Jacques Loeb Centre
for the History and Philosophy
of the Life Sciences
Prof. Ute Deichmann, Director

Ben-Gurion University of the Negev
P.O. Box 653, Beer-Sheva, 8410501, Israel
Building 39, Room -114 | Phone: +972-8-6472258
Email: jloebcentre@post.bgu.ac.il | Web: in.bgu.ac.il/en/loeb

Data, Knowledge, and Genomic Causality

Philosophy and scientific practice in informational biology the era of "big data" technology and computation

December 9-10, 2019

Ilse Katz Institute for Nanoscale Science and Technology (Bldg. 51)
BGU Marcus Family Campus

Abstracts

(By Author)

Fred Sanger (1918-2013), the biologist who worked out how to sequence proteins, RNA and DNA
George G. Brownlee (University of Oxford, UK)

Fred Sanger (1918-2013) was one of the most influential molecular biologists of the twentieth century. He devised methods for sequencing first proteins and then nucleic acids. He was awarded the Nobel Prize for Chemistry twice and is the only scientist to have achieved this distinction. A modest man he rarely spoke about his vision or the future direction of his science so it is difficult to pigeonhole his philosophy. He stated, on his award of his second Nobel Prize in Stockholm in 1980, 'scientific research is one of the most exciting and rewarding of occupations. It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure.' Sanger believed in the need for basic research to solve the outstanding problems facing any scientist. Just after completing his PhD, supervised by Albert Neuberger (1908-96) in 1943, he was introduced to insulin - the hormone secreted by the pancreas, by Professor Chibnall (1894-1988) who was the incoming Professor of Biochemistry in 1943 replacing Gowland Hopkins (1861-1947). This set Sanger on a path of research ideally suited to his temperament and research interests. In my biography of Fred Sanger (Cambridge University Press, 2014) I argue that Sanger was introduced to science through his medical father, (although he did not want to become a medic himself 'preferring the idea of having a single problem I could really get my teeth into') and an excellent scientific education, both at Bryanston school and at Cambridge University. Sanger was thus the product of an enlightened educational system encouraging scientific enquiry and research into novel problems.

Ironically, although Sanger did not want to become a medic, the impact of Sanger's DNA sequencing methods in medical research continues even today, as his method is the basis of most current high-throughput DNA sequencing in current use.

Correlations and causal relations in the study of the cellular and molecular basis of drug addiction

Ami Citri (Hebrew University of Jerusalem, Israel)

Causality is a big word. And it is often misapplied.

I will describe our attempts to minimize the gap between correlation and mechanistic insight, in the investigation of the cellular and molecular basis of mouse models of drug addiction.

Data, theory, and scientific belief in early molecular biology: Pauling's and Crick's conflicting theories about the genetic determination of protein synthesis and the solution to the 'secret of life'

Ute Deichmann (Ben-Gurion University of the Negev, Israel)

Opinions on the relationship between data and theory fall between the two poles of empiricism and anti-empiricism. On the one hand, the Duhem-Quine thesis holds that theories are so radically underdetermined by data that data are insufficient to determine what scientific beliefs a scientist should hold. On the other, proponents of big-data science have declared a new era of empiricism in which the volume of data accompanied by computational tools enables data to speak for themselves, free from theory.

Motivated by these contradictory arguments, I analyze, both historically and conceptually, the generation of two highly important conflicting theories in early molecular biology, namely Linus Pauling's structural and Francis Crick's informational theory of biological specificity and protein synthesis, both generated in the 1950s. My goals are:

- To explore the relationship between experimental data and theory in Pauling's and Crick's theories
- To show that both Pauling and Crick based their views on only a few, and nearly the same, experimental data
- To argue that:
    - Unlike in the Duhem/Quine 'underdetermination', theory, choice was possible at the time if direct experimental data was complemented by logic, causal analysis, and a broad scientific knowledge outside the field in question.
    - Pauling's and Crick's theories were not equivalent, because Pauling's theory lacked a causal connection to DNA.

- A reliance on data alone does not lead to a causal understanding of basic features of life.

Using big data to dissect the autism spectrum

Ally Eran (Ben-Gurion University of the Negev)

The promise of precision medicine lies in data diversity. More than the sheer size of biomedical data, it is the layering of multiple data modalities, offering complementary perspectives, that is thought to enable the identification of coherent patient subgroups with shared pathophysiology. We used autism spectrum disorder (ASD) to test this notion. We first integrated familial whole exome sequences with neurodevelopmental expression patterns to identify clusters of neurodevelopmentally coregulated, sexually dimorphic, ASD-segregating deleterious variation, consistent with our current understanding of the complexity, origin, and epidemiology of ASD. While the function of most clusters converged on previously described ASD etiologies, 20% of the identified clusters were enriched with lipid regulation functions. Collectively, these variants are predicted to lead to high cholesterol and triglyceride levels. Therefore, we used electronic medical records to examine blood lipid profiles of children with ASD as compared to neurotypical children. ASD was associated with elevated LDL (OR=1.78, 95% CI=[1.40-2.26], Fisher's P = 6.42×10-7), total cholesterol (OR=1.67, 95%CI=[1.35-2.06], Fisher's P=5.55×10-7) and triglycerides (OR=1.19, 95%CI=[1.02-1.40], Fisher's P=2.60×10-2). We further used healthcare claims data to compare the abundance of dyslipidemia diagnoses in individuals with ASD, unaffected siblings, and unrelated controls. Children with ASD had a strong enrichment of dyslipidemia diagnoses as compared to matched controls (OR=1.95, 95% CI=[1.80,2.11], P=1.94x10-65) as well as unaffected siblings (OR=1.76, 95% CI= [1.61 1.92], P=2.25x10-36). Finally, mouse models of dyslipidemia were found to exhibit striking phenotypic similarities to autism mouse models (power = 0.97). Taken together, this work identifies a robust subgroup of patients with dyslipidemia-associated ASD.

Inferring ancestral functions for highly conserved developmental genes in early animals and their ancestors

Douglas H. Erwin (Smithsonian National Museum of Natural History, U.S.A.)

Understanding causal relationships for events in deep time such as the origin of animals and their diversification during the Ediacaran-Cambrian radiation (c. 5570-520 million years ago) was long thought to be beyond the remit of paleontologists, who focused on pattern rather than process. In this case the challenges are multiplied by the nexus of environmental, ecological, genomic and developmental influences on the rise of early animals. Here I concentrate on one aspect of these challenges, inferring the ancestral role of deeply homologous genes shared across animals and in some cases with their closest ancestors. The first examples of deep homologies came from comparisons between vertebrates and flies, leading to inferences about the nature of the last common ancestor of these two clades, variously described as the protostome-deuterostome ancestor or the Urbilaterian. As information became available from more deeply branching clades, including cnidarians and sponges, the extent of functional evolution of deeply homologous genes became more evident. It is now clear that deep homology of genes provides limited insight into function in ancestral forms.

Optogenetics, causation and the enlarged biological community

Denis Forest (Université Panthéon-Sorbonne, France)

From its early beginnings, Optogenetics has been defined as the combination of two components, “genetic targeting of specific neurons” and “optical control of the targets within intact, living neural circuits” (Deisseroth et al., 2006). The key innovation has been the expression of an opsin, ChR2, discovered in a unicellular alga, Chlamydomonas reinhardtii, in the membrane of neurons, where it triggers action potentials in response to pulses of light (Boyden et al., 2005). Allowing researchers to go beyond observational knowledge and to “dissect” neural circuits, optogenetics matters because it “brings neuroscience closer to causality” (Insel, 2015).

In my talk, I will first make explicit what “causality” may be in the context of optogenetics. Then I shall deal with two main questions. The first is the relation between interventions on neurons made possible by this method, and the knowledge of corresponding causal relations. In particular, I suggest that assessments of optogenetics as an epistemic tool rely on a reference to scientific values (like accuracy, fruitfulness, and coherence). In this regard, the much-lauded spatial and temporal precision associated with optogenetics (Carter & De Lecea, 2011) is not necessarily synonymous with accuracy, because of the intricacy of neural networks (Otchy & al., 2015; Sullivan, 2016; Wolff & Ölveczky, 2018). The second question is the significance of the new relation established through optogenetics between microbiology and neuroscience (Deisseroth & Hegemann, 2017). The success of the transformation of biological tools (found in algae and Archaea) into epistemic tools is another sign of the reintegration of neural organisms into a larger biological community where similar low-level mechanisms contribute to diverging outputs.

Data is not information, information is not knowledge, knowledge is not wisdom, wisdom is not truth: The many methodological crimes of data-driven big science

Dan Graur (University of Houston, U.S.A.)

In the biomedical realm, "big data science" refers to studies that seek to extract meaningful insights from data sets that are claimed to be either too large or too complex to be dealt with in traditional scientific methodology. Big data usually involves many cases (rows) and many traits (columns) and are thought to offer great statistical power. Sadly, big data also lends itself to unjustifiable, sometimes fraudulent, and always hard-to-detect manipulation.

In this lecture I will use examples from the literature—libel laws be damned—to discuss several problematic practices in big-data studies: (1) the abandonment of Popperian hypothesis testing in favor of 17th-century Beconian data-driven discovery, (2) the employment of methods and experimental setups that intentionally inflate or deflate the estimates of interest, (3) the employment of overly complex models and statistical overfitting leading to apophenia (the experience of seeing meaningful patterns in random data), (4) favoring discovery of true positives (statistical sensitivity) over the ability to distinguish true positives from false positives (statistical specificity), (5) practicing the P-value cult, whereby statistical significance is confused with biological significance (the magnitude of the effect), (6) failing to recognize the existence of overlapping categories, (7) Using Bayesian statistics with informative priors, which has been shown to be indistinguishable from falsification, (8) employing arithmetical dirty tricks, (9) using “black boxes” (machine learning) to yield inferences unlikely to be applicable to any other data set, and (10) employing logical fallacies, such as “affirming the consequent.”

Some “crimes” related to "big data science" are due to external factors such as multi-authorship-distributed responsibility, ignorance of pre-big-science literature, and too much money.

Beyond ‘Big Data’ and ‘Big Model’: The role of abstraction in biology
Jeremy Gunawardena (Harvard Medical School, U.S.A.)

Two contrasting approaches characterise the interplay between experiments and theory in biology: Big Data, coupled to Bigger Computers; and Big Models. We will explore the role of abstraction in biology as offering, perhaps, a third way.

Identifying genomic changes that are associated with or cause phenotypic changes between species
Michael Hiller (Max Planck Institute of Molecular Cell Biology and Genetics, Germany)

Evolution has generated a fascinating diversity of phenotypes in different species. The main question that unites the research in our group is: What is the genomic basis of phenotypic differences between species? By focusing on differences between species rather than on differences within a species, our work aims at revealing genomic differences that are relevant for macroevolutionary changes. To this end, we have developed comparative genomic methods to detect functionally-relevant differences in the genomes of species and to detect associations between such genomic and phenotypic differences. I will discuss principles and which experiments can help to distinguish between genomic changes that are associated to a phenotypic change and genomic changes that may have contributed to phenotypic changes.

Spatial transcriptomics of the mammalian intestine

Shalev Itzkovitz (Weizmann Institute of Science, Israel)

The intestinal epithelium is a highly structured tissue composed of repeating crypt-villus units. Enterocytes, the nutrient absorbing cells, are born from stem cells that reside deep in crypts, and migrate for 3 days along the walls of villi until they are shed off from their tips. During their migration, they perform the diverse tasks of absorbing a wide range of nutrients while protecting the body from the harsh bacterial-rich environment. We applied spatial transcriptomics approaches to identify broad functional zonation of these migrating enterocytes and found that they constantly change their cellular states as they migrate, a phenomenon termed 'zonation'. Enterocytes first implement anti-microbial programs at the bottom of the villi, then shift to sequential absorption of distinct nutrients at different villus height, finally inducing an immuno-modulatory villus-tip expression program. Using similar approaches, we identified spatially stratified populations of mesenchymal cells that provide signaling cues that shape enterocyte zonation. Our work demonstrates design principles that dictate spatial division of labor in mammalian tissues.

Genomic causality in a bacterium that was lab-evolved to turn CO2 into sugar

Ron Milo (Weizmann Institute of Science, Israel)

In this talk, I demonstrate how a combination of rational metabolic rewiring, recombinant expression and laboratory evolution has led to the biosynthesis of sugars and other major biomass constituents, by a carbon dioxide fixation cycle engineered into the model gut bacterium E. coli. I will describe the genetic basis for the adaptation of E. coli to sugar synthesis from CO2. We find that only five mutations are sufficient to enable robust growth. All mutations are found either in enzymes that affect the efflux of intermediates from the autocatalytic CO2 fixation cycle towards biomass or in key regulators of carbon metabolism.

Causality and validation in single-cell genomics: Analysis of a major hematopoietic transition

Ellen V. Rothenberg (California Institute of Technology, U.S.A.)

T-cell development is a transformation of multipotent blood cell progenitors that enter the thymus to acquire a T-lineage identity and lose their access to alternative fates. The T-cell specification process involves a response to Notch pathway signaling sustained across at least ten cell cycles. Studies of populations sorted based on cell surface markers have identified multiple, apparently stepwise regulatory transitions, with activation of T-lineage supporting transcription factors, silencing of alternative lineage promoting factors, and numerous changes in chromatin accessibility all over the genome as the cells become committed to the T-cell fate. There is evidence from perturbation studies that specific transcription factors have strong impacts on some of these changes, as expected for nodes in a T-cell gene regulatory network. However, interpretations in terms of a comprehensive gene network model have still called for caution. The exact regulatory states of individual cells and their order of changes in the critical earliest stages of this process have been obscure until now, partly due to their strong similarity to other very undifferentiated, stem-cell-proximal hematopoietic precursors.

Data-driven models of embryogenesis

Stanislav Shvartsman (Princeton University, U.S.A.)

We aim to establish and experimentally test mathematical models of embryogenesis. While the foundation of this research is based on models of isolated developmental events, the ultimate challenge is to formulate and understand dynamical systems encompassing multiple stages of development and multiple levels of regulation. These range from specific chemical reactions in single cells to coordinated dynamics of multiple cells during morphogenesis. Examples of our dynamical systems models of embryogenesis – from the events in the Drosophila egg to the early stages of gastrulation – will be presented. Each of these will demonstrate what had been learned from model analysis and model-driven experiments, and what further research directions are guided by these models.

Epithelial-mesenchymal transition in malignant progression: The limited contribution of genomic analyses

Robert Weinberg (Whitehead Institute for Biomedical Research, U.S.A.)

As the carcinoma cells of primary tumor progress toward a state of high-grade malignancy, they accumulate a number of somatic mutations in "driver genes" that confer growth and survival benefits on the evolving cells that ultimately form the primary tumor. However, the final step in this progression involves the acquired abilities to invade and disseminate to distant sites in the body. In contrast to the previously acquired changes in primary tumor cells driven by somatic mutations, no recurring mutant alleles have been associated with these traits of high-grade malignancy. The failure to find such alleles has redirected focus toward non-genetic, i.e., epigenetic changes that may confer these traits.

In fact, the cell-biological program termed the epithelial-mesenchymal transition (EMT), once activated in these cells through non-genetic means, generates many of the cell-biological phenotypes that are associated with high-grade malignancy. The EMT program, which plays critical roles in normal embryonic morphogenesis and wound-healing, is appropriated by these carcinoma cells, being activated in these cells through signals that they receive from the adjacent recruited stroma. As will be described, once activated, the EMT program enables physical dissemination, increased resistance to anti-cancer therapies, and the formation of tumor-initiating cancer stem cells. In addition, as will be described, activation of this program confers on carcinoma cells an ability to resist the onslaught of attacks by host cytotoxic T-cells that is often provoked by checkpoint immunotherapy. Taken together, the actions of this non-genetic program may ultimately prove more important to malignant progression than the somatically mutated alleles that carcinoma cells accumulated during the course of primary tumor formation.