From Data to Knowledge to Action: Enabling an Initiative in "New Biology"

Chase Hensel
Computing Research Association

Erwin P. Gianchandani
Computing Research Association

Computing Community Consortium
Version 9: August 1, 2010[1]

In 2009, the National Academy of Science (NAS) published a report titled "A New Biology for the 21st Century"[2]. This report outlined the vital role that "New Biology" – an emerging interdisciplinary field incorporating computer science, mathematics, and engineering along with biology – will play in the future. The NAS report presents four areas of significant importance to the wellbeing of our nation – food, environment, energy, and health – and describes how New Biology can catalyze change in each of them, ultimately improving the daily lives of all Americans. **In each area, new data analytics approaches – i.e., new ways of data integration, analysis, and modeling – will be critical.** Here we detail the importance of data analytics to the New Biology movement, and we provide specific recommendations for related Federal investment in research and education.

New Biology: A paradigm shift

Over the past several decades, biologists have employed a *reductionist* approach, deconstructing biological systems into individual components and their associated interactions, yielding mechanistic insights such as the source of the activation or inactivation of a given protein. However, in the early 1990s, the development of high-throughput experimental technologies began to dramatically accelerate the rate and volume of data generation in biology, enabling complete genome sequences of organisms as well as rapid molecular-level measurements of gene expression, protein function, and protein-protein interactions, etc. Today, these technologies are facilitating the detection and measurement of large numbers of components and interactions at multiple scales simultaneously, allowing researchers to perform *integrative* systems-level studies with unprecedented breadth and depth.

At the same time, these new experimental technologies, the data they are generating, and the systems-level studies they are enabling are illustrating the unparalleled complexity that exists in biology. Consider, for instance, a cellular signaling network, i.e., the set of pathways that a living cell uses to transduce extracellular cues like the availability of water or oxygen into intracellular responses. Within every human cell, there are about 2,000 distinct genes involved in signaling, and on average, each gene can yield 20 unique proteins (i.e., the components of a cell that perform most life functions). In other words, these 2,000 signaling genes can synthesize 40,000 signaling proteins, each with its own function. These kinds of combinatorics illustrate the grand challenge of New Biology, i.e., being able to assimilate and mine large, noisy experimental data sets to understand and predict natural phenomena.

---

[1] Contact: Erwin Gianchandani, Director, Computing Community Consortium (202-266-2936; erwin@cra.org). For the most recent version of this essay, as well as related essays, visit http://www.cra.org/ccc/initiatives.
[2] http://books.nap.edu/openbook.php?record_id=12764&page=R1.

Research areas underlying New Biology

In general, there exist three principal sub-fields in New Biology: systems biology, computational biology, and synthetic biology. We summarize these fields below, highlighting the specific contributions of computing research to each one.

Systems biology: As the NAS report stated, "Improved measurement technologies and mathematical and computational tools have led to the emergence of a new approach to [address] biological questions termed 'systems biology' [that] strives to [integrate heterogeneous experimental data sets] and achieve predictive modeling [of biological systems]." Rather than pursuing the decades-old *reductionist* approach, interrogating individual components and reactions underlying a given system, systems biology attempts to *integrate* various biological structures and create predictive models representing systems-level functions and behaviors.

For example, in 2007, systems biologists published a genome-scale reconstruction of the human metabolic network[3]. This reconstruction catalogs all known gene, protein, and reaction relationships underlying human metabolism – the vital cellular process that is attributed to many human diseases – in a highly quantitative, structured, and chemically consistent manner. In other words, the reconstruction assimilates all existing experimental knowledge about the system, and enables a quantitative analysis of the "flows" through the network – much like a map of a highway system overlaid with quantitative data about traffic volumes. Nearly 1,500 genes spanning 2,000 proteins and 3,300 reactions were incorporated from nearly 1,600 different papers. The resultant *model* represents the set of all hypotheses about the network that have been reported in the literature to date and, in turn, can be used to *predict which genes are essential or inessential, and which ones are involved in mechanisms of chronic diseases like cancer and arthritis*. Ultimately, such a model *enables us to better understand the manifestation of human diseases and identify ideal drug targets to combat these illnesses*.

Computational biology: Whereas systems biology takes an integrative, systems-based approach, computational biology applies data mining, machine learning, graphics/visualization, and related computational techniques to specific biological questions. For instance, clustering algorithms have been applied to gene expression data to associate genes with similar functions. High-throughput gene expression assays are enabling us to measure the expression levels of thousands of genes simultaneously, across different conditions and over time. These assays result in incredibly large data sets: the expression of each gene requires multiple "probes," meaning that there are often 20 or more data elements per gene, and a routine experiment involving human cells measures 54,000 human gene transcripts concurrently. By clustering these data, we are able to make sense of the data and gain insight into gene function; genes that respond similarly to different stimuli are more likely to have related functions. Likewise, "compendium analyses" are used to study the mechanisms underlying drug function, by comparing the gene expression profiles of unknown drugs with databases of profiles of known drugs. Drugs with similar mechanisms are likely to have correlative gene expression footprints[4].

---

[3] Duarte, N.C., et al. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci.* **104**(6): 1777-82.

[4] Subramanian, A., et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**(43): 15545-50.

Synthetic biology: Synthetic biology involves building biological parts and structures. It relies upon systems biology for computational models that are then manipulated in different ways to simulate the behaviors of putative artificial structures. While synthetic biology is not directly a sub-field of computer science, it is "engineering" in every sense of the word – in fact, "molecular engineering" is a phrase that is sometimes used to describe synthetic biology and other related fields. Predictive modeling is critical for its success.

For instance, researchers recently engineered yeast to produce artemisinic acid, the precursor to the antimalarial drug artemisinin[5]. Malaria is a global health problem that threatens 300 to 500 million people and kills more than one million annually, and artemisinin is in short supply and unaffordable to most malaria sufferers. By modeling the metabolism of yeast and identifying key perturbations that would result in much more efficient synthesis of artemisinic acid, researchers were able to engineer yeast and can now generate the precursor in substantial quantities quickly and effectively. Subsequently, a simple and inexpensive purification process is used to obtain the desired product.

Key national challenges – and the promise of New Biology

As the examples above illustrate, and as the NAS report concluded, New Biology has the potential to drastically improve the quality of life in the U.S. in the areas of food, environment, energy, and health:

Food: According to the United Nations Food and Agriculture Organization, 923 million people were undernourished in 2007. Due to increasing population sizes and higher standards for food in developing nations, the growth of quality food is a rising challenge. New Biology will "deliver a dramatically more efficient approach to developing plant varieties that can be grown sustainably under local conditions." New Biology techniques will create models of plant growth at both cellular and molecular levels. Using these models, genetic changes in plants can be targeted in a highly predictable manner. This predictability "will make it faster and less expensive to develop plant varieties with helpful characteristics." This drastic increase in efficiency will greatly reduce the time from blackboard to blacktop for new plant varieties.

Environment: New Biology will help us "understand and sustain ecosystem function and biodiversity in the face of rapid change." Ecosystem restoration is a large interdisciplinary problem that will require biological models on both the micro- and macro-scales. These models will provide more accurate pictures of changes in the ecosystem and enable ecosystem engineers (i.e., the M.D.-Ph.D. equivalent for the environment) to mitigate the harmful effects of climate change.

Energy: "Making efficient use of plant materials – biomass – to make biofuels is a systems challenge, and this is another example of an area [in which] New Biology can make a critical contribution." Currently, each stage of the biofuel production process is inefficient. Various segments of the process can be made more efficient using New Biology, e.g., by modeling a system such as yeast and engineering a strain that is optimized for biofuel production. With the

[5] Dae-Kyun, R., et al. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**: 940-3.

increase in efficiency, biofuels could become a viable alternative to crude oil. There is currently a large debate about the role of biofuel, and the use of edible crops for food. Practical biofuels, which minimize land use, could diminish the concerns of the crops-for-food camp.

Health: "A goal of New Biology is to provide individually predictive surveillance and care." Within the next five years, DNA sequencing will be as cheap as a blood test, enabling doctors to provide patients with preventative care based on their genetic predispositions to specific diseases and traits. For example, the woman whose genome puts her at high risk for heart disease will be told to cut salt from her diet years before incidence of the disease, and the man who is genetically at high risk for colon cancer will start being screened for the disease 10 years earlier than normal. New Biology research directly enhances the quality and range of predictive – and preemptive – care.

A Federal agenda

The research and education challenges described above are consistent with the missions of multiple Federal funding agencies, including the National Institutes of Health (NIH), the Department of Energy (DoE) Office of Science and Advanced Research Projects Agency-Energy (ARPA-E), and the National Science Foundation (NSF), as well as other Federal agencies like the Environmental Protection Agency (EPA), the US Department of Agriculture (USDA), and the Food and Drug Administration (FDA), etc. While these agencies have sponsored several New Biology-related initiatives in the past decade, these efforts have largely been fragmented and lacking the focus on data analytics that is increasingly necessary.

For example, several years ago, the National Cancer Institute (NCI) launched the Cancer Bioinformatics Grid (caBIG), attempting to connect researchers and data through a shareable and interoperable infrastructure. caBIG provides standardized tissue specimens, complete with patient data (such as survival information), and supports open-source development for analyzing these specimens for biomarker discovery. Similar efforts are necessary to understand other fundamental diseases, such as diabetes, arthritis, and Alzheimer's.

Similarly, the Department of Energy (DoE) has provided funding for research through its Genomic Science Program, focusing most recently on bioenergy production, environmental remediation, and carbon cycling and sequestration. The President's budget request to Congress for FY 2011 includes $177 million for the Genomic Science Program, up from $166 million in FY 2010, and supports foundational genomics research, genomics analysis and validation, metabolic synthesis and conversion, computational biosciences, and bioenergy research centers. However, the "computational biosciences" research – including the development of models and computational tools to integrate diverse types of data from genomics, proteomics, and other experiments into single models that describe and predict the behavior of metabolic pathways and genetic regulatory networks – received just $8.3 million in FY 2010 and is budgeted at $12.7 million in the President's FY 2011 request to Congress. These amounts can support only a half-dozen research teams working in these areas – and barely scratch the surface of the terabytes of data that we have yet to mine.

We therefore call upon these Federal agencies to establish coordinated, multi-disciplinary, highly collaborative, and long-duration programs in New Biology, including a focus on the development of new tools to analyze the wealth of biological data and knowledge that we now have at our disposal. These programs must encourage teams of multiple senior investigators, graduate students, and senior personnel, spanning traditional "wet-laboratory" and more recent "dry-laboratory" researchers. And they must support the following advances:

- **Exploiting high-throughput assays for unbiased, global discovery and integrative, systems-oriented modeling/analysis.** Large-scale robust integration of biological data will require leaps in diverse computer science fields including pattern recognition, classification, machine learning, graphics/visualization, computer vision, distributed computing, data compression, encryption algorithms, database management, and Web services.

- **Modeling systems at multiple, interacting scales (e.g., molecular, cellular, organismal, epidemiological/ecological, and biosphere).** Models that will be developed will be enormously large, complicated in structure, and complex in function, and will include contributions from many disciplines. Multiple existing modeling platforms (such as agent-based models and intracellular network reconstructions, etc.) will need to be integrated. We need to ensure that such models are experimentally falsifiable; understandable to researchers from all contributing disciplines; reproducible by the rest of the research community; accessible to non-specialists (e.g., government bodies, media, etc.); and clear about their limits and uncertainties.

- **Providing interdisciplinary education, as well as also career structures that reward rather than penalize interdisciplinary research by faculty.** For both pedagogic and research purposes, funding is needed for multidisciplinary centers that span multiple departments and have their own resources, buildings, and career paths, etc. Examples of success stories in these areas include the Broad Institute of MIT and Harvard[6] as well as the Institute for Systems Biology (ISB)[7].

- **Successfully translating research into implementations with economic, ecological, and social benefits.** This translation will require new technologies to share data for research purposes, while guarding the rights of owners; monitor and optimize the use of new solutions; and guard against abuses of data.

Some research directions will inherently be agency-specific, such as data-analytics approaches for personal surveillance and care (NIH); for the development of biofuels more cost-effective than oil (ARPA-E); for ecosystem engineering (EPA); and for genetically-engineered crops (USDA). However, **other programs will require multiple agencies, as research into New Biology often falls at the intersection of NSF and NIH or NSF and DoE, etc**. For example, earlier this year, NSF and NIH announced first-ever joint New Biology programs, "New Biomedical Frontiers at the Interface of the Life and Physical Sciences" and "Transforming Medicine at the Interface of the Life and Physical Sciences." These programs seek to support

[6] http://www.broadinstitute.org/.
[7] http://www.systemsbiology.org/.

innovative new directions at the intersection of computing, biology, and engineering. They will provide principal investigators up to $1 million over five years, in areas spanning biological computing, theoretical modeling of intracellular processes, deep imaging technologies, and complex systems analysis, etc. **The NIH/NSF programs must be continued over the next several years – and expanded each year – to support dozens of new teams of researchers annually.**

The road ahead

The emergence of high-throughput experimental technologies in the past two decades has drastically altered how we study biology. Today, machines are generating large, heterogeneous, noisy data sets at astounding rates – far faster than any single person or computer can handle. Coupling these data with new and existing analytical approaches – including data mining, machine learning, and predictive modeling, etc. – has given birth to New Biology, which seeks to understand biological phenomena, such as mechanisms of disease or strategies by which organisms may be engineered to efficiently synthesize biofuels.

This emerging discipline has the potential to address the grand challenges the U.S. faces in food, environment, energy, and health. Already, we have come a long way from the days when experimental data were simply used to build predictive models; today, computational approaches are generating new insights and hypotheses that in turn are informing "wet-laboratory" experiments and driving our decisions.

But to fulfill the promise of New Biology – i.e., to generate landmark breakthroughs in areas of food, environment, energy, and health – we need to sustain this progress in the years ahead. We must enhance investment in basic research by the relevant Federal funding agencies, we must renew our commitment to New Biology education, and we must foster a close partnership between computer scientists, biologists, and engineers. Our competitiveness as a nation in the twenty-first century depends upon our success in this area.