The Axes of Life: A roadmap for understanding dynamic multiscale systems

Sriram Chandrasekaran, University of Michigan csriram@umich.edu
Nicole Danos, University of San Diego, ndanos@sandiego.edu
Uduak George, San Diego State University; ugeorge@sdsu.edu
Jin Ping Han, IBM TJ Watson Research Center, hanjp@us.ibm.com
Gerald Quon, UC Davis, gquon@ucdavis.edu
Rolf Müller, Virginia Tech, rolf.mueller@vt.edu
Yin-Phan Tsang, University of Hawaii at Manoa; tsangy@hawaii.edu
Charles Wolgemuth, University of Arizona; wolg@email.arizona.edu
(Authors ordered alphabetically)

Summary:

The biological challenges facing humanity are dynamic and multiscale, and are intimately tied to the future of our health, welfare, and stewardship of the Earth. Tackling problems in diverse areas including agriculture, ecology and health care require linking vast data sets that encompass numerous spatial and temporal scales. Here, we provide 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 spatiotemporal scales, factors and biological components.* To understand life, we must be able to integrate complex dynamic systems across diverse scales (Spatial, Temporal, Physical, Biological, and Evolutionary). In addition to multiple scales, extrinsic and intrinsic factors such as perturbations 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: biological components, internal and external factors that act on the system, and the scale of the system is necessary and sufficient to predict a system's behavior.

The traditional research paradigm focuses on fixing two axes and varying the third. For example, studying a bacterium exposed to external perturbation fixes both the scale and system, and modulates the factors. Yet most biological phenomena are multiscale exhibiting various degrees of emergence, self-organization, robustness, resilience and complexity. A classic example of a dynamic multiscale challenge is health care. Most diseases involve the dysfunction of biological processes at multiple scales from genes to proteins or organ systems and usually translates into decrease of complexity. The actions of those altered processes change the behavior of cells, which then lead to systemic effects within the body. Other examples that span multiple dimensions include predicting phenotype of an organism or community from genotypes, or predicting how global temperature change affects organismal behavior and ecosystem balance.

Another reason why axes are typically fixed in an experiment is one of practicality. 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, and the number of possible interactions explodes as multiple scales are considered (combinatorial complexity). While fixing axes improves the tractability of studying a system, it also limits the linking of data across scales. Theoretical and empirical frameworks are needed for looking across scales and making educated hypotheses about which connections across scales might be most fruitful to experimentally explore.

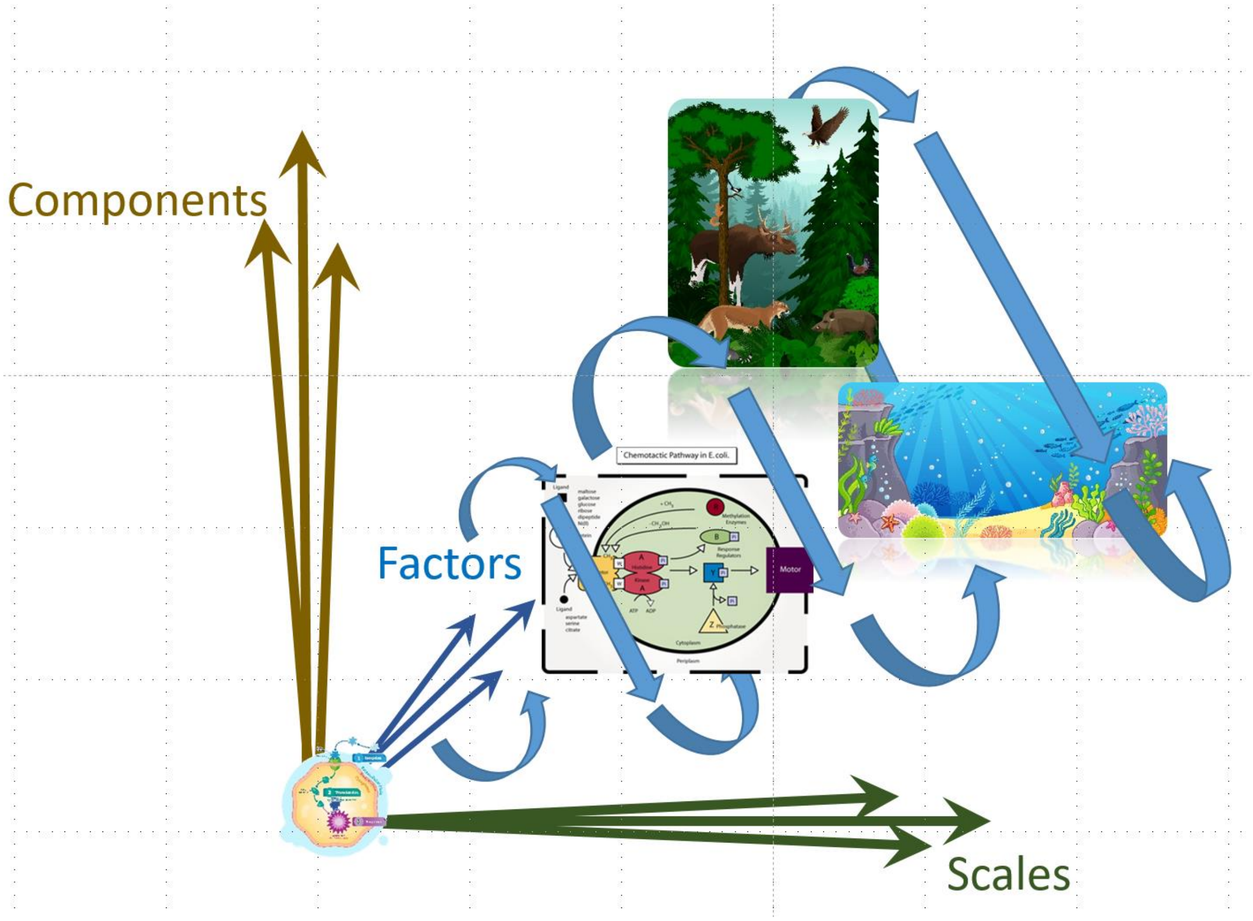

Figure 1. The axes of life. Biological systems span three independent axes spanning various scales, components and interactions, and influenced by intrinsic and external factors. Factors and interactions may operate at various scales. Characterizing a system based on these three axes is necessary to predict its behavior. (Diagram of the chemotactic pathway in E. coli modified from Falke et al In Annu Rev Cell Dev Biol. 13 (1997): 457-512.)

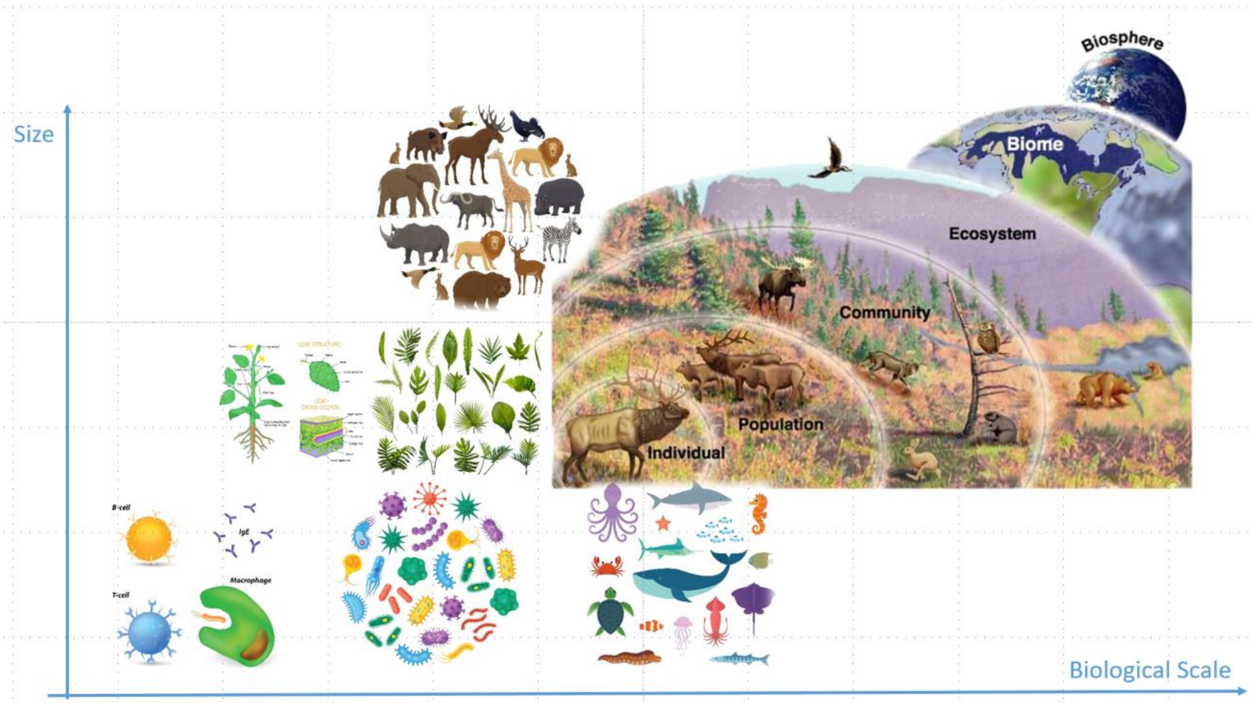

Figure 2. The scales of life. Multi-cellular organisms coordinate a diverse array of components into complex organizations that work together as a single living entity across its lifespan. These organisms then group together into larger, interdependent structures such as communities and ecosystems. [Six ecological levels]. Retrieved November 30, 2011, from http://buffonescience9.wikispaces.com/UNIT+6+-+Ecology+

How do we foster and enable new research that will effectively bridge across scales and biological components? Iterative dialog of experiment 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 multiscale systems. We will focus on four key impediments limiting our understanding of dynamic multiscale systems. This ultimately requires Iterative interactions between diverse disciplines and between Data, Methods and Theory.

- **Multidimensional Heterogeneous / Multimodal Data** -- Generating, curating and disseminating relevant and high quality heterogeneous / multimodal data across multiple disciplines, scales and components
- **Bridging Methods** -- Developing and applying methods that integrate this heterogeneous data to drive research in biological systems
- **Theoretical Frameworks for Integration** -- developing theoretical framework that links data and methods to research hypotheses
- **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

Data Generation and Management

An integrative approach to quality data generation and management has the potential to provide bridges between disciplines, breaking through structural and theoretical bottlenecks. We have identified several of these bottlenecks, including choosing data that is 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. These topics are described in more detail in a companion paper (#21).

Briefly, current data generation practices provide limited representation of all three axes. We recommend funding for studies that span all three axes - for example, studying the impact of both global temperature change and local release of a toxin on microbial metabolism, physiology and ecosystem biodiversity over a decade involves various scales (temporal, physical), factors (temperature, toxins) and components (molecules, microbes, plants).
This goes against the traditional view of fixing various factors and components; while this traditional approach has been fruitful, it nevertheless limits the creation of theories that span the axes of life.

Theoretical Frameworks for Synthesis

Before dynamic multiscale systems are modeled, a feasibility study should be conducted to ensure the system can be causally inferred and modeled (e.g. is predictable). That is, the degree of predictability of a system should be a principal criterion for prioritization. Equally important is also the identification of the causal factors or variables that can continuously or transiently (sporadically) influence the dynamics of biological systems. If a system is not predictable (given a pre-defined set of measurements), then it may not be worth studying/modeling (until we identify the input data and its critical (most relevant) variables needed to make it predictable). For example, predicting a phenotype (e.g. behavior) from genotype may not be feasible if the behavior is strongly driven by noise or environment.

Interpretability of the framework is not necessary here; 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 AI can be applied to identify the patterns and build mechanistic models.

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).

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, based on whether measurements on different scales can be made on the same entities, or when only 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. Otherwise, strategies from manifold alignment can be used to construct models of biological phenomena at individual scales, and then alignment performed to identify connections between scales.

Alternately, theoretical frameworks can be derived from first principles of evolution, chemistry, mathematics, computer science or physics. For example, the production of a metabolite may not be possible from mass-balance and thermodynamic principles. Similarly, 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, that then help focus efforts on finding approximate solutions to difficult problems.

Mathematical models have enormous potential to unravel the complex interactions of biological phenomenon occurring at different scales. They can be applied at different scales and also at system levels of organization. However, there are limited tools that allow the coupling of models defined at different spatial scales. Moreover, the modeling and analysis of interdependent biological systems also requires mathematical models capable to identify the causal influences and to capture either the Markovian or non-Markovian dynamics of some biological constituents. The implementation of new techniques that link multiple scales to study system level outcomes would be invaluable in understanding the complex interactions of biological systems.

In general, current theoretical methods lack the ability to transition between scales as we lack an underlying ‘objective’ for models. For example, do all living systems maximize their biomass production or energy efficiency or 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 poses many unique challenges. There are several fundamental concerns that must be addressed that are often taken for granted in traditional systems. For example, it is often difficult to define the objective, since the data sets 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 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 scale-level differences may occur in multiple aspects of the system under study, as exemplified by the conceptual Axes of Life outlined above involving scales, factors and biological systems.

As we build bridges between scales, 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 NSF jumpstart, 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 the cellular scale. The challenge then centers on the question:

How do we link scales? 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. In contrast, in empirical models that link mutations at the molecular scale to a physiological phenotype (e.g. breast cancer) the links are ‘imposed’ by the scientist.

The potential of applying AI to these challenges promoted a vigorous discussion at the JumpStart. 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 (#22). Briefly, some strategies to make AI transparent include linking traditional models based on biological or mathematical principles with deep learning models. Alternately, the structure of the AI model can be influenced by prior knowledge of the biological system.

All of the typical challenges with managing data are multiplied in integrative approaches, and flexible methods to deal with these challenges will be necessary. For example, how do we deal with noise? Noise operates at multiple scales (temporal, biological, etc), 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.

As we move toward bridging across scales, a critical first step involves selection of appropriate, tractable systems. Multiscale processes are complex (in the sense that their rate of change may not only exhibit various degrees of nonlinearities, but also a nontrivial 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 scales. 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 affects behaviors, such as motility. Along the same lines, multiscale neuronal analysis has revealed that brain regions exhibit long-range memory / non-Markovian and multifractal characteristics.

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, 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 questions.

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 (NLP) 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.

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 hadn't 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 contains a deep truth. 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.

Acknowledgements

We than David Goodsell and other participants in the NSF JumpStart for their insightful comments and feedback on this article.