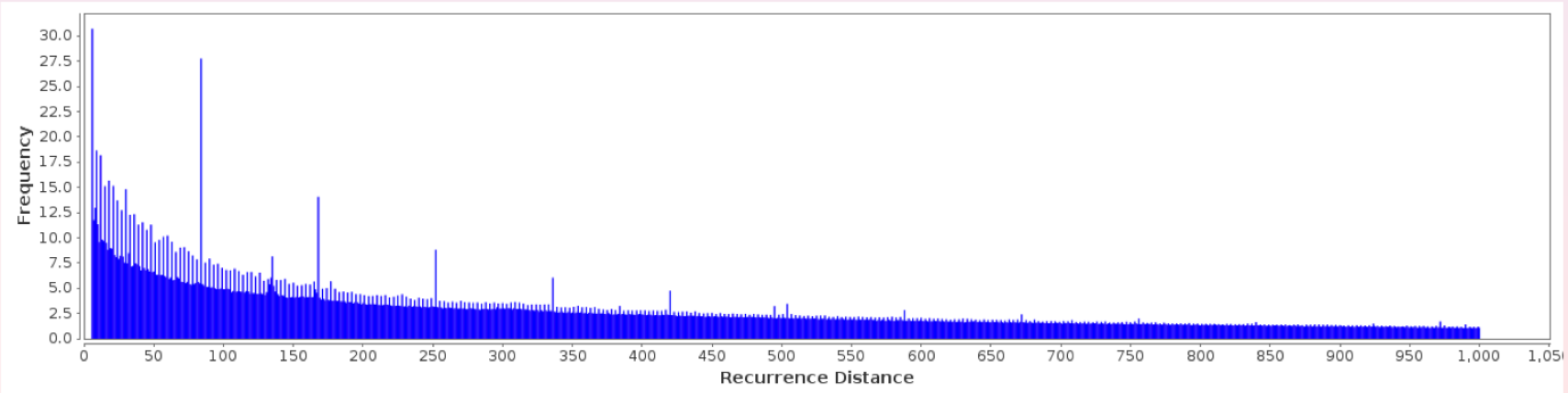


Repetitive elements in ex. regions distance multiple 84

k=4



k=6



Importance of grouped occurrences

Recurrence is a peculiar feature of words

Occurrences of words semantically relevant for a text tend to be grouped around some points

RDD(α) is a measure to visualize the non-randomicity of the word α distribution

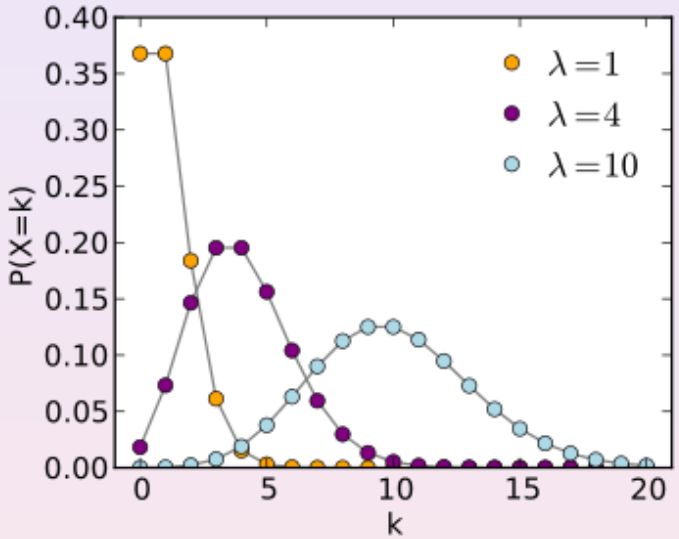
Segment multiplicity and recurrent words

A **Bernoullian genome** is generated by means of casual extraction (with insertion after extraction) from an urn containing balls of four colours.

Segment multiplicity. Let us consider the genome as the concatenation of equal length segments. Given a word α, this distribution assigns to each *n* the number of segments where α occurs *n* times.

Bernoullian (random) genomes

In a Bernoullian genome, such distribution (normalized, with the total num of segments) of word frequencies follows a Poisson prob distribution (of a certain variance λ): $Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$



The **waiting time** follows the Poisson law: the distance between two consecutive occurrences of a given α is an exponential of parameter h, $f(x) = he^{-hx}, \quad x \geq 0$, for some h.

Tandem repeats investigation

RDD (recurrence distance distribution): given a sequence α, to each *n* it is assigned the number of times that α occurs at distance *n* from its previous occurrence. Once normalized, the distribution above may be compared with (as corresponds to) the **waiting time** associated to a Poisson process.

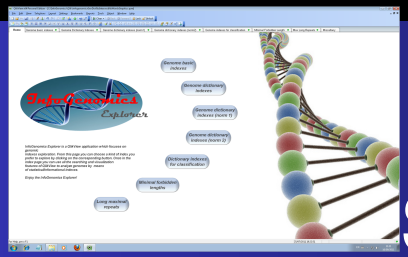
If *x* is the distance between two occurrences of a given string, $P(x) = f(x, h)$ is the number of times two occurrences appear at distance *x*. The probability of occurring at distance *x* is the ratio between the number of times it recurs at distance *x* and the multiplicity of α in G. A curve average may be then computed, over all sequences having the same length of α.

Information Correlation and RDD in Genomes

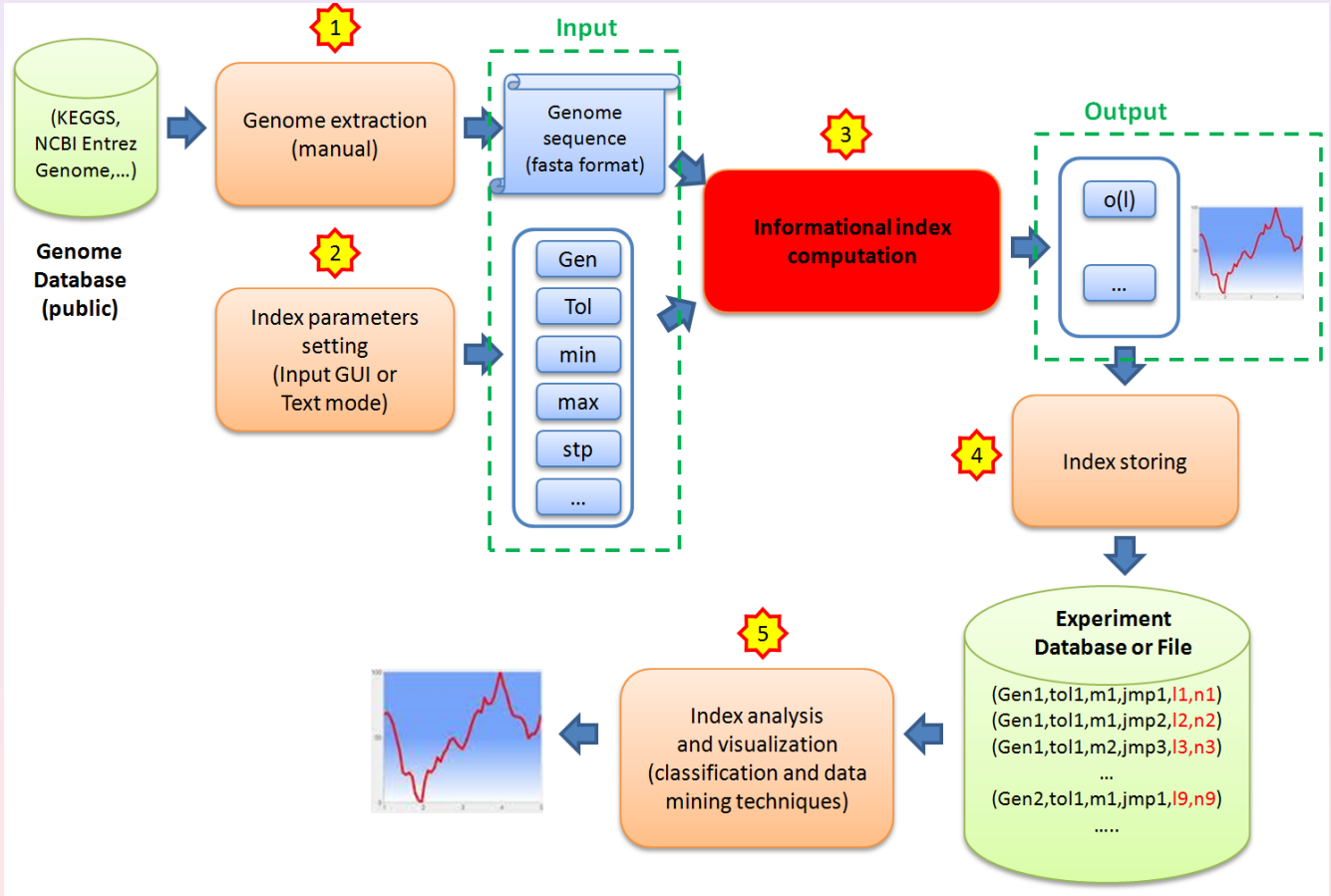
- Trifonof et al. : DNA correlation periodicities, 1980
- Shepherd : DNA periodicities in coding regions, 1981
- Eigen et al. : periodicity in Transfer-RNA, 1981
- Fickett :1982 non min. RDD periodicity in coding regions, 1982
- Li : Mutual information in DNA Strings, 1990
- Herzel et al. : Measuring 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), 2009-20013
- Computational Chemistry 2014

****

Slide courtesy of prof. Vincenzo Manca, Univ. Verona, IT

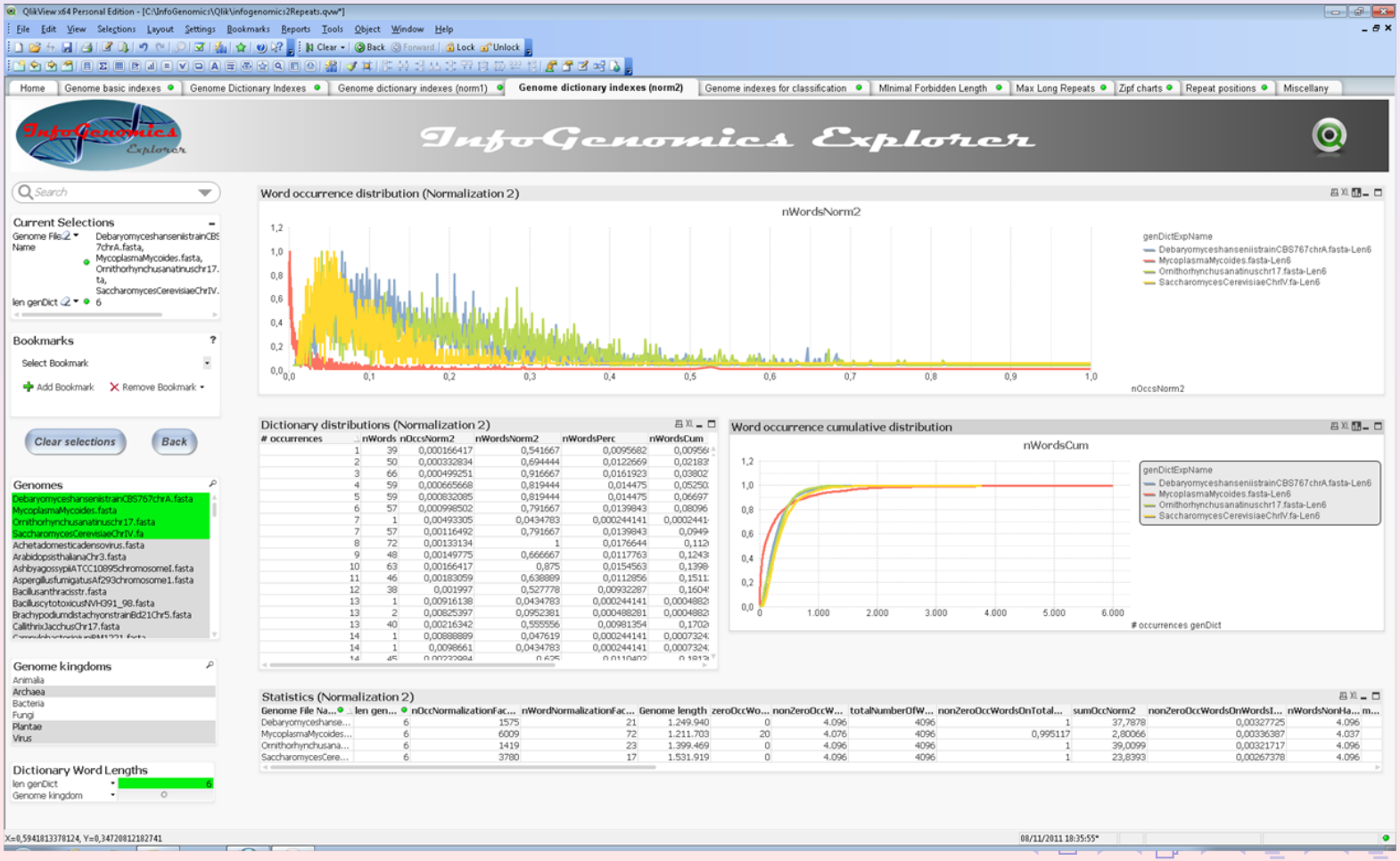


Software for massive computations



Infogenomics Explorer

Visualization and exploration of informational indexes by means of a Qlik®View application called InfoGenomics.



IGtools

Interactive graphical interfaces and CLI (batch analyses).
Advanced data structures and algorithms, for real genomes.

Conclusions

A method

- to represent and compare genomes (genomic profiles, dictionary intersections)
- to describe a genome by numerical information: statistics (amount, multiplicity, localization of repeats), informational index vectors
- to study gene networks, in general and along with a complete clusters analysis
- to find tandem repeats, and “good” genomic dictionaries.

Biological and computational annotations

Biological annotations of DNA elements, by the functional role they play in regulatory mechanisms (biological semantics).

Numerical annotation of genomic words, by the position in the genome, the total number of occurrences, the occurrences lying inside, outside, or between encoding regions, its CpG content, and more sophisticated informational indexes of text analysis.[8]

[8] G. Franco, *Perspectives in computational genomics*, Discrete and Topological Models in Molecular Biology, 2014.

Open problems

◇ How information is structured within genomes? Finding a "good" genomic dictionary (cardinality, word length k, sequence and average positional "coverage")

◇ Regularity properties, informational indexes characterizing (classes of) genomes: e.g., MRL, MFL

◇ Genome discrimination methods, namely for phylogenetic or medical purposes.

Applications beyond genomics

Phylogenetic methods

Cyber virus code analysis (by breaking down malware signatures)

Computational linguistics

The dictionary parameter k has to be appropriately chosen for each model (a range of k may be possibly required)

What's next?

Next Tue, Nov the 7th – two appointments:

- 9:00am, sala verde, lectio magistralis prof. Seth Lloyd (MIT): *Prospects in Quantum Machine Learning*;
- 4:30pm, room I, lecture on the use of IGtools.