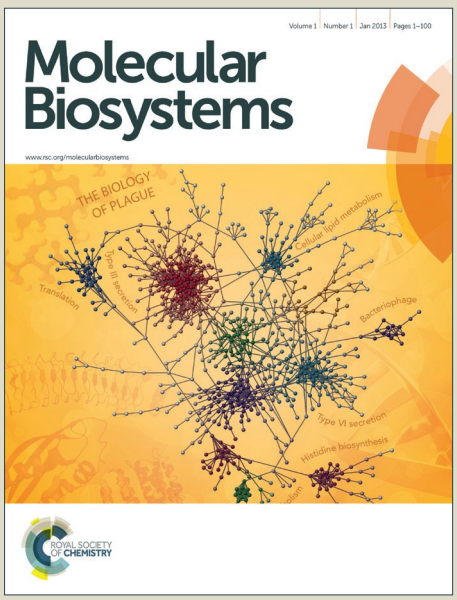
Molecular BioSystems

Accepted Manuscript

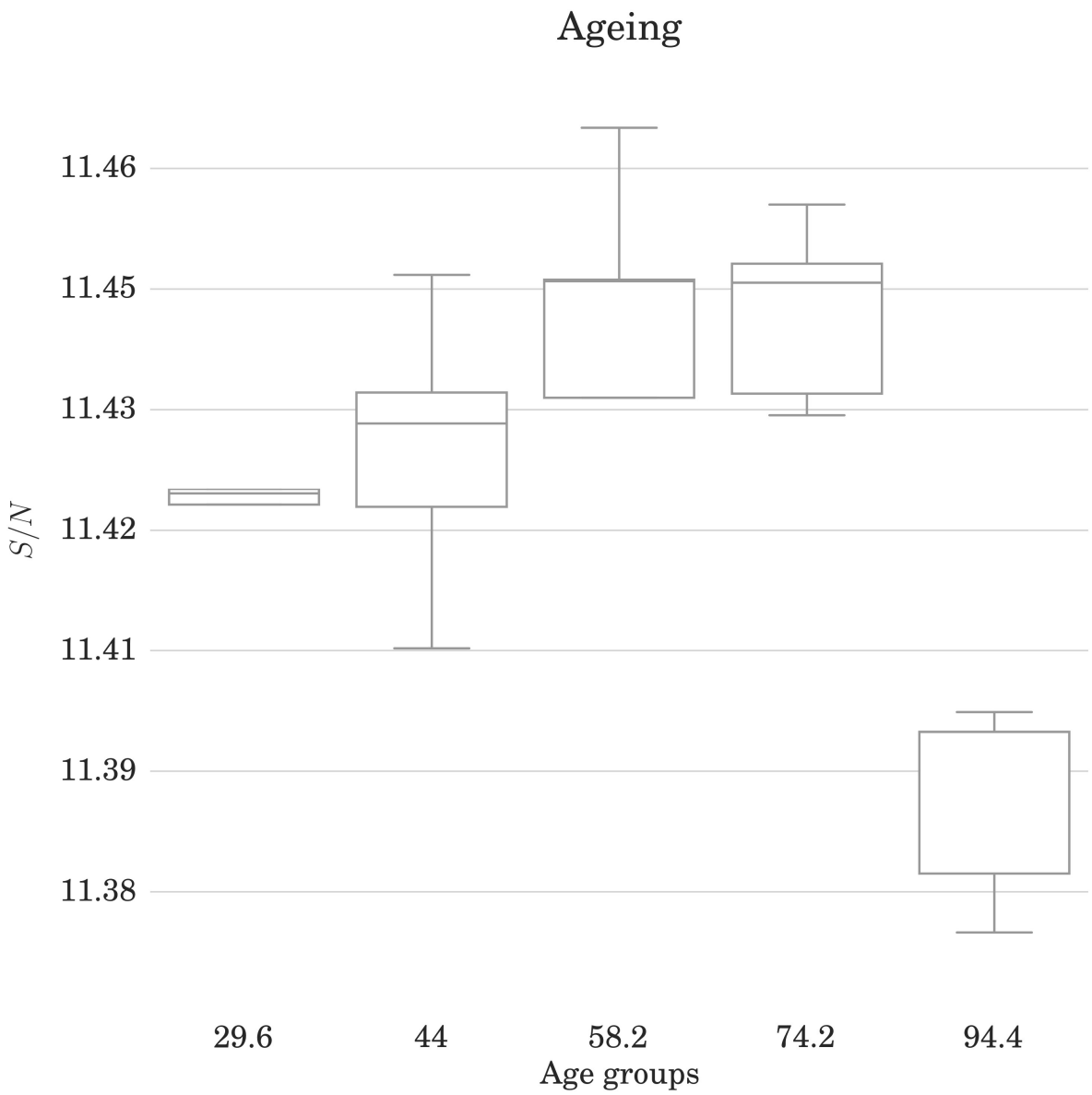


ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Multiscale characterization of aging and cancer progression by a novel Network Entropy measure

G. Menichetti[1], G. Bianconi[2], G. Castellani[1], E. Giampieri[1], D. Remondini[1]
[1]*Department of Physics and Astronomy and INFN, Bologna University, Viale B. Pichat 6/2 40127 Bologna, Italy*
[2]*School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom*



We characterize cancer and ageing cell states with a multiscale network entropy approach, measuring "parameter space" available to the cell.

Multiscale characterization of ageing and cancer progression by a novel Network Entropy measure

Giulia Menichetti[1], Ginestra Bianconi[2], Gastone Castellani[1], Enrico Giampieri[1] and Daniel Remondini[1]

[1]Department of Physics and Astronomy and INFN, Bologna University, Viale B. Pichat 6/2 40127 Bologna, Italy
[2]School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom

Abstract

We characterize different cell states, related to cancer and ageing phenotypes, by a measure of entropy of network ensembles, integrating gene expression profiling values and protein interaction network topology. In our case studies, network entropy, that by definition estimates the number of possible network instances satisfying the given constraints, can be interpreted as a measure of the "parameter space" available to the cell. Network entropy was able to characterize specific pathological conditions: normal versus cancer cells, primary tumours that developed metastasis or relapsed, and extreme longevity samples. Moreover, this approach has been applied at different scales, from whole network to specific subnetworks (biological pathways defined on *a priori* biological knowledge) and single nodes (genes), allowing a deeper understanding of the cell processes involved.

1 Introduction

Biological systems can be seen as complex systems that translate genomic information into phenotypes [1, 2]. A useful approach is to describe these systems as networks, with the system elements (e.g. genes, proteins) as nodes, and the relationships between them (e.g. transcription or protein-protein interaction) as edges [3, 4]. An important class of biological networks comprises the protein-protein interaction networks (PPI [5, 6, 7]): edges in these networks describe interactions between proteins that are part of the same physical complex or post-translational modifications mediating signal transduction flows. Networks of interacting proteins can be thought as characterizing the cell phenotypes given their genetic and transcriptomic profile.

These and other interactions are also encoded into functional pathways, such as signalling and metabolic pathways, as are mapped for example in KEGG database (Kyoto Encyclopaedia of Genes and Genomes, *www.genome.jp/KEGG*). In our study we are interested in the integration between the transcriptomic

and the interactomic data, thus the statistical properties of integrated PPI-signalling-mRNA expression networks seem to be good observables to investigate systemic pathologies such as cancer and ageing [8, 9]. This approach can be more informative than analyzing gene expression data on its own. Indeed, integrative PPI-mRNA expression studies have helped to tease out relevant patterns of expression variation in the contextual framework of signalling pathways and protein complexes [1, 10, 11].

Following the recent developments in statistical mechanics of complex networks, we have the chance to build up a thorough biological network model. Thanks to some suitable constraints encoding the most relevant network features, we can evaluate the information content of biological structures, and moreover, we can apply the same approach to time-dependent and time-independent data [12, 13].

As explained in the following, our approach relies on the theory of network ensembles with given topology (encoded in the degree sequence) and metrics (represented by distance between values assigned to the nodes). In our case, the information on PPI-signalling structure is embedded in the network topology, and is mapped onto our model by imposing a constraint on the degree sequence of the networks belonging to the ensemble. Moreover, mRNA expression profiles for each sample are introduced in the model as values on the nodes of the network (corresponding to genes) and the distance between gene values, corresponding to links in the PPI network, are collected in a distribution. Our model considers as a constraint for the calculation of network entropy the number of links per bin bsed on such distance distribution. For further details on the method, see Figure (1) and Supplementary Material Section, in which we show the results of this approach applied to a clear toy model, and the comparison with the results obtained with a similar approach previously appeared in literature [8, 10].

We studied two biological phenomena that encode different landscapes of cellular perturbation, namely cancer and ageing in humans, and whose datasets were characterized by a different experimental design (case-control studies and a time series built on samples of different age). Network entropy approach offers a new perspective to the study of such phenomena, highlighting a more systemic behavior of the cell beyond single-element analysis, but nonetheless it can be applied at several scales, from a whole-cell point of view (the full network) to single biological pathways characterizing the main cell processes like metabolism and signaling (subnetworks defined by a priori biological knowledge), up to single nodes (genes/proteins in the network).

2 Methods

2.1 PPI-signalling network

In order to define a network in which the nodes (namely proteins, measured by their mRNA transcription profile) could be adequately annotated both in terms of their biological function and their potential interactions, we considered only the genes that were annotated both in KEGG database and in Pathway-Commons (*www.pathwaycommons.org*) PPI network.

We started considering the protein-protein interaction network extracted

from the Pathway Commons database regarding *Homo Sapiens* proteins. The initial PPI network contained 11604 nodes and 420601 links: after self-interaction and redundant annotation removal we obtained a giant component of 11394 nodes and 420516 links. Since we used different gene expression datasets on different microarray platforms, we considered the intersection of the PPI protein IDs with the gene annotations of each microarray platform, considering only the genes that had also a known annotation in the KEGG database. In this way, each network could be further divided considering nodes annotated into each single KEGG pathway (see Supplementary Table 1). This procedure produced different networks for each considered platform, with a number of nodes ranging from 2000 to 3000.

2.2 Cancer datasets

The analysis has been performed onto four datasets by downloading the normalized data from GEO Omnibus (www.ncbi.nlm.nih.gov/geo). The first dataset (referred to as "Colon", GEO accession number GSE4183 [14]) is composed by 8 normal colon biopsies and 15 colorectal cancer samples. The second dataset is related to Ewing's sarcoma ("Ewing" dataset, GEO accession number GSE12102 [15]), consisting of 30 primary tumour samples and 7 metastatic samples.

Other two datasets refer to breast cancer samples: in the first we have primary tumour specimens that developed metastasis or not (97 and 28 samples respectively, referred to as "Met", GEO accession number GSE2990 [16, 17]), while in the second there are primary tumour biopsies that relapsed or not (107 and 179 samples respectively, referred to as "Rel", GEO accession number GSE2034 [18]).

Colon and Ewing datasets are both profiled on the Affymetrix U133 plus 2 microarray platform, and the intersection with the PPI network and the KEGG database resulted in a network with 2835 nodes. Rel and Met datasets are both profiled with the Affymetrix U133 A microarray platform, and the intersection with the PPI network and the KEGG database resulted in a network of 2618 nodes.

In each dataset, a restricted gene list (and a corresponding induced subnetwork) was obtained by performing a Student's T test for uncoupled samples over the two groups in which each dataset is divided into, in order to evaluate the behaviour of the network entropy measure for a subset of nodes that significantly behave differently in the two groups, as compared to the full set of available nodes in the network.

For the Colon dataset we applied a $P < 0.05$ significance threshold plus Benjamini-Hochberg post-hoc correction, obtaining a subnetwork of 312 nodes. For the Ewing, Rel and Met datasets we only applied a $P < 0.05$ significance threshold, obtaining a network with 136 nodes for Ewing, 151 and 313 nodes for Met and Rel datasets respectively, since almost no genes would have passed the post-hoc correction. This is probably due to the fact that in these datasets the differences between groups are less pronounced than in a normal-cancer comparison, as described in the related papers from which the data were collected.

Since we can calculate the network entropy value for each sample, we obtain 23 entropy values for Colon, 37 for Ewing, 125 for Met and 286 for Rel datasets, both for the full network (that will be used for single-node entropy calculation,

as described below) and the 5% significance gene selection.
Since the null distribution of network entropy values is not known in advance for arbitrary networks, we performed nonparametric Wilcoxon rank sum tests between the entropy values for each group.

2.3 Ageing dataset

We considered a cross-sectional study (time series) of 25 whole-genome expression profiles of T lymphocytes extracted from healthy males of ages spanning typical adult human lifespan (from 25 to 97 years, see [19] for further details). This dataset could be divided into 5 age groups with about 10 y between each group: A) 25-34 y (mean = 29.6 y); B) 43-46 y (mean = 44 y); C) 55-62 y (mean = 58.2 y); D) 70-79 y (mean = 74.2 y); E) 92-97 y (mean = 94.4 y). The gene expression dataset (obtained through a custom array, see [19]) after processing is composed by 13103 probes x 25 age samples. The intersection with the giant component of Pathway Commons data and the KEGG database results in a PPI network of 1976 nodes, used for single-node entropy analysis. A restricted gene list was obtained by performing a 1-way Anova over the age groups, in order to look for genes significantly changing expression profile in time. With a $P < 0.05$ significance threshold plus Benjamini-Hochberg post-hoc correction we obtained a subnetwork of 217 nodes. We applied the same significance threshold considered in the original paper in order to compare the results obtained by gene expression analysis and the results obtained by this network entropy approach.

We obtained 25 network entropy values (one for each sample) both for the whole network and for the 5% significance gene selection. Also in this case we applied nonparametric test for network entropy comparisons, namely Kruskal-Wallis test over the 5 age groups to define a subgroup of genes significantly changing expression profile over the whole time series, and Wilcoxon rank sum test for comparison between any two groups.

2.4 Entropy of network ensembles

In this paper, based on the formalism developed in [20], we apply the concept of Entropy of network ensembles to a real biological situation, extending in more detail the necessary formalism and implementing the algorithm to calculate the Entropy values and all the related variables. An extended example on a particular network model is shown in the Supplementary Materials Section. The concept of “ensemble" is inherited from statistical mechanics, where it indicates a large number of copies of a system, representing the possible “microstates" in which the real system might be, given a specified “macrostate".

A macrostate is characterized by a specific set of observables: for example, in an ideal gas we can calculate (by the entropy function) the number of atomic configurations (microstates) corresponding to the macroscopic system at a specific temperature (the macrostate). What statistical mechanics tells us is that, given the constraint of fixed temperature, the distribution of the possible microstates that maximizes entropy is a gaussian distribution for the velocities. This is thus the most likely situation (in terms of probability) that can happen in a real instance of an ideal gas with fixed temperature. In a similar way, a real network (with specific links and weights) can be seen as a specific instance of

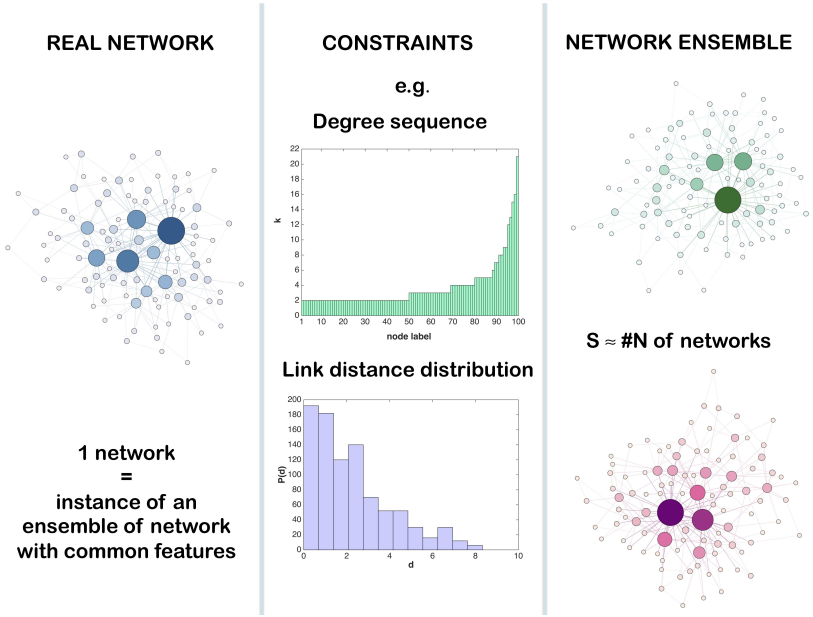


Figure 1: Description of the Network Entropy method. From left to right: given a real network (with specific number of nodes, links and weights), we consider it as an instance of a larger ensemble of networks sharing common features. For our study, we characterized as an ensemble all the networks with the same degree sequence and link distribution than the original network (the considered canonical ensemble satisfies these constraints on average). By calculating maximum entropy of the ensemble satisfying given constraints, we can count the number of networks belonging to the ensemble, thus estimate its extension in the space of all the possible networks. We remark that given the p_{ij} obtained by entropy maximization, we can generate any network belonging to this ensemble. In the right panel we show two networks generated from the same ensemble, similar but not identical to the original one (in terms of the defined constraints).

a larger set of networks (an *ensemble*) with similar features, such as a number of links, or degree sequence (i.e. the number of interactions of each node) that constitute the "macroscopic observables" characterizing our system, analogous to temperature for the ideal gas (see Fig. 1 for a pictorial description).

In our case, the system is a network of N nodes, described by an adjacency matrix A (a_{ij}, $i, j = 1, \ldots, N$) with weights d_{ij}, and the macrostate is specified by two sets of observables, thus a more complex situation than tipically studied in Statistical Mechanics (as compared to the single constraint introduced by fixing system temperature). The first set of observables is related to the network topological structure, and is given by the degree sequence of the PPI network, namely the N-dimensional vector of the connectivity degree of each node: $\{k_i\}, i = 1, \ldots, N$, with $k_i = \sum_j a_{ij}$. Since we consider a network (and calculate an entropy value) for each sample, this set of topological constraints

is equal for all the samples[1]. The second set of observables is related to the weights of the network, expressing metric relations between nodes: assigning to the nodes of each sample the values of mRNA expression of the corresponding genes in the selected microarray g_i^a(with index i ranging over all the nodes and index a ranging over all the samples of the dataset), we define the weights d_{ij}^a as the euclidean distance of the gene expression values:

$$d_{ij}^a = \sqrt{(g_i^a - g_j^a)^2} = |g_i^a - g_j^a| \tag{1}$$

We collect all these values into an histogram with Nb bins, with a number of bins equal to the square root of the number of nodes in the network i.e., $Nb = \sqrt{N}$. In our analysis Nb has range $15 - 20$ and the results appear robust against variations of ± 5 bins. For each couple of genes we have a particular distance value but not necessarily a link in the PPI network. The second set of network observables (i.e. constraints for entropy maximization) refers to the number of the PPI links whose distance values fall in a given bin. For each distance bin, we count the number of PPI links falling inside its boundaries, and impose that the network ensemble has the same average value of links per bin. We remark that this set of observables, based on expression profile, is specific for each sample, and so will be for the corresponding entropy values.

We define the entropy of a network ensemble as

$$S = -\sum_{i<j} p_{ij} \log p_{ij} - \sum_{i<j} (1 - p_{ij}) \log(1 - p_{ij}) \tag{2}$$

in analogy with the definition of entropy for a canonical ensemble, in which the constraints are not satisfied exactly but only on average by the members of the ensemble. The marginals p_{ij} represent the probability of having a link between node i and node j. In a generic graph of this ensemble, a link a_{ij} is present with probability p_{ij}, otherwise absent with probability $(1 - p_{ij})$.
We define the *spatial ensemble* as an ensemble of network obtained by imposing the constraints on the degree sequence $\{k_i\}$ and on the number B_l of PPI links belonging to each distance bin, $d_{ij} \in I_l$, described by the following equations:

$$k_i = \sum_{j}^{N} p_{ij}; \qquad i = 1, ..., N \tag{3}$$

$$B_l = \sum_{i<j}^{N} \chi_l(d_{ij}) p_{ij}; \qquad l = 1, ..., Nb \tag{4}$$

where N is the number of nodes in the network, Nb is the number of bins considered for the empirical distribution of distances, and χ_l is the characteristic function of each bin of width $(\Delta d)_l$: $\chi_l(x) = 1$ if $x \in [d_l, d_l + (\Delta d)_l]$, $\chi_l(x) = 0$ otherwise.

The probability matrix $\{p_{ij}\}$ is obtained by the constrained maximization of the entropy function, as described in the following equation:

$$\frac{\partial}{\partial p_{ij}} \left\{ S + \sum_{i}^{N} \lambda_i \left(k_i - \sum_{j} p_{ij} \right) + \sum_{l}^{Nb} g_l \left(B_l - \sum_{i<j}^{N} \chi_l(d_{ij}) p_{ij} \right) \right\} = 0$$

[1]Measured on the same microarray platform.

where λ_i and g_l are the the Lagrangian multipliers related to our constraints. For each (i, j) the resulting marginal probability is

$$p_{ij} = \sum_{l}^{Nb} \chi_l(d_{ij}) \frac{e^{-(\lambda_i+\lambda_j+g_l)}}{1+e^{-(\lambda_i+\lambda_j+g_l)}} = \sum_{l}^{Nb} \chi_l(d_{ij}) \frac{z_i z_j W_l}{1+z_i z_j W_l} \tag{5}$$

in which $z_i = e^{-\lambda_i}$, $W_l = e^{-g_l}$ are functions of the Lagrangian multipliers λ_i and g_l.

If we consider only the constraint on the degree sequence stated in Eq. (3) we obtain the entropy for the so called *configuration ensemble*. The number of constraints for the configuration ensemble is N, while for the spatial ensemble it is $N+Nb$. We remark that a significant difference between the network entropy calculated in the spatial and configuration ensembles reflects the relevance of the information encoded in the gene expression data integrated on the network, as will be the case for all of our analyses.

The canonical ensemble deriving from a real instance gives an entropy value that estimates the logarithm of the number of "typical" networks in this ensemble, i.e. those network that satisfy on average the given constraints.

Considering the link probabilities p_{ij} obtained for the full PPI network, it is also possible to define a single-node entropy measure for the i-th node. This single-node entropy takes exactly the form of a Shannon entropy for a string, thus it can be interpreted in the same framework of Information Theory. Since the probability values are all positive, and since we know that $\sum_j p_{ij} = k_i$, the connectivity degree of the i-th node, we can define the single-node entropy S_i as follows:

$$S_i = -\sum_j p'_{ij} \log p'_{ij} \qquad p'_{ij} = \frac{p_{ij}}{k_i} \tag{6}$$

Given the single node entropy values $\{S_i\}$ for each sample, we checked by a nonparametric Wilcoxon rank sum test for significant differences at a single node level between the groups of our datasets. Since we know the KEGG annotation for each gene of our network, we also performed a functional analysis of specific biochemical pathways, based on enrichment analysis of pathways by genes significantly changing their single-node entropy value S_i. In this way the entropy analysis could be scaled from the full PPI network to single-node and single-pathway level.

Taking advantage of the a priori biological knowledge available from the KEGG database, we remark that it is indeed possible to obtain several subnetworks of the initial PPI network: at a first level, the genes annotated in the PPI can be divided into 6 functional groups, that can be further subdivided into 42 metapathways, and again into 191 KEGG biological pathways (see Supplementary Tables 1 and 2). We decided to apply our analysis at the pathway level, in order to gain more information on the single known biological mechanisms described into the KEGG database.

For the calculation of the entropy values, of the link probabilities p_{ij} and of the Lagrangian multipliers, we developed an iterative algorithm (see Supplementary Materials Section for an extended description of the implementation and its performance): given a random starting guess for the value of the lagrangian multipliers $\{z_i\}$ and $\{W_l\}$, the p_{ij} values are calculated according to

Table 1: Cancer datasets: median values of the network entropy groups S_1 and S_2 as pictured in Fig. 2. With p_W we consider the p-value given by the Wilcoxon rank sum test.

	S_1	S_2	Size	p_W
Colon	13.0349	13.0680	312	$4.35 \cdot 10^{-4}$
Ewing	8.8057	8.7569	136	0.0023
Met	9.9483	9.9159	151	$4.12 \cdot 10^{-4}$
Rel	15.4700	15.4664	313	0.0197

Table 2: Ageing dataset: in the upper part of the table we show the median values for the five network entropy age groups as pictured in Fig. 3. With p_K we consider the p-value given by the Kruskal-Wallis test over the five age groups. In the lower table we show the results for the Wilcoxon rank sum test for each pair of groups.

S_1	S_2	S_3	S_4	S_5	p_K
11.4249	11.4331	11.4497	11.4495	11.3972	0.0028

p_W	group 1	group 2	group 3	group 4	group 5
group 1		0.5476	0.0952	0.0079	0.0079
group 2			0.4206	0.1508	0.0079
group 3				1	0.0079
group 4					0.0079

Eq. (5). These values are then substituted in the constraint equations (3) and (4) for the next calculation of the lagrangian multipliers, and the process is repeated upon convergence. We checked by random sampling that the application of the iterative algorithm for different initial guesses leads to the same final entropy values (since under these constraints it is a convex function that admits an unique maximum). In our analyses, the algorithm convergence threshold was set to 10^{-5}, and we remark that every significant change in entropy values was at least two orders of magnitude higher, thus the chosen precision is not affecting our results. This algorithm is available in Matlab and Python code (availabe as Supplementary Material).

3 Results

3.1 Cancer datasets

The first analysis consisted in comparing the entropy values for the samples belonging to the different classes (see Figure 2). For the Colon dataset (Fig. 2, Panel *a*) we see a significant increase of network entropy S between normal and cancer samples (P = 0.00043) when considering the selection of genes which expression profile differed between normal and cancer samples. At a full-network level, the same trend is observed, but the result is weakly non significant (P = 0.057). We interpret this result as an increase in cell deregulation when passing from normal to cancer cell, reflected in a higher "phenotypic space" available, since many regulation mechanisms (e.g. related to cell cycle, apoptosis or DNA

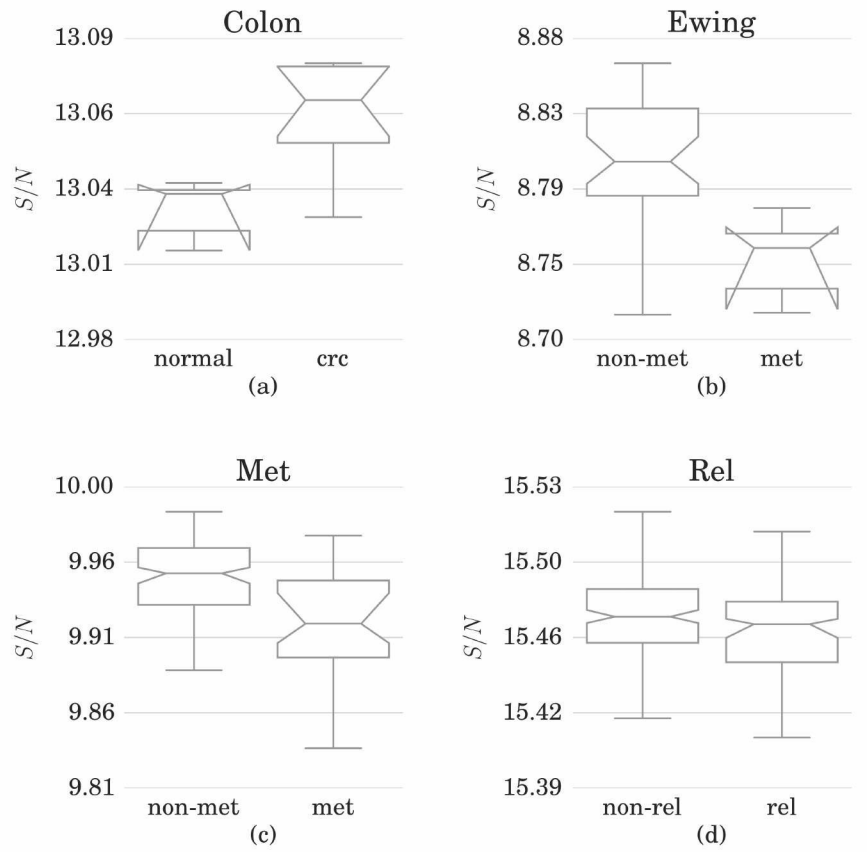


Figure 2: Boxplots for the network entropy values in the studied cancer databases. Panel a: colon cancer, normal vs. cancer samples. In this case cancer samples have a significantly higher entropy. Figure b, c, d: Ewing sarcoma, metastatic and relapsing breast cancer databases, respectively. In b, c, d cases primary tumour samples are compared with tumour samples that relapsed or developed metastasis during disease progression. In these cases entropy has a significantly higher value in the primary tumour groups.

repair) are lost in a cancer cell [22].

If we consider single-node entropy, we find 665 genes (over 2835) with a significant difference between normal and cancer samples (see Supplementary Table 2). The single genes with highest significance are involved in known cancer-related pathways, such as “Wnt", “Mapk", “Notch" and “Cell communication" pathways. The role of the genes which single-node entropy is differing significantly between normal and cancer samples can be better understood at a KEGG pathway level: a functional enrichment analysis based on the hypergeometric distribution (i.e. counting the number of genes with significant differences in single-node entropy for a particular pathway, given the total number of significant variations in the whole network) shows that 25 (over 191) pathways are

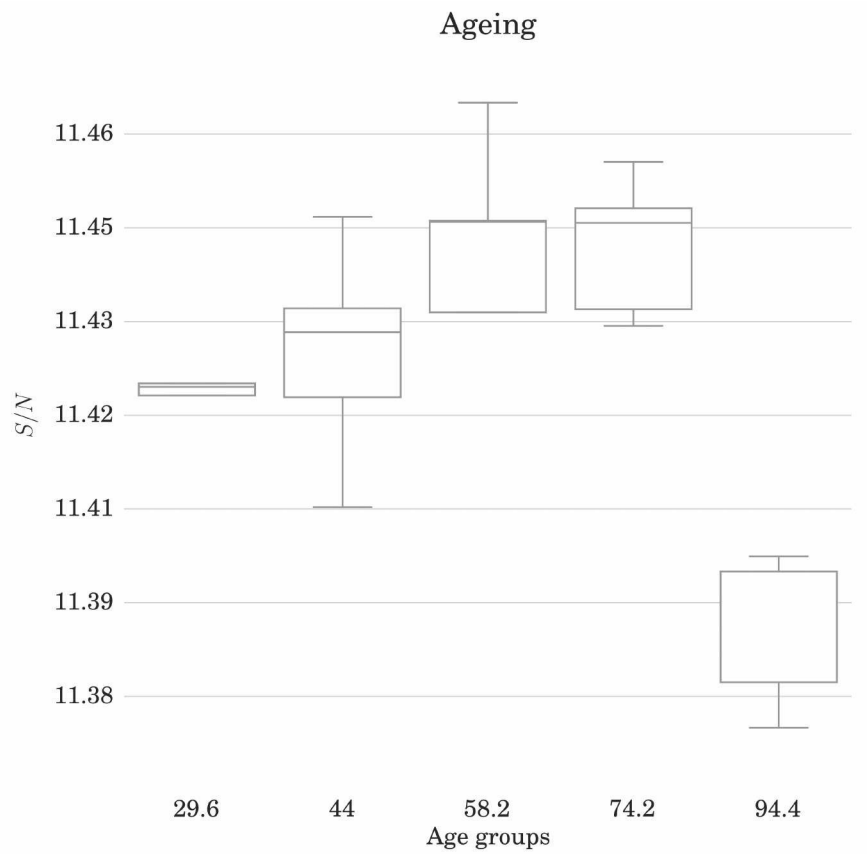


Figure 3: Boxplots for the network entropy values in the studied ageing database. Successfully aged people have a significantly lower entropy. We show the results for the Kruskal-Wallis test over the five age groups and for the Wilcoxon rank sum test for each pair of groups in Tab. 2.

significantly enriched ($P < 0.05$), among which "Oxidative phosphorylation", "Focal adhesion", "TCA cycle", "Cell communication", "Apoptosis", "Cell adhesion molecules" with a clear involvement in cancer progression both at signalling and metabolic level [22].

In the comparison between primary and secondary cancers (metastatic or relapsing) we find instead a significant decrease in network entropy (P = 0.014 for the Ewing dataset, P=0.00041 for MET dataset, P=0.02 for the REL dataset). In this case, the change from a primary cancer to a metastatic or relapsing state implies an evolutionary selection, since some specific steps need to occur, e.g. regarding epythelial-to-mesenchymal transition mechanisms [23] or adaptation to pharmacoresistance [24], or clonal selection induced by therapy [25]. Network Entropy seems a reliable measurement to quantify the reduction in phenotypic space as a result of an underlying evolutionary process. In the Ewing dataset, 142 genes have a significant difference in single-node entropy S_i ($P < 0.05$) be-

Table 3: Pathway analysis: number of significant genes and pathways based on the single node entropy variations. For the genes we applied a Wilcoxon rank sum test in the usual case-control setup. For the pathways we performed an enrichment analysis, highlighting those paths enriched by genes significantly changing their single-node entropy value.

	Significant genes	Significant pathways
Colon	665	25
Ewing	142	35
Met	342	48
Rel	331	23
Ageing	290	16

tween primary and metastatic samples, involved in many pathways, most related to lipid metabolism. A significance analysis at KEGG pathway level produces 35 significantly enriched pathways, such as "Glycolysis/Gluconeogenesis", "Pentose phosphate", "Galactose metabolism", "Glycosphingolipid biosynthesis", involved in energy metabolism, and also "Cell communication", "Focal adhesion" and "ECM-receptor interaction" that might be involved in metastatic processes such as cell migration. In the MET dataset, 342 genes have a significant difference in single-node entropy. Even if the cell type is different (primary breast cancer) many pathways are the same as for the Ewing dataset, in particular related to the lipid metabolism. Functional analysis highlights 48 enriched pathways, among which "Glycolysis/Gluconeogenesis", "Galactose metabolism", "Glycosphingolipid biosynthesis" as for Ewing dataset, and also pathways such as "Cell adhesion molecules" that can be again related to metastatic progression. For the REL dataset, 331 genes had a significant difference in single-node entropy, and 23 pathways were functionally enriched with a $P < 0.05$. Among these pathways, some of them are related to metabolism ("Ether lipid biosynthesis", "Biosynthesis of steroids", "Pyrimidine metabolism"), but also to specific functions such as "RNA polymerase", "DNA polymerase", "Proteasome", "Cell adhesion molecules" and "Metabolism of xenobiotics by cytochrome P450".

A comparison between significant pathways, obtained by enrichment analysis of genes significantly different on single-node entropy values or on gene expression values, are shown in Supplementary Table 3.

3.2 Ageing dataset

For the Ageing dataset, we exploited the time series design by applying a Kruskal-Wallis test over the age groups, in order to evaluate significant changes in network entropy over the whole life span. The trend for the five groups was significantly different (P=0.0028, see Fig. 3). In particular, among the 5 age groups, Wilcoxon testing revealed that only the oldest age group showed a significantly different behaviour, with a lower Network Entropy than the other age groups (see Table 2). The last age group is related to successfully ageing people, since their age is larger than average life expectancy, thus it represents a very selected group from an epidemiologic point of view. Its different value in network entropy could be explained in two ways, that our data do not allow to distinguish: first, the successfully ageing group represents a selection, in terms

of phenotype, over the human population. Thus the reduced entropy highlights their peculiar expression profile. As a second hypothesis, the oldest group shows a smaller plasticity in terms of the possible phenotypic profiles that the cells can assume. This aspect can be related to the "frail" phenotype [26, 27], for which old people are less capable of adaptation, both from a physiological and physical point of view. For the single-node entropy and functional enrichment analysis we considered a comparison between the youngest ("A") and the oldest ("E") age group, representing the two extremes of our time series: Wilcoxon test found 290 (over 1976) genes with a significant difference in S_i ($P < 0.05$). The KEGG pathways mostly enriched by significant genes are in part related to the specific cell type, i.e. lymphocytes ("T cell receptor signalling", "B cell receptor signalling", "hematopoietic cell lineage"), metabolic pathways ("Androgen and estrogen metabolism", "Biotin metabolism", "Histidine metabolism"), and pathways involved in cellular degradation/production machinery ("Proteasome"), in particular at the nucleolar level, such as "Ribosome" and "DNA polymerase" that are known to be altered during ageing [28, 29]. We remark that the pathways involved in a change in entropy, as shown above, are very different from the pathways obtained by an identical functional analysis performed on changes in gene expression, reflecting the different information encoded in network entropy at whole-cell and single-node level.

All the tables, with P values for single nodes and for KEGG pathways, are included as supplementary material (see Supplementary Table 2 and Supplementary Table 3). Moreover, the relations between significant pathways have been displayed as networks, with significant pathways linked by shared significant genes (shown in Supplementary File).

4 Conclusion

We introduce a measure of network entropy, based on a rigorous statistical mechanics definition, that integrates the topological information encoded in the protein interaction network with gene expression profiling. This measure is introduced to characterize different levels of cellular perturbation, namely the comparison between healthy and cancer samples, primary and metastatic cancer samples, and a time series of healthy samples with different ages across the whole human lifespan. This measure estimates the number of networks that satisfy given constraints, based on PPI network and gene expression profiles, and can be interpreted as the extent of the "parameter space" allowed to the cell in a given state in terms of gene expression plasticity, or also in terms of different cell phenotypes (e.g. cell clonality for the case of cancer).

Different case studies help to clarify this interpretation. Regarding the comparison between healthy and cancer cells, we observe an increase of network entropy, possibly due to a larger deregulation of the biological mechanisms and functions involved, and to an increase in cell phenotypical diversity. On the contrary, when we consider primary vs. metastatic (or relapsing) samples, network entropy shows a significant decrease, reflecting the canalization or the evolution (in terms of clonal extent or gene expression profile) necessary to achieve this specific state. In a time series of ageing people, we see a sharp decrease of network entropy for the successful ageing group (with an age larger than typical life expectancy) that could also in this case represent a sort of selection of specific

ageing phenotypes.

The formalism allows to define a measure of entropy at different scales, from single gene to biological pathways (as obtained from KEGG database), that highlight how the changes in entropy are specific for the biological function and the experimental design considered.

This method provides a different perspective on the analysis of gene expression data, integrating single-gene expression measurements and functional relationships between genes due to biological functions inside the cell. The entropy measure S seems sensitive enough to evaluate the effect of physiological perturbations, such as those occurring during the cellular ageing process, and also the differences between cancer subtypes before the progression to metastatic and relapsing phenotypes. The statistical significance of S resulted independent on network properties, such as the number of nodes, and increased when a selected subset was considered, thus reflecting the biological relevance of the data used.

Finally, we remark that this approach can be generalized to other systems as well, considering 1) different networks for the topological constraints, like transcription or metabolic networks, 2) different high-throughput measurements, for example methylation states or metabolite concentrations, and 3) different metrics to define the weights of the network, like correlation or mutual information, allowing to adapt this formalism to the specific experimental design and biological context.

Acknowledgement

DR, EG, GC and GM were supported by EU project MIMOmics n. 305280 and Italian CNR Flagship project Interomics.

References

[1] M. Pagel and A. Pomiankowski, *Evolutionary genomics and proteomics*. Sinauer Associates, 2008.

[2] J. De Las Rivas and C. Fontanillo, "ProteinÐprotein interactions essentials: Key concepts to building and analyzing interactome networks," *PLoS Computational Biology*, vol. 6, no. 6, p. e1000807, 2010.

[3] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, 02 2004.

[4] E. Alm and A. Arkin, "Biological networks," *Curr Opin Struct Biol*, vol. 13, no. 2, pp. 193–202, 2003.

[5] M. Vidal, M. Cusick, and A. Barabasi, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 989–98, 2011.

[6] E. Cerami, B. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic Acids Res*, vol. 39, pp. D685–D690, 2011.

[7] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. Jensen, and C. von Mering, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucl. Acids Res.*, vol. 39, no. S1, pp. D561–D568, 2011.

[8] A. Teschendorff and S. Severini, "Increased entropy of signal transduction in the cancer metastasis phenotype," *BMC Systems Biology*, vol. 4, no. 1, p. 104, 2010.

[9] F. Barea and D. Bonatto, "Aging defined by a chronologic-replicative protein network in saccharomyces cerevisiae: an interactome analysis," *Mech Ageing Dev*, vol. 130, no. 7, pp. 444–60, 2009.

[10] J. West, G. Bianconi, S. Severini, and A. Teschendorff, "Differential network entropy reveals cancer system hallmarks," *Scientific Reports*, no. 802, p. 10.1038/srep00802, 2012.

[11] W. N. van Wieringen and A. W. van der Vaart, "Statistical analysis of the cancer cell's molecular entropy using high-throughput data," *Bioinformatics*, vol. 27, no. 4, pp. 556–563, 2011.

[12] K. Anand and G. Bianconi, "Gibbs entropy of network ensembles by cavity methods," *Phys. Rev. E*, vol. 82, p. 011116, Jul 2010.

[13] G. Bianconi, P. Pin, and M. Marsili, "Assessing the relevance of node features for network structure," *PNAS*, vol. 106, p. 11433Ð11438, 2009.

[14] B. Gyorffy, B. Molnar, H. Lage, Z. Szallasi, and A. Ecklund, "Evaluation of microarray preprocessing algorithms based on concordance with rt-pcr in clinical samples," *PLoS One*, vol. 4, p. e5645, 2009.

[15] K. Scotlandi, D. Remondini, G. Castellani, M. Manara, F. Nardi, L. Cantiani, M. Francesconi, M. Mercuri, A. Caccuri, M. Serra, M. Knuutila, and P. Picci, "Overcoming resistance to conventional drugs in ewing sarcoma and identification of molecular predictors of outcome," *J. Clin. Oncol.*, vol. 27, pp. 2209–16, 2009.

[16] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 15 February 2006.

[17] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. A. Foekens, J. G. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart, and C. Sotiriou, "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade," *Journal of Clinical Oncology*, vol. 25, no. 10, pp. 1239–1246, April 1, 2007.

[18] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[19] D. Remondini, S. Salvioli, M. Francesconi, M. Pierini, D. J. Mazzatti, J. R. Powell, I. Zironi, F. Bersani, G. Castellani, and C. Franceschi, "Complex patterns of gene expression in human t cells during in vivo aging," *Mol. BioSyst.*, vol. 6, pp. 1983–1992, 2010.

[20] G. Bianconi, "Entropy of network ensembles," *Phys. Rev. E*, vol. 79, p. 036114, Mar. 2009.

[21] K. Anand and G. Bianconi, "Entropy measures for networks: Toward an information theory of complex topologies," *Phys. Rev. E*, vol. 80, p. 045102, Oct. 2009.

[22] D. Hanahan and R. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 2000.

[23] S. Brabletz and T. Brabletz, "The zeb/mir-200 feedback loopÑa motor of cellular plasticity in development and cancer?," *EMBO Reports*, vol. 11, no. 9, pp. 670–677, 2010.

[24] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, "Cancer drug resistance: an evolving paradigm," *Nat Rev Cancer*, vol. 13, pp. 714–726, 10 2013.

[25] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, pp. 306–313, 2012.

[26] L. Ferrucci, F. Giallauria, and D. Schlessinger, "Mapping the road to resilience: Novel math for the study of frailty," *Mechanisms of Ageing and Development*, vol. 129, pp. 677–679, 2008.

[27] L. Fried, L. Ferrucci, J. Darer, J. Williamson, , and G. Anderson, "Untangling the concepts of disability, frailty, and comorbidity: Implications for improved targeting and care," *Journal of Gerontology*, vol. 59, no. 3, pp. 255–263, 2004.

[28] E. Bellavista, M. Martucci, F. Vasuri, A. Santoro, M. Mishto, A. Kloss, E. Capizzi, A. Degiovanni, C. Lanzarini, D. Remondini, A. Dazzi, S. Pellegrini, M. Cescon, M. Capri, S. Salvioli, A. D'Errico-Grigionic, B. Dahlmannd, G. Grazi, and C. Franceschi, "Lifelong maintenance of composition, function and cellular/ subcellular distribution of proteasomes in human liver," *Mechanisms of Ageing and Development*, vol. 141-142, pp. 26–34, 2014.

[29] H. Lempiainen and D. Shore, "Growth control and ribosome biogenesis," *Current Opinion in Cell Biology*, vol. 21, p. 855Ð863, 2009.