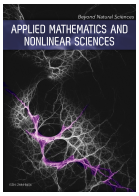
sciendo Applied Mathematics and Nonlinear Sciences 4(1) (2019) 223–230



Applied Mathematics and Nonlinear Sciences

https://www.sciendo.com

Intrachromosomal regulation decay in breast cancer

Guillermo de Anda-Jáuregui[a], Cristobal Fresno[a], Diana García-Cortés[a], Jesús Espinal Enríquez[a,b], Enrique Hernández-Lemus[a,b]*

a. Computational Genomics Division, National Institute of Genomic Medicine

b. Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México

Submission Info

Communicated by E. Ugalde
Received August 13th 2018
Accepted November 13th 2018
Available online June 28th 2019

Abstract

Biological systems exhibit unique phenotypes as the result of the expression of a genomic program. The regulation of this program is a complex phenomenon, wherein different regulatory mechanisms are involved. The deregulation of this program is at the centre of the emergence of diseases such as breast cancer. In particular, it has been observed that coregulation patterns between physically distant genes are lost in breast cancer.

In this work, we present a systematic study of chromosome-wide gene coregulation patterns in breast cancer as inferred by information theoretical measures over large (whole-genome expression in several hundred transcriptomes) experimental data corpora. We analyzed the chromosomal distance decay of correlations and found it to be with fat-tail distribution in breast cancer while being fundamentally constant in nontumour samples.

After model discrimination analyses, we concluded that the behaviour of the breast cancer distributions belongs to an intermediate regime between power law and Weibull distributions, with distinctive contributions corresponding to different chromosomes. This behaviour may have biological implications in terms of the organization of the gene regulatory program, and the changes found in this program between health and disease.

Keywords: Genomic regulation; Information theory; Cancer; Decay of correlations
AMS 2010 codes: 60-08; 62B10; 93A30

1 Introduction

Behind every biological system, there is a program governing the expression of different genes in the genome. In normal phenotypes, these programs will lead to a state of homoeostasis. Nevertheless, alterations to the regulatory mechanisms may lead to pathological conditions. One example of these pathologies is breast cancer.

Gene regulatory programs (GRPs) are composed of several genomic and epigenomic mechanisms interacting in a complex, nonlinear fashion. In recent years, genomic technologies have allowed the study of these

*Corresponding author.
Email address: ehernandez@inmegen.gob.mx

sciendo
ISSN 2444-8656 doi:10.2478/AMNS.2019.1.00020

systems. Such massive data may be studied and analyzed through the use of computational methods derived from Information Theory, in order to identify large-scale features related to the organization of the system.

We have studied the deregulation of GRPs in breast cancer in terms of the influence of physical distance between genes on coexpression. Our recent work has shown that there is a loss in gene coexpression between genes that are further away, either in different chromosomes (sometimes known as trans-regulation) or at greater distances within the same chromosome (sometimes known as cis-regulation) [4].

In this work, we characterize the relationship between intrachromosomal gene coexpression and distance. We identify a distance-dependent gene coexpression decay in a breast cancer phenotype, a phenomenon not observed in healthy breast tissue. We adjust these relationships to known models for heavy-tailed distributions. Through a model discrimination analysis, we observe that the distance-dependent gene coexpression decay may exist in an intermediate regime between a power law and a Weibull distribution, which may have important implications in terms of the organization of the GRP in health and disease.

2 Analysis

2.1 Probabilistic inference of GRPs

One relevant problem in contemporary computational biology is the probabilistic inference of the best (i.e., the maximum-likelihood or maximum-entropy) set of regulatory interactions between genes starting from a large –but still partial – data corpus Ω, as given for instance, in RNA sequencing experiments over whole-genome transcriptomes. We will call this problem the gene regulatory program deconvolution, GRPD. Solving the GRPD problem involves large-scale probabilistic inference in highly noisy data sets, and it thus remains a challenge to common probabilistic modelling approaches.

In order to circumvent these limitations, a number of algorithms – some of them with information theoretical foundations – have been developed. These approaches include mutual information maximization, Markov random fields, use of the data processing inequality, minimum description length, and Kullback–Leibler divergence. Also relevant to the GRPD problem are machine learning techniques, as well as Monte Carlo methods, variational methods, and hidden Markov models or stochastic linear dynamical systems. Common to these latter models is the fact that they are parametrically conditioned on the hidden state vector; the past, present, and future observations are statistically independent [8].

Information theoretical-founded approaches are useful for the tasks of feature selection and network inference. Feature selection methods applied to transcriptomics aim to improve molecular diagnosis and prognosis in complex diseases (such as cancer) by identifying a (minimum) set (a molecular signature) of features that characterize the underlying biological phenomenon. Network inference, in contrast, usually tries to present the full set of statistical dependencies between genes by means of a probabilistic graphical model of a gene regulatory network.

A first step towards solving either the feature selection or network inference subproblems of a GRPD is to have a detailed knowledge of the joint and marginal gene expression probability distributions. In what follows, we present an information theoretical approach to the GRPD problem via mutual information distributions.

2.2 The mutual information formalism

Let $X_i = \{X_1, X_2, \ldots, X_N\}$ be a set of N random variables, representing the expression levels of N different genes in a transcriptome. For each duplex $\mathbb{D}_{i,j} = (i, j)$ (representing a pair of genes) it is possible to define the mutual information function $I(X_i, X_j)$ as follows [3]:

$$I(X_i, X_j) = \sum_{i \in \mathscr{I}} \sum_{j \in \mathscr{J}} P(X_i, X_j) \log \frac{P(X_i, X_j)}{P(X_i)\, P(X_j)} \tag{1}$$

Here, $\mathscr{I}$ and $\mathscr{J}$ are the complete sampling spaces associated to the gene expression levels of genes i and j respectively –i.e. the sets of all possible values of X_i, and X_j, within a given (large) experimental data corpus Ω, associated with a general probability triple $(\Omega, \mathscr{F}, P)$. $P(X_i, X_j)$ is the joint probability distribution of X_i and X_j in Ω, whereas $P(X_i)$ and $P(X_j)$ are the marginal probability distributions of X_i and X_j, respectively. As it is widely know, the mutual information function $I(X_i, X_j)$ quantifies the statistical dependence between two given random variables X_i and X_j [3].

We can also define the off-diagonal mutual information, $I^{\dagger}$ as follows :

$$I^{\dagger}(X_i, X_j) = I(X_i, X_j) \cdot (1 - \delta_{ij}) \tag{2}$$

Here δ_{ij} is Kronecker's delta. $I^{\dagger}(X_i, X_j)$ equals the mutual information $I(X_i, X_j)$ everywhere, except in the set $i = j$, $\forall i$, $\forall j$. The purpose of $I^{\dagger}(X_i, X_j)$ is to eliminate self-information from our calculations. From now on, we will drop the † superscript and we will always refer to the off-diagonal mutual information in all of our further calculations.

2.3 Definition of GRPs

A GRP is the solution of a GRPD problem, i.e., the full set of interactions among genes that give rise to a transcriptional phenotype. Within the context of the theoretical and experimental settings we have just described, let us define what the solution of a GRPD problem is.

Definition 1. Here we define a **Gene Regulatory Program** as a graph $\mathscr{G}[I(X_i, X_j)]$ of all the mutual information functions for a given empirical transcriptomics sampling space Ω.

$\mathscr{G}[I(X_i, X_j)]$ contains all the information about the statistical dependencies among genes at a transcriptional level. $\mathscr{G}[I(X_i, X_j)]$ is thus a form of a **Markov Random Field** [10, 14], since it considers mutual information distributions at the pairwise sufficiency (hence, Markov) level [12].

2.4 Gene–distance dependency

Biological phenotypes are the result of a large set of regulatory interactions that are controlled by diverse genomic and epigenomic elements. Perturbation of these elements is involved in the origin and maintenance of the pathological phenotype observed in cancer. One of these elements is the spatial configuration of the genome. As we have previously mentioned, in a recent work [4], we have shown the existence of a major difference in the relationship between gene interactions and physical distance between breast cancer and healthy breast phenotypes.

Starting from the solution $\mathscr{G}[I(X_i, X_j)]$ of the breast-cancer GRPD program, we are in a position to study in a more formal and less constrained way the phenomenon of loss of long-range regulation that our group has been reporting to exist in breast cancer [4].

The functional $\mathscr{G}[I(X_i, X_j)]$ is composed of the complete gene–gene mutual information distributions. Two distinctive issues distinguish the present approach from the studies we have previously made on gene regulatory networks [1, 4, 6]. First of all, no mutual information threshold has been established, so that even interactions with very small statistical dependencies are considered. Second, no 'binning' has been applied to the data. By removing these common constraints – at the cost of a highly increased computational burden in already complex calculations– in our analysis, we are in a position to establish (as we will see later on) that the loss of long-range regulation in cancer that we have been reporting is not the product of thresholding nor of binning. We have already shown in our previous work [4] that this phenomenon is not a sampling-size or network inference algorithm artifact either. Since the concept of physical chromosomal distance is only reasonable in the context of genes within the same chromosome, we will from now on consider chromosome-wise GRPs $\mathscr{G}^k[I(X_i, X_j)]$, here $k = \{1, 2, \ldots, 22, x, y\}$ is an index working as the chromosome label.

By analyzing $\mathscr{G}^k[I(X_i, X_j)]$ it is possible to develop a deeper understanding on the way correlation structure

relates to functional features. For instance. in a previous work [6], we have observed the phenomenon of decay of long-range correlations. The visual inspection of the mutual information distribution dependence in the tumour plots in that work suggests a power-law decay of correlations.

2.5 Model discrimination analysis

Visual inspection, however, may be misleading us to attribute power-law behaviour to other long-tailed data distributions [2], and even common regression techniques may prove deceptive to establish functionality in heterogeneous variance settings [7, 25]. For these reasons, we decided to implement a comprehensive approach to model discrimination analysis.

To do this, we modeled the chromosomal gene-gene distance $d(i,j)$ dependence of the $\mathscr{G}^k[I(X_i,X_j)]$ by using the R implementation of Generalized Additive Models for Location Scale and Shape (GAMLSSs). GAMLSSs [16] are univariate distributional regression models. Under GAMLSSs, all the parameters of the assumed distribution for the response can be modelled as additive functions of the explanatory variables.

In the present context the use of GAMLSS allows us to perform differential goodness of fit analysis in the chromosome-wise gene-gene mutual information distributions. The data were adjusted to the following models: power law (PL), log-normal (LN), Weibull (WB) and Linear Exponential (LE), using polynomials up to order 4. This analysis generated $4 \times 4 \times 23 = 368$ models for the three high-likelihood distributions, under the constant, linear, logarithmic and polynomial approximations in 23 chromosomes –for the sake of this study, we have not considered chromosome y since the samples come from chromosomally female populations (xx), as data comes from breast cancer patients–.

3 Results and discussion

Once we have validated more formally the loss of long-range gene regulatory interactions in breast cancer, it is pertinent to analyze what is the distance dependence of the correlations. This may allow us to determine whether there is a preferred *correlation length* or any feature of this correlation structure that will provide us some hints as to what the actual biological mechanisms behind this decay of correlations are.

The inference of the chromosome-wise GRPs $\mathscr{G}^k[I(X_i,X_j)]$ is based the RNA sequencing data acquired, and pre-processed as in reference [4]. For the purposes of this discussion, we focus on the basal subtype of breast cancer because it is 1) usually of worse clinical significance, 2) it exhibited the most marked differences relative to the healthy phenotype in our previous work, and 3) it displays the lowest values of interchromosomal relationships [6]. It is important, however, to mention that all molecular subtypes of breast cancer show similar patterns as our example.

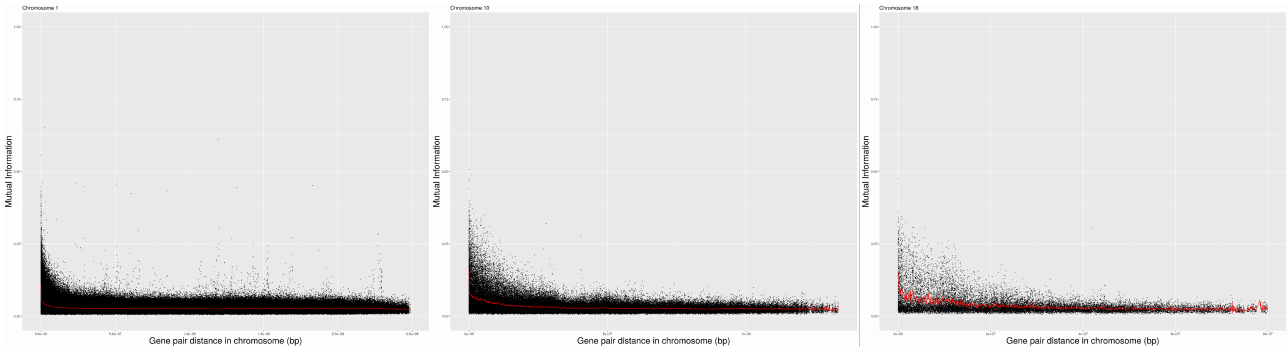


Fig. 1 Gene–gene mutual information–distance distributions. Depicted are chromosomes 1, 10, and 18 (large, medium, and small sized).

Figure 1 shows the full mutual information distributions for interactions in basal breast cancer and healthy breasts for chromosomes 1, 10, and 18. As can be observed, the strength of gene–gene correlations decays dramatically in cancer; meanwhile, for the healthy distribution, those values remain almost constant. As can be seen, the regulatory program in cancer is strongly altered (something that is, of course, widely known). One possible contribution to this deregulation relates to spatial chromosome rearrangements (i.e., changes in the 3D structure of chromatin).

In this regard, Figure 1 shows a long-tail decay of gene–gene mutual information correlations in the basal tumor distribution for chromosomes 1, 10, and 18 (upper panels), whereas a fundamentally constant (i.e., distance-independent) behaviour is observed in the nontumour distributions (lower panels).

As can be observed in Figure 1 and in Supplementary materials 1, the phenomenon of decay in the tumour intrachromosomal relationships occurs following a similar fashion in the different chromosomes. However, as can be seen in Figure 2 and in Supplementary materials 2, there are subtle differences in the decay regimes among chromosomes in breast cancer transcriptomes.

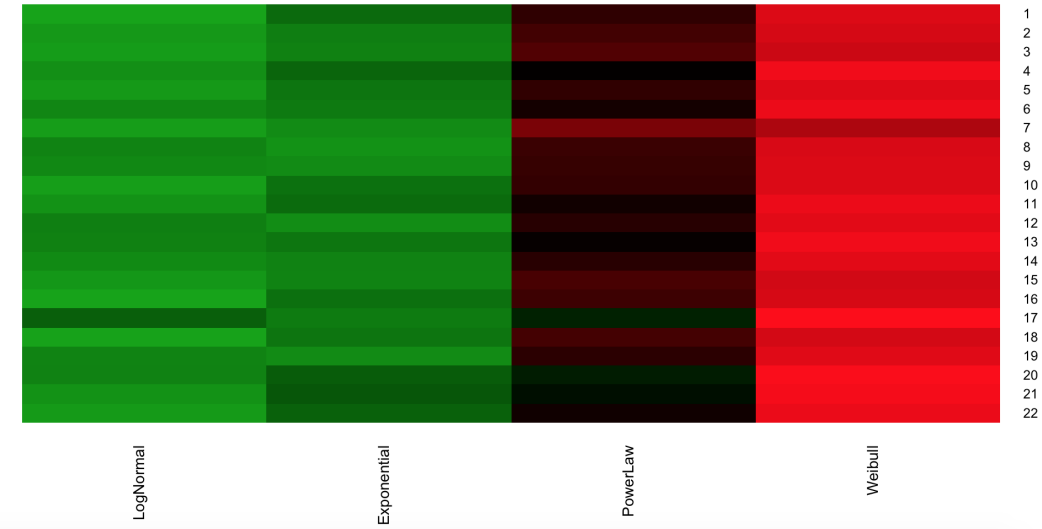


Fig. 2 Model validation. A heatmap depicting Z-scored relative likelihood – as measured by adjusted coefficient of determination R2 – for the four models. Rows correspond to chromosomes, while columns correspond to the four models for nonscaled distance. Green tones correspond to low-to-negative scores, while the dark and red colours represent positive-to-high scores.

3.1 Decay of gene–gene correlations

The model discrimination analysis we performed indicates that the best goodness of fit (by resorting to a combination of extensive determination coefficient (R^2) calculations and Bayesian information content predictors [16] – –See Supplementary Materials 2–) correspond to either the power-law or the Weibull distributions (see figure 2). The indicators for these distributions come so close, that it is not unreasonable to consider that the $\mathscr{G}^k[I(X_i,X_j)]$ in basal breast cancer indeed represent an intermediate regime with a mixture of power-law and Weibull distribution features. Indeed, it can be proved, by means of the Glivenko-Cantelli theorem [23] that the two distributions belong to the same Glivenko–Cantelli classes of functions [24] and then asymptotically converge as you get more data. As already discussed, the healthy breast distributions are almost distance independent, as can be seen in Supplementary materials 2.

Power-law decay of spatial correlations is a well-established phenomenon in the physical and biological settings [17,22]. Power-law decay may be resulting from the action of a generalized central limit theorem [5] for multiplicative growth processes [13,15,18,20], which may [26,27] or may not [21] result from self-organization

processes.

The Weibull distribution has been extensively used in recent times to characterize a variety of skew-distributed phenomena in the physical, biological, and economical sciences [9, 11, 19]. As in the case of the power law, the Weibull distribution may arise as a consequence of generalized central limit theorems, but in this case, for branching – instead of multiplicative growth – processes, as has been proved by Jo et al. [9] from the asymptotics of the Galton–Watson branching process of simple replicative systems. The authors have also shown that the branching process can be mapped into a process of aggregation of clusters. Weibull distributions may also arise in the context of percolation theory. Sornette [20] has shown how the existence of intermittency in continuum percolation changes the distribution from extreme exponential to a smoother Weibull-like form.

By considering the constructive processes associated with fat-tailed probability distributions – in particular, power law and Weibull-like decay – we can hypothesize that the behaviour of gene–gene correlations in breast cancer tumours may arise due to a combination of multiplicative growth, branching, and clustered aggregation processes. These hints are indeed relevant to arrive at the design of experimental strategies to probe what are the actual biological and physical processes behind the dramatic changes in the gene regulation patterns in cancer.

4 Conclusions

We have examined the phenomenon of decay of intrachromosomal regulation in breast cancer from an information theoretical perspective. In this work, we performed a systematic study of the phenomenon for each chromosome in breast cancer. By considering the whole GRP, defined in terms of mutual information, without resorting to any form of thresholding, binning, or any other method of feature selection or aggregation, we were able to model the relationship of intrachromosomal gene coexpression in terms of distance using the computationally demanding GAMLSS approach.

Comprehensive model discrimination analysis allowed us to identify that the distance-dependent gene co-expression decay lies in an intermediate regime between power law and Weibull distributions. This allows us to hypothesize that the divergence found between the healthy breast phenotype and breast cancer in terms of distance-dependent gene coexpression may arise from a combination of multiplicative growth, branching, and aggregation processes in the regulatory program. It can be argued that changes in the chromosomal structure, as well as changes in the topological associating domains in DNA, epigenomic phenomena, and chromosomal aberrations, may be among the more likely phenomena behind the loss of long-range regulation in cancer. However, a great deal of experimental as well as theoretical and computational data analysis must be done before arriving at a definite answer to this conundrum.

References

[1] Sergio Antonio Alcalá-Corona, Guillermo de Anda-Jáuregui, Jesús Espinal-Enriquez, Hugo Tovar, and Enrique Hernández-Lemus. Network modularity and hierarchical structure in breast cancer molecular subtypes. In *International Conference on Complex Systems,* pages 352–358. Springer, 2018.

[2] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[3] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[4] Jesus Espinal-Enriquez, Cristobal Fresno, Guillermo Anda-Jauregui, and Enrique Hernandez-Lemus. Rna-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Scientific Reports*, 7:1760, May 2017.

[5] Robert Fox and Murad S Taqqu. Central limit theorems for quadratic forms in random variables having long-range dependence. *Probability Theory and Related Fields*, 74(2):213–240, 1987.

[6] Diana García-Cortés, Guillermo de Anda-Jáuregui, Cristobal Fresno, Enrique Hernandez-Lemus, and Jesús Espinal Enriquez. Loss of trans regulation in breast cancer molecular subtypes. Bi*oRxiv,* page 399253, 2018.

[7] John A Gubner. Theorems and fallacies in the theory of long-range-dependent processes. *IEEE Transactions on Information Theory*, 51(3):1234–1239, 2005.

[8] Enrique Hernández-Lemus and Claudia Rangel-Escareño. The role of information theory in gene regulatory network inference. *Information Theory: New Research*, pages 109–144, 2011.
[9] Junghyo Jo, Jean-Yves Fortin, and MY Choi. Weibull-type limiting distribution for replicative systems. *Physical Review E*, 83(3):031123, 2011.
[10] Ross Kindermann. *Markov random fields and their applications*. American Mathematical Society, 1980.
[11] Jean Laherrere and Didier Sornette. Stretched exponential distributions in nature and economy: fat tails with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.
[12] Lina Merchan and Ilya Nemenman. On the sufficiency of pairwise interactions in maximum entropy models of networks. Journal of Statistical Physics, 162(5):1294–1308, 2016.
[13] Elliott W Montroll and Michael F Shlesinger. On 1/f noise and other distributions with long tails. *Proceedings of the National Academy of Sciences*, 79(10):3380–3383, 1982.
[14] John Moussouris. Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, 10(1):11–33, 1974.
[15] Mark EJ Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.
[16] Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.
[17] Georgios Roumpos, Michael Lohse, Wolfgang H Nitsche, Jonathan Keeling, Marzena Hanna Szymańska, Peter B Littlewood, Andreas Löffler, Sven Höfling, Lukas Worschech, Alfred Forchel, et al. Power-law decay of the spatial correlation function in exciton-polariton condensates. *Proceedings of the National Academy of Sciences*, 2012.
[18] Herbert A Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.
[19] D Sornette. Weibull-like failure distribution induced by fluctuations in percolation. *Journal de Physique*, 49(6):889–896, 1988.
[20] Didier Sornette. Multiplicative processes and power laws. *Physical Review E*, 57(4):4811, 1998.
[21] Didier Sornette. Mechanism for powerlaws without self-organization. *International Journal of Modern Physics C*, 13(02):133–136, 2002.
[22] HE Stanley, SV Buldyrev, AL Goldberger, S Havlin, C-K Peng, and M Simons. Long-range power-law correlations in condensed matter physics and biophysics. *Physica A: Statistical Mechanics and its Applications*, 200(1-4):4–24, 1993.
[23] Howard G Tucker. A generalization of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, 30(3):828–830, 1959.
[24] Aad Van der Waart. Asymptotic statistics, 1998.
[25] Wei Biao Wu, Yinxiao Huang, and Wei Zheng. Covariances estimation for long-memory processes. *Advances in Applied Probability*, 42(1):137–157, 2010.
[26] Cao Xiao-Feng, Deng Zong-Wei, and Yang Chun-Bin. On origin of power-law distributions in self-organized criticality from random walk treatment. *Communications in Theoretical Physics*, 49(1):249, 2008.
[27] CB Yang. The origin of power-law distributions in self-organized criticality. *Journal of Physics A: Mathematical and General*, 37(42):L523, 2004.