Applications of Network Entropy to gene expression time series data

G. Menichetti[1], G. Bianconi[2], E. Giampieri[1], G. Castellani[1], D. Remondini[1]
[1]Dept. of Physics, Bologna University and INFN, Bologna, IT
[2]Dept. of Physics, Northeastern University, Boston, USA
daniel.remondini@unibo.it

The statistical properties of integrated PPI-mRNA expression networks seem to be good observables in order to investigate systemic pathologies as cancer and ageing. It is widely recognized that approaches based on single-gene differential expression analysis are too simplistic and have been replaced by network-based methods, in a holistic or Systems Biology view. In particular, we suggest the using of Network Entropy (see [*K. Anand, G. Bianconi, Physical Review 2009 E 80, 045102(R)*]) and its related Lagrangian multipliers as global observables to study high-throughput gene expression time series data.

DATASETS

1) Time series gene expression dataset of human T cells, collected from peripheral blood of 25 healthy human donors of different age from 25 to more than 95 years, in order to characterize changes that occur throughout the entire adult lifespan. We equally divided the subjects into five age-classes: A (25-35 years old), B (40-50 years old), C (55-65 years old), D (70-80 years old) and E (> 90 years old).
2) The PPI network extracted by means of the meta-search engine called APID (*http://bioinfow.dep.usal.es/apid/index.htm*). Putting as a query 1600 gene dataset previously selected (see [*Remondini et al., Mol Biosys 2010, 6, 1983-1992*]), APID produces a protein-protein interaction network of 426 nodes, that is the final gene expression dataset we considered for our analysis.

CANONICAL DISTANCE NETWORK ENSEMBLE

The entropy of a canonical network ensemble is the logarithm of the number of graphs that satisfy a set of constraints in average.

$$S=-\sum_{i<j}^{N}\sum_{l}^{Nb}\chi_l\left(d_{ij}\right)p_{ij}\log p_{ij}-\sum_{i<j}^{N}\sum_{l}^{Nb}\chi_l\left(d_{ij}\right)\left(1-p_{ij}\right)\log\left(1-p_{ij}\right)$$

- ***p***$\boldsymbol{p_{ij}}$ probability of having an interaction between the protein ***i*** and the protein ***j***

- Fixed degree sequence vector ***k*** given by the adjacency matrix of the PPI network (“configuration ensemble”)

$$k_i=\sum_{j\neq i}^{N}p_{ij}$$

- Metrics: difference of the expression values of each gene

$$d_{ij}=\left|e_i-e_j\right|$$

- Distance binning with 20 bins (N_b=20)
- Fixed B_l links per bin

$$B_l=\sum_{i<j}^{N}\chi_l\left(d_{ij}\right)p_{ij}$$

MAXIMIZATION FUNCTION

$$F=S+\sum_{i}^{N}\lambda_i\left(k_i-\sum_{i<j}^{N}\sum_{l}^{Nb}\chi_l\left(d_{ij}\right)p_{ij}\right)+\sum_{l}^{Nb}g_l\left(B_l-\sum_{i<j}^{N}\chi_l\left(d_{ij}\right)p_{ij}\right)$$

LINK PROBABILITY DISTRIBUTION

$$p_{ij}=\sum_{l}^{Nb}\chi_l\left(d_{ij}\right)\frac{z_i z_j W_l}{1+z_i z_j W_l}$$

LAGRANGIAN MULTIPLIERS

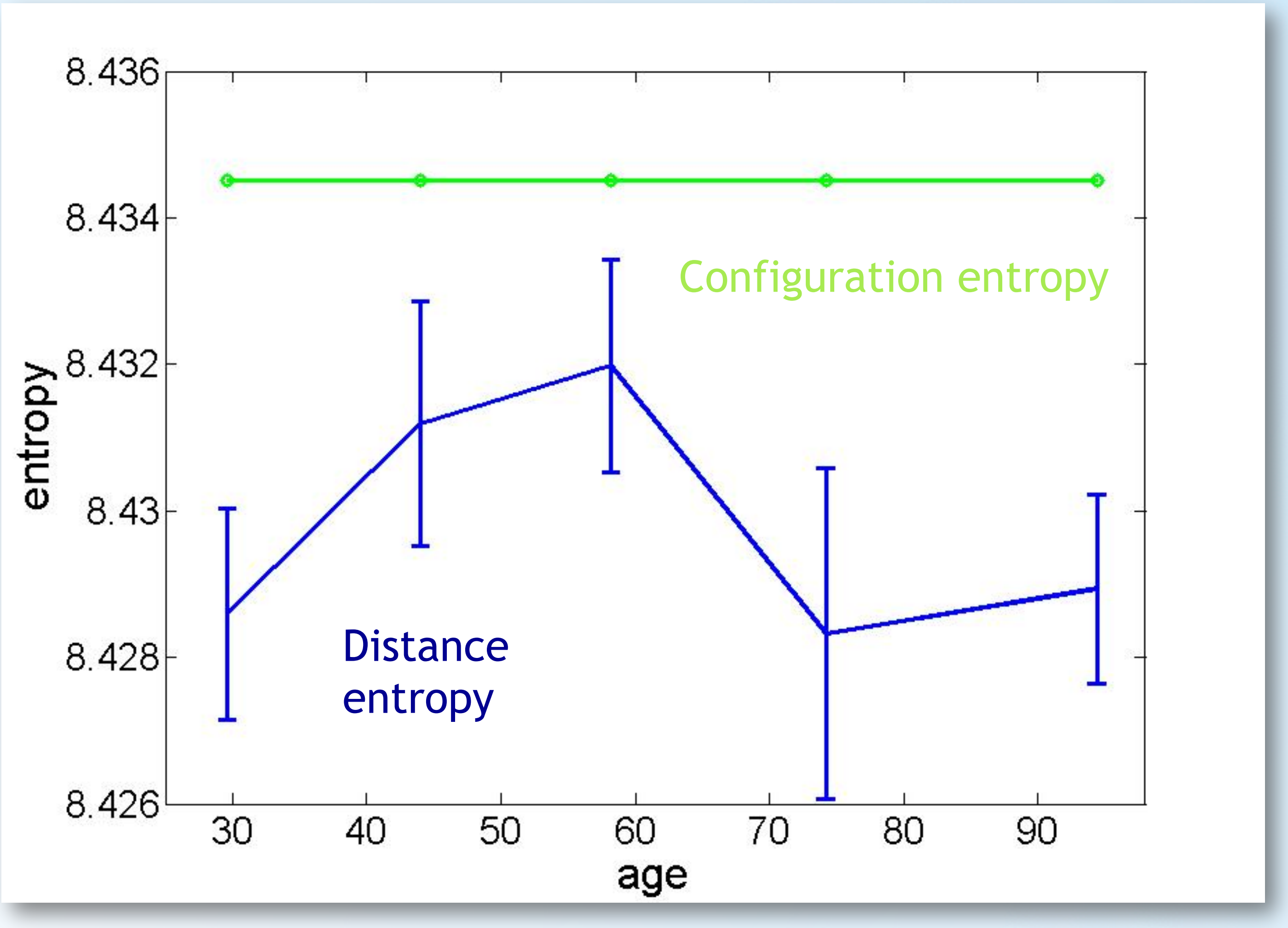
$$z_i=e^{-\lambda_i}$$

$$W_l=e^{-g_l}$$

ALGORITHM IMPLEMENTATION

We developed an iterative algorithm in MATLAB and C++ that calculates the value of the canonical entropy and of the Lagrangian multipliers depending on the different constraints we ask the network to satisfy.

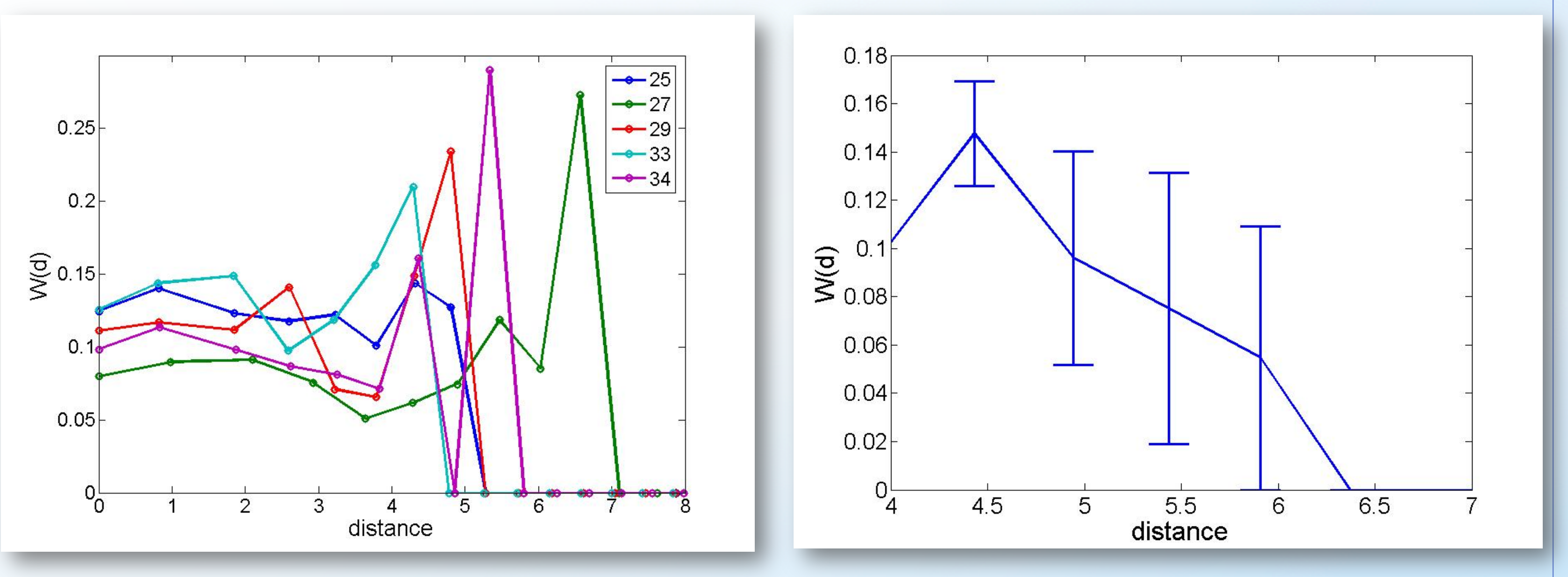
RESULTS: CANONICAL ENTROPY



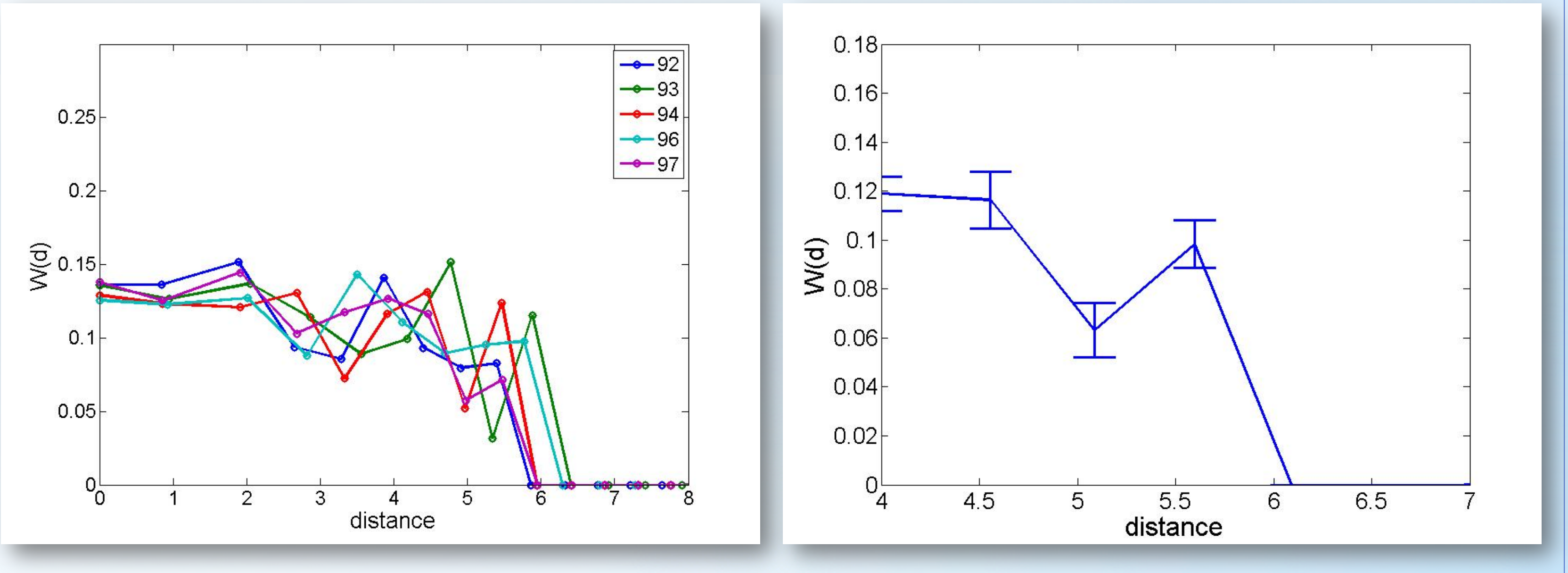
- Non-monotonic trend in time of the distance entropy
- The youngest mean value is similar to the oldest mean value
- The distance entropy is always lower than the configuration entropy, so the distance matrices given by the time series gene expression dataset introduce a significant amount of information in our ensembles and then reduce the possible configurations.

RESULTS: LAGRANGIAN VECTOR W

Age 25-34



Age 92-97



The vector of the Lagrangian multipliers $\boldsymbol{W_l}$ or ***W(d)*** is important because it gives us the statistical weight depending on the distance of nodes, that modulates the probability $\boldsymbol{p_{ij}}$ of having a link between the node ***i*** and the node ***j*** at distance $\boldsymbol{d_{ij}}$. This measure is the main difference between a “distance network ensemble” and a “configuration ensemble”.
We see that the average values are similar but the variances of the oldest people are quite smaller.