The New Biology

David C. Schwartz, Ph.D.
University of Wisconsin

I was contacted by Jim Kadonaga and asked to take a look into the future and describe what I saw at that edge—for which I call, "The New Biology." I believe right now that biology is in the midst of a major paradigm shift, and I think a lot of it is due to advances in information technology and the new systems that provide huge experimental data sets. Information technology is going to enhance the way that we do science in many ways, giving opportunities and challenges to a broad range of scientists, engineers, and mathematicians.

The New Biology has the three following components. It is about one third statistics, mathematics, and computer science. Another third is more in the physical sciences—physics, chemistry, and engineering. Finally, the balance is in the biological sciences and genetics. Different people will be attracted to different components and a major challenge at the university level is how to train young researchers to become the New Biologists.

The New Biologist must be able to commingle these different disciplines to address important biological problems. Teamwork is necessary since the systems that you are dealing with are very complex, and beyond the ability of an individual, or even a small laboratory to fully encompass. Some of these systems may involve complex instrumentation, but the level of complexity goes beyond instrumentation.

The fundamental nature of the systems we scientists deal with has entirely changed, and when I speak of systems, I do not mean instrumentation. I am going to talk about the very stuff that you put your fingers on, the very data that you look at, and especially the very phenomena you

use to interrogate your biological systems. Let me give you some examples of advances in the New Biology.

- We are all comfortable thinking about genes. Now, we are accustomed to thinking about many genes, and lately we think about whole genomes—not only one genome, but many genomes that span populations and species. Look at the databases, and you will see hundreds of microbial genome sequences—and this is just the beginning.
- We are used to dealing with many cells. Now we can deal with single cells, in a new way, with the intent of building capacious ensembles from complex measurements obtained from a large number of individually interrogated cells. The goal is to be able to perform sophisticated experiments within a single cell and to acquire data sets that encompass large cell populations, which simultaneously require and potentiate statistical interpretation. In other words, new approaches allow us to massively "test tube-ize" cells.
- Being able to do science with single molecules has become a very popular coin of the realm. It is not difficult, and it gives you a great deal of power in terms of forming your ensembles and represents the ultimate in the ongoing quest for ever-increasing levels of miniaturization.
- We are going from instruments that used to occupy a lot of lab space to chip-based instruments that can create labs on a chip. Here, we have scaled down large instruments to reside on a single chip. In this regard, what I urge people to do is if you are going to miniaturize, revel in the scale of matter that you are working in. Take advantage of the novel phenomena that this scale gives you. It is proven to be a very interesting problem in its own right in terms of great nanotechnology and physics.
- Achieving sub-Dalton resolution, means utilizing mass spectroscopy—it is just amazing what you can do with mass spectroscopy. When I was in graduate school in chemistry, I had friends working in mass spectrometry. One day, after too many beers, someone wondered what would happen if you put a protein in a mass spectrometer. Everyone had a good laugh. Nobody is laughing anymore.
- Now the best label is no label. So we are going from labeling substrates to not labeling, yet maintaining the ability to detect with some specificity; for example surface plasmon resonance approaches offer this advantage—my Wisconsin colleagues boast about the fact that they could do chips and other types of assays with no labels, thus enabling biomolecules to function in a more native environment.
- When I was in graduate school we were able to create a limited number of compounds and characterize them. Now, combinatorial chemical libraries hold tens of thousands of characterized compounds that can be further "functionalized" through series of biological assays.

I think the modes of inquiry have changed. We are used to hearing about discovery-based science using large-scale screens. This approach leverages chance resulting from the ability to cast a large experimental net. Earlier, we heard about the fantastic work that Regeneron is doing in this area from George Yancopoulos. Discovery is the ability to do large-scale screens that produce useful information. I dare say that we are not as smart as we think we are, so being able to toss the dice in a very systematic and controlled way helps us solve very tough scientific problems.

We are trained as scientists to do hypothesis-driven research. But we can go beyond this when we blend discovery with hypothesis. Chance favors the prepared mind—I think that is a quote from Louis Pasteur, who was an accomplished chemist as well as a pioneering microbiologist. In the New Biology, the ultimate experimental space is a single molecule, or a single cell, or a single-molecule system rapidly analyzed at high resolution over a large population. At the end of the day you want to nimbly create large multidimensional databases that directly help you solve problems that are biologically relevant. With all of this data, you are not going to analyze this on a spreadsheet, so the need for sophisticated statistical and computational approaches becomes acute.

Given large data sets has encouraged many of us to become very friendly with our local computer scientists, statisticians, and even mathematicians. And while it is interesting explaining our systems to these people, it is becoming a thing of the past. Are we faced with a problem of too much information? No, I do not think so. Instead, I think that we are suffering from too little information. We have enough information to tell us what we do not know. We have enough information to tease us; but we do not have enough information to close many stories. I think that we need to find new ways to augment and balance large-scale experimental design and analysis. Given the new experimental approaches and databases that we have, are we generating the type of experiments that make sense of these data?

When I say "sense of these data" I am describing chip data, for one example. Researchers want to make sense of expression profiles. In many cases, the interesting features in a complex dataset are going to be buried deep within the data and obscured by noise, and, consequently, may not be represented at all. Subsequent activities center on slow "conventional" ad hoc experimentation and analysis to confirm "results" and then the drilling down into a select group of candidates to do proper scientific investigation with the intent of making a story. Such activities usually require several graduate/postdoc years per candidate. So the idea is to design new experimental systems that intrinsically work with operations in cyberspace—you want systems that are designed to rapidly close

stories on a large number of candidates. This all begs the question, "can massive candidate identification be adequately addressed experimentally?"

Let's talk about theory and experiment. As a biologist you have ideas and you have hypotheses. But just having an idea and a hypothesis is never sufficient; you need to do experiments. So you utilize experimental spaces available to the New Biologist:

- sequence from an ABI-3700,
- large databases with other scientists' data,
- chip data,
- mass spectroscopy,
- and, if you have a lot of money, utilization of a sequencing center.

What happens is that you wind up in a loop. You come up with an idea, and may go to the databases; ultimately you sit in front of your computer. You might do some simulations, and then you come up with some candidates. And you might stay within this space a bit to refine what candidates you have. And you had better, because when you get down to the experimental end, things start slowing down a whole lot.

When I do one experiment, it seems to beget a lot more experiments. I tend to root into a problem; I get new ideas; I interact with the experimental matter: and I want to do more experiments. This process can get very slow. One solution is to have large, automated systems, enabling you to come up with even more experimental candidates that will keep a thousand postdocs busy. The loop starts out very quickly and then slows down. What is the problem; how do we expedite this loop that covers hypothesis generation and validation through experiments? We need to dramatically accelerate our ability to create and analyze complex experiments within the context of hypothesis generation.

I am a big fan of large, complex datasets. I have been using Bayesian inference techniques to analyze data. Not being a computer scientist or statistician, I am absolutely in awe of this technique. You define your system in terms of models, that describe your experimental variables, and errors—add to this a large experimental data set and then exhaustive analysis enables you to determine the best fit between your data and the "best" answer. The necessary component of this process is the availability of a large and sometimes complex dataset; obviously this approach will not solve all problems.

Let's take a look at how physicists deal with theory and experiment. Particle physicists work at very high energies, and their "city-sized" experiments produce incredibly large data sets that establish their functional loop between hypothesis and empirical data. At the European Or-

ganization for Nuclear Research (CERN) there is a cyclotron ring that is about 27 kilometers across. One experimental system at CERN handles as much information as the entire European telecommunications network does today—about 800 million-proton collisions a second, and they are able to cull from such data very rare events. They detect something of interest in one out of every hundred million, million collisions. If we were to equate this to something more biological, and this is not a perfect analogy, it means that the human genome dataset could be decoded in seconds, and lately we can envision systems that might be able to do this. Evaluating screens for point mutations means finding one mutation in a genome that is 30,000 times larger than the human genome; however, beautifully designed genetic screens can be designed for specific criteria, which similarly scale. Overall, these vignettes and comparisons provide some context on how far biological research needs to advance in terms of the generation and analysis of large datasets. In this regard, I think our problems will prove to be more difficult to solve.

We are talking then about the hard problem of generating large biologically relevant datasets and at the same time making scientific sense out of the results that come from these efforts. Up to now, we have largely relied on automation and multiplexing to do the experimental heavy lifting, and at times these activities are walled off from our hypothesis and simulation efforts. This makes moving through the loop slow and toilsome. What I think we need to do is to dig very deeply into basic physical and chemical phenomena to come up with new ways of interrogating biological matter in ways that productively impedance match what we are able to do in terms of IT structures and simulations. The idea is to engineer these components to work as a harmonious whole. We need to think about ways of making complex, multidimensional experiments as simple to perform as working a spreadsheet—and this is a very tough problem.

I was talking with Jeff Duyk and it turns out he has an article coming out in *Nature* where he talks about CAD/CAM—Computer-Aided Design/Computer-Aided Manufacturing. This is how CAD/CAM works. An engineer sits down at a computer terminal and in cyberspace designs an object. He defines the components of the design and analytical tools within the software to enable a suite of structural analysis tools to consider his design, or his "hypothesis." Next, the finished file is transferred to a computer controlled system that automatically creates the object, or performs the complex "experiment" for you. Basically you do not conduct the "experiment," but can directly interact with the results.

For example, a few years ago I found out about a device that rasters a laser beam, sort of like a laser printer, but instead of putting black toner onto a piece of paper it rasters over a pool photo-reactive polymer. The

process repeats the rastering process on top of successive layers of uncured liquid polymer. Eventually a complex object appears after a number of layers have been completed, producing something that you can hold and play with. It was as if you had machined the object, but you did not have to be a machinist. All you had to do is sit at the computer and create that object that you could put in your pocket. I find this truly amazing.

So why don't we use these approaches to develop biological CAD/CAM? Schematically, we can imagine the following: a scientist interactively runs through hypotheses by sitting at a computer system employing databases and simulation tools. The next step would be for the system to guide the scientist through complex experimental designs for which the intelligent system would "assemble" and run to validate hypotheses. The experiment "assembler" is a direct interface for cyberspace to directly conceive and control experimental activities.

Let me sketch out this concept of an experiment assembler in some detail. In the 1400s, printing in Europe was done though woodblocks, a time-consuming process. In order to meet the demand for the printed page, Gutenberg developed the letterpress and the concept of movable type, avoiding the need to make a whole new plate in order to print something like a newspaper or a book. What we need in terms of our experiments is something similar to movable type, with multiple copies of very large-scale experimental motifs that you can physically move around and bring into juxtaposition. This is actually a major problem in terms of sample handling, which, when solved, would obviate significant robotic intervention. In essence, you want to develop complex experimental systems that intrinsically self-assemble, or are very simple to manipulate in terms of bringing disparate experimental motifs, or components, together in a controllable way by the experiment assembler. (At the University of Wisconsin, we are developing components for the experimental assembler, in terms of systems that can either self-assemble or are simple to manipulate in terms of bringing massive numbers of experimental components together in a logical way.) Again, in terms of the scheme I am describing here, you would start with experimental motifs consisting of cells, peptides, nucleic acid, chemical libraries and so on. The assembler puts your experiment together, and is run creating a dense, multidimensional flux of data—this will certainly require radically new detection schemes. The data streams back into your system to validate aspects of your hypotheses, and to develop further experiments to refine this fit, or to spur new hypotheses. If we can do this, we can create many loops, and greatly accelerate the number and complexity of candidates that are truly developed in a complete biological way.

The origin of this talk stems from the fact that I am very jealous of my

colleagues in the industry. I am very jealous of the fact that they have the means to pursue very large screens and to create very large datasets. I want to be able to create this in my laboratory. Industry has done a great job in development and discovery, and academia has done a great job in hypotheses development. Both sides are crossing over into the domains of the other. The activities of the Markey Scholars are a testimonial to this process.

I believe, however, that industry will always have superior resources and organization to pursue these problems and the reasons are all obviously based on economic considerations. It is a lot different writing a RO1 and getting funded, as opposed to going after a drug that can yield perhaps a billion dollars in profit.