Biological Data in the 21st Century: The Challenge of Integration

Nicole Danos, University of San Diego, ndanos@sandiego.edu
David S. Goodsell *, the Scripps Research Institute and Rutgers University
goodsell@scripps.edu
Uduak Z. George, San Diego State University, ugeorge@sdsu.edu
Jin Ping Han, IBM TJ Watson Research Center, hanjp@us.ibm.com
Rolf Müller, Virginia Tech, rolf.mueller@vt.edu
Joan Segura, RCSB PDB UC San Diego, joan.segura@rcsb.org

*corresponding author

The Goal

Our ultimate goal is to integrate mathematical modeling and experimental approaches to better understand the Rules of Life at varying scales (temporal, spatial or biological organization). Although such integration exists within individual fields, e.g. cellular biology or organismal biomechanics, models spanning these fields are still limited. Barriers we identified include a disintegrated and incomplete data pool, which may be addressed through promotion of effort to develop common ontologies and data standards, continued exploratory studies to grow the knowledgebase of biology, and nimble methods for development of FAIR databases in nascent fields.

This paper was created as part of a group exploring larger issues of the integration of experimental and computational approaches in paper #4.

Potential

An integrative approach to data management has the potential to provide bridges between disciplines, breaking through bottlenecks that are currently inhibiting understanding. The time is right for this, because data acquisition through automation, among other techniques, has been maturing in many areas, and advanced technologies have enabled data centers to grow exponentially in many fields. Advanced methods are currently available for data collection, deposition, and curation in databases. Such data is incorporated into iterative models that link experiment with theory and modeling, but these methods are largely confined within individual disciplines. For example, the Protein Data Bank (PDB) currently plays a central role in the structural biology community, by providing a freely available repository of structures, fully curated and annotated [1, 2] (www.pdb.org). In this role, the archive has led to new discoveries in the basic mechanisms of biomolecular structure and function, and provided the basis for an entire field of biomolecular design and bionanotechnology. More recently, there has been a concerted effort to link the data archived at the PDB with appropriate data held in genome and protein sequence databases, as well as functional ontologies such as the Gene Ontology (GO) knowledgebase [3]. This has greatly expanded the scope of use of the PDB database, making the results more accessible and usable by users who are not experts in structural biology. Similar curation and linkage of isolated biological databases with other resources will enable the development of novel data analysis pipelines.

Data Challenges in Integrative Biology:
Traditional methods for gathering and interacting with data pose several challenges in integrative efforts. Foremost, we need to choose data that is relevant to our problem. This problem impacts study at many levels. At the simplest level, we need to choose a model system. We often have multiple types of data, which inform understanding in different (occasionally inconsistent) ways. The challenge for integrative model building then becomes to find consistent frameworks for incorporation and synthesis. Integrative problems are often dealing with systems that have wide ranges of spatial and/or temporal scale, so this challenge of data choice often involves critical decisions about whether or not the data has enough mutual dependency to be relevant in the integrated system under study.

Biological data is messy and, in many cases, not well curated, so it is imperative to have methods to assess the quality of data and incorporate these quality measures into the resultant integrative syntheses. This central need permeates the entire process of integrative science, requiring attention to statistical significance at the data collection stage, effective annotation at deposition, incorporation of quality measures during model building, and encoding of significant measures during analysis and dissemination of models.

We must also overcome the current barriers for making data accessible across disciplines. The concept of FAIR (findable, accessible, interoperable and reusable) principles has been an essential first step towards this goal. For example, in the structural biology community, pressure from funding agencies and from professional publishers has ensured that all publicly-funded atomic structures are deposited in a timely manner, but the same may not be true in other fields. Less mature databases may also not have the resources to create portals that are easily accessible and usable by non-expert users. In addition, important connections between disciplines are often hidden in manually-annotated metadata, and thus are not available to automated metastudies that seek to integrate information across heterogeneous datasets.

Exploratory studies play an essential role in discovery. We have not yet entered an era of scientific study where integration of existing data is the only challenge. Currently, the creation of multiscale mathematical models and their validation will require some data that does not yet exist. These gaps are particularly large at the intermediate scales that bridge more traditional fields of study, and for which effective experimental techniques are still being developed. For example, data on the behavior of complex tissues is lacking, even though the components of those tissues and higher-order physiology are both well characterized. Additionally, we are far from having described the existing biological diversity of Earth. In a similar way, current protein structural data covers most essential parts of the human proteome, however, this information scales poorly when attempting structural characterizations at the cellular scale: only computational models are available for the remainder of the proteome and how they are integrated into a functional cell. Looking at the history of science, paradigms are continually broken as scientists find new ways to look at nature. An increasingly complete view of the world will allow us to assess the universality of any Rules of Life that we discover.

Opportunities

Manual annotations, often included in metadata, play an oversized and woefully ill-defined role in integrative projects. A theme that came up over and over during this JumpStart is the central need for common ontologies. This is the only way to promote an effective dialog between disparate disciplines and develop standards for data exchange across different scales. In addition, annotation of data, in particular annotations that forge connections with other disciplines, are often altruistic, so integrative efforts would strongly benefit through specific promotion of this activity. We need to promote a culture that includes training in basic principles of annotation as part of the education process and includes a comprehensive annotation effort as part of required data management plans. Exchange of annotations necessitates a community conversation about the language used in their description, since many different fields use slightly different terminology for similar or related phenomena. Such an exercise would be inherently integrative because it will force a conversation about who can and who might want to use these resources, expanding the field of users and collaborators.

Data management plans as required by funding agencies are currently insufficient because they typically achieve the bare minimum of storing, preserving and potentially sharing data with other researchers, and do not promote the richer annotation and integration that is needed to promote interdisciplinary work. Unfortunately, the path to this goal is not clear, since, given the nascent nature of integrative studies, we are still developing the basic guiding principles needed to make data accessible and usable across scales and also across disciplines. This is particularly problematic in the context of integrating mathematical modelling with biological data. Many mathematical modelers and experimentalists working on similar topics are eager to consolidate their ideas and share data. Mathematicians who build mechanistic models need experimental data to test and validate their models. Experimentalists are interested in identifying available mathematical and computational tools to analyze their data. There are repositories that host mathematical models of biological and biomedical systems that have been published in the scientific literature, an example of such a repository is BioModels [4, 5] (https://www.ebi.ac.uk/biomodels). Currently available repositories allow mathematicians to share mathematical models and expand existing studies. However, an experimentalist with no mathematical background might find it difficult to utilize these freely available mathematical models. New data management plans are needed that will require users to think about the challenges posed here. In addition, a concerted effort is needed to develop additional repositories to help biologists to use these freely available mathematical models in their research. These can be achieved by making graphical user interface that allows users with limited programming knowledge to access and use existing mathematical models, and innovative search algorithms that do not rely on exact terminology since this is likely to vary between mathematicians and biologists.

In addition, we may need to rethink our approach to public databases. Mature databases (UniProt, PDB, PubMed, etc) are essential for central resources, but we need nimble methods to develop and deploy accessible databases for nascent fields. This creates the potential for the design of easily modified databases with enhanced search algorithms that return results at several levels of relevance, accounting for differences in terminology across fields and

incomplete metadata annotations. For example, databases at larger biological levels of organization, such as access portals for natural history collections or anatomical data such as MorphoSource, are much younger and thus would benefit from the expertise created in the development of other more mature databases. Therefore, a concerted investment in such an effort would move the entire individual field forward, making its data accessible not only to vertebrate biologists but also researchers from other fields looking to integrate their findings.

Broader Impacts

An accessible data pool will promote broader participation in biology by reducing the expertise required to utilize existing datasets. As a result, researchers from other STEM fields, academic or industry, will be able to apply their expertise in pursuit of the universal Rules of Life or the development of further methodologies. Additionally, it will lower the threshold for scientists at all levels of their career to engage in this research regardless of their financial resources. It will also eliminate unnecessary repetition of similar experiments by different laboratories, thereby enhancing better use of resources, time and efforts.

The increased size of datasets and the complexity of metadata associated with it requires the involvement of computer scientists in the creation and maintenance of such resources. These computer scientists will work on creating user-centered databases but are informed by the biological goal of identifying the Rules of Life across scales. A well curated data pool will enhance data mining across different spatial and time scales and aid in the identification of patterns defining the Rules of Life.

References Cited

1. Berman, H. et al. (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10 (12), 980.
2. wwPDB consortium (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res 47 (D1), D520-D528.
3. Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (1), 25-9.
4. Chelliah, V. et al. (2015) BioModels: ten-year anniversary. Nucleic Acids Res 43 (Database issue), D542-8.
5. Glont, M. et al. (2018) BioModels: expanding horizons to include more modelling approaches and formats. Nucleic Acids Res 46 (D1), D1248-D1253.