Bio*física* M a g a z i n e

COOL BIOPHYSICS

Systems biophysics

Jose M. G. Vilar, *Biofisika Institute (CSIC, UPV/EHU) and Ikerbasque, Bilbao (Spain)*

Mainstream biophysics has traditionally focused on biological systems as *single entities*, such as a macromolecule, a membrane, a cell, or a tissue. The objective is typically to study physical properties of the system, such as force-extension curves of macromolecules or elastic properties of cells, or to use physical approaches to obtain information about biologically relevant properties, such as the structure of macromolecular complexes. This single-entity view of biophysics that has proved to be so prolific, however, cannot capture the origins of *emergent behavior*. Systems biophysics, in contrast, emphasizes the focus on how the system properties emerge from the relations between constituent elements (Saiz & Vilar, 2006a). These types of approaches are needed, for instance, to study how mutations affect the molecular properties of the cellular components; how the mutated components affect different signaling pathways; and how these modified pathways confer cell-growth advantages during tumor progression and metastasis (Vilar & Saiz, 2013a).

Systems biophysics is not a new field *per se*. The study of emergent behavior in terms of the properties of the components has led to historical breakthroughs. A most notable example is the work of A. L. Hodgkin and A. F. Huxley on the ionic mechanisms underlying the initiation and propagation of action potentials in the squid giant axon (Hodgkin & Huxley, 1952), for which they were awarded the **Nobel Prize in Physiology or Medicine in 1963**. After a series of experiments, Hodgkin and Huxley developed a circuit model that was able to capture how the squid axon carried an action potential in terms of the electrical properties of the cell membrane, voltage-gated conductivities for different ions, and electrochemical gradients. This model has been exceptionally successful, not just in describing but also in predicting a large number of neuronal properties, to the extent that modern investigations have confirmed many aspects of the model that were assumptions at the time.

Two main types of challenges

There have been many developments since the pioneering work of Hodgkin and Huxley. What makes systems biophysics approaches so relevant today is the need, and the opportunity, to make sense of the data obtained in two complementary fronts: new sources of high-precision data and massive amounts of data.

On the *precision-data front*, there have been tremendous advances in the cellular imaging field that can couple cellular responses and perturbations to precise measurements of the intracellular state (Wartlick, et al., 2011). Many of these advances arose from the advent of fluorescent-protein reporters, which allow us to precisely correlate molecular events on real time with behavior at the single cell level. These technologies include, among many others, quantitative time-lapse fluorescence microscopy, fluorescence/Förster resonance energy transfer (FRET), fluorescence recovery after photobleaching (FRAP), fluorescence correlation spectroscopy (FCS), and single molecule imaging. They have been used to estimate quantities such as diffusion and transport coefficients of cellular components, binding kinetics, cellular localization, lifetimes of intracellular interactions, and stochastic fluctuations in the number of components (Sung & McNally, 2011). There have also been substantial advances in structural biology and in single molecule biophysics that have provided us with an atomic level description of many of the cellular components. These types of advances have increased the quality of the data, which we now have at hand to unprecedented levels. Yet, most of these data remain disconnected from each other and it is up to systems approaches to put them together into a functional description that could indicate how the system functions as a whole.

On the *massive-data front*, there have been major breakthroughs in automated technologies for the collection of data. These range from traditional proteomics and genomics analyses to high-throughput single-cell analyses (Aghaeepour, et al., 2013), such as multichannel flow cytometry (FCM), to new genome-wide functional screens, including RNA interference and diverse types of CRISPR screens. A most prominent example of our far-reaching abilities for gathering information is single-cell genome sequencing (Gawad, Koh & Quake, 2016). These automated technologies have brought the cartoon-like representations of cellular processes to exponentially growing webs of nodes and links that seem as close to completion as ever. The complexity of the emerging picture, however, makes it clear that all this information by itself is not sufficient to truly understand complex processes. In order to piece back together the experimental information into physiologically relevant descriptions, one needs constructive methods (Vilar, Guet & Leibler, 2003). Systems biophysics approaches have emerged as a promising tool for transforming molecular detail from different sources into a more integrated form of understanding complex behavior. I discuss below two examples of these two types of challenges.

The *lac* operon: a not-so-simple paradigm of gene regulation

The *E. coli lac* operon is the genetic system that regulates and produces the enzymes needed to metabolize lactose, including a lactose sensor (the *repressor*), a lactose transporter (the *permease*), and an enzyme that breaks lactose into simpler sugars (the *β-galactosidase*). It has been a paradigm in genetics since F. Jacob and J. Monod used it over 50 years ago to put forward the very basic principles of gene regulation (Jacob & Monod, 1961), for which they received the **Nobel Prize in Physiology or Medicine in 1965**. They postulated the existence of molecules that bind to specific sites in nucleic acids to control the expression of genes. In the lac operon, the response to lactose is controlled by the lac repressor, which can bind to the main *operator* and prevent the RNA polymerase from transcribing the genes. When lactose is present, however, this binding is strongly reduced and transcription can take place. This leads to the production of the β-galactosidase and the permease codified in the *lacZ* and *lacY* genes (Müller-Hill, 1996). The original idea of the lac repressor preventing transcription has been refined over the years to incorporate a complex hierarchy of events that extend from specific protein-DNA interactions to the combinatorial assembly of nucleoprotein complexes (Vilar & Saiz, 2013a).

Figure 1. Modular deconstruction of the lac operon and predictive accuracy. Components: the basic modules are the lac repressor dimers (n2) and operators (Oj). Interactions: repressor dimers interact with other dimers and with operator DNA to form tetramers in solution, to bind single DNA sites as dimers or tetramers, and to loop DNA as tetramers bound to two DNA sites. Just a few parameters are needed to characterize these interactions. System: The main, O1, and the two auxiliary, O2 and O3, operators are shown as yellow rectangles on the black line representing DNA. Binding of the lac repressor to O1 prevents transcription of the three lacZYA genes. The predicted vs. the observed repression level Vilar & Saiz, 2013b is plotted for all the available operator configurations, repressor oligomeric forms, and promoters (panel on the bottom left). The dashed lines represent a factor 1.7 higher and lower than the perfect prediction indicated by the continuous line. Adapted with permission from Vilar & Saiz, 2013b. © 2013 American Chemical Society.

During this time, it has become evident that systems biophysics approaches are needed to tackle the complexity of the molecular interactions in the control of the response to lactose. This complexity is already present in the mode of functioning of the lac repressor, which upon binding to *O1*, the main operator, prevents the RNA polymerase from binding to the promoter (Saiz & Vilar,
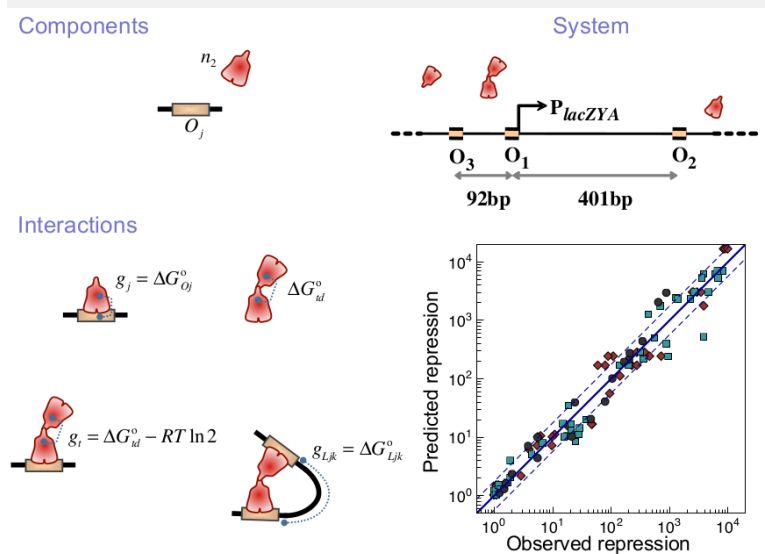
2006b). There are also two distal auxiliary operators, *O2* and *O3*, where the repressor can bind specifically without preventing transcription (**Figure 1**). These two additional sites were originally considered to be remnants of evolution, because they are orders of magnitude weaker than the main site and by themselves do not affect transcription substantially. In combination with the main site, however, they were shown to increase repression of transcription by almost a factor of 100. For over 20 years after the characterization of these sites, a long-standing question was how such weak sites could help the binding to a strong one. The reason for this counterintuitive effect turned out to be that the lac repressor can also bind as a bidentate tetramer to two operators simultaneously and loop the intervening DNA. Binding while *looping DNA* is difficult to analyze with traditional biochemical methods and required new biophysical approaches to characterize it (Vilar & Leibler, 2003).

> **Despite the apparent simplicity of the lac operon, it took over 50 years to have an effective biophysical characterization of this system**

This type of behavior, involving oligomeric transcription factors that can bind simultaneously single and multiple DNA sites, is a recurrent theme in gene expression, to the extent that transcription regulation through DNA looping is nowadays considered to be the rule rather than the exception (Alberts, et al., 2014). It is present in many bacterial operons, such as *ara*, *gal*, and *deo* operons, and in bacteriophages, such as *phage λ*. DNA looping plays an important role in mediating long-range interactions because it allows proteins bound to non-adjacent DNA sites to come close to each other. This strategy is widely used in eukaryotic enhancers, as in the case of the interactions between enhancers and promoters mediated by androgen and progesterone receptors, to integrate multiple signals into the control of the transcriptional machinery. It is also present in the tumor suppressor *p53*, the nuclear factor κB (*NF-κB*), the signal transducers and activators of transcription (*STATs*), the octamer-binding proteins (*Oct*), and the retinoid nuclear hormone receptor *RXR* (Vilar & Saiz, 2011).

The lac operon is well suited to test our current understanding about these types of systems and to develop new methods. The main reason is that it embodies the core elements present across many levels of transcription regulation, it offers the possibility of considering the actual mode of binding and regulation, and it has substantial amounts of experimental data available to contrast the hypothesis and results of the model. In short, there is no room for wiggling. Currently, it is possible to predict how the effects of a single-base pair mutation in the operator DNA would propagate trough all the series of events that lead to protein production from the lac operon (Vilar & Saiz, 2013b). This task proved to be challenging in several fronts. Firstly, it requires an efficient approach to connect the parts as a system to avoid getting into a combinatorial complexity problem, in which the number of potential states of the system grows

> **This approach accurately reproduces the observed transcriptional activity of the lac operon over a 10,000-fold range**

exponentially with the number of components (Vilar & Saiz, 2010). Secondly, the increase in components leads also to an increase in the number of parameters, but many of these parameters

are thermodynamically related to each other. Finally, the values of the parameters might be different under different experimental conditions.

To achieve such predictive capabilities, it was necessary to elucidate biophysical principles for integrating the prototypical complex interactions of transcription regulation into a manageable description. The key idea is to use a modular design with a decomposition of the free energy of the different states into additive contributions of the interactions (Vilar & Saiz, 2013a, Vilar & Saiz, 2013b). This approach allowed the whole system to be characterized in terms of a few parameters directly connected to the experimental data. It considers lac repressor dimers and operator sequences as elementary components. The behavior of the system is obtained starting off from the dimer assembly into tetramers, binding of dimers and tetramers to the different operators, and looping of DNA by the simultaneous binding of a bidentate tetrameric repressor to two operators (**Figure 1**). This approach accurately reproduces the observed transcriptional activity of the lac operon over a 10,000-fold range for 21 different operator setups (deletions and mutations), different repressor concentrations, and tetrameric and mutant-dimeric forms of the repressor (**Figure 1**). Incorporation of the calibrated model into more complex scenarios, taking into account stochastic transcription and translation, accurately captures the induction curves for key

cell-population distributions

maximum-entropy distributions

entropy difference: $\Delta S_i = \int P_i (\ln P_{Normal} - \ln P_{AML}) d\Gamma$

$\Delta S_{19} > 0$ for $P_i = P_{19}$

$\Delta S_{37} < 0$ for $P_i = P_{37}$

Figure 2. Diagnosing leukemia based on the entropy of distributions of cell molecular and morphological parameters. *The two-dimensional distribution of cells with given values of the logarithms of the side scatter (SS log) and the marker CD45 (CD45 log) intensities are shown for representative cell populations, with black and white representing high and zero densities, respectively (Vilar, 2014). The distributions for AML patient no. 37 (P37) and normal individual no. 19 (P19) are indicative of distributions that closely resemble the maximum-entropy distribution of their state, either AML (PAML) or normal (PNormal) states. Healthy and leukemia cell distributions are associated with positive and negative values of ΔSi, respectively. The subscript i in the equation indicates the patient number and Γ is a two dimensional variable Γ = (SS log, CD45 log).*

operator configurations and the temporal evolution of gene expression of growing cell populations (Vilar & Saiz, 2013a, Vilar & Saiz, 2013b). Despite the apparent simplicity of the lac operon, it took over 50 years to have an effective biophysical characterization of this system.
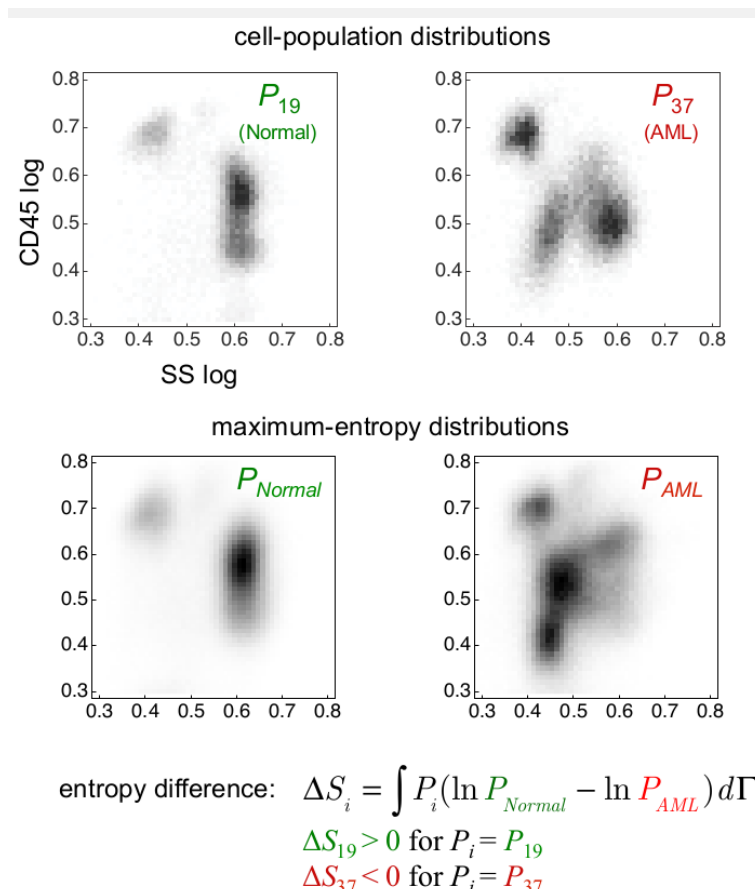
Automated diagnosis of leukemia based on entropy

High-throughput measurement technologies, such as flow cytometry (FCM), can characterize nowadays multiple properties of a single cell at a rate of thousands of cells per second (Aghaeepour, et al., 2013). Acute myeloid leukemia (AML) epitomizes the class of highly complex diseases that these technologies aim to tackle by using large sets of single-cell-level information. Achieving such a goal, however, has proved to depend critically not only on experimental techniques but also on approaches to interpret the data. Specifically, a central aspect of all data-intensive approaches is identifying the relevant quantitative features of the disease from the massive amounts of information produced.

Several machine-learning techniques have been developed to analyze the data in order to diagnose leukemia with different degrees of success (Aghaeepour, et al., 2013). It is also possible, however, to follow more biophysically inspired approaches. Along this path, it is important to take into account that FCM data do not measure the causes of the disease but just its effects in the cellular markers, which is reflected in the statistical properties of the cell populations. From very general principles, one can show that the probability distribution that best represents the healthy or AML state is the one with the largest entropy for each state (Vilar, 2014). From this characterization one can derive, for each patient, a measure of relative entropy as the difference between the patient's distribution and the reference distributions of AML and healthy states deduced from a reference dataset (**Figure 2**). This relative entropy allows the classification of each patient as healthy or AML positive with almost perfect accuracy, which lead this approach to rank first in the DREAM6 challenge (Aghaeepour, et al., 2013). This case illustrates how using biophysical information it is possible to efficiently identify the key features that are hidden within large amounts of data.

The overarching goal

Linus Pauling noted that "life is a relationship among molecules and not a property of any molecule". The ultimate goal of systems biophysics is precisely to work out those relationships. New tools, and especially new frameworks and conceptual developments, are still needed to accurately determine the cellular behavior in terms of the physical properties of the molecular interactions.

New approaches have to be able to describe the complex assembly dynamics of the multiple cellular components that carry out the cellular function

Even relatively simple systems, like the lac operon, have proved to be substantially more complex that originally speculated. Major challenges are still present on how to integrate thermodynamic and structural information with massive data in order to obtain at least information at the mesoscopic level. New approaches have to be able to describe the complex assembly dynamics of the multiple cellular components that carry out the cellular function over scales ranging from milliseconds to hours and days and they need to account for processes as diverse as protein-protein interaction, binding to DNA, transcription, translation, degradation and macromolecular

assembly of signaling complexes at membranes and scaffolds. Achieving this goal, at least partially, has important implications, as it is a prerequisite for the rational identification of therapeutic molecular targets and eventually for bridging prediction of clinical outcomes with molecular properties.

Jose M. G. Vilar

Biofisika Institute (CSIC, UPV/EHU) and Department of Biochemistry and Molecular Biology, University of the Basque Country, P.O. Box 644, 48080 Bilbao (Spain).

Basque Foundation for Science – Ikerbasque, 48011 Bilbao (Spain).

E-mail: j.vilar@ikerbasque.org

References

Aghaeepour N, Finak G. "FlowCAP Consortium"; "DREAM Consortium" (including JMG Vilar), Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH. "Critical assessment of automated flow cytometry data analysis techniques". *Nat Methods*, **2013**, 10: 228. DOI: 10.1038/nmeth.2365.

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, Walter P. "Molecular biology of the cell". 6th ed. *Garland Science*, New York, **2014**. ISBN: 9780815344322.

Gawad C, Koh W, Quake SR. "Single-cell genome sequencing: current state of the science". *Nat Rev Genet*, **2016**, 17: 175. DOI: 10.1038/nrg.2015.16.

Hodgkin AL, Huxley AF. "A quantitative description of membrane current and its application to conduction and excitation in nerve". *J Physiol*, **1952**, 117: 500. PMC1392413.

Jacob F, Monod J. "Genetic regulatory mechanisms in the synthesis of proteins". *J Mol Biol*, **1961**, 3: 318. DOI: 10.1016/S0022-2836(61)80072-7.

Müller-Hill B. "The *lac* operon : a short history of a genetic paradigm". *Walter de Gruyter*, Berlin ; New York **1996**.

Saiz L, Vilar JMG. "Stochastic dynamics of macromolecular-assembly networks". *Mol Sys Biol*, **2006a**, 2: 2006.0024. DOI: 10.1038/msb4100061.

Saiz L, Vilar JMG. "DNA looping: the consequences and its control". *Curr Opin Struct Biol*, **2006b**, 16: 344. DOI: 10.1016/j.sbi.2006.05.008.

Sung MH, McNally JG. "Live cell imaging and systems biology". *Wiley Interdiscip Rev Syst Biol Med*, **2011**, 3: 167. DOI: 10.1002/wsbm.108.

Vilar JMG. "Entropy of Leukemia on Multidimensional Morphological and Molecular Landscapes". *Physical Review X*, **2014**, 4: 021038. DOI: 10.1103/PhysRevX.4.021038.

Vilar JMG, Guet CC, Leibler S. "Modeling network dynamics: the *lac* operon, a case study". *J Cell Biol*, **2003**, 161: 471. DOI: 10.1083/jcb.200301125.

Vilar JMG, Leibler S. "DNA looping and physical constraints on transcription regulation" *J Mol Biol*, **2003**, 331: 981. DOI: 10.1016/S0022-2836(03)00764-2.

Vilar JMG, Saiz L. "CplexA: a Mathematica package to study macromolecular-assembly control of gene expression". *Bioinformatics*, **2010**, 26: 2060. DOI: 10.1093/bioinformatics/btq328.

Vilar JMG, Saiz L. "Control of gene expression by modulated self-assembly". *Nucleic Acids Res*, **2011**, 39: 6854. DOI: 10.1093/nar/gkr272.

Vilar JMG, Saiz L. "Systems biophysics of gene expression". *Biophys J*, **2013a**, 104: 2574. DOI: 10.1016/j.bpj.2013.04.032.

Vilar JMG, Saiz L. "Reliable prediction of complex phenotypes from a modular design in free energy space: an extensive exploration of the lac operon". *ACS Synth Biol*, **2013b**, 2: 576. DOI: 10.1021/sb400013w.

Wartlick O, Mumcu P, Kicheva A, Bittig T, Seum C, Julicher F, Gonzalez-Gaitan M "Dynamics of Dpp Signaling and Proliferation Control". *Science*, **2011**, 331: 1154. DOI: 10.1126/science.1200037.

EDITORS

Jesús Salgado

Jorge Alegre-Cebollada

Xavier Daura

Teresa Giráldez

ISSN 2445-4311

CONTACT

SBE - Sociedad de Biofísica de España

Secretaria SBE, IQFR-CSIC,

C/Serrano 119, 28006 Madrid

Email: sbe_secretaria@sbe.es

WEB: http://www.sbe.es

SPONSORS