Complexity, sequence distance and heart rate variability

M. Degli Esposti

desposti@dm.unibo.it

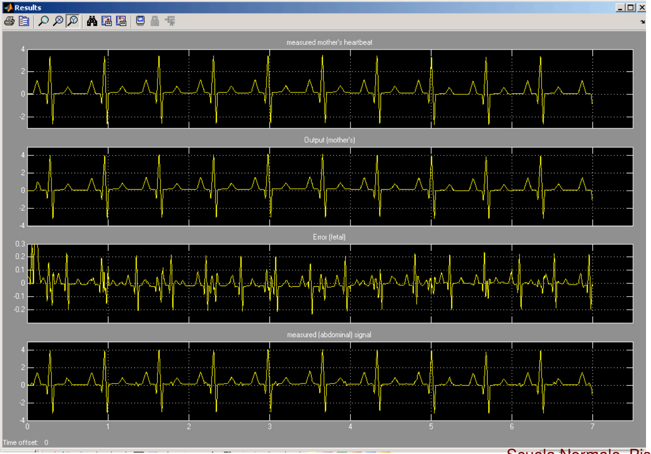
http://www.dm.unibo.it/˜desposti/

Dipartimento di Matematica
Università di Bologna

Scuola Normale, Pisa, Marzo 2009

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**

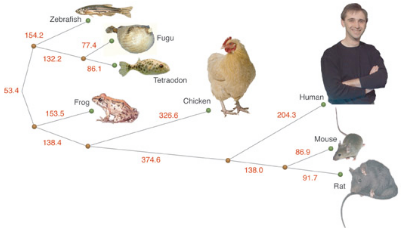
old group 13 healthy subject belonging to **gk** previously described.

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

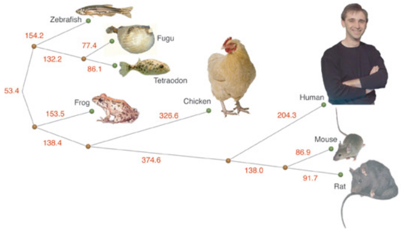
Genome Phylogeny Problem

- 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) ?



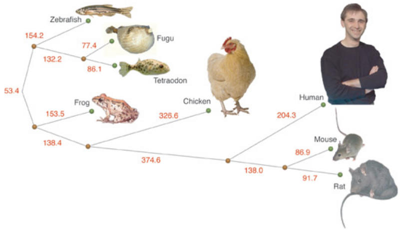
Genome Phylogeny Problem

- 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) ?



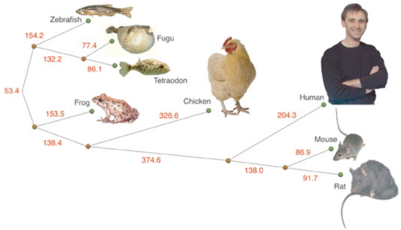
Genome Phylogeny Problem

- 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) ?

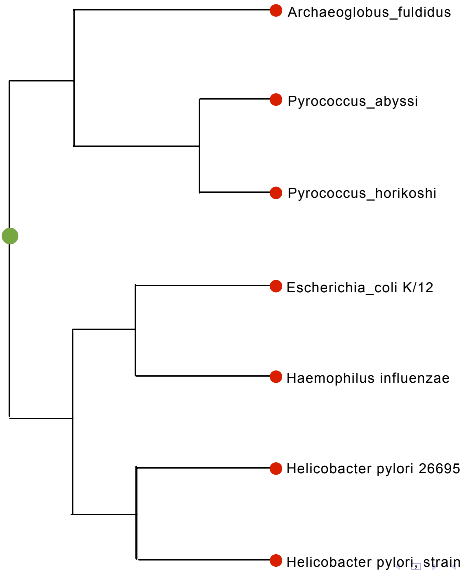


Complete Genoma

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



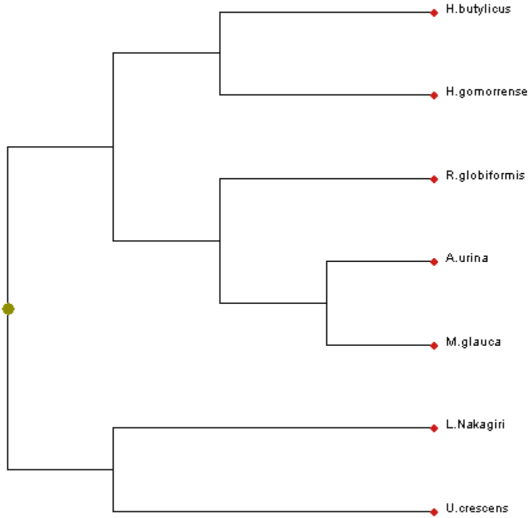
rRNA single Genes

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* ?

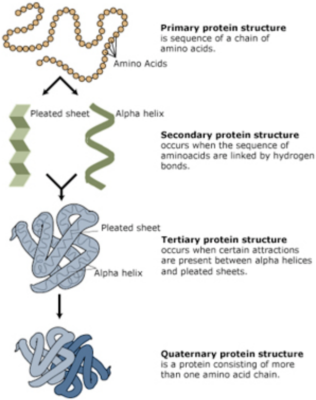


Image adapted from: National Human Genome Research Institute.

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

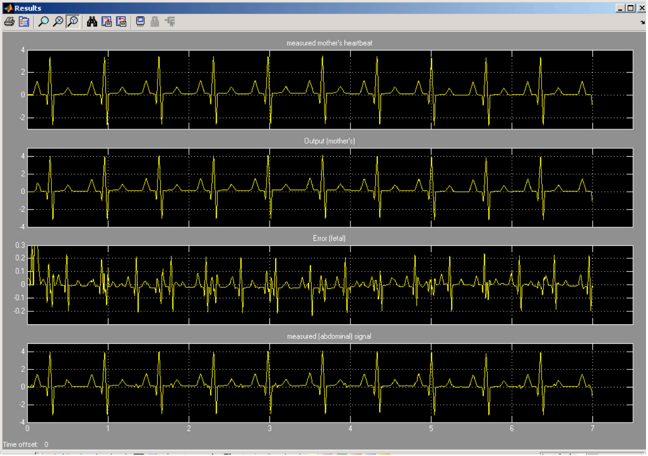
- During the period 1914-1928, Gramsci produced an enormous number of articles on different national newspaper.
- Most of these article (hundreds, if not thousands) are NOT signed
- Other possible authors : Bordiga, Serrati, Tasca, Togliatti...the aim is to recognize the articles really written by A. 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

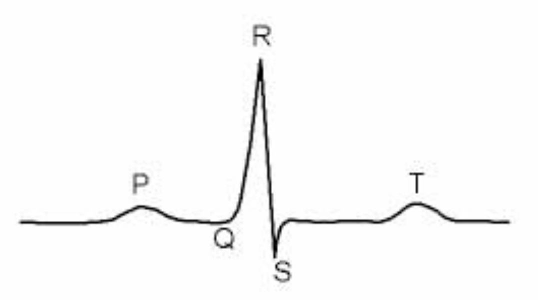
A. Gramsci (1891-1937), Journalist and founder of the Italian Comunist Party

- 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)*

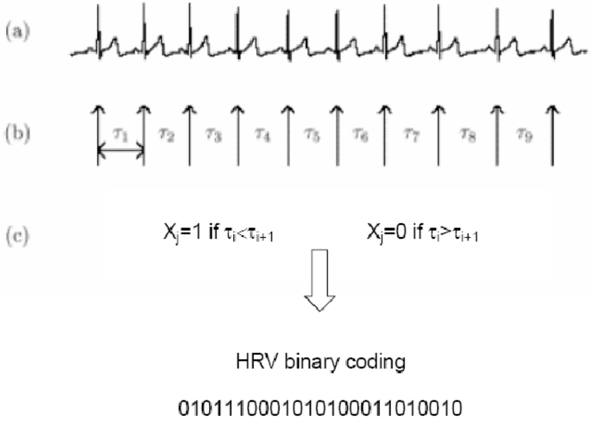
Let us go back to the heart signal...



the QRS complex



A binary HRV coding



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

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

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

similarity (pseudo) distance:

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

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

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:

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

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

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 $d(x, y) = 0 \iff x = y$
- $d(x, y) \lessapprox d(x, z) + d(z, y)$

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:

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

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

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:

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

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

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...

Processes and Entropy

- Discrete-time, stochastic ,stationary Process on $\mathcal{A}$: $X_1, X_2, \ldots, X_n \ldots\ldots.., X_j \in \mathcal{A}$
- *k-th order joint distribution*: $k = 1, 2, \ldots$

$$\mu_k(a_1, a_2, \ldots, a_k) := \mu_k(a_1^k) = \text{Prob}(X_1 = a_1, X_2 = a_2, \ldots, X_k = a_k)$$

- *conditional distribution*

$$\mu(a_k|a_1^{k-1}) := \text{Prob}(X_k = a_k \,|\, X_1^{k-1} = a_1^{k-1}) = \frac{\mu_k(a_1^k)}{\mu_{k-1}(a_1^{k-1})}$$

Processes and Entropy

- Discrete-time, stochastic ,stationary Process on $\mathcal{A}$:
  $X_1, X_2, \ldots, X_n \ldots ...., X_j \in \mathcal{A}$
- *k-th order joint distribution*: $k = 1, 2, \ldots$

$$\mu_k(a_1, a_2, \ldots, a_k) := \mu_k(a_1^k) = \mathrm{Prob}(X_1 = a_1, X_2 = a_2, \ldots, X_k = a_k)$$

- *conditional distribution*

$$\mu(a_k | a_1^{k-1}) := \mathrm{Prob}(X_k = a_k \,|\, X_1^{k-1} = a_1^{k-1}) = \frac{\mu_k(a_1^k)}{\mu_{k-1}(a_1^{k-1})}$$

Processes and Entropy

- Discrete-time, stochastic ,stationary Process on $\mathcal{A}$:
  $X_1, X_2, \dots, X_n \dots ...., X_j \in \mathcal{A}$
- *k-th order joint distribution*: $k = 1, 2, \dots$

$$\mu_k(a_1, a_2, \dots, a_k) := \mu_k(a_1^k) = \mathrm{Prob}(X_1 = a_1, X_2 = a_2, \dots, X_k = a_k)$$

- *conditional distribution*

$$\mu(a_k|a_1^{k-1}) := \mathrm{Prob}(X_k = a_k \,|\, X_1^{k-1} = a_1^{k-1}) = \frac{\mu_k(a_1^k)}{\mu_{k-1}(a_1^{k-1})}$$

Processes and Entropy

- Consistency:

$$\mu_k(a_1^k) = \sum_{a_{k+1}\in\mathcal{A}} \mu_{k+1}(a_1^{k+1}),\ a_1^k \in \mathcal{A}^k$$

- Bernoulli

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k)$$

- Markov Chain

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-1})$$

- Markov process with (fixed) memory ℓ and VLMP

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-\ell}^k)$$

Processes and Entropy

- Consistency:

$$\mu_k(a_1^k) = \sum_{a_{k+1}\in\mathcal{A}} \mu_{k+1}(a_1^{k+1}),\ \ a_1^k \in \mathcal{A}^k$$

- Bernoulli

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k)$$

- Markov Chain

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-1})$$

- Markov process with (fixed) memory ℓ and VLMP

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-\ell}^k)$$

Processes and Entropy

- Consistency:

$$\mu_k(a_1^k) = \sum_{a_{k+1} \in \mathcal{A}} \mu_{k+1}(a_1^{k+1}), \; a_1^k \in \mathcal{A}^k$$

- Bernoulli

$$\mu(a_k | a_1^{k-1}) = \mu_1(a_k)$$

- Markov Chain

$$\mu(a_k | a_1^{k-1}) = \mu_1(a_k | a_{k-1})$$

- Markov process with (fixed) memory ℓ and VLMP

$$\mu(a_k | a_1^{k-1}) = \mu_1(a_k | a_{k-\ell}^k)$$

Processes and Entropy

- Consistency:

$$\mu_k(a_1^k) = \sum_{a_{k+1}\in\mathcal{A}} \mu_{k+1}(a_1^{k+1}),\ \ a_1^k \in \mathcal{A}^k$$

- Bernoulli

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k)$$

- Markov Chain

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-1})$$

- Markov process with (fixed) memory ℓ and VLMP

$$\mu(a_k|a_1^{k-1}) = \mu_1(a_k|a_{k-\ell}^k)$$

Life is tuff....

- Knowing the μ_k's means knowing the process
- but usually we do NOT know the μ_k's....
- acutally, we do not EVEN know the "memory" (long or short correlations)
- even worst: often we just have one (short) realization $X_1, X_2, \ldots, X_n$

Life is tuff....

- Knowing the μ_k's means knowing the process
- but usually we do NOT know the μ_k's....
- acutally, we do not EVEN know the "memory" (long or short correlations)
- even worst: often we just have one (short) realization $X_1, X_2, \ldots, X_n$

Life is tuff....

- Knowing the μ_k's means knowing the process
- but usually we do NOT know the μ_k's....
- acutally, we do not EVEN know the "memory" (long or short correlations)
- even worst: often we just have one (short) realization $X_1, X_2, \ldots, X_n$

Life is tuff....

- Knowing the μ_k's means knowing the process
- but usually we do NOT know the μ_k's....
- acutally, we do not EVEN know the "memory" (long or short correlations)
- even worst: often we just have one (short) realization $X_1, X_2, \ldots, X_n$

Entropy

- n-th entropy

$$H_n := -\sum_{a_1^n \in \mathcal{A}^n} \mu(a_1^n) \log \mu(a_1^n)$$

- entropy rate

$$h_n := H_{n+1} - H_n \qquad h_{n+1} \leq h_n \leq \cdots \leq h_1$$

- Entropy

$$h = \lim_{n\to\infty} h_n =, \lim_{n\to\infty} \frac{1}{n} H_n$$

Entropy

- n-th entropy

$$H_n := -\sum_{a_1^n \in \mathcal{A}^n} \mu(a_1^n) \log \mu(a_1^n)$$

- entropy rate

$$h_n := H_{n+1} - H_n \qquad h_{n+1} \leq h_n \leq \cdots \leq h_1$$

- Entropy

$$h = \lim_{n\to\infty} h_n =, \lim_{n\to\infty} \frac{1}{n} H_n$$

Entropy

- n-th entropy

$$H_n := -\sum_{a_1^n \in \mathcal{A}^n} \mu(a_1^n) \log \mu(a_1^n)$$

- entropy rate

$$h_n := H_{n+1} - H_n \qquad h_{n+1} \leq h_n \leq \cdots \leq h_1$$

- Entropy

$$h = \lim_{n\to\infty} h_n =, \lim_{n\to\infty} \frac{1}{n} H_n$$

Entropy: two Theorems and one question

- Theorem A: a process $>\mu$ is k-th Markov *if and only if* $h = h_k$
- The Entropy Theorem: For ergodic μ and for almost all realizations

$$\lim_{n\to\infty} -\frac{1}{n} \log \mu(x_1^n) = h$$

- How to estimate h, without knowing μ ?

Entropy: two Theorems and one question

- Theorem A: a process $>\mu$ is k-th Markov *if and only if* $h = h_k$
- The Entropy Theorem: For ergodic μ and for almost all realizations

$$\lim_{n\to\infty} -\frac{1}{n} \log \mu(x_1^n) \ = \ h$$

- How to estimate h, without knowing μ ?

Entropy: two Theorems and one question

- Theorem A: a process $>\mu$ is k-th Markov *if and only if* $h = h_k$
- The Entropy Theorem: For ergodic μ and for almost all realizations

$$\lim_{n\to\infty} -\frac{1}{n}\log\mu(x_1^n) \;=\; h$$

- How to estimate h, without knowing μ ?

Interpretation of entropy

- Define:

$$S_n(\epsilon) = \left\{ x_1^n \in \mathcal{A}^n \,:\, 2^{-n(h+\epsilon)} \leq \mu(x_1^n) \leq 2^{-n(h-\epsilon)} \right\}$$

- typical sequences For each $\epsilon > 0$,

$$x_1^n \in S_n(\epsilon)$$

  eventually almost surely

- entropy typical cardinality bound

$$|S_n(\epsilon)| \leq 2^{n(h+\epsilon)}$$

Interpretation of entropy

- Define:

$$S_n(\epsilon) = \left\{ x_1^n \in \mathcal{A}^n \, : \, 2^{-n(h+\epsilon)} \leq \mu(x_1^n) \leq 2^{-n(h-\epsilon)} \right\}$$

- typical sequences For each $\epsilon > 0$,

$$x_1^n \in S_n(\epsilon)$$

eventually almost surely

- entropy typical cardinality bound

$$|S_n(\epsilon)| \leq 2^{n(h+\epsilon)}$$

Interpretation of entropy

- Define:

$$S_n(\epsilon) = \left\{ x_1^n \in \mathcal{A}^n \;:\; 2^{-n(h+\epsilon)} \leq \mu(x_1^n) \leq 2^{-n(h-\epsilon)} \right\}$$

- typical sequences For each $\epsilon > 0$,

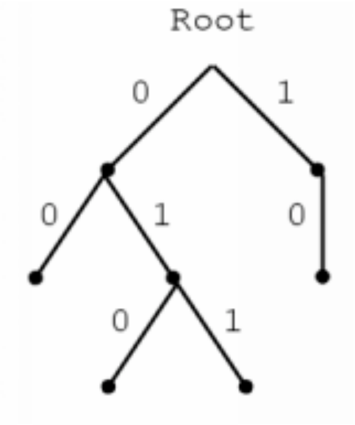
$$x_1^n \in S_n(\epsilon)$$

  eventually almost surely

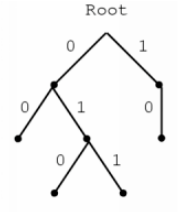
- entropy typical cardinality bound

$$|S_n(\epsilon)| \leq 2^{n(h+\epsilon)}$$

Entropy and Prefixcode



Entropy and Prefixcode



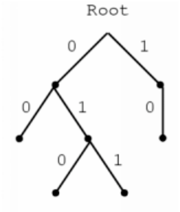
If:

$$E_\mu(L(\cdot|C)) = \sum_{a\in\mathcal{A}} L(a|C)\mu(a)$$

then for any prefix code C:

$$E_\mu(L(\cdot|C)) \geq h(\mu)$$

Entropy and Prefixcode



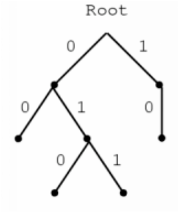
If:

$$E_\mu(L(\cdot|C)) = \sum_{a \in \mathcal{A}} L(a|C)\mu(a)$$

then for any prefix code C:

$$E_\mu(L(\cdot|C)) \geq h(\mu)$$

Entropy and Prefixcode



If:

$$E_\mu(L(\cdot|C)) = \sum_{a \in \mathcal{A}} L(a|C)\mu(a)$$

then for any prefix code C:

$$E_\mu(L(\cdot|C)) \geq h(\mu)$$

The Lempel-Ziv Algorithms: *zippatori*

LZ- parsing of

$$x_1 x_2 x_3 \cdots x_n \ldots$$

where:

the next word is the shortest word

Example:

$$1100101000100010 0\ldots$$

parses into:

$$1, 10, 0, 101, 00, 01, 000, 100, \ldots$$

We denote by $C(x_1^n)$ the cardinality of the parsing.

The Lempel-Ziv Algorithms: *zippatori*

LZ- parsing of

$$x_1 x_2 x_3 \cdots x_n \ldots$$

where:

the next word is the shortest word

Example:

$$11001010001000100...$$

parses into:

$$1, 10, 0, 101, 00, 01, 000, 100, ...$$

We denote by $C(x_1^n)$ the cardinality of the parsing.

The Lempel-Ziv Algorithms: *zippatori*

LZ- parsing of

$$x_1 x_2 x_3 \cdots x_n \ldots$$

where:

the next word is the shortest word

Example:

$$11001010001000100\ldots$$

parses into:

$$1, 10, 0, 101, 00, 01, 000, 100, \ldots$$

We denote by $C(x_1^n)$ the cardinality of the parsing.

The Lempel-Ziv Algorithms: *zippatori*

LZ- parsing of

$$x_1 x_2 x_3 \cdots x_n \ldots$$

where:

the next word is the shortest word

Example:

$$110010100010000100\ldots$$

parses into:

$$1, 10, 0, 101, 00, 01, 000, 100, \ldots$$

We denote by $C(x_1^n)$ the cardinality of the parsing.

The Lempel-Ziv Algorithms: *zippatori*

LZ- parsing of

$$x_1 x_2 x_3 \cdots x_n \ldots$$

where:

the next word is the shortest word

Example:

$$1100101000100 0100 \ldots$$

parses into:

$$1, 10, 0, 101, 00, 01, 000, 100, \ldots$$

We denote by $C(x_1^n)$ the cardinality of the parsing.

The LZ convergence Theorem

If μ is an ergodci process with entropy h, then almost surely

$$\frac{1}{n} C(x_1^n) \log n \to h$$

proof: Ornstein-Weiss observations

- For any partition into distinct word, "most" of the words are not much shorter of $(\log n)/h$
- For any partition into words that have been seen in the past, "most" of the words are not much longer than $(\log n)/h$

The LZ convergence Theorem

If μ is an ergodci process with entropy h, then almost surely

$$\frac{1}{n} C(x_1^n) \log n \to h$$

proof: Ornstein-Weiss observations

- For any partition into distinct word, "most" of the words are not much shorter of $(\log n)/h$
- For any partition into words that have been seen in the past, "most" of the words are not much longer than $(\log n)/h$

The LZ convergence Theorem

If μ is an ergodci process with entropy h, then almost surely

$$\frac{1}{n}C(x_1^n)\log n \to h$$

proof: Ornstein-Weiss observations

- For any partition into distinct word, "most" of the words are not much shorter of $(\log n)/h$
- For any partition into words that have been seen in the past, "most" of the words are not much longer than $(\log n)/h$

The LZ convergence Theorem

If μ is an ergodci process with entropy h, then almost surely

$$\frac{1}{n} C(x_1^n) \log n \to h$$

proof: Ornstein-Weiss observations

- For any partition into distinct word, ”most” of the words are not much shorter of $(\log n)/h$
- For any partition into words that have been seen in the past, ”most” of the words are not much longer than $(\log n)/h$

Entropy and Recurrence Times

$$R_n(x) \;=\; \min\left\{m \geq 1 \;:\; x_{m+1}^{m+n} = x_1^n\right\}$$

Theorem: for any ergodic process μ with entropy h, almost surely

$$\lim_{n\to\infty} \frac{1}{n} \log R_n(x) \;=\; h$$

Entropy and Recurrence Times

$$R_n(x) \;=\; \min\left\{m \geq 1 \;:\; x_{m+1}^{m+n} = x_1^n\right\}$$

Theorem: for any ergodic process μ with entropy h, almost surely

$$\lim_{n\to\infty} \frac{1}{n} \log R_n(x) \;=\; h$$

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

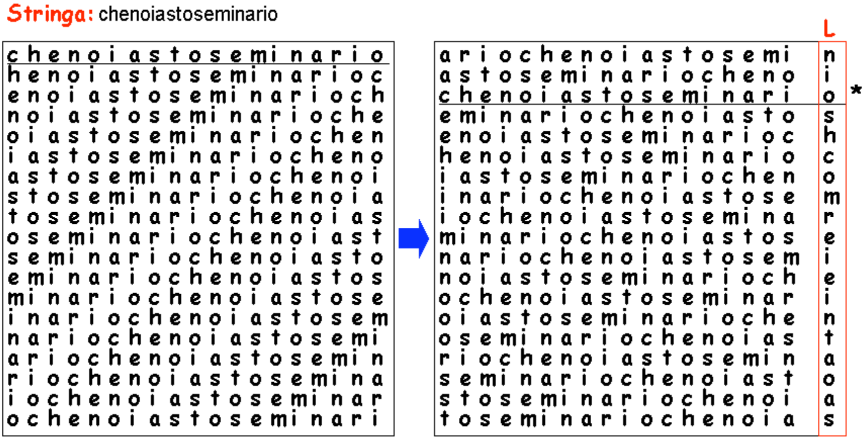
- The BWT is a permutation (easy to invert) and in particular it trasforms a finite memory Markovian sequence into a ”piecewise Bernoulli (i.i.d)” sequence
- It has been introduce in 1993 but up to now it has been used in computer science for compression: BWT(x) is ”more suitable” for compression through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

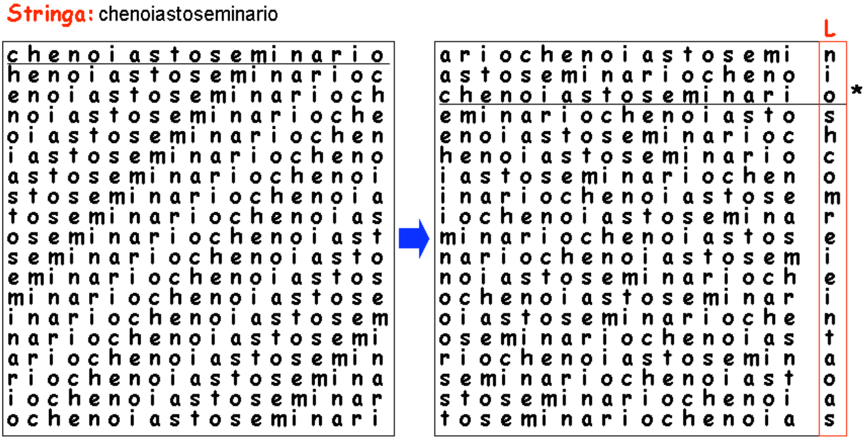
BWT

BWT(”chenoiastoseminario”):



BWT

BWT("chenoiastoseminario"):



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*

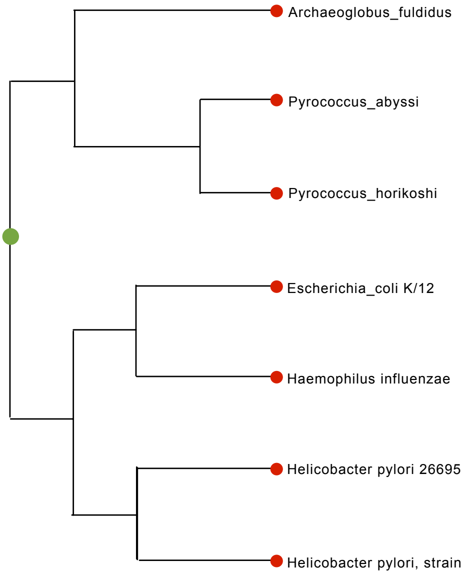
- X. Chen, S. Kwong, M.Li, *A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison* , (1999)
- d_{LZ} is NOT an *ad hoc* method
- NO *alignment* between sequences is required
- it can work for both complete genoma and single gene

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



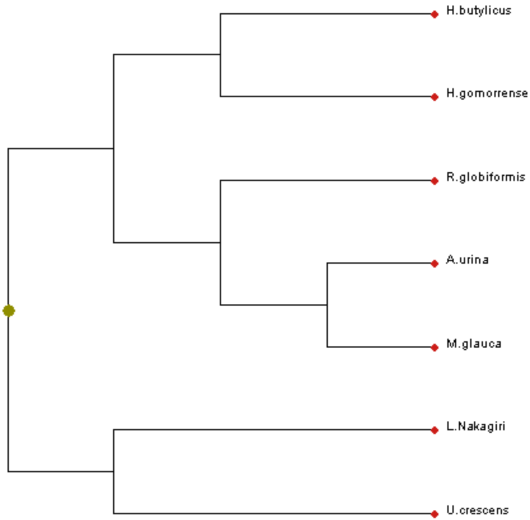
rRNA single Genes

Archaebacteria *H. butylicus* and *Halobaculum gomorrense*

Eubacteria *Aerococcus urina*, *M. glauca* strain B1448-1 and *Rhodopila globiformis*

Eukaryotes *Urosporidium crescens*, *Labyrinthula sp. Nakagiri*

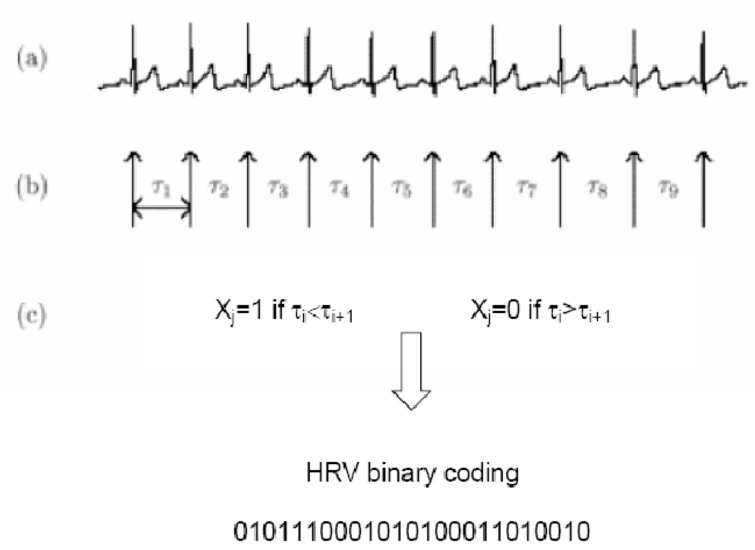
rRNA single Genes



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...

The basic ECG complex (fig. 1) represents the repetitive cycle of electrical activity in the heart, starting with the spread of stimulation through the atria (P wave) and ending with the return of stimulated ventricular muscle to its resting state (ST–T–U sequence).

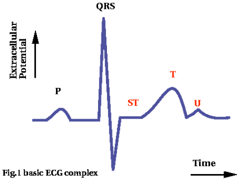
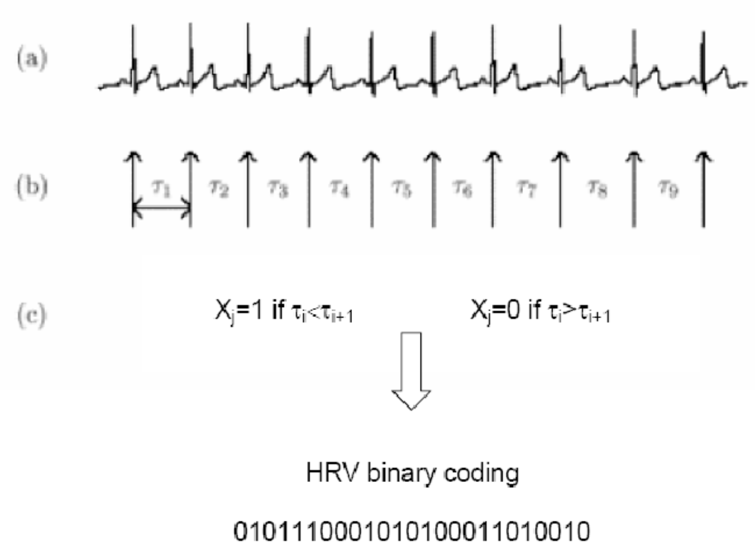


Fig.1 basic ECG complex

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

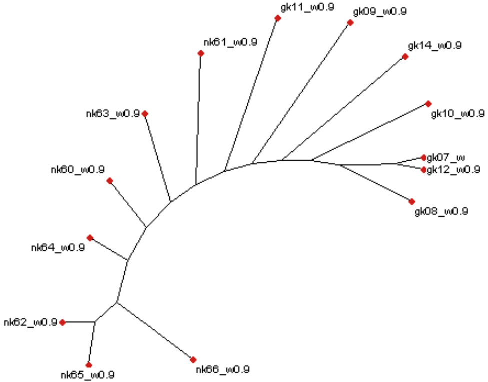
gk v.s. nk, 2

	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

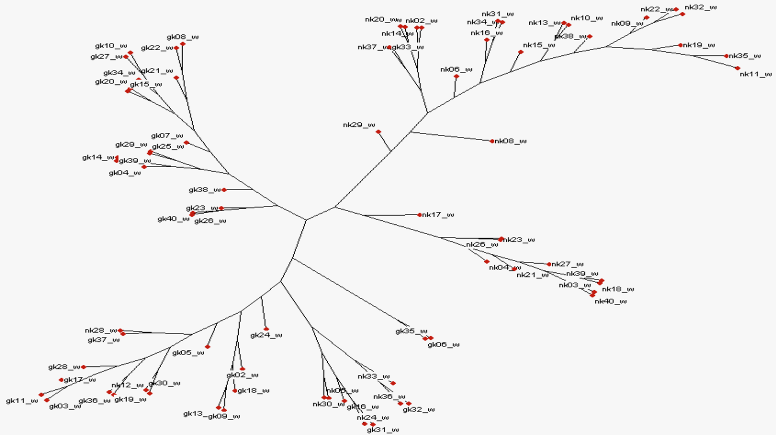
gk v.s. nk, 2

	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



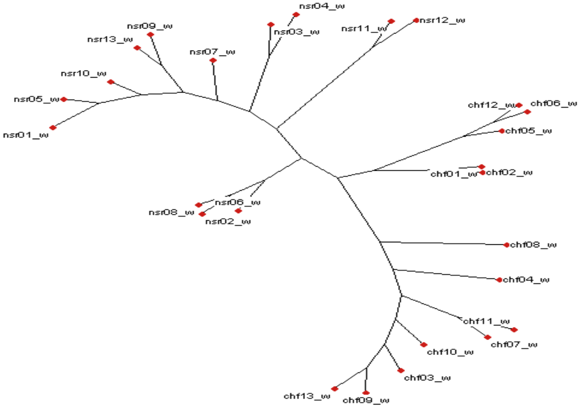
chf v.s. nsr

	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



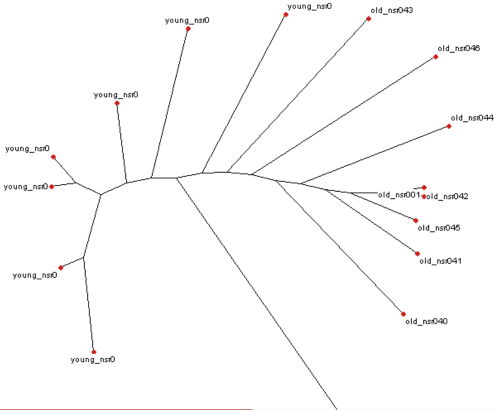
young v.s. old

	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	0,9573488
youngs	0,952945375	0,96150325	0,95560775	0,949488	0,95707025	0,957765875	0,953048625	0,959256875	0,948845857	0,9497952

young v.s. old


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

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	