Similarity distances and entropy methods: an application to biological signals and authorship attribution

Dipartimento di Matematica, Università di Pisa

M. Degli Esposti

desposti@dm.unibo.it

http://www.dm.unibo.it/˜desposti/

Dipartimento di Matematica
Università di Bologna

Pisa, 31 Gennaio 2008

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

1. Few ”Toy” problems
2. The Methods
   - Alphabets, Strings and Distances
   - Universal distances based on Kolmogorov Complexity
   - L-Z complexity, exhaustive parsing and the Similarity Metrics $d_{LZ}(\cdot)$
   - The N-Gram distance $d_G(\cdot)$
   - Kullbach-Leibler divergence and heuristic indicators
   - The Burrows-Wheeler Transform (BWT)
3. Applications and Results
   - Phylogenetic Tree Construction
   - Cardiac Sequences and HRV
   - Authorship Attribution
   - Una struttura matematica per i testi ?

ECG clustering

- ECG sequences (after a suitable coding): can we recognize and discriminate between different *pathologies* or *ages* of given ECG signals ?

Experimental Data

Data Set 1: **nk** v.s. **gk**

nk group made of 90 patients from the Department of Cardiology of Medical University in Gdańsk, Poland (9 women, 81 men, the average age is 57 $\pm$10) in whom the reduced left ventricular systolic function was recognized by echocardiogram.

gk group made of 40 healthy individuals (4 women, 36 men, the average age is 52 $\pm$ 8) without past history of cardiovascular disease, with both echocardiogram and electrocardiogram in normal range.

Experimental Data

Data Set 2: **young** v.s. **old**

old group 13 healthy subject belonging to **gk** previously described.

young group 13 healthy and rather young people (age between 20-40 years). These patients (3 men, 10 women) show no significant arrhythmias.

Data Set 3: **NYHA >Classification**, 20 patients distributed among the 4 NYHA classes

Experimental Data

Data Set 2: **young** v.s. **old**

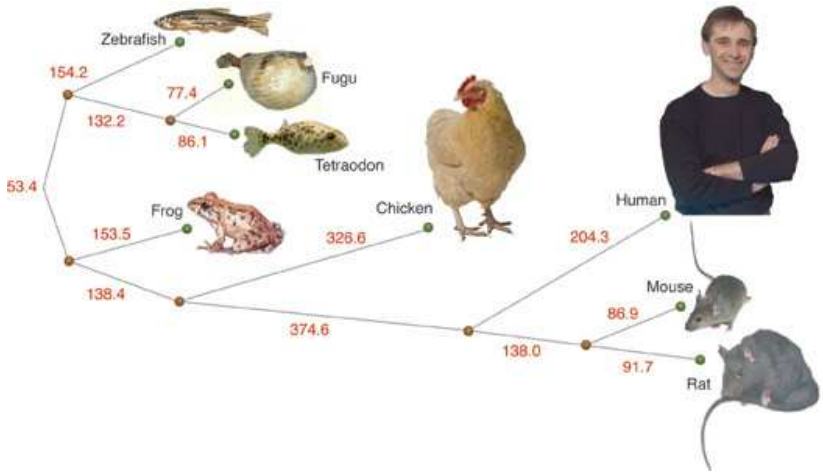
old group 13 healthy subject belonging to **gk** previously described.

young group 13 healthy and rather young people (age between 20-40 years). These patients (3 men, 10 women) show no significant arrhythmias.

Data Set 3: **NYHA >Classification**, 20 patients distributed among the 4 NYHA classes

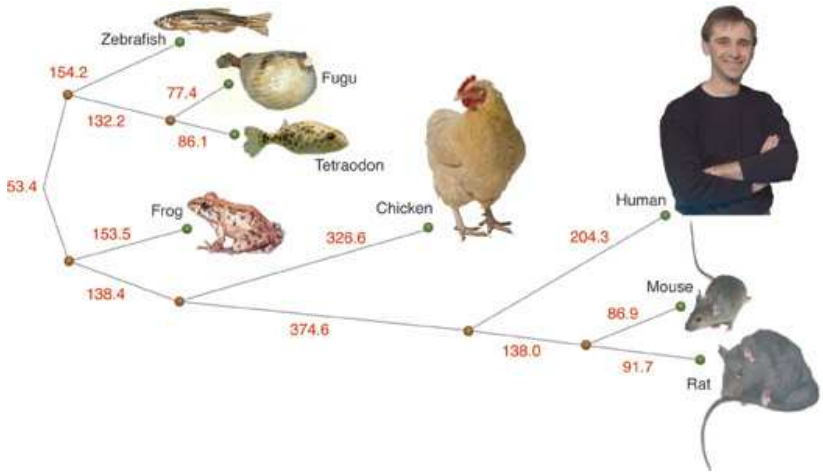
Genome Phylogeny Problem

- DNA sequences , $\mathcal{A} = \{A, C, G, T\}$: can we reconstruct phylogenetic trees using an alignment free distance d to measure the similarities between different genetic sequences (either single genes or complete genoma sequences) ?



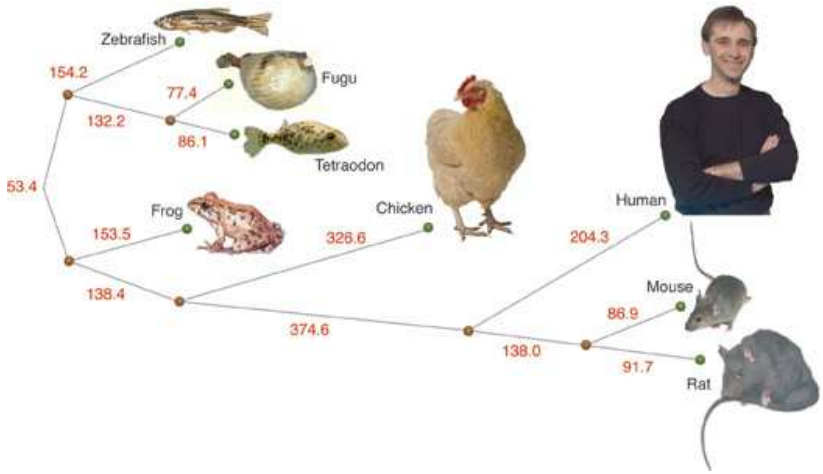
Genome Phylogeny Problem

- DNA sequences , $\mathcal{A} = \{A, C, G, T\}$: can we reconstruct phylogenetic trees using an alignment free distance d to measure the similarities between different genetic sequences (either single genes or complete genoma sequences) ?



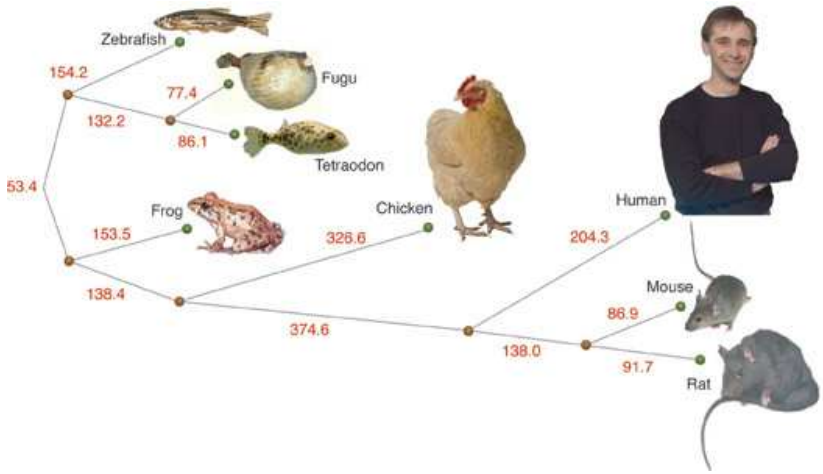
Genome Phylogeny Problem

- DNA sequences , $\mathcal{A} = \{A, C, G, T\}$: can we reconstruct phylogenetic trees using an alignment free distance d to measure the similarities between different genetic sequences (either single genes or complete genoma sequences) ?



Genome Phylogeny Problem

- DNA sequences , $\mathcal{A} = \{A, C, G, T\}$: can we reconstruct phylogenetic trees using an alignment free distance d to measure the similarities between different genetic sequences (either single genes or complete genoma sequences) ?

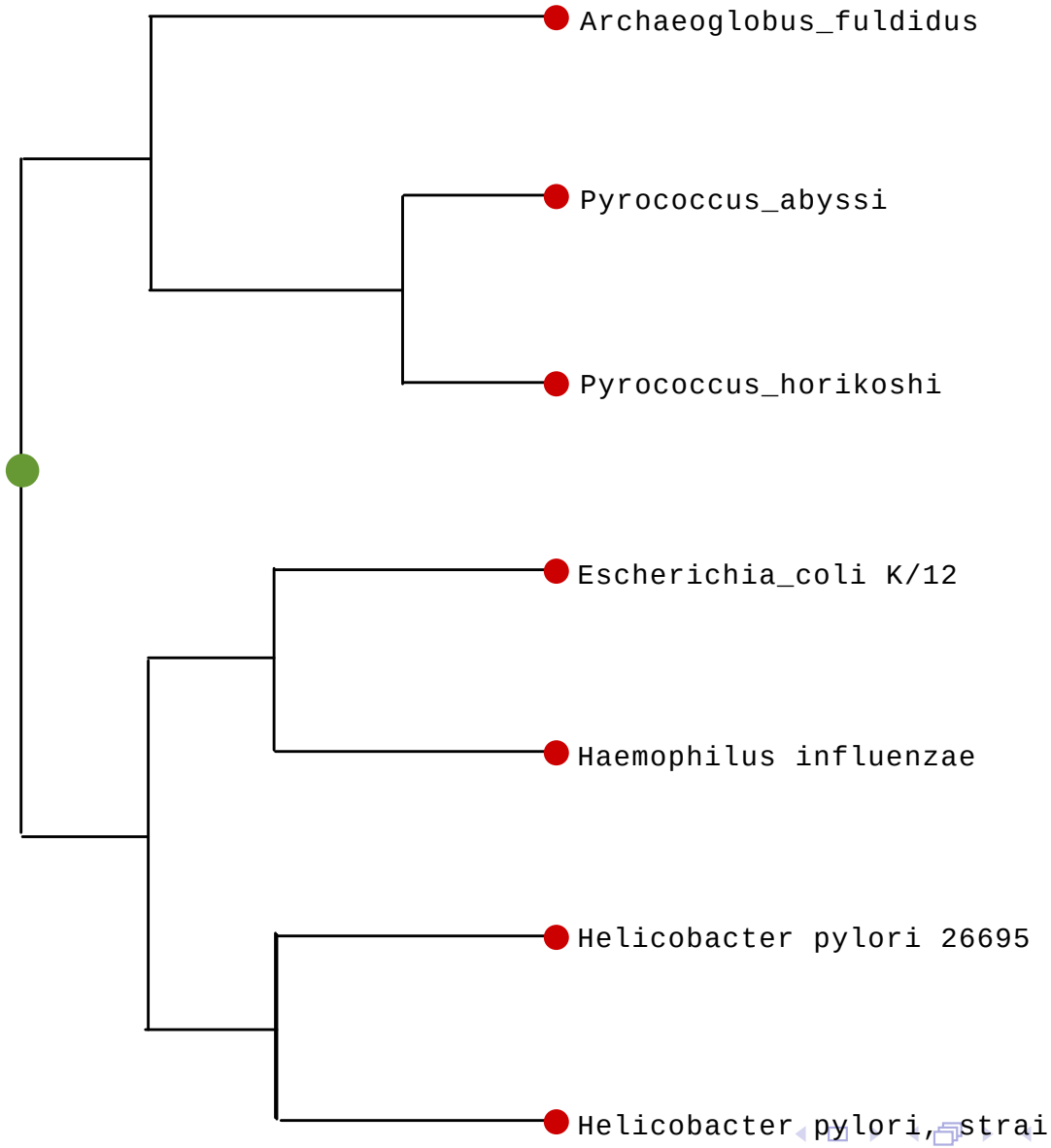


Complete Genoma

Archaea Bacteria *Archaeoglobus fulgidus*,*Pyrococcus abyssi* and *Pyrococcus horikoshii* OT3

Bacteria *Escherichia coli* K-12 MG1655, *Haemophilus influenzae* Rd, *Helicobacter pylori* 26695 and *Helicobacter pylori*, strain J99

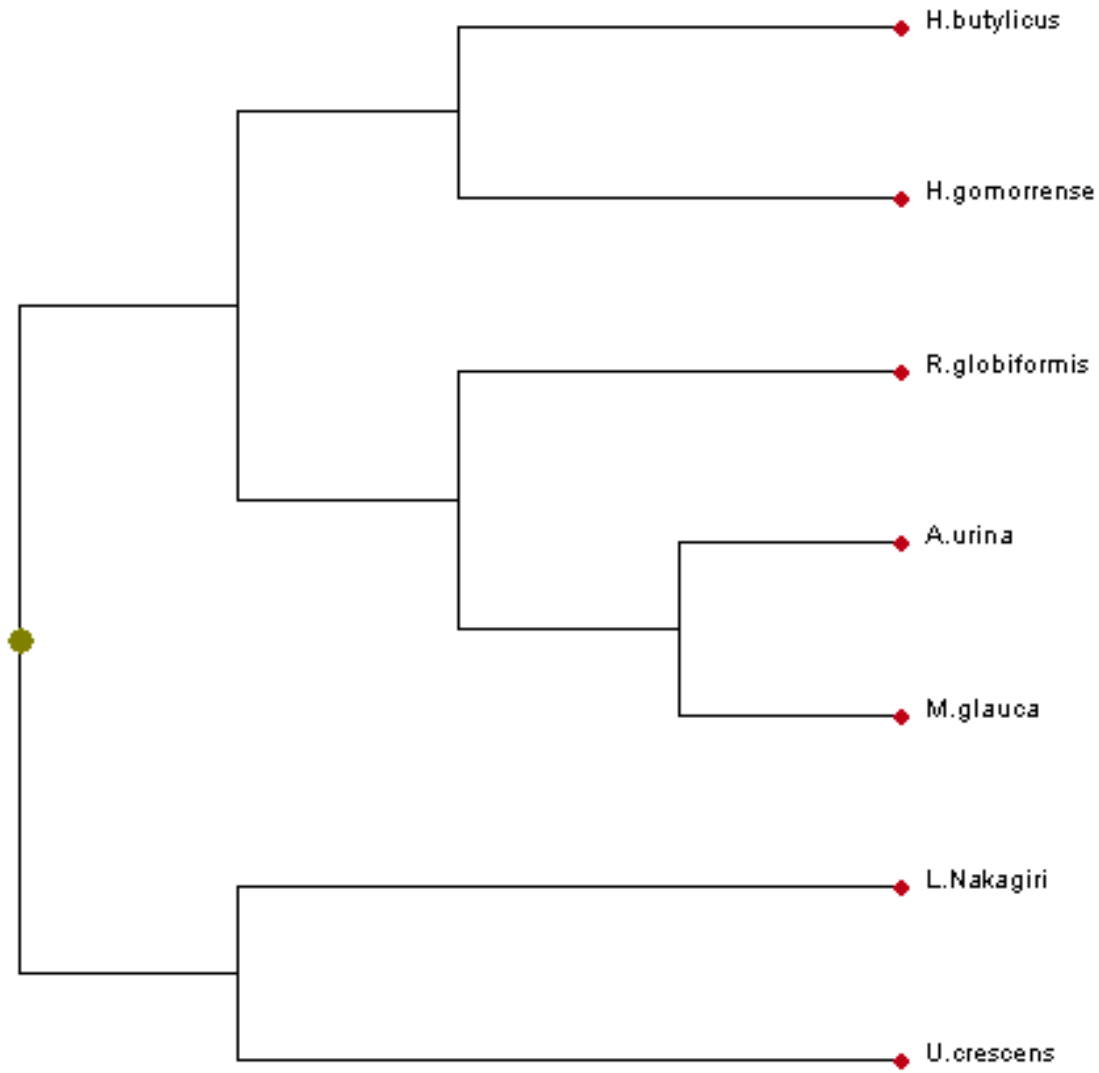
Complete Genoma



rRNA single Genes

Archaebacteria	*H. butylicus* and *Halobaculum gomorrense*
Eubacteria	*Aerococcus urina*, *M. glauca* strain B1448-1 and *Rhodopila globiformis*
Eukaryotes	*Urosporidium crescens*, *Labyrinthula sp. Nakagiri*

rRNA single Genes



Information and Protein

- Proteins: could we detect *different levels of similarities* (e.g. topology, functional similarity, homology...) either from the *primary aminoacid sequence* or from the *contact maps* ?

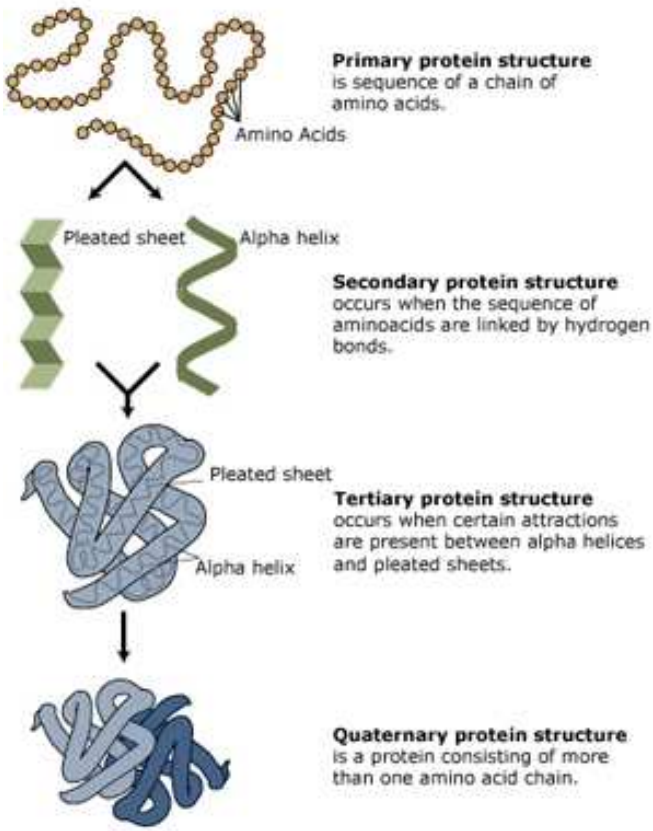


Image adapted from: National Human Genome Research Institute.

Authorship Attribution

- can we recognize the *style* of a writer ?

common scenario in Authorship Attribution

1. Consider n literary authors $A_1, A_2, \ldots, A_n$
2. For each authors A_k, assume we have a certain number (m_k) of texts $T_k(1), T_k(2), \ldots, T_k(m_k)$
3. Let now X be an unknown text: we assume X has been written from one of the author, but X is NOT contained in the reference set. The problem is to recognize, using quantitative methods, the author of the text X.

common scenario in Authorship Attribution

1. Consider n literary authors $A_1, A_2, \ldots, A_n$
2. For each authors A_k, assume we have a certain number (m_k) of texts $T_k(1), T_k(2), \ldots, T_k(m_k)$
3. Let now X be an unknown text: we assume X has been written from one of the author, but X is NOT contained in the reference set. The problem is to recognize, using quantitative methods, the author of the text X.

common scenario in Authorship Attribution

1. Consider n literary authors $A_1, A_2, \ldots, A_n$
2. For each authors A_k, assume we have a certain number (m_k) of texts $T_k(1), T_k(2), \ldots, T_k(m_k)$
3. Let now X be an unknown text: we assume X has been written from one of the author, but X is NOT contained in the reference set. The problem is to recognize, using quantitative methods, the author of the text X.

A real scenario....

D. Benedetto, E. Caglioti, V. Loreto “Language Tree and Zipping”, Physical Review Letters **88**, no.4 (2002)

Verga Giovanni:Eros
Verga Giovanni:Eva
Verga Giovanni: La lupa
Verga Giovanni: Tigre reale
Verga Giovanni: Tutte le novelle
Verga Giovanni: Una peccatrice
Svevo Italo: Corto viaggio sperimentale
Svevo Italo: La coscienza di Zeno
Svevo Italo: La novella del buon vecchio e ...
Svevo Italo: Senilità
Svevo Italo:Una vita
Salgari Emilio: Gli ultimi filibustieri
Salgari Emilio: I misteri della jungla nera
Salgari Emilio:I pirati della Malesia
Salgari Emilio: Il figlio del Corsaro Rosso
Salgari Emilio: Jolanda la figlia del Corsaro Nero
Salgari Emilio:Le due tigri
Salgari Emilio: Le novelle marinaresche di mastro Catrame

Tozzi Federigo: Bestie
Tozzi Federigo: Con gli occhi chiusi
Tozzi Federigo: Il podere
Tozzi Federigo: L’amore
Tozzi Federigo: Novale
Tozzi Federigo: Tre croci
Pirandello Luigi:.....
Petrarca Francesco:.....
Manzoni Alessandro:.....
Machiavelli Niccolo’:.....
Guicciardini Francesco:.....
Goldoni Carlo:.....
Fogazzaro Antonio:.....
Deledda Grazia:.....
De Sanctis Francesco:.....
De Amicis Edmondo:.....
D’Annunzio Gabriele:....
Alighieri Dante:....

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci* (2007-2008)

Another real scenario: Gramsci’s articles

A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. Gramsci...
- Quite positive results for the period 1915-1917 (!?!?)
- Joint collaboration with D. Benedetto, E. Caglioti e M. Lana, for the new *Edizione Nazionale delle Opere di Gramsci (2007-2008)*

una definizione di *stile*

STYLE:

Etymology: Middle English stile, style, from Latin stilus spike, stem, stylus, style of writing;

1 : DESIGNATION, TITLE

2 a : a distinctive manner of expression (as in writing or speech)

b : a distinctive manner or custom of behaving or conducting oneself, also : a particular mode of living

c : a particular manner or technique by which something is done, created, or performed

3 a : STYLUS b : GNOMON 1b c : the filiform usually elongated part of the pistil bearing a stigma at its apex

d : a slender elongated process (as a bristle) on an animal

4 : a distinctive quality, form, or type of something

5 a : the state of being popular : FASHION b : fashionable elegance c : beauty, grace, or ease of manner or technique

6 : a convention with respect to spelling, punctuation, capitalization, and typographic arrangement and display followed in writing or printing

una definizione di *stile*

STYLE:

Etymology: Middle English stile, style, from Latin stilus spike, stem, stylus, style of writing;

1 : DESIGNATION, TITLE

2 a : a distinctive manner of expression (as in writing or speech)

b : a distinctive manner or custom of behaving or conducting oneself, also : a particular mode of living

c : a particular manner or technique by which something is done, created, or performed

3 a : STYLUS b : GNOMON 1b c : the filiform usually elongated part of the pistil bearing a stigma at its apex

d : a slender elongated process (as a bristle) on an animal

4 : a distinctive quality, form, or type of something

5 a : the state of being popular : FASHION b : fashionable elegance c : beauty, grace, or ease of manner or technique

6 : a convention with respect to spelling, punctuation, capitalization, and typographic arrangement and display followed in writing or printing

Similarità e Stili

a Very general aim

- We aim to discuss here how some ideas and results from the area located in the intersection of the Theory of dynamical systems, la Statistical Mechanics and the Information theory can lead us towards the resolution of these problems (at least in some specific situations).

a Very general aim

- We aim to discuss here how some ideas and results from the area located in the intersection of the Theory of dynamical systems, la Statistical Mechanics and the Information theory can lead us towards the resolution of these problems (at least in some specific situations).

a Very general aim

- We aim to discuss here how some ideas and results from the area located in the intersection of the Theory of dynamical systems, la Statistical Mechanics and the Information theory can lead us towards the resolution of these problems (at least in some specific situations).

a Very general aim

- We aim to discuss here how some ideas and results from the area located in the intersection of the Theory of dynamical systems, la Statistical Mechanics and the Information theory can lead us towards the resolution of these problems (at least in some specific situations).

a Very general aim

- We aim to discuss here how some ideas and results from the area located in the intersection of the Theory of dynamical systems, la Statistical Mechanics and the Information theory can lead us towards the resolution of these problems (at least in some specific situations).

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

from where we belong...

- Can we develop some heuristic and universal methods to estimate the divergence (realtive entropy) between two Markovian sources with unknown memory and unknown distribution, from two arbitrary realization ?

Alphabets and Strings

$\mathcal{A}$ finite alphabet, $\mathcal{A}^* = \bigcup_n \mathcal{A}_n$

$$x = (x_1, x_2, \ldots, x_n), \quad x_j \in \mathcal{A}$$

- $\mathcal{A} = \{0, 1\}$: Bernoulli, HRV and Audio files
- $\mathcal{A} = \{A, C, G, T\}$
- $\mathcal{A} = \{a, b, c, \ldots, A, B, C, D, \ldots, ";", "!", ".", "?", \ldots\}$

Alphabets and Strings

$\mathcal{A}$ finite alphabet, $\mathcal{A}^* = \bigcup_n \mathcal{A}_n$

$$x = (x_1, x_2, \ldots, x_n), \quad x_j \in \mathcal{A}$$

- $\mathcal{A} = \{0, 1\}$: Bernoulli, HRV and Audio files
- $\mathcal{A} = \{A, C, G, T\}$
- $\mathcal{A} = \{a, b, c, \ldots, A, B, C, D, \ldots, ";", "!", ".", "?", \ldots\}$

Alphabets and Strings

$\mathcal{A}$ finite alphabet, $\mathcal{A}^* = \bigcup_n \mathcal{A}_n$

$$x = (x_1, x_2, \ldots, x_n), \quad x_j \in \mathcal{A}$$

- $\mathcal{A} = \{0, 1\}$: Bernoulli, HRV and Audio files
- $\mathcal{A} = \{A, C, G, T\}$
- $\mathcal{A} = \{a, b, c, \ldots, A, B, C, D, \ldots, ";", "!", ".", "?", \ldots\}$

Alphabets and Strings

$\mathcal{A}$ finite alphabet, $\mathcal{A}^* = \bigcup_n \mathcal{A}_n$

$$x = (x_1, x_2, \ldots, x_n), \quad x_j \in \mathcal{A}$$

- $\mathcal{A} = \{0, 1\}$: Bernoulli, HRV and Audio files
- $\mathcal{A} = \{A, C, G, T\}$
- $\mathcal{A} = \{a, b, c, \ldots, A, B, C, D, \ldots, ";", "!", ".", "?", \ldots\}$

Alphabets and Strings

$\mathcal{A}$ finite alphabet, $\mathcal{A}^* = \bigcup_n \mathcal{A}_n$

$$x = (x_1, x_2, \ldots, x_n), \quad x_j \in \mathcal{A}$$

- $\mathcal{A} = \{0, 1\}$: Bernoulli, HRV and Audio files
- $\mathcal{A} = \{A, C, G, T\}$
- $\mathcal{A} = \{a, b, c, \ldots, A, B, C, D, \ldots, ";", "!", ".", "?", \ldots\}$

similarity (pseudo) distance:

$$d : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}^+$$

- $d(x,y) = d(y,x)$
- $d(x,y) \geq 0$, con $d(x,y) = 0 \Longleftrightarrow x = y$
- $d(x,y) \lessapprox d(x,z) + d(z,y)$

d is a distance function able to detect and to enhance *similarity* among 2 or more symbolic strings, independently from the nature and from the origin of the similarities itself...

similarity (pseudo) distance:

$$d : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}^+$$

- $d(x,y) = d(y,x)$
- $d(x,y) \geq 0$, con $d(x,y) = 0 \Longleftrightarrow x = y$
- $d(x,y) \lessapprox d(x,z) + d(z,y)$

d is a distance function able to detect and to enhance *similarity* among 2 or more symbolic strings, independently from the nature and from the origin of the similarities itself...

similarity (pseudo) distance:

$$d : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}^+$$

- $d(x,y) = d(y,x)$
- $d(x,y) \geq 0$, con $d(x,y) = 0 \Longleftrightarrow x = y$
- $d(x,y) \lessapprox d(x,z) + d(z,y)$

d is a distance function able to detect and to enhance *similarity* among 2 or more symbolic strings, independently from the nature and from the origin of the similarities itself...

similarity (pseudo) distance:

$$d : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}^+$$

- $d(x,y) = d(y,x)$
- $d(x,y) \geq 0$, con $d(x,y) = 0 \Longleftrightarrow x = y$
- $d(x,y) \lessapprox d(x,z) + d(z,y)$

d is a distance function able to detect and to enhance *similarity* among 2 or more symbolic strings, independently from the nature and from the origin of the similarities itself...

similarity (pseudo) distance:

$$d : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}^+$$

- $d(x,y) = d(y,x)$
- $d(x,y) \geq 0$, con $d(x,y) = 0 \Longleftrightarrow x = y$
- $d(x,y) \lessapprox d(x,z) + d(z,y)$

d is a distance function able to detect and to enhance *similarity* among 2 or more symbolic strings, independently from the nature and from the origin of the similarities itself...

$K(y|x)$ e $H(Y|X)$

- $K(x)$: *Kolmogorov Complexity* for a given finite string $x = (x_1, \ldots, x_n)$.e.g. $K(x)$ is the length of the shortest file that can be obtained by *compressing* x using all possible reversible compression algorithms.
- $H(X) = \lim_{n\to\infty} \frac{K(x)}{n}$ *a.s* , *entropy* of a given stationary source with finite memory X.
- $K(y|x)$: *Conditional* Kolmogorov Complexity of y with respect a given string x; i.e. the shortest (prefix) program P such that, when x is given to the program P as input, the program prints y and then halts.
- $H(Y|X) = H(Y,X) - H(X)$: *Conditional Entropy*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

$K(y|x)$ e $H(Y|X)$

- $K(x)$: *Kolmogorov Complexity* for a given finite string $x = (x_1, \ldots, x_n)$.e.g. $K(x)$ is the length of the shortest file that can be obtained by *compressing* x using all possible reversible compression algorithms.
- $H(X) = \lim_{n\to\infty} \frac{K(x)}{n}$ *a.s* , *entropy* of a given stationary source with finite memory X.
- $K(y|x)$: *Conditional* Kolmogorov Complexity of y with respect a given string x; i.e. the shortest (prefix) program P such that, when x is given to the program P as input, the program prints y and then halts.
- $H(Y|X) = H(Y,X) - H(X)$: *Conditional Entropy*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

$K(y|x)$ e $H(Y|X)$

- $K(x)$: *Kolmogorov Complexity* for a given finite string $x = (x_1, \ldots, x_n)$.e.g. $K(x)$ is the length of the shortest file that can be obtained by *compressing* x using all possible reversible compression algorithms.
- $H(X) = \lim_{n\to\infty} \frac{K(x)}{n}$ *a.s* , *entropy* of a given stationary source with finite memory X.
- $K(y|x)$: *Conditional* Kolmogorov Complexity of y with respect a given string x; i.e. the shortest (prefix) program P such that, when x is given to the program P as input, the program prints y and then halts.
- $H(Y|X) = H(Y,X) - H(X)$: *Conditional Entropy*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

$K(y|x)$ e $H(Y|X)$

- $K(x)$: *Kolmogorov Complexity* for a given finite string $x = (x_1, \ldots, x_n)$.e.g. $K(x)$ is the length of the shortest file that can be obtained by *compressing* x using all possible reversible compression algorithms.
- $H(X) = \lim_{n\to\infty} \frac{K(x)}{n}$ *a.s* , *entropy* of a given stationary source with finite memory X.
- $K(y|x)$: *Conditional* Kolmogorov Complexity of y with respect a given string x; i.e. the shortest (prefix) program P such that, when x is given to the program P as input, the program prints y and then halts.
- $H(Y|X) = H(Y,X) - H(X)$: *Conditional Entropy*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

$K(y|x)$ e $H(Y|X)$

- $K(x)$: *Kolmogorov Complexity* for a given finite string $x = (x_1, \ldots, x_n)$.e.g. $K(x)$ is the length of the shortest file that can be obtained by *compressing* x using all possible reversible compression algorithms.
- $H(X) = \lim_{n\to\infty} \frac{K(x)}{n}$ *a.s* , *entropy* of a given stationary source with finite memory X.
- $K(y|x)$: *Conditional* Kolmogorov Complexity of y with respect a given string x; i.e. the shortest (prefix) program P such that, when x is given to the program P as input, the program prints y and then halts.
- $H(Y|X) = H(Y,X) - H(X)$: *Conditional Entropy*

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$K(y|x)$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$K(y|x)$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y) \approx C(x.y) - C(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y) \approx C(x.y) - C(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y) \approx C(x.y) - C(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

A Universal distance $d_K(x,y)$

- Li&Vitanyi (1997)

$$d_K(x,y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))}$$

but clearly $K(x|y)$ is NOT computable...

$$K(y|x) = K(x,y) - K(y) \approx K(x.y) - K(y) \approx C(x.y) - C(y)$$

$$d_C(x,y) = \frac{\max(C(x.y) - C(x), C(y.x) - C(y))}{\max(C(x), C(y))}$$

Parsing a string

A given parsing of a given string $x(1,n)$ is an arbitrary subdivision of x in substrings,

$$x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$$

The number of words of the parsing $c(x) = t$ is called the *complexity of the parsing*

Lempel and Ziv (1976): Exhaustive Parsing

This particular parsing is similar, but it does NOT coincide with the parsing implemented by the usual compression algorithms (parsing + coding) LZ77 ed LZ78

Parsing a string

A given parsing of a given string $x(1,n)$ is an arbitrary subdivision of x in substrings,

$$x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$$

The number of words of the parsing $c(x) = t$ is called the *complexity of the parsing*

Lempel and Ziv (1976): Exhaustive Parsing

This particular parsing is similar, but it does NOT coincide with the parsing implemented by the usual compression algorithms (parsing + coding) LZ77 ed LZ78

Parsing a string

A given parsing of a given string $x(1,n)$ is an arbitrary subdivision of x in substrings,

$$x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$$

The number of words of the parsing $c(x) = t$ is called the *complexity of the parsing*

Lempel and Ziv (1976): Exhaustive Parsing

This particular parsing is similar, but it does NOT coincide with the parsing implemented by the usual compression algorithms (parsing + coding) LZ77 ed LZ78

LZ Exhaustive Parsing: definitions

The defining rule of the exhaustive parsing is based on a construction process of the string from some of its proper substrings.

The process could be either a pure *reproduction* process (if there are no new original elements in the string with respect the substring) or a pure production process (if one allow the introduction of a new symbol at the end of the cloned substring)

LZ Exhaustive Parsing: definitions

The defining rule of the exhaustive parsing is based on a construction process of the string from some of its proper substrings.

The process could be either a pure *reproduction* process (if there are no new original elements in the string with respect the substring) or a pure production process (if one allow the introduction of a new symbol at the end of the cloned substring)

LZ Exhaustive Parsing: definitions

The defining rule of the exhaustive parsing is based on a construction process of the string from some of its proper substrings.

The process could be either a pure *reproduction* process (if there are no new original elements in the string with respect the substring) or a pure production process (if one allow the introduction of a new symbol at the end of the cloned substring)

LZ Exhaustive Parsing: definitions

The defining rule of the exhaustive parsing is based on a construction process of the string from some of its proper substrings.

The process could be either a pure *reproduction* process (if there are no new original elements in the string with respect the substring) or a pure production process (if one allow the introduction of a new symbol at the end of the cloned substring)

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.

For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.

For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|01110\mathbf{1}1001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001 0 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100110010 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100110010 is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Production and Reproduction

Reproduction Process $x(1,n)$ is reproduced from $Q = x(1,m)$ with $m < n$ if $x(m+1,n)$ is a substring of $x(1,n-1)$, namely if it exists an index $p < m$ such that $x(m+1,n) = x(p,p+n-m-1)$.
For example 1110011100111001 is reproduced from $Q = 111001110$

$$1110|01110|0111001$$

Production Process $x(1,n)$ is produced from $Q = S(1,m)$ with $m < n$ if $x(1,n-1)$ is reproduced from Q.
For example 1110011100111001**0** is produced from $Q = 111001110$

$$1110|01110|01110010$$

Exhaustive Parsing of a given symbolic string

An LZ-admissible parsing of complexity $c(x) = t$, $x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$ is a parsing for which the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is always a production process

- The LZ77 and LZ78 parsing are LZ-admissible

Given an LZ-admissible parsing, if for each $m = 1,\ldots,t-2$ the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is NOT a reproduction process but a true production one, then the parsing is called *exhaustive*.

- For any given finite string x, the exhaustive parsing is unique

Exhaustive Parsing of a given symbolic string

An LZ-admissible parsing of complexity $c(x) = t$, $x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$ is a parsing for which the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is always a production process

- The LZ77 and LZ78 parsing are LZ-admissible

Given an LZ-admissible parsing, if for each $m = 1,\ldots,t-2$ the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is NOT a reproduction process but a true production one, then the parsing is called *exhaustive*.

- For any given finite string x, the exhaustive parsing is unique

Exhaustive Parsing of a given symbolic string

An LZ-admissible parsing of complexity $c(x) = t$, $x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$ is a parsing for which the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is always a production process

- The LZ77 and LZ78 parsing are LZ-admissible

Given an LZ-admissible parsing, if for each $m = 1,\ldots,t-2$ the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is NOT a reproduction process but a true production one, then the parsing is called *exhaustive*.

- For any given finite string x, the exhaustive parsing is unique

Exhaustive Parsing of a given symbolic string

An LZ-admissible parsing of complexity $c(x) = t$, $x(1,n) = x(1,h_1)x(h_1+1,h_2)\cdots x(h_{t-1}+1,n)$ is a parsing for which the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is always a production process

- The LZ77 and LZ78 parsing are LZ-admissible

Given an LZ-admissible parsing, if for each $m = 1,\ldots,t-2$ the construction of $x(1,h_{m+1})$ from $x(1,h_m)$ is NOT a reproduction process but a true production one, then the parsing is called *exhaustive*.

- For any given finite string x, the exhaustive parsing is unique

Properties of the Parsing

- The complexity $c_E(x)$ of the exhaustive parsing is minimal among the complexity of all LZ-admissible parsing:

$$c_E(x) = \min_{H \text{ admissible history}} c_H(x)$$

- $c_E(yx) \leq c_E(y) + c_E(x)$
- $$c_E(x(1,n)) \leq \frac{n}{(1-\epsilon_n)\log n}$$
- For a given stationary ergodic source, we have:

$$\lim_{n\to\infty} P\left\{c_E(x) < \frac{Hn}{\log n}(1-\epsilon)\right\} = 0, \quad \forall \epsilon > 0$$

i.e.

$$c_E(x)\frac{\log n}{n} \approx H$$

Properties of the Parsing

- The complexity $c_E(x)$ of the exhaustive parsing is minimal among the complexity of all LZ-admissible parsing:

$$c_E(x) = \min_{H \text{ admissible history}} c_H(x)$$

- $c_E(yx) \leq c_E(y) + c_E(x)$
- $$c_E(x(1,n)) \leq \frac{n}{(1-\epsilon_n)\log n}$$
- For a given stationary ergodic source, we have:

$$\lim_{n\to\infty} P\left\{c_E(x) < \frac{Hn}{\log n}(1-\epsilon)\right\} = 0, \quad \forall \epsilon > 0$$

i.e.

$$c_E(x)\frac{\log n}{n} \approx H$$

Properties of the Parsing

- The complexity $c_E(x)$ of the exhaustive parsing is minimal among the complexity of all LZ-admissible parsing:

$$c_E(x) = \min_{H \text{ admissible history}} c_H(x)$$

- $c_E(yx) \leq c_E(y) + c_E(x)$
- $$c_E(x(1,n)) \leq \frac{n}{(1-\epsilon_n)\log n}$$
- For a given stationary ergodic source, we have:

$$\lim_{n\to\infty} P\left\{c_E(x) < \frac{Hn}{\log n}(1-\epsilon)\right\} = 0, \quad \forall \epsilon > 0$$

i.e.

$$c_E(x)\frac{\log n}{n} \approx H$$

Properties of the Parsing

- The complexity $c_E(x)$ of the exhaustive parsing is minimal among the complexity of all LZ-admissible parsing:

$$c_E(x) = \min_{H \text{ admissible history}} c_H(x)$$

- $c_E(yx) \leq c_E(y) + c_E(x)$
- $$c_E(x(1,n)) \leq \frac{n}{(1-\epsilon_n)\log n}$$
- For a given stationary ergodic source, we have:

$$\lim_{n\to\infty} P\left\{c_E(x) < \frac{Hn}{\log n}(1-\epsilon)\right\} = 0, \quad \forall \epsilon > 0$$

i.e.

$$c_E(x)\frac{\log n}{n} \approx H$$

A *computable "distance"* between strings

- Otu&Sayood (2003)

$$d_{LZ}(x,y) = d(x,y) = \frac{c_E(yx) - c_E(y) + c_E(xy) - c_E(x)}{\frac{1}{2}(c_E(yx) + c_E(xy))}$$

- $d(x,y) = d(y,x)$
- $d(x,x) \leq \frac{1}{c_E(x)} \approx 0$
-

$$d(x,y) \lesssim_{\log} d(x,z) + d(z,y)$$

A *computable "distance"* between strings

- Otu&Sayood (2003)

$$d_{LZ}(x,y) = d(x,y) = \frac{c_E(yx) - c_E(y) + c_E(xy) - c_E(x)}{\frac{1}{2}(c_E(yx) + c_E(xy))}$$

- $d(x,y) = d(y,x)$
- $d(x,x) \leq \frac{1}{c_E(x)} \approx 0$
- $$d(x,y) \lesssim_{\log} d(x,z) + d(z,y)$$

A *computable "distance"* between strings

- Otu&Sayood (2003)

$$d_{LZ}(x,y) = d(x,y) = \frac{c_E(yx) - c_E(y) + c_E(xy) - c_E(x)}{\frac{1}{2}(c_E(yx) + c_E(xy))}$$

- $d(x,y) = d(y,x)$
- $d(x,x) \leq \frac{1}{c_E(x)} \approx 0$
- $d(x,y) \lesssim_{\log} d(x,z) + d(z,y)$

A *computable "distance"* between strings

- Otu&Sayood (2003)

$$d_{LZ}(x,y) = d(x,y) = \frac{c_E(yx) - c_E(y) + c_E(xy) - c_E(x)}{\frac{1}{2}(c_E(yx) + c_E(xy))}$$

- $d(x,y) = d(y,x)$
- $d(x,x) \leq \frac{1}{c_E(x)} \approx 0$
- $$d(x,y) \lesssim_{\log} d(x,z) + d(z,y)$$

Another *statistical distance* : NGram distance d_G

- Kesley (2003)

Given $x = (x_1, \ldots, x_N)$ e $2 \leq n \leq 8$, $f(\omega)$ represents the empirical frequency in x of the n-gram $\omega = (\omega_1, \ldots, \omega_n)$.

100101101

$$00 = 1 \quad \Rightarrow \quad f(00) = \frac{1}{8}$$

$$01 = 3 \quad \Rightarrow \quad f(01) = \frac{3}{8}$$

$$10 = 3 \quad \Rightarrow \quad f(10) = \frac{3}{8}$$

$$11 = 1 \quad \Rightarrow \quad f(11) = \frac{1}{8}$$

Another *statistical distance* : NGram distance d_G

- Kesley (2003)

Given $x = (x_1, \ldots, x_N)$ e $2 \leq n \leq 8$, $f(\omega)$ represents the empirical frequency in x of the n-gram $\omega = (\omega_1, \ldots, \omega_n)$.

$$1\,00\,101101$$

$$00 = 1 \quad \Rightarrow \quad f(00) = \frac{1}{8}$$

$$01 = 3 \quad \Rightarrow \quad f(01) = \frac{3}{8}$$

$$10 = 3 \quad \Rightarrow \quad f(10) = \frac{3}{8}$$

$$11 = 1 \quad \Rightarrow \quad f(11) = \frac{1}{8}$$

Another *statistical distance* : NGram distance d_G

- Kesley (2003)

Given $x = (x_1, \ldots, x_N)$ e $2 \leq n \leq 8$, $f(\omega)$ represents the empirical frequency in x of the n-gram $\omega = (\omega_1, \ldots, \omega_n)$.

$$100101101$$

$$\begin{aligned} 00 = 1 \quad &\Rightarrow \quad f(00) = \frac{1}{8} \\ 01 = 3 \quad &\Rightarrow \quad f(01) = \frac{3}{8} \\ 10 = 3 \quad &\Rightarrow \quad f(10) = \frac{3}{8} \\ 11 = 1 \quad &\Rightarrow \quad f(11) = \frac{1}{8} \end{aligned}$$

Another *statistical distance* : NGram distance d_G

- Kesley (2003)

Given $x = (x_1, \ldots, x_N)$ e $2 \leq n \leq 8$, $f(\omega)$ represents the empirical frequency in x of the n-gram $\omega = (\omega_1, \ldots, \omega_n)$.

100101101

$$
\begin{aligned}
00 = 1 \quad &\Rightarrow \quad f(00) = \frac{1}{8} \\
01 = 3 \quad &\Rightarrow \quad f(01) = \frac{3}{8} \\
10 = 3 \quad &\Rightarrow \quad f(10) = \frac{3}{8} \\
11 = 1 \quad &\Rightarrow \quad f(11) = \frac{1}{8}
\end{aligned}
$$

Another *statistical distance* : NGram distance d_G

- Kesley (2003)

Given $x = (x_1, \ldots, x_N)$ e $2 \leq n \leq 8$, $f(\omega)$ represents the empirical frequency in x of the n-gram $\omega = (\omega_1, \ldots, \omega_n)$.

$$100101101$$

$$\begin{aligned} 00 = 1 \quad &\Rightarrow \quad f(00) = \frac{1}{8} \\ 01 = 3 \quad &\Rightarrow \quad f(01) = \frac{3}{8} \\ 10 = 3 \quad &\Rightarrow \quad f(10) = \frac{3}{8} \\ 11 = 1 \quad &\Rightarrow \quad f(11) = \frac{1}{8} \end{aligned}$$

$d_G(x, y)$

Given two finite strings $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, denote by f_1 and f_2 the frequency vectors of the n-grams in x and y ordered in decreasing order and truncated at a given fixed length L.

Given the two parameters n and L, we define (used in Authorship Attribution):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \left(\frac{f_2(\omega) - f_1(\omega)}{f_2(\omega) + f_1(\omega)} \right)^2$$

In the case of DNA, the following turns out to be more useful (A. Tomović, P. Janičić and V Kešelj (2005)):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \frac{|f_2(\omega) - f_1(\omega)|}{\sqrt{f_1(\omega) \cdot f_2(\omega)} + 1}$$

$d_G(x, y)$

Given two finite strings $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, denote by f_1 and f_2 the frequency vectors of the n-grams in x and y ordered in decreasing order and truncated at a given fixed length L.

Given the two parameters n and L, we define (used in Authorship Attribution):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \left(\frac{f_2(\omega) - f_1(\omega)}{f_2(\omega) + f_1(\omega)} \right)^2$$

In the case of DNA, the following turns out to be more useful (A. Tomović, P. Janičić and V Kešelj (2005)):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \frac{|f_2(\omega) - f_1(\omega)|}{\sqrt{f_1(\omega) \cdot f_2(\omega)} + 1}$$

$d_G(x, y)$

Given two finite strings $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, denote by f_1 and f_2 the frequency vectors of the n-grams in x and y ordered in decreasing order and truncated at a given fixed length L.

Given the two parameters n and L, we define (used in Authorship Attribution):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \left(\frac{f_2(\omega) - f_1(\omega)}{f_2(\omega) + f_1(\omega)} \right)^2$$

In the case of DNA, the following turns out to be more useful (A. Tomović, P. Janičić and V Kešelj (2005)):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \frac{|f_2(\omega) - f_1(\omega)|}{\sqrt{f_1(\omega) \cdot f_2(\omega)} + 1}$$

$d_G(x, y)$

Given two finite strings $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, denote by f_1 and f_2 the frequency vectors of the n-grams in x and y ordered in decreasing order and truncated at a given fixed length L.

Given the two parameters n and L, we define (used in Authorship Attribution):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \left(\frac{f_2(\omega) - f_1(\omega)}{f_2(\omega) + f_1(\omega)} \right)^2$$

In the case of DNA, the following turns out to be more useful (A. Tomović, P. Janičić and V Kešelj (2005)):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \frac{|f_2(\omega) - f_1(\omega)|}{\sqrt{f_1(\omega) \cdot f_2(\omega)} + 1}$$

$d_G(x, y)$

Given two finite strings $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$, denote by f_1 and f_2 the frequency vectors of the n-grams in x and y ordered in decreasing order and truncated at a given fixed length L.

Given the two parameters n and L, we define (used in Authorship Attribution):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \left(\frac{f_2(\omega) - f_1(\omega)}{f_2(\omega) + f_1(\omega)} \right)^2$$

In the case of DNA, the following turns out to be more useful (A. Tomović, P. Janičić and V Kešelj (2005)):

$$d_G(x, y) = \sum_{\omega \in f_1 \bigcup f_2} \frac{|f_2(\omega) - f_1(\omega)|}{\sqrt{f_1(\omega) \cdot f_2(\omega)} + 1}$$

n-gram in a literary text, $n = 6$

zione 185	uzione67	erche 55
della121	oluzio66	zioni 54
della 121	luzion66	o dell54
mente 114	non s65	stori54
quest101	voluzi65	perch53
sono 94	loro 63	perche53
azione92	hanno62	ione 52
e che 80	hanno 62	ente 50
che 79	socia61	come 50
non 74	social61	mento 50
amente74	rivolu61	e dell50
o che 71	ivoluz61	prole50
rivol68	e non 60	prolet50

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{ by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Kullbach-Leibler divergence (relative entropy)

q_z e p_x two Markovian sources with unknown memory length

$$
\begin{aligned}
D(q_z \parallel p_x) &\simeq \sum_i q_i \log \frac{q_i}{p_i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{\omega\in\mathcal{A}_n} q_z(\omega) \log \frac{q_z(\omega)}{p_x(\omega)} \\
&= \lim_{n\to\infty} -\frac{1}{n} \log p_x(Z^n) - H(q_z) \\
&= \sum_{s\in\mathcal{S}} q_z(s) \sum_{\alpha\in\mathcal{A}} q_z(\alpha|s) \log \left(\frac{q_z(\alpha|s)}{p_x(\alpha|s)} \right) \\
&= \text{by the Merhav-Zev Theorem} \\
&= \lim_{n\to\infty} \frac{1}{n} K_{78}(Z \parallel X) \log n - \frac{1}{n} K_{78}(Z) \log K_{78}(Z)
\end{aligned}
$$

Other K-L divergence estimator

- (Benedetto, Caglioti e Loreto (2002)), using gzip:

$$D(q_z \parallel p_x) \simeq \frac{\Delta_{xz_0} - \Delta_{zz_0}}{|z_0|}.$$

- (Cai, Kukarni and Verdu' (2006)), K-L estimation by the Burrows-Wheeler Transform (BWT)

Other K-L divergence estimator

- (Benedetto, Caglioti e Loreto (2002)), using gzip:

$$D(q_z \parallel p_x) \simeq \frac{\Delta_{xz_0} - \Delta_{zz_0}}{|z_0|}.$$

- (Cai, Kukarni and Verdu' (2006)), K-L estimation by the Burrows-Wheeler Transform (BWT)

BWT

- The BWT is a permutation (easy to invert) and in particular it trasforms a finite memory Markovian sequence into a ”piecewise Bernoulli (i.i.d)” sequence
- It has been introduce in 1993 but up to now it has been used in computer science for compression: BWT(x) is ”more suitable” for compression through an arithmetic coding, for example
- let us see how it works through an example....

BWT

- The BWT is a permutation (easy to invert) and in particular it trasforms a finite memory Markovian sequence into a ”piecewise Bernoulli (i.i.d)” sequence
- It has been introduce in 1993 but up to now it has been used in computer science for compression: BWT(x) is ”more suitable” for compression through an arithmetic coding, for example
- let us see how it works through an example....

BWT

- The BWT is a permutation (easy to invert) and in particular it trasforms a finite memory Markovian sequence into a ”piecewise Bernoulli (i.i.d)” sequence
- It has been introduce in 1993 but up to now it has been used in computer science for compression: BWT(x) is ”more suitable” for compression through an arithmetic coding, for example
- let us see how it works through an example....

BWT

- The BWT is a permutation (easy to invert) and in particular it trasforms a finite memory Markovian sequence into a ”piecewise Bernoulli (i.i.d)” sequence
- It has been introduce in 1993 but up to now it has been used in computer science for compression: BWT(x) is ”more suitable” for compression through an arithmetic coding, for example
- let us see how it works through an example....

BWT

BWT ("chenoiastoseminario"):

Stringa: chenoiastoseminario

c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o
h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c
e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h
n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e
o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n
i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o
a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i
s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a
t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s
o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t
s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o
e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s
m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e
i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m
n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i
a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n
r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a
i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r
o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i

															L	
a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n
a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i
c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o
e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s
e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h
h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c
i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o
i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m
i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r
m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e
n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i
n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e
o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i
o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n
o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t
r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a
s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o
s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a
t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s

*

BWT

BWT ("chenoiastoseminario"):

Stringa: chenoiastoseminario

c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o
h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c
e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h
n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e
o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n
i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o
a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i
s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a
t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s
o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t
s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o
e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s
m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e
i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m
n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i
a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n
r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a
i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r
o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i

															L	
a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n
a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i
c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o
e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s
e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h
h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c
i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o
i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m
i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r
m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e
n	a	r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i
n	o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e
o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a	r	i
o	i	a	s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n
o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t
r	i	o	c	h	e	n	o	i	a	s	t	o	s	e	m	i	n	a
s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s	t	o
s	t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a
t	o	s	e	m	i	n	a	r	i	o	c	h	e	n	o	i	a	s

*

the Cai, Kukarni and Verdu' (2006) algorithm

- Entropy indicator
- Divergence estimate

the Cai, Kukarni and Verdu' (2006) algorithm

- Entropy indicator
- Divergence estimate

d_{LZ} vs. *GeneCompress*

- X. Chen, S. Kwong, M.Li, *A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison* , (1999)
- d_{LZ} is NOT an *ad hoc* method
- NO *alignment* between sequences is required
- it can work for both complete genoma and single gene

d_{LZ} vs. *GeneCompress*

- X. Chen, S. Kwong, M.Li, *A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison* , (1999)
- d_{LZ} is NOT an *ad hoc* method
- NO *alignment* between sequences is required
- it can work for both complete genoma and single gene

d_{LZ} vs. *GeneCompress*

- X. Chen, S. Kwong, M.Li, *A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison* , (1999)
- d_{LZ} is NOT an *ad hoc* method
- NO *alignment* between sequences is required
- it can work for both complete genoma and single gene

d_{LZ} vs. *GeneCompress*

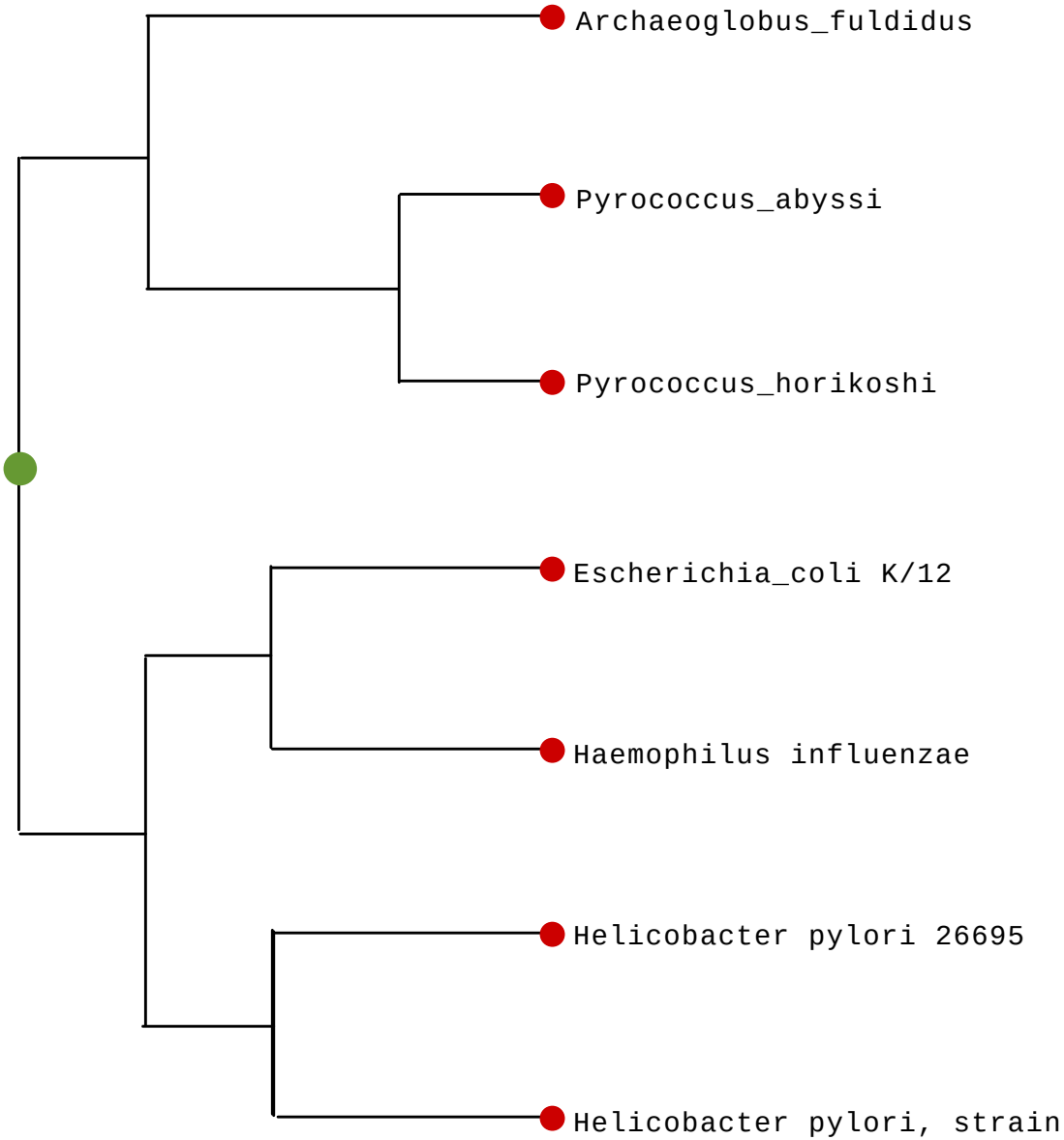
- X. Chen, S. Kwong, M.Li, *A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison* , (1999)
- d_{LZ} is NOT an *ad hoc* method
- NO *alignment* between sequences is required
- it can work for both complete genoma and single gene

Complete Genoma

Archaea Bacteria *Archaeoglobus fulgidus*,*Pyrococcus abyssi* and *Pyrococcus horikoshii* OT3

Bacteria *Escherichia coli* K-12 MG1655, *Haemophilus influenzae* Rd, *Helicobacter pylori* 26695 and *Helicobacter pylori*, strain J99

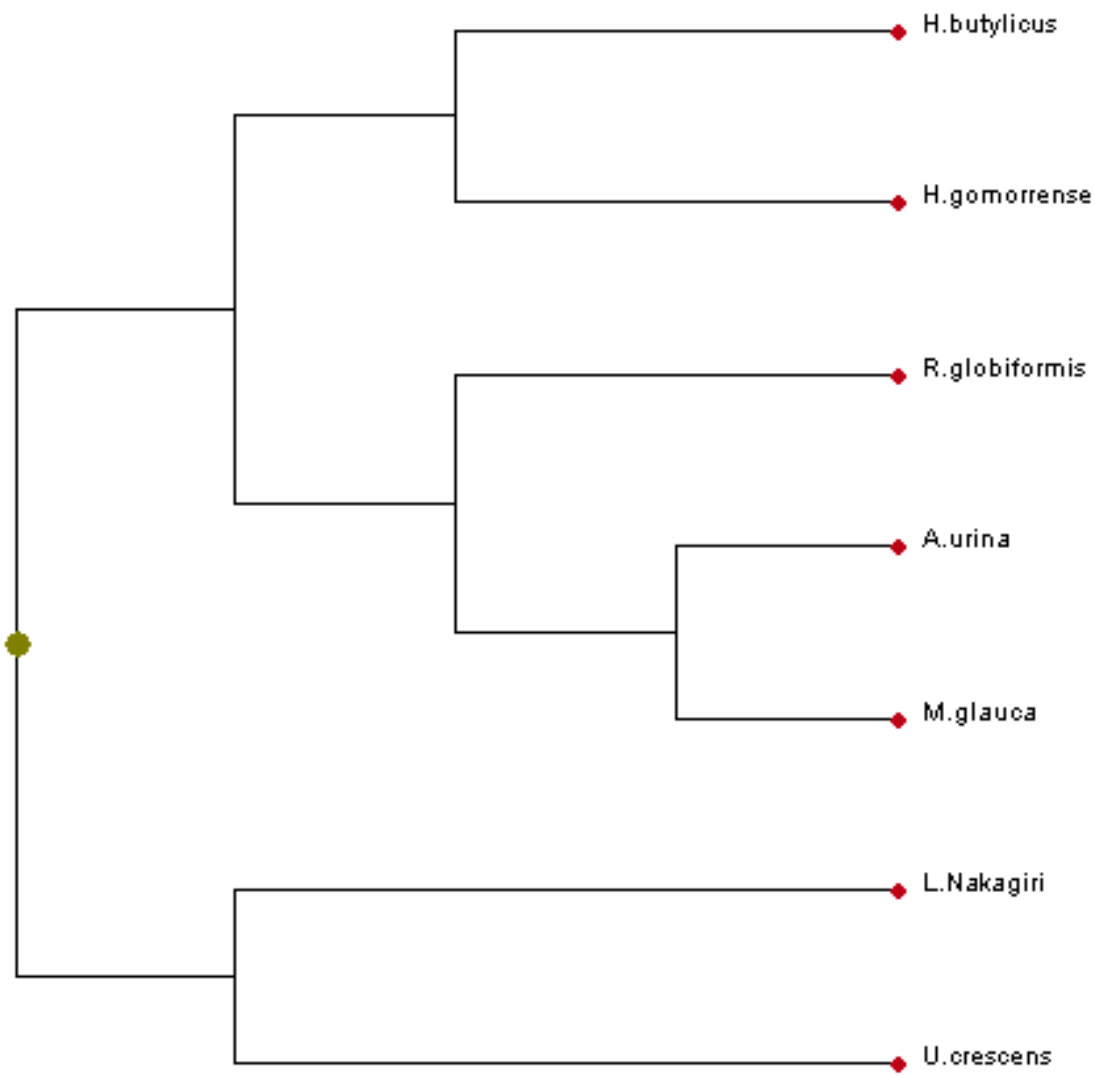
Complete Genoma



rRNA single Genes

Archaebacteria	*H. butylicus* and *Halobaculum gomorrense*
Eubacteria	*Aerococcus urina*, *M. glauca* strain B1448-1 and *Rhodopila globiformis*
Eukaryotes	*Urosporidium crescens*, *Labyrinthula sp. Nakagiri*

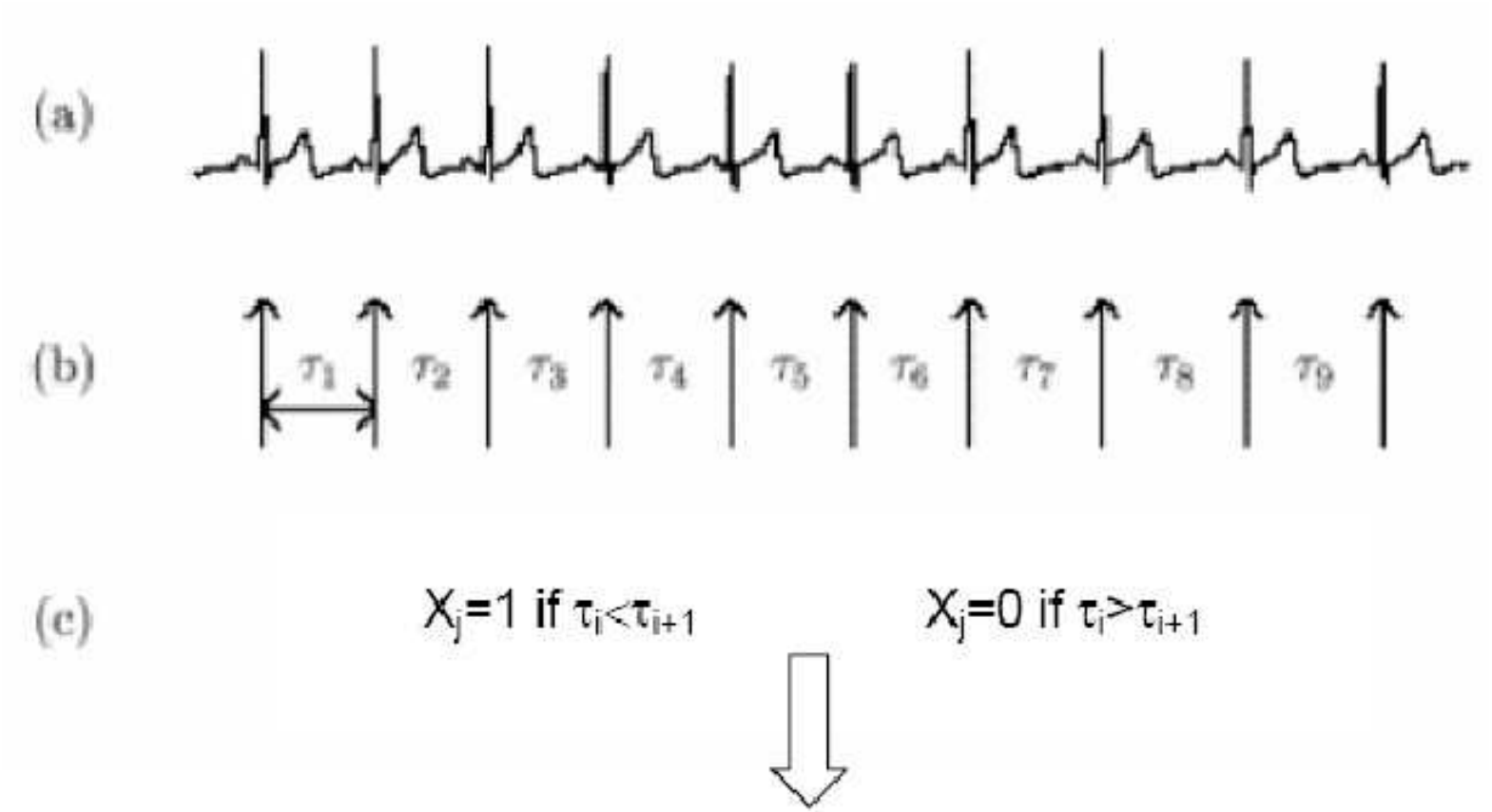
rRNA single Genes



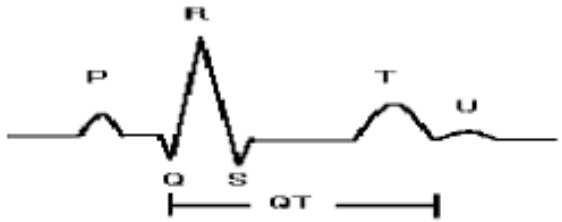
from the ECG sequenc to HRV...

- —, C. Farinelli, M. Manca, A. Tolomelli: *"A sequence distance measure for biological signals: new applications to HRV analysis"*, submitted to Physica A (2006).
- —, C. Farinelli e G. Menconi :*"Parsing complexity and sequence distance with applications to heartbeat signals"*, submitted 2007

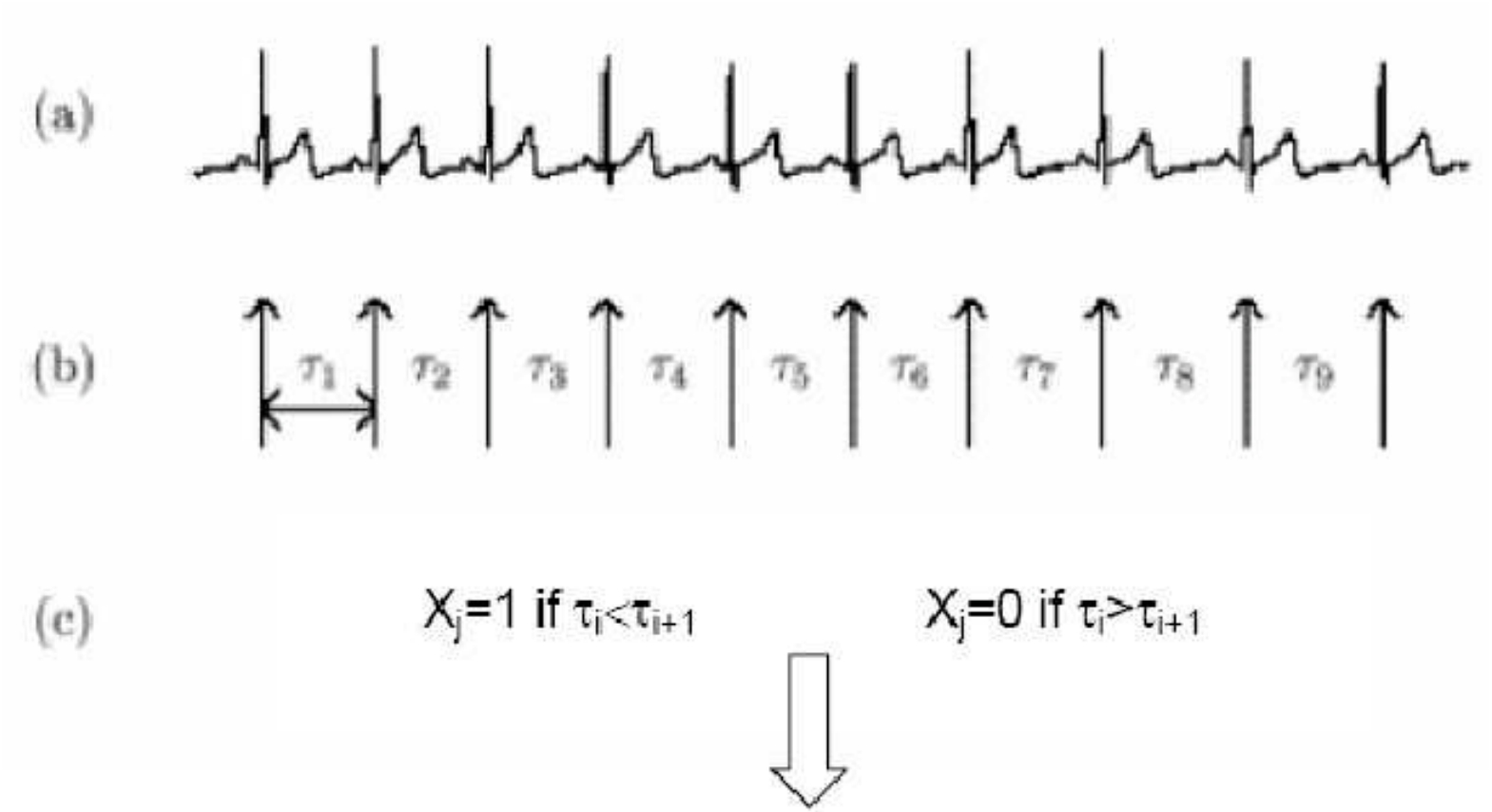
from the ECG sequenc to HRV...



from the ECG sequenc to HRV...



from the ECG sequenc to HRV...



Experimental Data

Data Set 1: **nk** v.s. **gk**

nk group made of 90 patients from the Department of Cardiology of Medical University in Gdańsk, Poland (9 women, 81 men, the average age is 57 $\pm$10) in whom the reduced left ventricular systolic function was recognized by echocardiogram.

gk group made of 40 healthy individuals (4 women, 36 men, the average age is 52 $\pm$ 8) without past history of cardiovascular disease, with both echocardiogram and electrocardiogram in normal range.

Experimental Data

Data Set 2: **young** v.s. **old**

old group 13 healthy subject belonging to **gk** previously described.

young group 13 healthy and rather young people (age between 20-40 years). These patients (3 men, 10 women) show no significant arrhythmias.

Data Set 3: **NYHA >Classification**, 20 patients distributed among the 4 NYHA classes

Experimental Data

Data Set 2: **young** v.s. **old**

old group 13 healthy subject belonging to **gk** previously described.

young group 13 healthy and rather young people (age between 20-40 years). These patients (3 men, 10 women) show no significant arrhythmias.

Data Set 3: **NYHA >Classification**, 20 patients distributed among the 4 NYHA classes

gk v.s. nk, 1

	gk_group	nk_group
gk02_nn	0,950977	0,955649
gk03_nn	0,9512	0,959749
gk04_nn	0,951591	0,957155
gk05_nn	0,949889	0,953167
gk06_nn	0,949679	0,958141
gk07_nn	0,951273	0,962977
gk08_nn	0,951308	0,962828
gk09_nn	0,949684	0,95644
gk10_nn	0,950085	0,959365
gk11_nn	0,949688	0,954517
gk13_nn	0,94936	0,95906
gk14_nn	0,949817	0,957204
gk15_nn	0,951751	0,964054
gk16_nn	0,949499	0,952967
gk17_nn	0,950058	0,956208
gk18_nn	0,951352	0,958267
gk19_nn	0,950012	0,957825
gk20_nn	0,953429	0,965333
gk21_nn	0,950678	0,959302
gk22_nn	0,950278	0,958852
nk10_nn	0,953073	0,952105
nk11_nn	0,955284	0,950414
nk12_nn	0,951612	0,954686
nk13_nn	0,955527	0,950697
nk14_nn	0,95358	0,958575
nk15_nn	0,952657	0,950346
nk16_nn	0,95545	0,952969
nk17_nn	0,975155	0,969354
nk18_nn	0,976497	0,964703
nk19_nn	0,952482	0,950202

gk v.s. nk, 1

	gk_group	nk_group
gk02_nn	0,950977	0,955649
gk03_nn	0,9512	0,959749
gk04_nn	0,951591	0,957155
gk05_nn	0,949889	0,953167
gk06_nn	0,949679	0,958141
gk07_nn	0,951273	0,962977
gk08_nn	0,951308	0,962828
gk09_nn	0,949684	0,95644
gk10_nn	0,950085	0,959365
gk11_nn	0,949688	0,954517
gk13_nn	0,94936	0,95906
gk14_nn	0,949817	0,957204
gk15_nn	0,951751	0,964054
gk16_nn	0,949499	0,952967
gk17_nn	0,950058	0,956208

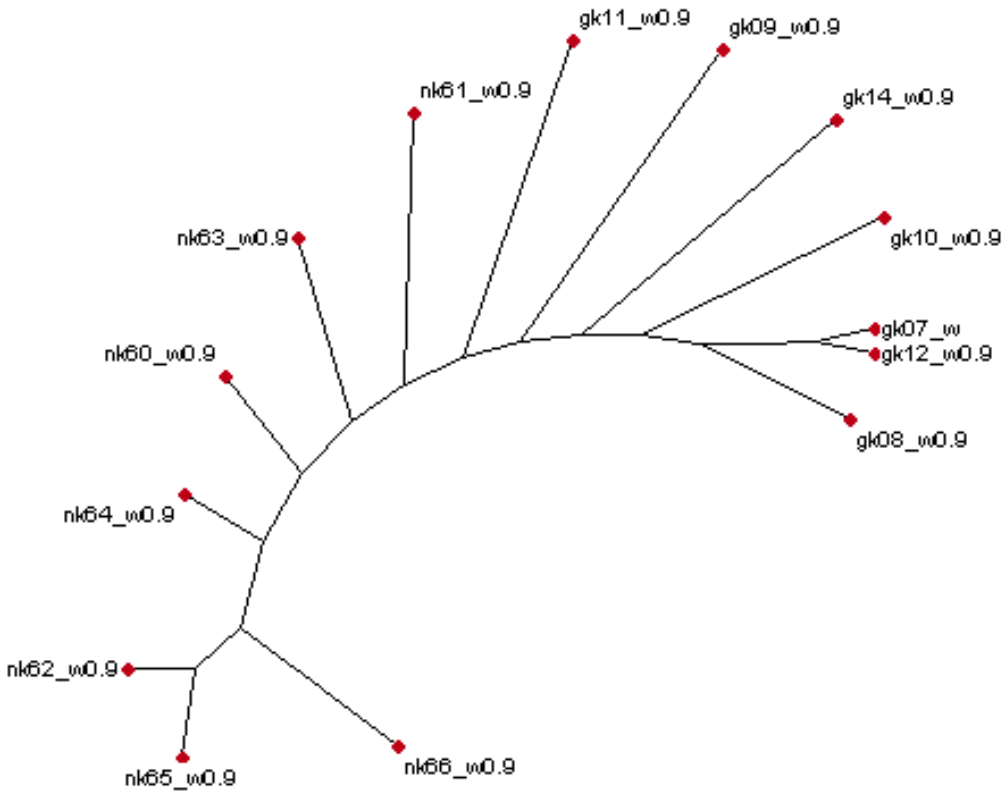
gk v.s. nk, 2

	gk_group	nk_group
gk02_ w	0,944999	0,949697
gk03_w	0,942169	0,949849
gk04_w	0,94477	0,949449
gk05_w	0,946066	0,947472
gk06_w	0,943874	0,953748
gk07_w	0,945075	0,960126
gk08_w	0,94387	0,955866
gk09_w	0,943006	0,951416
gk10_w	0,941327	0,954052
gk11_w	0,942418	0,945749
gk13_w	0,940751	0,948664
gk14_w	0,942632	0,954633
gk15_w	0,943504	0,956356
gk16_w	0,94459	0,947752
gk17_w	0,940355	0,949688
gk18_w	0,944521	0,950204
gk19_w	0,942666	0,946773
gk20_w	0,944984	0,960437
gk21_w	0,943947	0,955633
gk22_w	0,944009	0,95303
nk10_w	0,94555	0,94192
nk11_w	0,950804	0,942961
nk12_w	0,94292	0,943463
nk13_w	0,950983	0,941804
nk14_w	0,949428	0,952428
nk15_w	0,947493	0,944664
nk16_w	0,950896	0,944168
nk17_w	0,970349	0,962885
nk18_w	0,964134	0,948842
nk19_w	0,946231	0,942469

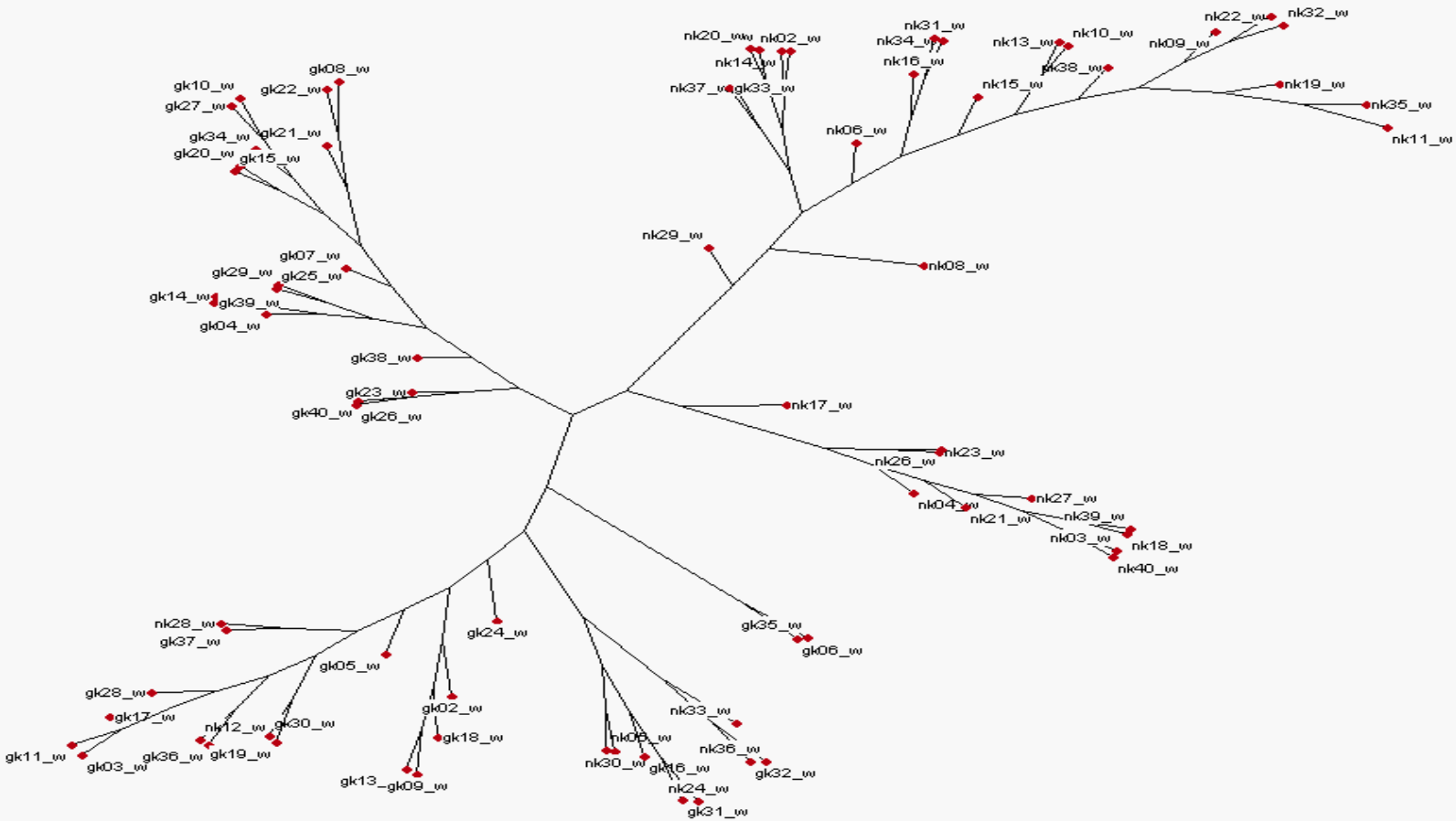
gk v.s. nk, 2

	gk_group	nk_group
gk02_ w	0,944999	0,949697
gk03_w	0,942169	0,949849
gk04_w	0,94477	0,949449
gk05_w	0,946066	0,947472
gk06_w	0,943874	0,953748
gk07_w	0,945075	0,960126
gk08_w	0,94387	0,955866
gk09_w	0,943006	0,951416
gk10_w	0,941327	0,954052
gk11_w	0,942418	0,945749
gk13_w	0,940751	0,948664
gk14_w	0,942632	0,954633
gk15_w	0,943504	0,956356
gk16_w	0,94459	0,947752
gk17_ w	0,940355	0,949688

gk v.s. nk: Alberi



gk v.s. nk: Alberi



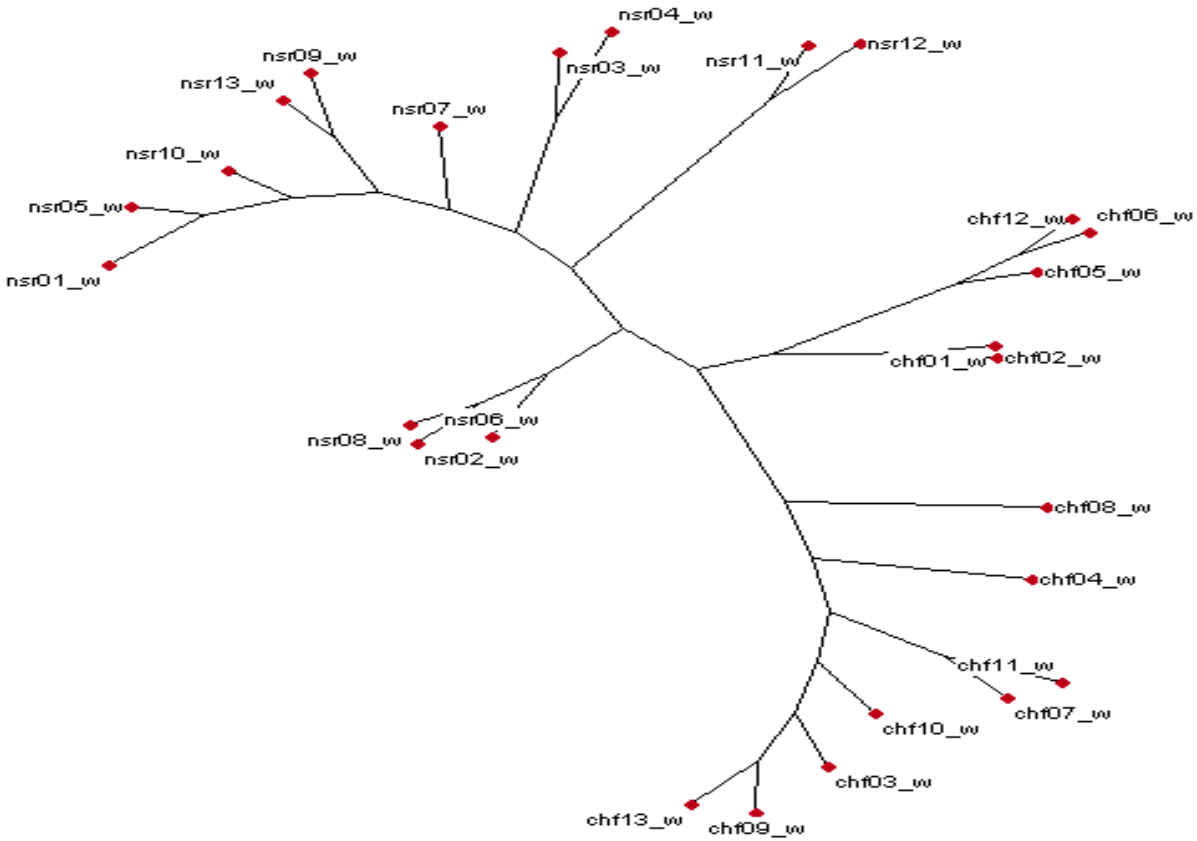
chf v.s. nsr

	chf	nsr
chf01_w	0,988736	0,993407
chf02_w	0,992512	0,994858
chf03_w	0,971186	0,996126
chf04_w	0,980403	0,991931
chf05_w	0,980736	0,992299
chf06_w	0,979843	0,9914
chf07_w	0,974151	0,993553
chf08_w	0,994647	0,99748
chf09_w	0,969402	0,994815
chf10_w	0,966486	0,992431
chf11_w	0,979891	0,99794
chf12_w	0,981962	0,992295
chf13_w	0,973136	0,996432
nsr01_w	0,994181	0,925976
nsr02_w	0,993675	0,928663
nsr03_w	0,993803	0,923911
nsr04_w	0,994018	0,935523
nsr05_w	0,994254	0,925418
nsr06_w	0,994561	0,930583
nsr07_w	0,993325	0,922587
nsr08_w	0,994585	0,938982
nsr09_w	0,994489	0,923555
nsr10_w	0,994857	0,926272
nsr11_w	0,994628	0,92443
nsr12_w	0,994004	0,931252
nsr13_w	0,994587	0,923272

chf v.s. nsr

	chf	nsr
chf01_w	0,988736	0,993407
chf02_w	0,992512	0,994858
chf03_w	0,971186	0,996126
chf04_w	0,980403	0,991931
chf05_w	0,980736	0,992299
chf06_w	0,979843	0,9914
chf07_w	0,974151	0,993553
chf08_w	0,994647	0,99748
chf09_w	0,969402	0,994815
chf10_w	0,966486	0,992431
chf11_w	0,979891	0,99794
chf12_w	0,981962	0,992295
chf13_w	0,973136	0,996432
nsr01_w	0,994181	0,925976
nsr02_w	0,993675	0,928663

chf v.s. **nsr**: Alberi



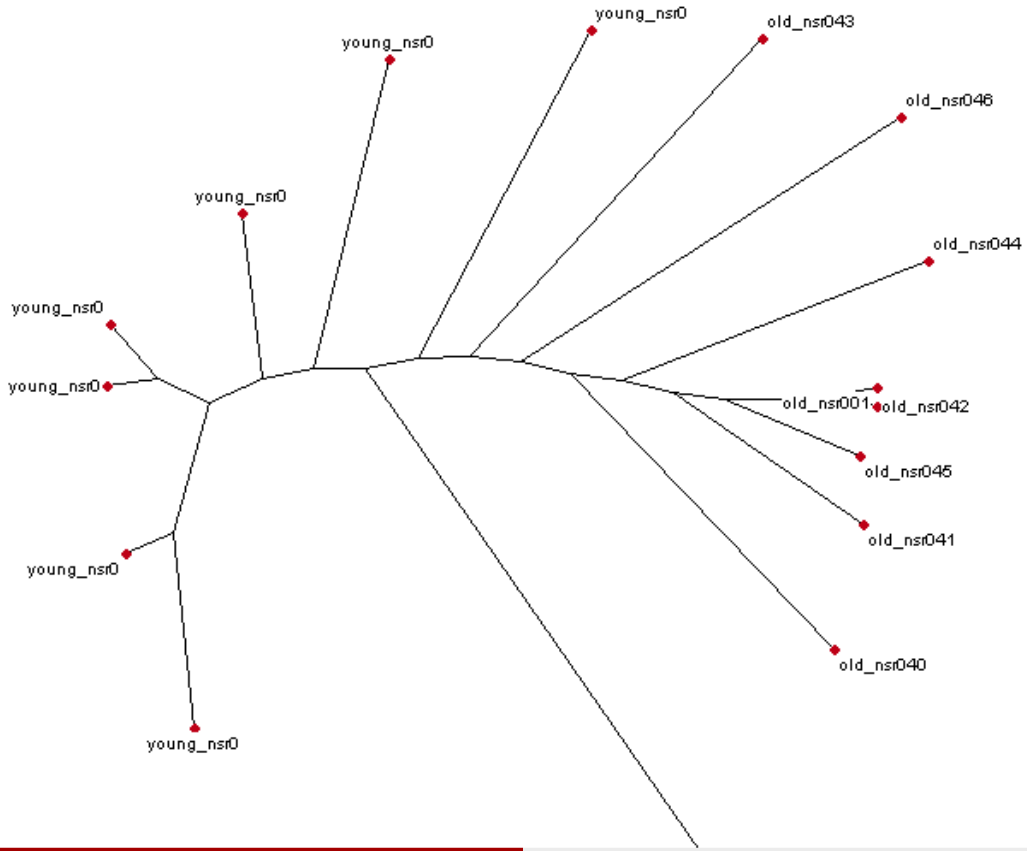
young v.s. old

	old_nsr001	old_nsr040	old_nsr041	old_nsr042	old_nsr043	old_nsr044	old_nsr045	old_nsr046	young_nsr047	young_nsr048	young_nsr049	young_nsr050	young_nsr051	young_nsr052	young_nsr053	young_nsr054
old_nsr001	0,000367	0,952024	0,95061	0,945933	0,949736	0,950958	0,948364	0,954146	0,949841	0,954005	0,955413	0,948662	0,953394	0,950865	0,9557	0,955683
old_nsr040	0,952024	0,00034	0,953121	0,951928	0,946163	0,949815	0,953287	0,949094	0,958728	0,962704	0,960268	0,955893	0,962963	0,956731	0,969789	0,96495
old_nsr041	0,95061	0,953121	0,000377	0,951346	0,949566	0,949914	0,950161	0,955327	0,955474	0,958248	0,958666	0,952243	0,955644	0,952609	0,958549	0,953429
old_nsr042	0,945933	0,951928	0,951346	0,000345	0,949109	0,951131	0,946419	0,951499	0,946238	0,951136	0,949203	0,947774	0,947583	0,949594	0,951523	0,952853
old_nsr043	0,949736	0,946163	0,949566	0,949109	0,000378	0,948612	0,951932	0,947139	0,954896	0,957349	0,957155	0,95058	0,96005	0,951108	0,964736	0,960688
old_nsr044	0,950958	0,949815	0,949914	0,951131	0,948612	0,000348	0,950339	0,951358	0,956038	0,958926	0,959561	0,95272	0,958239	0,954904	0,962013	0,959726
old_nsr045	0,948364	0,953287	0,950161	0,946419	0,951932	0,950339	0,000404	0,954754	0,949637	0,956214	0,953575	0,95223	0,950636	0,952009	0,953416	0,956672
old_nsr046	0,954146	0,949094	0,955327	0,951499	0,947139	0,951358	0,954754	0,000412	0,9564	0,960209	0,956874	0,954809	0,960467	0,955471	0,967215	0,96261
young_nsr047	0,949841	0,958728	0,955474	0,946238	0,954896	0,956038	0,949637	0,9564	0,000335	0,948902	0,945155	0,950148	0,946676	0,951359	0,948919	0,950762
young_nsr048	0,954005	0,962704	0,958248	0,951136	0,957349	0,958926	0,956214	0,960209	0,948902	0,000327	0,947119	0,948525	0,952556	0,949741	0,951433	0,950291
young_nsr049	0,955413	0,960268	0,958666	0,949203	0,957155	0,959561	0,953575	0,956874	0,945155	0,947119	0,000367	0,952881	0,949172	0,951173	0,951686	0,951869
young_nsr050	0,948662	0,955893	0,952243	0,947774	0,95058	0,95272	0,95223	0,954809	0,950148	0,948525	0,952881	0,000302	0,955884	0,945132	0,954209	0,951154
young_nsr051	0,953394	0,962963	0,955644	0,947583	0,96005	0,958239	0,950636	0,960467	0,946676	0,952556	0,949172	0,955884	0,000359	0,953737	0,948845	0,951921
young_nsr052	0,950865	0,956731	0,952609	0,949594	0,951108	0,954904	0,952009	0,955471	0,951359	0,949741	0,951173	0,945132	0,953737	0,000331	0,95448	0,951443
young_nsr053	0,9557	0,969789	0,958549	0,951523	0,964736	0,962013	0,953416	0,967215	0,948919	0,951433	0,951686	0,954209	0,948845	0,95448	0,00038	0,947724
young_nsr054	0,955683	0,96495	0,953429	0,952853	0,960688	0,959726	0,956672	0,96261	0,950762	0,950291	0,951869	0,951154	0,951921	0,951443	0,947724	0,00032
olds	0,950253	0,950776	0,951435	0,949623571	0,948893857	0,950303857	0,950750857	0,951902429	0,9534065	0,957348875	0,956339375	0,951863875	0,956122	0,952911375	0,960367625	0,958326375
youngs	0,952945375	0,96150325	0,95560775	0,949488	0,95707025	0,957765875	0,953048625	0,959256875	0,948845857	0,949795286	0,949865	0,951133286	0,951255857	0,951009286	0,951042286	0,950737714

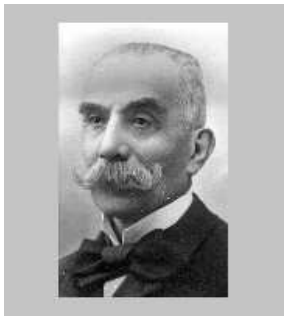
young v.s. old

	old_nsr001	old_nsr040	old_nsr041	old_nsr042	old_nsr043	old_nsr044	old_nsr045	old_nsr046	young_nsr047	young_nsr
old_nsr001	0,000367	0,952024	0,95061	0,945933	0,949736	0,950958	0,948364	0,954146	0,949841	0,95400
old_nsr040	0,952024	0,00034	0,953121	0,951928	0,946163	0,949815	0,953287	0,949094	0,958728	0,96270
old_nsr041	0,95061	0,953121	0,000377	0,951346	0,949566	0,949914	0,950161	0,955327	0,955474	0,95824
old_nsr042	0,945933	0,951928	0,951346	0,000345	0,949109	0,951131	0,946419	0,951499	0,946238	0,95113
old_nsr043	0,949736	0,946163	0,949566	0,949109	0,000378	0,948612	0,951932	0,947139	0,954896	0,95734
old_nsr044	0,950958	0,949815	0,949914	0,951131	0,948612	0,000348	0,950339	0,951358	0,956038	0,95892
old_nsr045	0,948364	0,953287	0,950161	0,946419	0,951932	0,950339	0,000404	0,954754	0,949637	0,95621
old_nsr046	0,954146	0,949094	0,955327	0,951499	0,947139	0,951358	0,954754	0,000412	0,9564	0,96020
young_nsr047	0,949841	0,958728	0,955474	0,946238	0,954896	0,956038	0,949637	0,9564	0,000335	0,94890
young_nsr048	0,954005	0,962704	0,958248	0,951136	0,957349	0,958926	0,956214	0,960209	0,948902	0,00032
young_nsr049	0,955413	0,960268	0,958666	0,949203	0,957155	0,959561	0,953575	0,956874	0,945155	0,94711
young_nsr050	0,948662	0,955893	0,952243	0,947774	0,95058	0,95272	0,95223	0,954809	0,950148	0,94852
young_nsr051	0,953394	0,962963	0,955644	0,947583	0,96005	0,958239	0,950636	0,960467	0,946676	0,95255
young_nsr052	0,950865	0,956731	0,952609	0,949594	0,951108	0,954904	0,952009	0,955471	0,951359	0,94974
young_nsr053	0,9557	0,969789	0,958549	0,951523	0,964736	0,962013	0,953416	0,967215	0,948919	0,95143
young_nsr054	0,955683	0,96495	0,953429	0,952853	0,960688	0,959726	0,956672	0,96261	0,950762	0,95029
olds	0,950253	0,950776	0,951435	0,949623571	0,948893857	0,950303857	0,950750857	0,951902429	0,9534065	0,957348
youngs	0,952945375	0,96150325	0,95560775	0,949488	0,95707025	0,957765875	0,953048625	0,959256875	0,948845857	0,949795

young v.s. old


young_nsr0
young_nsr0
young_nsr0
old_nsr043
old_nsr046
young_nsr0
old_nsr044
young_nsr0
young_nsr0
old_nsr001
old_nsr042
old_nsr045
old_nsr041
young_nsr0
old_nsr040
young_nsr0

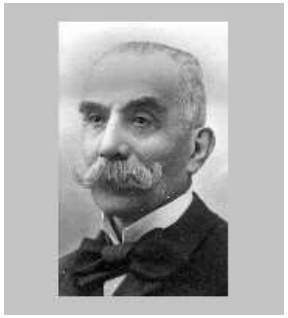
Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica. E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

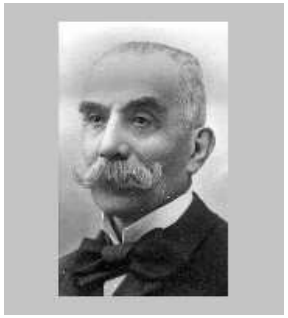
Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica.E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

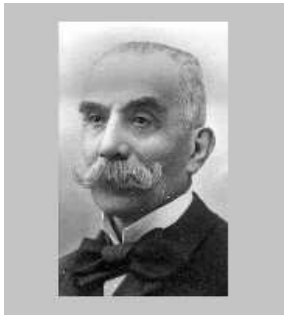
Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica.E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

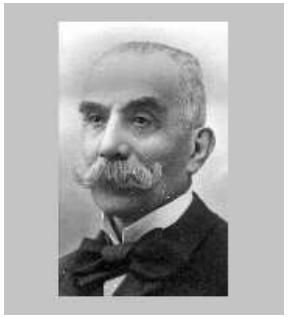
Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica.E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

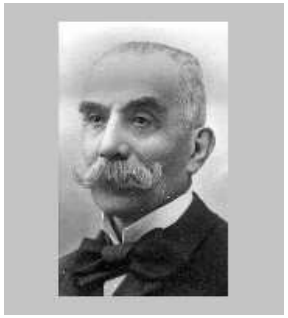
Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica.E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*



L'assunto di Mariotti è espresso nella dedica del volume:

La Retorica ha detto tanto bene di Dante, che io ebbi vaghezza di sapere che cosa ne pensasse l'Aritmetica, chiamata con verità dal Gibbon la nemica naturale della Retorica.E l'Aritmetica ne dice meglio che mai; com' ebbi a ragionarne all'Accademia dei Lincei, e come più distesamente ne ho scritto in questo libretto che forse non riuscirà disutile per la scienza e l'arte.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

La musica delle parole ha le sue leggi. Perchè non si cercano ?

Il Drobisch, filosofo tedesco, ha fatto uno studio della metrica di Virgilio col paragone di quella di Lucrezio e di altri, procedendo statisticamente e contando nei versi le varie ricorrenze e successioni dei piedi dattili e spondei, e delle cesure.

Il simile non si potrebbe fare per la Divina Commedia, comparandola con altri poemi?

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

La musica delle parole ha le sue leggi. Perchè non si cercano ?

Il Drobisch, filosofo tedesco, ha fatto uno studio della metrica di Virgilio col paragone di quella di Lucrezio e di altri, procedendo statisticamente e contando nei versi le varie ricorrenze e successioni dei piedi dattili e spondei, e delle cesure.

Il simile non si potrebbe fare per la Divina Commedia, comparandola con altri poemi?

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

La musica delle parole ha le sue leggi. Perchè non si cercano ?

Il Drobisch, filosofo tedesco, ha fatto uno studio della metrica di Virgilio col paragone di quella di Lucrezio e di altri, procedendo statisticamente e contando nei versi le varie ricorrenze e successioni dei piedi dattili e spondei, e delle cesure.

Il simile non si potrebbe fare per la Divina Commedia, comparandola con altri poemi?

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

La musica delle parole ha le sue leggi. Perchè non si cercano ?

Il Drobisch, filosofo tedesco, ha fatto uno studio della metrica di Virgilio col paragone di quella di Lucrezio e di altri, procedendo statisticamente e contando nei versi le varie ricorrenze e successioni dei piedi dattili e spondei, e delle cesure.

Il simile non si potrebbe fare per la Divina Commedia, comparandola con altri poemi?

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

La musica delle parole ha le sue leggi. Perchè non si cercano ?

Il Drobisch, filosofo tedesco, ha fatto uno studio della metrica di Virgilio col paragone di quella di Lucrezio e di altri, procedendo statisticamente e contando nei versi le varie ricorrenze e successioni dei piedi dattili e spondei, e delle cesure.

Il simile non si potrebbe fare per la Divina Commedia, comparandola con altri poemi?

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Dopo varie considerazioni di portata generale Mariotti inizia ad entrare nella statistica della Divina Commedia:

La geometria della Divina Commedia parte è meditata,parte è usata, senza che Dante stesso ne sia consapevole.

Il poema deve essere di tre cantiche, pensava Dante; i canti devono essere 100: 34 nell' Inferno, 33 nel Purgatorio, e 33 nel Paradiso. ... I 100 canti sono di vario numero di versi ... Ma in ogni cantica si ha, ragguagliatamente, un eguale numero di versi: nell' Inferno 4720, nel Purgatorio 4755, nel Paradiso 4758, che insieme fanno 14,233.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

Mariotti continua...:

Quante sono le parole che compongono il poema ?

Se uno contasse, ad una ad una, le parole di tutto quanto il poema, quante ne avrebbe?

Forse piacerà di saperlo.

L'Inferno ha 33,444 parole, il Purgatorio ha 33,379 parole, il Paradiso ha 32,719 parole. Tutto il poema ha 99,542 parole

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

E' molto interessante la sua descrizione del modo in cui effettuò i conteggi:

Si dirà siffatto conteggiare è esatto? Come è stato eseguito? Col copiare tutto il poema, ma per modo che, notomizzandolo, si potesse vedere la composizione delle parti del discorso e le loro proporzioni.Nel margine a sinistra di un foglio di carta ho notato le parti del discorso, cominciando, per comodo, dall' articolo e giù di grado in grado fino all' interiezione; talchè tutti gli articoli, i segnacasi, i vari pronomi, i nomi, i verbi nel loro vario essere, in somma tutte le parole dell'istessa natura si trovassero insieme al posto assegnato.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

E' molto interessante la sua descrizione del modo in cui effettuò i conteggi:

Si dirà siffatto conteggiare è esatto? Come è stato eseguito? Col copiare tutto il poema, ma per modo che, notomizzandolo, si potesse vedere la composizione delle parti del discorso e le loro proporzioni.Nel margine a sinistra di un foglio di carta ho notato le parti del discorso, cominciando, per comodo, dall' articolo e giù di grado in grado fino all' interiezione; talchè tutti gli articoli, i segnacasi, i vari pronomi, i nomi, i verbi nel loro vario essere, in somma tutte le parole dell'istessa natura si trovassero insieme al posto assegnato.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

E' molto interessante la sua descrizione del modo in cui effettuò i conteggi:

Si dirà siffatto conteggiare è esatto? Come è stato eseguito? Col copiare tutto il poema, ma per modo che, notomizzandolo, si potesse vedere la composizione delle parti del discorso e le loro proporzioni.Nel margine a sinistra di un foglio di carta ho notato le parti del discorso, cominciando, per comodo, dall' articolo e giù di grado in grado fino all' interiezione; talchè tutti gli articoli, i segnacasi, i vari pronomi, i nomi, i verbi nel loro vario essere, in somma tutte le parole dell'istessa natura si trovassero insieme al posto assegnato.

Filippo Mariotti (1833-1911):*Dante e la statistica delle lingue*

E' molto interessante la sua descrizione del modo in cui effettuò i conteggi:

Si dirà siffatto conteggiare è esatto? Come è stato eseguito? Col copiare tutto il poema, ma per modo che, notomizzandolo, si potesse vedere la composizione delle parti del discorso e le loro proporzioni.Nel margine a sinistra di un foglio di carta ho notato le parti del discorso, cominciando, per comodo, dall' articolo e giù di grado in grado fino all' interiezione; talchè tutti gli articoli, i segnacasi, i vari pronomi, i nomi, i verbi nel loro vario essere, in somma tutte le parole dell'istessa natura si trovassero insieme al posto assegnato.

una Tabella di Filippo Mariotti

	Inferno.	Purgatorio.	Paradiso.
1 Articoli	2,188	1,973	1,993
2 Articoli indeterminati.	237	189	133
3 Segnacasi.	1,495	1,598	1,522
4 Segnacasi articolati. .	870	883	1,007
5 Pronomi dimostr. sost.	1,416	1,199	1,123
6 Pronomi dimostr. agg.	861	791	1,016
7 Pronomi possessivi. . .	505	668	788
8 Pronomi personali. . .	1,701	1,651	1,247
9 Pronomi relativi	1,013	990	1,224
10 Participi	310	269	252
11 Gerundi.	203	285	206
12 Avverbi	2,596	2,849	2,471
13 Avverbi di due parole.	715	649	594
14 Particelle negative. . .	575	576	533
15 Preposizioni	1,268	1,467	1,380
16 Preposizioni articolate.	282	262	307
17 Preposizioni di due parole.	135	50	42
18 Congiunzioni	1,044	983	909
19 Congiunzione *e*	1,451	1,387	1,343
20 Nomi sostantivi	6,082	5,894	6,004
21 Aggettivi	1,866	2,069	2,280
22 Verbi	4,739	4,717	4,202
23 Verbi di due parole. .	1,866	1,966	2,128
24 Interiezioni.	26	14	5
	33,441	33,379	32,719

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle sue parole:

Ora questi numeri che dicono o che rispondono?

Innanzi tutto dimostrano una proporzione continua e uniforme delle parti del discorso nei canti del poema;

il che si conosce dividendo per cento i numeri dell' ultima colonna, e che meglio si vede scorrendo coll'occhio per quelle file di numeri negli specchietti delle tre cantiche, ...

Certo è che Dante, senza avvedersene, adoperava va un numero proporzionato e quasi uguale di sostantivi, di aggettivi, di verbi per ogni canto, e perfino della congiunzione e.

Ciò dà forse indizio di una legge della mente umana: che per natura, per educazione, per abito opera con regola costante ? Quali sono veramente le cause di questi fatti ?

Lo scopo della sua ricerca, dalle parole del Mariotti

Il cervello umano non può essere soggetto a certi esperimenti, che mostrino l'indole sua, perchè a volere sperimentare bisognerebbe disfarlo; e pero' conviene che sia studiato negli effetti. Ora Dante ha detto che sempre scriveva con ardore: ... " Io mi son un che, quando Amore spira, noto; ed a quel modo, Che detta dentro, vo significando".

Questo ardore è costante in tutti i canti, come è perenne e uniforme la proporzione delle parole, con cui li compone.

Lo scopo della sua ricerca, dalle parole del Mariotti

Il cervello umano non può essere soggetto a certi esperimenti, che mostrino l'indole sua, perchè a volere sperimentare bisognerebbe disfarlo; e pero' conviene che sia studiato negli effetti. Ora Dante ha detto che sempre scriveva con ardore: ... " Io mi son un che, quando Amore spira, noto; ed a quel modo, Che detta dentro, vo significando".

Questo ardore è costante in tutti i canti, come è perenne e uniforme la proporzione delle parole, con cui li compone.

Lo scopo della sua ricerca, dalle parole del Mariotti

Il cervello umano non può essere soggetto a certi esperimenti, che mostrino l'indole sua, perchè a volere sperimentare bisognerebbe disfarlo; e pero' conviene che sia studiato negli effetti. Ora Dante ha detto che sempre scriveva con ardore: ... " Io mi son un che, quando Amore spira, noto; ed a quel modo, Che detta dentro, vo significando".

Questo ardore è costante in tutti i canti, come è perenne e uniforme la proporzione delle parole, con cui li compone.

Lo scopo della sua ricerca, dalle parole del Mariotti

Il cervello umano non può essere soggetto a certi esperimenti, che mostrino l'indole sua, perchè a volere sperimentare bisognerebbe disfarlo; e pero' conviene che sia studiato negli effetti. Ora Dante ha detto che sempre scriveva con ardore: ... " Io mi son un che, quando Amore spira, noto; ed a quel modo, Che detta dentro, vo significando".

Questo ardore è costante in tutti i canti, come è perenne e uniforme la proporzione delle parole, con cui li compone.

Lo scopo della sua ricerca, dalle parole del Mariotti

Il cervello umano non può essere soggetto a certi esperimenti, che mostrino l'indole sua, perchè a volere sperimentare bisognerebbe disfarlo; e pero' conviene che sia studiato negli effetti. Ora Dante ha detto che sempre scriveva con ardore: ... " Io mi son un che, quando Amore spira, noto; ed a quel modo, Che detta dentro, vo significando".

Questo ardore è costante in tutti i canti, come è perenne e uniforme la proporzione delle parole, con cui li compone.

Le conclusioni del Mariotti...

L'ammirata precisione del pensiero dantesco, conosciuta per mezzo dei numeri, può invitare gli studiosi ad applicare il computo delle quantità e perciò la statistica delle lingue, alle opere dei parlatori e degli scrittori, per trarne utili conseguenze di scienza e d'arte.

Un'osservazione preliminare

Qualunque *misura quantitativa* è per definizione suscettibile di trattamento matematico; qui però ci limitiamo agli strumenti matematici che abbiano due caratteristiche:

1. sono astratti: trattano cioè il testo come una sequenza di simboli, e dunque non solo non prendono in considerazione aspetti grammaticali, ma non distinguono, nelle modalità di trattamento, le lettere dai simboli di punteggiatura o di spaziatura;
2. hanno delle motivazioni "matematiche" per sperare nella propria efficacia: dietro al metodo c'è un modello matematico che potrebbe essere adatto a descrivere i testi, e se anche non lo fosse potrebbe comunque dare indicazioni utili.

Un'osservazione preliminare

Qualunque *misura quantitativa* è per definizione suscettibile di trattamento matematico; qui però ci limitiamo agli strumenti matematici che abbiano due caratteristiche:

1. sono astratti: trattano cioè il testo come una sequenza di simboli, e dunque non solo non prendono in considerazione aspetti grammaticali, ma non distinguono, nelle modalità di trattamento, le lettere dai simboli di punteggiatura o di spaziatura;
2. hanno delle motivazioni "matematiche" per sperare nella propria efficacia: dietro al metodo c'è un modello matematico che potrebbe essere adatto a descrivere i testi, e se anche non lo fosse potrebbe comunque dare indicazioni utili.

Un'osservazione preliminare

Qualunque *misura quantitativa* è per definizione suscettibile di trattamento matematico; qui però ci limitiamo agli strumenti matematici che abbiano due caratteristiche:

1. sono astratti: trattano cioè il testo come una sequenza di simboli, e dunque non solo non prendono in considerazione aspetti grammaticali, ma non distinguono, nelle modalità di trattamento, le lettere dai simboli di punteggiatura o di spaziatura;
2. hanno delle motivazioni "matematiche" per sperare nella propria efficacia: dietro al metodo c'è un modello matematico che potrebbe essere adatto a descrivere i testi, e se anche non lo fosse potrebbe comunque dare indicazioni utili.

Assunto metodologico

In accordo con questi punti, il testo viene considerato come una sequenza astratta di simboli: le 26+26 lettere (minuscole e maiuscole), le vocali accentate, (e se servono anche gli opportuni simboli per lettere di altri alfabeti), le cifre, i segni di interpunzione, le parantesi, lo spazio, il simbolo di a-capo.

Inoltre le caratteristiche che verranno misurate saranno legate non alle frequenze delle parole ma alle frequenze di successioni di simboli: gli n-grammi.

Assunto metodologico

In accordo con questi punti, il testo viene considerato come una sequenza astratta di simboli: le 26+26 lettere (minuscole e maiuscole), le vocali accentate, (e se servono anche gli opportuni simboli per lettere di altri alfabeti), le cifre, i segni di interpunzione, le parantesi, lo spazio, il simbolo di a-capo.

Inoltre le caratteristiche che verranno misurate saranno legate non alle frequenze delle parole ma alle frequenze di successioni di simboli: gli n-grammi.

Assunto metodologico

In accordo con questi punti, il testo viene considerato come una sequenza astratta di simboli: le 26+26 lettere (minuscole e maiuscole), le vocali accentate, (e se servono anche gli opportuni simboli per lettere di altri alfabeti), le cifre, i segni di interpunzione, le parantesi, lo spazio, il simbolo di a-capo.

Inoltre le caratteristiche che verranno misurate saranno legate non alle frequenze delle parole ma alle frequenze di successioni di simboli: gli n-grammi.

Assunto metodologico

In accordo con questi punti, il testo viene considerato come una sequenza astratta di simboli: le 26+26 lettere (minuscole e maiuscole), le vocali accentate, (e se servono anche gli opportuni simboli per lettere di altri alfabeti), le cifre, i segni di interpunzione, le parantesi, lo spazio, il simbolo di a-capo.

Inoltre le caratteristiche che verranno misurate saranno legate non alle frequenze delle parole ma alle frequenze di successioni di simboli: gli n-grammi.

Per un matematico analizzare dei testi vuol dire tre cose:

1. pensare, in modo del tutto irragionevole ma matematicamente suggestivo, che l'autore è un generatore matematico di simboli (d'ora in poi "sorgente"), e che i suoi testi disponibili sono solo "esempi casualmente generati" (in termini probabilistici un "campione")
2. se una qualche struttura matematico/probabilistica esiste per l'autore come sorgente o per il singolo testo, essa determina quantitativamente tutti gli oggetti statisticamente misurabili nel testo
3. attraverso le misure di tali quantità si può dunque risalire alle caratteristiche della sorgente/autore

Per un matematico analizzare dei testi vuol dire tre cose:

1. pensare, in modo del tutto irragionevole ma matematicamente suggestivo, che l'autore è un generatore matematico di simboli (d'ora in poi "sorgente"), e che i suoi testi disponibili sono solo "esempi casualmente generati" (in termini probabilistici un "campione")
2. se una qualche struttura matematico/probabilistica esiste per l'autore come sorgente o per il singolo testo, essa determina quantitativamente tutti gli oggetti statisticamente misurabili nel testo
3. attraverso le misure di tali quantità si può dunque risalire alle caratteristiche della sorgente/autore

Per un matematico analizzare dei testi vuol dire tre cose:

1. pensare, in modo del tutto irragionevole ma matematicamente suggestivo, che l'autore è un generatore matematico di simboli (d'ora in poi "sorgente"), e che i suoi testi disponibili sono solo "esempi casualmente generati" (in termini probabilistici un "campione")
2. se una qualche struttura matematico/probabilistica esiste per l'autore come sorgente o per il singolo testo, essa determina quantitativamente tutti gli oggetti statisticamente misurabili nel testo
3. attraverso le misure di tali quantità si può dunque risalire alle caratteristiche della sorgente/autore

Scrittori Markoviani ?

- Una approssimazione di “ordine 0”:

  mZmJMux,1UrsN.ul3HEpf7.hy!7WForèÈ;1tSàgMfÈFXsa7WX9FXfùrOO

- L’approssimazione al “primo ordine”:

  illfmbaoaocnn e aai,sfrmrta eeoiddmaoo’iVar legeq arnoh everl dl slB lanl

- Approssimazione di “secondo ordine” (catena Markoviana): il nuovo carattere si ottiene scegliendolo in funzione del precedente.

  Loncueresono astantà chedali co le prora Lafra Seoccoro do li, fi dunqu No o ch

Scrittori Markoviani ?

- Una approssimazione di “ordine 0”:

  mZmJMux,1UrsN.ul3HEpf7.hy!7WForèÈ;1tSàgMfÈFXsa7WX9FXfùrOO

- L’approssimazione al “primo ordine”:

  illfmbaoaocnn e aai,sfrmrta eeoiddmaoo’iVar legeq arnoh everl dl slB lanl

- Approssimazione di “secondo ordine” (catena Markoviana): il nuovo carattere si ottiene scegliendolo in funzione del precedente.

  Loncueresono astantà chedali co le prora Lafra Seoccoro do li, fi dunqu No o ch

Scrittori Markoviani ?

- Una approssimazione di “ordine 0”:

  mZmJMux,1UrsN.ul3HEpf7.hy!7WForèÈ;1tSàgMfÈFXsa7WX9FXfùrOO

- L’approssimazione al “primo ordine”:

  illfmbaoaocnn e aai,sfrmrta eeoiddmaoo’iVar legeq arnoh everl dl slB lanl

- Approssimazione di “secondo ordine” (catena Markoviana): il nuovo carattere si ottiene scegliendolo in funzione del precedente.

  Loncueresono astantà chedali co le prora Lafra Seoccoro do li, fi dunqu No o ch

Scrittori Markoviani ?

- ecco un esempio di testo generato con un modello del decimo ordine

  La pietra fondamentale nel contegno delle due alleate, quando si è convertito, è sempre da creare

- Nell'approssimazione del primo ordine la divisione in parole somiglia a quella della lingua italiana; in quella del secondo ordine le sillabe sono sostanzialmente corrette, e sono credibili l'inizio e la fine delle parole; l'approssimazione di ordine dieci riproduce le singole parole e rispetta le regole grammaticali.
- Si può supporre, e molti lo hanno infatti supposto, che le differenze "stilistiche" tra autori debbano tradursi in differenze numeriche per le frequenze degli n-grammi.

Scrittori Markoviani ?

- ecco un esempio di testo generato con un modello del decimo ordine

  La pietra fondamentale nel contegno delle due alleate, quando si è convertito, è sempre da creare

- Nell'approssimazione del primo ordine la divisione in parole somiglia a quella della lingua italiana; in quella del secondo ordine le sillabe sono sostanzialmente corrette, e sono credibili l'inizio e la fine delle parole; l'approssimazione di ordine dieci riproduce le singole parole e rispetta le regole grammaticali.
- Si può supporre, e molti lo hanno infatti supposto, che le differenze "stilistiche" tra autori debbano tradursi in differenze numeriche per le frequenze degli n-grammi.

Scrittori Markoviani ?

- ecco un esempio di testo generato con un modello del decimo ordine

  La pietra fondamentale nel contegno delle due alleate, quando si è convertito, è sempre da creare

- Nell'approssimazione del primo ordine la divisione in parole somiglia a quella della lingua italiana; in quella del secondo ordine le sillabe sono sostanzialmente corrette, e sono credibili l'inizio e la fine delle parole; l'approssimazione di ordine dieci riproduce le singole parole e rispetta le regole grammaticali.
- Si può supporre, e molti lo hanno infatti supposto, che le differenze "stilistiche" tra autori debbano tradursi in differenze numeriche per le frequenze degli n-grammi.

Scrittori Markoviani ?

- ecco un esempio di testo generato con un modello del decimo ordine

  La pietra fondamentale nel contegno delle due alleate, quando si è convertito, è sempre da creare

- Nell'approssimazione del primo ordine la divisione in parole somiglia a quella della lingua italiana; in quella del secondo ordine le sillabe sono sostanzialmente corrette, e sono credibili l'inizio e la fine delle parole; l'approssimazione di ordine dieci riproduce le singole parole e rispetta le regole grammaticali.
- Si può supporre, e molti lo hanno infatti supposto, che le differenze "stilistiche" tra autori debbano tradursi in differenze numeriche per le frequenze degli n-grammi.

Authorship Attribution

- D. Benedetto, E. Caglioti, V. Loreto “Language Tree and Zipping”, Physical Review Letters **88**, no.4 (2002)

Verga Giovanni:Eros
Verga Giovanni:Eva
Verga Giovanni: La lupa
Verga Giovanni: Tigre reale
Verga Giovanni: Tutte le novelle
Verga Giovanni: Una peccatrice
Svevo Italo: Corto viaggio sperimentale
Svevo Italo: La coscienza di Zeno
Svevo Italo: La novella del buon vecchio e ...
Svevo Italo: Senilità
Svevo Italo:Una vita
Salgari Emilio: Gli ultimi filibustieri
Salgari Emilio: I misteri della jungla nera
Salgari Emilio:I pirati della Malesia
Salgari Emilio: Il figlio del Corsaro Rosso
Salgari Emilio: Jolanda la figlia del Corsaro Nero
Salgari Emilio:Le due tigri
Salgari Emilio: Le novelle marinaresche di mastro Catrame

Tozzi Federigo: Bestie
Tozzi Federigo: Con gli occhi chiusi
Tozzi Federigo: Il podere
Tozzi Federigo: L’amore
Tozzi Federigo: Novale
Tozzi Federigo: Tre croci
Pirandello Luigi:.....
Petrarca Francesco:.....
Manzoni Alessandro:.....
Machiavelli Niccolo’:.....
Guicciardini Francesco:.....
Goldoni Carlo:.....
Fogazzaro Antonio:.....
Deledda Grazia:.....
De Sanctis Francesco:.....
De Amicis Edmondo:.....
D’Annunzio Gabriele:....
Alighieri Dante:....

Authorship Attribution

	Il bugiardo
La bancarotta	0,926528
La bottega del caffè	0,935032
La buona moglie	0,936331
Il fiasco del maestro Chieco	0,941575
Giovanni Episcopo	0,943557
Clizia	0,944401
Schopenhauer e Leopardi	0,944434
...66 brani di 30 diversi autori	

REGOLA DEL MAX

	Il bugiardo
La bancarotta	0,864323
La bottega del caffè	0,877063
La buona moglie	0,892676
Giovanni Episcopo	0,893
Il fiasco del maestro Chieco	0,896035
Bestie	0,900281
Il conte di Carmagnola	0,902791
...66 brani di 30 diversi autori	

Authorship Attribution

	CONSOLATORIA
Ricordi	0,926361
Discorsi politici	0,931585
Considerazioni intorno ai discorsi del Machiavelli	0,933536
Principe	0,940348
Memorie di famiglia	0,943995
Dell'arte della guerra	0,947862
La monaca di Monza	0,949762
Il conte di Carmagnola	0,95162
Amore e ginnastica	0,952718
Clizia	0,953221
...66 brani di 30 diversi autori	

REGOLA DEL MAX

	CONSOLATORIA
Ricordi	0,879473
Discorsi politici	0,885806
Principe	0,902807
Considerazioni intorno ai discorsi del Machiavelli	0,903456
La monaca di Monza	0,913806
Amore e ginnastica	0,914569
Una peccatrice	0,915366
Il podere	0,917611
Dell'arte della guerra	0,918185
Corto viaggio sperimentale	0,920031
Le novelle marinaresche di mastro Catrame	0,921291
...66 brani di 30 diversi autori	

Authorship Attribution

	TIGRE REALE
Eva	0,917454
Eros	0,917882
Una peccatrice	0,924972
Giovanni Episcopo	0,930097
Amore e ginnastica	0,930228
Il fiasco del maestro Chieco	0,930319
L'amore	0,930405
Corto viaggio sperimentale	0,933624
Elias Portolu	0,934806
Il libro delle vergini	0,935552
...66 brani di 30 diversi autori	

REGOLA DEL MAX

	TIGRE REALE
Eva	0,85592
Una peccatrice	0,872881
Corto viaggio sperimentale	0,87339
L'amore	0,875529
Eros	0,875805
Amore e ginnastica	0,877452
Il libro delle vergini	0,878297
La monaca di Monza	0,882008
Le novelle marinaresche di mastro Catrame	0,890908
Giovanni Episcopo	0,892107
...66 brani di 30 diversi autori	

Gramsci 1

A. Gramsci (1891-1937), Giornalista e Fondatore del Partito Comunista Italiano

- Iniziamo con un test controllato, dove abbiamo essenzialmente combinato insieme le distanze precedentemente descritte,oltre ad un opportuno sistema di voto.
- Definendo un opportuno Indice di Gramscianità , abbiamo poi provato a discriminare gli articoli di Gramsci (BLUE) da quelli NON Gramsciani (RED)

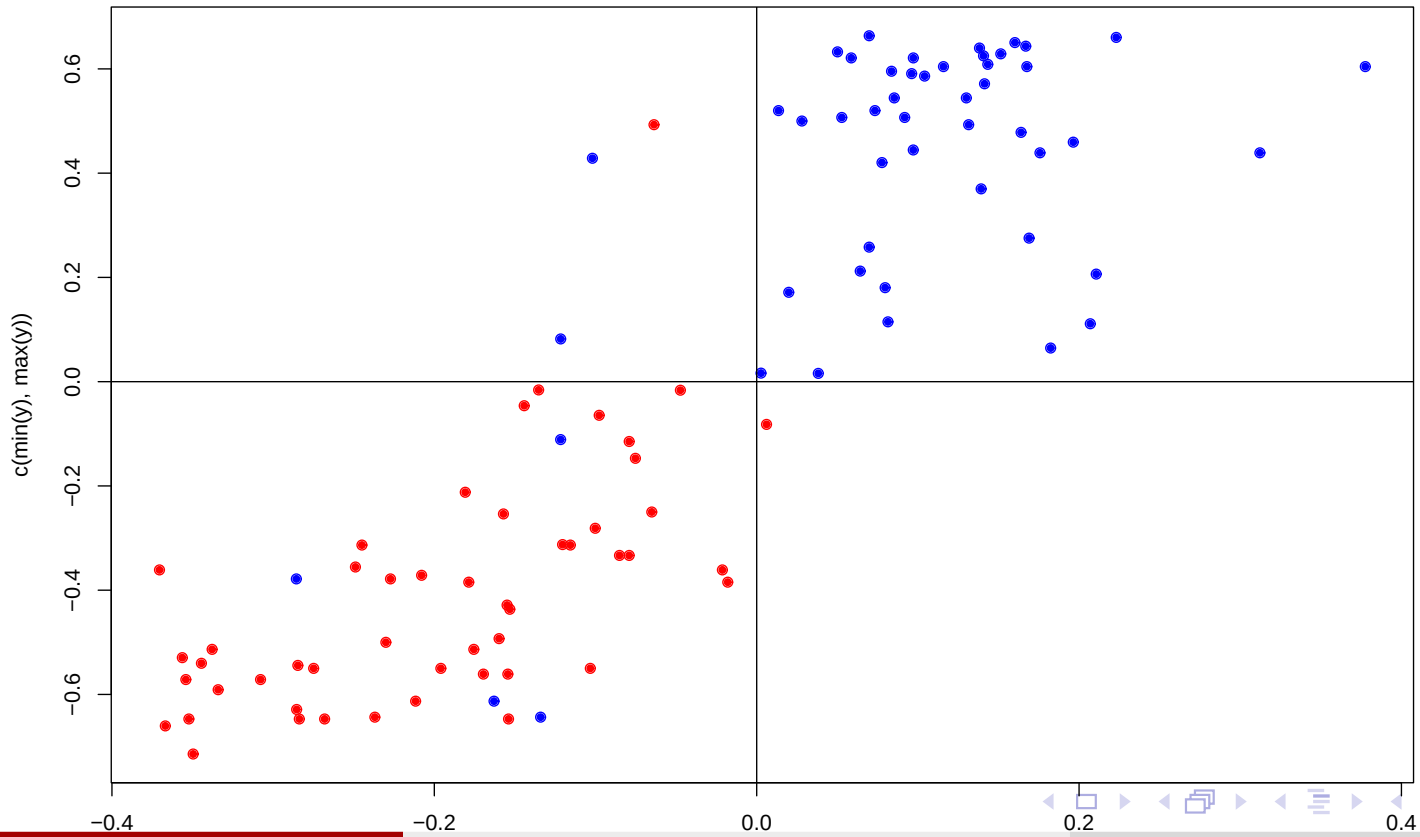
Gramsci 1

A. Gramsci (1891-1937), Giornalista e Fondatore del Partito Comunista Italiano

- Iniziamo con un test controllato, dove abbiamo essenzialmente combinato insieme le distanze precedentemente descritte,oltre ad un opportuno sistema di voto.
- Definendo un opportuno Indice di Gramscianità , abbiamo poi provato a discriminare gli articoli di Gramsci (BLUE) da quelli NON Gramsciani (RED)

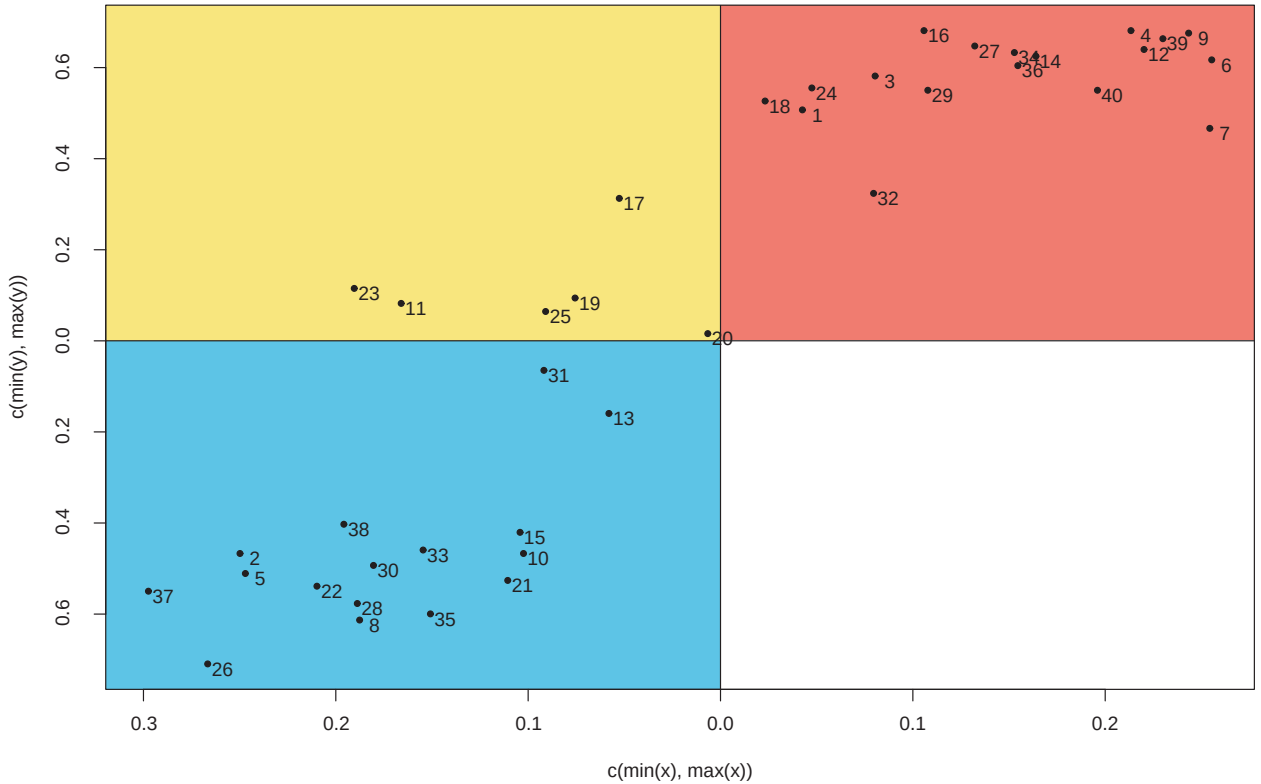
un Test controllato

A. Gramsci (1891-1937)



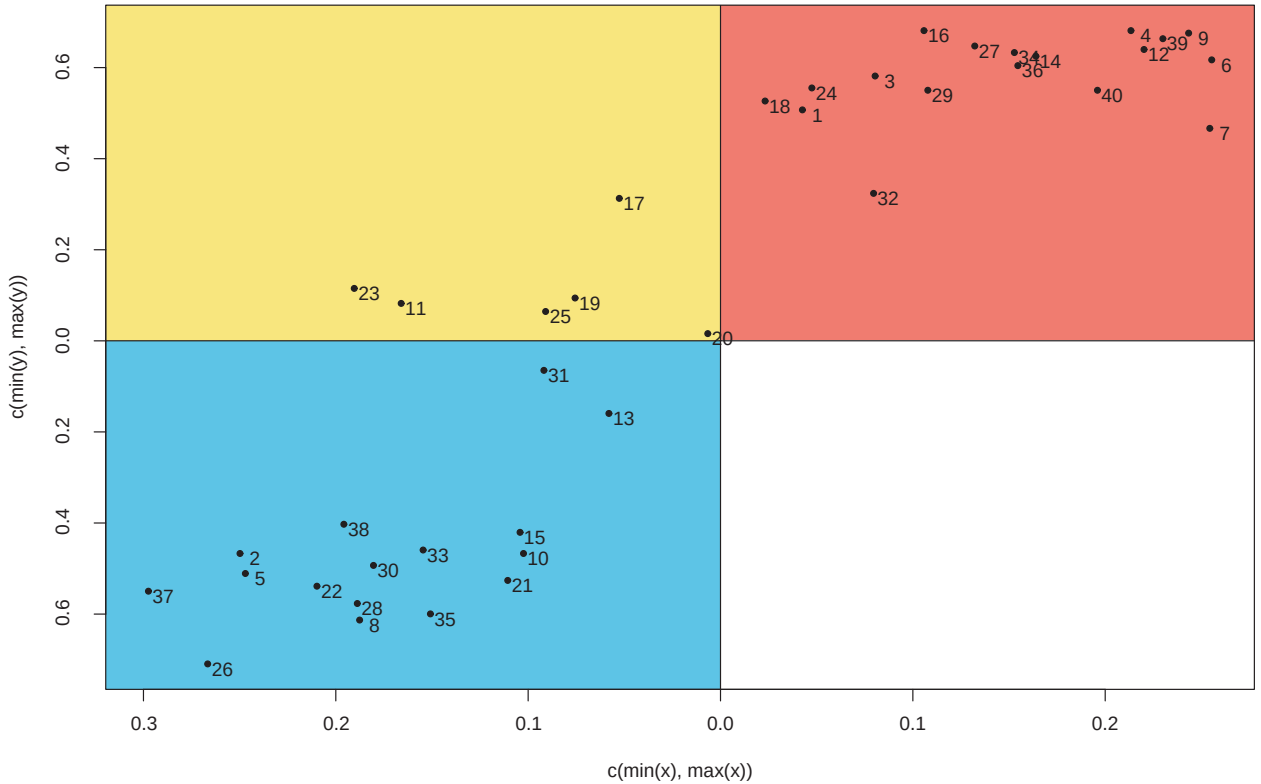
un Test Cieco

A. Gramsci (1891-1937)

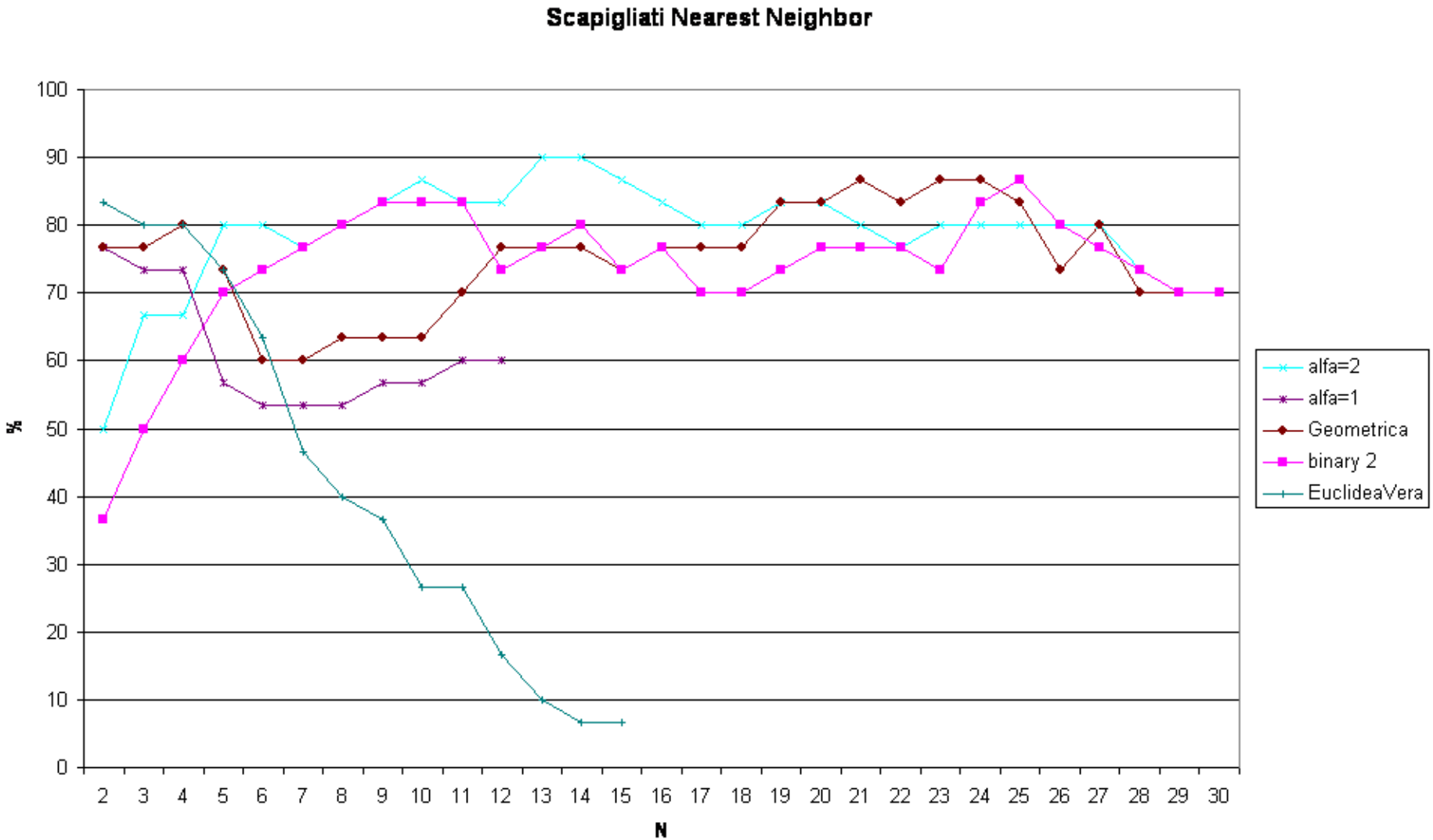


un Test Cieco

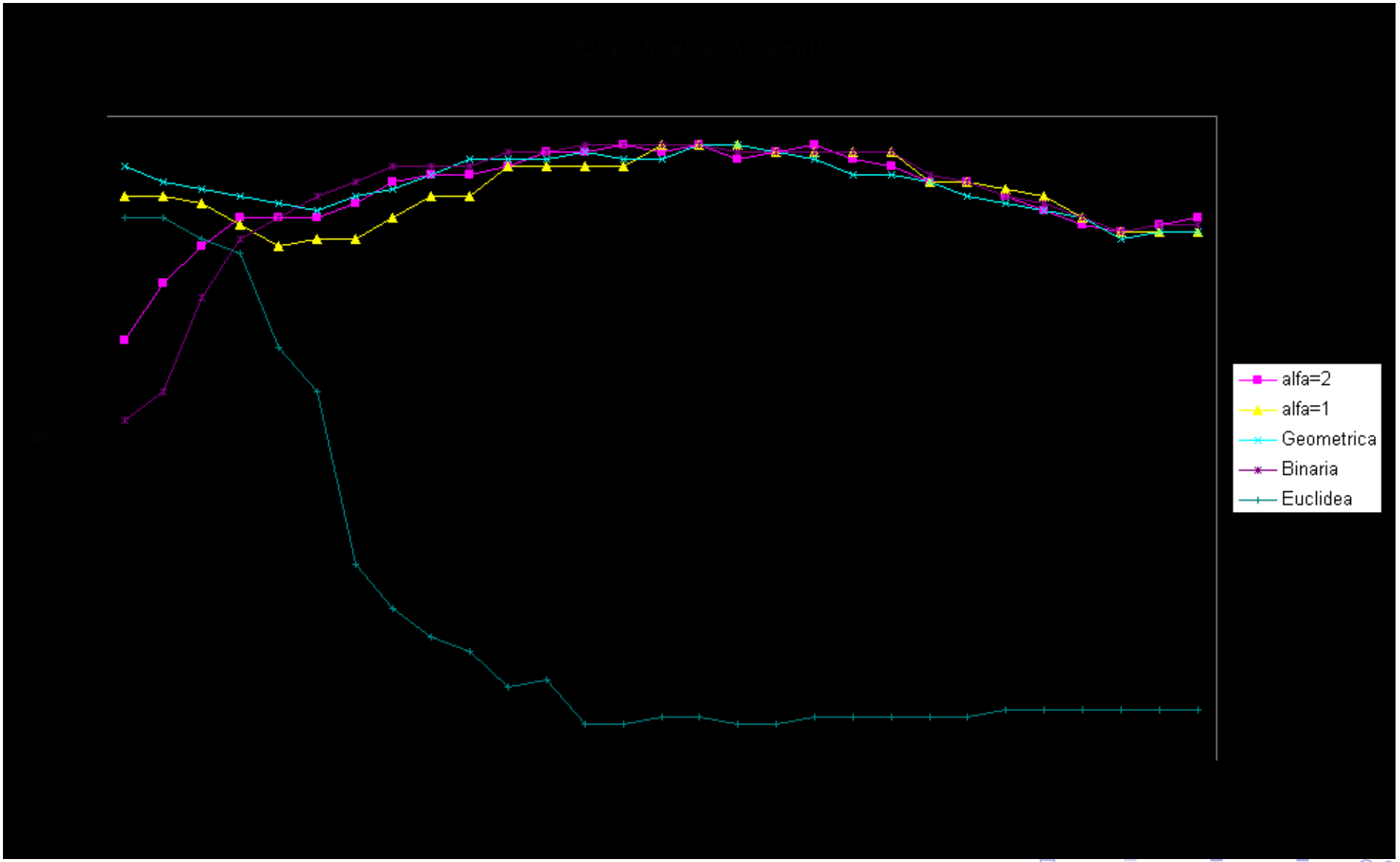
A. Gramsci (1891-1937)



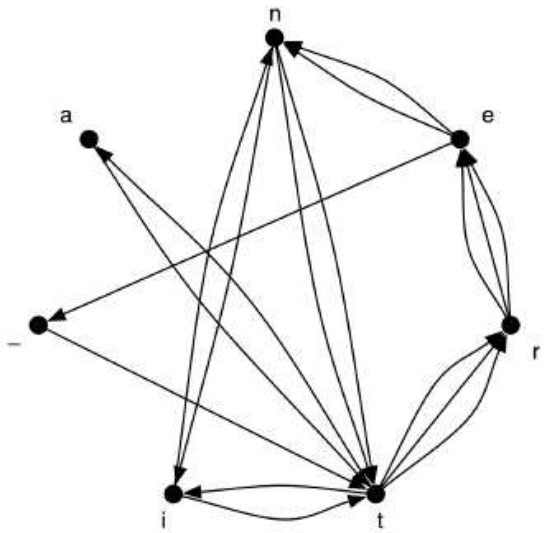
Gli Scapigliati

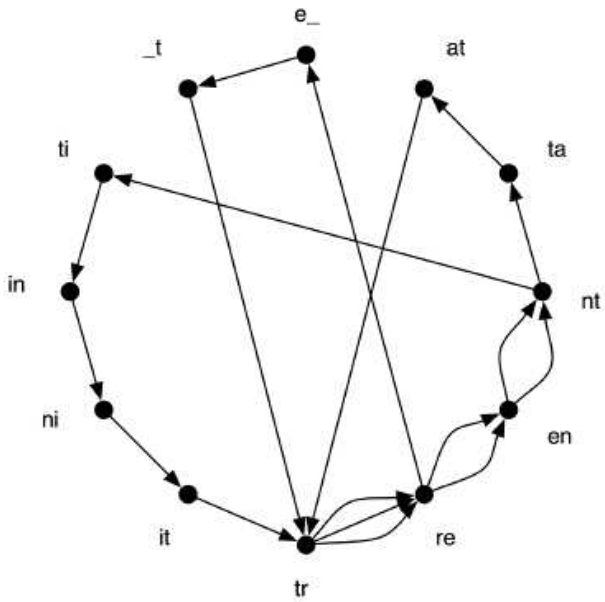


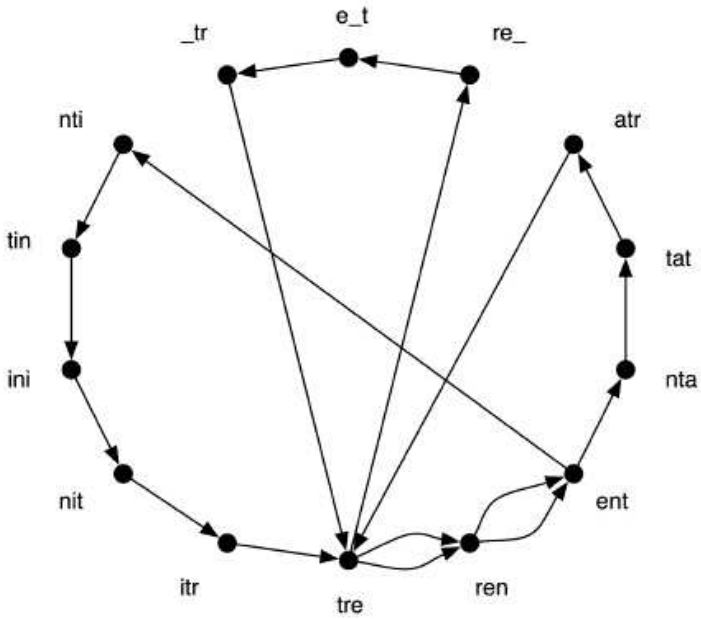
I Veristi



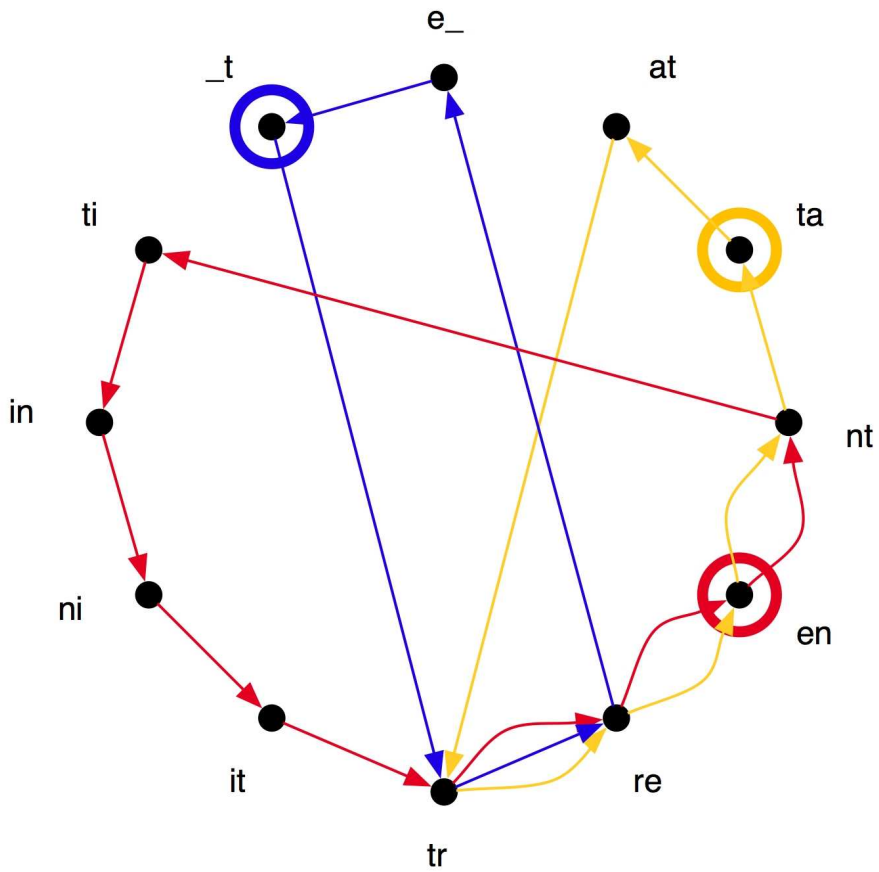
Testi Artificiali e Grafi Euleriani







Testi Artificiali e Grafi Euleriani



Testi Artificiali e Grafi Euleriani

So, so you think you can tell
heaven from hell, blue skies from pain,
can you tell a green field
from a cold steel rail?
A smile from a veil?
Do you think you can tell?

Did they get you to trade,
your heroes for ghosts?
Hot ashes for trees?
Hot air for a cool breeze?
Cold comfort for change?
And did you exchange
a walk on part in the war,
for a lead role in a cage?

How I wish, how I wish you were here,
we're just two lost soles swimming in a fish bowl,
year after year,
running over the same old ground,
but have we found the same old fears,

wish you were here.

Testi Artificiali e Grafi Euleriani

So, so you were here,
we're just two lost soles swimming in a cage?

How I wish, how I wish you think you exchange
a walk on part in the same old fears,
wish you think you to trade,
your heroes for a lead role in a fish bowl,
year after year,
running over the war,
for a cold steel rail?
A smile from a veil?
Do you tell a green field
from a cool breeze?
Cold comfort for trees?
Hot air for ghosts?
Hot ashes for change?
And did you can tell?

Did the same old ground,
but have we found they get you can tell
heaven from hell, blue skies from pain,

can you were here.

Testi Artificiali e Grafi Euleriani

Sempre caro mi fu quest'ermo colle,
e questa siepe, che da tanta parte
dell'ultimo orizzonte il guardo esclude.
Ma sedendo e mirando, interminati
spazi di là da quella, e sovrumani
silenzi, e profondissima quiete
io nel pensier mi fingo, ove per poco
il cor non si spaura. E come il vento
odo stormir tra queste piante, io quello
infinito silenzio a questa voce
vo comparando: e mi sovvien l'eterno,
e le morte stagioni, e la presente
e viva, e il suon di lei. Così tra questa
immensità s'annega il pensier mio:
e il naufragar m'è dolce in questo mare.

Testi Artificiali e Grafi Euleriani

Sempre caro mi fu quest'ermo colle,
e questa
immensità s'annega il pensier mi fingo, ove per poco
il cor non si spaura. E come il vento
odo stormir tra questa voce
vo comparando: e mi sovvien l'eterno,
e le morte stagioni, e la presente
e viva, e il suon di lei. Così tra questa siepe, che da tanta parte
dell'ultimo orizzonte il guardo esclude.
Ma sedendo e mirando, interminati
spazi di là da quella, e sovrumani
silenzio a queste piante, io quello
infinito silenzi, e profondissima quiete
io nel pensier mio:
e il naufragar m'è dolce in questo mare.