Analysis of Regulatory and Epigenetic Stochasticity in Development and Disease

Prepared by Andrew Feinberg, John Hopkins University & Roy Wollman, UCLA

(summarizing comments from workshop participants)

A workshop related to the analysis of stochasticity in epigenetics was held at IPAM on March 1-3, 2017. Experts in the application of novel analytical and experimental tools gathered to explore recent advances and barriers to progress related to the mathematical understanding of the relatively new biological field of epigenetics. The workshop served as a forum for scientists and engineers with an interest in computational biology to explore the role of stochasticity in regulation, development and evolution, and its epigenetic basis. Stochasticity has been transformative in physics and in some areas of biology. It also promises to fundamentally transform modern genetics and help to explain critical biological behavior, such as differentiation and cancer.

Epigenetics refers to information used to control gene expression that is not part of DNA sequence per se, yet it is transmitted during cell division. Through the control of gene expression, epigenetic regulation distinguishes stem cells from somatic cells, one organ from another and even identical twins from each other. In contrast to DNA sequence, the epigenome is relatively susceptible to modifications by the environment and can be subject to stochastic perturbations over time, adding to phenotypic diversity in a population. There is increasingly strong evidence that in epigenetic regulation "stochasticity" is hardly synonymous with "noise", which often refers to variation that obscures a "true signal" (e.g., measurement error) or which is structural, as in physics (e.g., quantum noise). Rather, "stochastic regulation" refers to purposeful, programmed variation; the fluctuations are random but there is no true signal to mask.

Talks at the conference included several exciting areas at the interface of mathematics and biology. Key themes were related to the role of stochasticity in the control of gene expression, biological limits imposed by stochastic fluctuations, and genomic inference of epigenetics states in the presence of stochastic fluctuations.

Current Progress

1. Beneficial role of stochastic gene expression

Andrew Feinberg (Johns Hopkins) discussed a new model in which natural selection will favor the emergence of genetic loci for epigenetic variation that can occur randomly or in response to environmental signals and affect phenotypes in which the environment changes unpredictably but often enough. This idea has led to a unifying model of cancer in which increased epigenetic stochasticity allows rapid selection for tumor cell survival at the expense of the host.

Gabor Balazsi (Stony Brook) discussed gene networks and cell evolution, and how one can decouple mean and noise in yeast protein expression. The noisy population has a selective advantage at high stress. Synthetic circuits with different topologies were used to experimentally validate the beneficial role of variable gene expression under stress conditions.

Leor Weinberger (UCSF) discussed bet hedging, an important driver of HIV latency. The HIV LTR promoter that regulates activity of the virus can toggle on and off. This is a paradigm of a stochastic switch, and Professor Weinberger discussed how stochasticity is harnessed for fate control. Mathematical modeling and experimentation have led to novel drug screens that have practical implications for HIV treatment.

2. Fluctuation and control of biological systems

Andre Levchenko (Yale) asked how one generates form from environmental information and internal cell states. He reviewed work showing that maximum information transfer occurs with heterogeneity in a population of cells involved in signaling, and discussed new data relating increased stochasticity and survival of tumor cells in the setting of stress, such as hypoxia.

Johan Paulson (Harvard) discussed limits on noise reduction in signaling. A series of high profile papers have explored the interface of mathematics and experimentation in understanding the enhancement of Shannon information through stochastic processes, as well as the role of anabolic efficiency (i.e. combining elements such as proteins) in creating large fluctuations. Recent work shows how stochastic antagonism can also lead to bistability in biological systems.

Adam Arkin (UC Berkeley) discussed biological control at ecological level. He presented work on the molecular determinants of microbial community stability. Analysis of microbial community composition shows how interactions between species can act as stabilizing factor that could introduce stability and prevent stochastic fluctuation from dominating community composition. This work has significant implication in the area of synthetic ecology and the use of engineered microbial agents in health and biotechnology.

3. Inference of epigenetics states

Garrett Jenkinson (Johns Hopkins) presented a statistical/computation approach to modeling a particular type of epigenetic data (DNA methylation sequencing data) using the Ising model of statistical physics. The approach borrows principles of the Ising model and applies them to genomic sequencing.

John Goutsias (Johns Hopkins) showed how to quantify stochasticity in DNA methylation using a normalized version of Shannon's entropy and demonstrated that this measure of stochasticity can be used to identify the boundaries of topologically associated domains (TADs), highly conserved structural features of the genome whose loci tend to frequently interact with each other, with much less frequent interactions being observed between loci of adjacent domains.

Don Geman & Laurent Younes (Johns Hopkins) discussed the "High d, low n problem" which results in much of contemporary genomic results being simply wrong, because of the relatedness of supposedly independent measures. This problem is confounded even further by the role of stochasticity, which can differ among sample groups, e.g. cancer versus normal.

Domitilla Del Vecchio (MIT) discussed a mathematical model of cell fate reprogramming that presents a potential hypothesis supporting the current low efficiency of the process (well below 1%), demonstrating that the structure of the GRN may be implicated in reprogramming failure. She introduced the design of a synthetic genetic feedback controller circuit for controlling cell fate while overriding the natural GRN. Stochastic models based on the chemical master equation were analytically solved for the steady state probability distribution, illustrating the advantageous properties of feedback overexpression control versus standard prefixed overexpression control toward controlling cell fate.

Challenges for future research on Epigenetic Stochasticity:

Key challenges to future research where discussed throughout the meeting:

1. Interplay between variability and stress.

A general theme of the meeting and the field, is the role of stochasticity in emerging properties across biological systems, particularly under stressful conditions, like anaerobic metabolism in yeast and oxidative stress in cancer. Feinberg/Goutsias/Jenkinson talked about the role of epigenetic stochasticity in unexpected emergent properties in development, such as pseudo-phase transitions from stem cells to differentiated cells, and normal cells to cancer cells. Moreover, stress-induced variability in cancer drives metastasis in a way very similar to micro-organisms. Also related is work by Paulsson showing that large fluctuations in control systems provide strong constraints on their target. In bacteria, epigenetic cell fate decisions between growth and cell division generate bistability whereas stochastic antagonism produces bistability.

2. Evolution of epigenetic stochasticity

How does cell growth and division interact with stochasticity to give rise to emergent properties in development and large cell populations? Stochasticity distributes cells on phenotypic and fitness landscapes that on their end can alter stochasticity. The stochastic cellular dynamics on phenotypic and fitness landscapes gives rise to a multiscale modeling challenge where molecular interactions at the lowest scale determine cellular properties that influence population dynamics at a larger scale.

3. Biology as information science

Biology is an information science par excellence, making it essentially different from chemistry or physics, although chemistry and physics certainly underlie elementary biological processes and functions. Information storage, passage and replication, and computation based on information exchange between the cell genome and the environment, are key to understanding biological structure and function, and, ultimately, endowing Darwinian Theory with its more complete description. As such, the use of information theory is critical to precisely understand how biological function is conditioned on environmental changes, and how this conditioning and cellular decision making are executed through diversified cell population-level responses. The new frontiers in this analysis over the coming decade will be to combine information theory applied to biological systems with control theory and the analysis of the complex cellular decision making undertaken over the multidimensional 'landscapes' of cellular states.

Information flow has multiple roles in biology. First, the demands of evolution require the preservation of heritable information over geological time while permitting enough drift to provide selectable substrate while conditions change. Second, to adapt at shorter timescales organisms need to sense and respond to their environment effectively. Third, these must be accomplished in energetically efficient ways, which include strategies for clonal division of labor or the ability to allow for error rates below a certain threshold.

There is a grand challenge in quantifying information flow in cellular networks and across evolutionary time (which, in microbes, is relevant on the timescale of our lifetimes with phenomena like antibiotic resistance). We have an amazing opportunity with large scale measurements of genotype, physiological measures and complex phenotypes to derive laws beyond unstructured statistical models that govern cellular behavior and dynamics. This requires a theory of experimental design: what interventions, under which set of perturbation in which statistical format. But most importantly, it requires development of new mathematical approaches and models that fit measurements obtained by these experimental designs.

4. Inclusion of mechanistic knowledge in genome-scale statistical inference.

Identifying robust disease biomarkers from tabula-rasa machine learning is virtually impossible in the ultra "high d, small n" domain scenario encountered in functional genomics and

epigenomics. It is therefore not surprising that few if any biomarkers for complex diseases, such as cancer, have made the transition from "blackboard to clinic." Perhaps the most feasible strategy is to introduce mechanistic-based constraints into modeling and learning. Put differently, the only way to confront the bias-variance trade-off in high-dimensional learning is to introduce the natural "biases" encoded in knowledge about large-scale systems biology, e.g., epigenetic mechanisms and gene regulatory networks.

The Ising model of epigenetic methylation discussed in the meeting serves as an illustrative example of mechanistic constraints providing statistical modeling benefits. The data source provides constraints on the information available to modeling; WGBS reads contain knowledge about means and nearest-neighbor correlations. From a maximum entropy perspective, these constraints lead to the 1D Ising model from statistical physics. However, the relatively large number of parameters cannot reliably be estimated from the relatively low number of observations. Therefore, known biochemical properties and behaviors of methylation must be incorporated to constrain the statistical modeling in order to reduce estimation variance. The highly censored nature of the experimental data source provides numerous complicating features to the statistical estimation problem from both a mathematical and a practical computational perspective. Namely, the usually convex maximum likelihood estimation problem in an exponential family of models, such as the Ising model, is converted into a non-convex problem. Additionally, the problem requires not only rapid computation of the partition function, but also marginalization of unobserved variables. Once again, the constrained structure of the nearest-neighbor Ising model is exploited, here to come up with factorizations of the probability distribution which allow for rapid linear time recursive algorithms that solve the aforementioned problems that, in the general case without any structure, would have geometric computational complexity.

Summary

The concept of stochasticity has strongly influenced one branch of science after another over the past two centuries: genetics and physics in the 19th century (and even more profoundly in the 20th century), and communications theory and machine learning since the middle of the 20th century. We anticipate that it will have a similar influence on the emerging field of epigenetics.