Journal of Physics: Conference Series **475** (2013) 012006 doi:10.1088/1742-6596/475/1/012006

Data integration, systems approach and multilevel description of complex biosystems

Enrique Hernández-Lemus

Computational Genomics Department, National Institute of Genomic Medicine (INMEGEN), Periférico Sur 4809 Col. Arenal Tepepan, Tlalpan, 14610 México City, México.
Also at Center for Complexity Sciences, National Autonomous University of México (UNAM), Torre de Ingeniería, Circuito Escolar s/n, 04510, México City, México

E-mail: ehernandez@inmegen.gob.mx

Abstract. Recent years have witnessed the development of new quantitative approaches and theoretical tenets in the biological sciences. The advent of high throughput experiments in genomics, proteomics and electrophysiology (to cite just a few examples) have provided the researchers with unprecedented amounts of data to be analyzed. Large datasets, however can not provide the means to achieve a complete understanding of the underlying biological phenomena, unless they are supplied with a solid theoretical framework and with proper analytical tools. It is now widely accepted that by using and extending some of the paradigmatic principles of what has been called *complex systems theory*, some degree of advance in this direction can be attained. We will be presenting ways in which by using data integration techniques (linear, non-linear, combinatorial, graphical), multidimensional-multilevel descriptions (multifractal modeling, dimensionality reduction, computational learning), as well as an approach based in systems theory (interaction maps, probabilistic graphical models, non-equilibrium physics) have allowed us to better understand some problems in the interface of Statistical Physics and Computational Biology.

1. Introduction

Living matter is characterized by an entangled web of complex physicochemical processes. The origin and nature of these phenomena is extremely disparate. There are many space and time scales involved, but also a multitude of information layers intertwined, sometimes even in recursive ways. In view of this extremely involved conundrum, one is to ask *How to disentangle this plethora of complex processes as to acquire some actual understanding of the mechanisms underlying biological function?*.

Intending to answer this question, contemporary biology -in particular molecular biology, genomics and proteomics, as well as electrophysiology and imaging- is entering a new era; one characterized by extremely fast advances in its capacity of generating data in huge amounts; looking to dig-in deeper into the molecular basis of biological function. Research in biology is now turning from purely descriptive accounts to process-driven, multi-scale analyses now commonly termed *systems biology*. The rising paradigm is thus the study of biological phenomena as *complex systems*. In these analyses, an integrative vision becomes mandatory, since data alone are insufficient to create a *real* understanding of the complex processes occurring in living

matter. It is not completely clear how to accomplish such a task. However, research in complex biological systems has so far revealed foundational principles such as self-organization, criticality, and robustness, but the biological implications of these phenomena are still to be revealed to a full extent.

In what follows, we will be presenting different aspects of the application of the methods of complex systems theory to problems in biological phenomena. We will be doing this on a case study basis. Section 2 deals with the first case, data integration in cancer genomics. In Section 3 we analyze the second case related with a system's approach to the coupling between metabolism and transcriptional anomalies in the onset of breast cancer. Section 4 is devoted to the general problem of dimensionality reduction -third case- as exemplified by whole genome gene expression analysis, a quite common computational biology study. The fourth case is presented in section 5 and it elaborates on multidimensional models, in particular applied to multichannel electrophysiology. Finally, Section 6 presents some brief conclusions and perspectives.

2. Case I: Data Integration

Modern high-throughput genomic experiments have paved the way to a large-scale characterization of living organisms. This usually involves the generation and, more importantly, the interpretation of data at an unprecedented scale. Computational tools and mathematical algorithms have been created in order to integrate, organize and also *mine* the gargantuan wealth of information generated. Technologies for the detection of several kinds of genomic alterations have been developed and applied to analyses of almost any living organism, even to cancer genomes. Cancer research in particular, has proven that studies based on a single technology platform result extremely limited in scope when compared with the extent of knowledge that can be acquired when using different platforms all together. For this reason, there is a need for systematic methodologies to facilitate data management, visualization and integration. The goal of these methodologies should be to permit a proper analysis of the biological implications for the findings, without sacrificing mathematical and statistical rigour and computational efficiency.

2.1. Data integration schemes in genomics

With view to the construction of an integral view of genomic alterations, a data driven combinatorial approach has been proposed [1]. It is based on the enumeration of all possible genomic alterations scenarios that may be present in a N-platform integrative analysis relying on a so-called *three-state model* applied to the set of statistically significant genes (3-MDI). Each scenario is represented as a sequence $S_1, S_2, \ldots S_N$ of states, where S_k denotes the state of a gene for platform k. Each state is defined to take values in {-1, 0, 1} -akin to spin-1 models in lattice ferromagnetism- interpreted as {`Down, NoChange, Up`}. This list represents the universe of hypotheses that describe structural variations in the genome as well as transcription activity in coding and non-coding regions. Hypotheses can be chosen for their clear biological relevance but also for their quantitative importance. We may find that a large set of genes follow a particular scenario or that genes commonly share a set of more specific scenarios leading to other important questions to be answered.

Under the 3-MDI algorithm [1], every dataset-platform is subjected to an independent low-level analysis, genes are then classified according to the proper significance levels for each technology while careful attention is given to the wide spectrum of dynamic ranges between these different data sources. The platforms $\mathcal{P}$ selected in [1] included {`RNA, miRNA, Methylation, NGS, CNV`} [Figure 1], from these, a list of highly relevant genes was generated in each set. Genes in every list were coded according to two of the three states in the three-state model. In every list of top-m_k genes can be either up or down. All k lists

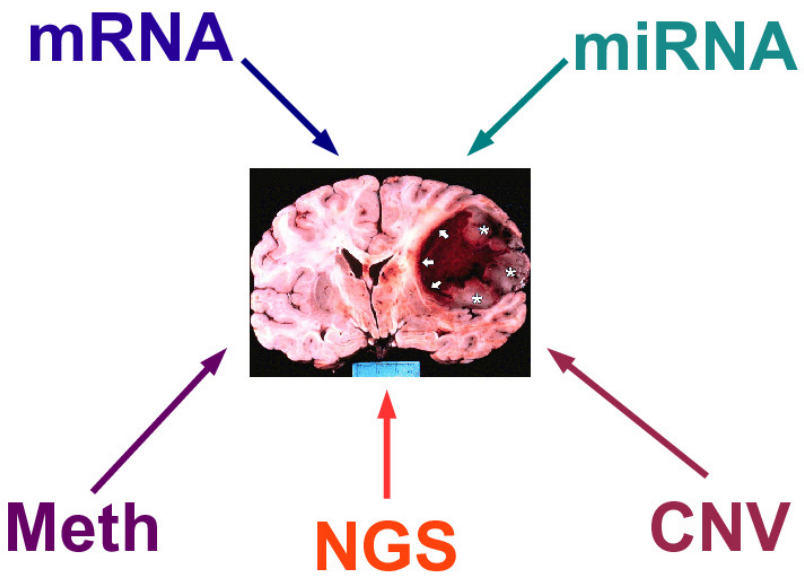

Figure 1. Multiple platform whole-genome analysis sheds light in Glioblastoma Multiforme via **data integration** [1].

are combined and basic set theory $P_i \bigcup P_j = (P_i \bigcap P_j) \bigcup (P_i \backslash P_j) \bigcup (P_j \backslash P_i)$ adds zeros when $(P_i \backslash P_j)$ indicates a zero in P_j. The combinatorial analysis can classify all possible scenarios searching for genes in 2,3,4, or all 5 platforms simultaneously. Under this approach there exist 3^k possible scenarios for a k-platform analysis assuming a three-state model. The discretization scheme proposed in 3-MDI is a finite alphabet classification scheme. This means that all of the possible states of the system correspond to a realization within the given lexicon (here the states are labeled -1 meaning subregulated, hypomethylated or deleted, 0 meaning no change and +1 meaning over-expressed, hypermethylated or duplicated (multiplicated)). Finite alphabet classification schemes have been shown to be equivalent to a class of machine learning algorithms called *Supervised learning* [2], so that that most (in theory, all) of the supervised machine learning techniques and algorithms could be applied to train our three-state model with either experimental or simulated data. This will be useful when assessing the findings of integrative genomics studies by experimental sources or with synthetic data.

2.2. Multidimensional plots

Visualization of large-scale data becomes a key aspect of the analysis, it allows to distinguish possible biological hypotheses. Important tool in this regard are multidimensional plots. *Circular multilayered plots* (a.k.a. *Circos* plots) for instance, are useful visualization aids in which a series of concentric rings are used to render the data. The circular layout is ideal for showing how different positions within our data domain relate to one another. This relationship can be quantitative, binary or any other classification scheme. Modifying the rings or *ribbons* that represent relationships, the progression and orientation of the circular segments, and so on, allow us to show a large number of connections between data points (or positions) in a clear cut and *intelligible* lay out. Circos plot have helps us to identify the most

mutated chromosomes, identifying genes in each chromosome and hence combine the differential expression and methylation for the set under study as it can be seen in, for instance in Figure 2 and Figure 3 of reference [1]. The advantage of using tools such as circos plots in managing high throughput data is that they are able to be easily incorporated into data acquisition, analysis and reporting pipelines. In the field of high throughput genomics and proteomics, the usual setting is that a data pipeline is implemented as a multi-step process. On this process, data is analyzed by multiple independent tools, each passing their output as the input to the next step. Another, more comprehensive example of the afore-mentioned *data acquisition and analysis pipeline* using the multidimensional plot philosophy (and in particular circos plots) is COSMIC (the Catalogue Of Somatic Mutations In Cancer) [4, 5]. The project page states that *"...There is a vast amount of information available in the published scientific literature about these changes. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers..."*.

2.3. Why, what and how?

- **Why?:** Because biological systems are intrinsically multidimensional; they are characterized by many layers of information and many levels and scales of description. A single reductionist scheme fail to capture the essence of living phenomena.
- **What?:** Here we propose novel data visualization schemes in order that some hidden connections (e.g., between *trans*- gene regulation by microRNAs and *cis*- DNA methylation as coordinated mechanisms of expression) could be unraveled.
- **How?:** In this particular case [1], we used two different and to a certain extent complementary strategies for data integration. In the one hand we performed a combinatorial classification based in a three state model for different data classes. In a way this would function as a discretization scheme akin to, say, a spin-1 lattice study in ferromagnets. In the other hand, we made use of multidimensional plots to show and reveal in a *geometrical* way, some unexpected connections between these discrete states.

2.4. What do we learn?

According to the popular belief in gene regulation theories, the effect of high DNA methylation levels in a gene region is to diminish the expression of the genes in such region. Hence, a *questionable* subset in our whole genome studies of *Glioblastoma multiforme* was the one with copy number loss yet marked as hypermethylated and over-expressed (see Table 1 of reference [1]). In the case of some genes (like RRM2 and CNGA3) that are hyper-methylated their over-expression could be explained by copy number amplification of the gene. However, in other cases, the expression level may also be affected by miRNA regulation as it seems to be the case of CD74 which presents no-significant DNA copy number variations. This result seem to be counterintuitive, and even controversial according to the current biological understanding of gene regulation phenomena, but it is present in many samples within the Glioblastoma multiforme dataset studied (around 500 samples were analyzed), so we can be pretty sure it is not an artifact of the analysis, but rather a new side of gene regulation yet to be completely understood.

3. Case II: Systems Approach

3.1. An archetypal example: Cancer Genomics and Metabolism

Thermodynamic studies at the transcriptional, epigenetic, and metabolic levels have pointed out to energetics as playing a non-trivial role in the onset and development of malignancy. We will focus on the relationship between transcriptional de-regulation of a set of genes that present both transcription factor (TF) and metabolic activity while at the same time have been associated with the presence of breast cancer. In order to unveil its regulatory and thermodynamical

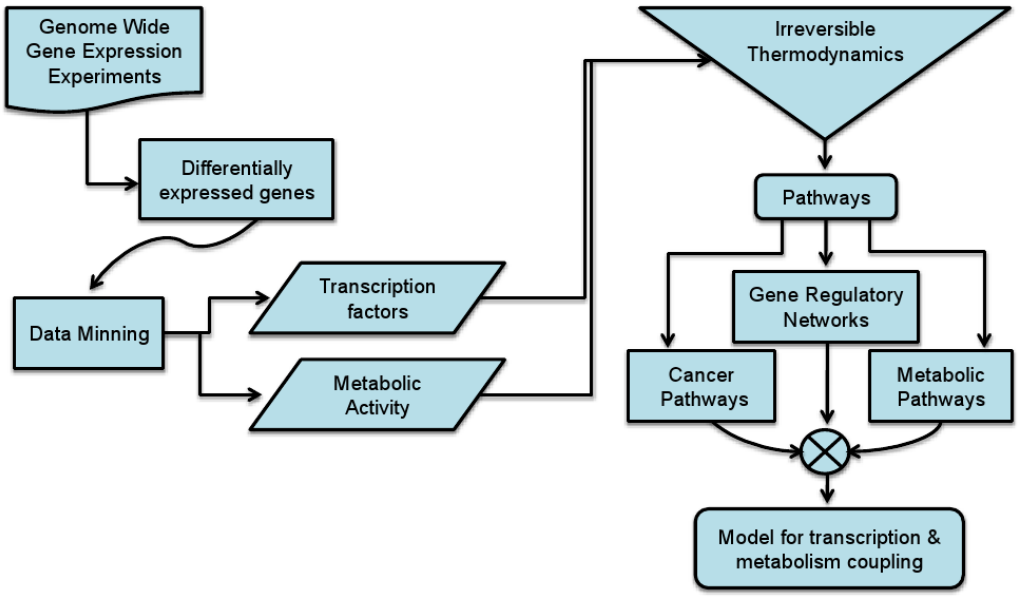

Figure 2. A **System's approach** considering both the mesoscopic thermodynamics and the molecular transcriptional regulation interactions, provide a deeper understanding of the relationship between cancer and metabolism [6].

behavior by means of gene expression data obtained from genome-wide analysis experiments in RNA from biopsy-captured tissue of both primary breast cancer and normal breast.

The role of gene interaction networks have also been extensively mentioned in relation to cancer phenomenology, it has been claimed that these network effects are, in fact much more important than individual gene contributions. Some of these networks are indeed related to energetic and metabolic processes, tyrosine-related deregulation, and immunity weakening. One usually think of tumor cells as having successful mechanisms to evade normal control and cell regulation of proliferation and apoptosis. Alterations in gene expression have become a better (but far from completely) understood component of normal development and disease progression. TFs have become a promising target for therapy. In brief, gross alterations in TF regulation would result in cascade triggering affecting both the whole cell cycle and the metabolic activity thus resulting in possible development of cancer. Many people have come to conclude that *cancer* is a *transcriptional disorder* disease, while, as we have mentioned other authors have recently turned their attention to the metabolic and energetic component, hence a possible connection between these two approaches could be found in the *energetic deregulation* $\rightarrow$ *transcriptional disorder* leading both to cascade triggering and metabolic disorders related to neoplasm formation and development. For these reasons it will result fruitful to model the role of TFs at both the energetics (thermodynamic) level and the network approach [6] [for more details of the whole system's approach taken refer to Figure 2].

3.2. How does the coupling occurs?

The complex dynamics behind even relatively simple models of transcription [16] demonstrate the necessity for a non-equilibrium thermodynamical characterization that includes the possibility to deal with fluctuations in small systems [15]. Systems outside the realms of the thermodynamic limit are characterized by large fluctuations, hence stochastic effects are to

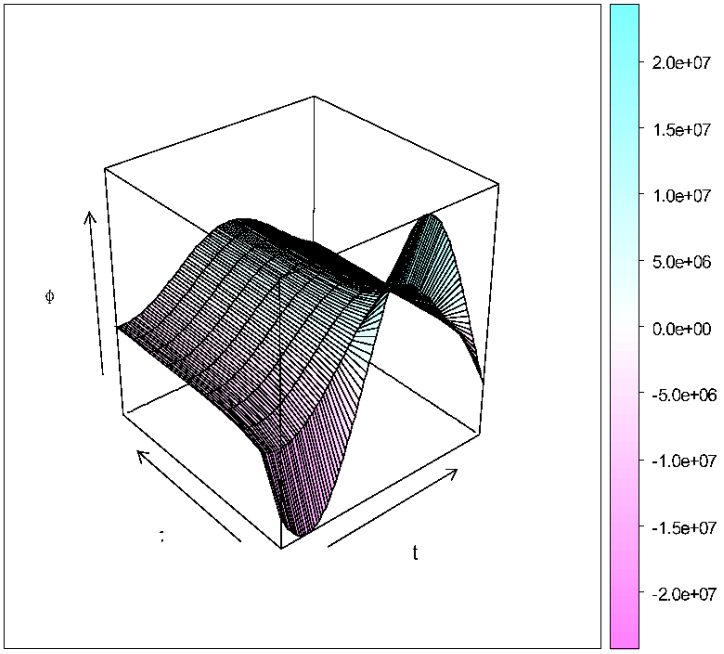

Figure 3. The coupling between energy influx and transcriptional bursting is shown to occur at low relaxation times [7].

be considered [7]. One particular instance in which stochastic effects take importance is the so-called *transcriptional burst* scenario [see Figure 3].

3.3. Why, what and how?

- **Why?:** Because, although it has been known for decades that energetic deregulation occurs in cancer cells and that it is responsible to a certain extent for neoplastic transformation and metastasis, no detailed physicochemical study was performed at the molecular and subcellular level to a whole genome extent, in order to understand the thermodynamic basis of transcriptional anomalies.
- **What?:** We develop a non-equilibrium thermodynamic formalism to study transcriptional regulation [15, 16] then we show how specific details of the energy influx may be taken into account [7]. We are applying such ideas to a specific network-based whole genome study of gene expression profiles in breast cancer samples [6] .
- **How?:** By taking into account, on one side, non-equilibrium coupling between subcellular metabolism and transcription via the related chemical potentials. We also considered the role that *transcription factors* and in particular *master regulator genes* have in the complex energetic landscape behind such couplings.

3.4. What do we learn?

Most of it is still work in progress. However, a core of four genes (MNDA, POU2AF1, SMAD3 and specially MEF2C) seem to be revealing as novel master regulators in transcriptional bursting

associated with breast cancer. It is specailly encouraging since homologues of these genes (in particular of MEF2C, a family called the MADS genes) have been proven to be playing a key role in regulation between proliferation and differentiation in plants via processes akin to those correspondent in animal cells.

4. Case III: Dimensionality Reduction

Gene expression analysis is probably one of the most fertile fields to develop new biological physics theories and techniques from the theory of complex systems. The complex mechanisms behind whole genome transcriptional regulation has attracted many physicists, for the subtleties of the associated physicochemical mechanisms, for the intricacy of the implicit stochastic dynamics and also for the challenges presented in the analysis and inference of regulatory networks, both from the mathematical and computational standpoints. We have gained great insight into different aspects of such phenomena [8, 9, 10, 11, 12, 13, 14, 15, 16]. In order to do so, however researchers need to possess strong computational and mathematical skills, such as those developed in computational physics, statistical mechanics and high energy physics to successfully manage such enormous amounts of data.

With regards to the dimensionality problem, the question is that whole genome expression experiments for higher species comprise several tens of thousands of variables (namely, the mRNA concentration or gene expression for each gene) and for technical, economic, and logistical reasons, the number of experimental samples runs on the order of a few hundreds at most. As a result, it is not feasible to use classical statistical approaches. Instead, some tools developed and used in statistical physics have been applied to this end, ranging from information theoretical approaches [8, 9, 17] to maximum entropy techniques [10, 19].

4.1. The Compressive sampling scenario

The basic aspects of the mathematical theory and the computational implications of a recently developed technique called Compressive Sampling, as well as some possible applications within the scope of Computational Genomics, and Computational Biology in general are considered. The central idea is that most of the information sampled from the experiments turns out to be discarded (for being non-useful) in the final stages of biological analysis, hence it would be better if we could find an algorithm to remove selectively such information in order to get rid of the computational burden associated with processing and analyzing such huge amounts of data. Here we show that a possible algorithm for doing so it is precisely Compressive Sampling [20].

4.2. Why, what and how?

- **Why?:** Because current (and even more future) datasets in genomics -and also in other instances, such as proteomics, electrophysiology, high energy physics, etc.- are huge (several Gigabytes in length often, even some Terabytes now and very likely in the Petabyte range in the near future) and at the same time they are extremely noisy and redundant, so that most of this wealth of data is NOT recovered as usual information in the final stages of analysis. So, as of today its data management is a burden and they constitute in important challenges in the analysis stages.
- **What?:** We propose as a proof-of-concept the implementation of non-linear data compression techniques at the acquisition stage in the case of whole genome gene expression experiments for a thousand-plus sample dataset. The outlined implementation is based in the algorithm known as *Compressive Sampling*.
- **How?:** Compressive sampling is an algorithm able to cope and defeat with usual limitations in data management and information recovery as given by Shannon-Nyquist sampling

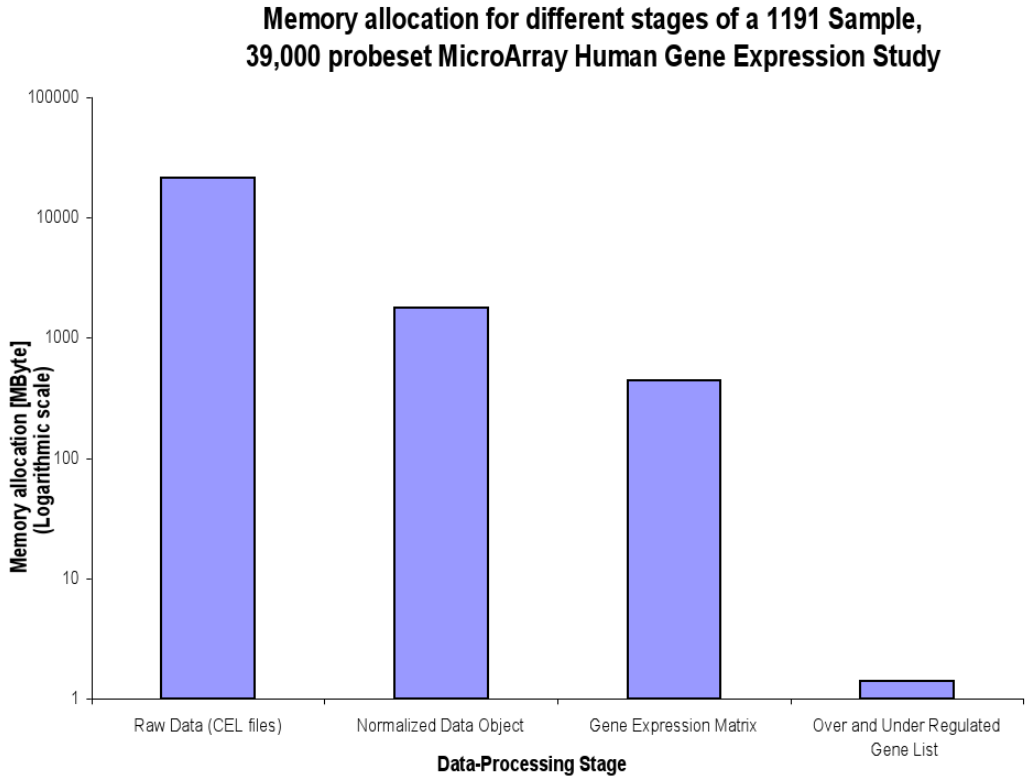

Figure 4. Memory allocation reduction in a 1191-sample whole genome gene expression analysis as depicted in [20]. Notice the logarithmic scale on y-axis.

theorem. It does so by using the fact that many large datasets that we a priori know that are compressible (i.e. redundant) possess a *sparse* matrix representation in some basis -in general orthonormal- so that l_0-norm and l_1-norm based minimizations are computationally tractable optimization problems yielding a singular value decomposition (SVD) submatrix much smaller than the original. Compressive sampling can be used even in random matrices so we can be dealing with noisy datasets.

4.3. What do we learn?

In principle, we are answering the question: how to get all the relevant information contained in a huge analysis dataset, say a 21 Gb raw gene expression experimental database [20] by *just* analyzing much smaller objects, in the mentioned case a 1.4 Mb matrix [see Figure 4]? This could be done after applying compressive sampling optimization and SVD of the original dataset. Sparsity of the original dataset matriz representation then plays a fundamental role in how well we can estimate signals in the presence of noise (i.e. how well we deal with *shrinkage*, and *soft-thresholding*), but also in compressibility by means of transform coding and in the ultimate case in how well we solve inverse problems (say gene network reconstruction [8, 9, 17, 19]). This si so because dimensionality reduction facilitates modeling via simple models/algorithms.

In brief, due to noise-uncertainty we can use the l_1 norm as a proxy for sparsity and hence use nonlinear sampling by convex programming instead of traditional (linear) Shannon sampling, thus largely reducing the need for individual samples. This is a very good deal if we recall that getting, processing and analyzing, a large number of good quality samples is a common burdensome challenge in biomedical research.

5. Case IV: Multidimensional models

5.1. Synchronization of spontaneous spinal cord potentials

The analysis of the interaction and synchronization of relatively large ensembles of neurons is fundamental for the understanding of complex functions of the nervous system. It is known that the temporal synchronization of neural ensembles is involved in the generation of specific motor, sensory or cognitive processes. Also, the intersegmental coherence of spinal spontaneous activity may indicate the existence of synaptic neural pathways between different pairs of lumbar segments. In this study we present a multichannel version of the detrended fluctuation analysis method (mDFA) to analyze the correlation dynamics of spontaneous spinal activity (SSA) from time series analysis. This method together with the classical detrended fluctuation analysis (DFA) were used to find out whether the SSA recorded in one or several segments in the spinal cord of the anesthetized cat occurs either in a random or in an organized manner [22].

5.2. Why, what and how?

- **Why?:** Within biological physics, we are often confronted with the non-linear analysis of the correlation structure for multidimensional time series with fractal or multi-fractal support. Most biosignals possess a complex and subtle arrangement of patters, most of them hidden behind a noisy and apparently random envelope. In some cases, for instance, synchronization phenomena are disguised within the multifractal spectrum and it is almost impossible to find them and analyze them properly.
- **What?:** In order to cope with this problem to a certain extent we are generalizing the well known detrended fluctuations analysis (DFA) of time series to deal with multidimensional signals arranged in the form of a time series vector. In this way, instead of analyzing a vector arrangement of Hurst exponents we only have single spectrum to analyze. However this spectrum contains the entire correlation structure of the multidimensional time series [see Figure 5].
- **How?:** Since, in our particular case [22] we were studying the central nervous system as an intrinsically synchronized dynamic system, as well as the role of synchronization of multiple neuronal areas in information processing, it was evident the need for a nonlinear correlation analysis that involves time series of *several inputs*. That is, to assess the long range correlation structure of *multichannel data* as a single entity. To that end we developed an *inner product*-based metric to consider the relation between the Hurst exponent and the singularity spectrum as a dual transformation. As consequence we now have a componentwise self-similar process whose vector 2-norm is taken as a time series metric and then the Hurst spectrum is calculated as in classic DFA [22].

5.3. What do we learn?

By using mDFA we found a significant synchronicity between potentials recorded from pairs of segments either rostrally (up-stream) or caudally (down-stream) to a spinal lesion. This may represent evidence for the hypothesis that left and right groups of neurons in different segments are both interconnected by pathways running in the same as well as in the opposite sides, an important recent paradigm in neurophysiology. Changes of the fractal correlation

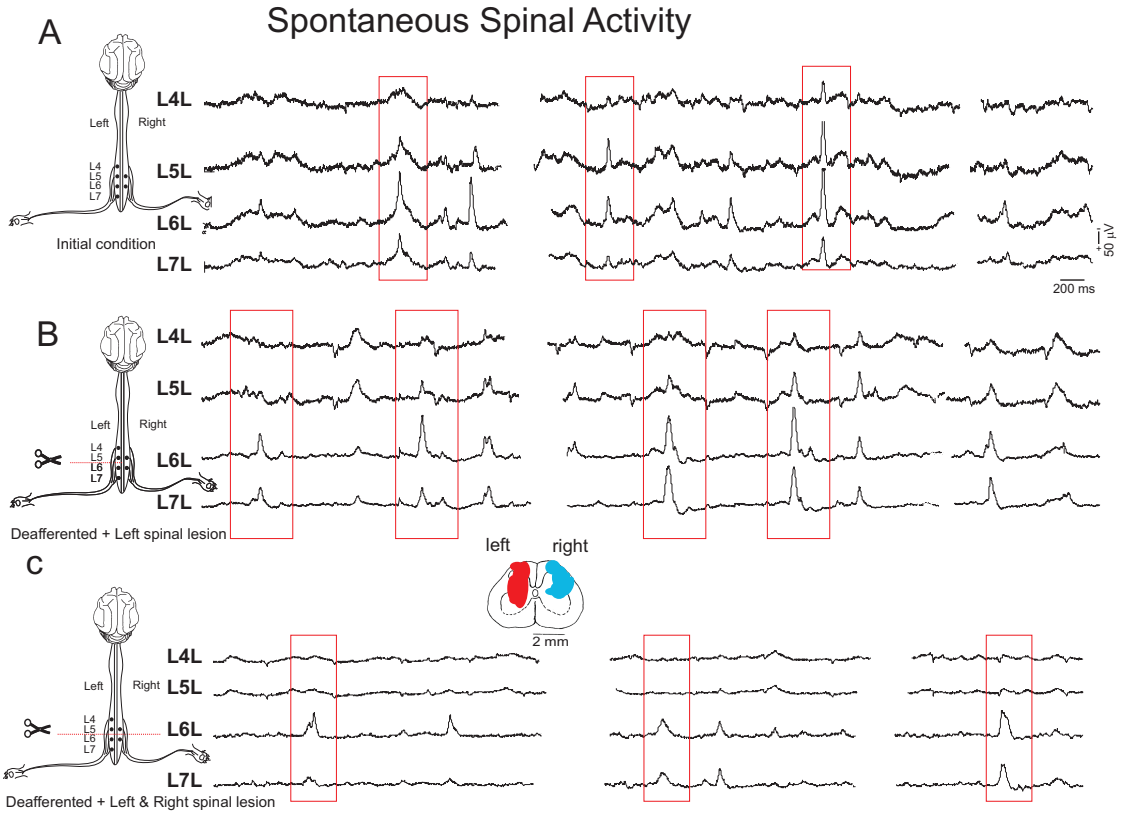

Figure 5. Synchronicity in the Spontaneous Spinal Activity can be accounted for by means of a multidimensional non-linear correlation analysis [22].

structure are believed to have a functional origin since it has been proposed that the segmental synchronizations of dorsal root reflexes and dorsal root potentials are mediated by the spinal cord connections through the lateral funiculi. In brief, it seems that not only brain neurons are connected in a complex network fashion but also spine neurons present network-like behavior in their correlation structure. mDFA also revealed an already envisioned temporal synchronization in the activity of multiple segmental signals, thus revealing spinal communication as a complex coupled phenomenon. We could also study how the intact neuroaxis and spinal damage caused both decorrelation and desynchronization of SSA signals.

6. Conclusions and perspectives

Data integration, multilevel analysis and other methods of complex system's theory and statistical physics have proved to be useful tools in the analysis of complex biological phenomena and are a means to acquire deeper understanding of the underlying phenomena. The role that such disciplines have acquired in the biological sciences is growing in importance [24], as is evident, in numerous research efforts in biology (see for instance reference [25]) and in the creation of the Office of Physical Sciences Oncology in the National Cancer Institutes in the U.S. [26] among a myriad other projects involving methods of statistical physics on one side and biology on the other. This mini-review presents the application of some of these ideas in our own

research, namely in the case of data integration in cancer 'omics [1], system's biology approach to the metabolic-transcriptional coupling in breast cancer [6, 7], high throughput data management and compression in computational biology [20] and multidimensional correlation analysis in computational neuroscience [22]. The presentation here is by no means exhaustive. Further, deeper understanding may be obtained by consulting the original sources in [1, 6, 7, 20, 22] and references therein.

Acknowledgments

We gratefully acknowledge support by grant: PIUTE10-92 (ICyT-DF) [Contract 281-2010], as well as federal funding from the National Institute of Genomic Medicine (México). We also thank all of our collaborators and students involved in the reviewed works.
Disclaimer: Figures 2-5 have been used under the *Creative Commons* license of their corresponding original sources (either BY or C0).

References

[1] Baca-López, K, Correa-Rodríguez, M D, Flores-Espinosa, R, Garcia-Herrera, R, Hernández-Armenta C I, Hidalgo-Miranda A, Huerta-Verde A J, Imaz-Rosshandler I, Martínez-Rubio A V, Medina-Escareño A, Mendoza-Smith R, Rodriguez-Dorantes M, Salido-Guadarrama I, Hernández-Lemus E and Rangel-Escareño C 2011 A three-state model for multidimensional genomic data integration *Proceedings of the ninth international conference for the Critical Assessment of Massive Data Analysis CAMDA*
[2] Vapnik V N 2000 The Nature of Statistical Learning Theory 2nd ed. (Springer Verlag)
[3] http://circos.ca
[4] http://www.sanger.ac.uk/genetics/CGP/cosmic/
[5] Forbes S A, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague J W, Futreal P A and Stratton M R 2008 The Catalogue of Somatic Mutations in Cancer (COSMIC) *Current Protocols in Human Genetics* 57:10.11.110.11.26.
[6] **Baca-López K, Hidalgo-Miranda A, Mayorga M, Gutiérrez-Nájera N and Hernández-Lemus E *Non-equilibrium thermodynamics and network analysis of transcription factor regulation in breast carcinomas*, PLoS Biology (in preparation)**
[7] Hernández-Lemus E and Correa-Rodríguez M D 2011 Non-equilibrium hyperbolic transport in transcriptional regulation *PLoS ONE* **6**(7): e21558 http://dx.plos.org/10.1371/journal.pone.0021558.
[8] Margolin, A A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla-Favera R and Califano A 2006 ARACNe: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context *BMC Bioinformatics* **7**(Suppl I), S7 doi:10.1186/1471-2105-7-S1-S7
[9] Bansal M, Belcastro V, Ambesi-Impiombato A and di Bernardo D 2007 How to infer gene networks from expression profiles *Molecular Systems Biology* **3** 78
[10] Berg J 2008 Dynamics of gene expression and the regulatory inference problem *Europhysics Letters* **82** 28010
[11] Berg J 2008 Out-of-equilibrium dynamics of gene expression and the Jarzynski equality *Physical Review Letters* **100** 188101
[12] Berg J and Stauffer D 2009 Adaptive gene regulatory networks *Europhysics Letters* **88** 48004
[13] Benecke A 2008 Gene regulatory network inference using out of equilibrium statistical mechanics *Human Frontier Science Program Journal* **2**(4) 183-88
[14] Sánchez A and Kondev J 2008 Transcriptional control of noise in gene expression *Proceedings of the National Academy of Sciences, USA* **105**(13) 5081-86
[15] Hernández-Lemus E 2009 Non-Equilibrium Thermodynamics of Gene Expression and Transcriptional Regulation *Journal of Non-Equilibrium Thermodynamics* **34**(4) 371-394
[16] Hernández-Lemus E 2010 Extended Irreversible Thermodynamics of Gene Regulation *Journal of Non-Newtonian Fluid Mechanics* **165** 1029-32
[17] Baca-López K, Hernández-Lemus E and Mayorga M 2009 Information-theoretical analysis of gene expression data to infer transcriptional interactions *Revista Mexicana de Física* **55**(6) 456-66
[18] Subramanian A, Tamayo P, Mootha K V, Mukherjee S, Ebert B L, Gillette M A, Paulovich A, Pomeroy S L, Golub T R, Lander E S and Mesirov J P 2005 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles *PNAS* **102**(43) 15545-50
[19] Hernández-Lemus E, Velázquez-Fernández D, Estrada-Gil J K, Silva-Zolezzi I, Herrera-Hernández M F and Jiménez-Sánchez G 2009 Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms *Physica* A **388** 5057-69

[20] Hernández-Lemus E 2011 On the Application of Compressive Sampling Techniques to High Throughput Data in Computational Genomics *Theoretical and Applied Informatics* **23** 3
[21] D'haeseleer P 2005 How does gene expression clustering work? *Nature Biotechnology* **23** 12
[22] Rodríguez E E, Hernández-Lemus E, Itzá-Ortíz B A, Jiménez I and Rudomín P 2011 Multichannel detrended fluctuations analysis reveals synchronized patterns of spontaneous spinal activity in anesthetized cats *PLoS ONE* **6**(10) e26449 http://dx.plos.org/10.1371/journal.pone.0026449.
[23] Segal E, Friedman N, Koller D and Regev A 2004 A module map showing conditional activity of expression modules in cancer *Nat Genet* **36**(10) 1090-98
[24] Hernández-Lemus E 2011 Biological Physics in Mexico: Review and New Challenges *Journal of Biological Physics* **37**(2) 167-84
[25] Bisson G, Bianconi G and Torre V 2012 The dynamics of group formation among leeches *Frontiers in Physiology* **3** 133 doi: 10.3389/fphys.2012.00133
[26] http://physics.cancer.gov/