Vision Paper: How might we find generalizable 'rules of life' that govern how a large number of signals control integrative biological function?

Contributors:

Bradley Evans, Donald Danforth Plant Science Center, St. Louis, MO. bevans@danforthcenter.org

Eben Gering, Department of Biological Sciences, Nova Southeastern University, Fort Lauderdale, FL. ebeng@nova.edu

Haiyan Hu, Department of Computer Science, University of Central Florida, Orlando, FL. haihu@cs.ucf.edu

Rahul Kulkarni, Department of Physics, University of Massachusetts Boston, Boston, MA. rahul.kulkarni@umb.edu

Arianna Tamvacakis, Integrative Systems Neuroscience, University of Arkansas, Fayetteville, AR. tamvacakis@gmail.com

Rajanikanth Vadigepalli, Daniel Baugh Institute for Functional Genomics/Computational Biology, Department of Pathology, Anatomy, and Cell Biology, Thomas Jefferson University, Philadelphia, PA. rajanikanth.vadigepalli@jefferson.edu

Contact: rajanikanth.vadigepalli@jefferson.edu

Preface:

The present Vision Paper has been developed from a series of discussions at the National Science Foundation Reintegrating Biology Virtual Jumpstart meeting held in December 2019. The initial discussion group was charged with identifying several antecedent questions ("*how might we …?*"), answering which may lead to a comprehensive approach to tackle the broad question of "*How do biological entities interact and communicate throughout life?*". The antecedent questions that were formulated in the initial discussion were rated based on the scale and scope of the question ("moon shot, mars shot, and jupiter shot"). Those questions with the largest scope and difficulty in approach were considered in further discussions to develop a vision for how the scientific community could productively pursue these questions. The present document contains a summary of discussion on the generalizable concepts that can emerge from the analysis and interpretation of biological signals.

The document is organized along the following questions:

1. What is the big question?
2. What is the exciting science needed to comprehensively answer the big question?
3. What are potential domain-specific translations of the big question?
4. What is the potential impact?
5. Why is it possible to pursue this question now?
6. What are some of the state of the art technologies that enable the pursuit?
7. What are the key barriers and challenges that need to be overcome?

8. What might be the broader impacts?
9. What disciplines might be needed for this pursuit?
10. How does the pursuit of above scientific questions reintegrate biology?
11. Who should care about this vision? Why?
12. What institutional changes are needed to make the proposed vision a reality?

At the end of the document, a summary of the discussion is included to provide additional context on how one might tackle the big question and pursue exciting scientific opportunities in an attempt to answer the big question comprehensively.

1. What is the big question?

> *How might we discover generalizable "rules of life" governing the use, function, manifestation, regulation, propagation, encoding, and decoding of biological signals across a wide array of biological contexts?*

Answering this question has important implications for our understanding of development, homeostasis, renewal and regeneration at all scales of biology.

2. What is the exciting science needed to comprehensively answer the big question?

The big question and science of finding meaning in a cacophony of a large set of signals ("signalome") gives rise to multiple related questions that drive exciting scientific opportunities for exploration. Note that several of these questions have been pursued in a limited scope within specific biological contexts across many scales. At present, however, biologists working within disparte subdisciplines conceptualize and study signalling in very different ways. The present vision document is aimed at outlining the scope of new opportunities and unaddressed questions in integrative biology at a previously unapproachable scale and resolution, enabled by technological advancements in large-scale measurements as well as computational analysis and modeling.

1. How do the constituents of a biological collective broadcast, census, and process information to drive the function of an integrative whole?
2. How might we identify the most salient subset of signals controlling a given phenotype? Are there generalizable approaches/heuristics/rules to figure out which parts of a signalome are relevant to a phenotype/context?
3. How might we figure out which signals have an outsized impact on phenotype at larger scale?
4. How can we identify the information within a multivariate array of potential signals that matters to a receiver? Which aspects of the whole can/does the individual entity decode from accessible components of the signalome?
5. How does the receiver transduce a signal and prioritize responses from a wide range of possibilities? What are the intrinsic properties/states of the receiver that condition/bias responses?

6. What are the properties of the signal subset that can inform the individual entities about the functionality/capacity/amount of the whole? Alternatively, What information is encoded in the subset of the signalome relevant to integrative function of the whole?
7. How are multiple signals coordinated across modes to achieve a functional objective at a larger scale?
8. In which systems (and for what purposes) is mechanistics/functional understanding of input/output sufficient? When should we instead be asking all of Tinbergen's four questions (signal ontogeny, evolution, adaptive value, function/mechanism)?
9. How might we identify the generalizable rules to be discovered on how, when, and why do signaling systems fail to elicit appropriate response at a given scale, or fail to lead to appropriate overall functional phenotype at a larger scale?
10. Scientific opportunities for methodological developments: How can we develop a systematic, ideally generalizable, approach that is as minimally subjective as possible for interpreting the large scale signalome in any given context? How to scale machine learning approaches to address these questions? How to make systems dynamics and nonlinear systems analysis methods routinely usable by integrative biologists working on any scale or system?

Refer to the discussion notes at the end of the document that explores these questions in detail.

3. What are potential domain-specific translations of the big question?

Cell biology example:

Consider the problem of liver regeneration where upon injury that damages a portion of the liver, or in the case of surgical removal of a part of the liver, the major constituent cells (hepatocytes) re-enter cell cycle and repopulate and/or grow the organ to nearly the same levels as prior to the injury/resection, over a broad range of injury that can incapacitate or remove up to 90% of the liver. In this context, the domain-specific translation of the big question is: *How to identify key drivers among the hundreds of signals that are activated upon liver injury that ultimately govern the highly repeatable, evolutionarily conserved, integrative process of liver regeneration?*

Driving sub-questions: How do we analyze the relationships between the signals across multiple modalities (genes, proteins, secreted molecules, mechanical cues, electrical activity, etc.) to reduce the dimensionality of the problem to a lower dimensional subset of signals that act as major determinants of liver regeneration phenotype? Which combination of signals in the liver microenvironment serves to perform organ-scale coordination such that, in principle, the signals impacting every cell carry information on the metabolic/functional capacity of the whole organ, so as to yield a collective response that regrows the liver to the physiologically-required (homeostatic?) liver mass/size? How does the intrinsic cellular network state (e.g., as is apparent in the spatial organization of various cellular functions in the liver lobules) govern how the cells respond to the signals triggered upon liver injury, such that an apparently appropriate fraction of response is mounted across the whole organ depending on the level of injury?

Host-pathogen interaction example:

Many pathogens and parasites can evade or subvert host immune systems; some of these organisms can further alter host traits to produce “extended phenotypes” that promote disease transmission at the host’s expense. Consider the infamous example of host manipulation by *Toxoplasma gondii*, an obligate intracellular parasite of mammals and birds worldwide (including an estimated 1-2 billion humans. Within some of these hosts (e.g. rodents), *T. gondii* infiltrates a wide array of tissues to induce behavioral modifications (e.g. attraction to urine odors of feline predators) that likely function to promote trophic transmission to the parasite’s definitive hosts. The infected host’s phenotype thus becomes an emergent property of interacting biological processes encoded and decoded by multiple parties with competing evolutionary interests.

Within the arena of host-pathogen interactions, there are at least two domain-specific translations of our big question: 1) *(How) do infectious agents coordinate their influence over biological processes of hosts?*, and 2) *(How) do these interactions vary across contexts (e.g. host taxon, host sex, host & parasite genotypes, or host condition?*

Driving sub-questions:

What information streams allow pathogens and parasites (e.g. asexual clones) to locate, infiltrate and disperse within a host to control dynamic aspects of host physiology or behavior (e.g. modulation of host hormones, neural activity, and gene expression)? (How) does the constellation of hosts with which parasites co-evolve impact sensory strategies and conditional responses to the local environment within a host? Can rules of life derived from other domains predict how pathogens and parasites evade or subvert host defenses (e.g. molecular mimicry, antigenic variation, etc.)? (When) does host susceptibility to infectious disease reflect trade-offs (e.g. with autoimmunity or other costs of resistance) vs. maladaptation? (How) can information systems of infectious agents be disrupted with minimal harm (or positive effects) to essential processes of hosts?

Population biology example:

In this context, the domain-specific translation of the big question is: *How are key feature of populations (e.g. distributions, dynamics, and behaviors) structured by the signals, sensory systems, signalling environments, and operating characteristics of their constituents?*

Driving subquestions: What roles do uncertainty, sensory bias, and signalling environments play in the flow of information among individuals, and in the emergent features of collectives? (How) do simple decision rules of individuals give rise to complex collective behaviors; can these patterns be predicted from rules of other domains of life? To what extent do the signals determine the key players in a population that significantly influence the behavior of the collective? How does heterogeneity of individuals within a population control how signals are processed across the whole population? Are there generalized insights to be transferred and applied from single cell biology where large heterogeneity is the norm and yet the functional variation at a tissue or organ scale is more constrained?

Communities & ecosystems example:

In this context, the domain-specific translation of the above-stated big question is: *Does information flow through self-assembled communities and ecosystems in similar ways to lower levels of biological organization (e.g. among genes, cells, tissues, individuals, and co-evolving species) to promote or disrupt homeostasis?*

A few driving questions are: What data (at regional or global scales) is necessary and sufficient to identify key regulators of ecosystem-level processes (e.g. nutrient cycling)? What signals regulate functional feedbacks between biotic and geochemical components of ecosystems? Do rules of information flow and signal transduction scale up from lower levels of biological organization (e.g. cells), or do self-assembling ecosystems require different models altogether?

Neuroscience example:

The nervous system is arguably a signaling organ: it communicates information from the external environment and the internal environment of an individual, and produces behaviors that are themselves communicating signals. In this context, the domain-specific translation of the above-stated big question is: *are there generalizable mechanisms by which neural components select salient signals amongst noise, at different levels of the nervous system?* The levels span gross behavior, brain-scale networks, local circuits, synaptic signal, membrane protein fluctuations, gene expression fluctuations, etc.

In particular, several driving sub-questions arise, including: how is the salience of the signal to the cellular function determined? what regulatory network mechanisms filter the noise? Can any "rules" that govern the signal processing in nervous systems be applied to signaling and communication in organisms which do not have a nervous system (slime molds, plants, etc.)?

Behavioral science example:

Behaviors are themselves signals, and at the same time the expression of a response to another signal from the external or internal environment. While in some ways such signals are extremely species- and context-specific, are there also some general rules with which we can think about signaling and receiving salient signals in the context of behaviors. Signal detection theory provides a utilitarian framework for understanding how the detectability and information content of signals. The domain-specific translation of the big question is: *Is there a generalizable approach to determine how an individual determines whether a signal is salient, and if there are defined sets of parameters that can be applied to predict potential behavioral outputs after the signal is received?*

Driving sub-questions: In which cases can the salience principle be applied to diverse, species-specific behaviors, or are behavioral signals always dependent on a number of contexts? i.e., in what cases are behavioral signals a general "rule of life" and in what cases are they only a "rule of life" for a species, environment, or other specific? Within an individual, what is the interplay between external (environmental) and internal (hormones, neurotransmitters, etc) signals that is relevant to a given behavior. Are there general rules? Can these general

rules be applied beyond an individual, to a group of individuals. Can these be applied to only one conceptual idea of behavior, or are there some rules that can be developed for signaling that can effectively define behaviors of animals, cells, genes, etc.? Do the ideas here scale with complexity? Are more complex signals necessarily subject to the same rules of life as relatively “simple” signals?

Note: Behavioral scientists use the term "signals" much more narrowly than is considered here. For example, running from a predator, falling asleep, etc. would provide "cues" about state, but are distinguished from "signals" intended to manipulate the receiver behavior. It is important to consider such differences in terminology as one looks for generalizable rules across levels of analysis.

4. What is the potential impact?

Answers to these questions and approaches developed along the way have implications to our understanding of living systems at all scales of biology. The techniques, methodology and informatics methods used to dissect the complexity of the ‘signalome’ can be applied to other problems with similarly complex information streams. Application of the methods and strategies used for reintegration of the informative signals used in guiding tissue engineering, disease management, etc. examples above can be reused and repurposed for integrating other complex, multi-stream data that can now be collected but not efficiently combined and interpreted. A few non-exhaustive set of cases are summarized below.

1. Control points in disease for therapy
2. Coordinated phenotype formulation for tissue/organism engineering
3. Policy implications in ecosystems management
4. Anticipation and management of pest and pathogen outbreaks and invasion of ecosystems by non-native species. The benefits would be broad in scope and would pertain to humans, animals, crops and wild systems.
5. Everything talks in multiple languages simultaneously - The benefit might be the ability to speak multiple languages and have redundancy in the system. Another way to look at it is to understand what is the optimal amount of nonessential communication or signaling ‘machinery’ that should be retained to allow for adaptation to new threats or stimuli. The adaptability or flexibility that systems have is governed by the redundancy or slack within the system that is available to take on new roles. This has implications for synthetic biology approaches.
6. Development of a set of rules or principles that can be used to describe the components needed to orchestrate the development and organization of structures at different scales ranging from molecular systems, cells, tissues or organismal communities.
7. We would gain the ability to appreciate the contributions of individual components of a biological system or organismal community and avoid the mistake we as humans have often made by disregarding perceived unimportant components as non-essential or dispensable. Perhaps the importance of unappreciated components is due to the context or because of some other shortsightedness. With a set of rules that have general application we may be able to better evaluate systems upon inspection or be

able to create systems that better recapitulate the function as intended as for the case of tissue engineering, environmental remediation or others.

5. Why is it productive to pursue this question now?

As discussed in the multiple domain-specific examples outlined in the preceding sections, we have now obtained significant amounts of data characterizing different system components. A key next step towards understanding the function of the integrative role is developing a quantitative understanding of the role of signaling-based communication in integrating the different parts.

Potential mechanisms focusing on signal integration by cells to elicit phenotypic responses are being explored in diverse biological systems (e.g., quorum sensing in bacteria, competing endogenous RNAs (ceRNA) hypothesis in cancer). Focusing on such selected well-studied examples can lead to insights into developing general approaches for dealing with the problem of connecting specific signals from the signalome to observed phenotypes.

We can measure changes in levels of (nearly) all genes, all proteins, all genomic loci, etc. individually in hundreds to thousands of cells, in multiple organs. The technological developments in large-scale molecular and cellular data acquisition, including in living systems, have led to an unprecedented ability to measure the molecular and cellular state at high-throughput and high-resolution. Single-cell genomics approaches provide data at the single-cell level giving us access to probability distributions across populations of cells. Analyzing these probability distributions can potentially be a critical component of characterizing the phenotype generated by the cacophonous signalome.

New initiatives (e.g. NEON) are providing the data needed to characterize ‘signalomes’ of ecosystems. If salient signals and feedbacks can be identified, we can use them to manage biological communities - preserving or enhancing ecosystem services they provide.

Network science, theory, tools have advanced and applicable to very large datasets. State-of-art computational modeling methods such as deep learning and network modeling - can be done now at large scale. Advances in machine learning approaches can be used in combination with the large datasets that are now available to address the problem.

Signaling systems are being affected by human activities (e.g. altered signal propagation environments, competitive ‘noise’ of synthetic molecules in organisms, etc.). Examples: environmental pollution with pharmaceuticals affects behavior and health of natural populations. We need to understand how signalling ‘works’ to minimize adverse effects.

6. What are some of the state-of-the-art technologies that enable the pursuit?

State-of-the-art technologies are available in various biological domains. For example, the National Ecological Observatory Network (NEON) is an NSF supported continental-scale

observation facility project. NEON aims to collect long-term open access ecological data to facilitate the study of coupled biotic and abiotic environmental change. Currently, NEON provides hundreds of open access data products and thousands of archived specimens collected across the US.

Recently emerging whole-body and whole-organ clearing and imaging techniques have been advanced to the level of single-cell resolution, which brings us ever closer to organism-level systems biology. Besides, current technology development has made it possible to predict variation in behavior across individuals and species. We see many examples of behavioral convergence in divergent lineages, as well as rapid evolutionary changes within human-altered environments (Sih et al. 2011). Thanks to evolutionary developmental biology, we can also now explore their underpinnings at the mechanistic level to discover parallelism and idiosyncrasies in the genes, pathways, or circuits that mediate shared features of different taxa (e.g. Bolnick et al. 2018, Young et al. 2019).

Additionally, next generation sequencing technologies (NGS) enable whole genome-scale, epigenetic, and transcriptomic measurements. As a consequence, omics of various kinds have been deposited in public databases such as NCBI, the European Molecular Biology Laboratory (EMBL), ENCODE and others. Single cell NGS make it a reality to analyze gene expression data both at the population and single-cell level. Such analyses reveal that genetically identical cells in a homogeneous environment can have phenotypic variation, just as clonal organisms reared in homogeneous conditions can become behaviorally differentiated (Bierbach et al. 2017).

Moreover, revolutionary computational modeling methods such as deep learning and network modeling - can be done now at large scale. All the current technologies and rapid data accumulation enable researchers to formulate problems they never thought they could solve before: what are the drivers for the switch between different phenotypes for a cell? How can we identify these based on the data? How do we attack the emerging problem of science communication: sharing/potentiating discovery in an over-crowded literary/research ecosystem (too much ‘noise’ in the current scientific community/literature)?

7. What are some of the key barriers and challenges that will need to be overcome?

Conceptual:

1) **We are often in the dark about the evolutionary histories of signalling and sensory systems.** This context can illuminate whether and how historical selection pressures (e.g. third parties, energetic trade-offs, propagation media, targeted receptors, etc.) have shaped signalling and sensory systems.
2) **Defining and observing homeostasis (or other phenotypes):** Perturbation at one level may produce robustness at higher levels (examples: individuals deploy tolerance vs. resistance to combat pathogens, divergent neuronal morphologies can produce convergent response to stimuli, etc.). Findings may or may not generalize across spatial or temporal scales
3) **Investigators at different scales discuss and understand signals, information, and responses in different ways.** There is disagreement about the importance of signal history

and ‘intent’. On one hand, it is important that a signal can be produced and received without consideration of intent. However, the distinction is helpful in terms of understanding how signals evolve, and whether or not signals are likely to be shaped by constraints/trade-offs in their design.

4) **Even if we measure a lot of what we can measure, are we measuring what is relevant? How to know?** This is a problem of ‘known unknowns’ versus ‘unknown unknowns’. Of relevance, after the successful release of the first comprehensive human genome sequence through the human genome project, the scientific community realized that we did not find the 200K genes we hoped for, and that opens up new challenges and questions, etc., requiring a fundamental reformulation of how one goes about unpacking biological systems. Similar lessons learned from ENCODE project as well in terms of the complexity of transcriptional regulation at the genome-scale.
5) **Our definition of the integrated whole that we seek to understand may be limiting.** Constituents of integrated systems (e.g. cells in organs) can be individually connected to other systems in ways that we are not looking for, yet contribute to function.

Methodological:

6) **We need to account for interactions among affected agents in high-resolution (e.g. molecules, cells, populations).** But we do not always know which parties to observe (who is sending and receiving signals?). Agents can signal to themselves, to targeted receivers, to harmless ‘bystanders’, and/or nefarious exploiters (pathogens, predators, etc.). Are there general methods for identifying all parties on either end of a signal?
7) **Lab models may not reflect natural systems** in which signals and responses evolved, or that we seek to understand and manipulate. We will need to characterize the context-dependency of communication to predict or alter how information streams impact biological phenomena.
8) **We lack methods for synthesizing different kinds of data,** e.g., those that make up an interactome. Even as we ‘catch up’ on integrating datasets (e.g. gene expression and chromatin structure in signalling cells of 3D tissue), we will continue to discover new layers of information that modulate biological processes. Analytical and collaborative pipelines must be flexible enough to integrate new kinds of data as their relevance emerges.
9) **Correlations or lackthereof across multiple data types is still hard to interpret,** e.g., levels of genes versus proteins - several layers of regulation in the middle. Multiple data types are not obtained under the exact same conditions, might be from different labs, following different protocols.
10) **Methodological advancements that show promise but are not yet operational at scale**, e.g., genome-scale mechanistic modeling to relate the big molecular data to systems dynamics is not quite feasible - but progress is being made (e.g., whole cell modeling of Mycoplasma). Another example is the simulation of an entire city at the resolution of thousands of individuals for examining the effects of various zoning policies on public health. Cannot quite simulate entire ecosystems yet in a realistic manner.
11) **Tools for translating large volumes of data into meaningful biological insight are still in their infancy.** Our ability for collecting high-content and high throughput data types using

a variety of technologies ranging from optical and spectroscopic imaging, chemical analysis and macromolecular sequencing and quantitation (genome sequencing/transcriptomics/ proteomics) techniques is not limiting. There is however, a limitation that affects the various analytical and 'omics technology discipline which is that all of these suffer the limitation of translating the large volumes of data into meaningful biological insight. Furthermore, it is often difficult or practically impossible to integrate the various data types. In order to facilitate translation of these big-data or 'omics scale/systems biology-type data to information that will advance the biological disciplines emphasis should be placed on developing a common "language" to enable data integration going forward as well as on developing "translation" methods to allow for usage of the vast amount of data that has been collected and is currently available in the public domain

12) **Integration of various data types is limited due to incompatible ontologies.** e.g., databases and accession/naming systems can differ in their completeness and quality, as well as in the code and convention for identifying biological components. For example database A used to annotate an RNASeq dataset uses a different naming system than database B used for annotating the proteomics data and also from that of database C used for annotating metabolite data. This situation limits the ability to correlate these data and therefore to analyze the growing number of experiments for which these complementary data exist at a high level and researchers instead often opt to "cherry-pick" data based on system-specific knowledge and hypotheses. If we aim to decipher correlations and patterns that can be used to identify and enumerate signaling pathway and mechanism categories or types to establish common "rules of life" that will move biology and science forward as a whole we must develop the ability to make full use of the existing and to be collected data.

Technical:

13) **Modelling high dimensional data risks overfitting errors, but can also trade-off statistical power and miss causal relationships,** e.g. salient signals and/or transduced responses.

14) **We lack a sufficient understanding of all the parts in the interactome**. This is needed to mine interactions that are relevant to a response of interest. Do we have the technological tools (e.g. pipelines and high performance clusters) to collect and analyze interactions across scales? e.g., can measure changes in all the phosphorylation states of thousands of proteins, not clear what to look for in such dataset - much mining done manually, driven by "prior knowledge" of which proteins are relevant (biased approaches).

15) **Opportunities for experimental replication/perturbation are limited at highest levels of organization.** We can only observe a single global ecosystem, whereas tissue regeneration can be studied in many replicated model organisms**.** This challenge can also be seen as a motivating rationale for identifying 'rules' that could scale up levels of organization and across biological systems. It is likely that someday computational models become sufficiently accurate to account for ecosystems as a whole, and getting there likely involves generalizing the learnings and insights from studies at lower scales.

8. What might be the broader impacts?

Public Health:
An ability to detect relevant vs irrelevant signals is a means for us to understand multiple facets of human health and disease.
Individualized or personalized medicine can benefit, based on an increased understanding of individual variability at the genetic and environmental levels
Predicting the spread of infectious diseases at multiple scales from local communities to world-wide.
Using big data approaches to better predict who will become sick, who will respond best to treatments, and what other factors affect disease outcome are potential applications, i.e., personalized prognosis of wellness.

Environmental impacts:
Understanding how homeostatic mechanisms impact different aspects of a given environment may help us to predict future effects of climate change. We may better predict factors that may lead to loss of species or critical impacts of loss of habitats. This initiative may identify predictive factors in large-scale biological phenomena. For example, are homeostatic mechanisms found in evolution? How might one harness those for sustainable management of ecosystems?

Beyond Biology and Health:
Can approaches that are applied to biological phenomena be useful to other disciplines? For example, could general models developed through this initiative be applied to economics, in order to predict how or whether the flux of financial resources around the world will impact specific aspects of global economies, or to predict which factors will become relevant in a financial crisis and to predict the outcomes of such global economic events. An ability to understand homeostatic mechanisms generally can be applied to AI and machine learning, as well.

There are many applications beyond biology for which secure, adaptive, and robust communication is essential. Recent studies demonstrate that nature holds underexploited possibilities for bioinspired design principles (e.g. Brady et al. 2015). Identifying the general and scalable rules outlined in this vision paper could therefore translate into other realms - abetting, for example: a) the effective dissemination of information in the presence of misinformation or propaganda, b) the coordination of responses to natural disasters, acts of violence, and other emergencies to efficiently and reliably converge on a state of safety c) engineering autonomous vehicles and other robotic systems to effectively parse relevant signals from noise during operations, and d) maintaining cybersecurity.

9. What disciplines might be needed for such a reintegration of biology?
All domains of Biology - Molecular and Cell Biology, Physiology, Plant Biology, Neuroscience, Immunology, Microbiology, Ecology, Biochemistry, etc.
Genomics, Proteomics, Omics of various kinds

Applied Mathematics and Physics
Network science, Computer sciences, Graph theory, Information theory
Systems Dynamics and Nonlinear systems analysis
Engineering and Instrumentation
Data sciences and Statistics
Humanities
Ethics

10. How does the pursuit of above scientific questions reintegrate biology?

We aim to establish fundamental rules with which to understand, and even predict, which signals are salient in a given context and how those signals are affected by homeostatic processes. Undertaking this task will require the foundation of an interdisciplinary field which will include experts from a range of scientific disciplines, and will necessitate understanding how signals are sent and interpreted in a wide range of areas. We need to understand for any given system the signals that should be broadcast, who and at which times the signals should be sent and who should be able to receive and act upon the respective signals.

11. Who should care about this vision and why?

Policy makers - so that priorities can be established for tackling the grand challenges.
Institutional administrators - so that appropriate resources can be allocated to drive multidisciplinary engagement.
Biologists - so that they can collect and interpret data at much larger-scale than is routinely done and therefore conduct truly integrative biology.
Practitioners in the needed fields - so that they can be recruited to tackle the grand challenges.
Curriculum developers - so that they can incorporate the multidisciplinary ideas into didactic elements and learning frameworks.
Trainees - so that they can seek learning experiences that span appropriate disciplines and aimed at the grand challenges.

12. What institutional changes are needed to make the proposed vision a reality?

There are several institutional and national scale efforts needed to actualize an effective pursuit of the proposed vision. For a detailed discussion on this topic, we direct the reader to another Vision Paper in the present series that is specifically focused on the evolving the institutions towards reintegrating biology. A summary is included here to highlight the multidisciplinary aspects, which are essential to tackle the big question and to pursue the scientific opportunities outlined here.

While it is becoming more common to provide incentives for interacting with colleagues in other sub-disciplines, a comprehensive approach is needed. Some of the examples discussed include: revising promotion and tenure standards to reward early career researchers

for collaborative team-based research; promoting a culture of engagement of faculty, students, researchers, and administrative leadership across departments and colleges; make co-advisement of students across disciplines the norm, not an exception; promote collaborations across biological scales and not just disciplines, e.g., between cell biologists and ecologists, neuroscientists and humanities, pathologists and population biologists, etc.; train faculty and students for communicating science to a broad audience so that they can effectively engage policy makers and society.

At the National Science Foundation level, it is essential to publicize and communicate widely to the institutional leadership of the present national scale efforts to reintegrate biology. Develop policy frameworks and highlight best practices to lead the path towards transforming institutional processes and structures towards enabling multidisciplinary engagement and training as the foundational core to which everything else is connected. Specifically, the federal and other funding agencies should realign their programs to consider multidisciplinary science as an essential requirement to influence the institutional administrations, whose priorities are partly governed by financial aspects.

On the training front, there is a dire need to take a big leap in our collective thinking about what it means to train in biology. It is no longer sufficient to formulate curricular transformation towards multidisciplinary aspects as needing so-called biologists to take electives on various technical skills (e.g., data sciences, machine learning and artificial intelligence, modeling and simulation). We need to get to a point where multidisciplinarity should be an implicit and default expectation of biology. A step towards that goal is to teach biology in truly multidisciplinary way through and through in order to redefine what it means to be a 'biologist'. Importantly, the biology curriculum should include training in philosophy of science to empower the learners with conceptual thinking that is integrative and focuses on generalized principles.

Below section contains the detailed discussion notes addressing the present topic, and is included in a largely unstructured form, reflecting the wide range of ideas that were brought to bear on the question: which research directions and scientific questions need to be pursued to comprehensively answer the big question?

How might we discover “rules of life” governing the (use, function, manifestation, regulation, propagation, encoding, decoding) of biological signals across a wide array of biological contexts?

If such rules exist, this would indicate that there are universal constraints on signaling, even in the face of evolutionary changes across large timespans in the complexity and variability of individuals and populations. If our search does not turn up broadly generalizable rules, then

what are the contexts and variations that influence when a given signal is important, and when it is just noise?

How do the constituents of a biological collective broadcast, census, and process information to drive the function of an integrative whole? How might we identify most salient subset of signals for a given phenotype?

How to go from the high-resolution, high-throughput, high-dimensional data that we can collect to the subset that impacts the phenotype of interest? When we can measure everything, how to identify what matters for the function of the integrative whole?

Can we classify the types of signals and signaling events is such a way that we can then explore these as different phenomena for commonality across organisms and at different scales. For instance, one type would be the single phospho site or master transcription factor. How many types can be enumerated? Molecular, Electrical, Mechanical, Magnetic? (birds), Electromagnetic (e.g., luminescence), Sound (is this mechanical?), Heat/Temperature (is this a stimulus/cue or a signal?), … … ?

All the molecules in a cell are fair game for inclusion into a signalome. Normally several hundreds of molecules are secreted or transported across cell boundaries.

How might we categorize the signal-response pair sets? How to do this in a way that leads to more general understanding of what the signal-response interaction means in a specific context/function?

Much has been done in molecular and cell biology to probe biological systems as signal-response pair sets (e.g., Alliance for Cell Signaling efforts, NIH Common Fund LINCS program). Extensive data available on the effects of knock-out, knock-in, over/under-expression of various molecules on cellular, tissue, and organismal phenotypes. In bacterial and unicellular eukaryotic systems, such approaches have been taken at genome-scale, i.e., systematic deletion of individual and pairs of genes in E. coli, yeast, etc. with extensive phenotyping of cellular scale response (e.g., survival, growth, metabolic capacity). Genetic approaches have been scaled similarly to develop large cohorts of genetically perturbed worm, fly, mouse (others?) along with extensive phenotyping of these laboratory models. Is this large corpus of data and knowledge useful for developing rules and principles of how to mine omics data on signals for predicting integrative function in a specific context?

Compare this across evolutionary hierarchy

E.g., host-pathogen interactions; symbiosis;

How might we figure out which signals have an outsized impact on phenotype at larger scale?

What has impact across scales? Scale-jumping effect…. Phospho site on one protein having an impact on whole cell behavior, a minor subset of cells affecting response/phenotype of whole organ. E.g., master regulators (transcription factors) in cancer.

Another would be the type where no one signal among perhaps hundreds, but some number of them must be detected to elicit a response.

How can we figure out what is the information in the subset of signals that matters to a receiver?

is the signal encoding the status of the sender? (e.g., infection in a specific location, metabolic need of muscle, food location signal sent by an ant, danger signal sent by a deer in a herd, sexual interest status of a male peacock, etc.

is the signal encoding the status of the whole (larger than the individual senders?) e.g., immune cells in circulation, overall activity of a motor fiber, resonance and synchrony in quorum sensing, metabolic capacity of whole liver/kidney/lung (as reflected in the level of urea, oxygen, salts, etc. in the circulation)

How do we integrate signalome with interactome? Are there rules to govern the hierarchy of functional signals in different contexts? Interactome the other-half of the signalome, and we need both to make sense of integrated function.

In the intracellular context, which signals carry 'history' - i.e., beyond instantaneous information about level. For example, modeling has shown that instantaneous levels of mRNA in single cells can be informative of the temporal history of signaling kinase activity upstream to gene regulation (Makadia et al., 2015). This contrasts with experimental data of Cheong et al. (2011) that found that downstream signaling effectors have at best binary on-off decoding ability on whether a receptor has been activated.

What are the properties of the signal subset that can inform the individual entities about the functionality/capacity/amount of the whole?

By definition, this has to be integrated (cumulative) signal that sums up over signals from many entities. e.g., bioluminescence in quorum sensing; injury-graded cytokine signals in a tissue; loudness of a sound indicative of number of individuals shouting in the group.

Compensation is an inherent aspect of Homeostasis that can complicate how we can interpret the signals that inform about the whole - i.e., notable differences in interpreting the response of closed loop versus open loop systems. In a system with feedback interactions, the dynamics of signalome encode not just the status of the whole and the individual entities, but also how the status is being achieved i.e., dependent on which feedback loops are triggered/operational.

Organ scale: Liver regeneration example

What constitutes "metabolic demand" at the molecular scale? No single metabolite or protein tracks liver volume or mass (due to compensation that is inherent in homeostatic systems). e.g., glucose homeostasis can be achieved in principle by storage versus consumption. Each mode of achieving homeostasis would change the signalome (e.g., levels of circulating lactate upon glucose consumption by muscle versus glucose storage in liver would correspond to distinct identify and nature of the circulating signals).

Intracellular scale: Competing endogenous RNAs - decoy targets - that titrate away microRNAs - to enable switch of functionality of microRNAs to affect phenotypic switching
This provides a potential mechanism for connecting integration of external signals to changes in phenotype once a threshold is exceeded.

Neuronal correlates of learning example: Neurons change production of certain genes and proteins in response to changes in stimuli, and restructure their dendritic projections and synaptic connections in a dynamic way. This phenomenon is a means of adjusting cellular physiology in order to adapt to changes outside the organism. The intracellular response to learning is a compensation in response to changes.

Organism scale: Homeostasis occurs in development: a specific individual can develop normally despite one or more defective genes, or an environment that is deficient in something normally required. Are there universal signals that are modulated in these cases in order to facilitate normal development, or would these always necessarily be specific to what was lacking? Does such modulation constitute a homeostatic mechanism that helps an individual survive, and are these modulatory signals necessarily inherited, or are their a signal that is more like learning at an organismal scale?

Multi-organism scale: Consider the constantly evolving host-pathogen interactions that exist for instance in plant and bacterial pathogens. Organisms that are separated by millions or billions of years of evolution retain the molecular patterns that enable sampling and selection for evading host immune responses by the bacterium and adaptation by the plant in developing resistance traits.

How multiple signals across modes need to be coordinated to achieve an objective? What are the challenges in making sense of such data at large-scale?
Correlations between signals across modalities - would that teach us about rules/principles in ways that correlations between signals within a given modality
    e.g., dog barking to communicate to another animal
What is the parallel for multiple modes of signal communication in cellular/molecular scales?
Multi-organ coordination via hormonal signals
    Insulin and glucose homeostasis
Neural signals
Mechanical signals
Multi-modal communication in animal groups - theoretical approaches

Scientific opportunities for methodological developments
How to scale machine learning approaches to address these questions? Dimension reduction approaches used in machine learning can provide pointers to combinations of input signals that most significantly impact the phenotype of interest.
Network science and Graph theoretical approaches?

Systems dynamics and nonlinear systems analysis methods useable at large scale? Any good examples? Physiome project, and Virtual Physiological Human efforts come to mind. As do the genome-scale metabolic modeling and whole cell modeling efforts.
How can data science methods for storage/annotation help?

A selection of relevant literature from the discussions:

Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: an appreciation and an update. *Trends in ecology & evolution*, *28*(12), 712-718.

Bolnick, D. I., Barrett, R. D., Oke, K. B., Rennison, D. J., & Stuart, Y. E. (2018). (Non) parallel evolution. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 303-330.

Brady, P. C., Gilerson, A. A., Kattawar, G. W., Sullivan, J. M., Twardowski, M. S., Dierssen, H. M., ... & Ibrahim, A. (2015). Open-ocean fish reveal an omnidirectional solution to camouflage in polarized environments. *Science*, *350* (6263), 965-969.

Bierbach, D., Laskowski, K. L., & Wolf, M. (2017). Behavioural individuality in clonal fish arises despite near-identical rearing conditions. *Nature communications*, *8*, 15361.

Laider and Johnstone (2013) Animal signalling. *Current Biology*, *23 (18)*, R829-R833.

Ng, Wai-Leung, and Bonnie L. Bassler. "Bacterial quorum-sensing network architectures." *Annual review of genetics* 43 (2009): 197-222.
https://www.annualreviews.org/doi/abs/10.1146/annurev-genet-102108-134304
(quorum sensing)

Salmena, Leonardo, et al. "A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?." *Cell* 146.3 (2011): 353-358.
https://www.sciencedirect.com/science/article/pii/S0092867411008129
(ceRNA hypothesis)
Di Leva, Gianpiero, and Carlo M. Croce. "miRNA profiling of cancer." *Current opinion in genetics & development* 23.1 (2013): 3-11.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632255/
(miRNA profiles and cancer phenotypes)

Makadia HK, Schwaber JS, Vadigepalli R. Intracellular Information Processing through Encoding and Decoding of Dynamic Signaling Features. *PLoS Comput Biol.* 2015 Oct 22;11(10):e1004563.
https://doi.org/10.1371/journal.pcbi.1004563
(information encoding decoding in a gene regulatory network)

Cheong R, Rhee A, Wang CJ, Nemenman I, Levchenko A. Information transduction capacity of noisy biochemical signaling networks. *Science*. 2011 Oct 21;334(6054):354-8.
https://science.sciencemag.org/content/334/6054/354.long
(information encoding decoding in a signaling network)

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, …, Ding L. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018 Apr 5;173(2):371-385.e18.
https://doi.org/10.1016/j.cell.2018.02.060
(Comprehensive Characterization of Cancer Driver Genes and Mutations)

Capp, Jean-Pascal. "Tissue disruption increases stochastic gene expression thus producing tumors: Cancer initiation without driver mutation." *International journal of cancer* 140.11 (2017): 2408-2413.
https://www.ncbi.nlm.nih.gov/pubmed/28052327
(signaling, stochastic gene expression and cancer)

Leonard AS, Dornhaus A, Papaj DR 2011 'Forget-me-not: complex floral signals, inter-signal interactions and pollinator cognition' Current Zoology 57: 215-224

Leonard AS, Dornhaus A, Papaj DR 2011 'Why are floral signals complex? An outline of functional hypotheses', in: Evolution of Plant-Pollinator Relationships. Patiny, S. (ed.) Cambridge University Press

Sih A, Ferrari MC, Harris DJ. (2011). Evolution and behavioural responses to human-induced rapid environmental change. *Evolutionary applications*, *4*(2), 367-387.