Integrative and Comparative Biology

Integrative and Comparative Biology, volume 61, number 6, pp. 2011–2019
doi:10.1093/icb/icab114

Society for Integrative and Comparative Biology

JUMPSTART NSF INVITED

The Axes of Life: A Roadmap for Understanding Dynamic Multiscale Systems

Sriram Chandrasekaran [*,1] Nicole Danos,[†] Uduak Z. George,[‡] Jin-Ping Han,[§] Gerald Quon,[¶] Rolf Müller,[‖] Yinphan Tsang[‖‖] and Charles Wolgemuth[**]

*Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA; [†]Department of Biology, University of San Diego, San Diego, CA, USA; [‡]Department of Mathematics & Statistics, San Diego State University, San Diego, CA, USA; [§]IBM TJ Watson Research Center, Ossining, NY, USA; [¶]Department of Molecular and Cellular Biology, University of California—Davis, Davis, CA,USA; [‖]Department of Mechanical Engineering, Virginia Tech, Blacksburg, VI, USA; [‖‖]Department of Natural Resources and Environmental Management, University of Hawai'i at Mānoa, Honolulu, HI, USA; **Departments of Physics and Molecular and Cellular Biology, University of Arizona, Tucson, AZ, USA

[1]E-mail: csriram@umich.edu

Synopsis The biological challenges facing humanity are complex, multi-factorial, and are intimately tied to the future of our health, welfare, and stewardship of the Earth. Tackling problems in diverse areas, such as agriculture, ecology, and health care require linking vast datasets that encompass numerous components and spatio-temporal scales. Here, we provide a new framework and a road map for using experiments and computation to understand dynamic biological systems that span multiple scales. We discuss theories that can help understand complex biological systems and highlight the limitations of existing methodologies and recommend data generation practices. The advent of new technologies such as big data analytics and artificial intelligence can help bridge different scales and data types. We recommend ways to make such models transparent, compatible with existing theories of biological function, and to make biological data sets readable by advanced machine learning algorithms. Overall, the barriers for tackling pressing biological challenges are not only technological, but also sociological. Hence, we also provide recommendations for promoting interdisciplinary interactions between scientists.

Introduction

How do we define life quantitatively? All living systems fall into a multidimensional space defined by scales, factors, and biological components. To understand life, we must be able to integrate complex biological processes across diverse scales—Physical (e.g., Spatial and Temporal), Chemical, and Biological (Box 1). In addition to multiple scales, extrinsic factors such as environmental stressors and noise can impact a system. Finally, the response of a system depends on its components, from molecules, cells, individuals, communities, populations to ecosystems. We posit that knowledge of these three dimensions, that is, biological components, factors that act on a system, and the scale of a system (Fig. 1), is necessary and sufficient to predict a system's behavior. The axes framework can serve as a universal vocabulary for defining and comparing biological systems.

Most biological phenomena span multiple dimensions of the axes of life, exhibiting various degrees of emergence, self-organization, robustness, resilience, and complexity (Kauffman 1992; Stelling et al. 2004; Mazzocchi 2008; Wolf et al. 2018). A classic example of a challenge involving multiple dimensions of the axes is found in healthcare. Most diseases involve the dysfunction of biological components at multiple scales from genes to organ systems usually in response to external stressors like infection or diet. The actions of those altered processes change the behavior of cells, which then lead to systemic effects within the body over time.

Multi-dimensional problems are ubiquitous in biology. For example, an integrative analysis of numerous genetic components and environmental factors is needed to tease out the effects of nature and nurture in development (Robinson 2004). This can help

Advance Access publication May 28, 2021

Box 1: Terminologies

The axes of life: A framework for comparing biological systems based on their components, scale, and factors acting on the system.

Scale: A physical (e.g., micron and seconds), chemical (e.g., molecular weight), or biological (e.g., number of generations) unit of measurement.

Component: A distinct biological unit with a specific function in a system (that can be acted upon by evolution) (e.g., protein, gene, or organ).

Factors: These are external forces or agents that act on a system and change its position in the axes of life (e.g., diet, drugs, social interactions, and climate change).

Multi-dimensional/multi-axes: This refers to challenges, datasets, or models that span all three axes' dimensions (e.g., climate change).

Multi-scale: This is a subset of multi-dimensional systems that span multiple scales (spatial and temporal)

answer why similar mutations during development sometimes lead to different disease symptoms after birth (Kammenga 2017). Other examples that span multiple dimensions include predicting from genotype the phenotype of an organism or organismal community, and predicting how global temperature change affects organismal behavior and ecosystem balance. Even seemingly simple biological processes, such as fish swimming, are multi-dimensional, involving diverse biological components (muscles and nerves), physical scale (muscle fibers and whole-body mechanics), and biological factors (group swimming behaviors on neural stimulation of muscles) (Massarelli et al. 2017; Mekdara et al. 2018; Tytell et al. 2018).

The traditional research paradigm focuses on fixing two axes and varying the third. For example, studying a bacterium exposed to an external perturbation fixes both the scale and components, and modulates the factors. Axes are also typically fixed in an experiment to assign causality and to reduce complexity (Platt 1964). Even within one Axis of Life, the existence of interaction effects (e.g., epistatic genetic effects on phenotype) are well documented but hard to study due in part to small datasets. The number of possible interactions explodes as multiple scales are considered (combinatorial complexity). While fixing axes improve the tractability of studying a system, it also limits the linking of data across scales or components. Theoretical and empirical frameworks are needed for looking across the axes and making educated hypotheses about which connections across scales might be most fruitful to experimentally explore.

How do we foster and enable new research that will effectively bridge across the axes? Iterative dialog of experiments and computation will allow us to determine generalizable principles to predict responses of biological systems. Here we propose a framework for predicting the behavior of such multi-dimensional systems. We will focus on combating four key impediments limiting our understanding of dynamic multiscale systems. This ultimately requires iterative interactions between diverse disciplines and between Data, Methods, and Theory. This includes:

- multidimensional data generation and management—generating, curating, and disseminating relevant and high-quality data across multiple disciplines, scales, factors, and components;
- theoretical frameworks for synthesis—developing theoretical framework that synthesizes data to drive experimental hypotheses;
- methods to bridge the axes—developing and applying methods that integrate multi-dimensional datasets and models to drive research in biological systems; and
- interaction across disciplines—to foster these goals, a culture of science is needed that educates, supports, and values integrative and interdisciplinary approaches.

Here we will address key questions in every step of this process of understanding dynamic multiscale systems.

Multi-dimensional data generation and management

The first step in understanding a system is to define its location on the axes of life. This requires data generation methods for characterizing its components, scale, and factors that influence the system behavior. An integrative approach to quality data generation and management has the potential to provide bridges between disciplines, breaking through structural and theoretical bottlenecks. There are several bottlenecks identified, including choices of data that are appropriate to the system under study, accessibility, and comprehensibility of appropriate data by interdisciplinary communities,

the need for incorporation of quality measures at all steps from data collection to model generation, and the continued need of exploratory experimental work to support and drive integrative approaches.

Briefly, current data generation practices provide limited representation of all three axes. For example, in conservation ecology, incorporating data across realms (i.e., terrestrial, aquatic, and coastal, and marine) is necessary to provide a holistic view of the ecosystem and possible strategies in conserving and managing the system effectively (Tsang et al. 2019). Traditionally, researchers have been trained to focus on collecting the measurements within a specific discipline. However, the ecosystem is rarely isolated in operation, the connections among ecosystem realms are explicit and inexplicit.

To overcome the limitation, we recommend funding for collaborative studies that span all three axes. As a concrete example of how this might be done, we give a hypothetical research design from a field where these effects are ever-present, the function of the brain. To understand the development and maintenance of memory, we need to know how external cues (axes: factors operating at organismal and seconds scale) alter neuronal firing and the alteration of dendritic spine morphology, potentially impacting learning and memory (axes: components and cellular-level/minutes scale). The morphological changes, along with biochemical activity, then need to be examined over the course of days to months, during and in between learning activities, to determine what alterations occur and how they are mediated. With advances in the fields of genetically expressed fluorescent probes and intra-vital imaging, it is conceivable that in the near future it will be possible to affix a microscope to the head of a mouse, in such a way as to investigate these processes during un-anesthetized actions.

Other examples of hypothetical multi-dimensional research design include a consortium working together to study the impact of both global temperature change and local release of a toxin on microbial metabolism and ecosystem biodiversity over a decade; this study spans diverse scales (temporal, physical), factors (temperature and toxins), and components (molecules, microbes, and plants). Similarly, quantifying the fitness of a genome-wide knockout library of *Escherichia coli* against short- and long-term treatments of antibiotics, spans all three axes.

The multi-dimensional studies do not necessarily have to be large-scale consortium efforts. For example, studying the effect of a honeybee transcriptional regulator on neuronal transcription, brain metabolism, and colony social behavior as a function of diet, spans all three axes. Notably, a recent study on the behavior of honeybees incorporates datasets that span these three axes (Jones et al. 2020). The authors analyze numerous biological components (transcripts and chromatin modifications), quantified their variation based on social behavior of individual honeybees, thus linking molecular and organismal scales, and performed this study in various queenless colonies that exhibited considerable variation in bee behaviors (factors). This led them to understand the role of plasticity in gene regulatory networks on evolution of social behavior (Jones et al. 2020).

These recommendations on data generation go against the traditional view of fixing various factors or components, and experimenting in a controlled environment. Varying a single factor at a time is an essential part of strong inference (Platt 1964; Beard and Kushmerick 2009). For example, traditional studies on transcriptional regulation rely on perturbing transcription factors individually to identify causal regulation. However, recent studies have used information theoretic tools to tease out the effects of hundreds of transcription factors without the need for perturbing one factor at a time (Chandrasekaran et al. 2011; Marbach et al. 2012; Jones et al. 2020). While this traditional approach has been fruitful, it nevertheless limits the creation of theories that span the axes of life. With the advance in methodologies and computational power, tackling the complexity that span the axes is possible, and the study systems are closer to reality.

Theoretical frameworks for synthesis

Before multi-dimensional systems are modeled, a feasibility study should be conducted to ensure the system can be causally inferred and simulated (e.g., is predictable). By "predictable," we do not imply that the system's behavior can be forecasted with 100% accuracy. Rather, quantifying the extent of predictability can ensure that we have identified some of the causal factors that can continuously or transiently influence the dynamics of biological systems. For example, predicting the phenotype of an organism is not possible unless relevant variables that influence the phenotype (genome sequence and environmental factors) are determined. If a system is not predictable given a predefined set of measurements, then it may not be worth studying until we identify the input data and its critical variables needed to make it predictable.

Through iterative model-driven experimental data generation, any system can be made more predictable. This strategy was used to improve the regulatory network models of *E. coli* and *Saccharomyces cerevisiae* (Covert et al. 2004; Chandrasekaran and Price 2013). However, in some cases, our ability to predict a phenotype from genotype may be limited if the phenotype is strongly driven by noise (Eldar and Elowitz 2010; Chalancon et al. 2012).

Interpretability of the framework is not necessary at this stage; for example, black-box neural networks in conjunction with techniques like cross-validation can be used to broadly determine whether a system is predictable. Once the predictability of a system is established, techniques such as interpretable Artificial Intelligence (AI) can be applied to identify the patterns and build mechanistic models (Ribeiro et al. 2016; Yu et al. 2018).

Theoretical frameworks for reasoning about the predictability of systems should be generalizable and nevertheless make specific predictions/hypotheses about each problem. Frameworks for determining the predictability of a system can be either derived from fundamental principles or empirical (data-driven) (Horgan 1995). An empirical framework for defining the predictability of a system can be any method that takes one or more measurements as input, and predicts one or more output measurements. Empirical frameworks for integrating heterogeneous datasets can be broadly grouped into two categories (1) whether measurements on different scales can be made on the same entities or (2) when different sub-populations or biological replicates can be measured. When measurements at different scales can be made on the same entities, strategies from the fields of multimodal learning can be applied (Min et al. 2017). Otherwise, strategies from manifold alignment can be used to construct models of biological phenomena at individual scales, and then alignment is performed to identify connections between scales (Welch et al. 2017). This recognition and development of such strategies are particularly important, as many previous studies and efforts have collected data at local or individual system scale and within a single discipline.

Having an adequate approach and framework to bridge and integrate the existing data across disciplines will allow the best use of precedent knowledge. For example, assessing the impact of climate change on the stream habitats that support stream fish population requires linking organism and ecosystem scales. Previous studies have accumulated abundant biological survey data from local and state studies. At the same time, US Geological Survey has been continuously recording long-term hydrological data, such as daily streamflow and stream temperature data nationwide. Recently, Tsang et al. developed a framework to integrate these local and national efforts (Tsang et al. 2021). This study showed that data from either scale alone could not predict the future climate impact on the changes on stream hydrologic and thermal habitat conditions, and the possible impact on the fishes species they support (Tsang et al. 2021). Similar approaches and concepts can be applied when dealing with ecosystems level problems.

Alternately, theoretical frameworks can be derived from first principles of evolution, chemistry, mathematics, computer science, or physics. For example, studying the biosonar system of bats involves studying how the physical attributes of the ears transform the incoming ultrasonic echoes to encode sensory information, and how the ear shapes of bats have diversified in the course of evolution in 1400 different species (Müller 2010; Gao et al. 2011; Ma and Müller 2011). Understanding this process involves integrating data from various physical and temporal scales (acoustics), diverse ear components, and how bat behaviors (factors) actively modulate ear structure. These first principles can also provide limits on predictability. For instance, predicting electron transfer in proteins may not be completely possible based on the Heisenberg uncertainty principle. Another set of examples are the computer science proofs of computational complexity ("NP-hard problems") for 3D structure prediction problems (Torrisi et al. 2020), which then help focus efforts on finding approximate solutions to difficult problems.

Mathematical modeling and simulations have enormous potential to unravel the complex interactions of biological phenomena occurring at different scales (Wooley and Lin 2006; Voit 2019). An emerging type of mathematical model called multi-scale models has allowed the linking of models at different biological scales, from molecular-scale processes like protein folding to entire organisms (Walpole et al. 2013; van Gestel and Tarnita 2017). For example, multiscale modeling has been used to understand the interactions across different scales that are necessary for development of organs and diseases (Schnell et al. 2007; George et al. 2015). However, there are limited tools that allow the coupling/integration of models across axes dimensions. Moreover, the modeling and analysis of interdependent biological systems also require mathematical models capable of identifying the causal influences and capturing either the Markovian or non-Markovian dynamics of some biological constituents. The implementation of

new techniques that link multiple axes dimensions to study system-level outcomes would be invaluable in understanding the complex interactions of biological systems.

Another key challenge with new high throughput technologies is that they generate thousands of correlations between biological components. We currently lack mechanisms to rapidly validate these correlations and assign causality (National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Board on Life Sciences 2020). This requires theoretical frameworks that bring together orthogonal high-throughput datasets that can then be cross-verified against each other to uncover functional or causal relationships. Some recent attempts at this process include a study that used a mathematical model of *E. coli* to reconcile thousands of transcriptomics, proteomics, and enzyme kinetic measurements with physiological measurements (Macklin et al. 2020), and a systems biology approach called GEMINI that reconciles transcriptional regulatory interactions from high-throughput studies with metabolic data (Chandrasekaran and Price 2013), Bridging across axes will enable us to harness a wider range of datasets for rapid evaluation of correlations and determine causality.

In general, current theoretical methods lack the ability to transition between axes dimensions as we lack an underlying "objective" for models (Feist and Palsson 2016). For example, do all living systems maximize their biomass production, energy efficiency, degree of emergence, self-organization, complexity, and intelligence? These principles can be represented mathematically but may not be accurate biologically. It is unclear if given a genome sequence and environmental factors as inputs to such a model, a complex cell or human being would appear naturally as an output.

Most of the mechanistic models of biological systems have not been validated against the system they describe. This is sometimes due to the inability to generate relevant data for model testing and validation, but it may also be due to a lack of access to data because it is not publicly available. This brings us back to the problem of not having better curated and publicly available data that can be accessed across researchers working in different disciplines.

Methods to bridge the axes

Development and application of methods for integrative projects pose many unique challenges. Several fundamental concerns must be addressed that are often taken for granted in traditional systems. For example, it is often difficult to define appropriate objectives for studies bridging the systems, since the datasets being integrated may approach the system from orthogonal directions. The levels of spatial/temporal scale may be so different that connections are not obvious.

The ground-level challenge is to define the starting scale of input data and final scale of our integrative model, and then look to a theoretical framework and practical methods to build bridges between these levels. These differences may occur in multiple aspects of the system under study, as exemplified by the conceptual Axes of Life outlined above (Fig. 1) involving scales, factors, and components of biological systems.

As we build bridges between axes, we also need to define and incorporate the granularity of the approach, defining points along the range of scale of the model/system that are necessary to include. As part of the National Science Foundation (NSF) JumpStart meeting, the team repeatedly brought up the challenges (and potential) of integrating work across wide scale ranges, and whether it is possible to ignore features and intermediate scale levels in the approach. For example, for predicting a physiological-scale phenotype (e.g., cancer) from molecular-scale genotype (DNA sequence) one may ignore explicit modeling of the cellular scale. The challenge then centers on the questions: How do we link across scales or components? Can we infer anything about scales that cannot be measured? Will links emerge naturally, or do we need to forge undiscovered links in our method/model? For example, in mechanistic models of metabolism, the phenotype (growth of a cell) naturally emerges from interactions between molecular components at a lower scale (Karr et al. 2012; O'Brien et al. 2015). In contrast, in empirical models that link mutations at the molecular scale to a physiological phenotype, such as cancer (Vogelstein et al. 2013) or social behavior (Jones et al. 2020), the links are "imposed" by the scientist.

The potential of applying AI to these challenges promoted a vigorous discussion at the JumpStart meeting. In particular, development and application of transparent approaches to look inside the current AI Black box was identified as a central goal, and is described in more detail in another paper. Briefly, some strategies to make AI transparent include linking traditional models based on biological or mathematical principles with machine learning models (Yang et al. 2019; Zampieri et al. 2019; Oruganty et al. 2020). For example, by linking a machine learning model of antibiotic action with a

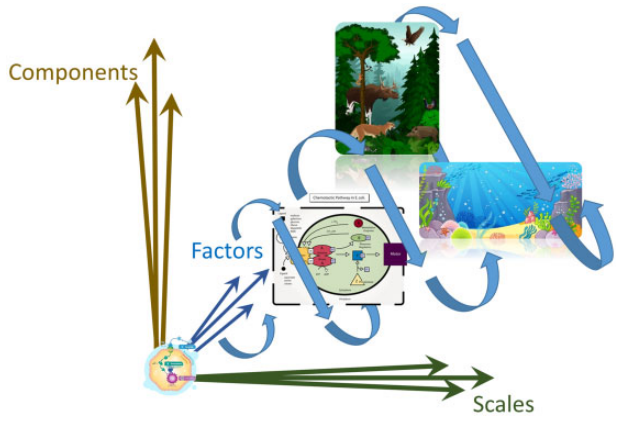

Fig. 1. The axes of life. Biological systems span three orthogonal axes spanning various scales (size and time), number of components and interactions, and influenced by external factors. Factors and components may operate at various scales. Characterizing a system based on these three axes is necessary to predict its behavior. The components axis can be considered a measure of biological complexity of the system as defined in different biological disciplines. For example, in ecology, the complexity of a system is proportional to the number of species and the number of interactions among them. Similarly, the complexity of the gene regulatory network is a function of the number of transcriptional regulators and their interactions (Szathmáry et al. 2001) (Diagram of the chemotactic pathway in *E. coli* modified from Falke et al. (1997)).

mathematical model of host immunity, Cicchese et al. were able to integrate molecular- and cellular-scale datasets from both pathogen and host (components), and create a mechanistic picture of the impact of various antibiotic treatments (factors) on pathogen clearance inside the lung infection site from a few days to a month (temporal scale) (Cicchese et al. 2021). Alternative strategies to make AI transparent include altering the structure of the AI model directly based on prior knowledge of the biological system (Ma et al. 2018).

All of the typical challenges with managing data are multiplied in integrative approaches, and the flexibility of methods to deal with these challenges will be necessary. For example, how do we deal with noise? Noise operates at multiple spatio-temporal scales and on various components, each of which must be quantified in unique experimental manners, and will need to be mined and integrated in a consistent way into the resultant synthesis. Methods to handle small and inconsistent datasets are also essential, since integrative efforts are often focused on data-poor nascent fields that are under rapid development and may require integration of results from multiple groups (Sung et al. 2012).

As we move toward bridging across the axes of life, a critical first step involves selection of appropriate, tractable systems. Most biological processes are complex (in the sense that their rate of change may not only exhibit various degrees of nonlinearities, but also a non-trivial combination of Markovian and non-Markovian dynamics). To make headway, we should seek systems that are simple enough that we can isolate specific behaviors and processes, while still being complex enough to require observations that span the axes dimensions. Overall, the advantages of focusing on a few model systems like *E. coli* should be carefully weighed against its limitations. For example, we may miss novel biological phenomena seen in exotic systems like archaea or aplysia that can lead to fundamental new insights.

One field where these questions might be currently addressable is neurobiology, where the action of individual neurons influences organismal behavior. The nematode *Caenorhabditis elegans* possesses a relatively simple neural architecture that can be easily visualized, and some researchers have already begun to explore how activating light-sensitive ion channels affect behaviors, such as motility (Sejnowski et al. 2014). Along the same lines, multiscale neuronal analysis has revealed that brain regions exhibit long-range memory/non-Markovian and multifractal characteristics.

Work by Eric Tytell and collaborators attempts to use a version of our proposed framework to study seemingly simple biological processes—fish swimming. Yet our understanding of this process has been impeded by its multi-dimensionality. The Tytell lab's work focuses on using a well-understood model organism, the lamprey. Because of the relative simplicity of the lamprey's structure (cylindrical body geometry, a non-segmented notochord, cuboidal muscle blocks, and a well-characterized spinal central pattern generator) the lab has been able to integrate isolated muscle experiments with mathematical modeling to quantify the effects of neural muscle stimulation on body mechanics (Tytell et al. 2018). They have applied computational-fluid dynamics modeling to quantify the effects of body mechanics on swimming behavior (Hamlet et al. 2015) and the feedback of swimming behavior on neural stimulation of muscles (Massarelli et al. 2017). The lab is even extending the studies to multiple individual fish swimming together (Mekdara et al. 2018). This integrative approach examines a biological problem across diverse biological scales, components, and factors. This has been facilitated by the collaboration among modelers and experimental scientists and has the

potential to inform the evolution of other body forms over time.

Similarly, Jones et al.'s study highlighted earlier utilizes cutting edge technologies to bridge the axes (Jones et al. 2020). They used an automatic behavior monitoring system to track individual honeybees in a colony and used convolutional neural networks (a type of machine learning algorithm) to quantify individual behaviors. These behaviors at the organismal scale were then linked to molecular scale measurements of gene expression and gene regulation for each individual bee using a gene regulatory network model built using an information theoretic approach called ASTRIX (Chandrasekaran et al. 2011). The authors then predicted individual behaviors solely based on the expression of transcription factors using another machine-learning algorithm called Random Forests. This study spanning multiple axes dimensions was made possible thanks to a diverse team comprising entomologists, bioengineers, genome scientists, data scientists, and bioinformaticians.

Interaction across disciplines

In addition to all the above, we agree there are barriers when bringing together all disciplines, institutions, departments, programs, and even sources of funding to deal with all the above barriers. These barriers exist because of the differences among all disciplines, such as language, terminology, and definition.

It could also be because of self-imposed barriers that limit interactions among the disciplines. Our tendency to gravitate toward like-minded individuals reduces cross-pollination that could bolster advances in interdisciplinary science. These interaction barriers also arise from academic cultural differences and from the physical separation of different disciplines that occurs at most institutions. In addition, different disciplines may approach similar problems from different perspectives, which causes a separation in focus when different disciplines try to answer similar questions. The agencies and sources of funding set their priorities, while researchers are driven to different emphasis and goals in question.

One short-term goal that can be achieved is to develop a general "match-making" system for helping researchers identify possible collaborators with complementary expertise (but similar research interests). Such a system would facilitate interdisciplinary collaboration in an equitable way (e.g., less-established scientists with fewer connections can still identify new collaborations). Here, we propose that Google Scholar be combined with techniques from network science and natural language processing to automatically generate "page ranks" of related collaborators to a given individual.

Another goal that can be achieved is to create more interdisciplinary journals that are topic-related instead of methods or discipline related. This would allow researchers working on similar topics across different disciplines to have a common venue in which to publish and stay informed of advances in their area. Another goal would be to organize interdisciplinary research meetings/workshops and bring together people from different disciplines to work on similar topics.

Conclusion

Approximately three million years ago a sequence of genetic mutations began occurring that would eventually lead to the evolution of humans. Those molecular level events produced an organism with altered behaviors that over the course of millions of years would lead to an altered global climate, the development of novel plants, and animals (e.g., corn and dogs), and the destruction of others (e.g., the woolly mammoth and the dodo). The impact of these mutations cannot be understood without a full comprehension of the interplay across biological components, factors, and scales. How do we begin to understand these complex interdependencies?

Here we introduce the axes of life framework that may be broadly applicable in characterizing biological systems based on their components, scale, and external factors acting on the system. We recommend methods for data generation and integration that span the axes of life and enable the discovery of universal biological principles. Our proposed framework and guidelines do have certain limitations. In practice, the axes framework is limited to systems for which a reasonable estimate of the underlying components and system properties are available. Similarly focusing on data generation and integration in a few well-studied model systems might prevent us from discovering novel biological phenomena in exotic species.

Finally, there is one cautionary note. Great breakthroughs are rarely the result of actively trying to make a great breakthrough. Rather, they often come from asking questions that had not previously been asked or choosing to look at something that no one had looked at before. For example, the theory of evolution was not developed because Darwin sought to discover a rule of life; it came because his travels exposed him to an array of observations that enabled

him to deduce a common unifying thread. The Special Theory of Relativity was a result of Einstein asking himself what it would look like if he were to run along with a beam of light. If we try too hard to ask big questions, we may miss the smaller question whose answer may contain a deep truth. Last but not least, we must also question whether our current funding paradigm provides sufficient freedom to allow researchers to follow their instincts, to allow their curiosities to guide them toward discovery, instead of requiring them to select problems that have an easily sellable significance and high likelihood of success.

Acknowledgments

We thank David Goodsell and other participants in the NSF JumpStart for their insightful comments and feedback on this article. This work is the outcome of a workshop on reintegrating biology that was sponsored by the National Science Foundation.

Competing interests

Authors declare no competing interests.

Data and materials availability

All datasets are available in the manuscript.

References

Beard DA, Kushmerick MJ. 2009. Strong inference for systems biology. PLoS Comput Biol 5:e1000459.

Chalancon G, Ravarani CNJ, Balaji S, Martinez-Arias A, Aravind L, Jothi R, Babu MM. 2012. Interplay between gene expression noise and regulatory network architecture. Trends Genet 28:221–32.

Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE. 2011. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. Proc Natl Acad Sci U S A 108:18020–5.

Chandrasekaran S, Price ND. 2013. Metabolic constraint-based refinement of transcriptional regulatory networks. PLoS Comput Biol 9:e1003370.

Cicchese JM, Sambarey A, Kirschner D, Linderman JJ, Chandrasekaran S. 2021. A multi-scale pipeline linking drug transcriptomics with pharmacokinetics predicts in vivo interactions of tuberculosis drugs. Sci Rep 11:5643.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. 2004. Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–6.

Eldar A, Elowitz MB. 2010. Functional roles for noise in genetic circuits. Nature 467:167–73.

Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA. 1997. The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. Annu Rev Cell Dev Biol 13:457–512.

Feist AM, Palsson BO. 2016. What do cells actually want? Genome Biol 17:1–2.

Gao L, Balakrishnan S, He W, Yan Z, Müller R. 2011. Ear deformations give bats a physical mechanism for fast adaptation of ultrasonic beam patterns. Phys Rev Lett 107:214301.

George UZ, Bokka KK, Warburton D, Lubkin SR. 2015. Quantifying stretch and secretion in the embryonic lung: implications for morphogenesis. Mech Dev 138:356–63.

Hamlet C, Fauci LJ, Tytell ED. 2015. The effect of intrinsic muscular nonlinearities on the energetics of locomotion in a computational model of an anguilliform swimmer. J Theor Biol 385:119–29.

Horgan J. 1995. From Complexity to Perplexity. Sci Am 272:104–9.

Jones BM, Rao VD, Gernat T, Jagla T, Cash-Ahmed AC, Rubin BE, Comi TJ, Bhogale S, Husain SS, Blatti C, Middendorf M, et al. 2020. Individual differences in honey bee behavior enabled by plasticity in brain gene regulatory networks. Elife 9:e62850.

Kammenga JE. 2017. The background puzzle: how identical mutations in the same gene lead to different disease symptoms. FEBS J 284:3362–73.

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Jr, Assad-Garcia N, Glass JI, Covert MW. 2012. A whole-cell computational model predicts phenotype from genotype. Cell 150:389–401.

Kauffman SA. 1992. Origins of order in evolution: self-organization and selection. In: Understanding origins. Dordrecht, Netherlands: Springer Netherlands.

Ma J, Müller R. 2011. A method for characterizing the biodiversity in bat pinnae as a basis for engineering analysis. Bioinspirat Biomim 6:026008.

Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T. 2018. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 15:290–8.

Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Agmon E, DeFelice MM, Maayan I, et al. 2020. Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. Science 369:eaav3751.

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G, et al.; DREAM5 Consortium. 2012. Wisdom of crowds for robust gene network inference. Nat Methods 9:796–804.

Massarelli N, Yau AL, Hoffman KA, Kiemel T, Tytell ED. 2017. Characterization of the encoding properties of intraspinal mechanosensory neurons in the lamprey. J Comp Physiol A Neuroethol Sens Neural Behav Physiol 203:831–41.

Mazzocchi F. 2008. Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. EMBO Rep 9:10–4.

Mekdara PJ, Schwalbe MAB, Coughlin LL, Tytell ED. 2018. The effects of lateral line ablation and regeneration in schooling giant danios. J Exp Biol 221:jeb175166.

Min S, Lee B, Yoon S. 2017. Deep learning in bioinformatics. Brief Bioinform 18:851–69.

Müller R. 2010. Numerical analysis of biosonar beamforming mechanisms and strategies in bats. J Acoust Soc Am 128:1414–25.

{National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Board on Life Sciences}. 2020. Next steps for functional genomics: proceedings of a workshop. Washington (DC): National Academies Press.

O'Brien EJ, Monk JM, Palsson BO. 2015. Using genome-scale models to predict biological capabilities. Cell 161:971–87.

Oruganty K, Campit SE, Mamde S, Lyssiotis CA, Chandrasekaran S. 2020. Common biochemical properties of metabolic genes recurrently dysregulated in tumors. Cancer Metab 8:5.

Platt JR. 1964. Strong Inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146:347–53.

Ribeiro MT, Singh S, Guestrin C. 2016. Why should i trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York (NY): Association for Computing Machinery. p. 1135–44.

Robinson GE. 2004. Genomics. Beyond nature and nurture. Science 304:397–9.

Schnell S, Grima R, Maini PK. 2007. Multiscale modeling in biology - New insights into cancer illustrate how mathematical tools are enhancing the understanding of life from the smallest scale to the grandest. Am Sci 95:134–42.

Sejnowski TJ, Churchland PS, Movshon JA. 2014. Putting big data to good use in neuroscience. Nat Neurosci 17:1440–1.

Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J. 2004. Robustness of cellular functions. Cell118:675–85.

Sung J, Wang Y, Chandrasekaran S, Witten DM, Price ND. 2012. Molecular signatures from omics data: from chaos to consensus. Biotechnol J 7:946–57.

Szathmáry E, Jordán F, Pál C. 2001. Molecular biology and evolution. Can genes explain biological complexity? Science 292:1315–6.

Torrisi M, Pollastri G, Le Q. 2020. Deep learning methods in protein structure prediction. Comput Struct Biotechnol J 18:1301–10.

Tsang Y, Infante DM, Wang L, Krueger D, Wieferich D. 2021. Conserving stream fishes with changing climate: assessing fish responses to changes in habitat over a large region. Sci Total Environ 755:142503.

Tsang YP, Tingley RW, Hsiao J, Infante DM. 2019. Identifying high value areas for conservation: accounting for connections among terrestrial, freshwater, and marine habitats in a tropical island system. J Nat Conserv 50:125711.

Tytell ED, Carr JA, Danos N, Wagenbach C, Sullivan CM, Kiemel T, Cowan NJ, Ankarali MM. 2018. Body stiffness and damping depend sensitively on the timing of muscle activation in lampreys. Integr Comp Biol 58:860–73.

van Gestel J, Tarnita CE. 2017. On the origin of biological construction, with a focus on multicellularity. Proc Natl Acad Sci USA 114:11018–26.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr, Kinzler KW. 2013. Cancer genome landscapes. Science 339:1546–58.

Voit EO. 2019. Perspective: dimensions of the scientific method. PLoS Comput Biol 15:e1007279.

Walpole J, Papin JA, Peirce SM. 2013. Multiscale computational models of complex biological systems. Annu Rev Biomed Eng 15:137–54.

Welch JD, Hartemink AJ, Prins JF. 2017. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol 18:19.

Wolf YI, Katsnelson MI, Koonin EV. 2018. Physical foundations of biological complexity. Proc Natl Acad Sci U S A 115:E8678–87.

Wooley JC, Lin HS. 2006. Catalyzing inquiry at the interface of computing and biology Washington, DC: National Research Council.

Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO, Walker GC, et al. 2019. A white-box machine learning approach for revealing antibiotic mechanisms of action. Cell 177:1649–61.

Yu MK, Ma J, Fisher J, Kreisberg JF, Raphael BJ, Ideker T. 2018. Visible machine learning for miomedicine. Cell 173:1562–5.

Zampieri G, Vijayakumar S, Yaneske E, Angione C. 2019. Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput Biol 15:e1007084.