Information-theoretical analysis of gene expression data to infer transcriptional interactions

K. Baca-López[a,b], E. Hernández-Lemus[a,c], and M. Mayorga[b]
[a]*Departamento de Genómica Computacional, Instituto Nacional de Medicina Genómica, Periférico Sur No. 4124, Torre Zafiro 2, Piso 6 Col. Ex Rancho de Anzaldo, Álvaro Obregón 01900, México, D.F., México,*
[b]*Facultad de Ciencias, Universidad Autónoma del Estado de México, Av. Instituto Literario 100 Ote. Centro 50000, Toluca, Estado de México, México*
[c]*Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Torre de Ingeniería, Piso 6, Circuito Escolar s/n Ciudad Universitaria, Coyoacán, 04510, México, D.F., México.*

Recibido el 4 de agosto de 2009; aceptado el 6 de octubre de 2009

The majority of human diseases are related with the dynamic interaction of many genes and their products as well as environmental constraints. Cancer (and breast cancer in particular) is a paradigmatic example of such complex behavior. Since gene regulation is a non-equilibrium process, the inference and analysis of such phenomena could be done following the tenets of non-equilibrium physics. The traditional *programme* in statistical mechanics consists in inferring the joint probability distribution for either microscopic states (equilibrium) or mesoscopic-states (non-equilibrium), given a model for the *particle* interactions (*e.g.* the potentials). An *inverse problem* in statistical mechanics, in the other hand, is based on considering a *realization* of the probability distribution of micro- or meso-states and used it to infer the interaction potentials between particles. This is the approach taken in what follows. We analyzed 261 whole-genome gene expression experiments in breast cancer patients, and by means of an information-theoretical analysis, we deconvolute the associated set of transcriptional interactions, *i.e.* we discover a set of fundamental biochemical reactions related to this pathology. By doing this, we showed how to apply the tools of non-linear statistical physics to generate hypothesis to be tested on clinical and biochemical settings in relation to cancer phenomenology.

Keywords: Cancer genomics; information theory; molecular networks.

La mayoría de las enfermedades humanas están relacionadas con la interacción de muchos genes, y con condicionantes ambientales, lo que las hace fenómenos complejos. El análisis de las interacciones bioquímicas relacionadas se basa frecuentemente en la consideración de las relaciones de regulación genética. Puesto que la regulación genética es un proceso fuera del equilibrio, la inferencia y el análisis de ésta puede hacerse siguiendo los principios de la termodinámica irreversible y la mecánica estadística fuera del equilibrio. El enfoque tradicional de la mecánica estadística es inferir la distribución de probabilidad conjunta para los estados del sistema en términos de un modelo para las interacciones. Un problema inverso en mecánica estadística consiste en considerar una realización de la distribución de probabilidad y emplearla para inferir las interacciones entre las partículas. Tomamos este enfoque para analizar 261 experimentos de expresión de mRNA de genoma completo, en pacientes con cáncer de mama y, a través de una medida basada en la teoría de la información descubrir el conjunto de interacciones transcripcionales asociadas. Mostramos cómo aplicar las herramientas de la física estadística no-lineal para generar hipótesis (es decir, el conjunto de interacciones inferidas) que pueden ser probadas en ensayos clínicos y bioquímicos con relación a la fenomenología del cáncer.

Descriptores: Genómica del cáncer; teoría de la información; redes moleculares.

PACS: 87.10.Vg; 87.16.Yc; 87.18.Cf; 89.75.Hc; 89.70.Cf

1. Introduction

The forms and functions of living cells, also called *cellular phenotypes* are known to be determined by the interplay of many genes and their products such as proteins, enzymes and so on. Given this fact, the identification of rules of behavior at the genome-wide level is essential to elucidate both normal cell function and pathological phenotypic conditions at whole-system scale. A usual tool to track down this phenotypic diversity is gene expression analysis. Since the process of gene expression by itself is often regulated by different genes and their products, statistical associations abound between genetic transcripts abundance (*e.g.* messenger RNA segments [mRNAs]). These associations could be behind the mechanisms of cell function. However, one hard-to-grasp issue is that the process of gene expression by itself is a complex one, both from the biochemical and thermodynamical points of view [1, 2]

The complex description given by the set of interactions consists, generally, in identifying gene correlations from experimental data through the use of theoretical models and computational analysis. The discovery of such an interaction's set involves the solution of an inverse problem (a deconvolution) that basically tries to uncover the interactions from the properties and dynamics of observable behavior in the form of, for example, RNA transcription levels in a characteristic gene expression profile.

Genome-wide transcriptional profiling, also called Gene Expression Analysis (GEA) has permit us to go far ahead of

studying gene expression at the individual-level, by providing global information about functional inter-relations between genes, mRNAs and the associated regulatory proteins. GEA have increased our understanding of the dynamics and interplay between different processes involved in gene regulation and have pointed out to previously unappreciated biological functional relations, such as the coupling between nuclear and cytoplasmic transcription and metabolic processes [2]. GEA also revealed extensive communication within regulatory units, for example in the organization of transcription factors into regulatory motifs. It is these kinds of regulatory interactions what one is ultimately looking for in GEAs.

Several stages are involved in the regulation of gene expression-transcription: such as mRNA processing, nuclear export, translation and degradation. These steps were usually analyzed *in isolation* by using conventional biochemical techniques like the PCR reaction and electrophoresis [1]. This point of view has left the impression that such stages are independent. In the past molecular biology research was focused on the mechanisms underlying individual gene expression or in the best scenario the behavior of a small set of genes, rather than exploring regulatory mechanisms that can influence many genes at one time.

Systematic studies of genome-wide binding patterns made evident the existence of a great deal of coordinate regulation among transcription factors (*i.e.* genes that catalyze or inhibit the expression of other genes, either by themselves or by means of their produced proteins). Factors that combinatorialy regulate (on a concomitant way) a particular gene also often coordinately regulate the expression of other genes, potentially even themselves or each other. For an introduction to gene expression phenomena and transcriptional regulation from a physicochemical point of view in Ref 3.

1.1. Cancer

Cancer is the random, uncontrolled and accelerated proliferation of cells due to genetic abnormality. This genetic abnormality usually starts with sets of mutated genes that either suppress or stimulate the cell's cycle continuity.

Normally, cells grow and divide to form new cells as the body needs them. When cells grow old, they die, and new cells take their place. Sometimes, this process goes wrong. New cells form when the body does not need them, and old cells do not die when they should. These extra cells can form a mass of tissue called a growth or tumor. Tumors (or neoplasms) can be benign or malignant. Benign tumors are not considered cancer, because they are rarely life-threatening, can be removed and usually do not grow back. Also, cells from benign tumors do not invade the tissues around them nor spread to other parts of the body. In the other hand, malignant neoplasms are cancer, may be life-threatening. And although they often can be removed, in many cases they grow back. Also carcinomas are able to invade and damage nearby tissues and organs and to spread (*metastasize*) to other parts of the body. Cancer cells spread by breaking away from the original (primary) tumor and entering the bloodstream or lymphatic system. The cells invade other organs and form new tumors that damage these organs. The spread of cancer in this form is what we call *metastasis*.

Cancer neoplasms correspond to malignant cells originated in glandular or epithelial lineages due to states of disordered genetic behavior. One particular point of focal interest is the de-regulation in the mechanisms that control the transcription of mRNA under normal conditions. A usual scenario is given by the so called *transcriptional bursts* which are stages of unusually high levels of mRNA synthesis within affected cells.

In the case of breast cancer, it forms in tissues of the breast, usually the ducts (tubes that carry milk to the nipple) and lobules (glands that make milk). It occurs in both men and women, although male breast cancer is rare. When breast cancer cells spread, the cancer cells are often found in lymph nodes near the breast. Also, breast cancer can spread to almost any other part of the body. The most common are the bones, liver, lungs, and brain. When metastatic processes arise, the new tumor has the same kind of abnormal cells as the primary tumor. For example, if breast cancer spreads to the bones, the cancer cells in the bones are actually breast cancer cells. The disease is metastatic breast cancer, not bone cancer. For that reason, it is treated as breast cancer, not bone cancer.

Given the large evidence of the genetic origins of cancer, a usual experimental tool to its study is the use of genome-wide high-thoughput gene expression analysis. In the following, we will demonstrate how to apply the tools of statistical physics to extract relevant information for such kind of experimental studies.

2. Gene expression data analysis

In recent times, the use of high density oligonucleotide arrays has become widely used in several instances in the molecular biomedical research community. The system, also known as GeneChip®-technology made use of oligonucleotides, usually of 25 base-pairs in longitude that are used to probe genes. Each gene is generally represented by a set of 16-20 pairs of those oligonucleotides known as probe sets. One of each pair of these oligos is known as the perfect match (PM) probe and correspond to an exact segment of the complementary sequence of the associated gene, whereas the other one, known as the mismatch probe (MM) is made by changing the middle (13th) base in order to look up for the effects of non-specific binding [4].

mRNA experimental samples are prepared, labeled with a fluorescent dye (see Fig. 1) and hybridized to the arrays (chips) (Fig. 2). Then the chips are scanned with a laser and images are produced (Fig. 3) and analyzed to obtain an intensity value asociated to each probe. The intensity of the fluorescent signal of a probe is related to the concentration of the mRNA molecule corresponding (tagged) by this probe [3].

Of particular interest results the question of how to combine the data for the set of 16-20 PM-MM pairs to define a measure of *expression* that represent in an optimal way the amount of the associated mRNA species [5]. This is not a trivial issue since, as one could see for the physicochemical procedures just sketched, there are a lot of variables involved in the analysis (several orders of magnitude more than the number of experimental samples) and the resulting signals are very noisy. These facts imply that the usual *frequentist* approach to probability and statistics has to be modified to deal with GEA data.

In this work we analyzed genome wide expresion data obtained with Affymetrix HGU133Plus2 human gene expression chip under procedure GPL570 for 261 Microarrays (MAs) of Breast Cancer (BRCA) and Normal tissue samples from several independent experiments (from the following NIH-NCBI/GEO accession keys: GSE7904 (62 samples), GSE5460 (129 samples), GSE5764 (30 samples), GSE3744 (40 samples). All arrays were processed within the same protocol (GPL570) and in all cases unnormalized fluorescence raw data files (.CEL files) were used. The preprocessing was done according to the RMA [6] algorithm implemented in R/BioConductor [8] (see Sec. 2.2) and statistical tests for differential expression profiles were done in the FlexArray/Nanuq platform [9] (see Sec. 2.3).

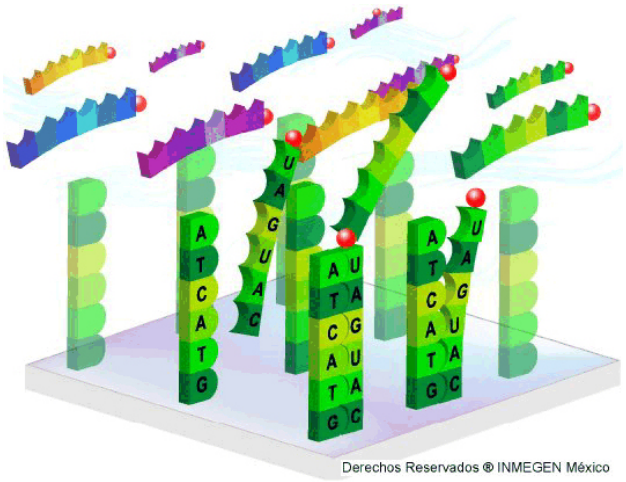

FIGURE 1. mRNA from the samples is marked with a fluorescent dye molecule [red spots] in solution and then put in contact to the surface of the GeneChip® .

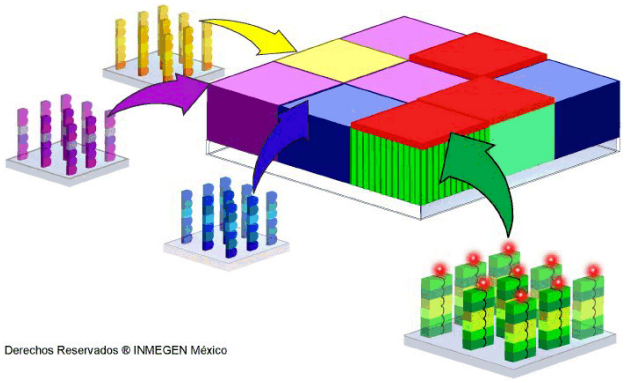

FIGURE 2. Tagged mRNA molecules (gene probes) hybridize in the chip's surface to complementary probes in localized regions (probe-sets) that now shine (in red here for pedagogical purposes) under the scanner's laser.

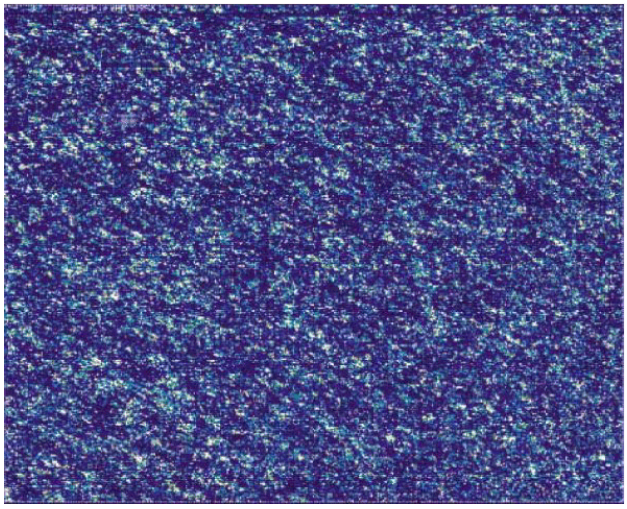

FIGURE 3. Scanning photograph of an actual GeneChip® as read by an Affymetrix unit.

2.1. Across laboratories comparisons

In this work (and in many other GEA-related works whose goal is to infer relationships under an integrative approach - the so-called *Systems Biology* paradigm-) we made use not just of a single experimental data source but of many. We did so, because in order to obtain the statistical power to make a reasonable deconvolution of the great information amount as given by high-througput technologies, one needs the biggest possible number of experimental samples. As is the case the number of *in-house* experiments that an individual facility could run is often limited by financial and technical matters [10].

Of course given the complex nature of the experimental setups for GEA [3], a question raises on the validity of using data from multiple sources (the so-called *across-labs* problem). Eventhough experimental settings may be very similar, each laboratory will usually end up with different results in the form of p-value distributions or gene lists. Under this conditions many people take these results and perform a statistical meta-analysis of the different data [10,11] to combine in a somehow systematic manner information from different labs. Of course, in order to be combined across studies, quantititative estimates should refer to the same measure or quantity, should be standarized to the same scale and should possess some inherent measure of variability.

A usual approach to this kind of meta-analysis is the performing of hierarchical Bayesian partitions and permutation of processed data [12]. A somehow better alternative involves gathering *raw* experimental data (.CEL files in the

case of Affymetrix GeneChips) from several sources and then pre-process them and normalize them together [11]. This of course implies a bigger amount of computational and human resources but is in our view a much *cleaner* and better alternative. Also Bayesian or information theoretical estimates of the joint probability distributions inferred from these *pools* of data are more sound, because systematic and fixed effects tend to cancel [12]. We then decided to proceed accordingly and thus we only included in this paper, raw data of chips processed under the very same protocol in different laboratories. We then did all the pre-processing and array normalization ourselves.

2.2. Pre-processing: Background correction and normalization

By analyzing the statistical behavior of the PM and the MM probes under controlled experimental conditions [5] some facts are revealed. For example, for large values of genetic abundance the differences between PM and MM probes have a bimodal distribution with the second mode occurring for negative differences. This effect has been related with heteroscedasticity (unequal variances in the distributions) [6].

Another strong challenge in GEA is related to how to effectively dissociate actual gene expression values from experimental noise. The hybridization noise has been found [14] to have very strong dependence on the expression level, with different characteristics for the low and high expression values. The hybridization noise characteristics at the high expression regime are mostly Poisson-like, whereas its characteristics for the small expression levels are more complex, probably due to cross-hybridization. Thus, in order to correctly assess the statistical relevance of the measured gene expression differences between two experiments, it is crucial to characterize the fluctuation caused purely by experimental measurement. It is known that noise depends strongly on the expression level. Therefore, an expression-dependent distribution function is needed to characterize the variability between replicates [14].

A related source of undesired variation is that, on increasing mRNA concentration levels, the distance of the average PM intensity to the background noise increases. The levels of background intensity could thus mask the effects of some mean-valued expression levels, *i.e.* the average *shining* effect could hide a relatively important signal. In order to optimize the *signal-to-noise-ratio* (SNR) a background correction has been proposed [5]. We will consider a model for the PM probes including both *true* signal and background noise in the following form $PM_{ijn} = bg_{ijn} + s_{ijn}$. If we assume that each array has a common average background level $\mathbb{E}(bg_{ijn}) = \beta_i$ it is possible (but very naïve) to consider removal of the background effect by substracting the β_i, $PM_{ijn}^{corrected} = PM_{ijn} - \beta_i$. A better alternative for improving the SNR is to consider the background correction as $\mathbb{B}(PM_{ijn}) = \mathbb{E}(s_{ijn}|PM_{ijn})$. This background correcting procedure is based on the consideration of the $\mathbb{B}$-transform that, as stated above, consists on adjusting the background noise via the conditional expectation of the signals on the PM values. The usual way to do so is by considering exponentially distributed signals and normally distributed background noises [5].

Also, in the vast majority of the applications of GeneChips one wishes to learn how mRNA concentration profiles differ in response to genetic, cellular and environmental differences. One important instance is when large (or small) expression of a given gene or set of genes may cause an illnes (such as cancer), thus resulting in variation between diseased and normal tissue (a so called case-control comparison). However, observed intensity levels also depend on sample preparation, manufacture of the arrays, and lab processing of such arrays (dye labeling, hybridization and scanning). These are called sources of *obscuring variation* [5].

Due to these facts, unless arrays are correctly *normalized* comparing data from different arrays can lead to misleading results. For example, in Fig. 4 we present a scatter plot of intensity of two different chips one from a breast cancer patient and the other from normal tissue. As we can see, it shows

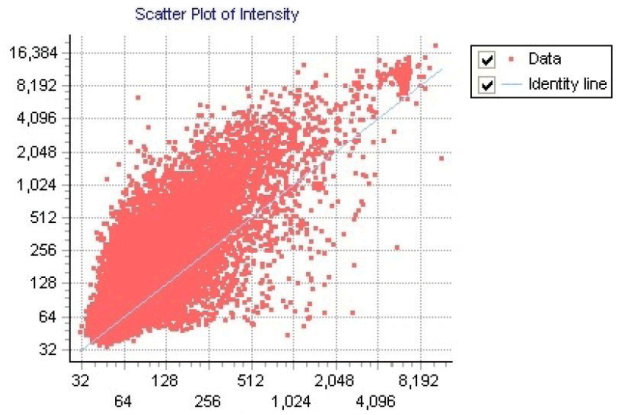

FIGURE 4. Scatter plot of intensity of a breast cancer patient (y-axis) versus normal tissue (x-axis) mRNA intensity levels.

FIGURE 5. QQ-plot for the quantile normalized distributions from the same data as in Fig. 4. Deviations from the identity line are over- or under- expressed genes.

a very noisy pattern, mostly because the data is not properly normalized. Nevertheless, since we are looking for statistically significant deviations from the identity line (no expression change) unnormalized data may lead to wrong conclusions. Several methods have been proposed to normalize the arrays and it has been found [6] that *quantile normalization* (QN) performs best. The goal of QN is to make the distribution of probe intensities the same for arrays within a given category (quantile), in such case a quantile-quantile plot (QQ plot) will be given as an identity line (see Fig. 5).

A generalization for the usual QQ-plot for n data vectors could be given in terms of a *projection operator* formalism, as follows: if all n data vectors (in this case n is the number of arrays) have the same distribution, then plotting the quantiles in n dimensions gives a straight line along the direction given by the unit vector

$$\mathbb{D} = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

This suggests thus that we could make a set of datasets having the same distribution if we *project* the points of our n dimensional quantile plot onto the diagonal $\mathbb{D}$.

If we denote by $\vec{q}_k = (q_{k,1}, \ldots q_{k,n})$ the vector of the k-quantiles for n arrays and $\mathbb{D}$ the unit-diagonal, the associated projection that make all quantiles lying in the diagonal is thus written as:

$$proj\,[\mathbb{D}|\vec{q}_k,] = \left(\frac{1}{n}\sum_{j=1}^{n} q_{k,j}, \frac{1}{n}\sum_{j=1}^{n} q_{k,j} \ldots \frac{1}{n}\sum_{j=1}^{n} q_{k,j}\right) \quad (1)$$

In other words, we *quantile-normalize* the arrays to the same distribution by taking the mean quantile and substituting it instead of the data item in the original dataset. This QN process could be computationally implemented in the following form [7] :

1. Given a dataset consisting of n arrays of lenght p, form the matrix $\mathfrak{X}$ of dimension $p \times n$ where each array is a column
2. Sort each column of $\mathfrak{X}$ to get $\mathfrak{X}_{sort}$
3. Take the means across rows of $\mathfrak{X}_{sort}$ and assign this means to each element in the row to get $\mathfrak{X}^{\dagger}_{sort}$
4. get $\mathfrak{X}$ normalized by rearranging each column of $\mathfrak{X}^{\dagger}_{sort}$ to have the same ordering as the original $\mathfrak{X}$.

We could rephrase Eq. 1 and the above algorithm in terms of non-linear transforms of the original distributions. QN is a special case of a transformation $x_i^{\dagger} = F^{-1}\left(G(x_i)\right)$ where estimates of G are given by the empirical distribution of each array and estimates of F followed the empirical distribution of the averaged sample quantiles (G and F are the standardize normal Gaussian and Fisher density distributions, respectively) [7].

A minor drawback of the bitransformed QN scheme just sketched is that it could fail to represent appropriately the tails of the distribution. However, in practice, given the fact that probeset expression measures are usually computed using multiple probes, this problem is highly diminished. Also after analyzing the error distributions, some researchers have proposed expression measures based only on the PM probes [13].

A general scheme has been proposed, called the Robust Multi-array Average (RMA) algorithm [6] which

1) background-corrects the arrays using an *expectation-of-the-signal* transformation ($\mathbb{B}$-transform),
2) normalizes the arrays by a QN and
3) Fits a linear model to summarize the probe intensities for each probeset.

RMA is less noisy than all other measures at lower mRNA concentrations, has a smaller spread (and thus is better tailored to detect differentially expressed probe-sets) and has greater sensitivity. For all the reasons above we decided to perform RMA correction and pre-processing of the raw gene expression data before performing any additional analyses.

2.3. Statistical tests for differential expression

As we have seen whole-genome GEA has turned out to be a technology that now is capable of providing genome-wide patterns of gene expression across many different conditions. The basic level of analysis of these patterns consists in categorize whether observed differences in expression (see for example, Fig. 6) are significant or not. Traditional statistical methods are unsatisfactory due to the lack of a systematic framework that can accommodate noise, variability, and low replication often typical of microarray data.

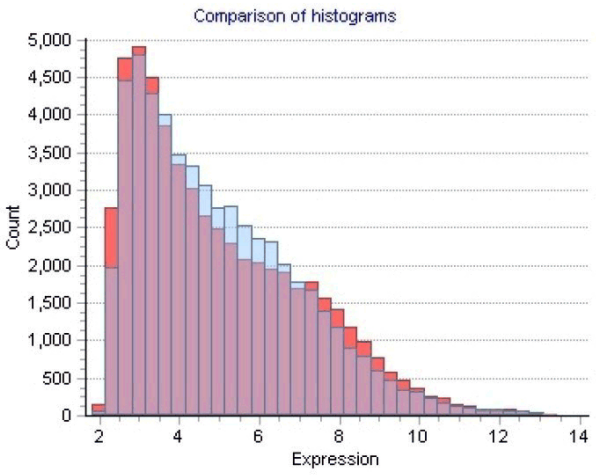

FIGURE 6. Histogram of gene expression after RMA preprocessing: dark-pink bars corresponds to breast cancer and translucent baby-blue bars to normal tissue. We are interested in single-colored regions, that exhibit differential expression between these two conditions.

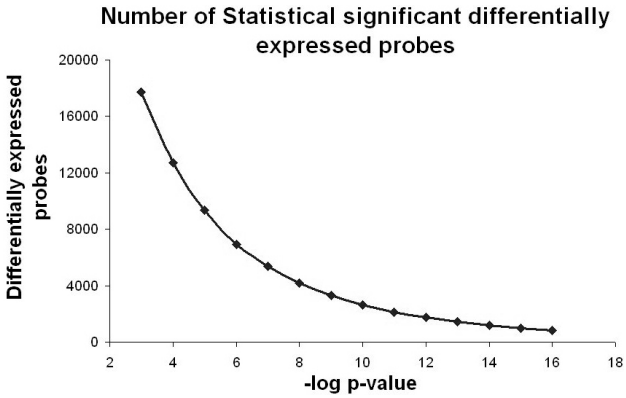

FIGURE 7. Number of statistically significant differentially expressed gene probes (NGP) vs p-value. Parametric t-tests were done to asses significance. $NGP = 29381 \times e^{-0.2318\,[-\log_{10} p-value]}$; $R^2 = 0.9906$.

Due to this fact, Baldi and Long [15] developed a Bayesian probabilistic framework for MA data analysis. Baldi-Long analysis (also called a cyberT-test) consists in modeling log-expression values by independent normal distributions, parameterized by corresponding means and variances with hierarchical prior distributions. From this data one derives point estimates for both parameters and hyperparameters, and regularized expressions for the variance of each gene by combining the empirical variance with a local background variance associated with *neighboring* (in parameter space) genes. An additional hyperparameter, inversely related to the number of empirical observations, is used to determine the *strength* of the background variance. These point estimates, combined with a statistical t-test (a *regularized* t-test), provide a systematic inference algorithm that compares favorably with the more widely used simple t-test or fold methods, and partly compensate for the lack of replication within the actual GEA framework.

Once we have properly pre-processed data, we are now in position to use it to look up for regulatory interactions. To do that we will apply the tools of non-linear statistical physics, in particular a quasi-Hamiltonian formalism based on the consideration of information-theoretical measures of correlation as surrogates for physical interactions as we will show in the following section.

3. Non-linear analysis

The deconvolution of the gene regulation interactions set will be based on a information-theoretical optimization of the Joint Probability Distribution (JPD) of gene-gene multicorrelations as given by gene expression experimental data. This could be implemented as follows. The JPD for the stationary expression of all genes, $P(\{g_i\})$, $i = 1, \ldots, N$ may be written as follows [16]:

$$P(\{g_i\}) = \frac{1}{Z}\, exp^{H_{gen}} \tag{2}$$

$$H_{gen} = -\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j) - \sum_{i,j,k}^{N} \Phi_{i,j,k}(g_i, g_j, g_k) - \ldots \tag{3}$$

Here N is the number of genes, Z is a the partition function, the Φ's are empirical *interaction potentials*. A truncation procedure in Eq. (3) will be used to define an approximate *coarse-grained hamiltonian* H^{aprox} that aims to describe statistical properties of the system. A set of *particles* (genes) Ω, interacts with each other when the potential Φ_Ω between such set is non-zero. The relative contribution of Φ_Ω is taken as proportional to the strength of the interaction between this set.

A closer look of Eq. (3) reminds us of the typical problem of the BBGKY hierarchy in statistical mechanics. Unfortunately, there is not such well defined criteria to truncate the geometric series expansion in multiple p-way interaction potentials $\Phi_{i_1,i_2,\ldots,i_p}(g_{i_1}, g_{i_2}, \ldots, g_{i_p})\,;\ \forall i$ as in, for example the diluted regime in the BBGKY hierarchy (see, for example [18]).

In the current genomics literature, sample sizes of order 10^2 (the usual maximum size available in most present-day studies) are generally taken as sufficient to estimate 2 - way marginals, whereas 3-way marginals [*e.g.* triplet interactions $\Phi_{i,j,k}(g_i, g_j, g_k)$] require about an order of magnitude more samples, a sample size unattainable under present circumstances. Being this the case, one is usually confronted with a 2-way hamiltonian of the form:

$$H^{aprox} = -\sum_{i}^{N} \Phi_i(g_i) - \sum_{i,j}^{N} \Phi_{i,j}(g_i, g_j) \tag{4}$$

Given these facts, the deconvolution of the set of biochemical interactions consists in the *inverse-problem* of determining the complete set of relevant 2-way potentials $\Phi_{i,j}(g_i, g_j)$ consistent with the JPD [Eqs. (2) and (3)] that defines all known constrictions, *e.g.* the values of the stationary expression of genes g_i as given by the set of $\Phi_i(g_i)$'s and non-committal with every other restriction in the form of a marginal [17]. A recently developed approach is the use of statistical and information theoretical models to describe the interactions [19].

If we consider a 2-way interaction hamiltonian, all gene pairs i,j for which $\Phi_{i,j} = 0$ are said to be non-interacting. This is true for genes that are statistically independent, $P(g_i, g_j) \approx P(g_i)\, P(g_j)$, but it is also valid for genes that do not have a direct interaction but are connected via other genes *i.e.* $\Phi_{i,j}$ = 0 but $P(g_i, g_j) \neq P(g_i)\, P(g_j)$. Several metrics such as Pearson Correlation, Square Correlation and Spearman Ranked coefficients over the sampling universe have been used, but the performance of these methods is usually

poor as suffers from a big number of false positive predictions. To overcome some of the limitations of these methods, we inferred gene-gene regulatory interactions by means of a local pattern-sharing approach [20].

3.1. Information theoretical measures

In the nonlinear analysis of highly complex data, special attention should be given to the discovery of hidden information within the repetitive appearance of certain basic patterns embedded in the given signals. In order to detect and somehow quantify such underlying structures several methods have been developed. A very promising approach is based on the consideration of the linguistic properties of the symbolic dynamics associated with the time series under consideration. This approach based on the quantification of the so called Information-Based Similarity Index (IBS) [20] initially developed to work out the complex structure generated by the human heart beat time series. Nevertheless, IBS has proved to be a very powerful tool in the comparison of the dynamics of highly nonlinear processes. In the particular case to be considered here, the data provided is the distribution of gene expression intensities for the set of differentially expressed genes in the group of cancer and normal tissue samples.

A promising approach to understnad these kinds of interactions is given if we consider that the correlations in the system are given by *communication channels* (either real or abstract) for the bio-signals. Thus, Information Theory (IT) could play a useful role in identifying entropic measures between pairs $\{g_i, g_j\}$ of genes within the sampling universe as potential interactions $\Phi_{i,j}$. IT can also provide with means to test for the MaxEnt distribution, by considering, for example the Kullback-Leibler (KL) divergence (also called multiinformation) or the Connected Information as criteria of iterative convergence to the MaxEnt PDF in the same sense that the cumulant distribution leads to the specification of usual PDFs [21].

Within the present context the symbolic sequence represent the expression values of a single gene (say gene k-th) all along the sampling universe (of size M), as given by a gene-expression vector $\Gamma = \overrightarrow{g_k} = (g_{k_1}, g_{k_2}, \ldots, g_{k_M})$. It is possible to classify each pair of successive points into one of the following binary states I_n, if $(\Gamma_{n+1} - \Gamma_n) < 0$ then $I_n = 0$; in the other case $((\Gamma_{n+1} - \Gamma_n) > 0)$ $I_n = 1$. This procedure maps the N step real-valued time series $\Gamma(i)$=$\{\Gamma_1, \Gamma_2, \ldots, \Gamma_N\}$ into an $N-1$ step binary-valued series $I(i)$. Once we have this series, it is possible to define a binary sequence of length m (called an m-bit word). Each of the m-bit words w_k represents a unique pattern of fluctuations in a given time series. For every unitary time-shift λ, the algorithm makes a different collection W_λ of m-bit words over the whole time series, $W_\lambda = \{w_1, w_2, \ldots, w_n\}_\lambda$. It is expected that the frequency of occurrence of these m-bit words will reflect somehow the underlying dynamics of the original (real-valued) time series.

It is in this point that one is to recall that in studies of natural languages it has been observed that authors have different *word usage*, *i.e.* characteristic preferences for the frequency of use of every word, this fact has been reflected on a statistical linguistic phenomena known as Zipf's law [22]. This statistical principle has been applied in a wide variety of systems from natural language [23] to DNA sequences [24, 25] and even musical structures [26]. In order to apply this concept to symbolic sequences, one should consider the frequency of every m-bit word and then sort them in descending order by frequency of occurrence, in this way we are able to write down a probability distribution function in the *rank-frequency* representation (RF-PDF). This RF-PDF represents the statistical hierarchy of symbolic words of the original time series [20]. In the theory of stochastic processes, two given symbolic sequences (chains or strings) are said to be *statistically equivalent* if they give rise to similar (or even identical) probability distribution functions.

Following the very same order of ideas, Yang and coworkers [20] defined a measure of similarity (akin to statistical equivalence) between two time series by plotting the rank number of every m-bit word in the first time series with the rank for the same m-bit word in the second time series. Obviously if the two RF-PDFs are statistically equivalent, then the scattered points will lie *almost surely* in the diagonal line. In this sense, the average deviation of these points from the diagonal (*i.e.* $\theta = 45°$) is a good measure of the distance (or dissimilarity) between these two time series.

Of course since the time series are supposed to be finite, the m-bit words are not equally likely to appear. The method introduces the likelihood of each word by defining a weighted distance Δ_m between two given symbolic sequences σ_1 and σ_2 as follows:

$$\Delta_m(\sigma_1, \sigma_2) = \frac{1}{2^m - 1} \sum_{k=1}^{2^m} |R_1(w_k) - R_2(w_k)| F(w_k) \quad (5)$$

$F(w_k)$ is the normalized likelihood of the m-bit word k, weighted by its given Shannon entropy, *i.e.*:

$$F(w_k) = \frac{1}{Z} \left[-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k) \right] \quad (6)$$

in this case, $p_i(w_k)$ and $R_i(w_k)$ represent the probability and rank of a given word w_k in the i-th series. The normalization factor in Eq. (6) is the total Shannon's entropy of the ensemble and is calculated as

$$Z = \sum_k \left[-p_1(w_k) \log p_1(w_k) - p_2(w_k) \log p_2(w_k) \right].$$

$\Delta_m(\sigma_1, \sigma_2)$ is called the Information Based Similarity Index (IBS) between series σ_1, and σ_2. One notices that $\Delta_m(\sigma_1, \sigma_2) \in [0, 1]$; $\forall \sigma_1, \sigma_2$; $\forall m$. In fact one is able to consider $\Delta_m(\sigma_1, \sigma_2)$ as a probability measure. In the situation in which $\lim \Delta_m(\sigma_1, \sigma_2) \to 1$ the series are absolutely dissimilar, whereas in the opposite case given by $\lim \Delta_m(\sigma_1, \sigma_2) \to 0$ the two series become identical (in the

statistical sense). One can then approximate the value of the interaction potentials $\Phi(g_i, g_j)$ between genes g_i and g_j as follows. If one is to consider interaction as given by correlation or information flow, one can notice that high values of Δ_m imply stronger dissimilarity, hence lower correlation and since Δ_m is a probability measure, one can define the complementary measure $\Delta^*_m = 1 - \Delta_m$ and then one can approximate $\Phi(g_i, g_j) \approx \Delta^*_m(g_i, g_j)$

4. Results

4.1. Over and under expressed genes

After pre-processing the gene expression data according to Sec. 2.2, we proceeded to perform statistical tests (Sec. 2.3) to look up for differentially expressed genes, *i.e.* genes that are present in a much higher (lower) concentration in diseased samples as compared to normal samples. The resulting number of differentially expressed gene probes depends of course on the statistical bounds of confidence as given, for example in the form of p-values. In Fig. 7 we can see the dependence of the number of statistically significant differentially expressed gene probes (NGP) with the p-value of the cut-off on a parametric t-test (Baldi-Long cyber T-test). A list of the top 20 over and under expressed genes in lobular and ductal breast cancer with respect to normal tissue is given in Table I.

TABLE I. **Top 20 Over and Under regulated genes in Breast Cancer.** Genes marked with an asterisk have been previously reported associated with breast cancer (PRBC)

Over regulated genes	PRBC	Under regulated genes	PRBC
KRT14	*	CTHRC1	*
DST	*	COL11A1	*
KRT15	*	LYZ	*
C2orf40		COL1A2	*
OXTR	*	COL10A1	
PIGR		RRM2	*
MYH11	*	COL11A1	*
KRT17	*	ASPN	*
KRT5	*	CDC2	*
KRT17	*	TOP2A	*
CNN1	*	COL8A1	*
CCL28		COL10A1b	*
SFRP1	*	KIAA0101	*
SBEM	*	RRM2	*
DMN		CXCL10	*
AK5	*	LY96	*
NTRK2	*	IGHV4-31	*
PTN	*	PRC1	*
FOSB	*	COL6A2	*
WIF1	*	MIF	*

The list of over- and under- regulated genes in Table I shows interesting features. 16 out of the 20 over-expressed genes have been previously reported associated to breast cancer. Also, 17 out of the top 20 under-regulated genes have been linked to breast cancer neoplasms (complete list of references available upon request). A statistical enrichment analysis of gene Ontology (Biological Process) [39] categories among the genes in Table I, also reveals interesting tendencies. The list, small as it appears, includes statistically significant enrichment of the following biological processes (we include permutation test p-values): multicellular organismal development (4.22×10^{-6}), anatomical structural development (1.29×10^{-5}), tissue development (5.74×10^{-5}), extracellular matrix, organization and biogenesis (2.53×10^{-3}), cell surface receptor and linked signal transduction (1.35×10^{-2}), positive regulation of retroviral genome replication (1.81×10^{-2}), epithelial cell proliferation (4.02×10^{-2}), apoptotic chromosome condensation (4.02×10^{-2}), negative regulation of apoptosis (4.57×10^{-2}) and negative regulation of programmed cell death (4.67×10^{-2}). These processes are linked to cell growth and proliferation, cellular communication, DNA damage and suppression of apoptosis. All of these functions are well known hallmarks of cancer.

4.2. Threshold determination

Most of the means of interaction inference are based on the evaluation of one or several values for the correlations (usually given in the form of probability or information theoretical measures). The problem is that there is no well defined criteria as to what is the *right value of the cut-off* in, for example the p-value, the IBS measure or another quantitative indicator of interaction. For if one takes a *too-stringent* criteria one is possibly having left-out an important interaction and if your cut-off is *too-loose* you will end up having a lot of false positive links. The usual approach to solve a problem like this is starting with a set of known interactions, generate a dataset with an open range of cut-off values and then choose the value that preserve the real interactions giving rise to a minimal error set. Unfortunately, for most interesting cases there is no set of already known interactions and, if it indeed exists, it is generally very poor.

Here we propose an alternative way to tackle the thresholding problem. The design was done specifically for genomic expression data but it is probably adaptable to other kinds of data. We developed a mixed approach based on the quantification of a *global* statistical criteria for which it is possible to generate permutation p-values, and also a *local* pattern-sharing IBS analysis characterized by two quantities, the IBS value itself and the size m of the m-bit-word window.

We calculated the IBS index for the gene expression vectors (GEVs) of the genes and for m values from 5 to 10 (the

algorithmic complexity of the calculations grows exponentially so m = 10 was an upper bound for our calculations given the large sample number of 261 whole genome experiments). For the sake of stringency we only retained as interactions those pairs of genes whose GEVs have $\Delta_m^* \geq 0.85$ corresponding to the higher degrees of correlation. As it can be seen in Fig. 8, an asymptotic regime is attained for these networks for values of m = 5-8 depending on the IBS (Δ_m) cut-off (or equivalently the Δ_m^* threshold).

TABLE I. **Top 20 Validated Gene-Gene Interactions.** All values of $\Delta_m^* > 0.85$, Function refers to gene pairs with known biological function, in some cases, previous reference to breast carcinoma (PRBC) has been given.

Gene$_i$	Gene$_j$	Function	PRBC
TBCB	NINJ2	Cell adhesion	–
PI15	SASH1	Tumor Suppression	[27]
GSTM3	RASSF4	Tumor Suppression	[28, 30]
LOC152217	MANEAL	Carcinogenesis	[31]
DEPDC1B	DUSP4	Chromosome instability	[32]
MCM4	CKLF	Chromosome instability	[33]
MCM4	ABCB10	Tumor Suppression	[34]
RBX1	ABCA5	Cell dysruption	–
GSTM3	ARSB	Tumor Suppression	[28]
CDC6	TRERF1	Hormone resistance	[36]
CDC7	CLPX	p53 inactivation	[37]
MCM4	MOXD1	Chromosome instability	[33]
KIAA0251	ARHGEF11	Angiogenesis	–
GAS2L3 /MTPN	LMNB1	Cytotoxicity	–
CDC7	COPE	Tumor Suppression	[37]
MELK	BET1	Proliferation	–
DEPDC1B	KRT15	Chromosome instability	[32]
BUB1B	RPL10A	Genomic Instability	[38]
CDC7	PHYHIP	Tumor Suppression	[37]
RPL3	ARPC5	Chromosomal instability	–

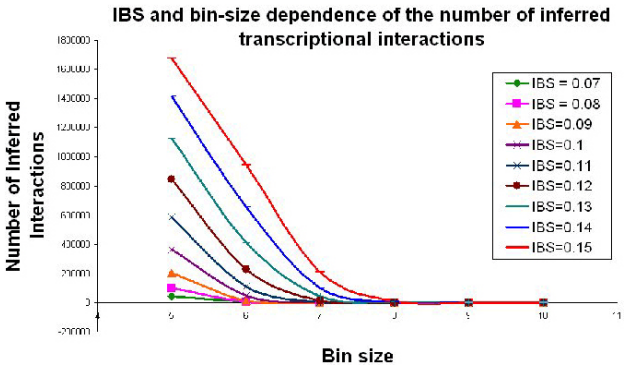

FIGURE 8. Dependence of the number of inferred transcriptional interactions on the information measure (IBS) and the bin size (pattern window size).

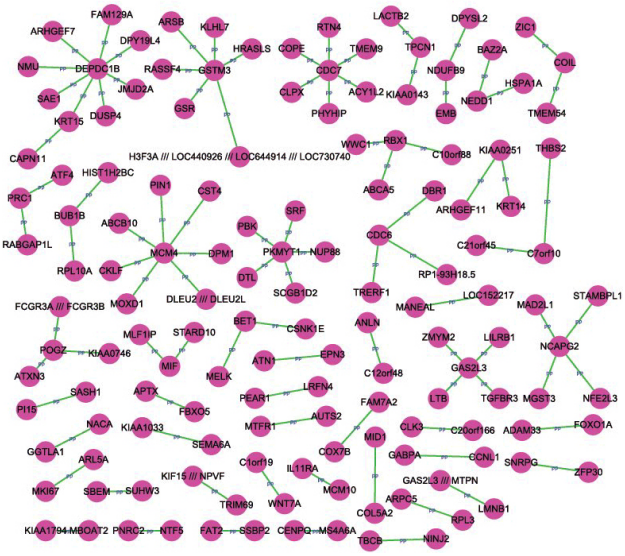

FIGURE 9. Transcriptional regulation network for breast cancer top gene-gene interactions.

4.3. Candidate interactions

Gene pairs with a high degree of correlation ($\Delta_m^* \to 1$) are considered candidate interactions. These are the most valuable result of this research since they are the starting point for the discussion and validation of biological pathways and processes to be tested in the laboratory and clinical studies in breast cancer. The complete list of interactions according to their Δ_m^* and p-value is available upon request. For p-values lesser than 1×10^{-11} and $\Delta_m^* \geq 0.85$ gene-gene interactions are showed as a network in Fig. 9. A list of some of the more important interactions discovered / confirmed by this research is given in Table II.

The importance of inferring transcriptional regulation interactions become evident by considering the examples given in Table II we have already seen that GEA reveals gene over-and- under-regulated that are associated with some common biological processes that are dysfunctional in cancer cells. Nevertheless, by considering table II. We saw that transcriptional interactions inferred by means of our statistical-mechanical formalism are related to highly specific features of Cancer such as chromosomal instability, tumor suppression, and carcinogenesis. For 14 out of the 20 showed interactions there are even experimental reports linking them specifically to breast carcinoma. Further studies of the associated (known or even unknown) biochemical pathways will further improve our understanding of both the structure and function of the cellular mechanisms related to breast cancer.

5. Conclusions and perspectives

Gene regulation, as we could see, is a complex phenomenon. A great number of so-called *reverse engineering techniques* have been developed to discover the relations and interactions between genes and/or sets of genes. In the particular case of

information theory based methodologies, most of them consider global correlations between GEVs (measured either on a simultaneous or dynamic way) such as mutual information, Kullback-Liebler divergences and so on. This approach has been, of course a fruitful one.

Apart from global measures of information theoretical correlation, there is a need for a local analysis of correlation. We then proposed here a complementary study based in the local *pattern-sharing* analysis between GEVs. For this end we applied a measure (IBS) that quantifies pattern-sharing of two data series under an information theoretical framework [20]. We applied IBS as a measure of interaction between two genes. The rationale behind this analysis is that the higher the degree of pattern sharing between two GEVs the larger its mutual correlation will be. If one is looking for the more relevant interactions, one should take a view on the large scale pattern-sharing (*i.e.* larger values of the bin size m). However, to account for more subtle interactions, then a closer look at pattern sharing should be adequate (smaller values of m).

As we already stated, microarray gene expression experiments were performed to compare the patterns of expression between breast carcinoma and normal tissue. From the set of statistically-significant differentially expressed genes, we found a set of transcriptional interactions as a means to discover a set of functional relations pertinent to breast cancer physiopathology. The knowledge of such functional relations is essential to design and perform further investigations leading to the discovery of therapeutic targets. In this sense, statistical physics-inspired analysis of this kind could be seen as a powerful hypotheses generation methodology in the context of functional genomics.

In the present case, we were able to reconstruct an ensemble of transcriptional interactions between genes that we prioritized in relation with ductal and lobular breast carcinomas in human samples. This group of highly validated interactions forms the theoretical basis to conform a gene regulatory network (after biological validation and annotation into biochemical-pathway charts).

Present findings point out to the existence of transcriptional modules that are related to such *cancer-only* processes like tumor suppression loss, chromosomic instability and apoptotic inactivation in the one hand; and to more general, *normal*-cellular functions processes like cell adhesion and dysruption, hormone resistance, proliferation and angiogenesis in the other. In some transcriptional interactions these two effects are mixed-up but there are some instances in which *module-separability* may lead to the discovery of pharmacological targets with less cytotoxicity to the normal cells than the usual chemotherapeutic drugs.

In conclusion, information theoretical methods result in powerful tools to reconstruct gene functional relationships. Problems like low signal to noise ratios, under-sampling, non-linear interactions, delays, etc., could be coped-with in a systematic manner under the Information Theoretical Formalism. The interaction's set obtained was validated by several *in silico* analysis. However, the most outstanding result up to date was the discovery of some regulatory interactions among genes previously related to breast cancer and the biological implications of the correlated gene clusters.

1. M.V. Rockmanm and L. Kruglyak, *Nat. Rev. Genet.* **7** (2006) 862.
2. S. Komili and S. Silver, *Nat. Rev. Genet.* **9** (2008) 38.
3. E. Hernández-Lemus, *Jou. of Non-equil. Thermodyn.* (2009), (in press)
4. D.J. Lockhart *et al.*, *Nature Biotech.* **14** (1996) 1675.
5. R.A. Irizarry *et al.*, *Biostatistics* **4** (2003) 249.
6. R.A. Irizarry *et al.*, *Nucleic Acids Research,* **31** (2003) 4
7. B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed, *Bioinformatics* **19** (2003) 185.
8. http://cran.r-project.org/
9. http://genomequebec.mcgill.ca/FlexArray/
10. J.R. Stevens and R.W. Doerge, *Comp Funct Genom* **6 116** (2005) 122.
11. J.R. Stevens and R.W. Doerge, *BMC Bioinformatics* **6** (2005) 57. doi:10.1186/1471-2105-6-57.
12. J.K. Choi, U. Yu, S. Kim, and O.J. Yoo, *Bioinformatics* **19** (2003) i84.
13. D. Hekstra, A.R. Taussig, M. Magnasco, and F. Naef, *Nucleic Acids Research* **31** (2003) 1962.
14. Y. Tu, G. Stolovitzky, and U. Klein, *Proc. Natl. Acad. Sci. USA* **99** (2002) 14031.
15. P. Baldi and A.D. Long, *Bioinformatics* (17) (2001) 509.
16. A.A. Margolin *et al.*, *BMC Bioinformatics* **7** (2006). doi:10.1186/1471-2105-7-S1-S7.
17. E.T. Jaynes, *Phys. Rev.* **106** (1957) 620.
18. C. Cercignani, R. Illner, and M. Pulvirenti, *Applied Mathematical Sciences* **106** (Springer-Verlag 1994).
19. H. Li and M. Zhan, *EURASIP Journal on Bioinformatics and Systems Biology*, (2007) doi:10.1155/2007/49478.
20. A.C. Yang, S.S. Hseu, H.W. Yien, A.L. Goldberger, and C.K. Peng, *Phys Rev Lett* **90** (2003) 108103.
21. N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, Elsevier, The Netherlands, 1997).
22. G.K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press Inc., Cambridge, 1949).
23. B.B. Mandelbrot, *An informational theory of the statistical structure of languages*, in Communication Theory, ed. W. Jackson, Betterworth, (1953) 486.
24. H. Stanley *et al.*, *Physica A* **273** (1999) 1.

25. B. Cantú-Bolán and E. Hernández-Lemus, *Rev. Mex. Fís. E* **51** (2005) 118.

26. L. Dagdug, J. Alvarez-Ramirez, C. Lopez, R. Moreno, and E. Hernández-Lemus, *Physica A* **383** (2007) 570.

27. *SASH1: a candidate tumor suppressor gene on chromosome 6q24.3 is downregulated in breast cancer.* Oncogene. 2003 May 15;22(19):2972-83.

28. *Glutathione S-transferase M1, M3, P1, and T1 genetic polymorphisms and susceptibility to breast cancer.* Cancer Epidemiol Biomarkers Prev. 2001 Mar;10(3):229-36.

29. *Transformation of MCF-10A cells by random mutagenesis with frameshift mutagen ICR191: a model for identifying candidate breast-tumor suppressors.* Mol Cancer. 2008 Jun 5;7:51.

30. *Candidate tumor-suppressor genes on chromosome arm 8p in early-onset and high-grade breast cancers.* Oncogene. 2004 Jul 22;23(33):5697-702.

31. M. Katoh and M. Katoh, *Int J Oncol* (2006) **28** 1243.

32. *Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors.* Int J Cancer. 2006 Sep 1;119(5):1052-60.

33. *Genetic screen for chromosome instability in mice: Mcm4 and breast cancer.* Cell Cycle. 2007 May 15;6(10):1135-40. Epub 2007 May 5.

34. *CMTM5 exhibits tumor suppressor activities and is frequently silenced by methylation in carcinoma cell lines.* Clin Cancer Res. 2007 Oct 1;13(19):5756-62.

35. *Expression of centromere protein F (CENP-F) associated with higher FDG uptake on PET/CT, detected by cDNA microarray, predicts high-risk patients with primary breast cancer.* BMC Cancer. 2008 Dec 22;8:384.

36. *Relevance of breast cancer antiestrogen resistance genes in human breast cancer progression and tamoxifen resistance.* J Clin Oncol. 2009 Feb 1;27(4):542-9. Epub 2008 Dec 15.

37. *Cdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivation.* Neoplasia. 2008 Sep;10(9):920-31.

38. M. Scintu *et al.*, *Genomic instability and increased expression of BUB1B and MAD2L1 genes in ductal breast carcinoma*, Cancer Lett. 2007 Sep 8;254(2):298-307. Epub 2007 May 10.

39. http://www.geneontology.org