IL NUOVO CIMENTO Vol. 37 C, N. 4 Luglio-Agosto 2014
DOI 10.1393/ncc/i2014-11804-5

Communications: SIF Congress 2013

Towards a network entropy formalism for gene expression data analysis

D. Remondini(*)

Dipartimento di Fisica e Astronomia DIFA, Università di Bologna - Bologna, Italy
INFN, Sezione di Bologna - Bologna, Italy

ricevuto il 31 Gennaio 2014; approvato il 23 Aprile 2014

Summary. — In many situations it has been widely recognized that studies based on single-gene differential expression statistical analyses are too simplistic, since they do not consider the complexity of the underlying biological system, mainly based on specific interaction between genes (at a transciptomic, proteomic and metabolomic level). For this reason, novel approaches are sought aiming to exploit network-based methods, in a Systems Biology perspective, capable of integrating single-probe measurements with biological information at a whole-genome scale. We describe a method, based on Statistical Mechanics and Network Theory, that goes into this direction, combining what is actually known about gene-gene interactions at a protein level and high-throughput mRNA data obtained in different experimental conditions. We will provide a framework for a phenomenological interpretation of Entropy of a Network Ensemble based on an Information Theory approach.

PACS 05.10.-a – Computational methods in statistical physics and nonlinear dynamics.
PACS 87.18.Vf – Systems Biology.
PACS 87.18.Vd – Genomics.

1. – Introduction

In the last 15 years, an enormous revolution has occurred in the medical and biological fields. New experimental techniques have allowed to collect information at a very fine scale (for example at the level of single genes or exons, single proteins, and even single DNA mutations) in a rigorously quantitative manner. These data have been made available to the whole scientific community thanks to open-access databases (see for example www.ncbi.nlm.nih.gov/geo/ and www.ebi.ac.uk/arrayexpress/ for transcriptomics and epigenomics data). Moreover, thanks to the availability of these

(*) E-mail: daniel.remondini@unibo.it

techniques, large-scale experiments have been performed allowing to build databases of relations between elementary constituents of the cells (like genes, proteins, metabolites: `www.broadinstitute.org/cmap/` for protein-protein interactions, or `humanmetabolism.org/` for the human metabolic reaction network).

This huge amount of data and information have become amenable of novel analysis and modeling approaches, more common to biophysicist rather than to a biologist, encompassing, among the others, advanced data analysis techniques, stochastic modeling and statistical mechanics. These approaches can be gathered under the big hat of Systems Biology, a highly interdisciplinary field of study that requires a strong interaction between biologists and physicists, also at the experimental design level, in order to gain new perspectives on classical biological and medical problems (*e.g.* understanding of complex biological processes by means of mathematical models [1], but also design of new therapeutic protocols based on such models [2]).

In Physics, since the second half of the 19th century, systems composed of many (interacting or not) elements have been described by the Statistical Mechanics formalism. Even if single elements have their own degrees of freedom, their collective behaviour can be analyzed with a probabilistic approach, and in many cases it can be shown that the larger the system, the smaller the deviation of its single elements from this "average" behaviour. On the other hand, Biology has always been the realm of extreme uncertainty, for several reasons: the difficulty to control all the experimental conditions with sufficient detail, the individual variability of the samples studied (being them single cells or whole organisms), but also for the lack of experimental techniques capable of a reliable quantitative measure over a large number of probes. Nowadays these measurements are available, so it has come the time to apply these physical approaches to these problems. In particular, since the interactions between biological elements is one of their characterizing features, the network formalism has been widely applied in this context [3-5]. A very recent synthesis between Statistical Mechanics and Network Theory has led to the definition of a measure of Entropy for Network ensembles [6], that is the basis of our approach described in the next section.

2. – Statistical Mechanics of network ensembles

The basis of a Statistical Mechanics approach to physical systems is a description of the interactions between its elements: this is provided by the Hamiltonian of the system, that describes how the energy is distributed in the system. The study of physical systems historically starts with a phenomenological approach, formalized in the laws of Thermodynamics and all the relations between thermodynamic variables, and only in a second time it is rigorously embedded in the theoretical approach provided by Statistical Mechanics. This to say that many theoretical results could be easily translated into useful observables and predictions (as in the case of heat capacity for ideal mono- and bi-atomic gases) since a great phenomenological work was preceding the theoretical achievements, and mathematical constraints such as constant energy (for the Microcanonical ensemble) or constant temperature (for the Canonical ensemble) could be easily understood for a physical system. Given a system with many elements, even if each single element must obey deterministic laws (stated by the Hamiltonian), the incomplete knowledge of each single element state leads to a probabilistic approach, in which every system state has a certain probability to be realized. To describe this behaviour, the Partition Function is introduced, describing the probability of the systems to be in a certain state given the

temperature T. When the system is in thermal equilibrium this reads

$$p(x,v) = \frac{e^{-\beta H(x,v)}}{Z}\,; \qquad Z = \int e^{-\beta H(x,v)}\,\mathrm{d}x\,\mathrm{d}v. \tag{1}$$

It can be demonstrated that in a situation of constant energy (or more realistically of constant temperature) the systems evolves towards a state of maximum entropy, a thermodynamic function defined as

$$S = -\int p(x)\log p(x)\mathrm{d}x. \tag{2}$$

This is not the case for Network Theory, since in many cases a clear Hamiltonian function is not available to describe the system. A network describes the structure of the interaction between elements, by means of an adjacency matrix A that takes into account the existence of a *link* a_{ij} between two *nodes* labeled as i and j. Our approach is better interpreted from an Information Theory point of view [7], that is based on an exact counting of the possible states of the "system" given specific constraints, and thus starts from the definition of the Entropy rather than from the Hamiltonian of the system. Also in this case Entropy maximization characterizes the system, and the Hamiltonian function (with its physical meaning) is substituted by one or more constraints on the variables describing the system (*e.g.* with respect to some statistical moments of the distribution of the system variables).

The combination of entropy S and the constraints on it leads to the definition of a free energy function F', here written in comparison with the thermodynamic free energy function F:

$$F = E - TS, \tag{3}$$

$$F' = -S + \frac{E'}{T'}\,. \tag{4}$$

In the first equation we have the classical Helmoltz free energy, while in the second we have emphasized([1]) the dependence of the function on entropy, and (as we will see in following examples) E' terms represent the constraints on S and $1/T'$ is the related Lagrange multiplier.

The point of applying this approach to networks derives from the following assumption: each real network (*e.g.* given by the hyperlinks in the WWW, by people having friendship relations or exchanging emails, or by genes and proteins interacting inside a biochemical process) is a *realization* (the most probable realization) of a general ensemble of network satisfying specific constraints. Thus, starting from a real instance of a network ensemble, the mathematical formulation of the constraints that define such ensemble is the foundation onto which the Information/Statistical approach is built. The problem is that there is not a clear phenomenology of networks, since they can be related to very different systems (people, proteins, stocks, web pages, and so on) and a simple "physical" interpretation of its observables is not available. After introducing the exact formalism

([1]) Note that F' has not the dimension of an energy, since the variables do not have the same phenomenological meaning as in F.

of entropy of network ensembles, we will consider a set of examples that will lead to the definition of our specific problem.

3. – Network Entropy: definition and application to transcriptomics

Contrarily to Statistical Mechanics, in which the Hamiltonian of the system and the physical observables of the system lead "naturally" to a definition of physical state and probability of a physical state [8], different starting points can be considered for defining the probability of having a specific network. As described in [6], we will characterize the probability of having a particular instance of a network as the joint probability of having all the single links composing the network:

$$p(A) = \prod p_{ij}, \qquad p_{ij} = p(a_{ij} = 1). \tag{5}$$

In a canonical approach, similar to the one described previously, Entropy for a network ensemble is calculated over the possible instances of the network, that are related to the probabilities of having (or not having) each single link in each possible network. If (as will be our case) we consider only symmetric([2]) unweighted networks (so that for each link we can have only two probabilities, $P_{ij} = p(a_{ij} = 1)$ and $p(a_{ij} = 0) = 1 - P_{ij}$), the probability for eaving each graph can be calculated, and also the network entropy accordingly:

$$P(A) = \prod_{i<j} p_{ij}^{a_{ij}} (1 - p_{ij})^{|1-a_{ij}|}, \tag{6}$$

$$S = -\sum_{i<j} \left(p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij}) \right). \tag{7}$$

At difference with Gibbs entropy, this entropy is a bivariate function (indexes i and j for each couple of nodes), and takes directly into account the presence as well as the absence of a link. The claim is that the real instance of the network observed is the one maximizing the entropy of the ensemble: it can be demonstrated that if entropy function has a set of constraints that is linear in its variables (the p_{ij}) it can have only one maximum [9], that is found by derivation (or numerically in case there is no simple analytical solution):

$$S_{MAX} \to \frac{\mathrm{d}S}{\mathrm{d}p_{ij}} = 0. \tag{8}$$

The different values of entropy will depend on network size but, most importantly, on the constraints imposed on it, derived from the information we want to include in our network model. Also in this case there is a substantial difference between classical Statistical Mechanics and Networks: in the former case, the constraints are typically intensive (*i.e.* independent on the number of elements) while for networks many relevant observables (*e.g.* connectivity degree, defined as $k_i = \sum_j a_{ij}$) are defined at a single-node level.

([2]) That restrict the analysis to the upper triangular part of A.

Since one important feature of a network is its topology (often summarized by the distribution of connectivity degrees $p(k)$) it can be taken into account in the constraints for the computation of maximum entropy by imposing the full degree sequence of the network (*i.e.* the series of all connectivity degrees $k_i \forall i = 1, \ldots, N$). In this case the "free energy" (entropy + constraints) becomes

$$S = -\sum_{i<j} (p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij})) + \sum_{i}^{N} \lambda_i \sum_{j}^{N} p_{ij}, \tag{9}$$

in which the λ_i are the Lagrange multipliers for the N constraints (analogous to the role of $\beta = 1/KT$ for the classical canonical ensemble). This formalism allows to calculate the entropy for a nework ensemble with a fixed number N of nodes and a fixed degree sequence $\{k_i\}$.

In order to apply this approach to real data, we want to include in the definition of the ensemble some specific features on the nodes derived by real experimental observations (that thus would distinguish one sample from the other, even if the topology of the network remains the same).

In the application to gene expression data, we consider the genes as nodes, and the gene expression values (measured for example by high-throughput microarray techniques) as values associated to the nodes specific for a particular realization (*e.g.* the expression profile of a cell under a specific condition, such as healthy, cancer, irradiated, treated with chemicals, etc.). Defining a metrics (*e.g.* Euclidean) we can calculate the distance between nodes i, j in terms of their expression level g_i, g_j: $d_{ij} = \sqrt{(g_i - g_j)^2}$, and this values can be interpreted as *weights* on the link between nodes i, j. Also this information can be embedded as a constraint in the calculation of the maximum entropy of the network ensemble, and we can summarize our approach as follows:

- each sample (a pool of cells from a specific tissue or organ of the human organism([3])) is represented by a *network ensemble*;
- in this network ensemble genes are *nodes*, and *links* are given by protein-protein interactions (PPI, as annotated in Connectivity Map, a repository from Broad Institute that collects information on protein-protein interactions from multiple datasets, *www.broadinstitute.org/cmap/*);
- PPI topology (equal for alle the samples) is imposed on the network ensemble by a constraint on the connectivity degree (in our case $N = 10000$ nodes for the full-genome network);
- information about gene expression profile (specific for each sample) is imposed on the network ensemble by a constraint on the number of links that have a weight in a specific interval (binning of the empirical distance distribution)

The "free energy" can thus be written as follows, with an additional constraint for the

([3]) But it could be applied to any organism for which the same information are available.

TABLE I. – *Entropy values for the two groups and P value for the statistical test.*

Probe set	Entropy group 1	Entropy group 2	P-value
Whole-genome	97577	97565	0.06
5% probe selection	19.02	19.00	0.0028

weighted link distribution with respect to the previous definition:

$$S = -\sum_{i<j} \left(p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij})\right) + \sum_{i}^{N} \lambda_i \sum_{j} p_{ij} + \sum_{N} b\gamma_i \left(B_i - \sum_{i<j}^{N} \chi_l(d_{ij}) p_{ij} \right).$$

Maximization of this function produces a maximum entropy value for each sample analyzed, that integrates information about single gene expression profile and interaction between nodes that are relevant from a biological point of view. Since it is not possible to obtain analytic results, the function is maximized by numerical implementation of an iterative algorithm (as described in [10]).

As an example for the application of this formalism, we consider a dataset of breast cancer samples (publicly available in th GEO Omnibus repository with the GEO accession number GSE2990) obtained from about 130 patients: 97 samples are primary tumours, and 28 are tissues from primary tumours that subsequently developed metastasis. We performed the analysis of entropy variations in the two groups with two different set of genes: the first with the whole genome profile available (corresponding to about 10000 genes mapped onto the PPI) and the second over a selection of about 400 probes whose expression level was different between the two groups (calculated by a Student's T-test with a 5% significance threshold).

As it can be seen in table I, the two groups have a different distribution of entropy values, with a lower value for the metastatic group. This result is similar both at a whole-genome level and in the significant gene selection, with an increase of statistical significance for the gene selection subset. We remark that this result cannot be obtained by an analysis of single probes, and it is not related to simple statistical parameters of the gene profile distribution such as mean or variance for each group (data not shown). The question is how to gain some biological insight from this result: since entropy gives a measure of the number of networks that belong to a certain ensemble (*i.e.* that satisfy the given constraints), we can think that in this example entropy measures the number of possible cell states available to the cell, given a specific gene expression profile and a network of interactions between genes. With this hypothesis in mind, we can try to interpret the lower value of entropy in the second group as follows: cancer cells that will become metastatic must undergo specific transformations, as a sort of "natural selection" from an evolutionary point of view, thus their expression profile must be more controlled, and more bounded (in terms of the possible values that can be assumed the genes). A lower entropy represents a smaller volume of "phenotypic space" available to the cell, with the cell phenotype defined as the set of observable characteristics (morphological, biochemical and functional) related to its genotype (our observables, the set of gene expression values) and to the environment (not considered in this analysis).

4. – Conclusions

We describe a framework for the application of Statistical Mechanics concepts to Network Theory. In particular, we can define a measure of entropy for network ensembles that satisfy some constraints, related to network topology and to features associated to the links. This measure has been applied to high-throughput transcriptomics data, allowing to calculate a value of entropy for each sample. An interpretation of this entropy values, in terms of the biological application, is under study, since a phenomenological interpretation of entropy is not as simple as in classical Statistical Mechanics, in which the physical meaning of the thermodynamics variables was already known before the rigorous formalization operated by the work of Gibbs and Boltzmann. The first results seem to point to a measure of the "parameter space" available to the cell, with implications for its phenotype given the genomic profile that characterizes its state.

* * *

The author acknowledges INFN Gruppo IV PIECES, and EU MIMOMICS 305280 Projects.

REFERENCES

[1] Remondini D. *et al.*, *Physica A*, **392** (2013) 336.
[2] Lamb J. *et al.*, *Science*, **313(5795)** (2006) 1929.
[3] Tieri P. *et al.*, *Bioinformatics*, **21** (2005) 1639.
[4] Remondini D. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **102(19)** (2005) 6902.
[5] Ravasz E. *et al.*, *Science*, **297** (2002) 1551.
[6] Bianconi G., *Phys. Rev. E*, **79** (2009) 036114.
[7] Cover T. M.and Thomas J. A., *Elements of Information Theory* (Wiley) 2006.
[8] Huang K., *Statistical Mechanics* (Wiley) 1987.
[9] Pressé S. *et al.*, *Rev. Mod. Phys.*, **85** (2013) 1115.
[10] Menichetti G., *Statistical mechanics approaches to networks: the role of entropy with applications to gene expression time series data*, Master Thesis, Bologna University (2010).