Spectral Methods for the Analysis of DNA Promoters

Roberto Livi

CSDC - Dipartimento di Fisica

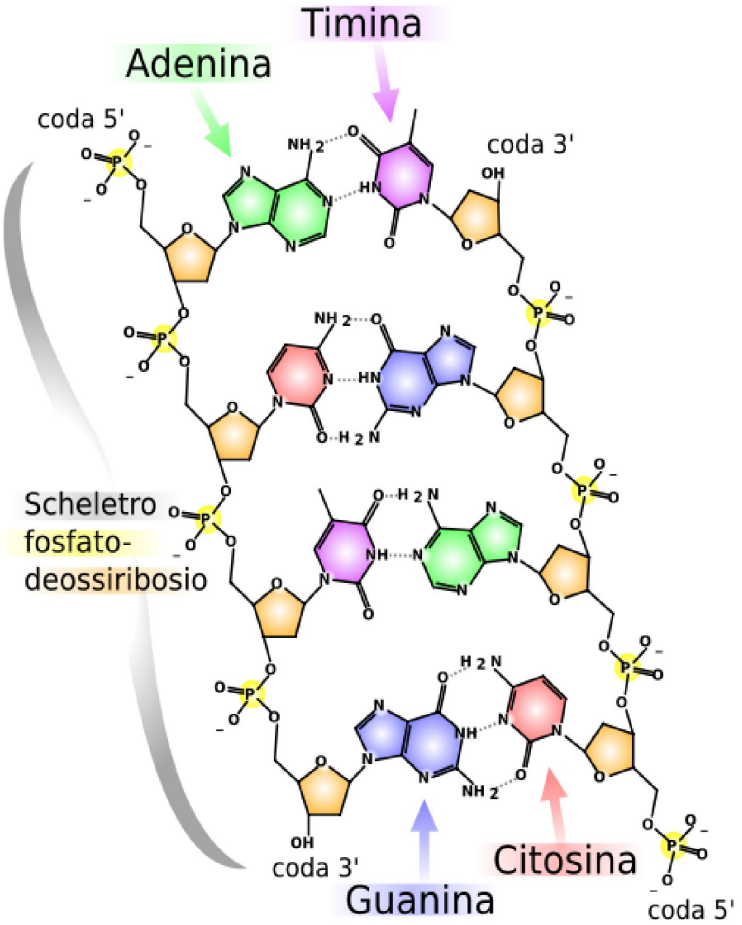
Universita' di Firenze, Italy

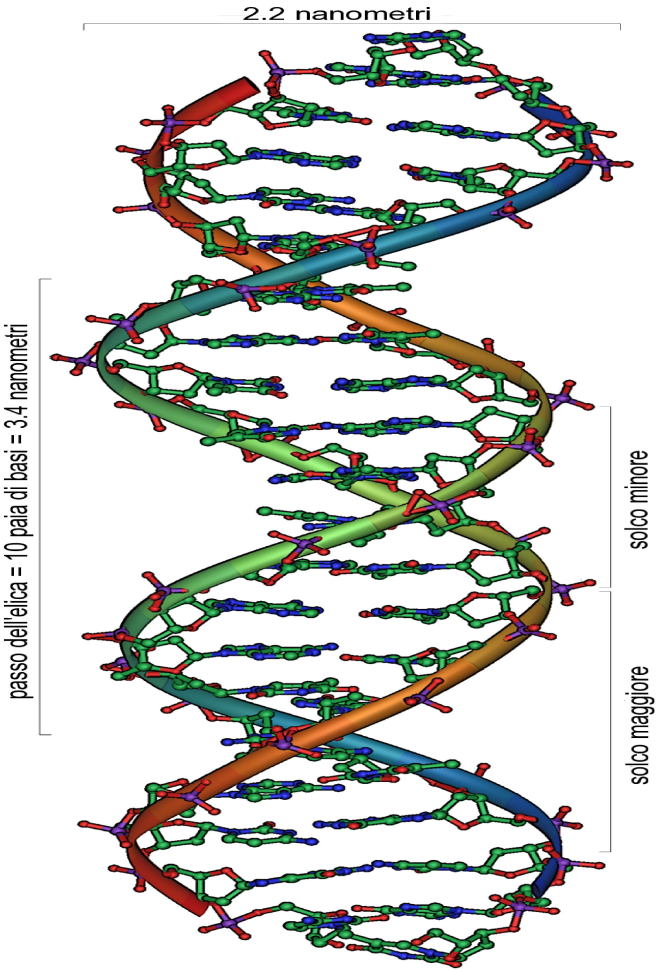
In collaboration with L. Pettinato, E. Calistri, F. Di Patti and S.Luccioli

"Strolling on Chaos, Turbulence and Statistical Mechanics" Rome Sept. 22-24/2014
In honor of Angelo Vulpiani 60th Birthday

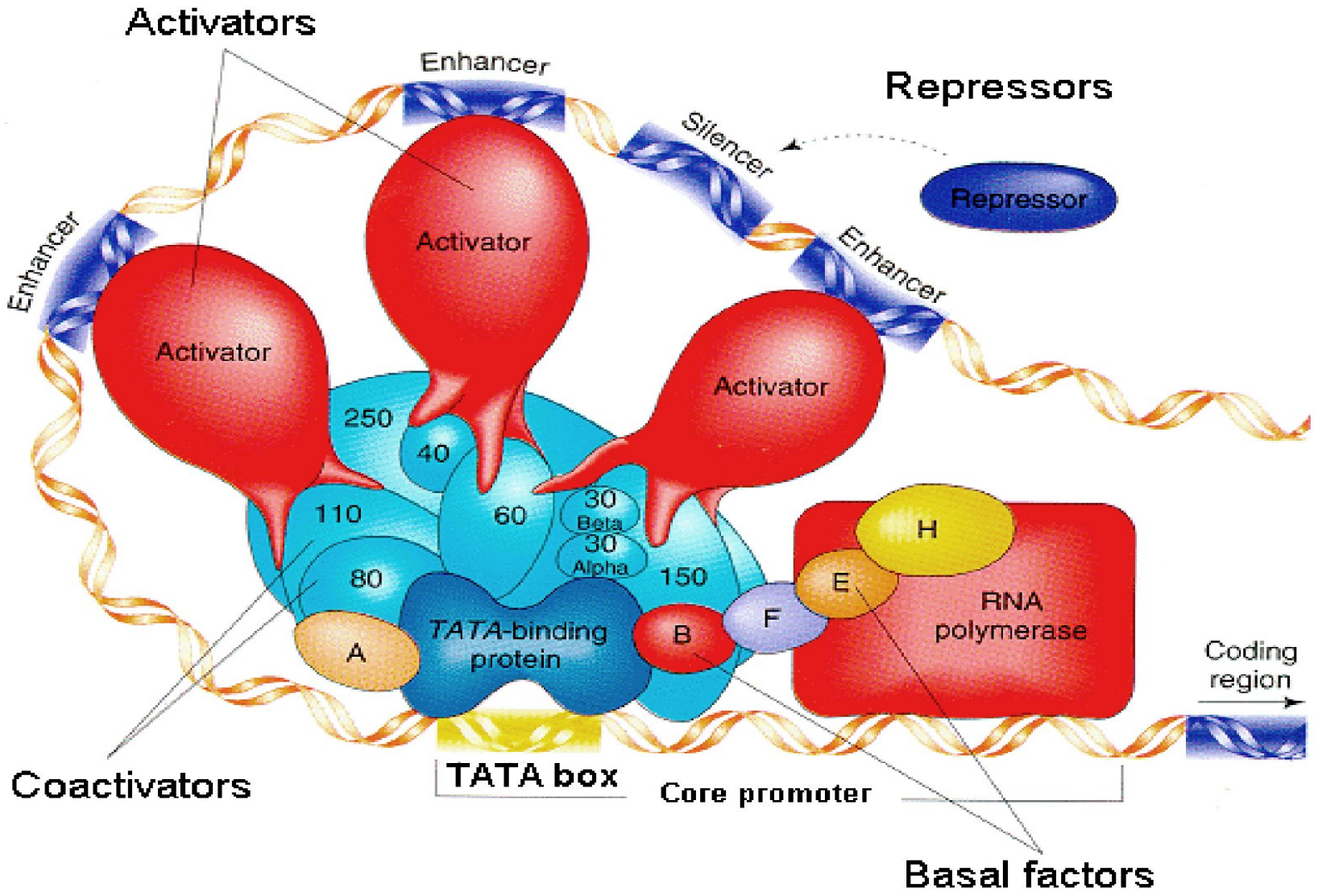
DNA contains the information necessary for the development of a living organism and allows for the transmission of this information to future generations

This is determined by its peculiar structure





Promoters play a crucial role in determining the expression and the control of genes



DNA double strand can be viewed as a sequence of symbols written in a quaternary alphabet A,T,C,G.

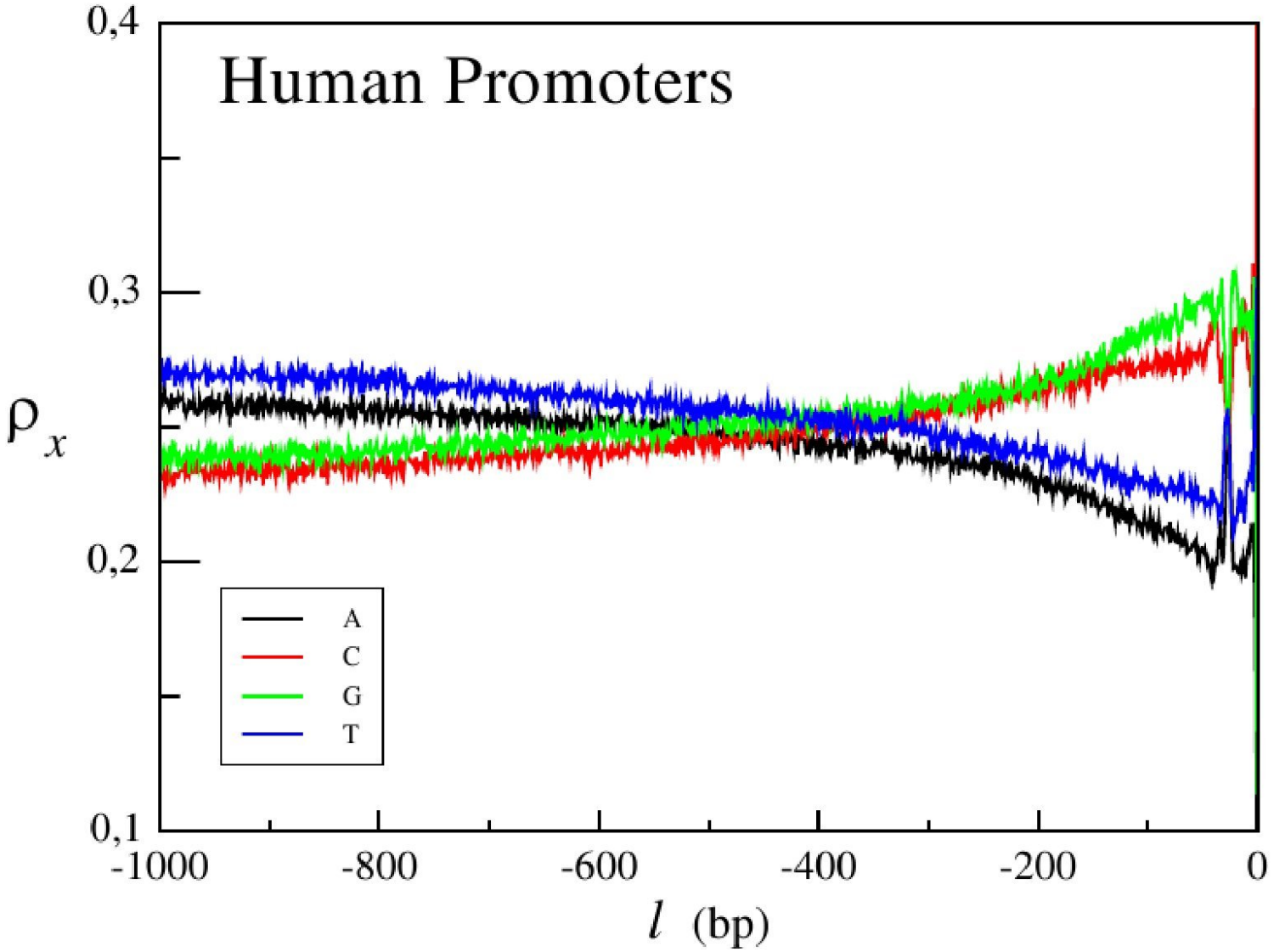
Promoters are the strings of 1000 nucleotides preceeding the transcription start site of genes.

Is it possible to recover some information encoded in promoters?

Entropic analysis based on Shannon and Lempel-Ziv algorithms doesn't help that much (although more refined methods could be more effective).

So, let's turn to a more basic tool : base compostion analysis

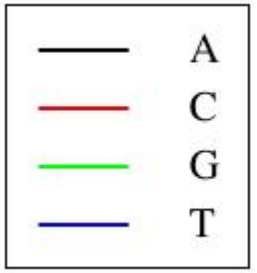
E. Calistri, R.L. and M. Buiatti, *Evolutionary trends of GC/AT distribution patterns in promoters,* Molecular Philogenetics and Evolution, 60 (2011), 228-235 and *Variation and constraints in species-specific promoter sequences,* JTB 2014



Differentiation between TATA and TATA-less promoters extending over 1000 basis

TATA-box is made of 8 basis (!)

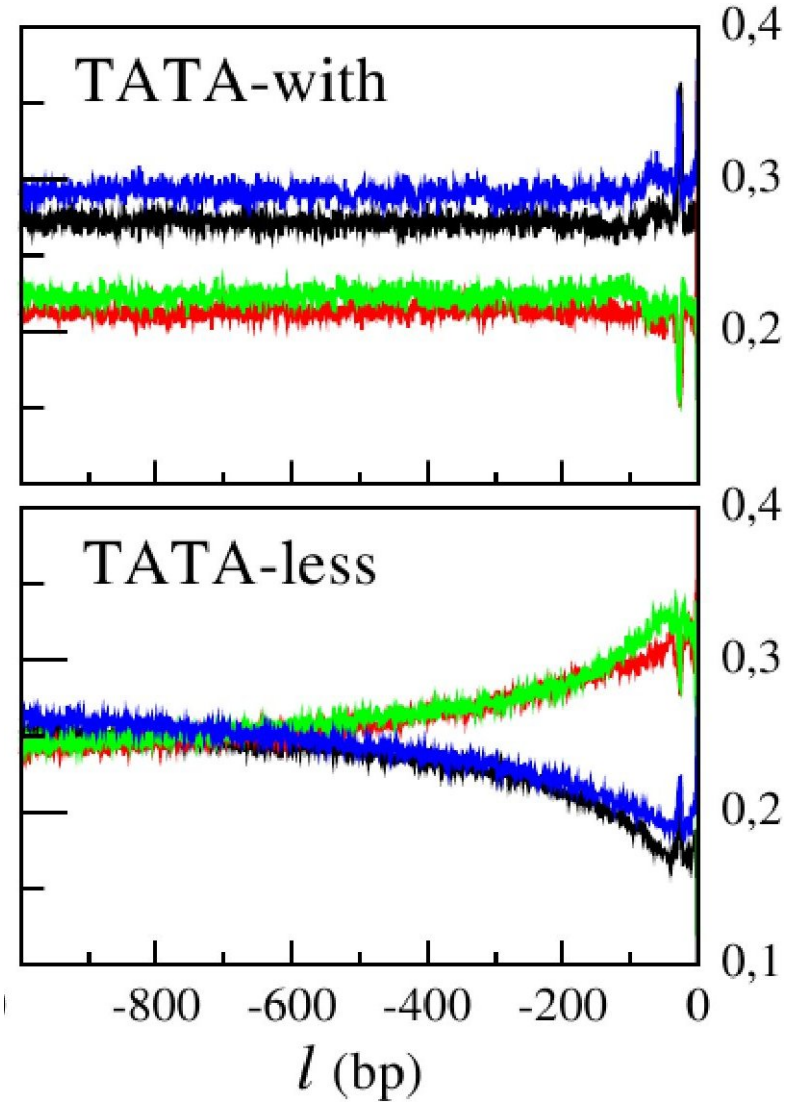
HWHWWWWR



TATA (tissue specific genes)
8100 : 1350 (S) + 7750 (A)

TATA-less (housekeeping genes)
23000 : 6000 (S) + 17000 (A)

Homo Sapiens



These results suggest to investigate more precise questions:

1) Can promoters be grouped into clusters depending on their structure according to a general *a priori* criterion ?

2) Can one point out in each of these clusters typical nucleotide subsequences that can establish a relation between structure and function ?

Spectral methods allow to answer both questions

CLUSTERING

Similarity between promoters structure can be computed by standard alignement algorithms, like Needeleman-Wunsch (global alignement) and Smith-Waterman (local alignement)

Sequenza 1: tctattgcagatttgtgtgaccaagc

Sequenza 2: agcgtcgcacgtttgaatttggcacc

Allineamento:

```
----tctattgcagatttg----tgtgaccaagc
    ||    |||..||||    ||..|||
agcgtc----gcacgtttgaatttggcacc----
```

The nontrivial aspect of this procedure is the optimization of the score to be attributed to aligned sequences and gaps

One obtains a Similarity Matrix **S** : it is symetric and introduces a metric in the promoter sample.

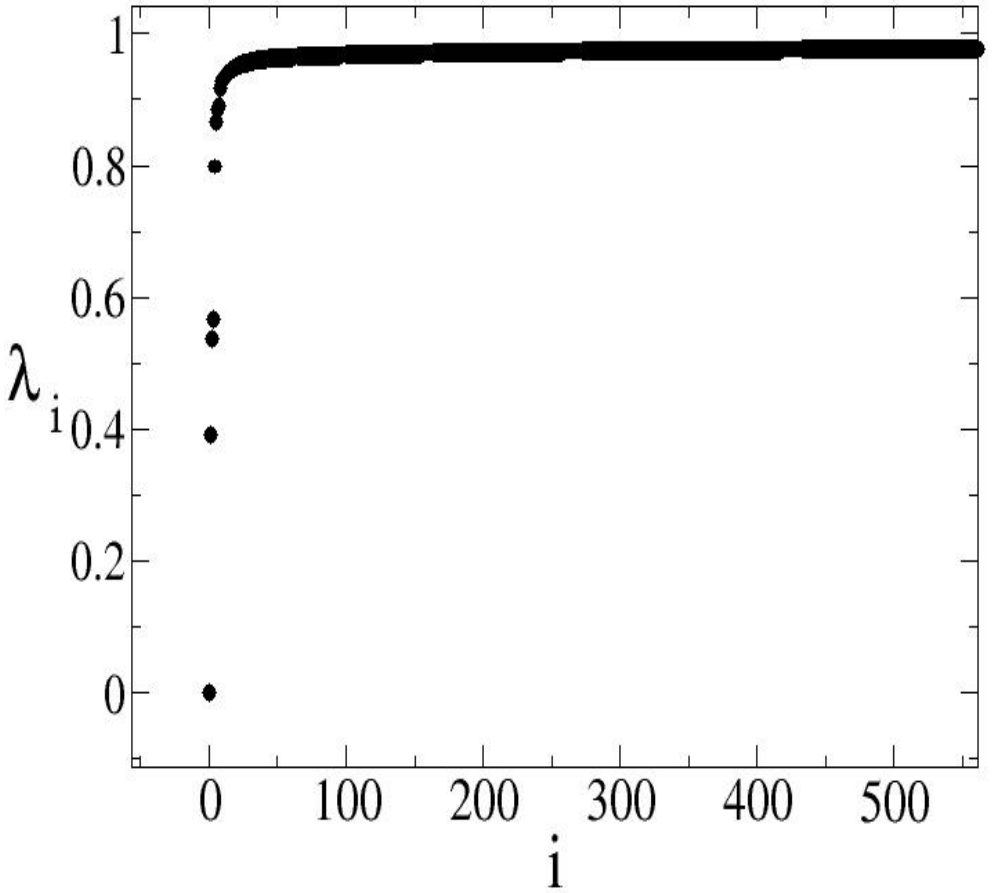
Then one can construct the associated Laplacian Matrix **L** , that yields the

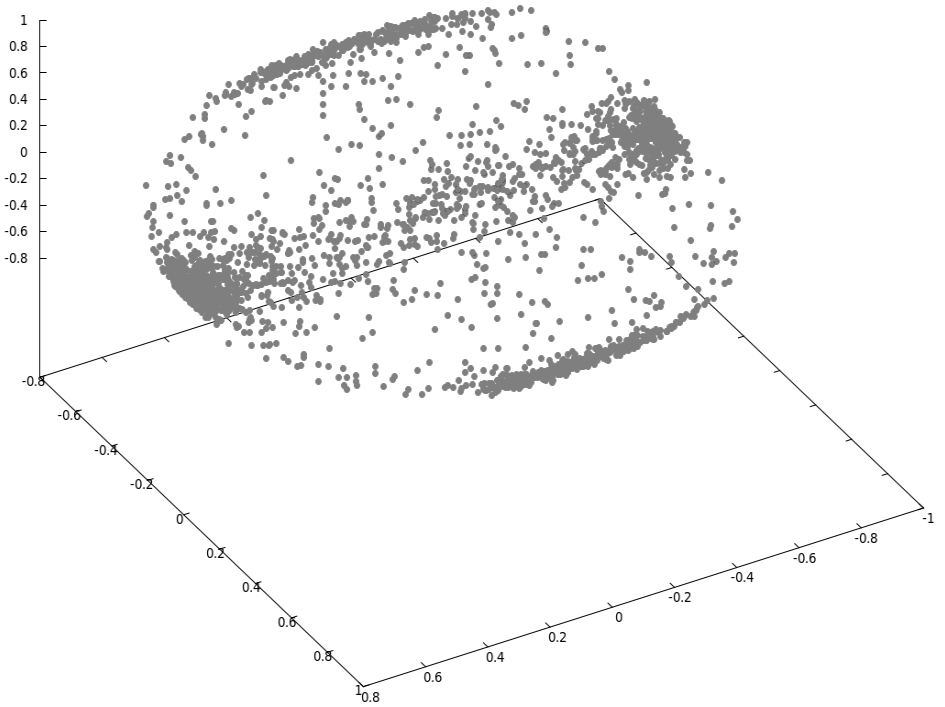
SPECTRAL CLUSTERING METHOD

Homo Sapiens

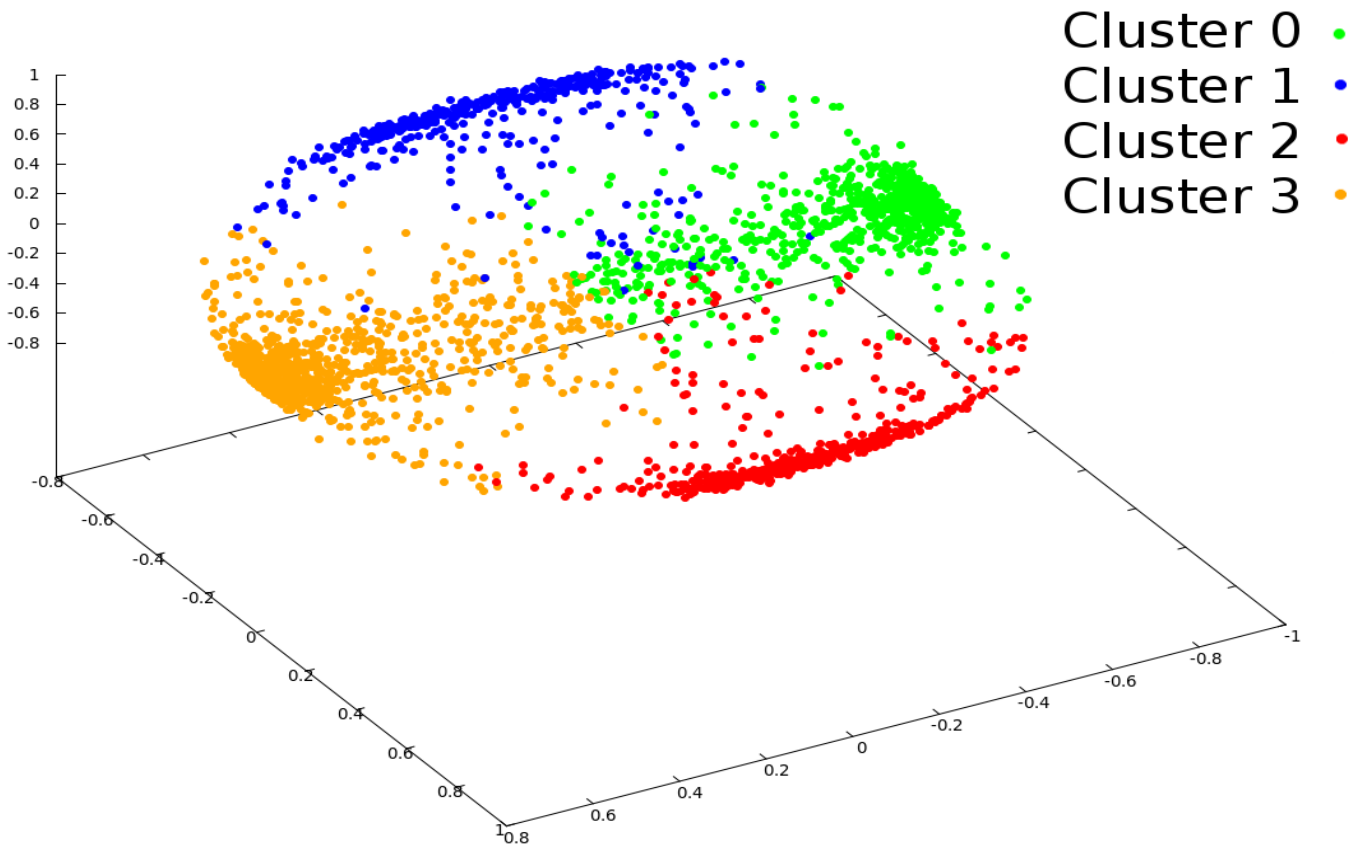
Eigenvalues of **L**

Eigenvectors of **L**

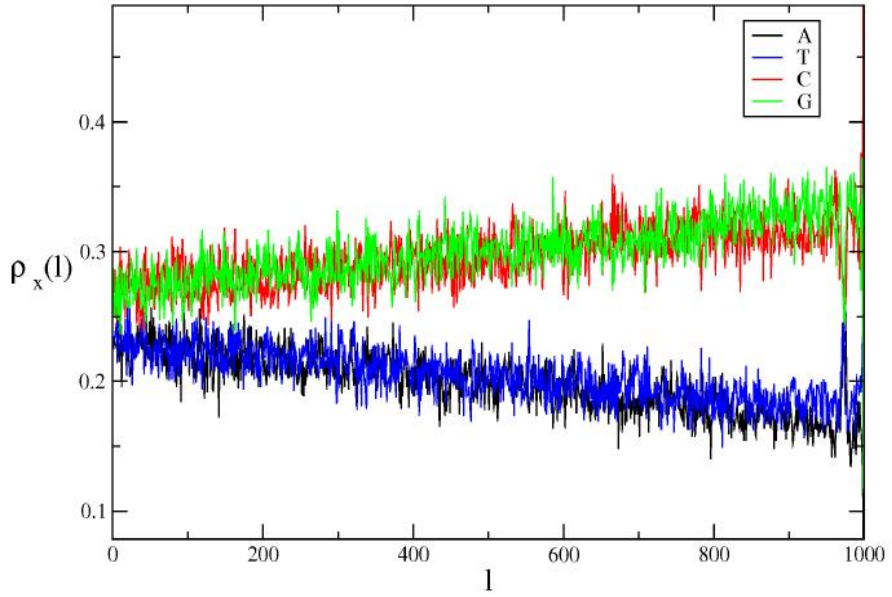


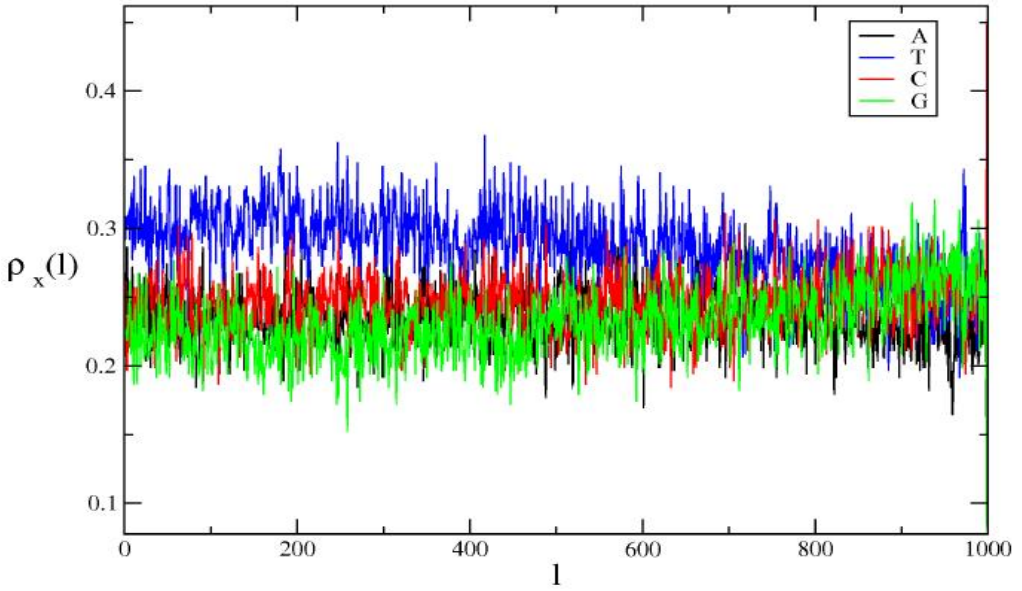


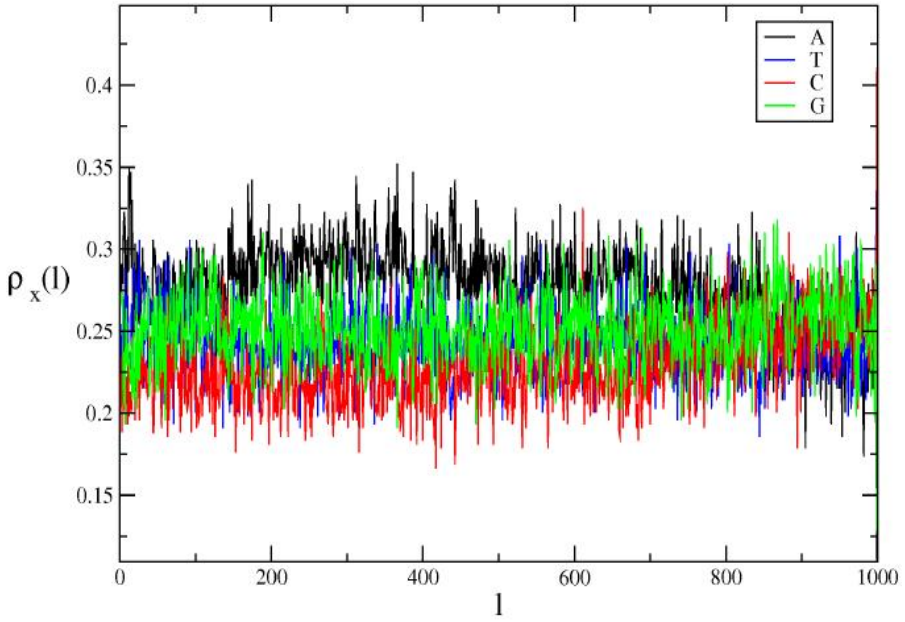
A K-means algorithm is finally employed for grouping the promoters into clusters

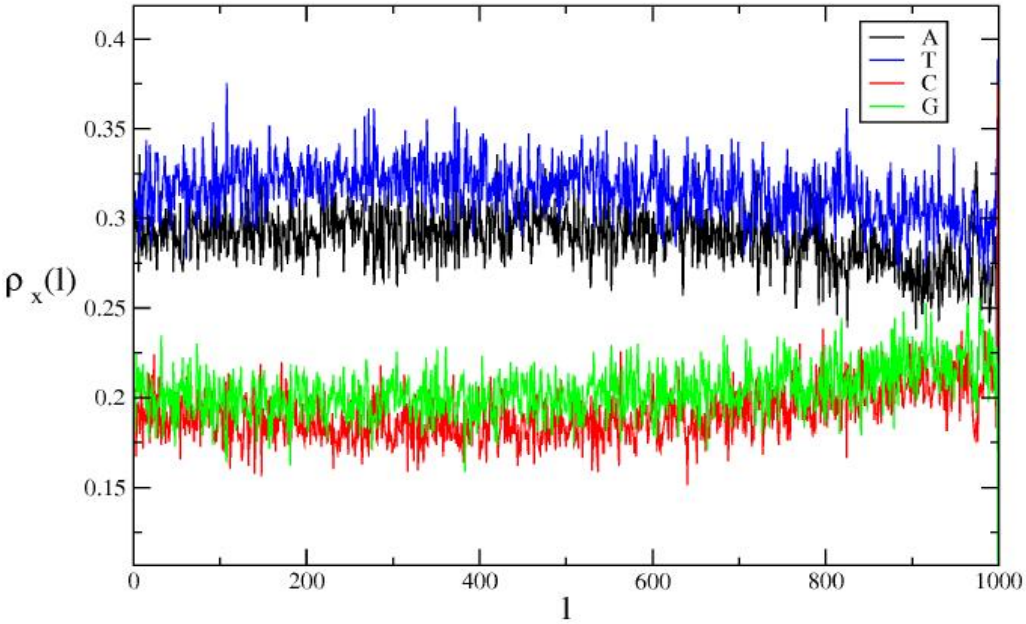


Base Composition of the four clusters of Homo Sapiens



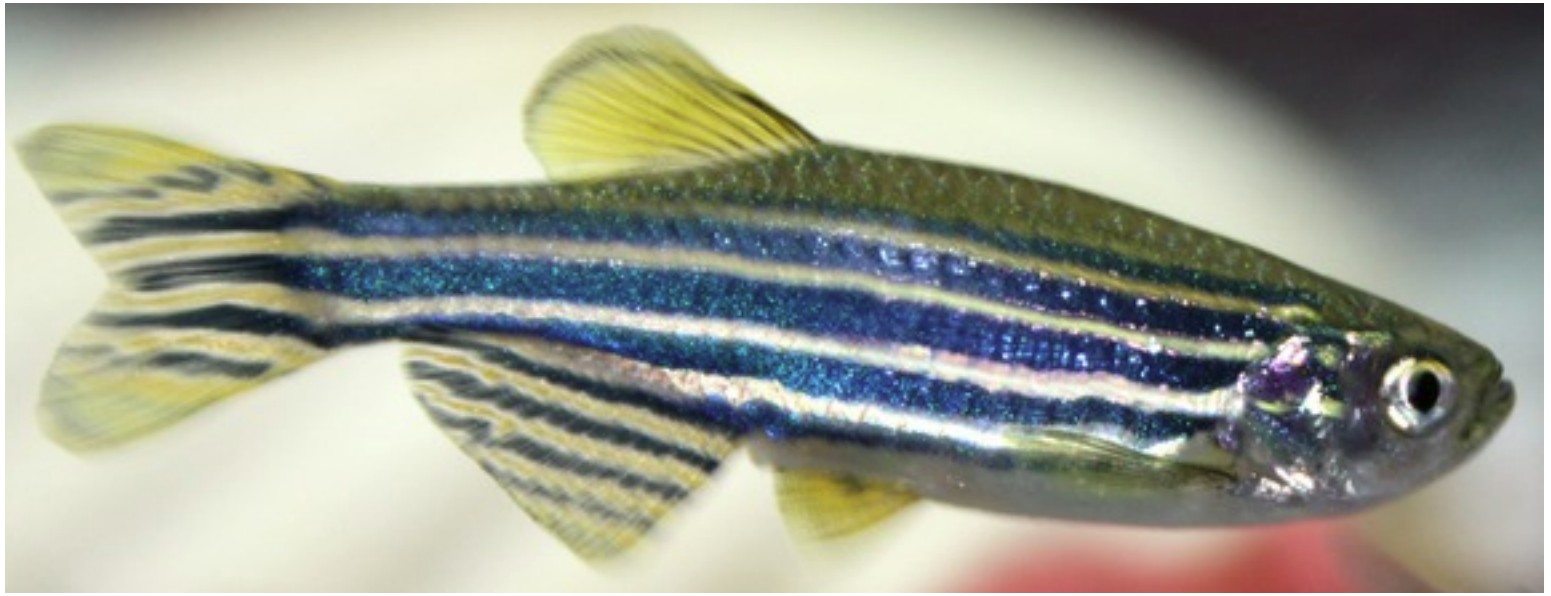




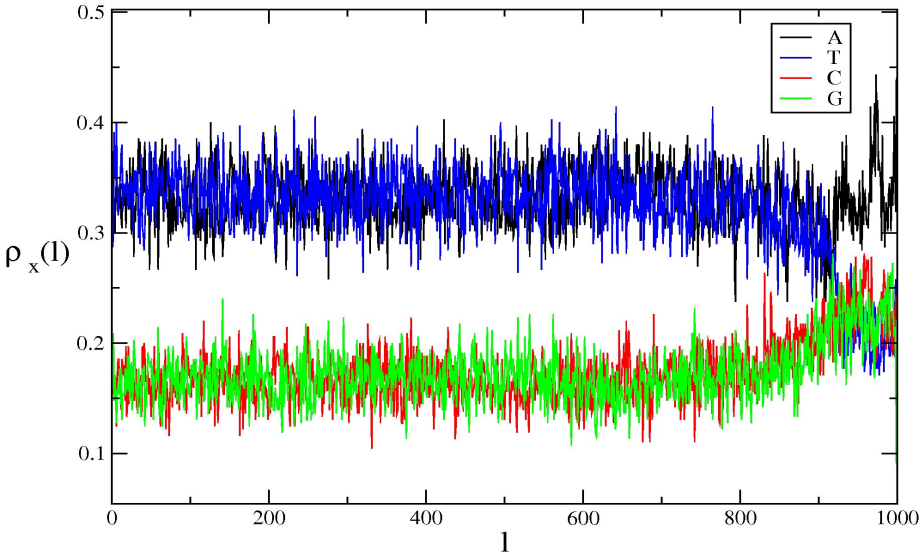


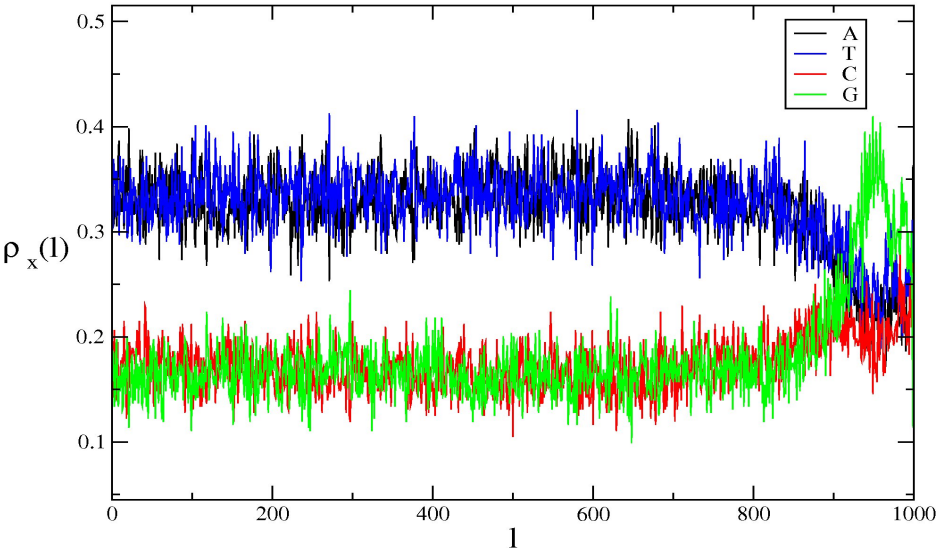
Comparison with other species.

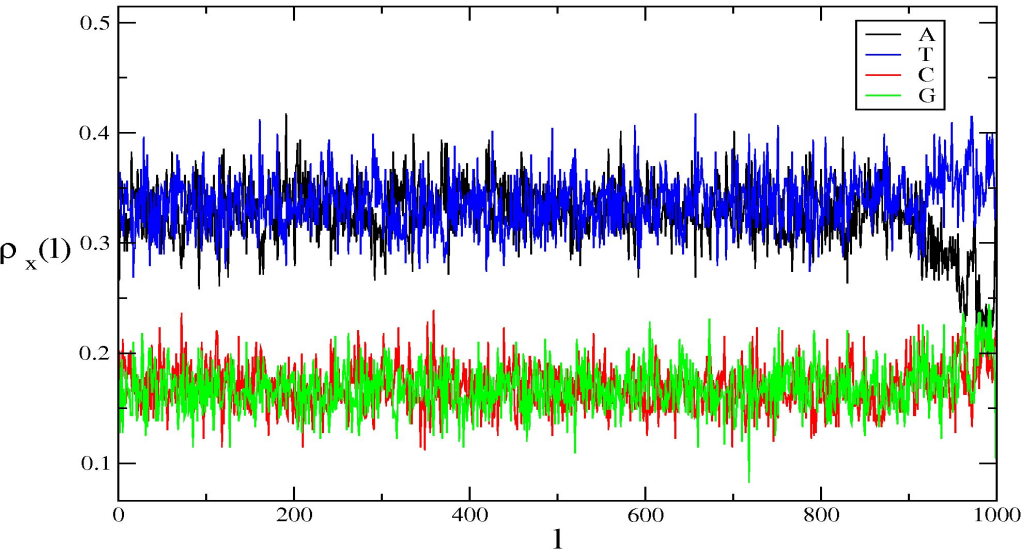
Danio Rerio

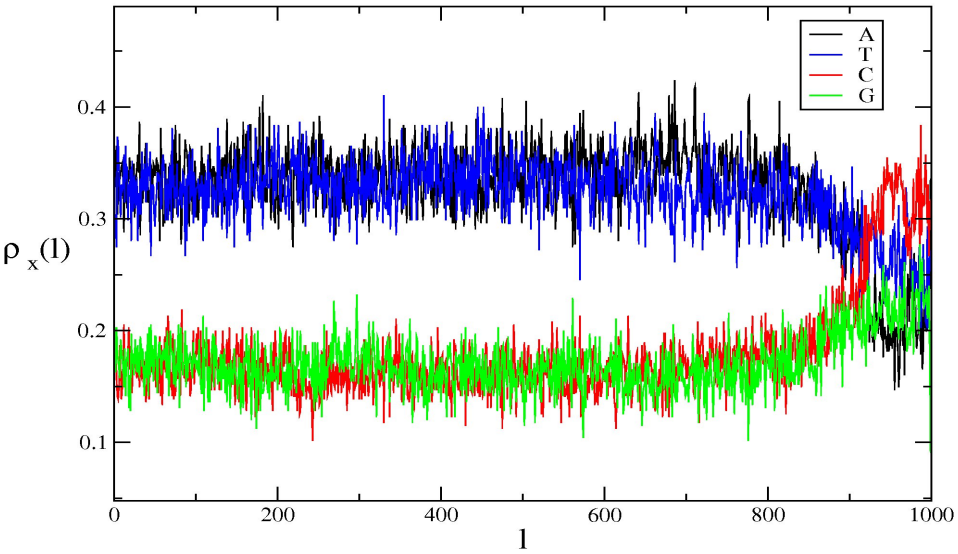


Promoters are dominated by A/T basis and alignment is effective when Performed over the last 100 basis: one obtains 4 clusters dominated by A,T (majority of TATA) and C,G (majority of TATA-less)







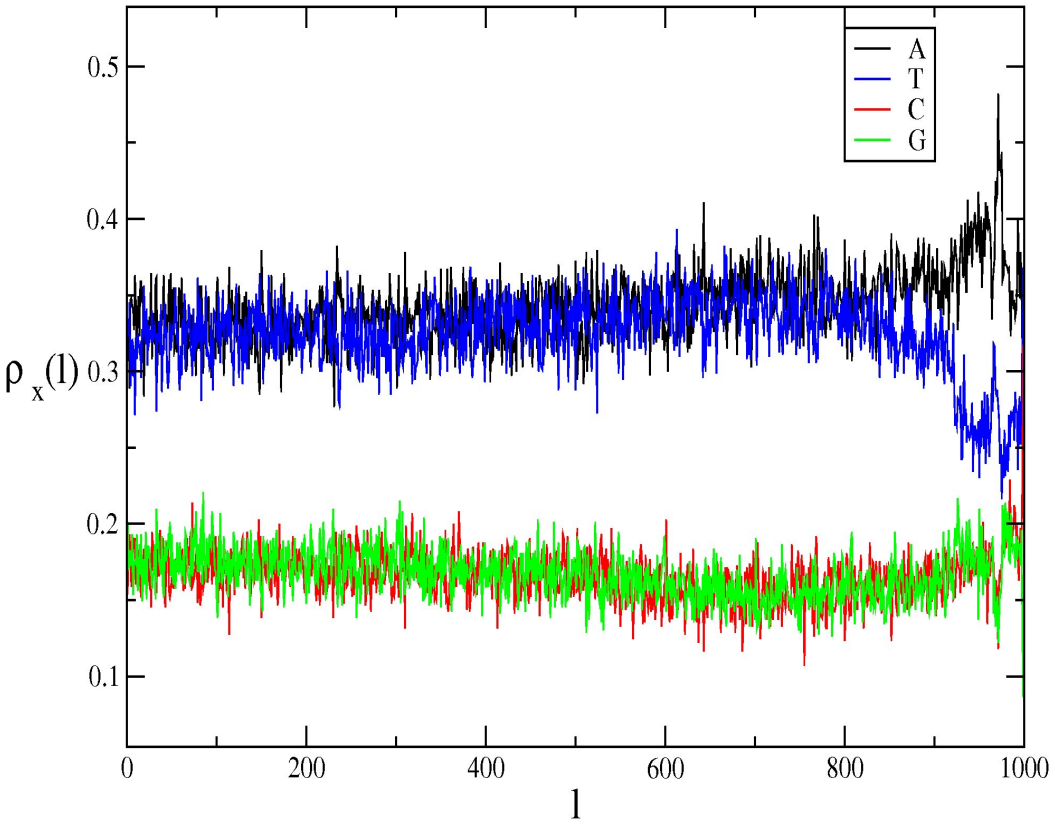


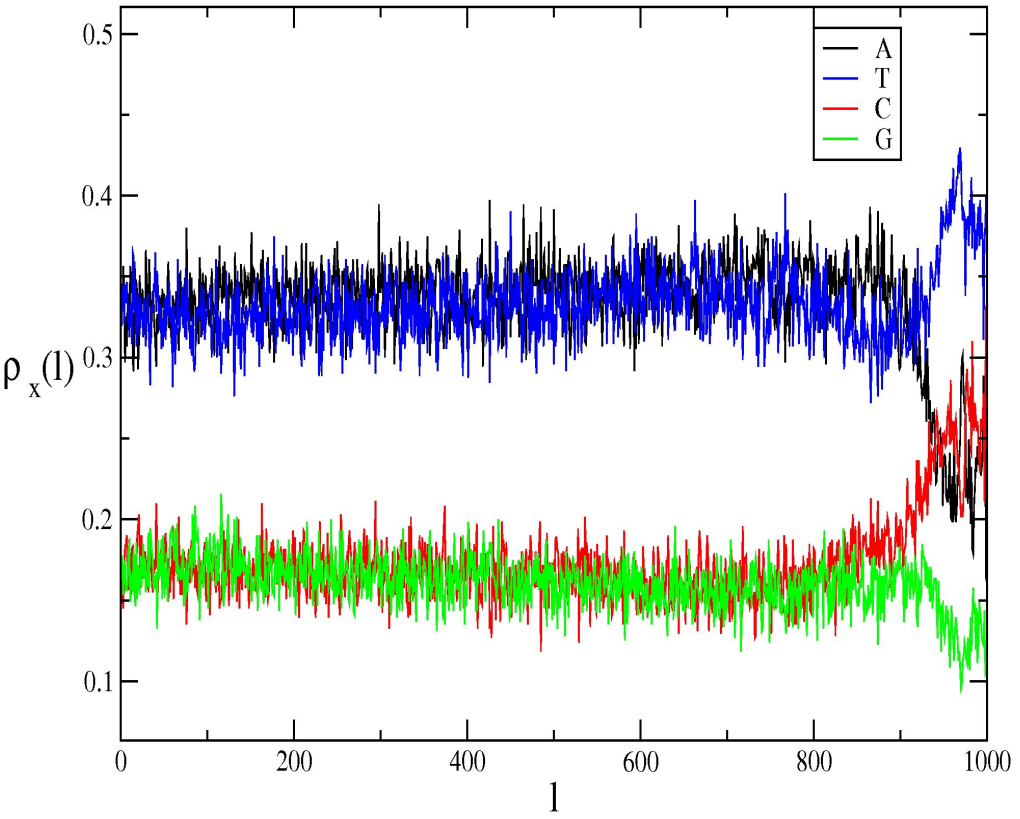
Arabidopsis Thaliana

“Strolling on Chaos, Turbulence and Statistical Mechanics” Rome Sept. 22-24/2014

In honor of Angelo Vulpiani 60th Birthday

As for Danio Rerio , promoters are dominated by A/T basis and the alignement is made over the last 100 basis.
One obtains 2 clusters characterized by an A gradient (majority of TATA) and a C&T gradient (majority of TATA-less)





Different densities of nucleotides along promoters are associated to the presence of "regular subsequences" (motives), where the nucleotides form (quasi)-periodic structures over some finite length, like in the TATA-box.

More generally, one could say that promoters exhibit a mix of ordered and disordered subsequences.

One can work out a spectral procedure for identifying these motives and possibly relating them to gene expression:

- low-affinity regions favouring transcription site recovery through 1-d diffusion (Sela & Lukatsky, 2011)

- structural properties associated to specific regulation functions

Inhomogeneous Disorder in Promoters

acccagacactccgcgcgctccctcgcgcttttgcttgcccgcccgccgccgcctgcttttccctcctgctttcg
ccggtcgccttctccttctttttttttttttttttttttttttttttttttttttttttttttttttttggtaccatagagttgctctgaa

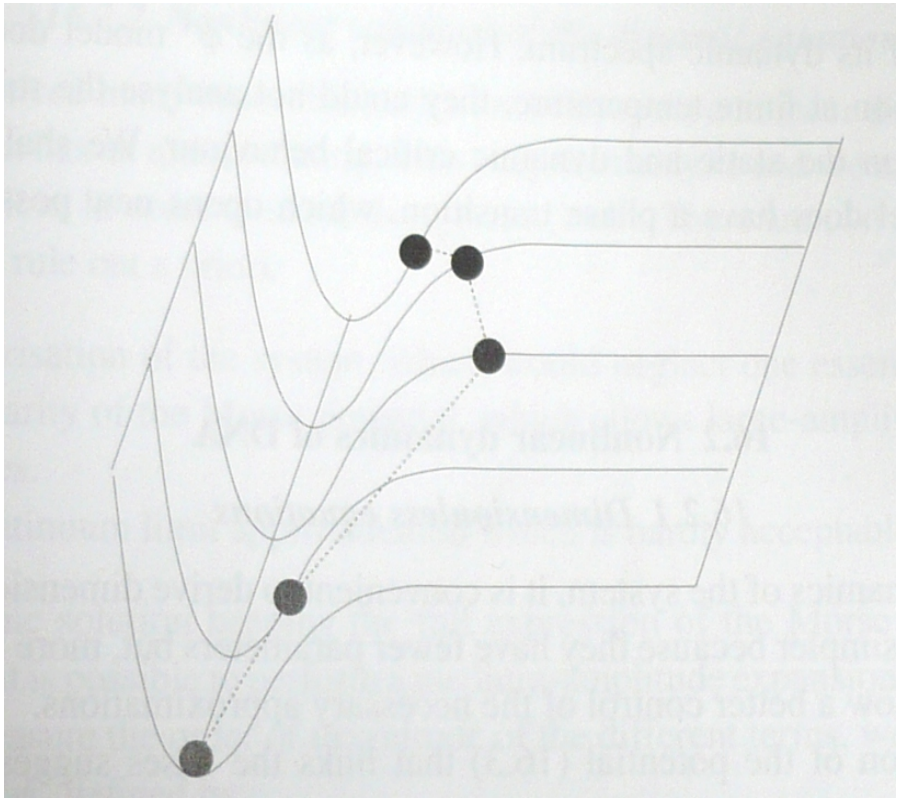
DISORDER yields **LOCALIZATION**

ORDER yields **EXTENSION**

"Strolling on Chaos, Turbulence and Statistical Mechanics" Rome Sept. 22-24/2014
In honor of Angelo Vulpiani 60th Birthday

Peyrard-Bishop Potential:
n.n. stacking interaction along the DNA strand plus inter-strand coupling between nucleotides (local dicotomic disorder due to H-bonds)

$$U = \sum_i \left[\frac{K}{2} \left(1 + \rho e^{-\alpha(y_{i+1}+y_i)} \right) (y_{i+1} - y_i)^2 + d_i (e^{-a_i y_i} - 1)^2 \right]$$

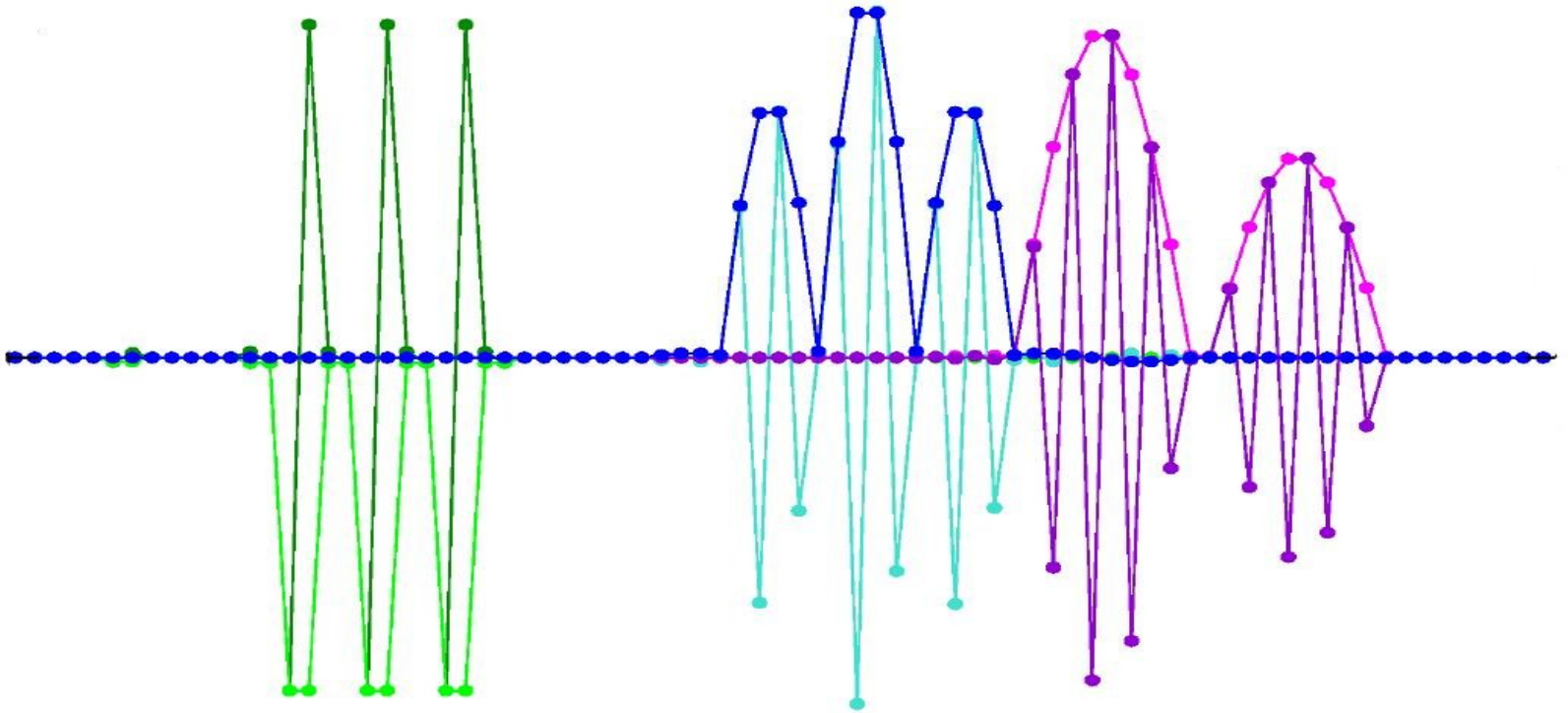


Small oscillation regime:
Hessian matrix

$$H_{i,j}^{MIN} = \left[\frac{\partial^2 U}{\partial y_i \partial y_j} \right]_{y_i=0}$$

Eigenvalues and eigenvectors

Regular subsequences are characterized by eigenvectors that are significantly different form zero over the subsequence extension and as many as the subsequence length in lattice units



cctcttgcaagcatcagccttggaaacaaaacaaaacaaaacaaaaaaaacctttttttatggagcagcc
1101001100110010111001100010000100001000010000000011000000001101101111

Indicators for identifying (normalized) extended eigenvectors of the Hessian Matrix

Center of mass

$$x_k^{cm} = \frac{\sum_{i=1}^{N} |e_k(i)| \cdot i}{\sum_{i=1}^{N} |e_k(i)|}$$

Variance

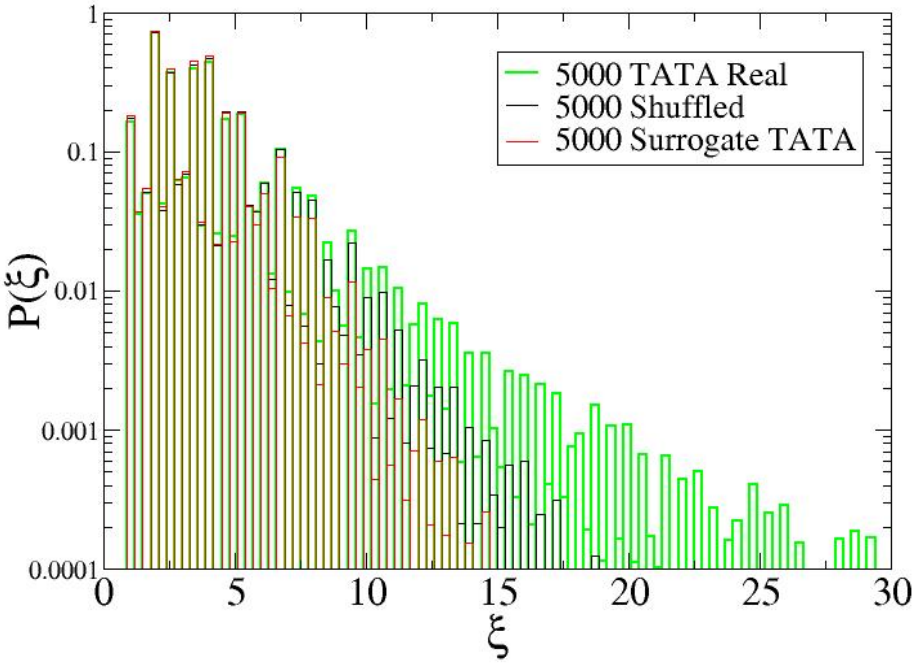
$$\Delta = \sqrt{\left(\frac{\sum_{i=1}^{N} |e_k(i)| \cdot i^2}{\sum_{i=1}^{N} |e_k(i)|} \right) - (x_k^{cm})^2}$$

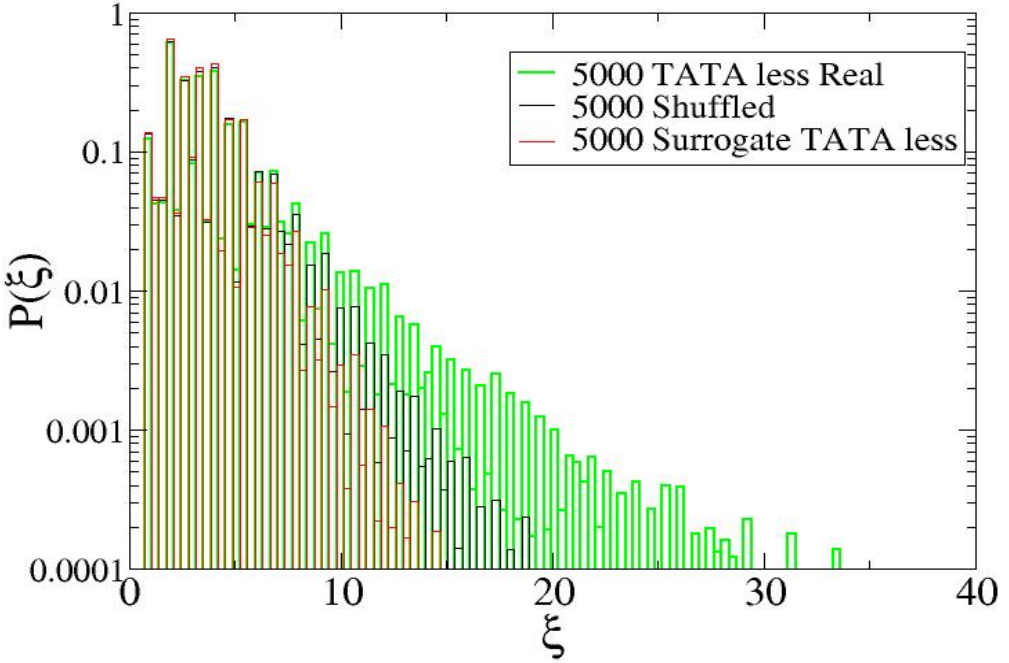
Participation Ratio

$$\xi_k = \left(\sum_{i=1}^{N} |e_k(i)|^4 \right)^{-1}$$

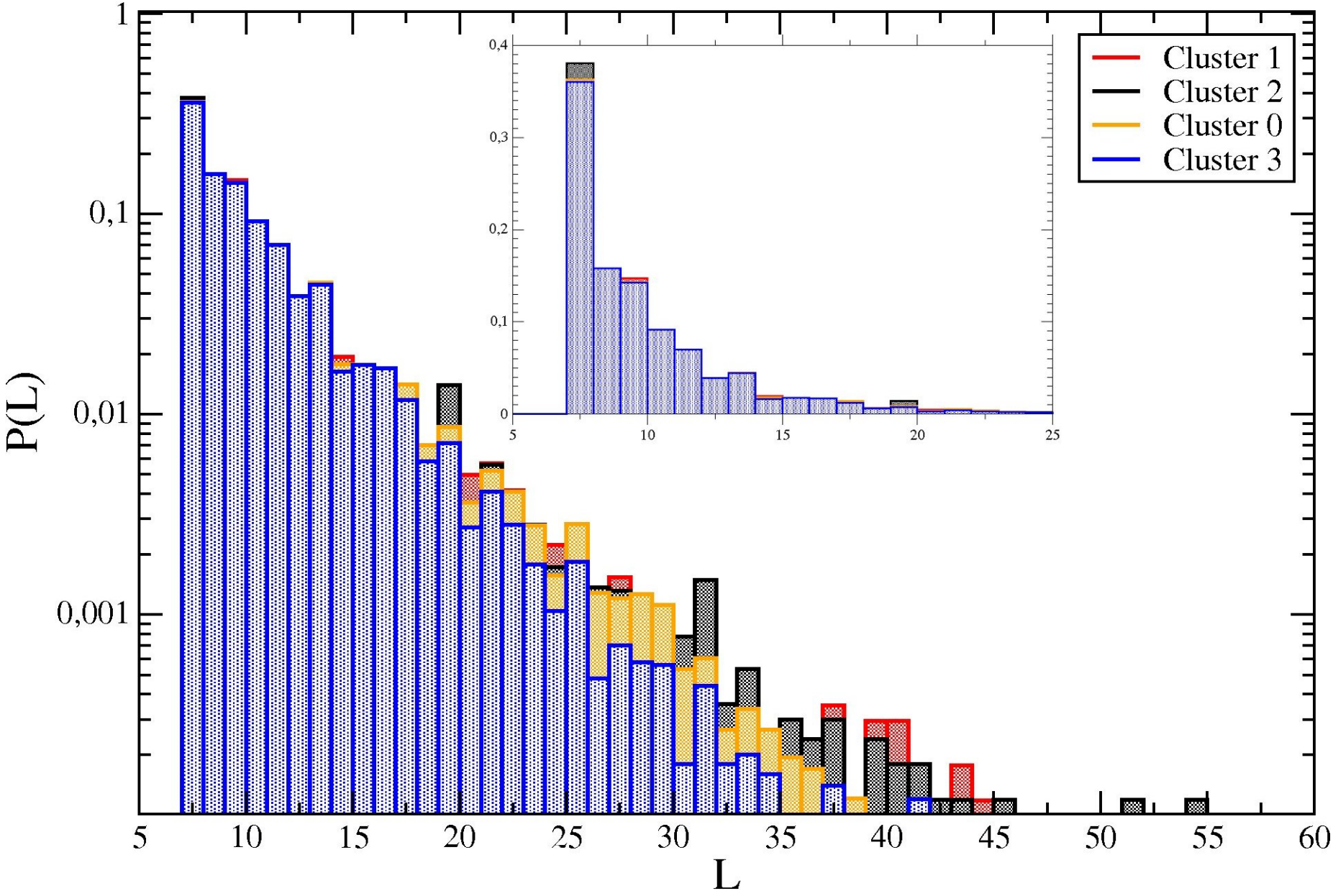
localized $\xi \simeq 1$ extended $\xi \simeq N$

Probability distribution of the participation ratio: comparison with surrogate and shuffled sequences

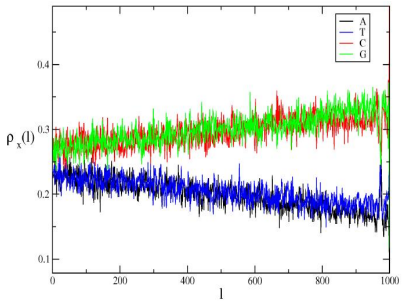




Distribution of regular sequences in the 4 HS clusters

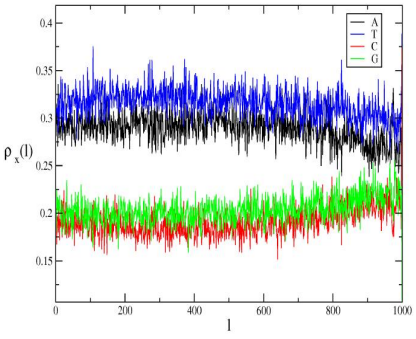


Identificaton of regular subsequences



Cluster 0 seq. 0111110	
Conteggi	Sequenza
126	TGGGGGA
122	TGGGGCT
121	TCCCCCA
116	TCCCCCT
113	AGCCCCA
...	
8	ACGCGGT
8	ACGCCGT
7	ACGGCGT
7	ACCGCGT
6	TCGGCGT

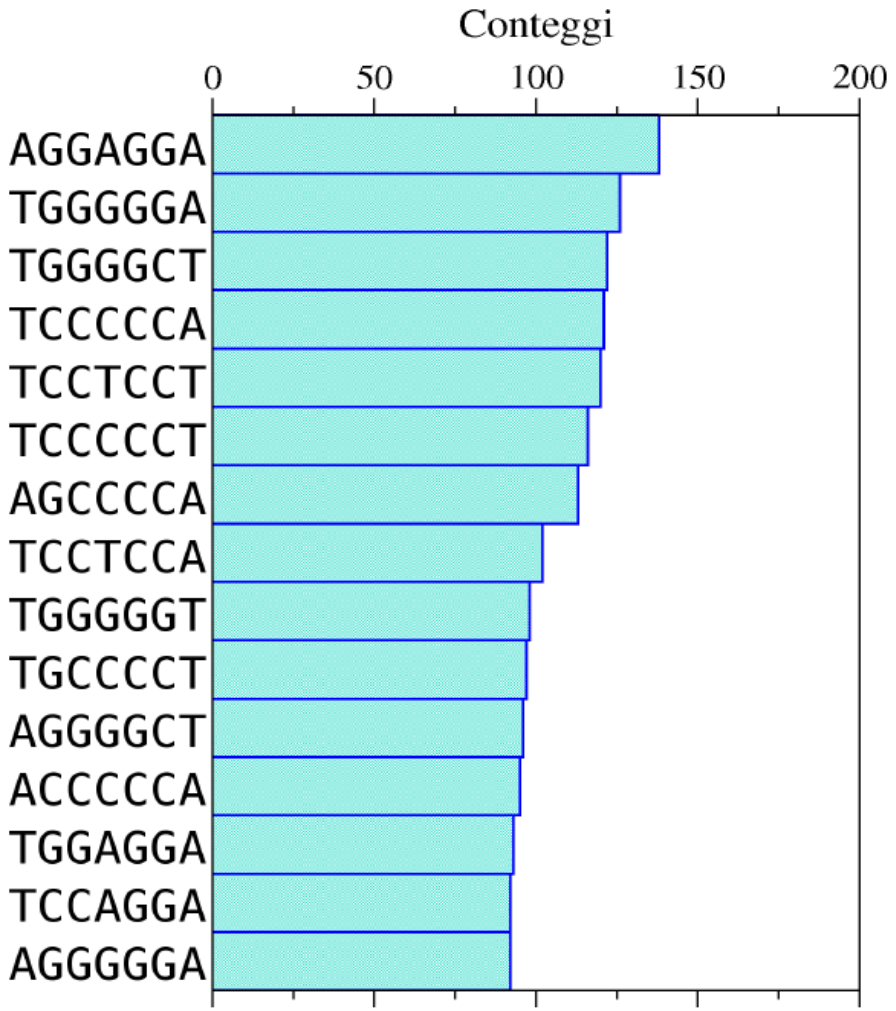
Cluster 3 seq. 10000001	
Conteggi	Sequenza
80	CTTTTTTC
73	CATTTTTC
72	GAAAAATG
66	CATTTTTG
62	GTTTTTTC
...	
9	GAATTATC
9	CAATTAAC
8	GTTATAAC
7	GATATATC
6	GATATAAC



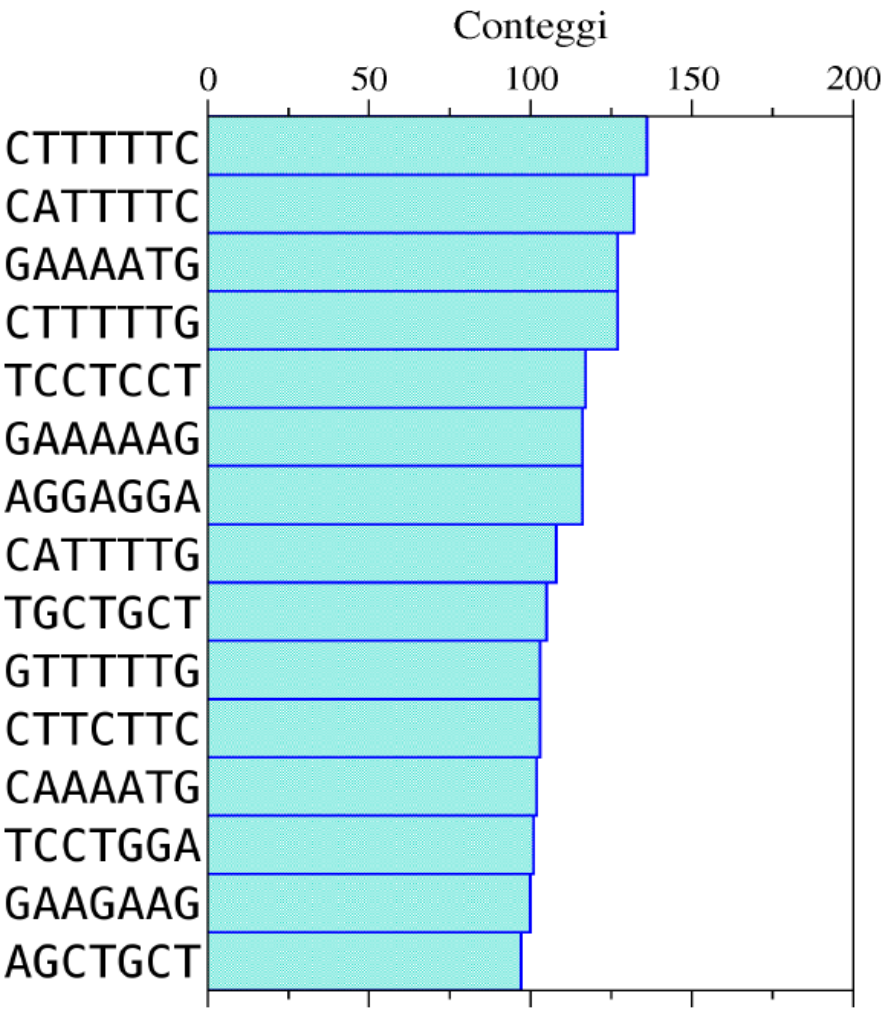
Quaternary sequences exhibit different frequencies

In HS clusters 0 and 3 the most frequent subsequences are of lenght 7 and appear in 10-15% of the promoters

Cluster 0

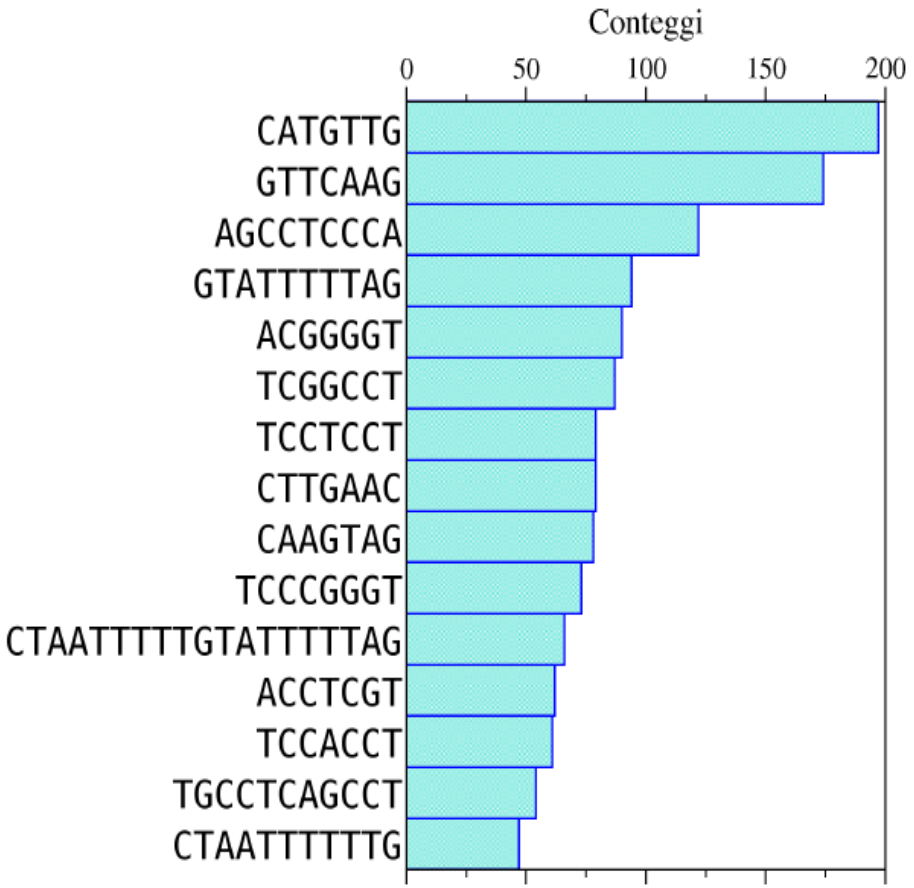


Cluster 3

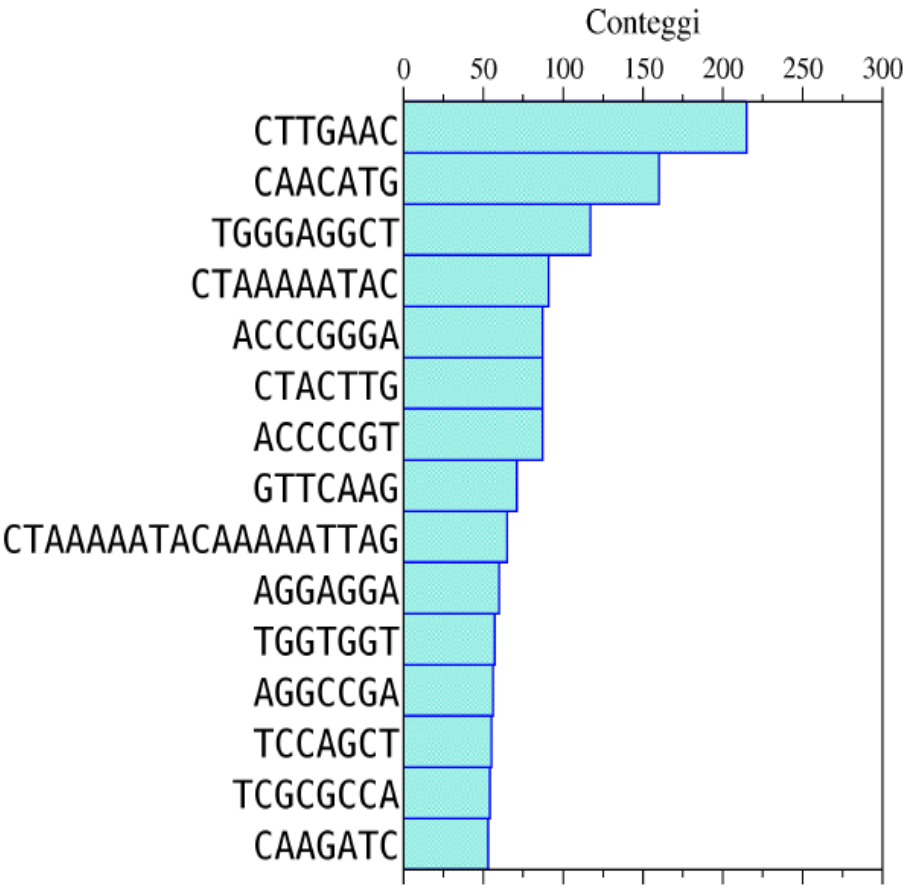


In HS clusters 1 and 2 the most frequent subsequences are typically larger and the most common appear in 50% of the promoters (complementary): correlated to transposons

Cluster 1



Cluster 2



The highly expressed subsequences in HS-Clusters 1 and 2 are typically located far from the TSS and are correlated to transposons and gene regulation (SP1 and AML1-a) or morphogenesis (CdxA)

Some highly expressed subsequences in HS-Cluster 0 (TATA-less rich cluster) are located everywhere along the promoter and typically do not correspond to specific functions (low-affinity ?)

In HS-Cluster 3 (TATA rich cluster) there are no highly expressed subsequences, while most of them are found to be associated to specialized regulation functions, like those belonging to the TATA family

CONCLUSIONS

The spectral methods discussed in this seminar amount to a general protocol for identifying clusters of promoter sequences and the regular subsequences, correlated to their regulatory functions in any living organisms whose DNA has been sequenced.

A relation with evolutionary trends in the selection of the base composition of promoters had already been conjectured by BCA and has been confirmed by these methods, although a more detailed and systematic analysis is still in progress.

L. Pettinato, E. Calistri, F. Di Patti, R.L. and S. Luccioli, PlosOne, 9 e85260 (2014)

PERSPECTIVES

Promoters Clustering suggests new directions of investigation

- The structure of genetic networks can be reconsidered by attributing a "cluster tag" to annotated genes, according to their promoters: quite interesting preliminary results

- More refined entropic indicators confirm that information content in promoters is mainly stored in the regular motives characterizing the different clusters (positional entropies JTB 2014 and Marsili et al. JSM 2013 (work in progress))

- Dynamical studies (promoters modelled as nonlinear chains) indicate that energy transport in this inhomogeneous disordered sequences exhibits quite unexpected features (work in progress)