**Large-Scale Collaborative Research in Biological Imaging**

(Manfred Auer, Life Sciences Division, Lawrence Berkeley National Laboratory)

**Abstract:**

21st Century biologists are producing an enormous amount of data that is not utilized to its full capacity due to the lack of appropriate tools that facilitate its interactive, collaborative exploration and analysis. Collaborative visualization and analysis platforms will enable a holistic approach to understanding the biological puzzles for they will allow biologists with different expertise to virtually sit together and assemble the puzzle. Such advanced technologies will allow scientists to study the complexity of biology systems in adequate detail, resulting in reality-based modeling and simulations. These simulations may propel biology into an era of quantitative biology and turn biology into an engineering discipline with defined systems design specifications.

**Biological imaging at the crossroads:**

For the most part of the 20th century biology had a reductionist approach. Complex biological processes could not be studied in their entirety and within their cellular context and, consequently, they were subdivided into small enough problems that were then studied on their own often at minuscule detail. This approach made sense then due to the lack of sophistication of the technologies that were available. Thus, a day in a biologist life focused on a particular protein, its enzyme kinetics, possibly its interaction target, its shape, size, structure and composition. However, even the most comprehensive knowledge about a particular regulatory protein could not answer the question of how its malfunction leads to such devastating diseases as cancer, or complex neurological and psychiatric disorders.

A first step towards higher complexity was taken in the early 1990s when it was realized that proteins rarely function in isolation, but that they are organized into so-called macromolecular machines. These machines can be thought of as an assembly line where certain tools (the enzymes) catalyze a particular chemical reaction, e.g. the addition or removal of a phosphate, which allows a protein to be recognized by a new set of proteins. This realization shifted the focus from individual protein function to the macromolecular complex function and resulted in increased complexity of systems that were studied. However, even the study of one particular assembly line couldn’t explain how the entire factory of a cell worked, as many molecular machines need to work together to sustain life.

The 21st century has brought a new challenge, which is related to the vast and heterogeneous data being created. This challenge is rooted in two main events. One is the advent of genomics and the determination of the human genome, which started a new era in biology based on high-throughput measurement. Genomics, Proteomics, Structural Genomics and now Metabolomics have added enormous information to the biological puzzle, but most of it is buried in databases that are hard to access and not particularly user-friendly, thus somewhat meaningless to a regular biologist that lacks bioinformatics expertise. The field of Systems Biology was created with the goal of collecting as much information as possible and try to make sense of it. This has yielded spectacular insight into metabolic pathways and allowed the rejuvenation of another field: metabolic engineering, also known as Synthetic Biology. Biologists begin to interact with bioinformatitians and computer scientists and these collaborations have resulted in spectacular outcomes in pharmacology and biofuel production.

The other big event that is revolutionizing biology started in the 20th century with the development of biological imaging. Many complementary optical techniques have been developed such as Raman Microscopy (in its various flavors) and Fourier Transform InfraRed (FTIR) Microscopy, which allow the identification of a whole set of macromolecules including proteins, carbohydrate, DNA and lipids. Even more recently mass spectrometry imaging allows for spatial detection of biological molecules. Last but not least, electron microscopy has witnessed a renaissance with the advent of sophisticated sample preparation protocols allowing faithful preservation. Modern EM imaging techniques now allow the study of the spatial organization of macromolecular complexes inside entire cells and in tissues, allowing to place respective macromolecular players into the subcellular, organelle, cellular and tissue context. Such data sets can now easily reach tens of Gigabytes and in the very near future with the advent of 32k by 32k by 10k pixels/voxels may reach Terabytes per data set.

Arguably, it is the knowledge of the parts list (obtained by OMICS) and its spatiotemporal organization (through the various forms of microscopy) that will allow biologists to determine function at a holistic level. However, none of the microscopy techniques alone will be enough as they work at different length, resolution and complexity scale. Hence, what is needed is to combine all these techniques and link the spatiotemporal information in a smart way with database knowledge. We strongly believe that this integration needs to be centered around the 3D images rather than around the Omics databases. After all we learn about the ins and outs of a city not from reading through its phone book but from walking through the streets and making intuitive inferences from spatiotemporal patterns found while exploring the 3D world.

Today, advances in 3D imaging techniques allow scientists to easily collect tens of gigabytes of 3D volume data. However, the inspection of the resulting volumes remains a tedious and time-consuming process that requires the slicing and manual segmentation of the volumes as well as the integration of experimental data from many sources. Typically, different pieces of data are collected by different research groups focused on particular aspects of a common problem. Therefore, principal investigators and their teams need to travel to where the data is analyzed in order to provide their input for comprehension and interpretation of the data. *Imaging biologists spend months to analyze just a subset of these 3D volumes. Besides the complexity of the data itself, there is a lot of planning involved to get one or more collaborators to travel to the lab to analyze the data. Usually these scientists cannot spend more than a couple of days visiting the lab, which is not enough to analyze all the data. Even though the data sets are rich in content, lots of this information is wasted due to the lack of appropriate technologies.*

With increased complexity and hence the need of prolonged analysis, such scenarios will become impossible.

**A vision for the next decade**

We expect an explosion of 3D volumes over the next decade as advances in hardware and software will allow the automated data collection and 3D reconstruction of several dozens of 3D datasets per day. Such datasets will contain enormous amounts of information. Interpreting these high-content datasets will require the expertise of a large number of biologists that will be experts in different aspects or components of the datasets. Hence such datasets will need to be analyzed, ideally simultaneously by various investigators from all over the world that will collaborate to decipher all the information contained in the large 3D volumes with multiplex information. In order to realize the potential of true collaborative analysis on complex multidimensional datasets, the 3D data will need to be stored in a common repository to make it available to the scientific community, as is common for Omics type of data. As the 3D imaging data becomes centralized, an infrastructure must be developed to enable the remote analysis and interactive visualization of the data as well as the interactive collaboration of groups of scientists located at different sites. Such repository should provide three levels of access to the data. The first level represents the raw data, the second level comprises the results of data analysis and annotations, and the third level consists of integrating different data sources such that data from multiple datasets can be simultaneously queried.

A day in the life of a biologist will involve teams of experts that are located in different parts of the world. Luckily, the data is centralized and they have a platform for interactive visualization of very large datasets, creation of simplified models on the fly, annotation of the models, and integration of diverse data.

This platform will include the following features:

* Supports a collaborative multi-display setting environment. The exploration of large volumes needs to be intuitive, real-time, and allow annotation, feature extraction and interpretation on the fly.
* Supports immersive visualization of large, complex 3D volumes with multiplex information display and video-game-like fly-through capabilities.
* Supports a sophisticated user interface that enables automatic feature extraction based on user specifications.
* Supports data simplification from voxel-based space to adequate geometrical objects, from voxels to semantic objects.
* Handles the different data formats and types of information.
* Permits to overlay these data sets onto one another, through data set registration and multi-characteristics data representation.

In the next 10-20 years, models of reality will replace the density-based multidimensional maps, and will serve as reality-based constraints for actual modeling and simulation of biological processes. Such simulations based on real data rather than first principle assumptions will help to catapult biology into the world of engineering, where precise shape/structures will be replaced by design specifications that are compatible with the respective biological function and thus take into account the design robustness that is impressively found in biology.

Clearly this is at least a 10-year vision, possibly a 20-year vision, but it is one where every single small piece can be accomplished if adequately supported.