An INFOrmational GENOMICS approach

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

Giuditta Franco

Department of Computer Science, University of Verona, Italy

Verona, 9-11 March 2016

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).

Traditional **alignment-based methods** are applied for multiple sequence analysis and comparison, with phylogenetic and pharmacogenomics applications.

Recent **alignment-free methods**[1] are based on empirical studies of frequencies of DNA k-mers (genome factors of length k), out of whole genomes/proteomes, where *information theory* is applied.

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

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**[2], 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.

[2] A. Castellini et al., Pattern Recognition in Bioinformatics, PRIB2011.

Informational genomics

Entropic measures (in both informational and biophysical terms) and information-theoretic methods have been widely applied to study genome organization[3].

In statistical mechanics, thermodynamic entropy is given by $H(X) = -k_B \sum_i n_i \ln n_i$, k_B Boltzmann constant and n_i number of microstates associated to the macrostate X. A measure of the amount of information needed to define the detailed microscopic state of the system. Micr vs macro info.

Average uncertainty/ignorance: $H(X) = \sum_i^n p(a_i) log \frac{1}{p(a_i)}$, $a_i \in A$. Maximum $\log_2 n$ reached when symbols are equally likely $\frac{1}{n}$, and minimum zero when probability is (0,1). $L(C) = \sum_i p(a_i) l(a_i) \geq H(X)$ (UD code C, I Shannon theorem)

[3] Chang et al. Shannon information in complete genomes. J. Bioinform and Comput Biol. 3:587-608, 2005. Chun-Ting et al. Segmentation algorithm for DNA sequences. Physical Rev. E. 2005; 72:041917.

Informational genomics: main ingredients

Genomes are information source, whose words are random variable values, codes, data encodings, and a probability distribution is given by frequences [Sims et al. PNAS, 2008 (FFP,12-mers)). Robins et al. Journal of Bacteriology 2005 (fingerprints)].

Infogenomics employs methods from FLT and information theory to define genomic indexes, sequence similarity measures, gene networks [A. Castellini et al. BMC Genomics 2012.].

Statistical encoding of data is initially given by a fixed length (then prefix, uniquely decodable, instantaneous) code for a given word length k. Frequences are computed on k-mers as normalized multiplicities. Divergence of Kulback-Leibler is computed to measure the distance between real and *random* genomes [Boholin et al. BMC genomics 2012].

Motivation, open problems

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

◇ Regularity properties, informational indexes characterizing (classes of) genomes

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

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

Lectures outline

- A brief description of (eukaryotic) genome structure and functioning (from ENCODE project)
- A dictionary based alignment-free approach:
  - Multiplicity-comultiplicity **genomic profiles**, and UCE (ultraconserved elements) in **dictionary intersections**
  - Systematic **analysis of repeats** variation in number, length, multiplicity, and localization (inside, outside genes)
  - Cluster analysis on repeat sharing **gene networks**, and genomic Recurrent Distance Distribution (RDD)
- Tools and references

ENCODE project: transcript as heredity unit

Revolutionary Human Genome Project (concluded in 2001)[4]: less than 2% is encoding, 21.000 genes (average 3000, dystrophin 2.4 million bases), 20.000 RNA genes.
Low-cost, rapid sequencing technologies (after pilot project, 2007).

ENCODE (Encyclopedia of DNA Elements), focus on 98% of junk/dark DNA, including latent viruses (8%) and 18000 pseudogenes (keeping a regulatory effect). 440 scientists, 32 labs, 40 papers, 300 million dollars investment by NHGRI.
80% is covered by **regulatory elements** necessary to **promote, inhibit, silence** gene activity, more than half of it is transcribed in different RNA types for **synthesis, processing, transport, modification** and translation activities.

[4] NCBI, UCSC, EMBL-EBI websites.

ENCODE project: transcript as heredity unit

Revolutionary Human Genome Project (concluded in 2001)[4]: less than 2% is encoding, 21.000 genes (average 3000, dystrophin 2.4 million bases), 20.000 RNA genes.
Low-cost, rapid sequencing technologies (after pilot project, 2007).

ENCODE (Encyclopedia of DNA Elements), focus on 98% of junk/dark DNA, including latent viruses (8%) and 18000 pseudogenes (keeping a regulatory effect). 440 scientists, 32 labs, 40 papers, 300 million dollars investment by NHGRI.
80% is covered by **regulatory elements** necessary to **promote, inhibit, silence** gene activity, more than half of it is transcribed in different RNA types for **synthesis, processing, transport, modification** and translation activities.

[4]NCBI, UCSC, EMBL-EBI websites.

Recent results from ENCODE

1640 genome-wide public datasets[5], for different types of cell, with a complete catalogue of:

- annotated human transcripts, identifying different types of RNAs: mRNA, (8800 short and 9600 at least 200b long) ncRNA, sRNA, rRNA, 32 (73-93b long) tRNAs, miRNA;
- functional elements, like promoters[6], millions of genetic switches, transcription factors, protein binding regions[7], transcription start sites (TSS), transcriptional repressors[8]

[5] regions of transcription and transcription factor association, chromatin accessibility and histone modification, by DNase

[6] http://epd.vital-it.ch/

[7] Es. CCCTC-binding factor

[8] Es, 15000-40000 binding sites of CTCF, 11-zinc finger protein

Recent results from ENCODE

1640 genome-wide public datasets[5], for different types of cell, with a complete catalogue of:

- annotated human transcripts, identifying different types of RNAs: mRNA, (8800 short and 9600 at least 200b long) ncRNA, sRNA, rRNA, 32 (73-93b long) tRNAs, miRNA;
- functional elements, like promoters[6], millions of genetic switches, transcription factors, protein binding regions[7], transcription start sites (TSS), transcriptional repressors[8]

[5] regions of transcription and transcription factor association, chromatin accessibility and histone modification, by DNase

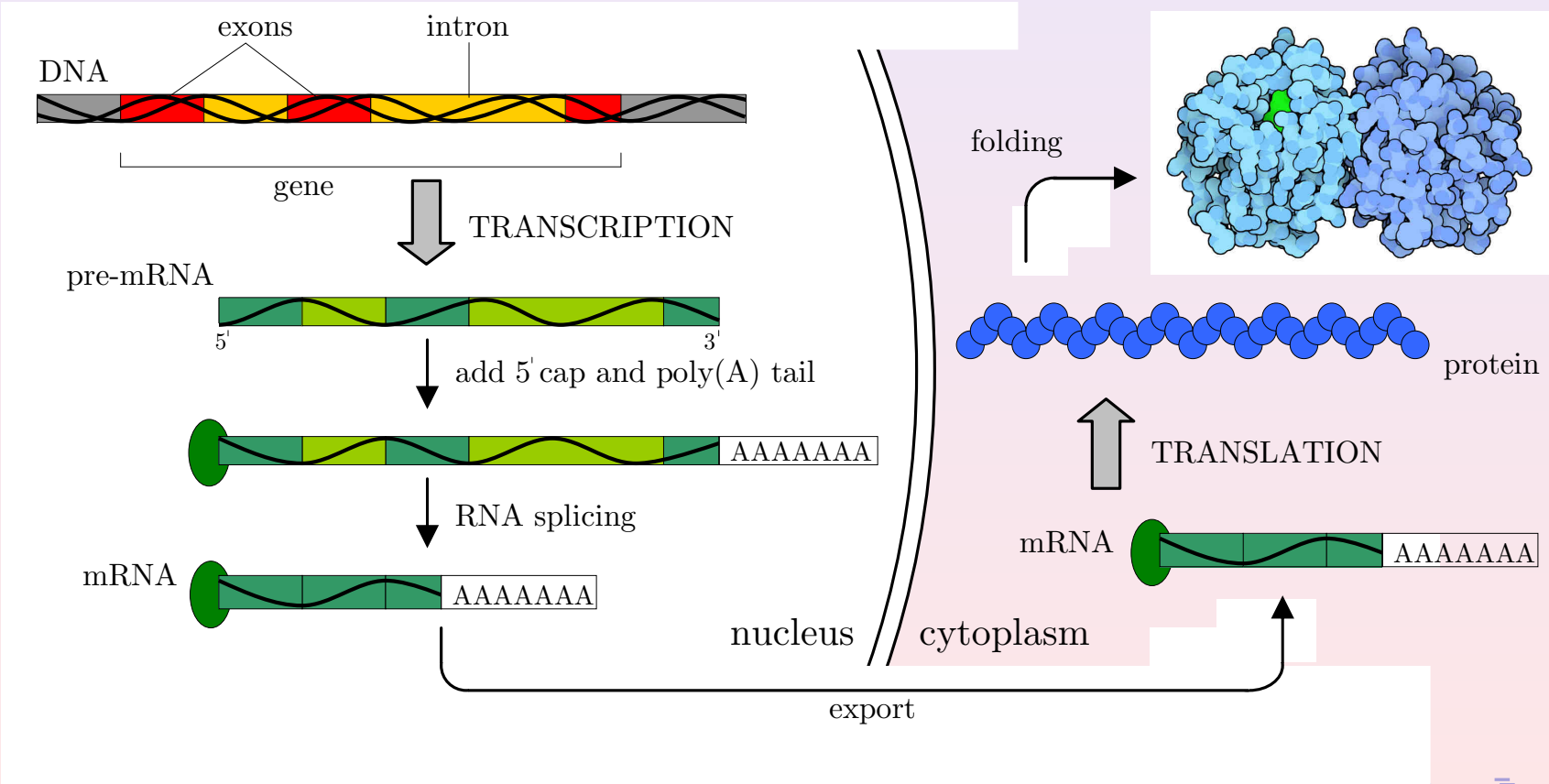
[6] `http://epd.vital-it.ch/`

[7] Es. CCCTC-binding factor

[8] Es, 15000-40000 binding sites of CTCF, 11-zinc finger protein

Central dogma (named by F. Crick)

Double helix wrapped around *histones*, forming *nucleosomes* (30nm), assembled in 46 *chromosomes*: 3m of 3 million bases in the cell nucleus (10^{-5}m). Secondary and tertiary structures.



Biological and computational annotations

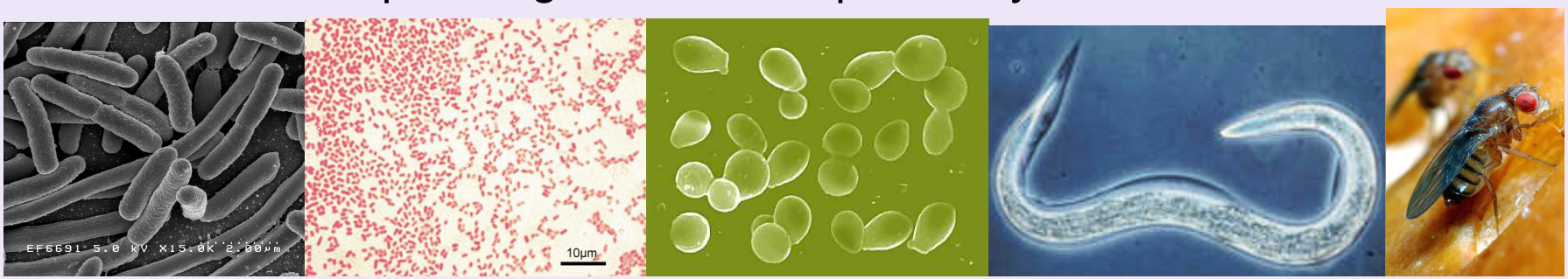
Biological annotations of DNA elements, by the functional role they play in regulatory mechanisms (biological semantics).

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.[9]

[9] G. Franco, *Perspectives in computational genomics*, Discrete and Topological Models in Molecular Biology, pp 3-22, 2014.

Our work focused on real genomes

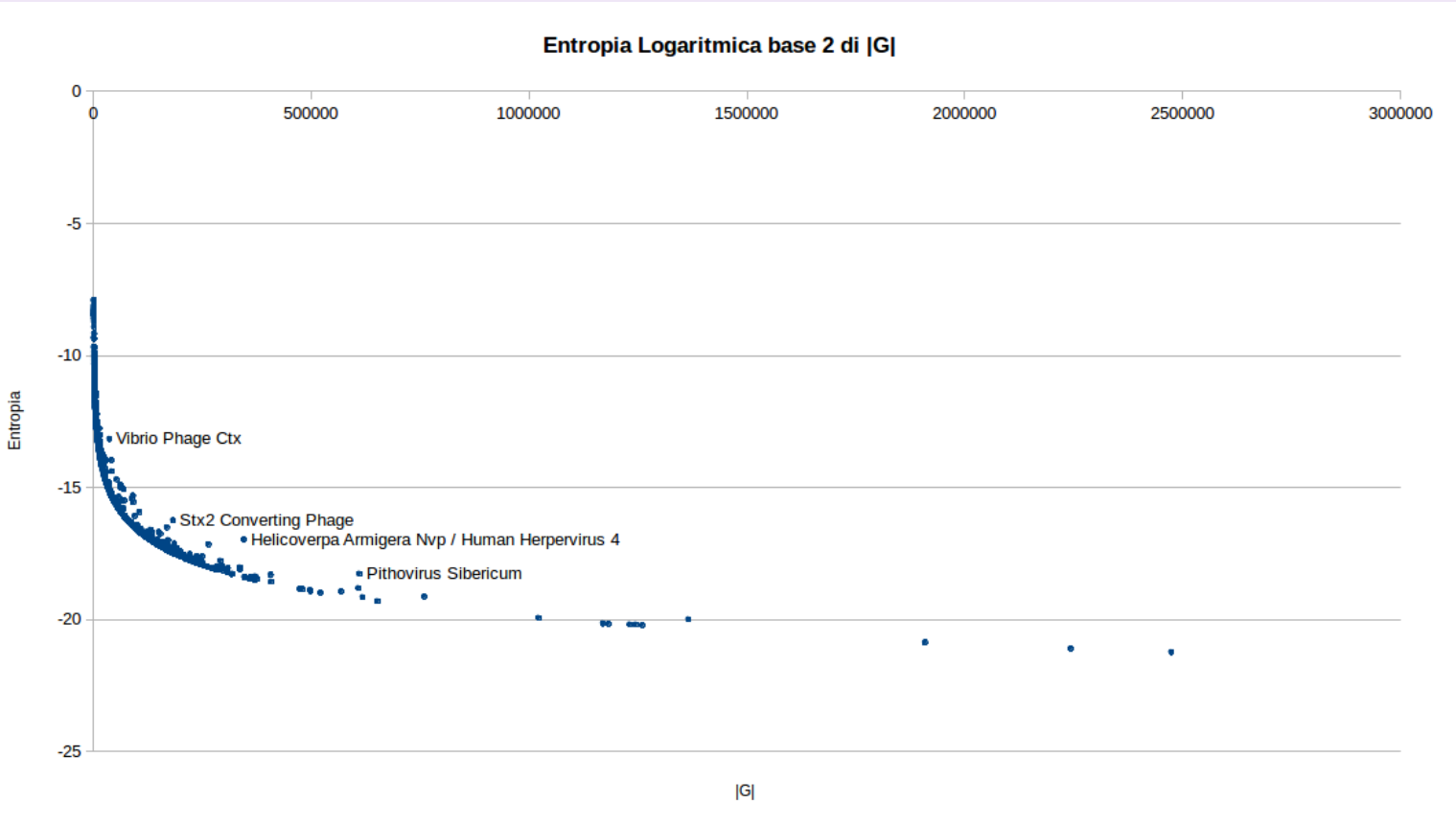
Dictionaries of 60 specific genomes, deeper analysis on 12



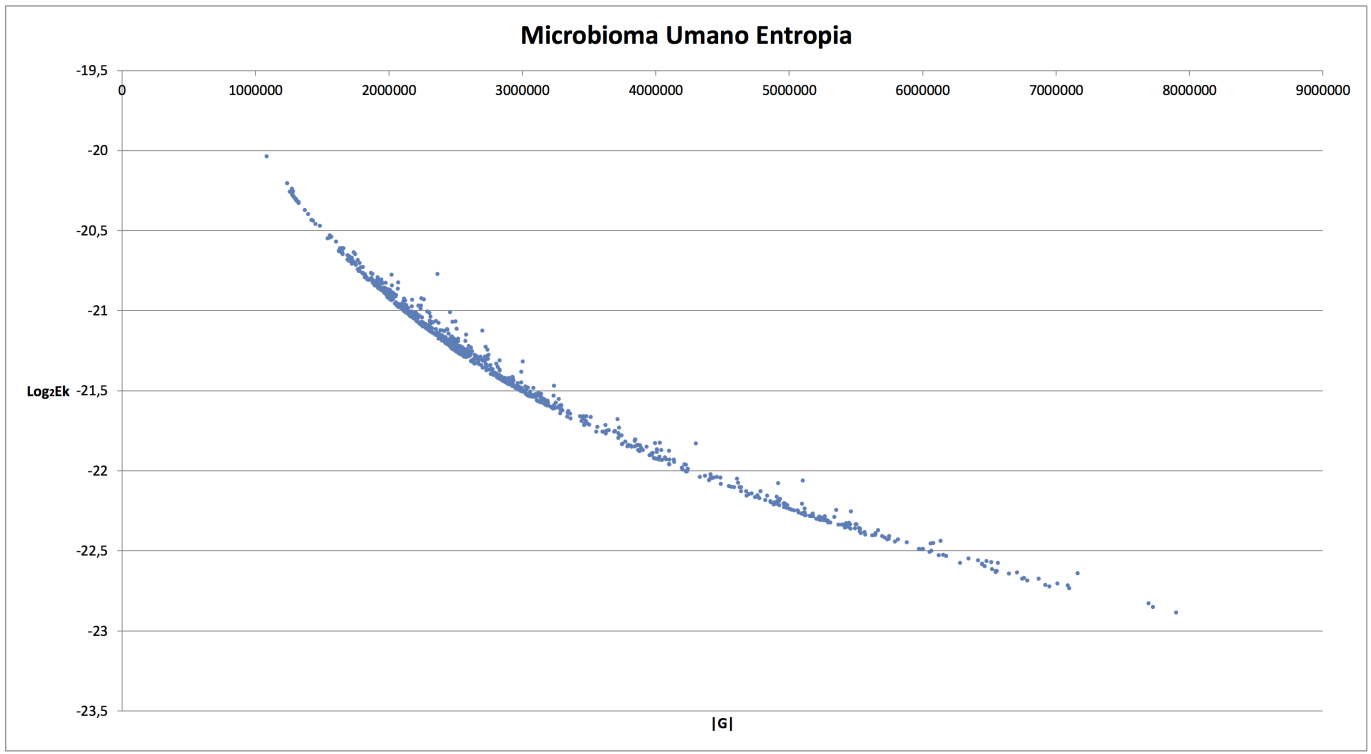
Organism Genome	Length	Genes	Type
Nanoarchaeum equitans	490,885	585	Minimal archaeum
Mycoplasma genitalium	580,076	476	Minimal bacterium
Mycoplasma mycoides	1,211,703	1,016	Venter's experiment bacterium
Haemophilus influenzae	1,830,138	1,717	First sequenced bacterium
∗*Escherichia coli*	4,639,675	4,685	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

Entropy for virus with k=log(|G)

Infogenomics analysis have been carried out also for human genes, chromosomes, 2109 bacteria, 917 from microbiome, 2221 archea, and 4383 virus, having *short* genomes.



Entropy for microbiome, with k=log(|G)



Which entropy?

Empirical entropy: given $T_k(G)$ (genomic multiset of words long k), measure the probability of each word as its frequency (multiplicity divided by |G|-k+1), and compute the Shannon entropy for the genomic dictionary D_k:

$$\sum_{\alpha \in D_k} p(\alpha) \log_2 \frac{1}{p(\alpha)}$$

Initial alphabet/code on which computing the entropy is a choice - the source is there, probability distribution on codes may completely changes.

Modeling empirical entropy: (open) problems

Fixed length codes (what about entropy on a variable length code?), with a length depending on the genome length (what about an entropy for fixed range of k for all organisms?).

Entropy has a maximum, minimum, flexction point - how to choose the appropriate dictionary in order to have an “informative entropy”?

For long k all k-mers occurs once, they are equally likely, then the entropy for long genomes is maximum (if k depends on the genome length n as in our case), and approximates $log_4 n$.

First 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

V. Brendel at al. Nucleic Acids Research 12, 1984.

First 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

V. Brendel at al. Nucleic Acids Research 12, 1984.

First 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

V. Brendel at al. Nucleic Acids Research 12, 1984.

First 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

V. Brendel at al. Nucleic Acids Research 12, 1984.

First 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.

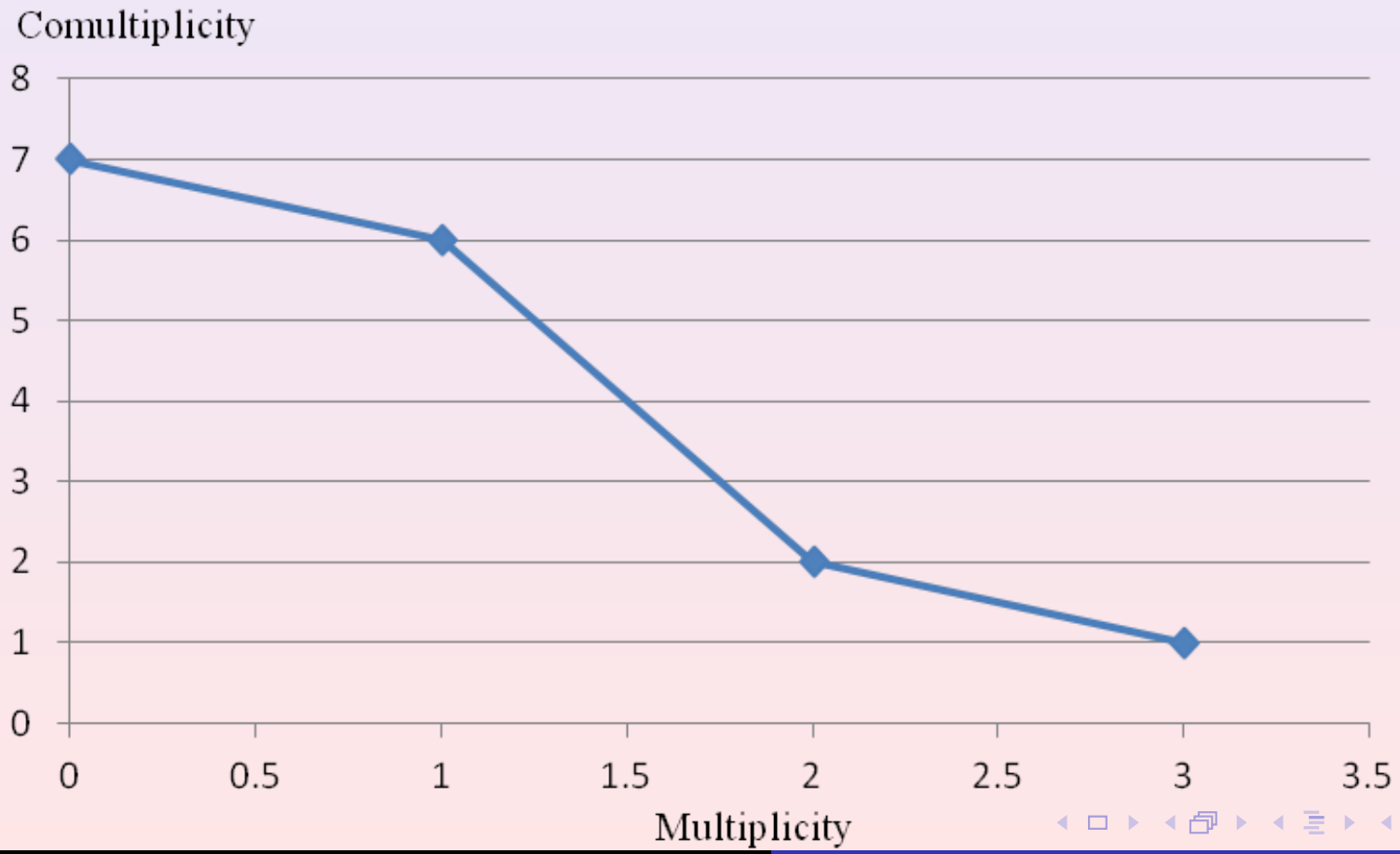
- $H_k(G)$: dictionary of **hapax k-words**, occurring once in G, $R_k(G)$ the set of **k-repeats**, occurring in G at least twice

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

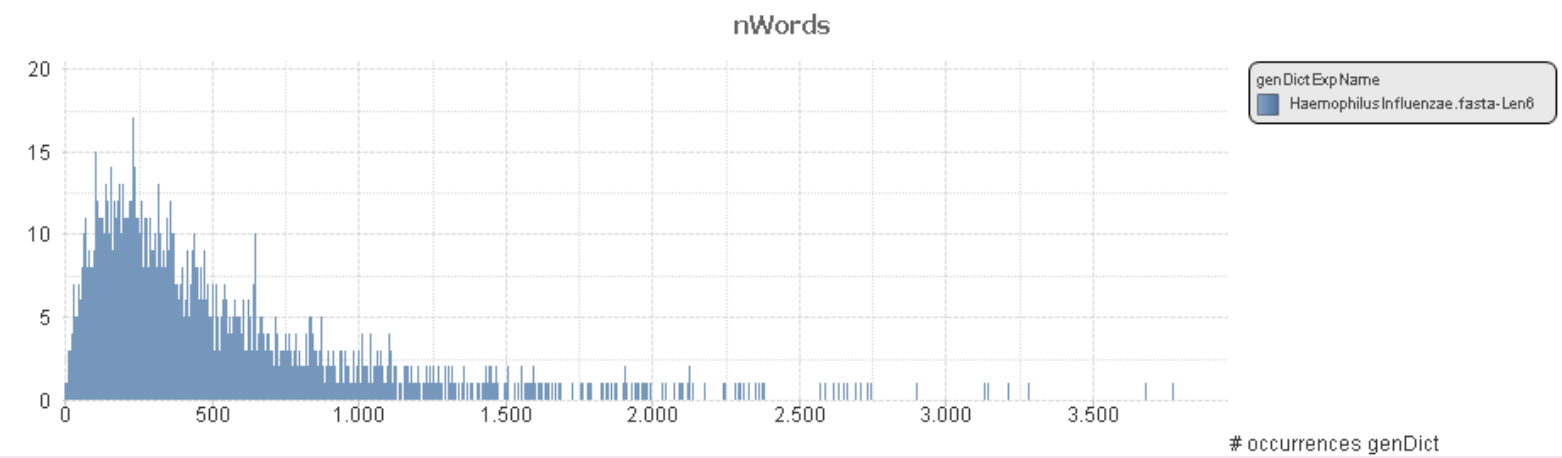
V. Brendel at al. Nucleic Acids Research 12, 1984.

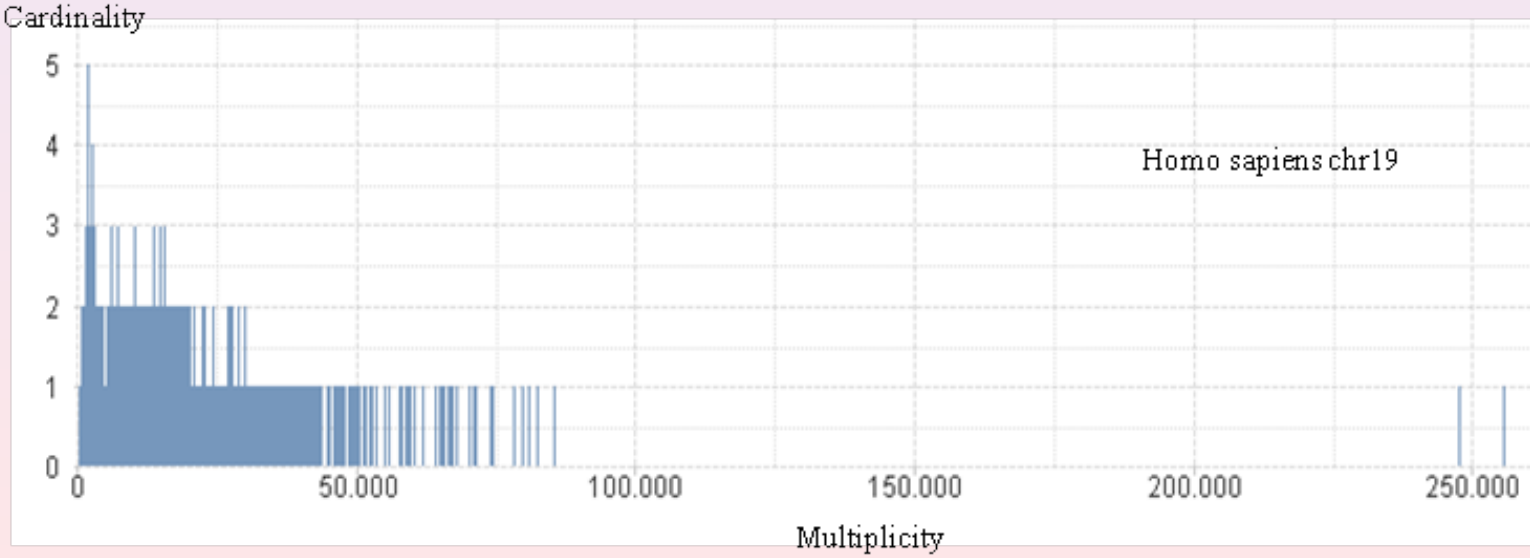
Multiplicity-comultiplicity diagrams

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



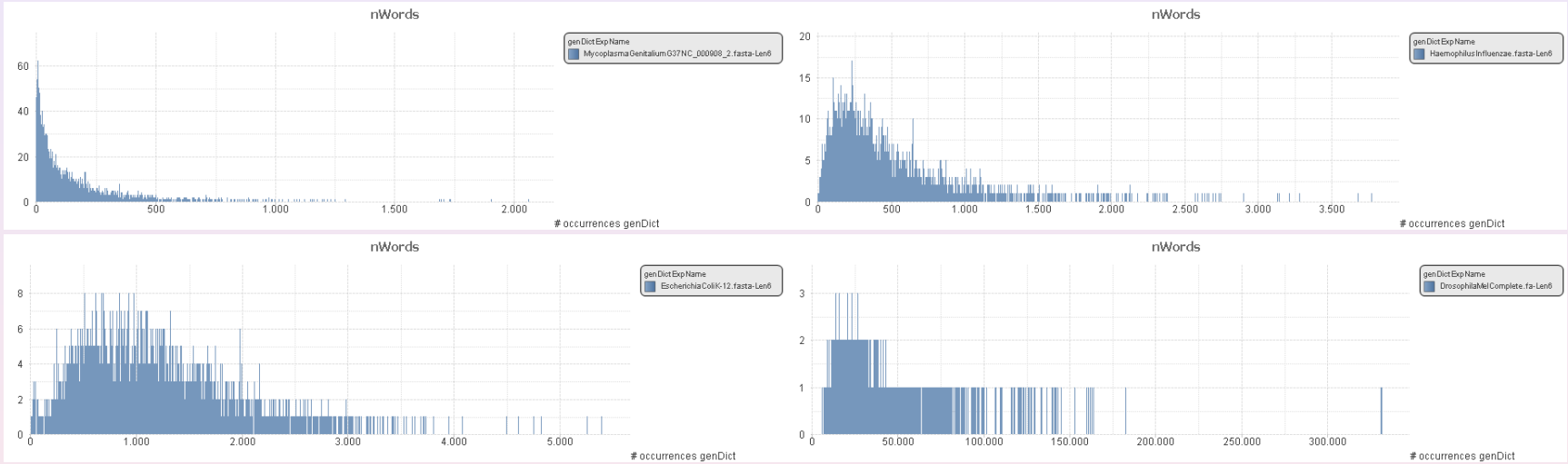
Genomic profiles





A related string problem

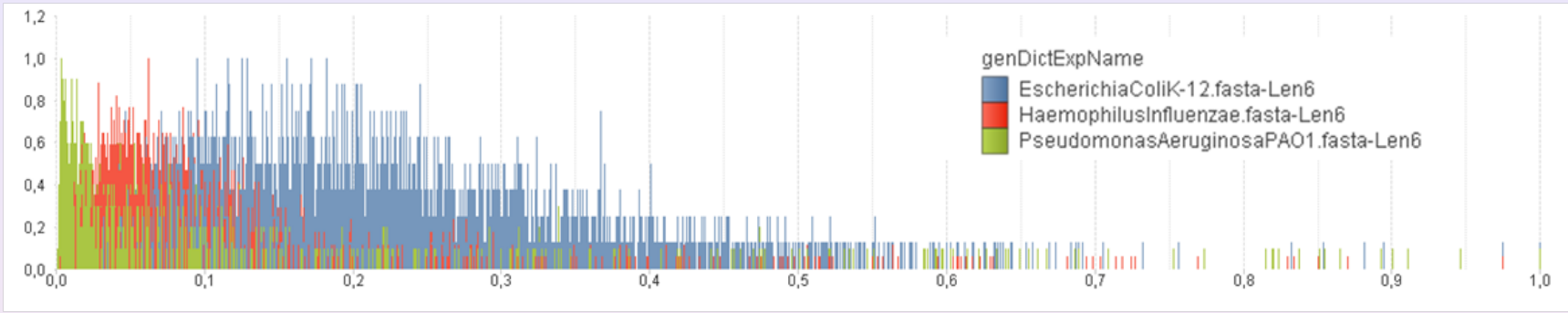
Related to the *word assembly problem*: genome reconstruction method,

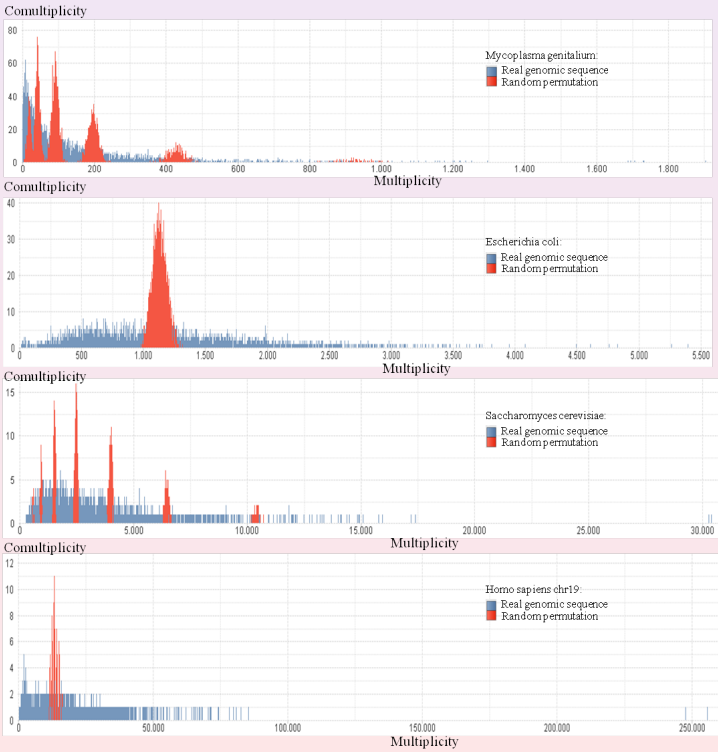


M. genitalium, H. influenzae, E. coli, D. melanogaster.

These studies could improve existent genome reconstruction algorithms, by providing estimations about repeatability of reads according to their own length.

Comparison of genomic profiles - examples

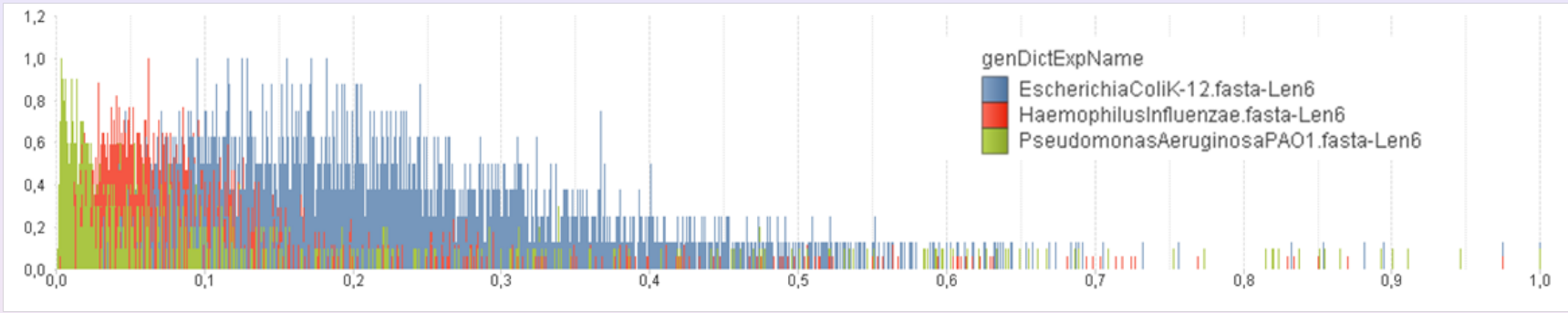


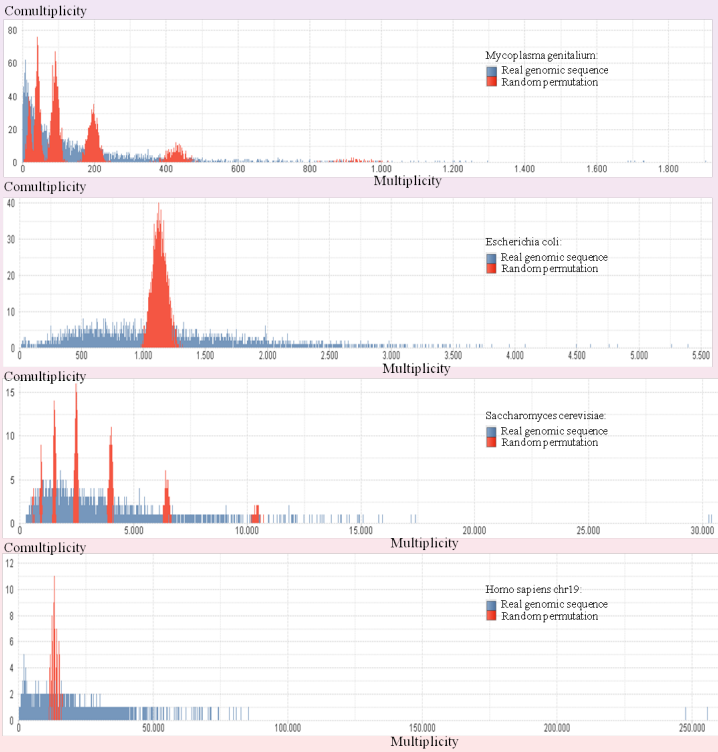


Above, *E. coli*, *H. influenzae*, *P. aeruginosa* are compared by their normalized genomic (6-)profiles.

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

Comparison of genomic profiles - examples

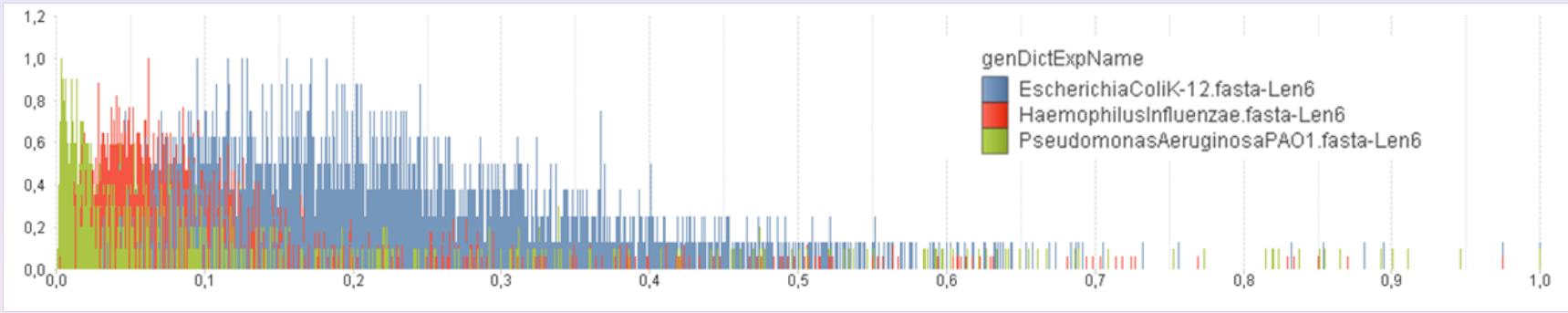




Above, *E. coli*, *H. influenzae*, *P. aeruginosa* are compared by their normalized genomic (6-)profiles.

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

Comparison of genomic profiles - examples

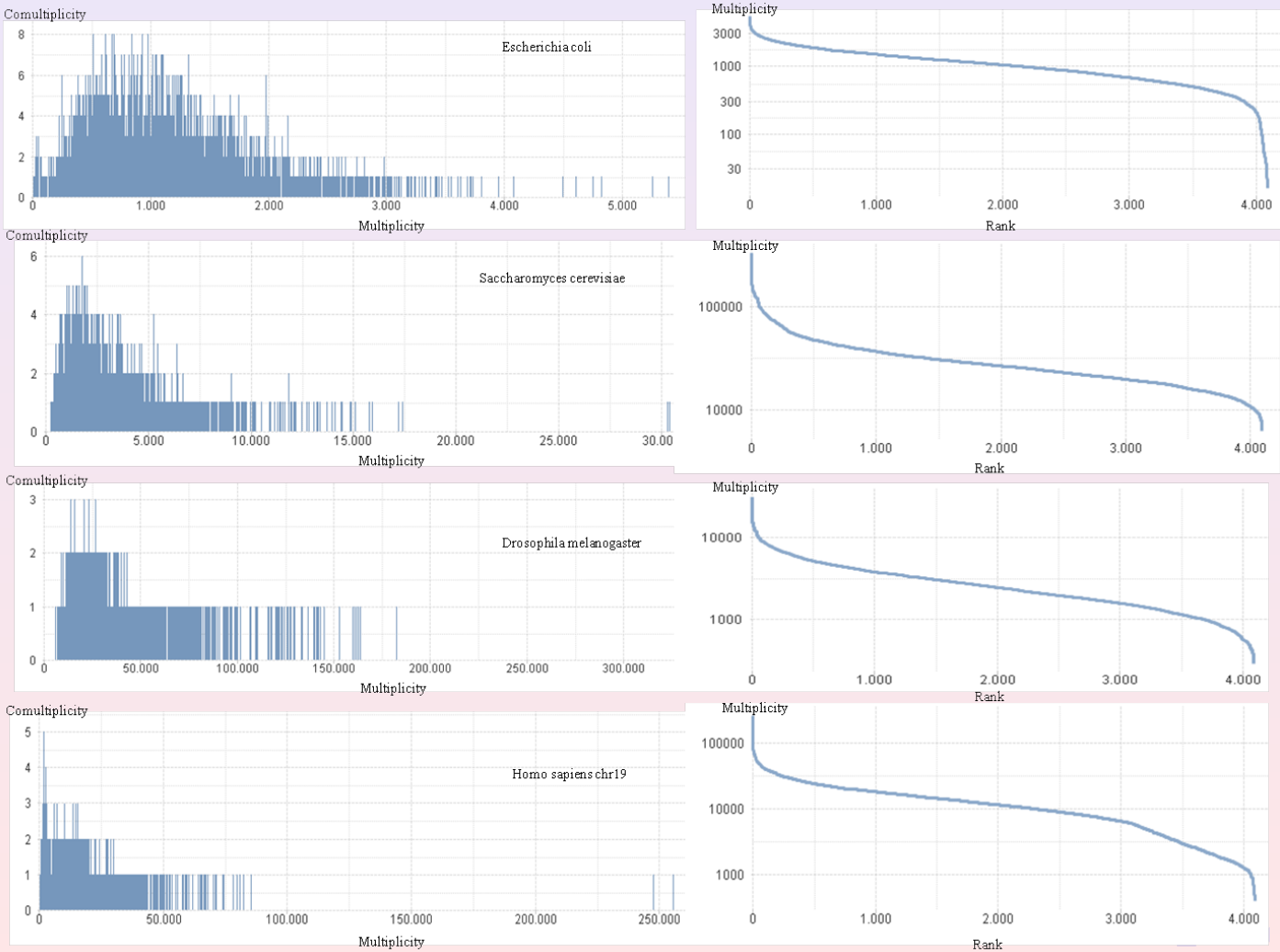


Above, *E. coli*, *H. influenzae*, *P. aeruginosa* are compared by their normalized genomic (6-)profiles.

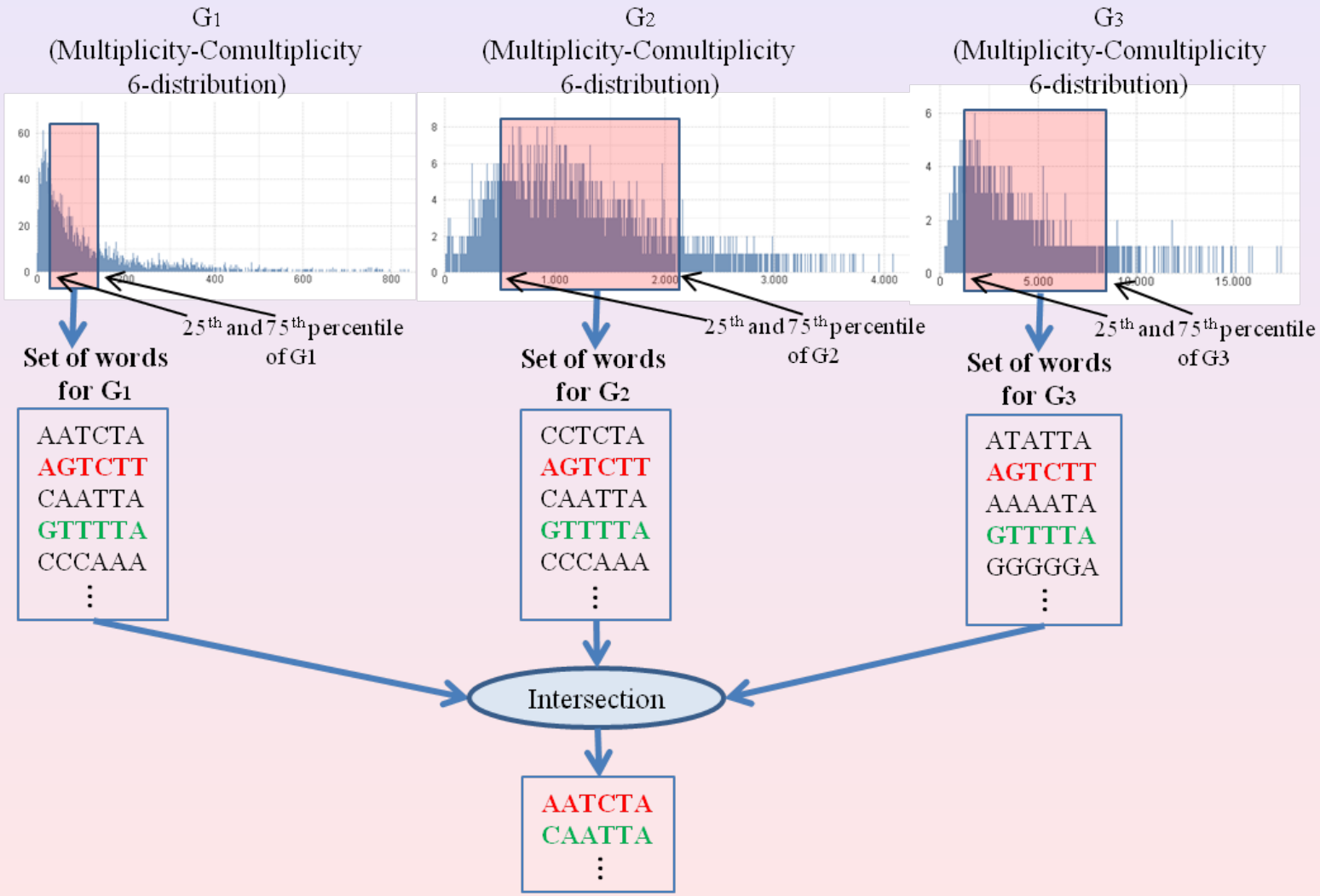
Aside, *M. genitalium*, *E. coli*, *S. cerevisiae*, *H. sapiens chr 19*, are compared with one random permutation (in red) by the multiplicity-comultiplicity profiles.

Genomic profiles and Zipf curves

Zipf curves measure word frequencies in natural languages

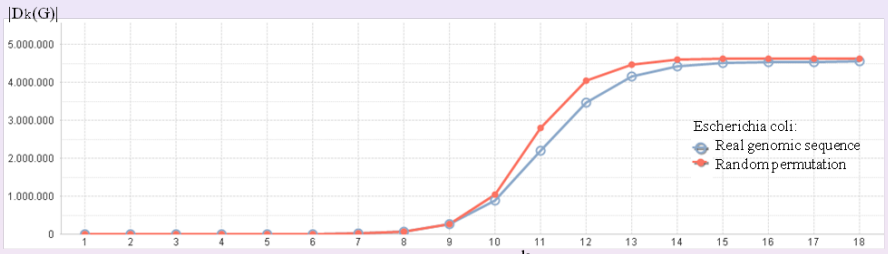


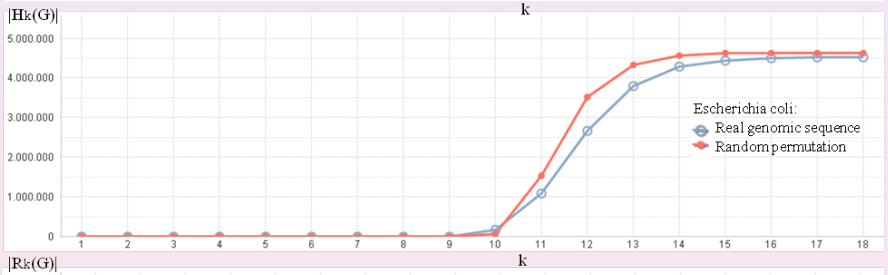
Fairly occurring motifs common to genomes (UCE)

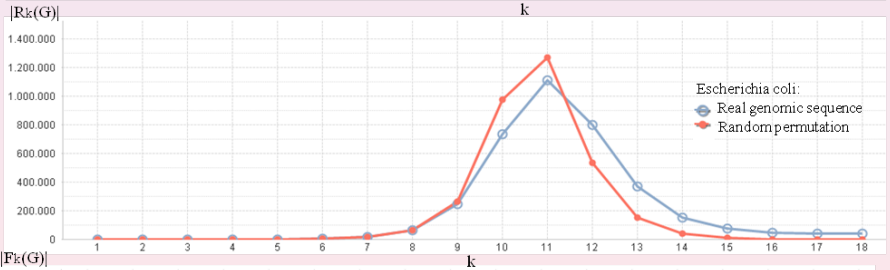


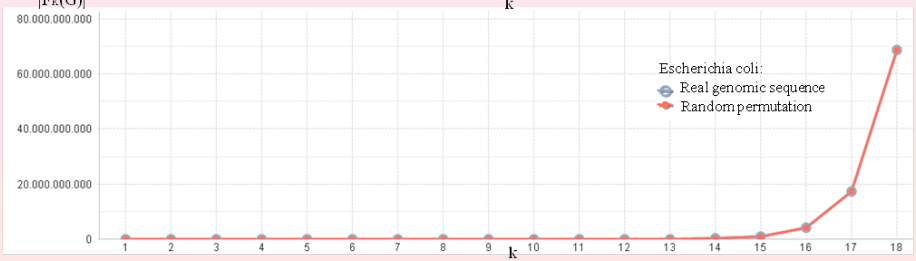
Sizes of *k*-dictionaries: : real vs random genomes

E. coli 's genomic dictionaries

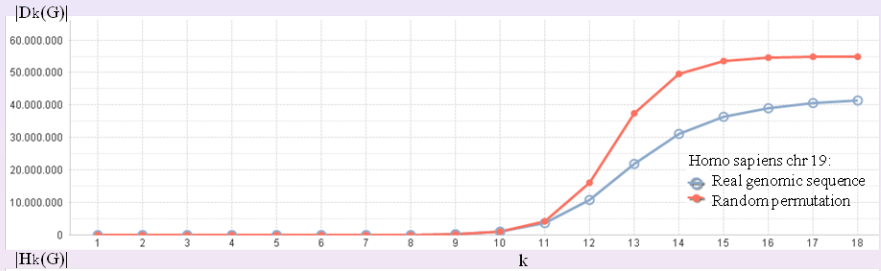


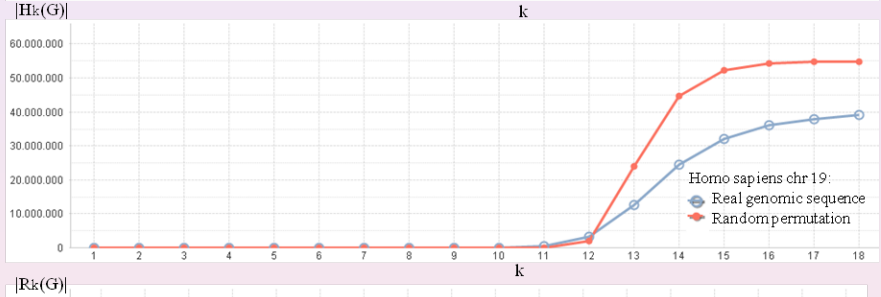


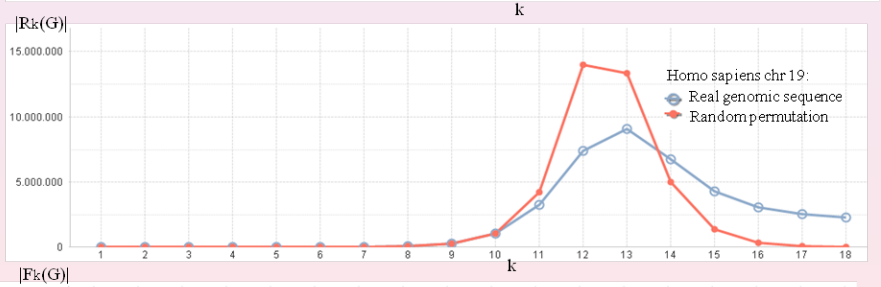


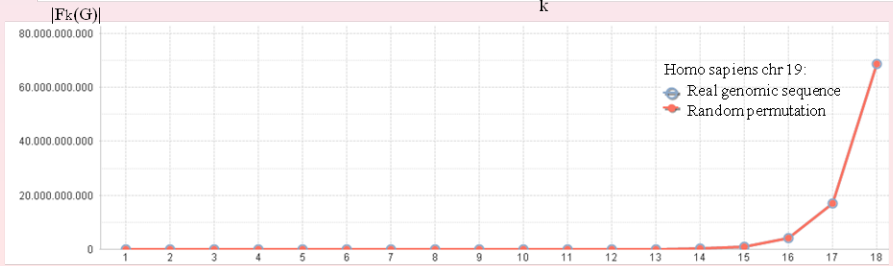


H. sapiens chr19 's dictionaries



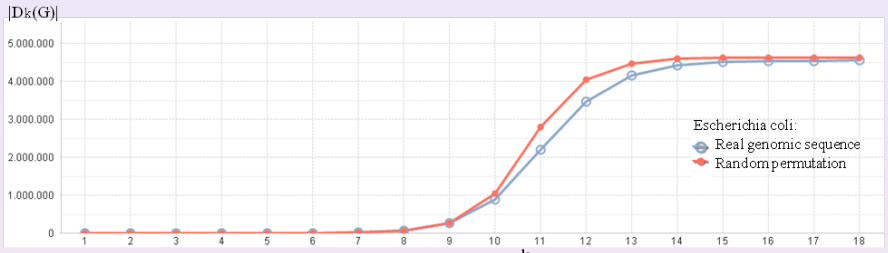


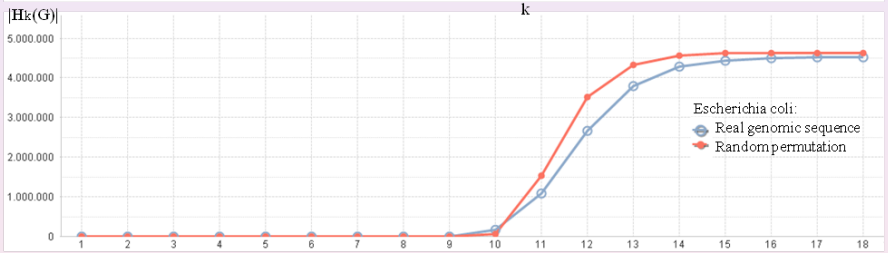


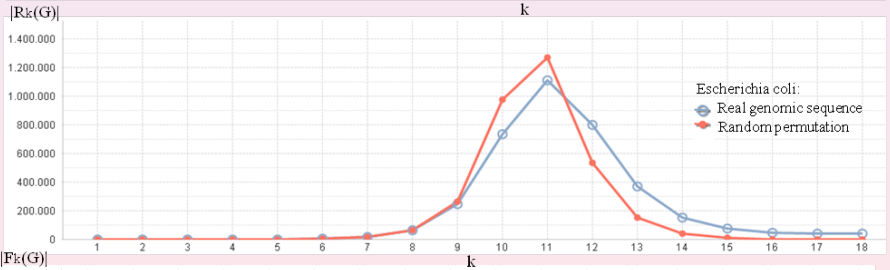


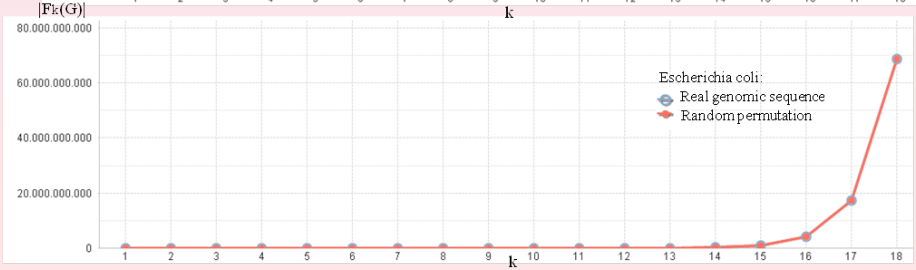
Sizes of *k*-dictionaries: : real vs random genomes

E. coli 's genomic dictionaries

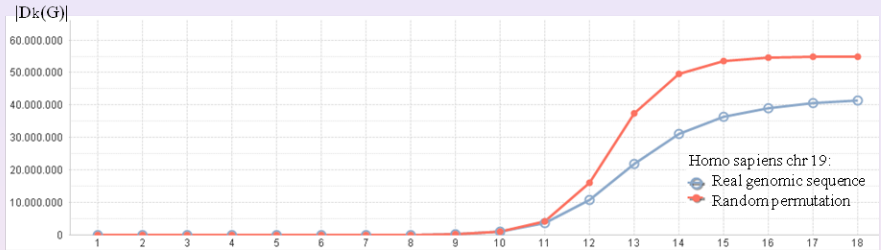


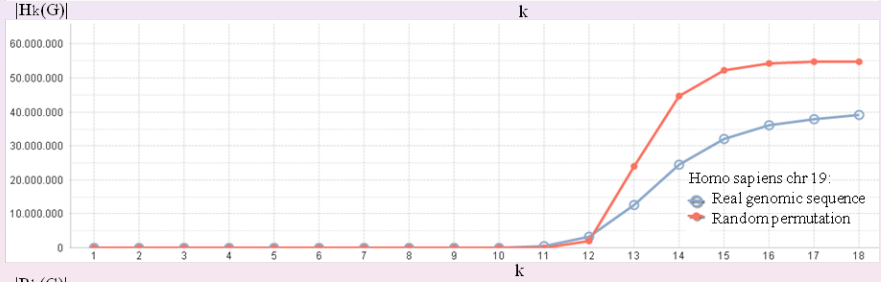


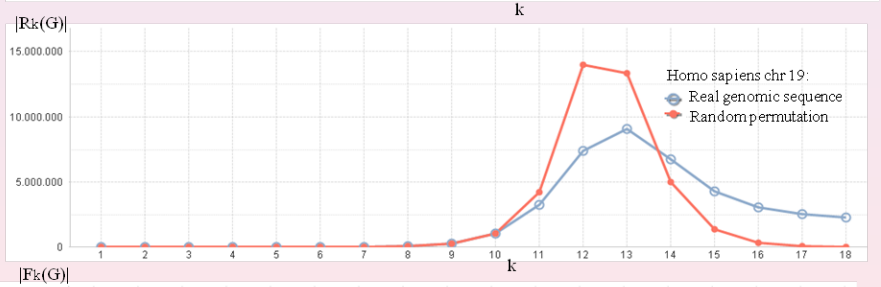


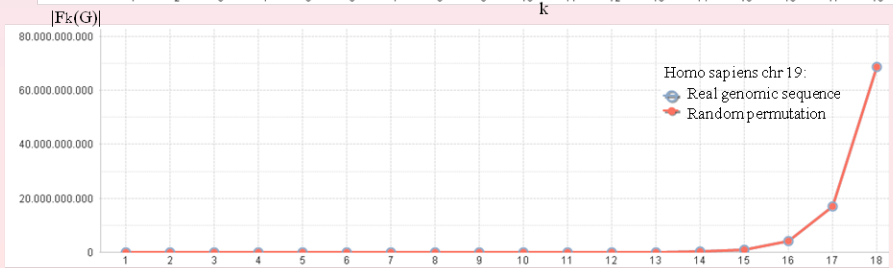


H. sapiens chr19 's dictionaries



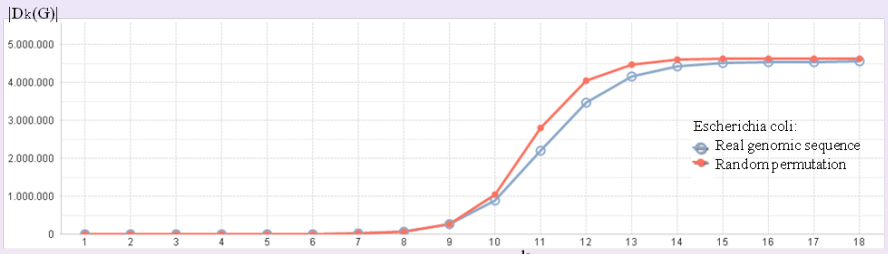


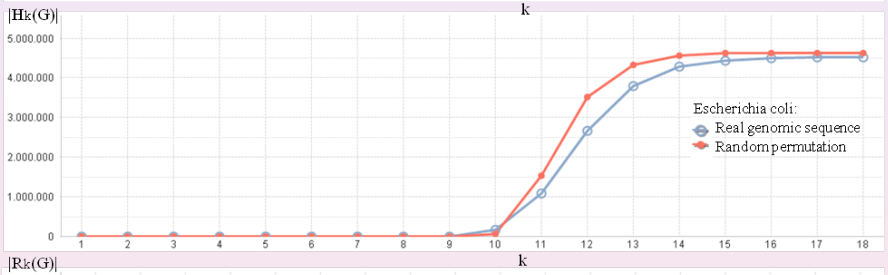


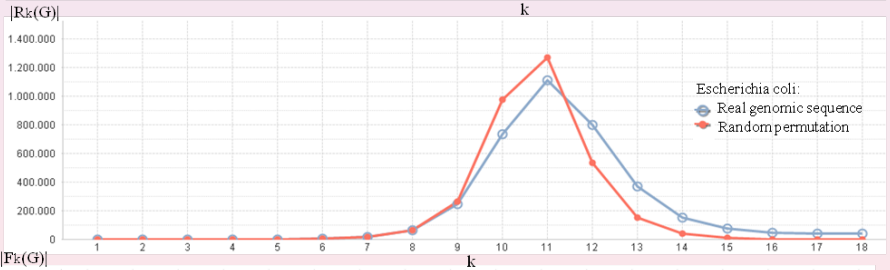


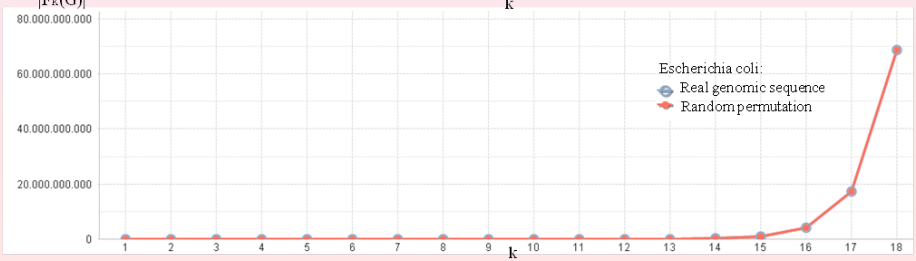
Sizes of *k*-dictionaries: : real vs random genomes

E. coli 's genomic dictionaries

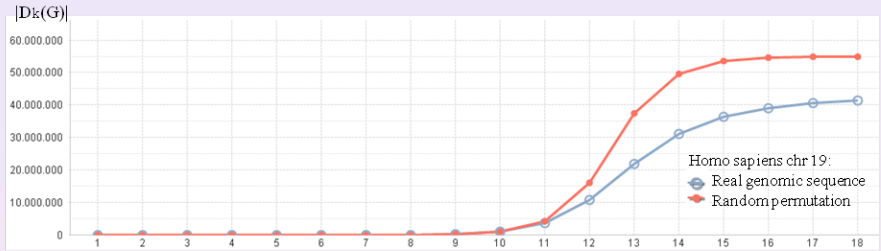


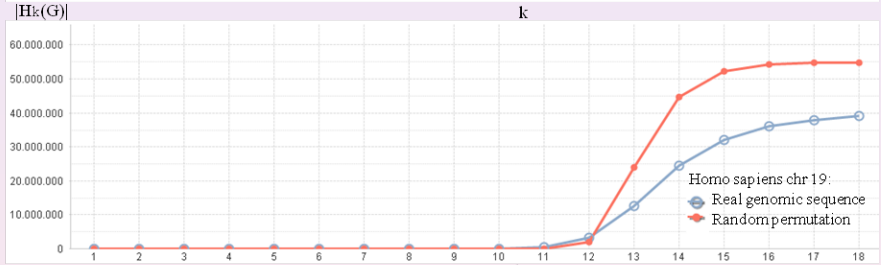


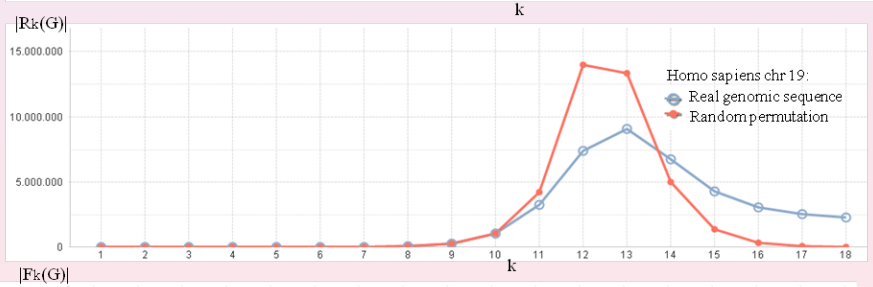


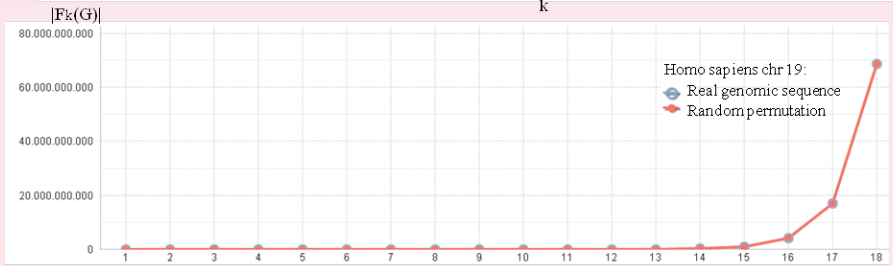


H. sapiens chr19 's dictionaries



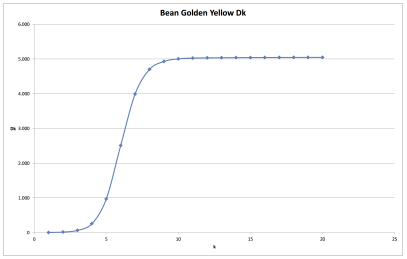


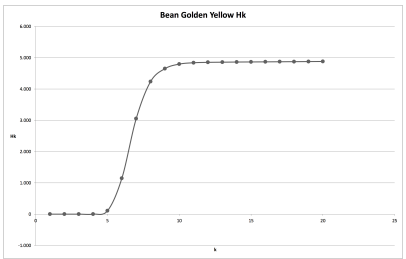


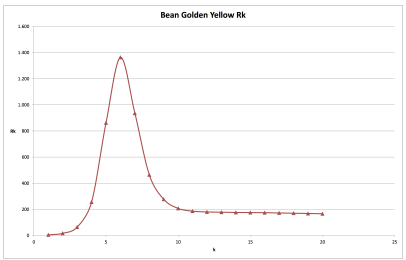


Sizes of *k*-dictionaries: : virus and bacterium

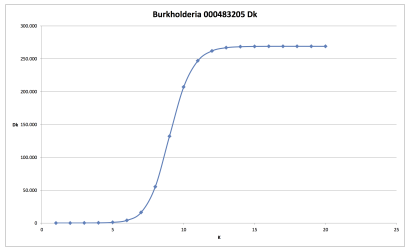
BeanGoldenYello 's dicts

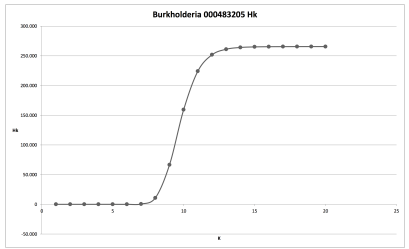


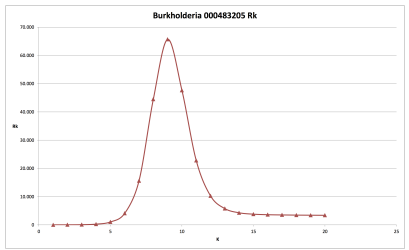




Burkholderia 's dicts

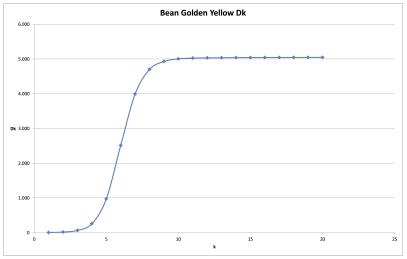


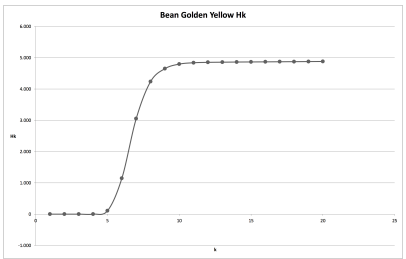


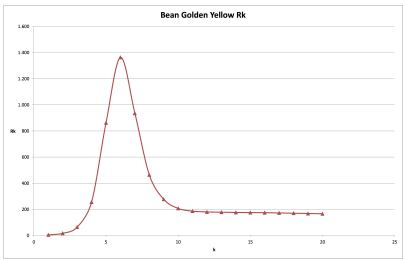


Sizes of *k*-dictionaries: : virus and bacterium

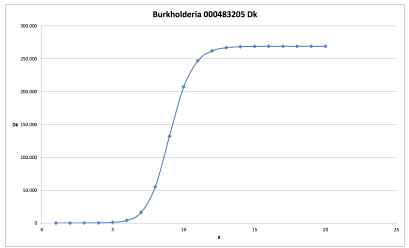
BeanGoldenYello 's dicts

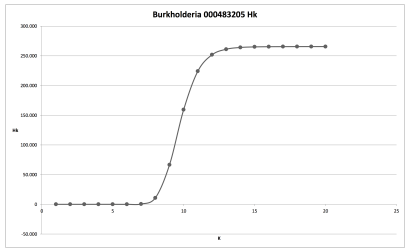


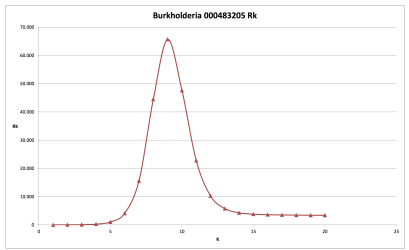




Burkholderia 's dicts

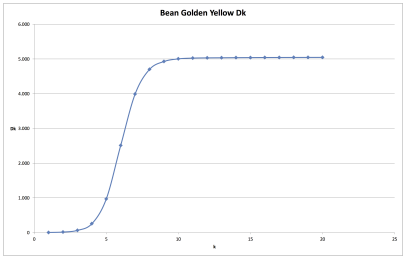


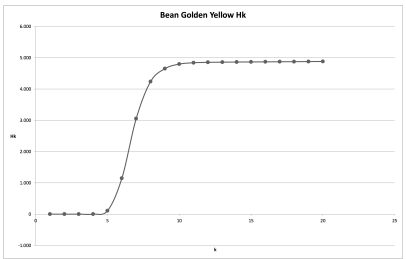


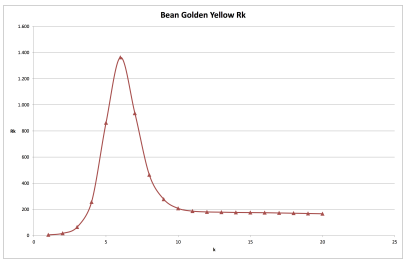


Sizes of *k*-dictionaries: : virus and bacterium

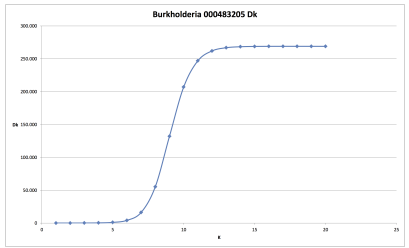
BeanGoldenYello 's dicts

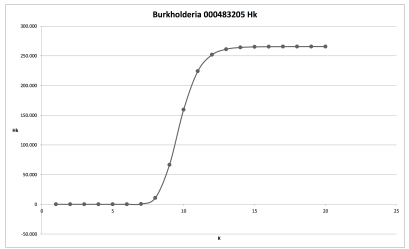


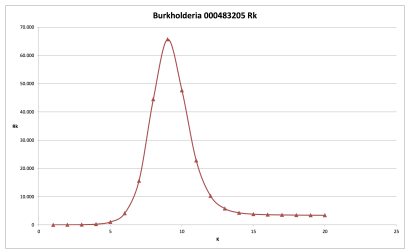




Burkholderia 's dicts







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.

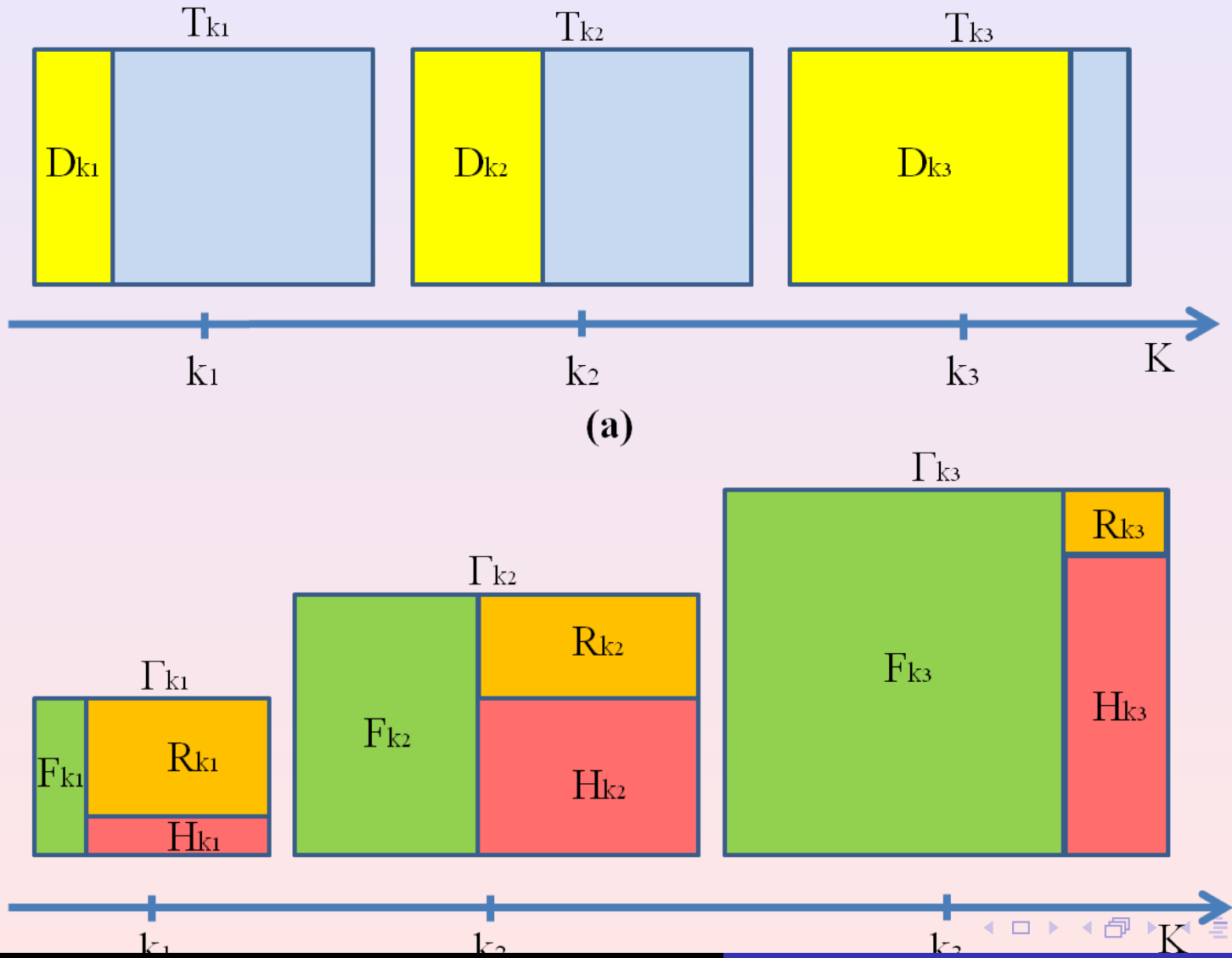
Phase transitions for hapax/repeat cardinality ratio

$DT_k = \frac{|D_k(G)|}{|T_k(G)|}$ (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)

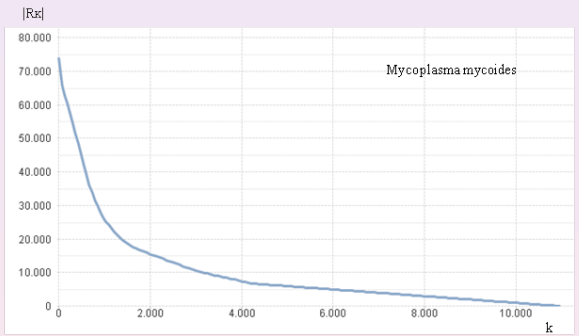
Analysis of (relatively) long repeats reduction

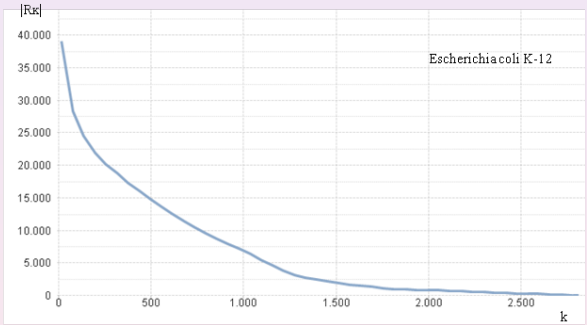


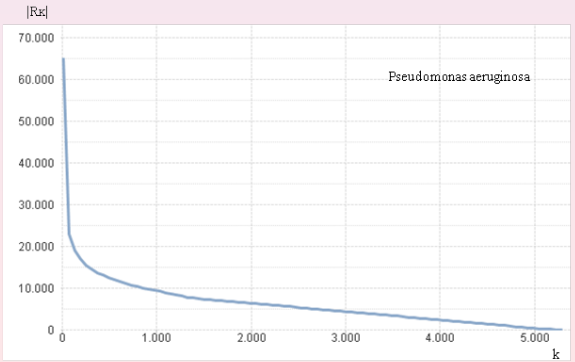
Informational indexes

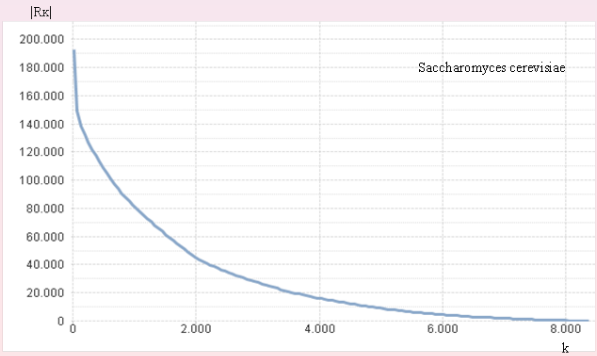
Minimal Hapax Length (genome itself is an hapax, any word including an hapax is an hapax), *Maximal Repeat Length* (any subword of repeat is a repeat), *Minimal Forbidden Length*

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





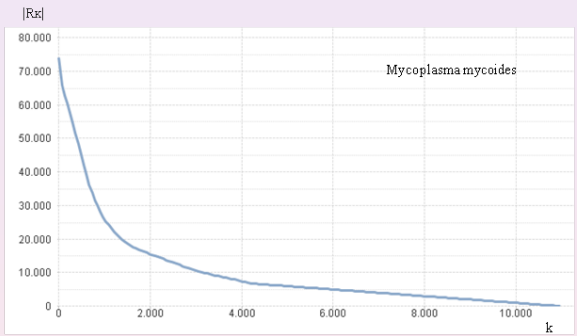


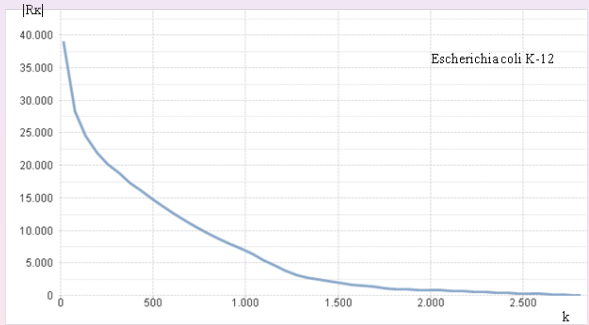


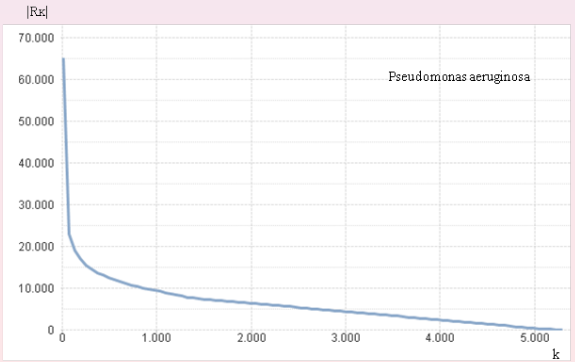
Informational indexes

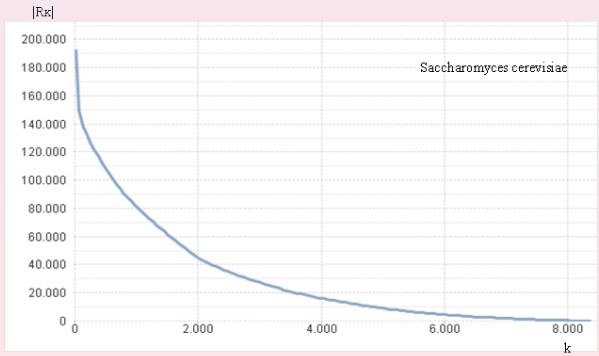
Minimal Hapax Length (genome itself is an hapax, any word including an hapax is an hapax), *Maximal Repeat Length* (any subword of repeat is a repeat), *Minimal Forbidden Length*

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





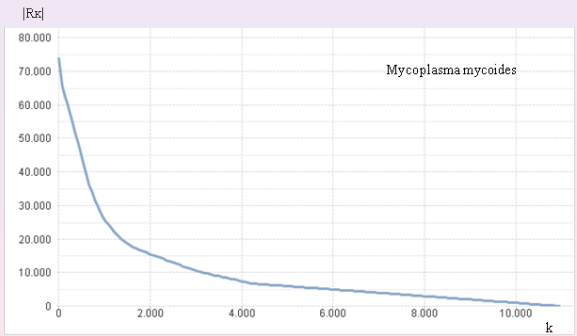


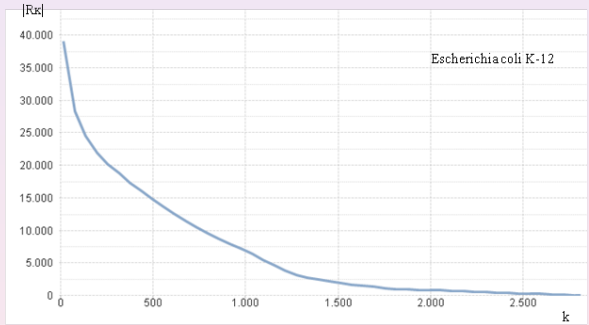


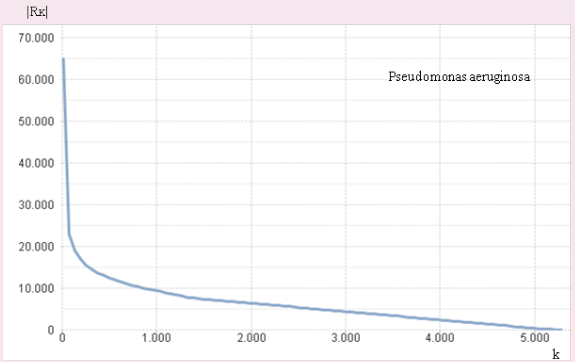
Informational indexes

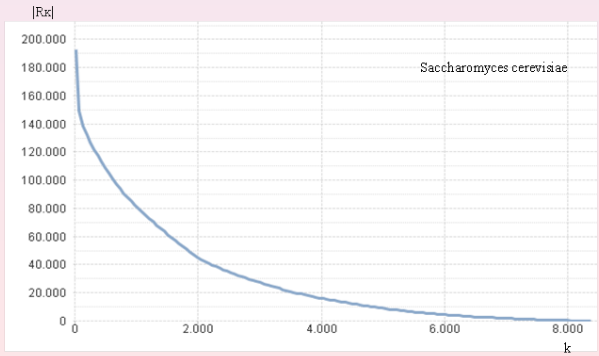
Minimal Hapax Length (genome itself is an hapax, any word including an hapax is an hapax), *Maximal Repeat Length* (any subword of repeat is a repeat), *Minimal Forbidden Length*

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









Informational indexes: some properties

For any genome G (of length n):

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

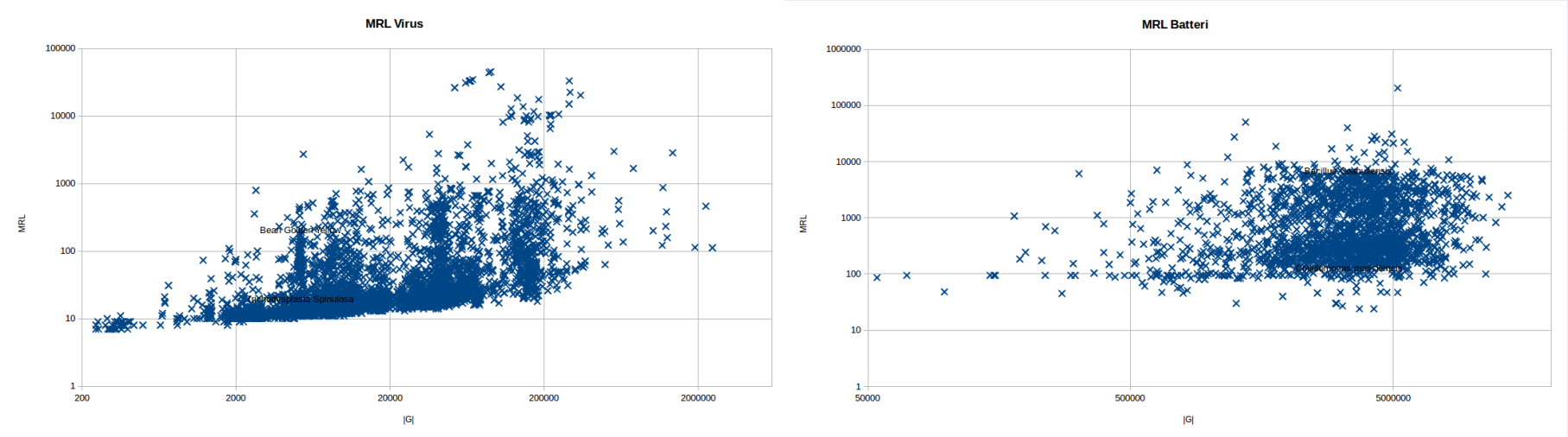
$AR_k = \frac{|T_k \backslash H_k|}{|R_k|}$ average k-factors repeatability

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

$MFL \leq MHL + 1$

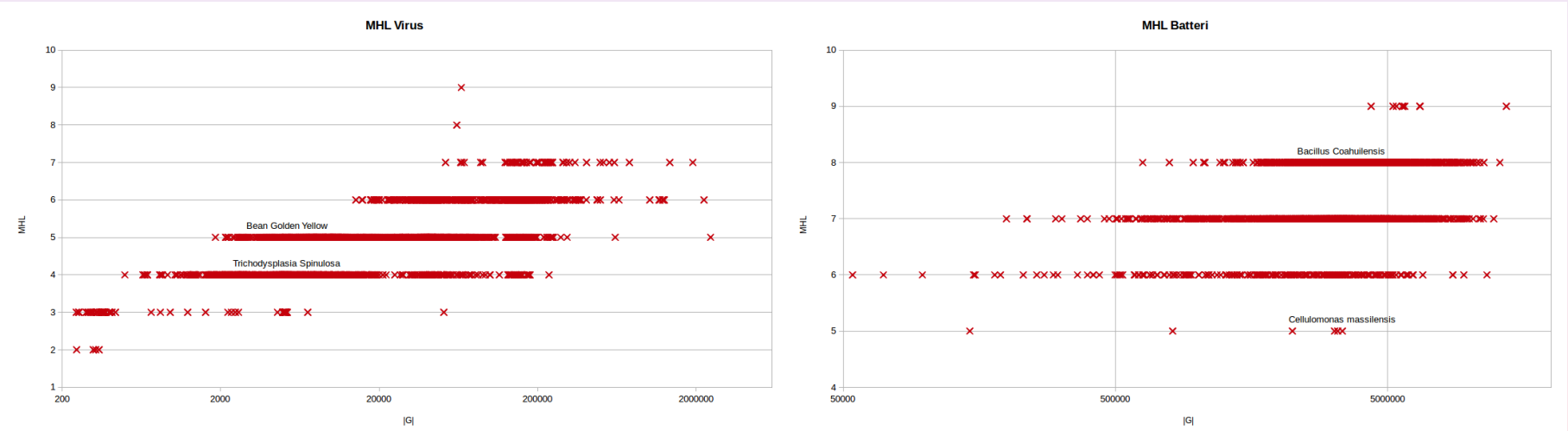
MRL of virus and bacteria

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



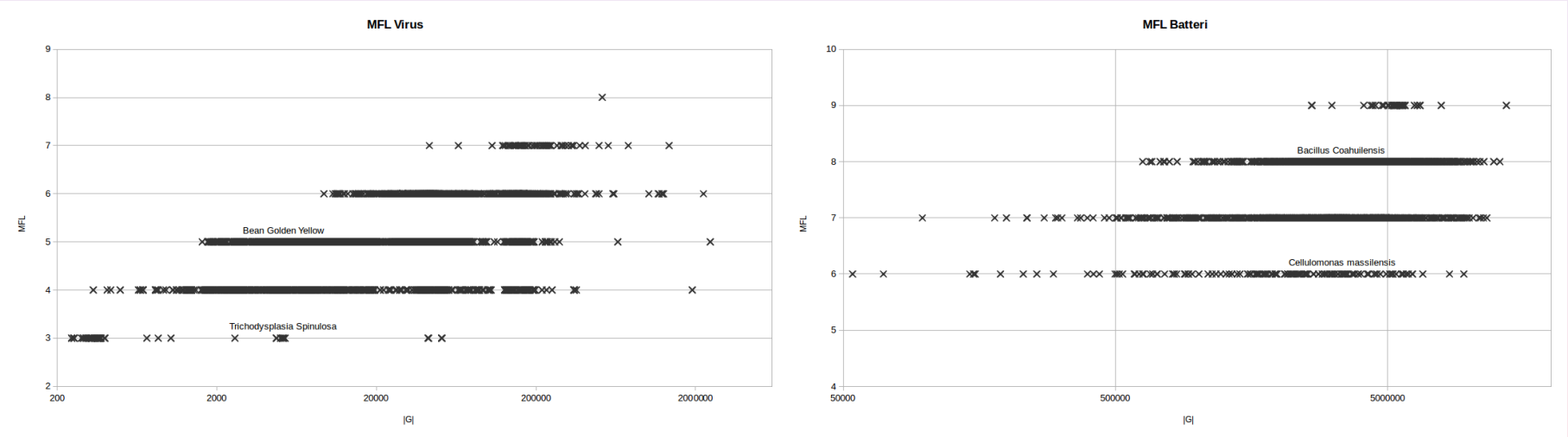
MHL of virus and bacteria

Major part of viruses has MHL and MFL between 4 and 6, bacteria between 6-8.



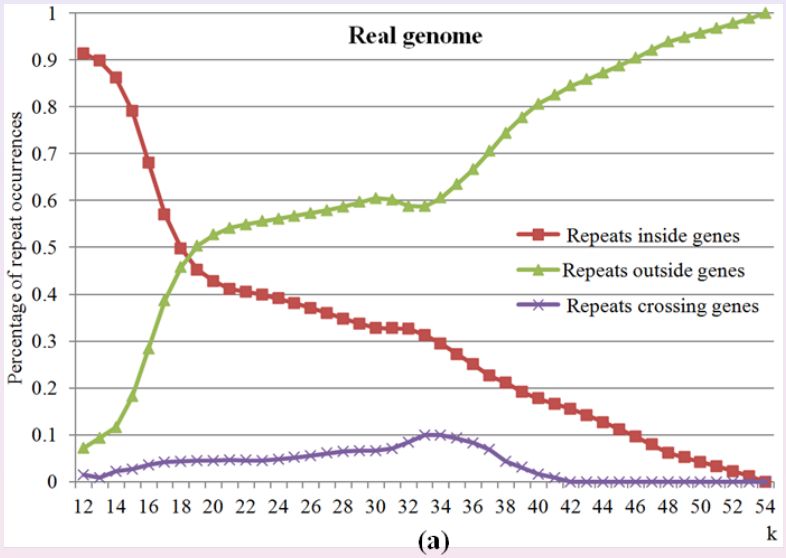
MFL of virus and bacteria

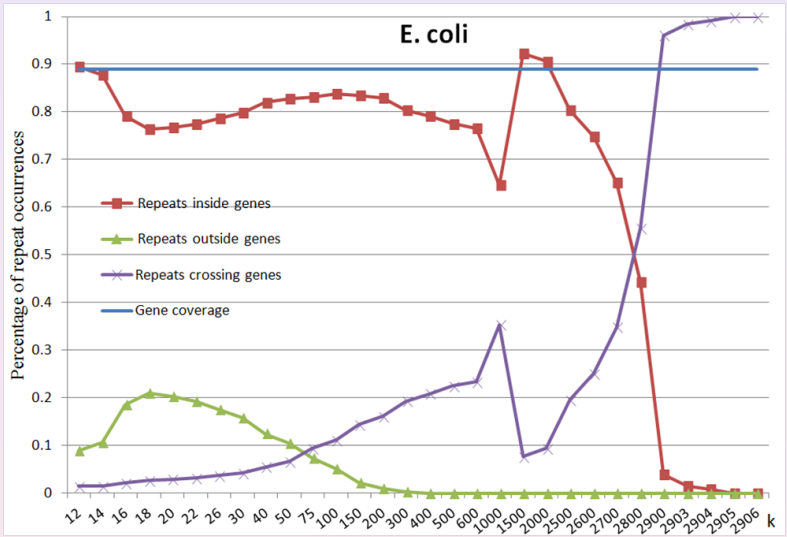
Clusters were found out: MHL=MFL+1, MFL-1, MFL+2.

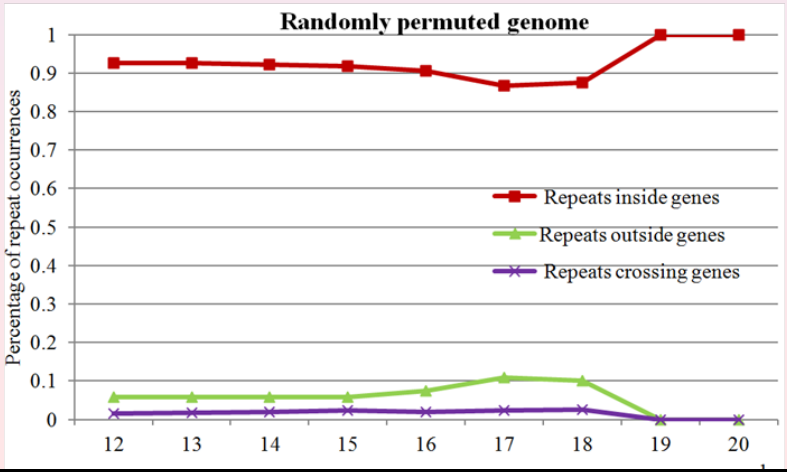


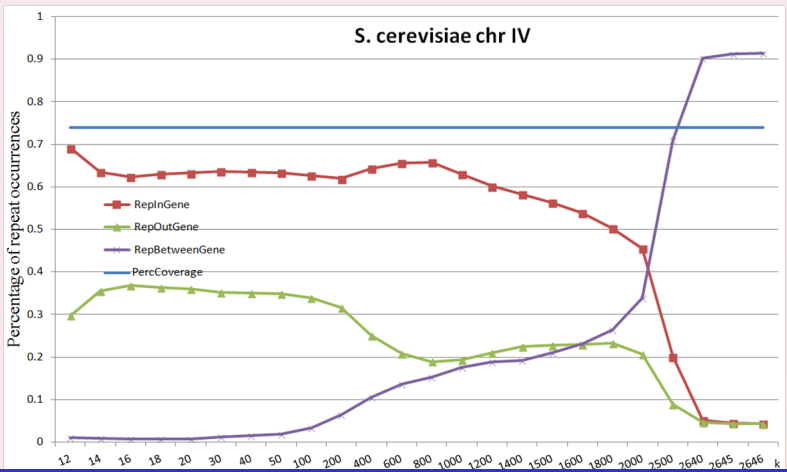
Repeat localization (inside, outside, across genes)

Percentage of *k*-repeat occurrences, in function of the length *k*



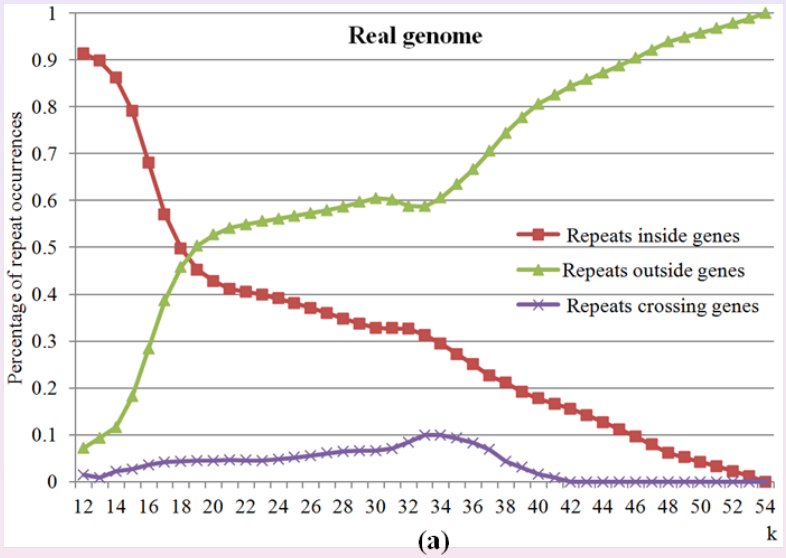


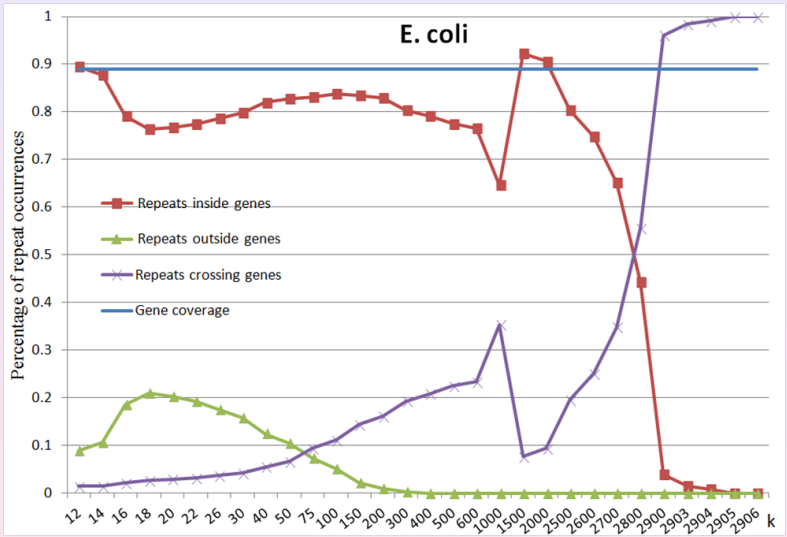


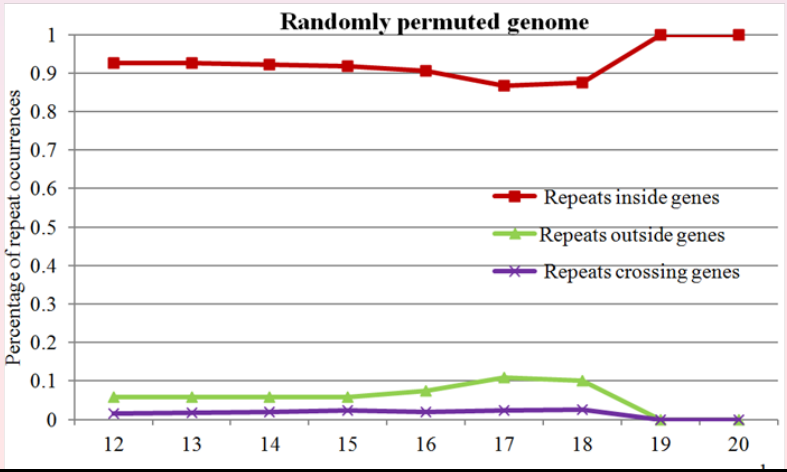


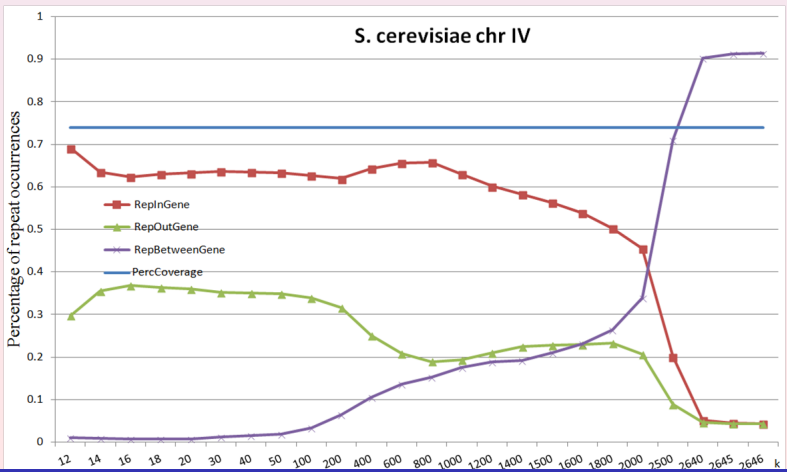
Repeat localization (inside, outside, across genes)

Percentage of *k*-repeat occurrences, in function of the length *k*



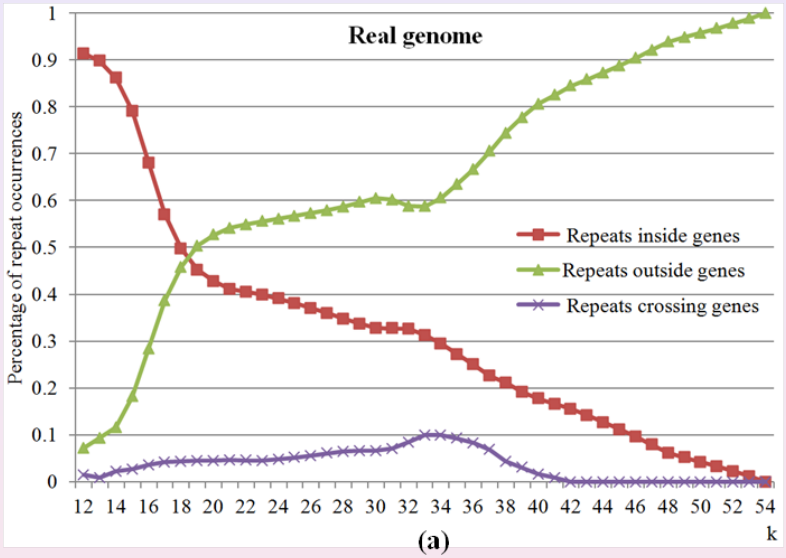


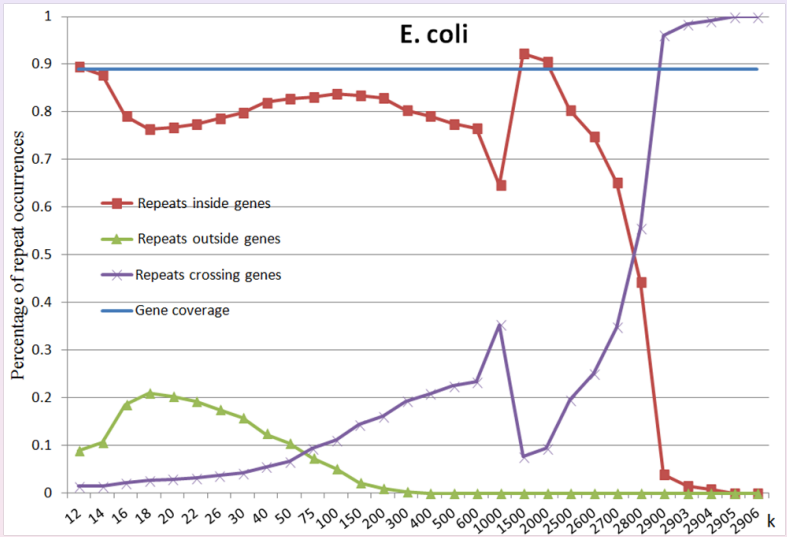


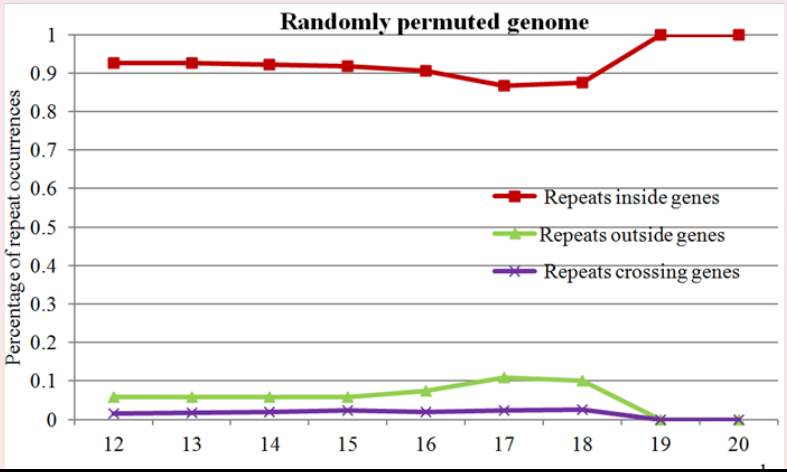


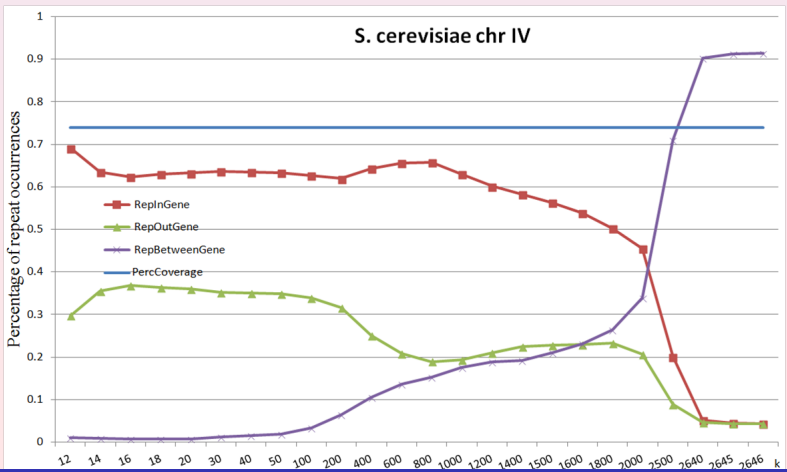
Repeat localization (inside, outside, across genes)

Percentage of *k*-repeat occurrences, in function of 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,

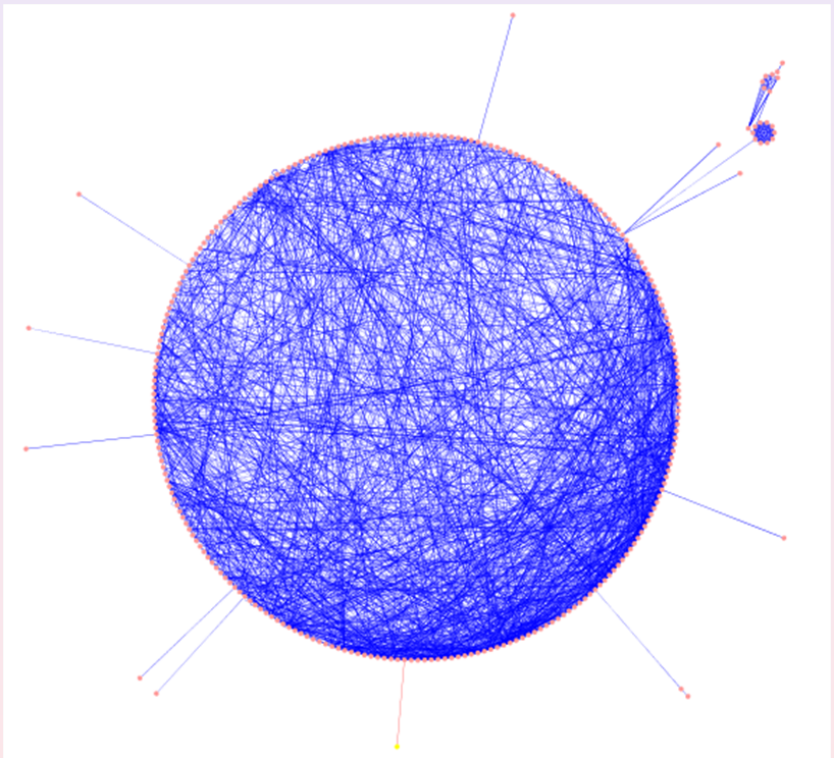
- two genes are connected (in E_k) if they share at least one k-repeat. The set of shared k-repeats is the edge label.

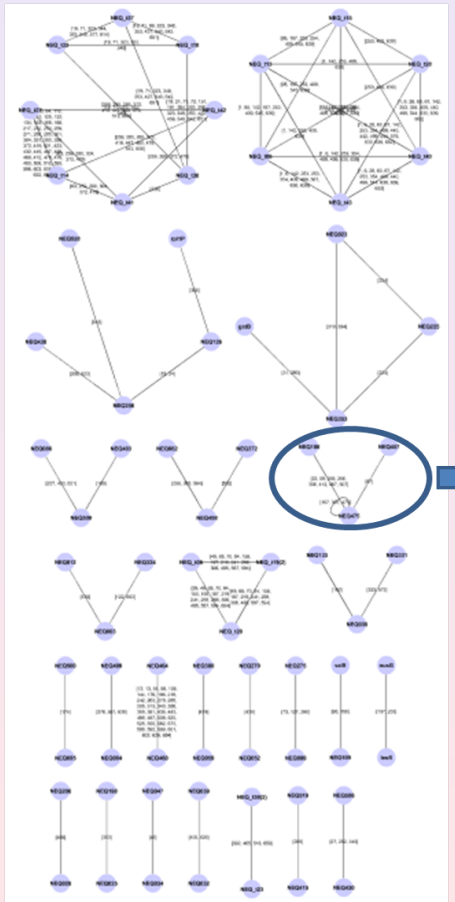
By sketching the k value, nodes and edges in G_k decrease (as isolated nodes are deleted), until the network disappears[10]. A break-point ($k = 18$) where the networks pass from having a big connected component to being a set of "vanishing" clusters.

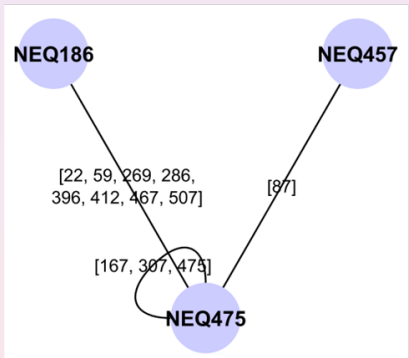
[10] In *N. equitans*, G_{54} is empty, while MR = 139

Genetic repeats within a specific genome

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

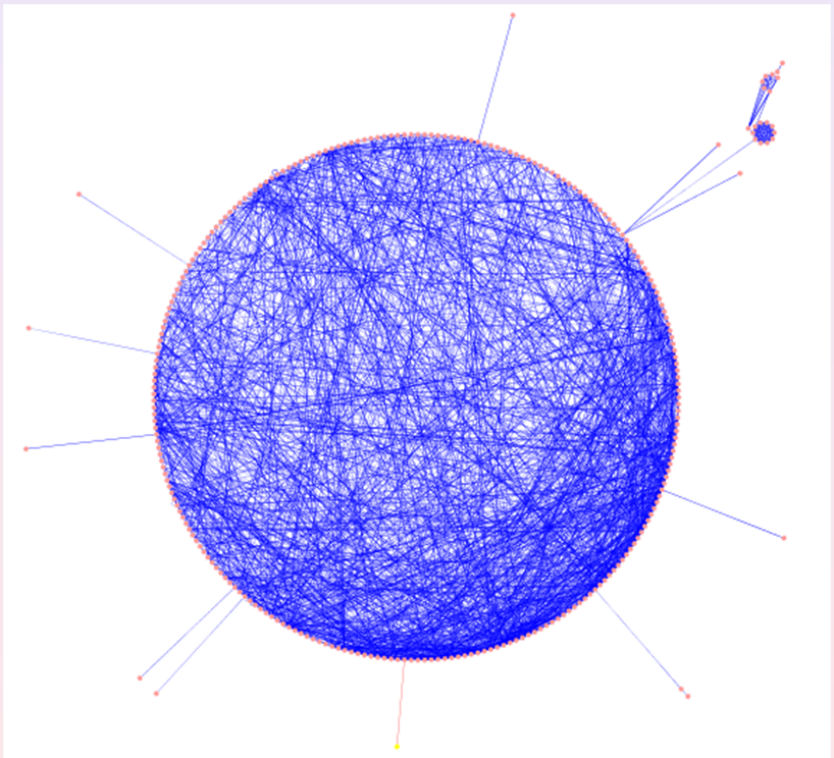


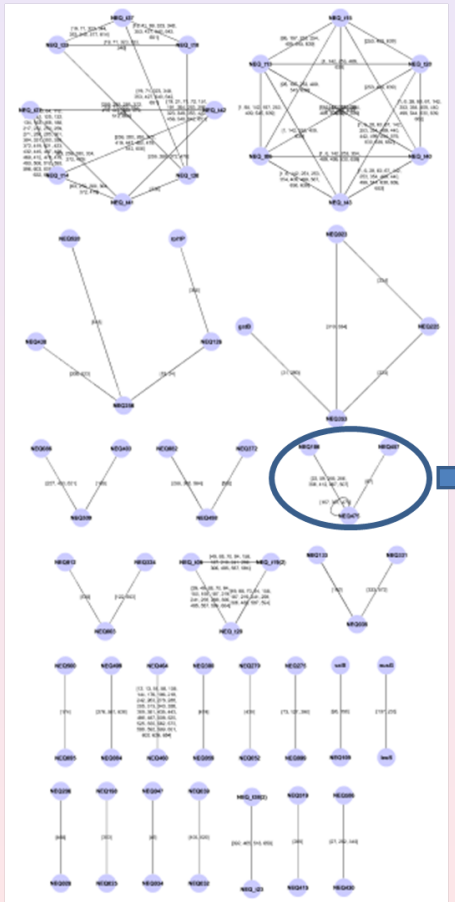


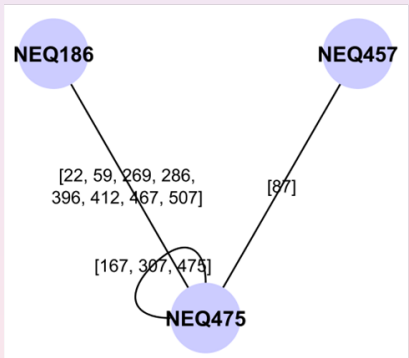


Genetic repeats within a specific genome

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

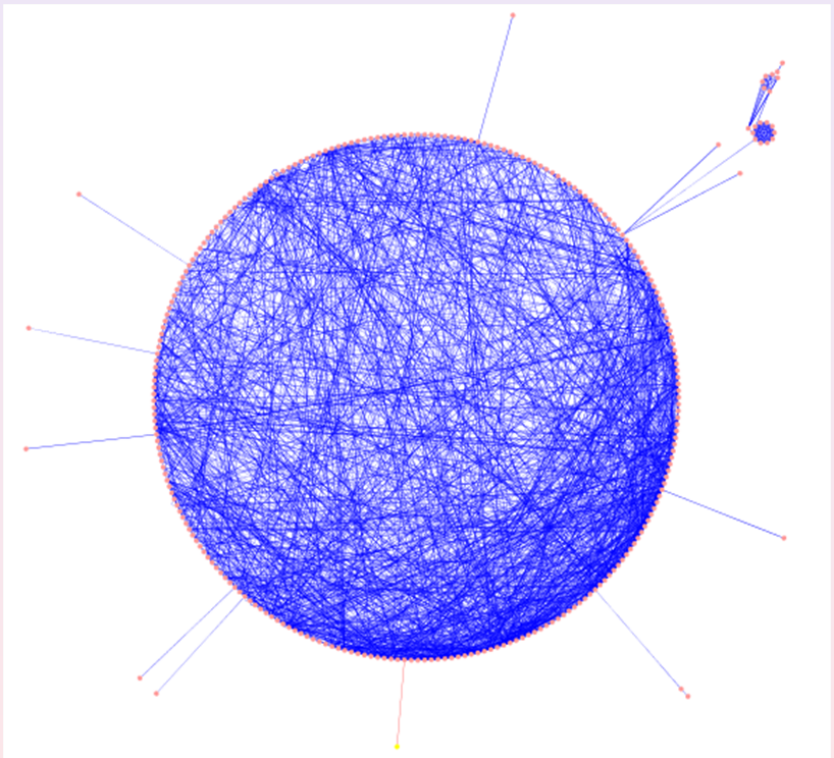


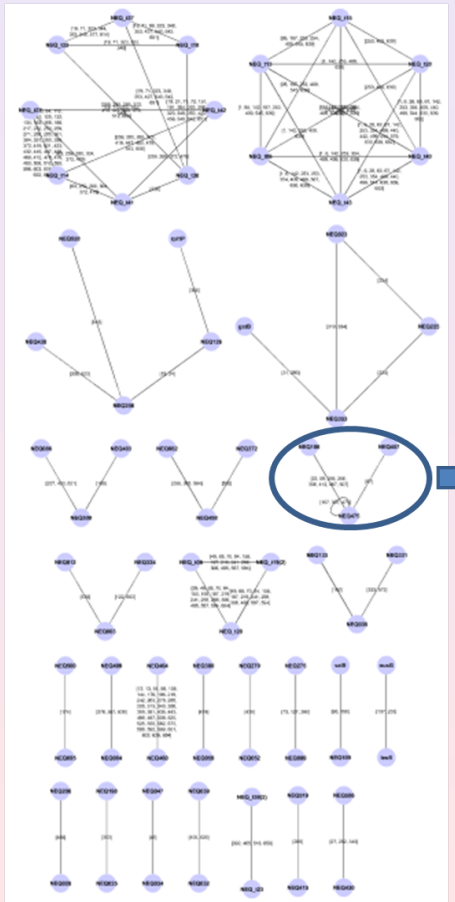


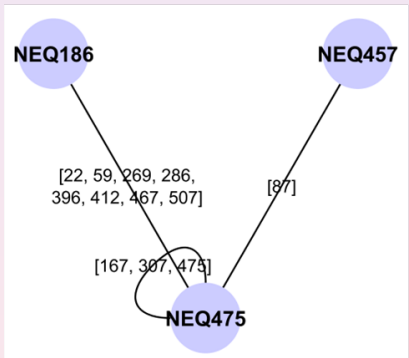


Genetic repeats within a specific genome

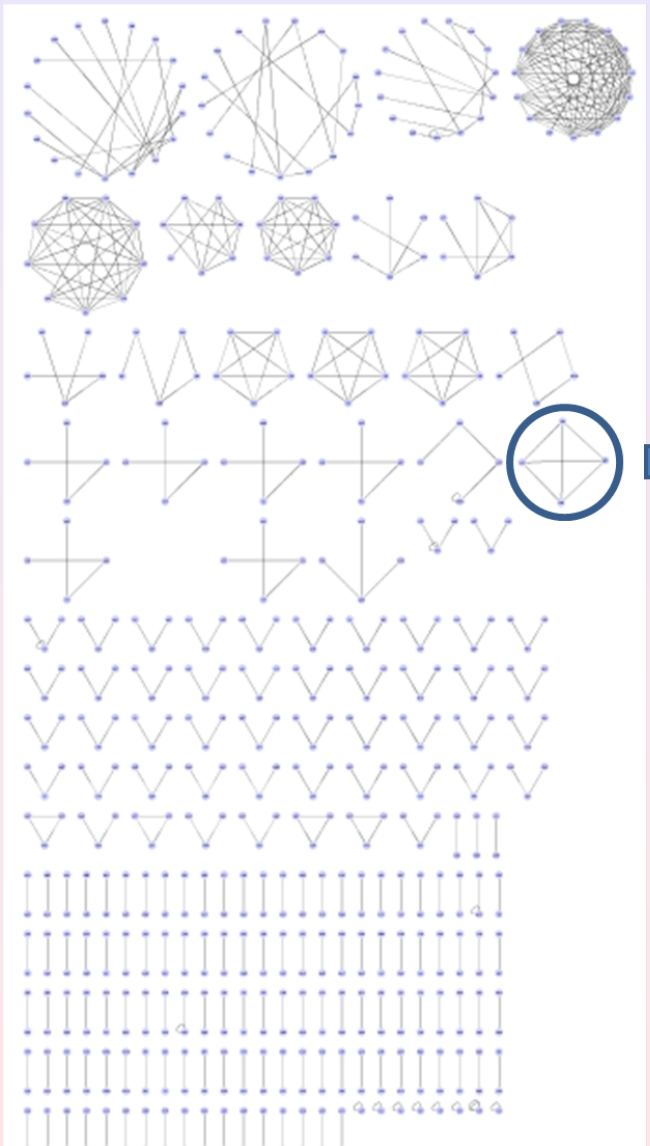
N. equitans 's gene networks G_{14} - G_{18}

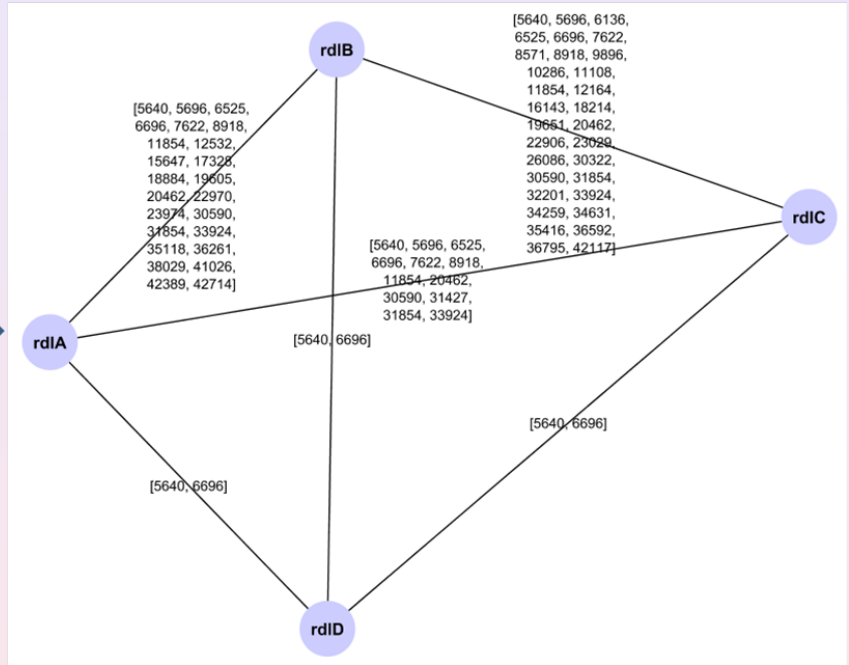






Another example: *E. coli* 's gene network G_{18}



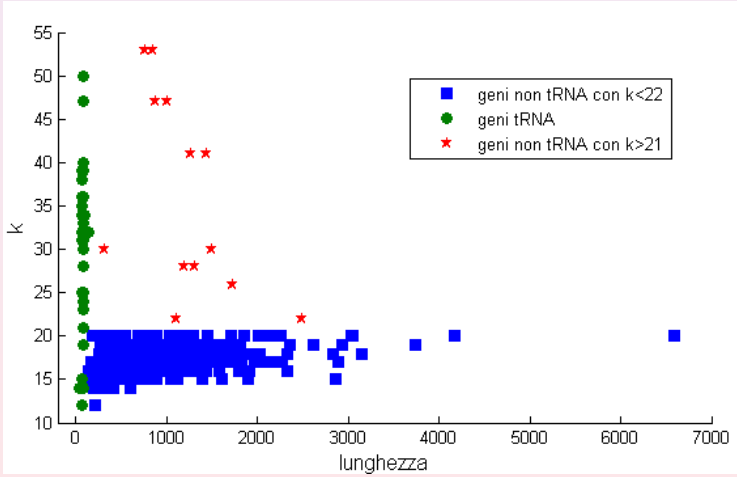


A first investigation of repeat sharing gene networks

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.[11]

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



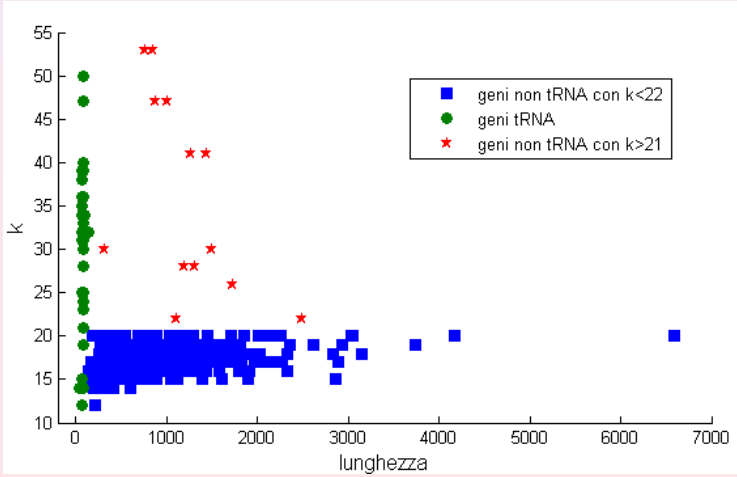
[11] A. Castellini et al., Natural Computing 2015

A first investigation of 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.[11]

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



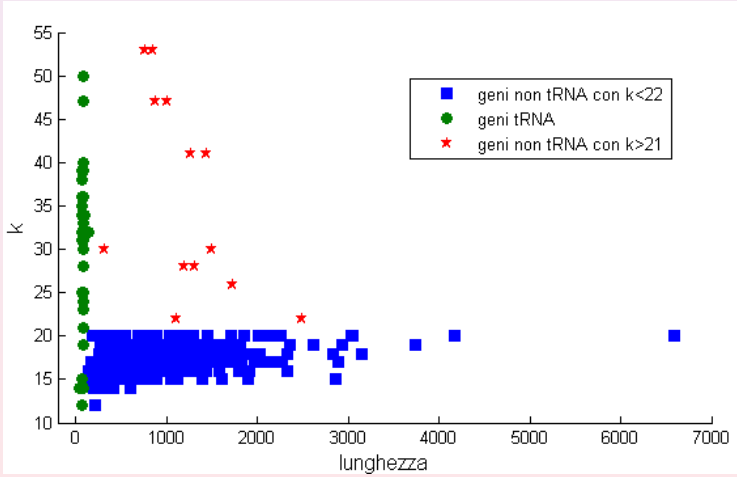
[11] A. Castellini et al., Natural Computing 2015

A first investigation of repeat sharing gene networks

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

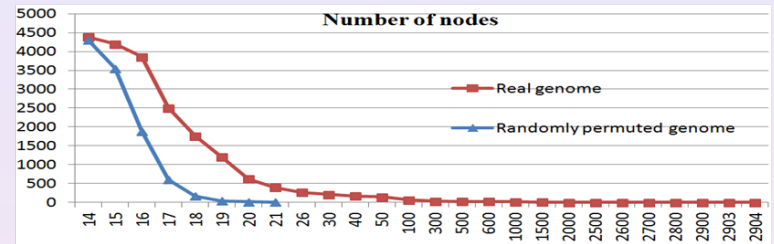
Highly connected genes, for k long enough, have similar functionality and turned out involved in important biological pathways, such as DNA repair and replication.[11]

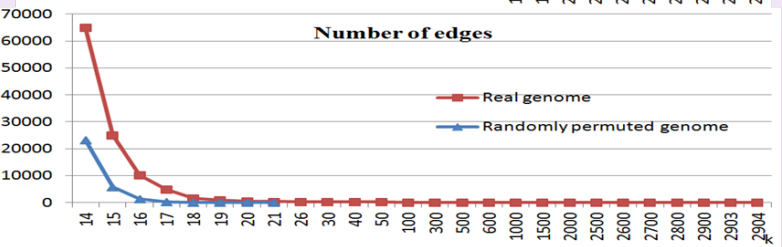
Gene length compared with repeat length k, to measure edge significance (genes shorter than 100 encode for tRNA)

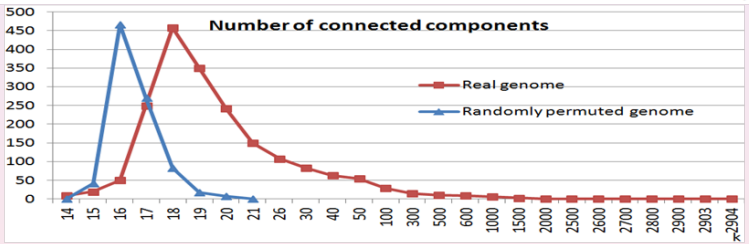


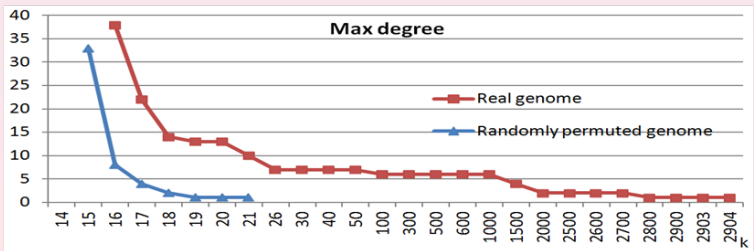
[11] A. Castellini et al., Natural Computing 2015

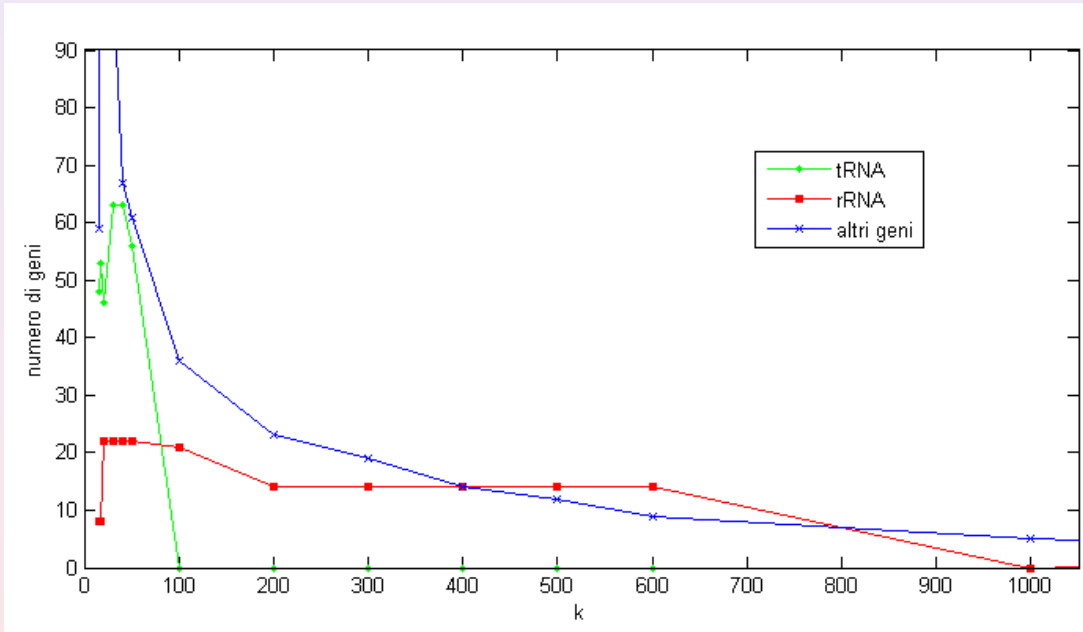
Basic E.coli network analysis



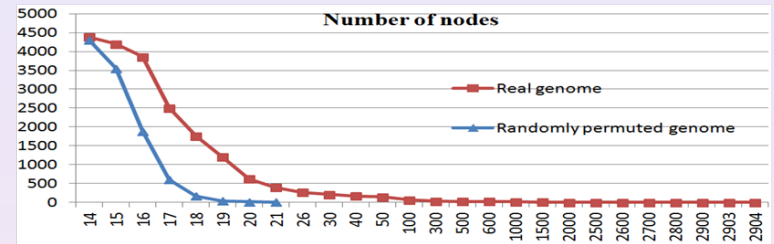


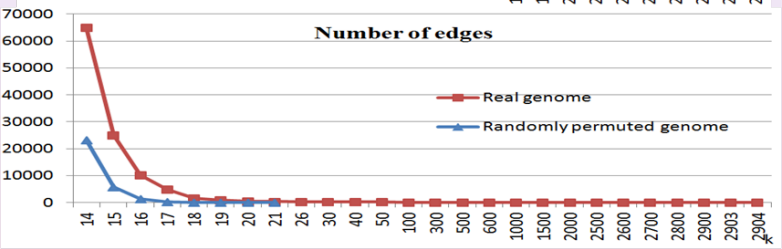


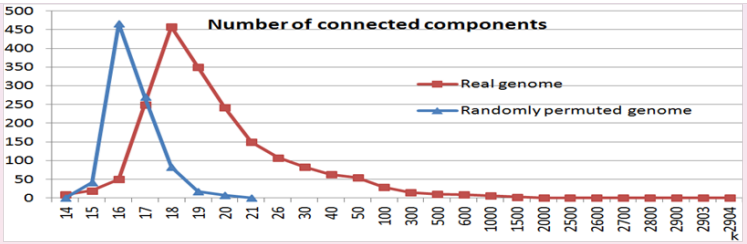


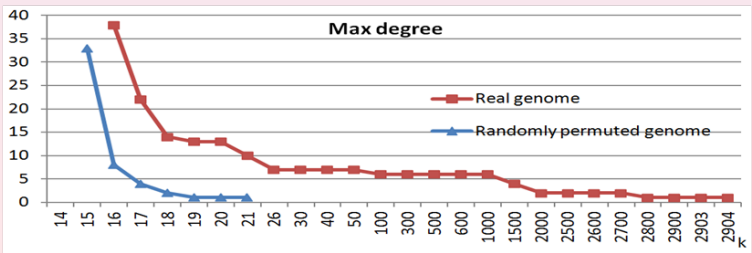


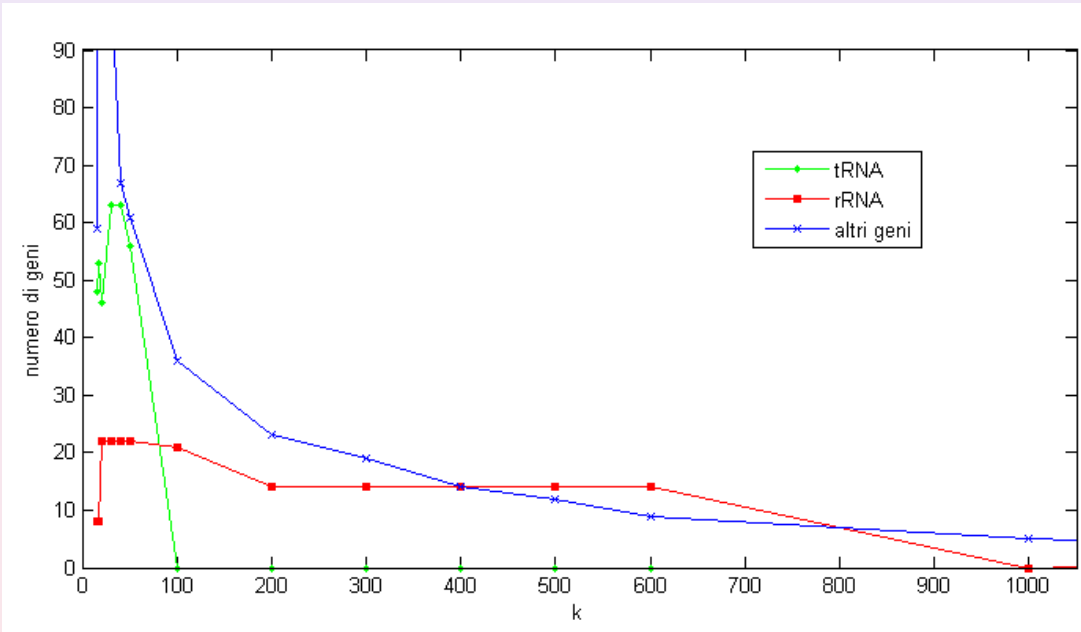
Basic E.coli network analysis



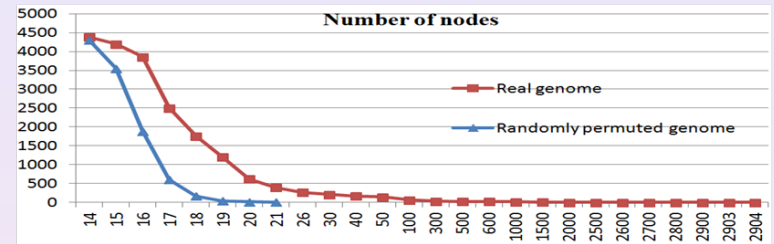


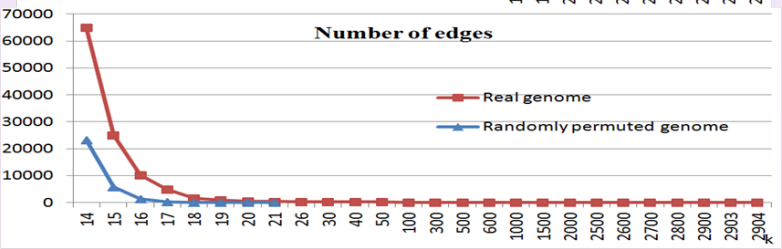


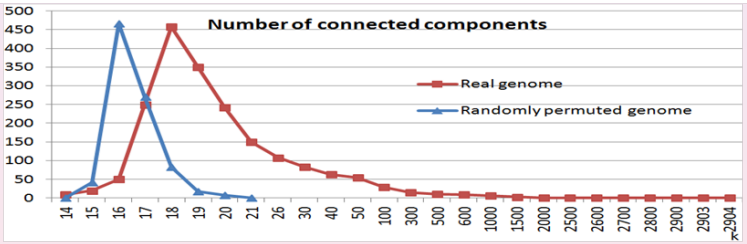


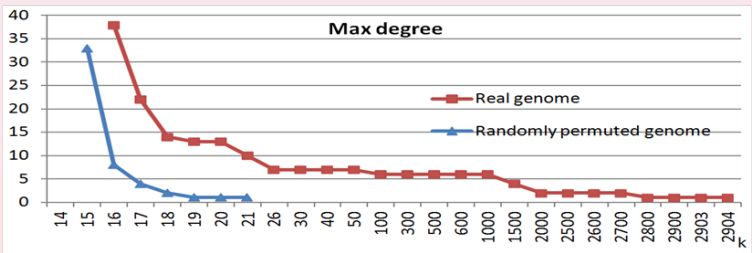


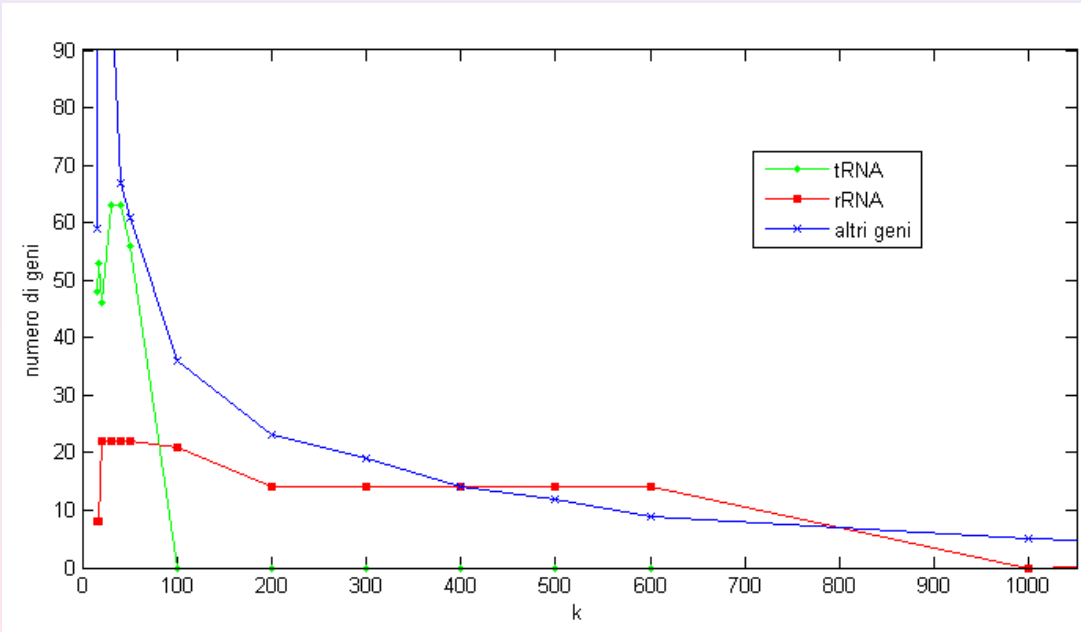
Basic E.coli network analysis





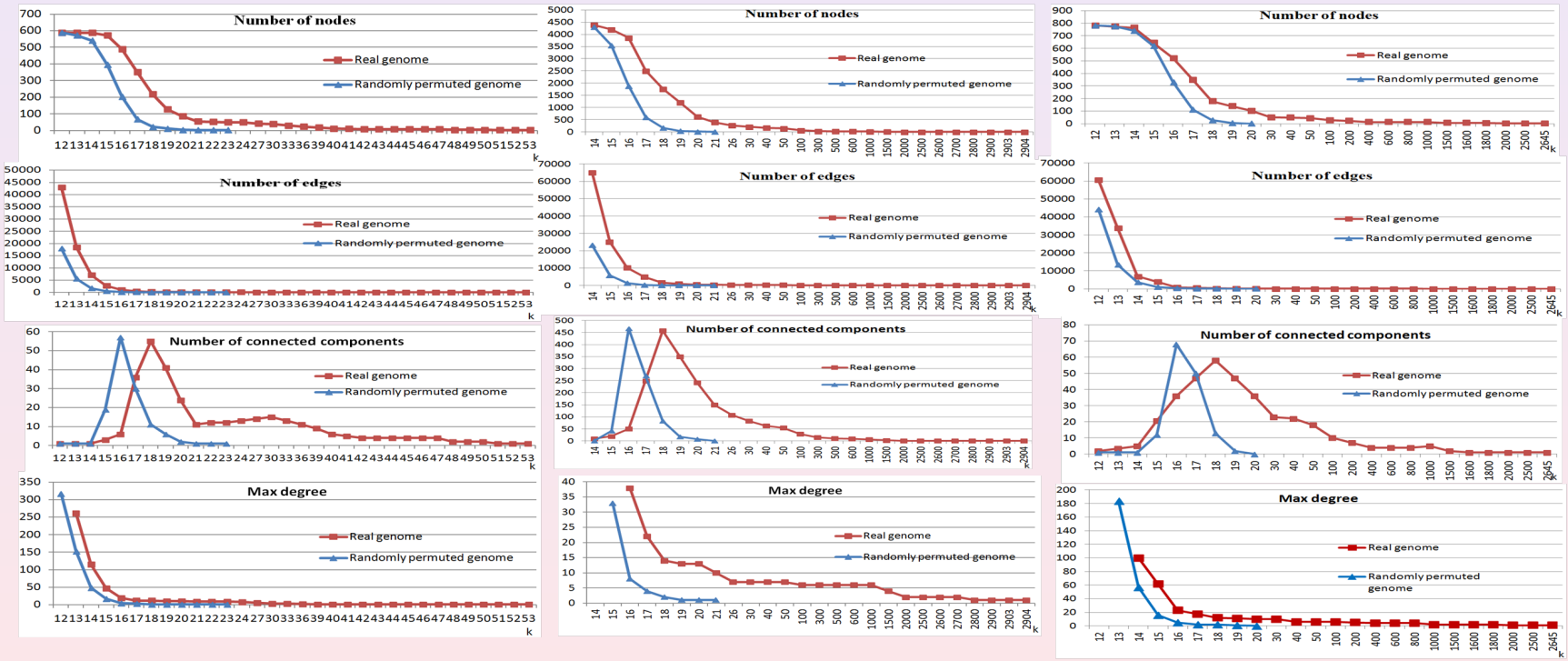






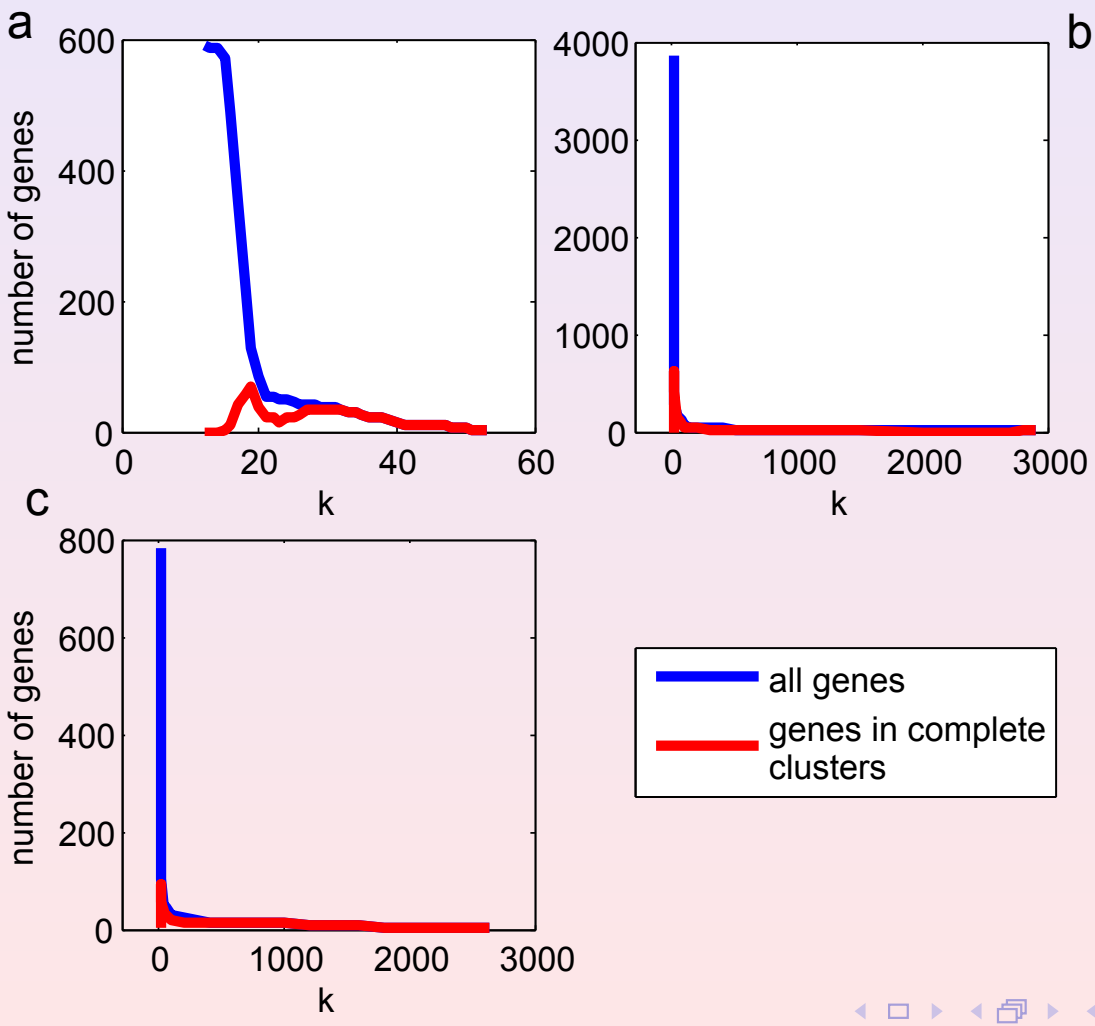
Similar network patterns

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

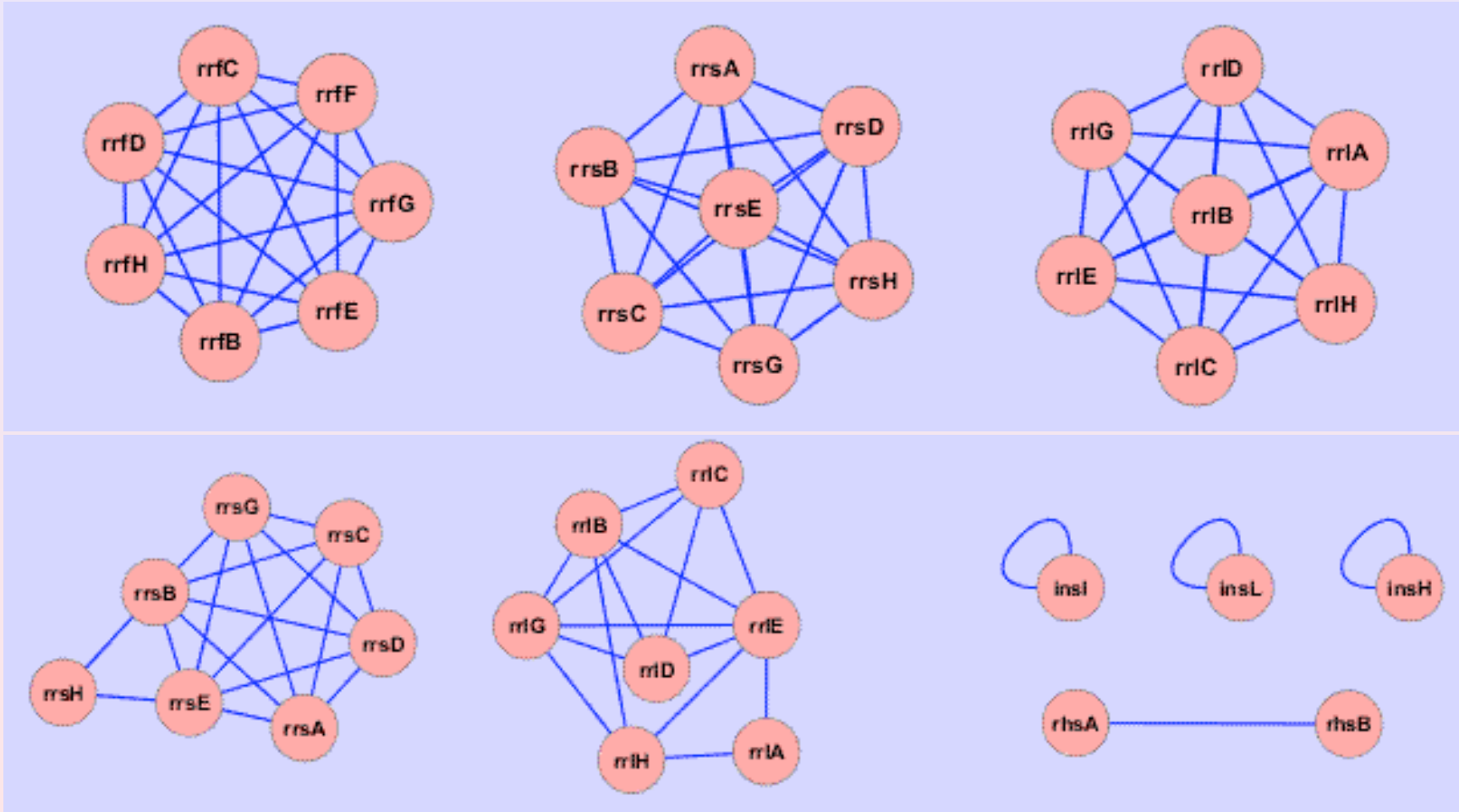


Complete clusters (N. equitans, E. coli, S. cerevisiae)

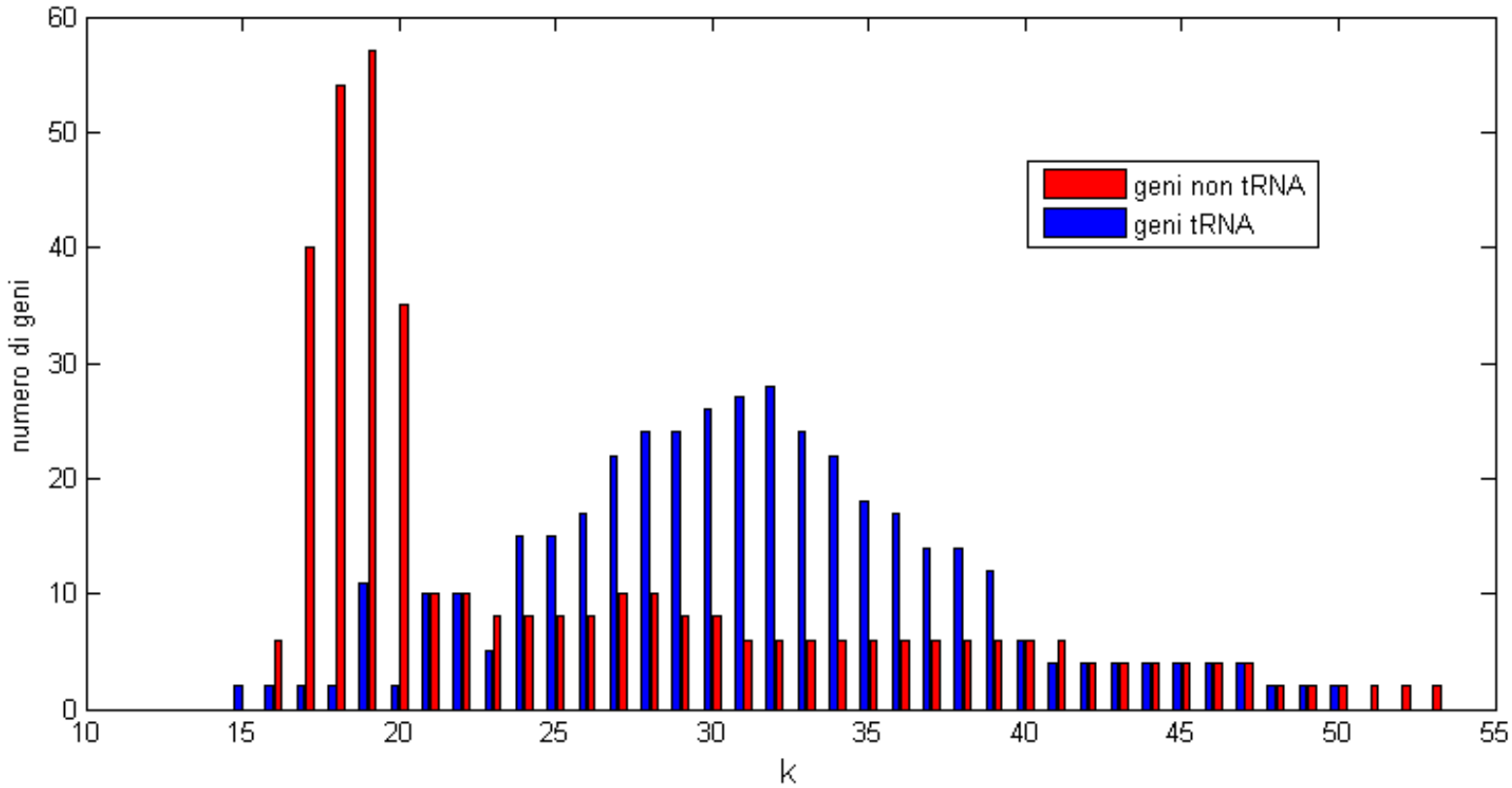
For longer repeats, gene networks tend to aggregate in cliques.



Examples of cliques (in E. coli)

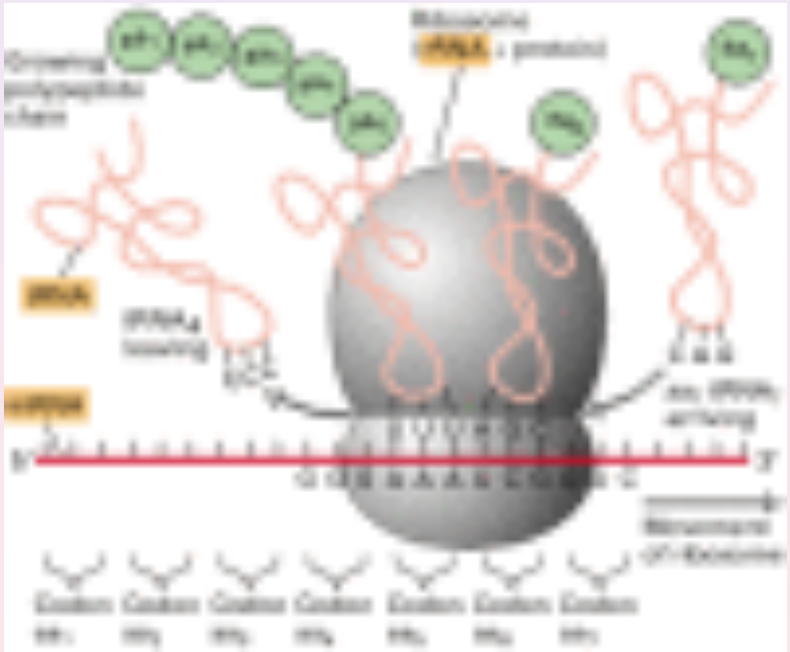


Clique amount (N. equitans)



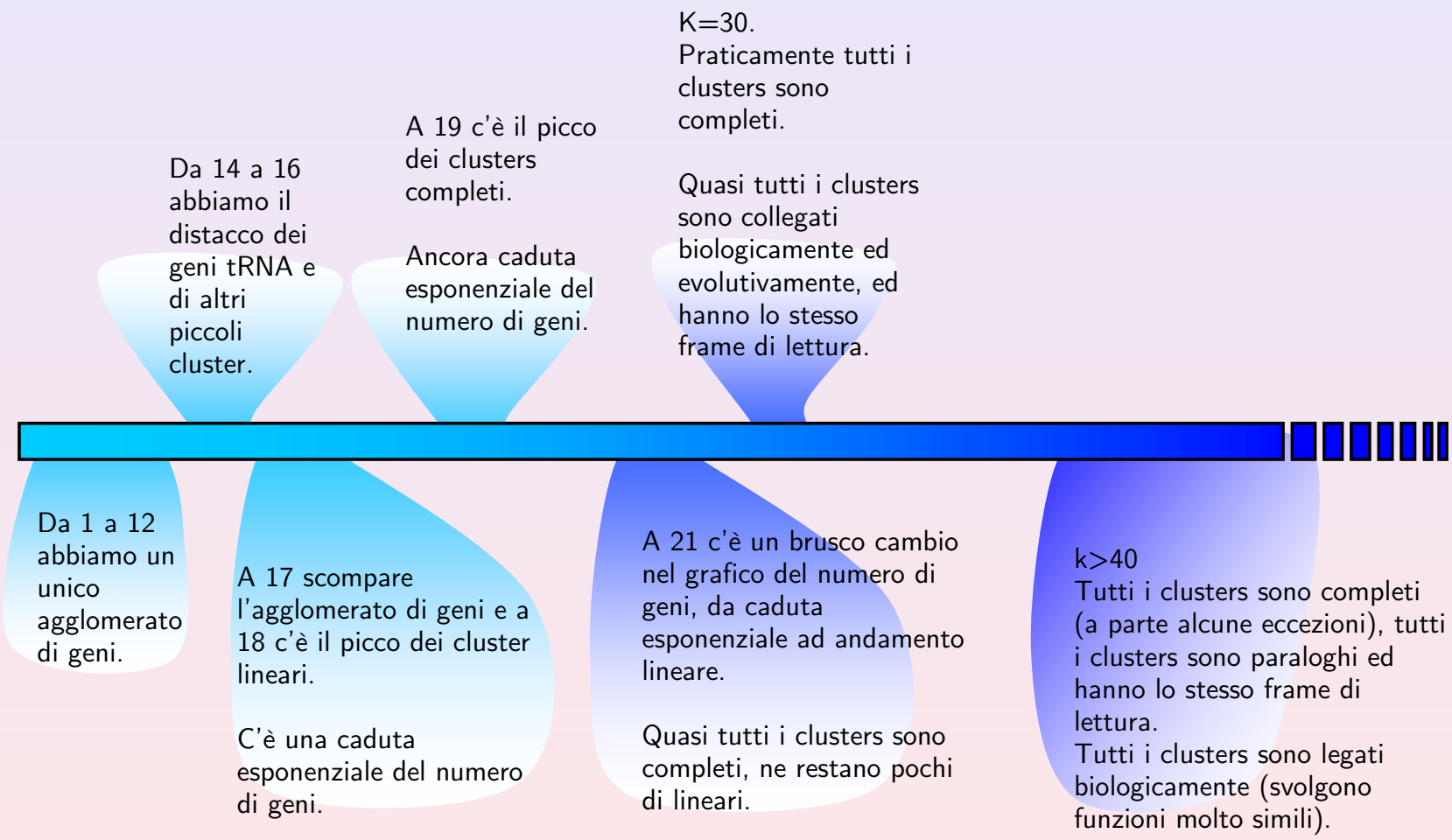
Genetic code

A code: function from T = {RNA triplets} into the set S of 21 aminoacids. Genetic code (codons $\rightarrow$ aminoacids) is degenerate, actually a very redundant fixed length code [12].



[12] Achaz et al. Origin and fate of repeats in bacteria. NAR 30 (13):2987-2994, 2002.

Clique analysis (by A. Milanese)



Repeat-wise clique analysis (A. Milanese)

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

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

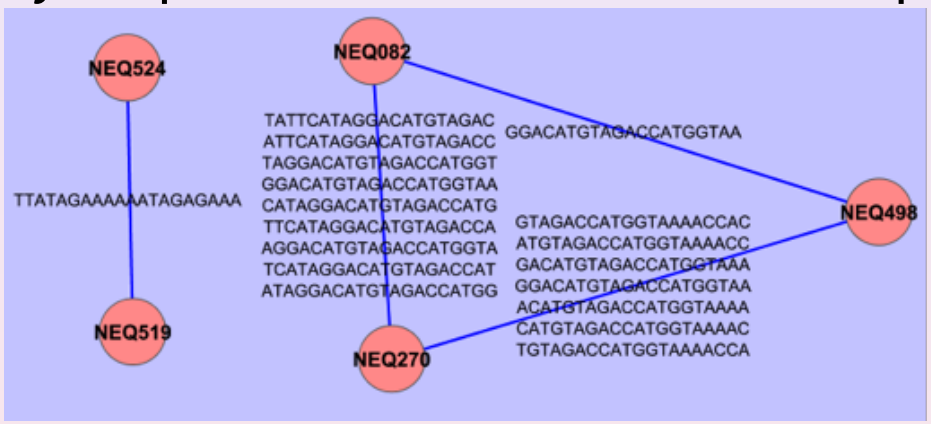
$k = 13, \ldots, 20$: some repeats are present only in couples of genes, only few have a biological role (14-15: first repeats present even in non-tRNA genes)

$k > 21$: 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.

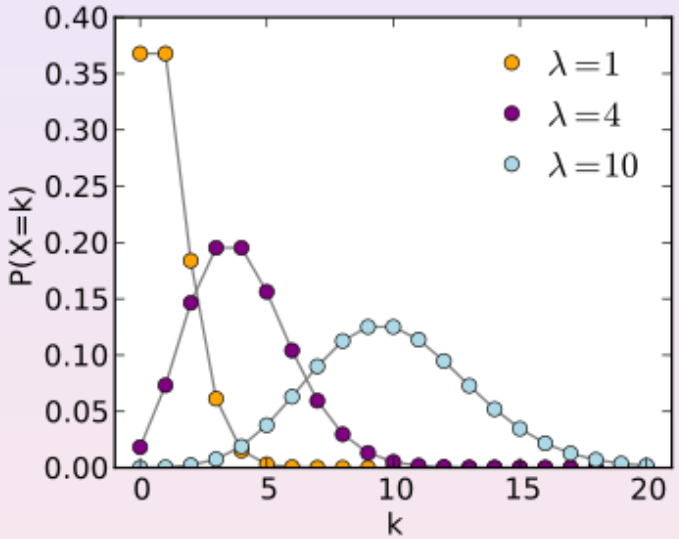
Segment multiplicity and recurrent words

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

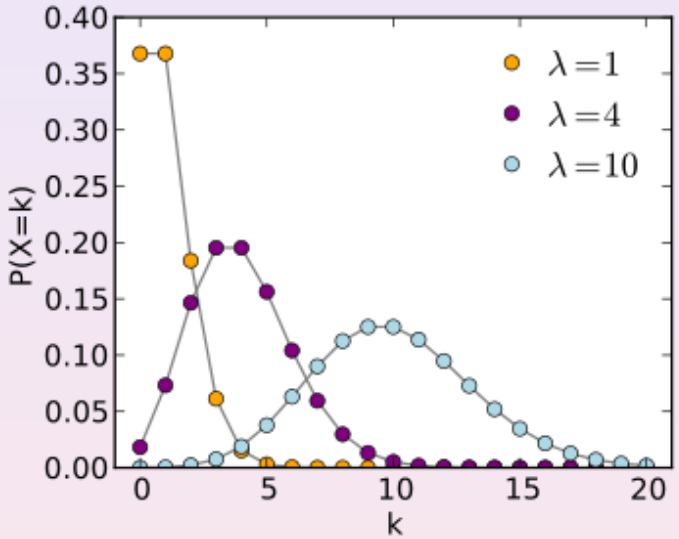
In a Bernoullian genome, such distribution (normalized, with the total num of segments) of word frequences follows a Poisson prob distribution (of a certain variance λ): $Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$



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.

Bernoullian (random) genomes

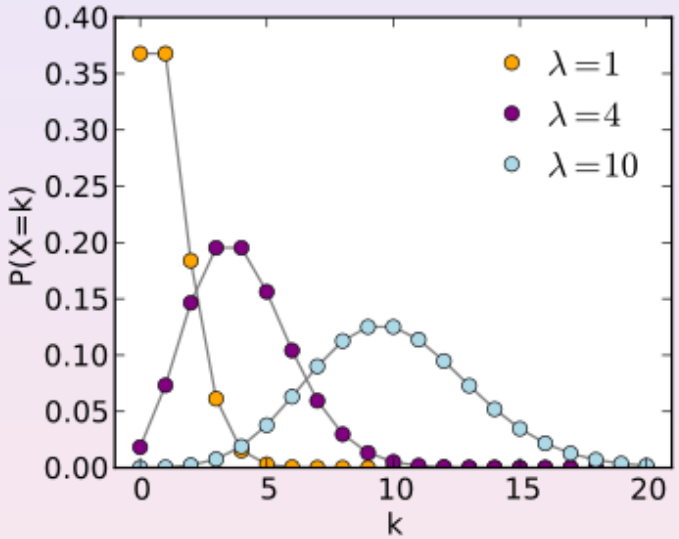
In a Bernoullian genome, such distribution (normalized, with the total num of segments) of word frequences follows a Poisson prob distribution (of a certain variance λ): $Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$



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.

Bernoullian (random) genomes

In a Bernoullian genome, such distribution (normalized, with the total num of segments) of word frequences 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.

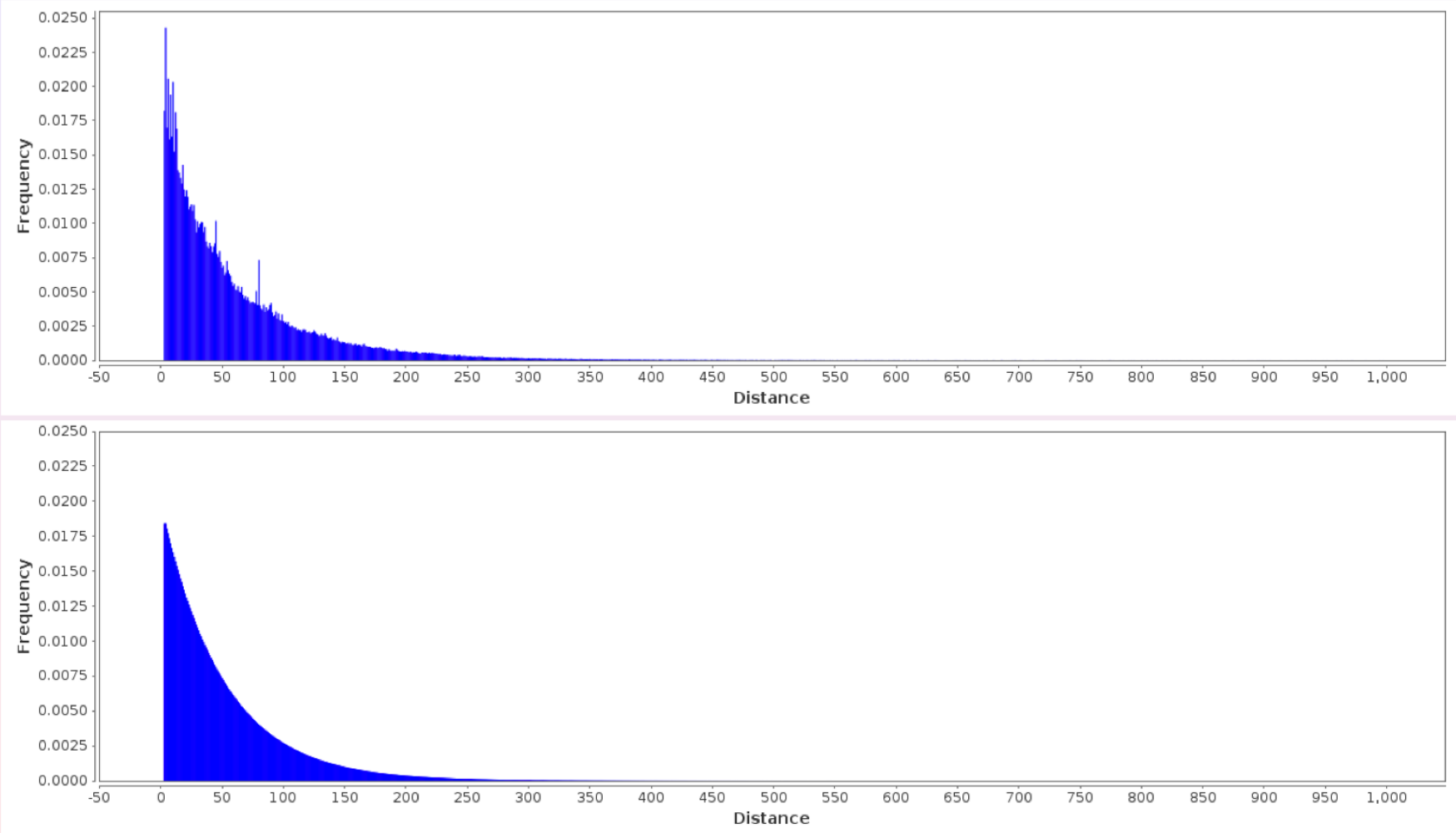
Tandem repeats investigation

RDD (recurrence distance distribution)[13]: 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.

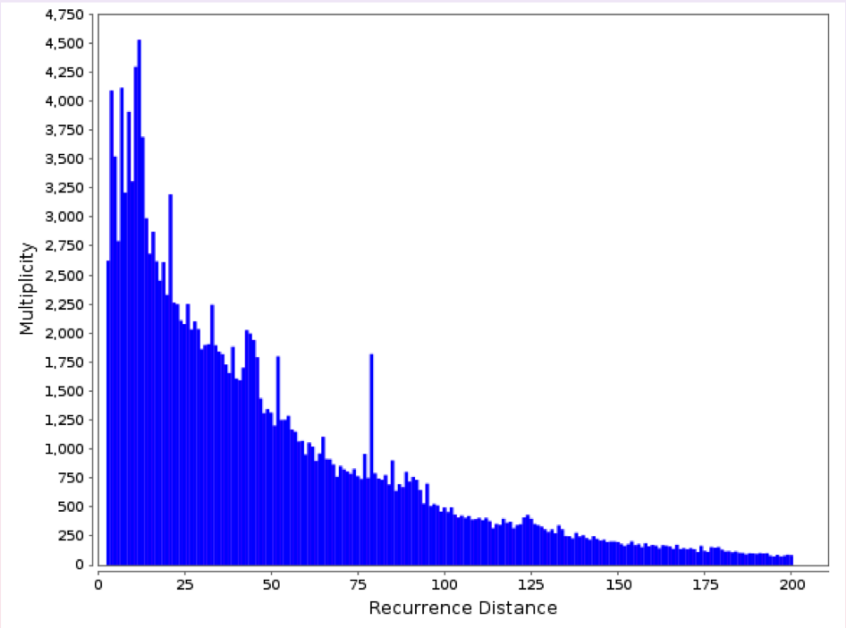
If *x* is the distance between two occurrences of a given string, $P(x) = f(x, h)$ is the number of times two occurrences appear at distance *x*. The probability $RDDG(\alpha)$ of occurring at distance x is the ratio between the number of times it recurs at distance x and the multiplicity of α in G.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

[13] V. Bonnici et al. American J. Bioinf and Comp Biol 2015.

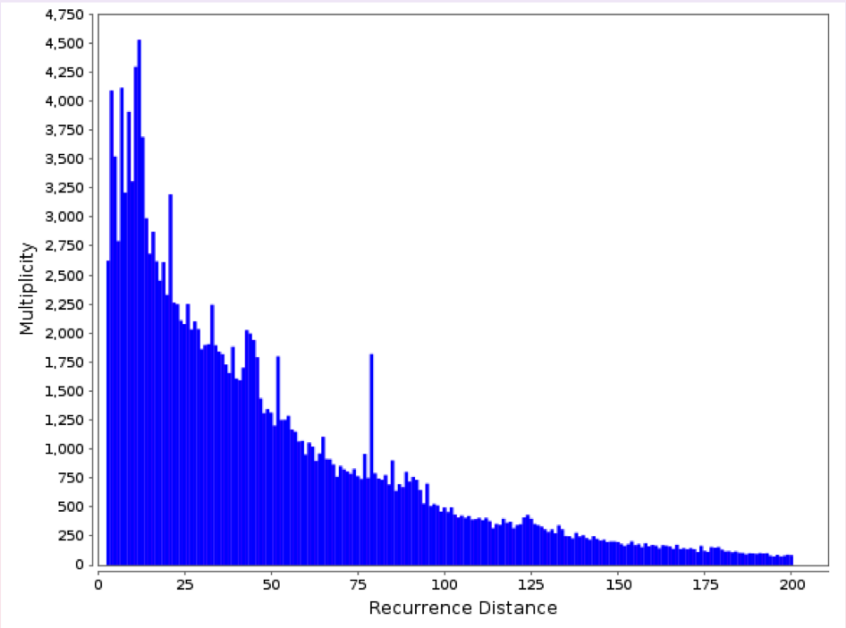
ATG in human chr 22 vs extimated Exp Distr



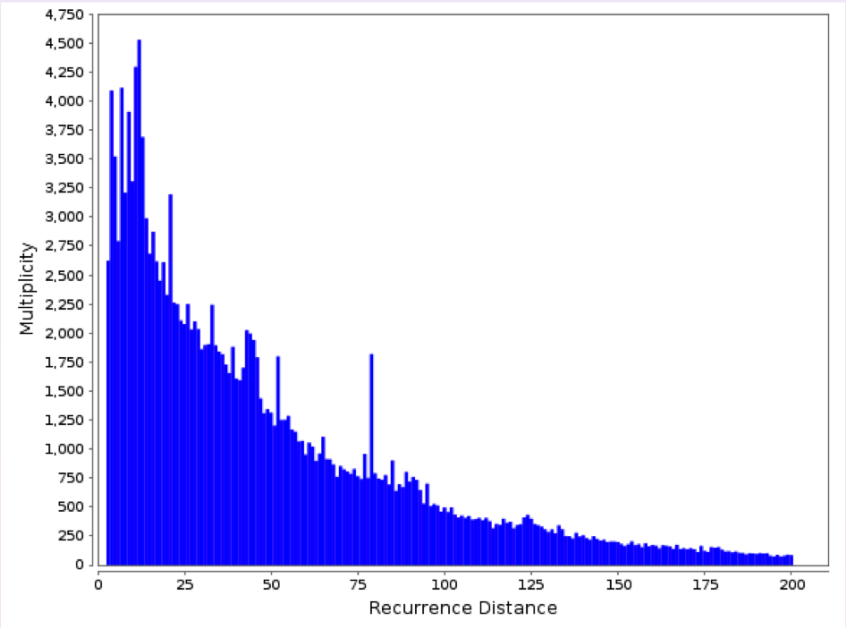
ATG in human chr 22 and sequences at distance 81



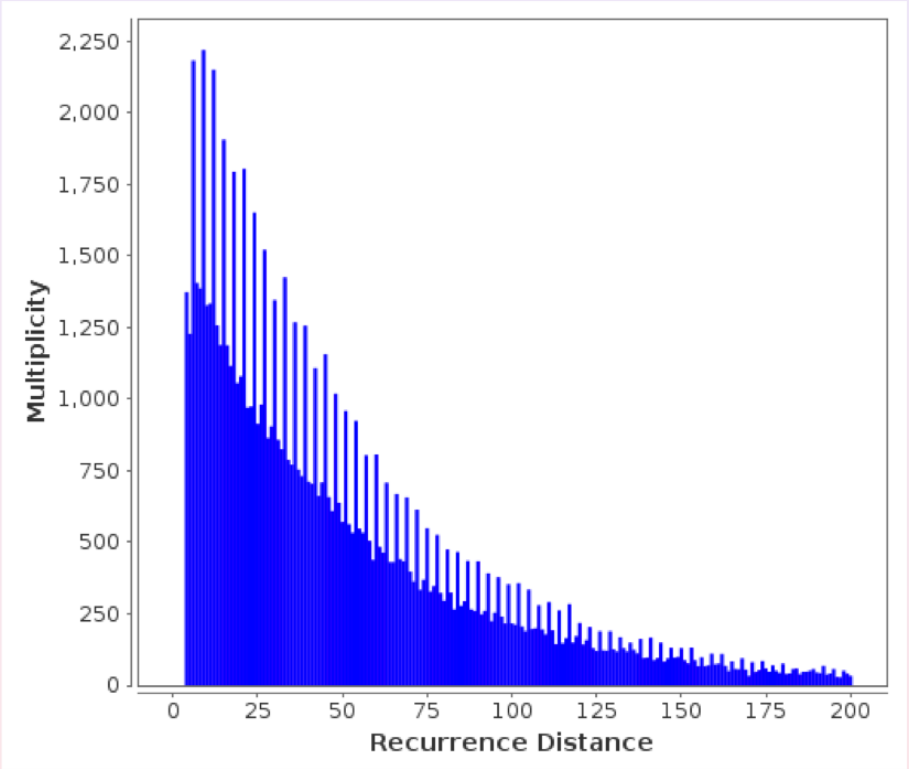
ATG in human chr 22 and sequences at distance 81

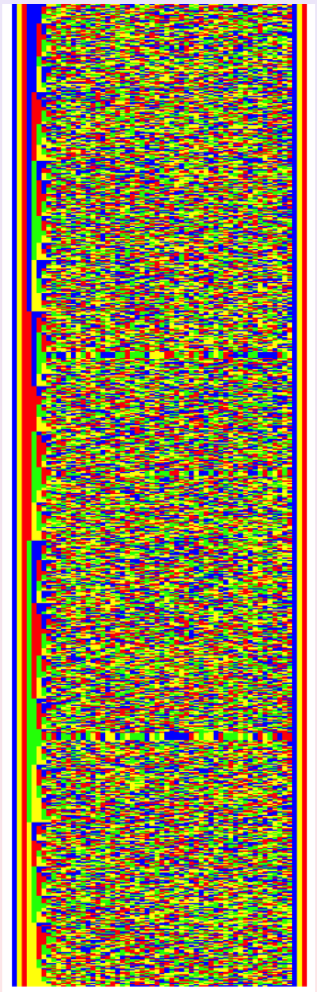


ATG in human chr 22 and sequences at distance 81

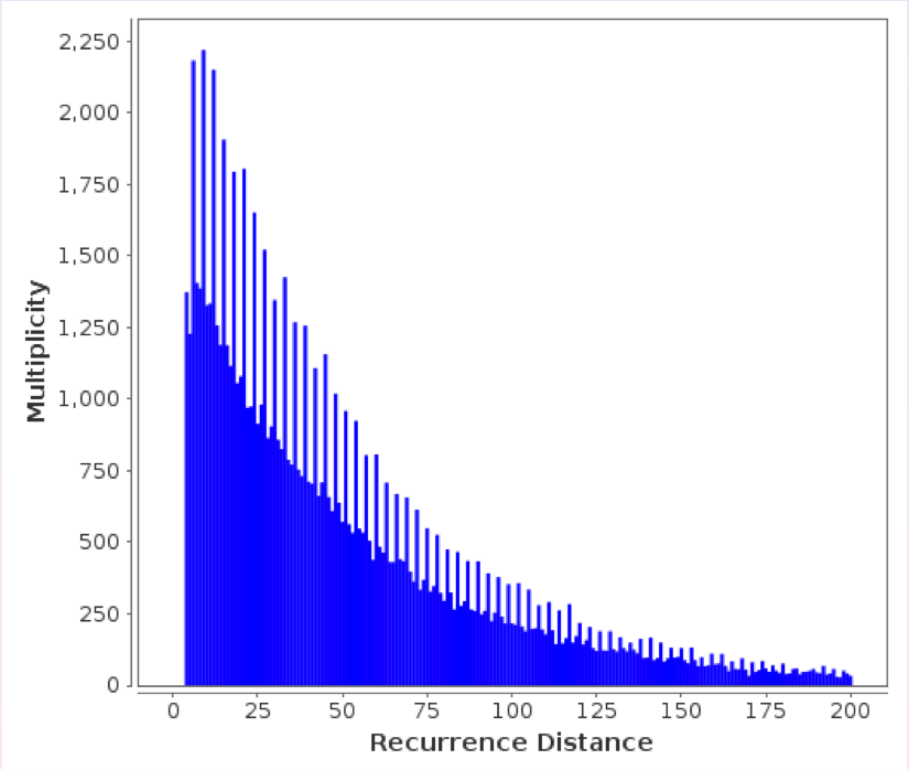


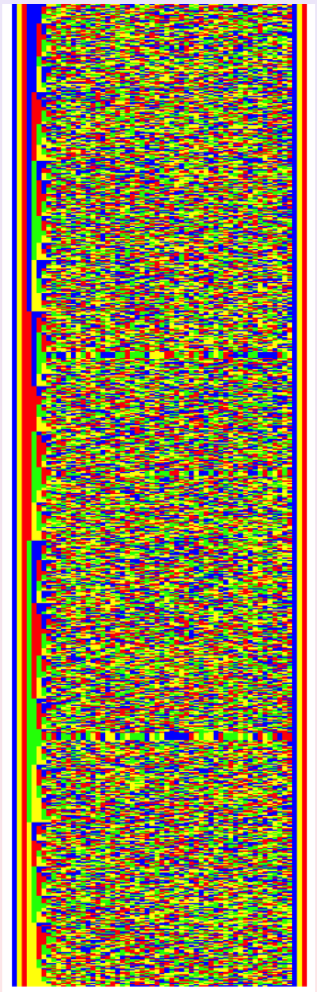
ATC in E. coli: peacks with non-repetitive elements



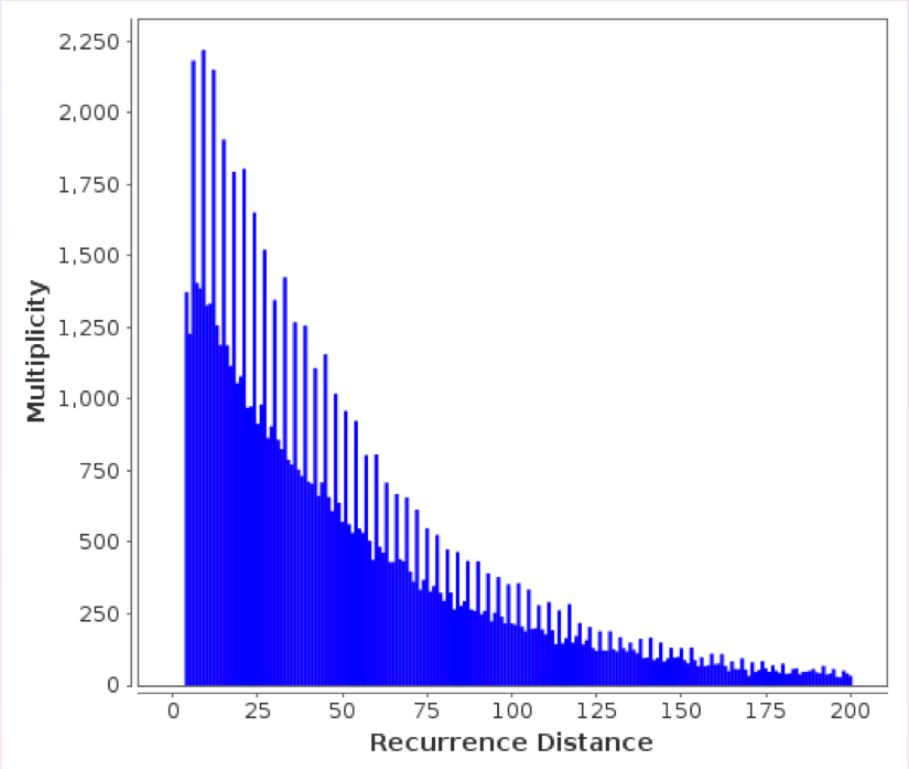


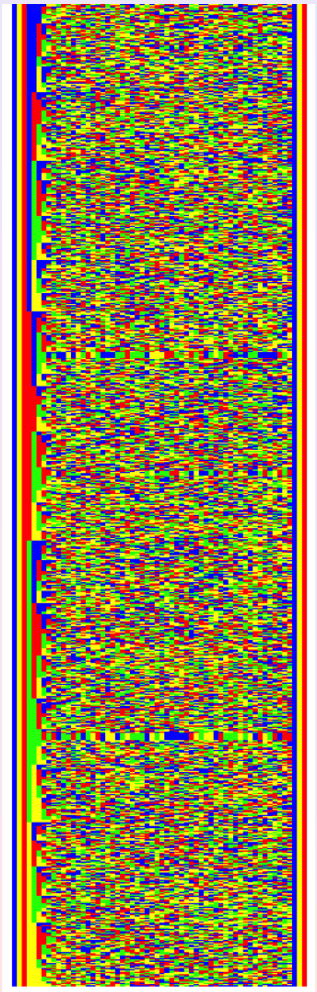
ATC in E. coli: peacks with non-repetitive elements



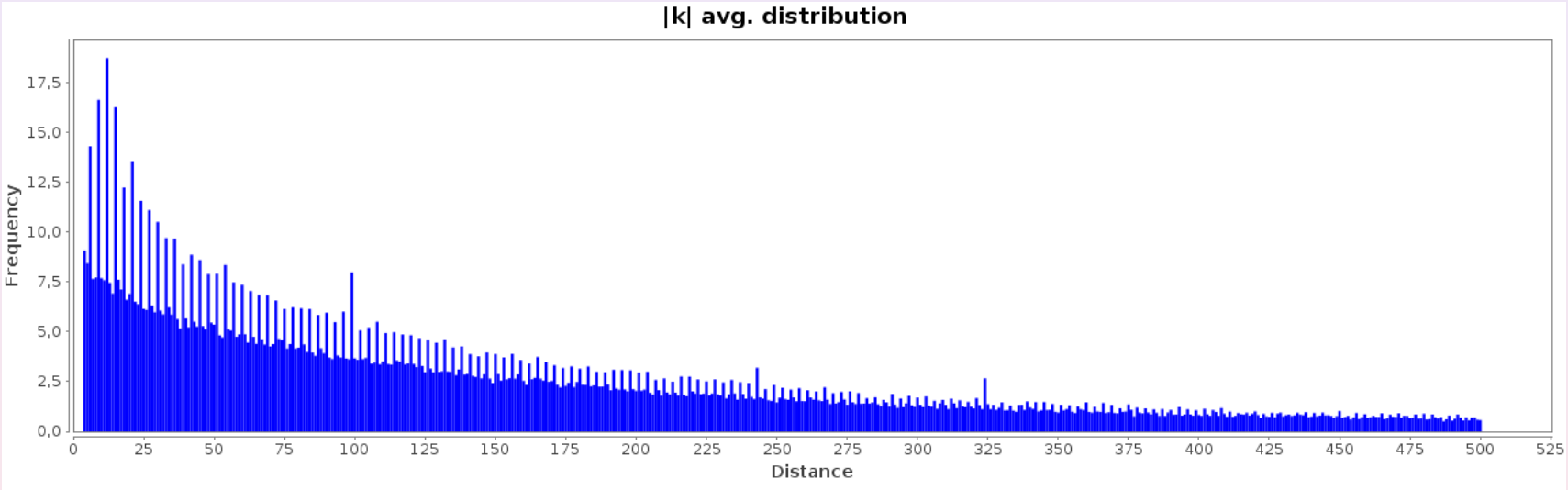


ATC in E. coli: peacks with non-repetitive elements



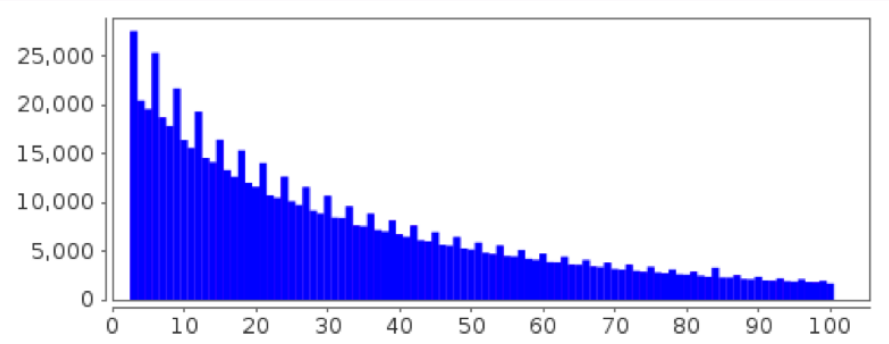


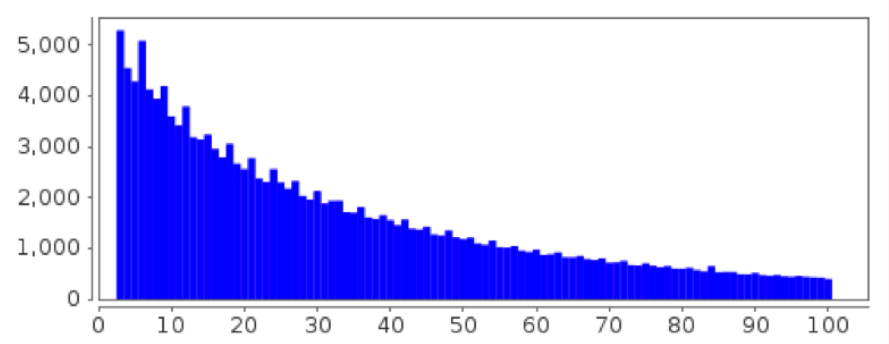
Average RDD for k=4, Emiliana huxley virus86



RDD in exonic regions, k=3

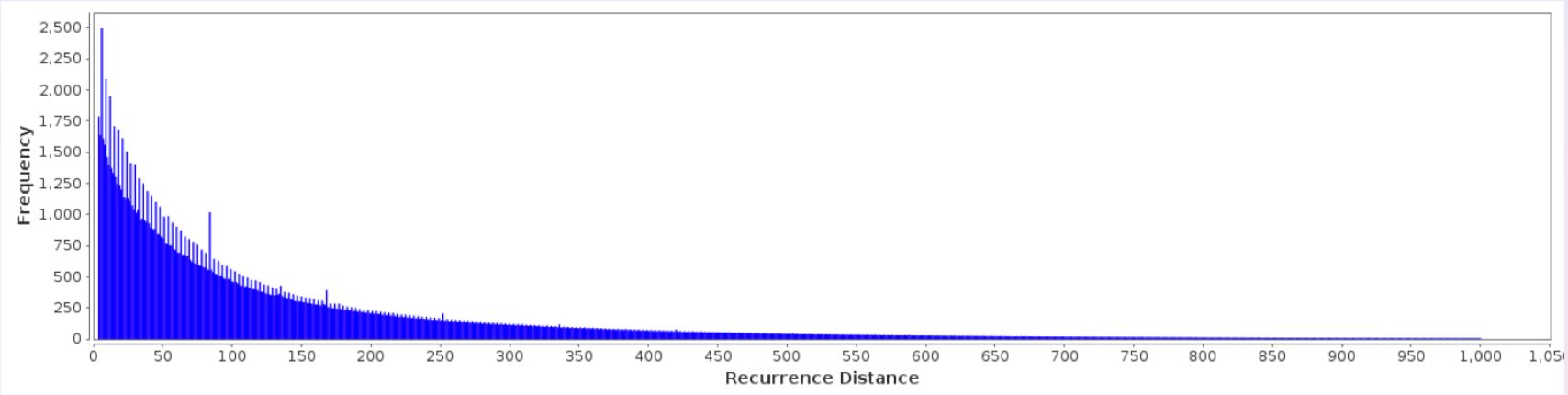
Peacks at distance 3 disappear in transcripts (introns + exons) - they are a *codonic language*



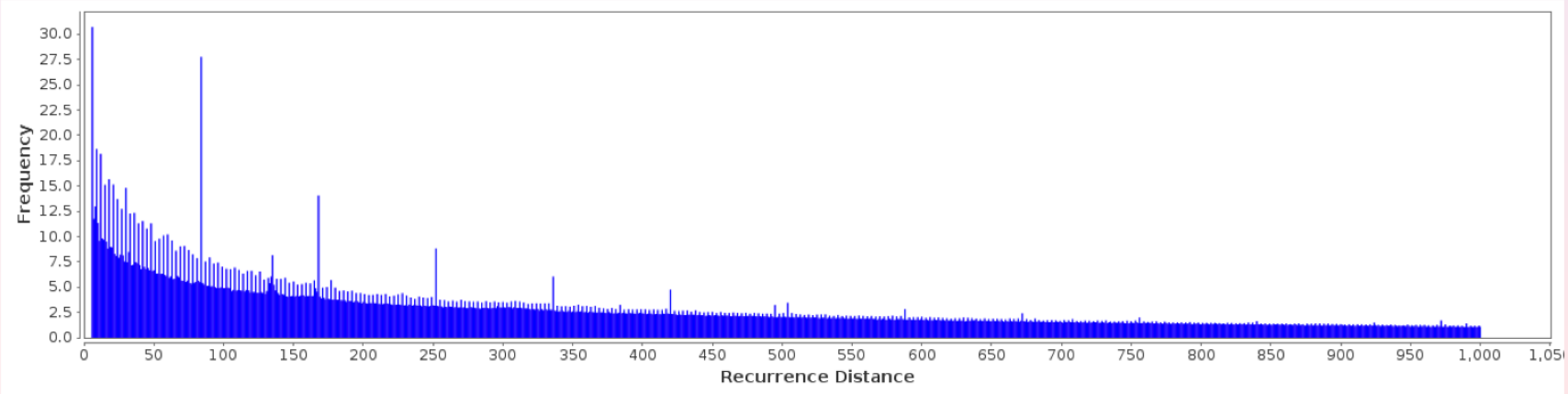


Repetitive elements in ex. regions distance multiple 84

k=4



k=6

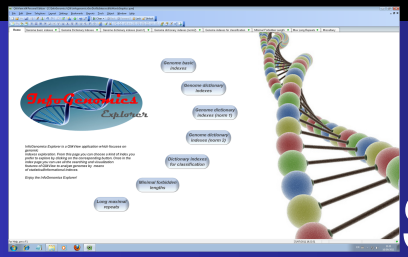


Grouped occurrences

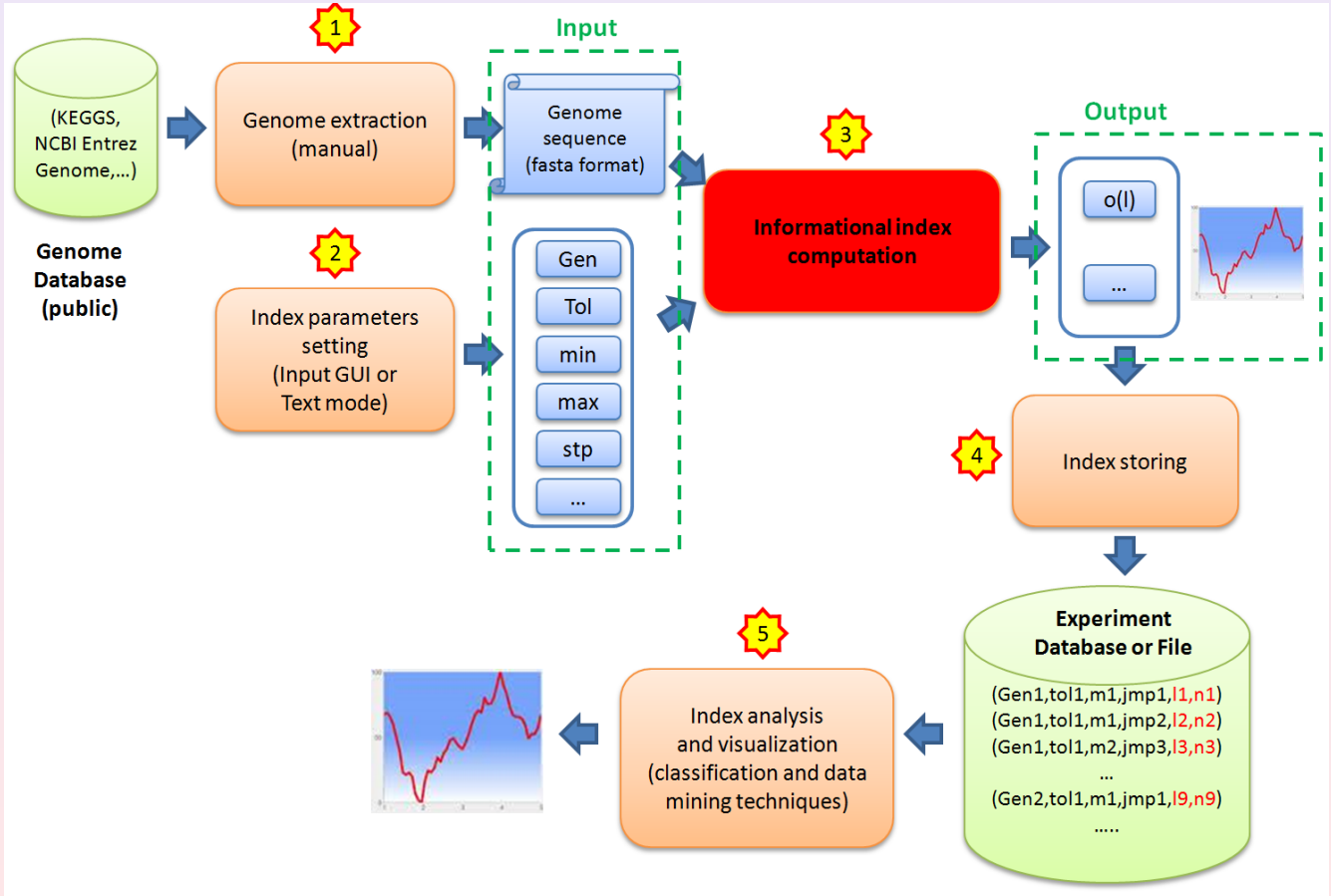
recurrence is a peculiar feature of words

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

RDD(α) is a measure to see the non-randomicity of the word α distribution
Divergence of Kulback-Leibler is computed to measure the distance between real (RDD) and *random* (exponential) genomes in terms of (average) k-mers distribution, and it is used to extract significant variable length words for a good alphabet.

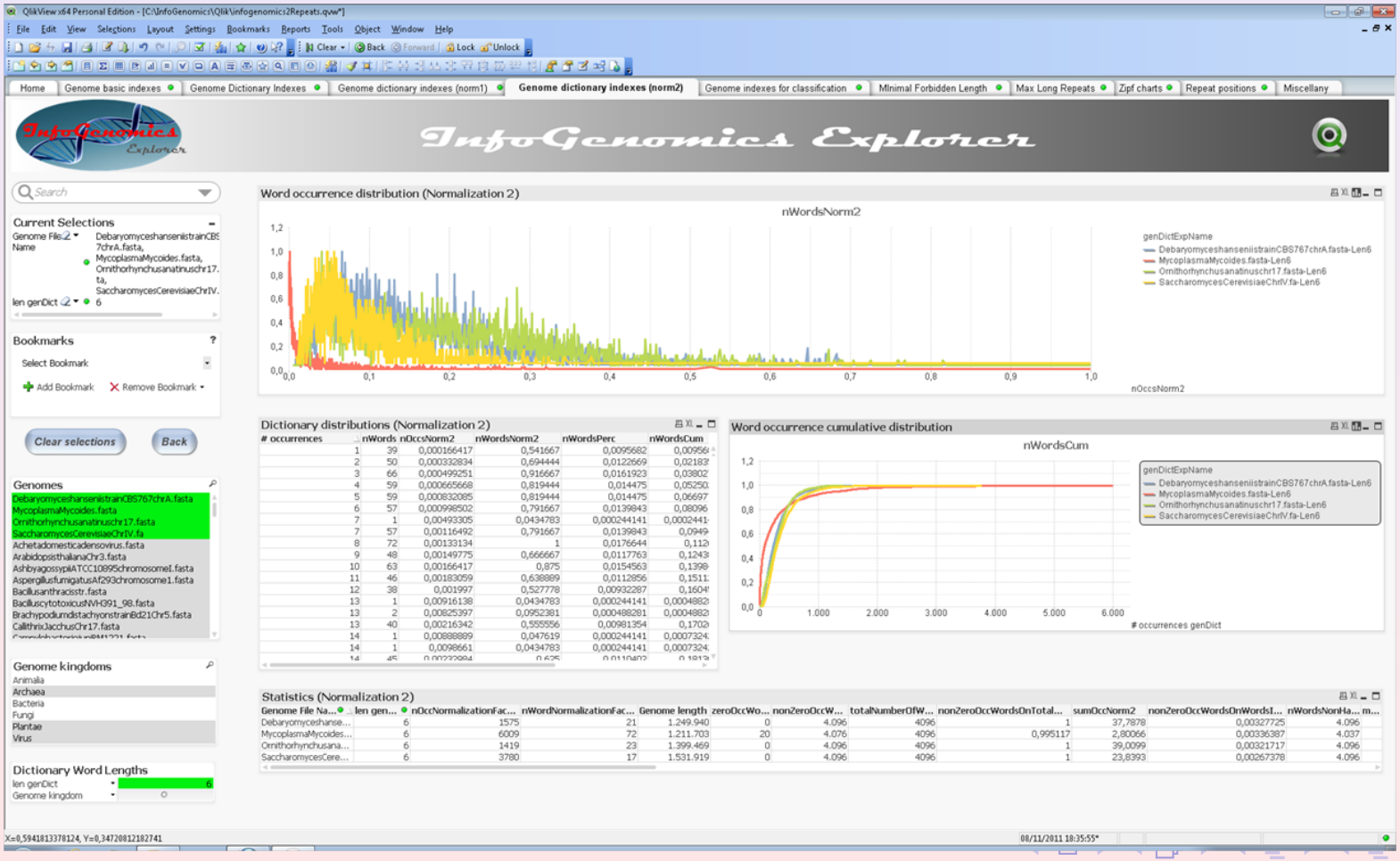


Software for massive computations



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, Zipf diagrams, 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.

Conclusions

A method

- to represent and compare genomes (genomic profiles, Zipf diagrams, 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.

Conclusions

A method

- to represent and compare genomes (genomic profiles, Zipf diagrams, 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.

Conclusions

A method

- to represent and compare genomes (genomic profiles, Zipf diagrams, 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.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.

Related References


V. Brendel and H. Busse (1984). Genome structure described by formal languages, Nucleic Acids Research, 12(94):2561–2568.


B. Chor and D. Horn and N. Goldman and Y. Levy and T. Massingham (2009). Genomic DNA k-mer spectra: models and modalities, Genome Biology, 10: R108.


Y. Fofanov, et al. (2004). How independent are the appearances of n-mers in different genomes?, Bioinformatics 20(15):2421–2428.


G. Marçais, C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics 27(6):764-70, 2011.


D. B. Searls (2002), The language of genes, Nature, 420: 211–217.


S. Vinga, J. Almeida (2003). Alignment-free sequence comparison - a review, Bioinformatics 19(4):513–523.


N. Whiteford, et al. (2008). Visualising the repeat structure of genomic sequences. Complex Systems, 17(4):381–398.


M. Lynch (2002). The Origins of Genome Architecture, Sinauer Associates Inc.


J. K. Percus (2007). Mathematics of Genome Analysis, Cambridge Studies in Mathematical Biology.