Bioinformatics: 
New Technology Models for Research, 
Education, and Service

Gary K. Allen, University of Missouri
Overview

The field of bioinformatics has experienced extraordinary growth over the past decade, in large part due to the vast amounts of complex data generated by the Human Genome Project and the concomitant need for sophisticated approaches to data storage, retrieval, analysis, and sharing. Bioinformatics remains one of the most active scientific areas nationally and internationally. Research and education in life sciences are increasingly dependent on bioinformatics and advanced information technology infrastructures to acquire, manage, interpret, and distribute extremely large data sets generated by high-throughput experimental approaches. This research bulletin provides an overview of the opportunities, challenges, and strategies for information technology organizations to optimally facilitate bioinformatics and broader life sciences development efforts in higher education.

Highlights of Bioinformatics Technology Models

Today, bioinformatics approaches are used to decipher data and to organize information gathered from traditional biology. Compared to the fields of physics and chemistry, biology is still immature in terms of its ability to make predictions based on general principles. The ultimate goal of biology research is to develop knowledge that will allow predictive approaches to all of the life sciences. In this regard, the data acquisition, management, and analysis techniques of bioinformatics are key, as they provide a lens through which to view accumulated large-scale data sets and knowledge. The ultimate goals of bioinformatics are to present a complete representation of the cell and the organism and to computationally predict systems of higher complexity, such as the interaction networks in cellular processes and the phenotypes of whole organisms.

To fully accomplish these goals, a sophisticated information technology infrastructure is needed. Bioinformatics investigators need direct access to a network of distributed databases, tools, and computing power connecting desktop machines to supercomputers located across the globe. Moreover, contemporary biological investigations increasingly operate across scales of complexity that place further demands on an information technology environment. For example, as scientists study everything from a specific gene to a clinical symptom of disease related to a function of that gene, they need technology that allows them to do specific analysis as well as to help the project team visualize activity and results. They also need technologies to support research collaborations at a distance.¹

What Is Bioinformatics?

Bioinformatics applies principles of information sciences and technologies to complex life sciences data. Both bioinformatics and computational biology grew out of the traditional fields of health informatics, biostatistics, information sciences, mathematics, computer sciences and related engineering fields, molecular biology and genetics, and behavioral sciences. The National Institutes of Health (NIH) defines bioinformatics as
“[r]esearch, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data” and computational biology as the “development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems,” noting that although they are distinct, “there is also significant overlap between these two subspecialties.” Increasingly, the term bioinformatics is used inclusively to refer to areas of endeavor encompassing both bioinformatics and computational biology, and this convention will be used for this research bulletin.

**Evolution of Bioinformatics and Life Sciences**

The recent evolution of bioinformatics as a recognized field resulted from technological developments in both biology and information technologies. The 1980s saw the growth of high-throughput techniques such as high-speed nucleic acid sequencing and polymerase chain reactions that increased, geometrically, the volume of data generated by biological experimentation.

Relevant developments in information technology include new algorithms and methods of data retrieval and analysis, including tools to analyze data and technologies that facilitate efficient storage and retrieval of large data sets. Bioinformatics grew as a result of the Internet and, more recently, Internet2, which enable creation and shared access to large repositories of biological data and offer a vehicle for rapid publication of research results. The Internet also provides a way for researchers to access supercomputing resources for very complex computing tasks and distributed computing architectures such as grid computing.

Early bioinformatics activities mainly involved analysis of single biological entities such as DNA and protein sequences or analysis of three-dimensional protein structure. In contrast, contemporary approaches include methods for analyzing large numbers of genes or proteins simultaneously, such as the identification of clusters of related genes and networks of interacting proteins. With the complete genome sequences for an increasing number of organisms at hand, bioinformatics is beginning to provide both conceptual and practical methods for detecting systemic functional behaviors of the cell and the organism.

Advances in methodologies applied to biological research have done more than merely increase the volume of a relatively few data types; they have also resulted in the significant diversification of biological data. Contemporary genomics research investigates not only a complete set of genes and their precise locations in the chromosome but also gene similarity relationships within the genome and across species. Experimental technologies include those used to evaluate the expression of genes (transcriptomics) and those used to understand how the structure and function of proteins and their modifications allow them to contribute to life processes (proteomics).
What It Means to Higher Education

Supporting bioinformatics in higher education can be viewed as a challenge, but for creative information technology units, it can provide leveraging opportunities for a new model of cross-discipline support.

Information Technology and Bioinformatics: Challenges and Opportunities

In contrast to the more mainstream tools typically required by most campuses, information technologies to support bioinformatics activities must address some special needs. Specific areas that require careful planning, deployment, and maintenance include:

- **High-performance computing.** Activities such as database searching, sequence analysis, modeling, structural and functional analysis of biopolymers, molecular dynamics, statistical pattern recognition, and visualization are computationally intensive. Reasonable performance for many of these complex problems is only available with multiprocessor-clustered compute engines. The information technology infrastructure should be able to deploy at least a moderately sized high-performance computing resource, in conjunction with means for very large jobs to be processed remotely on supercomputers at, for instance, one of the National Supercomputing Centers. An alternative solution is grid computing, in which the processing power of geographically disparate computers is harnessed over high-speed network connections. Regardless of the specific architectures used, the computing infrastructure must be reliable, highly available, secure, flexibly adapted to multiple purposes, and scalable to allow for expansion as needs demand.

- **Very large data storage facilities.** Contemporary life sciences research generates extremely large amounts of data. From 10 megabytes to more than one gigabyte of raw data are typically spawned from a single high-volume sequencing or microarray experiment. Data analysis can yield even larger files, and project-specific databases are frequently created to support data-mining and discovery efforts. Effective storage strategies for bioinformatics are increasingly utilizing enterprise-class storage area network (SAN) and network attached storage (NAS) architectures to contain and effectively manage these data. There is significant value in the balanced, thoughtful use of high-cost, high-performance “production” storage for very large, complex analysis jobs, with less expensive storage subsystems for “near line” data storage. Offline storage repositories are needed for disaster recovery as well as archival purposes. Maximum return on investment for this offline storage might involve more traditional approaches, such as tape-based systems.

- **High-performance networks.** Efficient transport of the large data volumes requires high-speed networks. These data paths must run between the data-generating devices (for example, sequencers) and local computational resources, as well as between the campus edge and remote computational and
data repositories (for example, to download updates of DNA and amino acid sequences from GenBank). Some campuses have developed separate research networks to segregate these massive data flows from their administrative networks and to provide an environment for research and innovation in networking approaches supporting extremely data-intensive research and scholarship.

- **Specialty application software and management tools.** Bioinformatics software tools are extremely diverse, ranging from those originally developed decades ago in command-line, mainframe environments to newly developed software suites for desktop use. Support must be provided across a disparate array of software solutions.

- **Non-standardized desktop computing platforms.** Practicing bioinformaticians need desktop computers that support the specific tools they need for their craft. Researchers’ preferences (and, in many cases, requirements) for a particular operating system argue against widespread standardization of desktop computing appliances.

- **Integration.** Despite the heterogeneity described above, it is critical to provide integrated computational, networking, and data services to optimize research efficiencies. Creating integrated databases, with requisite metadata and annotation to facilitate data sharing, is a major challenge facing life scientists. Other challenges include developing process and workflow management solutions such as laboratory information management systems for sample tracking and billing through core analytical facilities. Increasingly, Web-based portals to analytical tools are being effectively deployed to remove obstacles associated with multiple operating system computing environments.

- **Staff.** Specially trained personnel, whose background frequently will include cross-disciplinary training in life sciences, are increasingly needed to maintain technology systems and to provide training to researchers needing access to esoteric analytical tools.

These requirements can be viewed as a challenge and an opportunity for information technology support organizations in higher education. For example, high-performance computing devices are expensive, but if deployed in conjunction with processes to facilitate shared use, these devices can be scheduled with fewer periods of inactivity and at lower total cost as compared to multiple devices maintained by individual researchers or labs. Acquiring these high-end systems also represents an opportunity for development of vendor partnerships. Since the systems are frequently at the leading edge of manufacturer production lines, these relationships can provide mutual benefits. For example, performance metrics deriving from campus-based computational systems dedicated to real-world tasks can be of significant value to vendors, who, in exchange for this benchmarking, may then be able to provide these or other products or services at more competitive rates.
Sharing resources also helps to reduce complexity of the information technology environment and minimize the staffing costs associated with its management. Other opportunities include using this infrastructure in support of campus research and education activities outside of bioinformatics, such as using high-performance networks to support distance education. Thus, these environments can leverage campus investments, avoiding unnecessary duplication of infrastructure that is expensive to procure and maintain.

**Bioinformatics: A Model for Interdisciplinary Studies**

Contemporary life sciences is increasingly an interdisciplinary effort, as evidenced by the emergence of academic research and educational centers in which faculty teams from across the natural and physical sciences are brought together to create synergistic investigative and scholarly groups. Not only are the individual teams within these centers diverse, but more and more frequently the targets of their endeavors span the entirety of the life sciences, including developmental biology, plant and animal sciences, environmental sciences, bioengineering, and human and veterinary biosciences.

These centers represent a significant departure from relatively small, single investigator projects. Multidisciplinary, large-scale, long-term life sciences research projects, perhaps the best example of which is the Human Genome Project, are viewed as having significant potential to efficiently and effectively analyze complex biological systems. A critical consideration in the design and management of these centers, from their architecture to their participating faculty and staff, is identification of means to foster collaboration and information exchange. Moreover, these centers invariably include bioinformaticians as a core contributing segment, representing the functional “glue” of these enterprises.

In response to these dramatic changes in how contemporary biological research is performed and to further drive the development of the “new biology,” many funding agencies have recently funded research programs targeting large-scale, cross-disciplinary programs. In each of the examples listed in Table 1, bioinformatics occupies a critical role.
Table 1. Role of Bioinformatics for Different Agencies

<table>
<thead>
<tr>
<th>Agency</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Institutes of Health</td>
<td>The NIH Roadmap defines a framework for the next 20 years of biomedical research in which bioinformatics plays a central role, for example, in developing pharmacogenetics and personalized medicine.</td>
</tr>
<tr>
<td>Department of Homeland Security</td>
<td>The DHS requires bioinformatics-dependent identification of biological weapons, epidemiological and environmental monitoring, and prevention and early warning systems for agroterrorism.</td>
</tr>
<tr>
<td>National Science Foundation</td>
<td>The NSF’s Tree of Life and Biocomplexity initiatives rely heavily on bioinformatics for both data acquisition and analysis.</td>
</tr>
<tr>
<td>Department of Energy</td>
<td>The DOE’s Genomes to Life project is developing and using bioinformatics methods and tools to model complex biological systems for use in energy production, mitigation of global climate change, and environmental remediation.</td>
</tr>
<tr>
<td>Department of Agriculture</td>
<td>The USDA is partnering with other agencies in developing genome projects for food crops and animals, as well as in the development of databases and techniques to track infectious diseases of domestic animals, including zoonotic diseases.</td>
</tr>
<tr>
<td>Institute of Medicine</td>
<td>The IOM has recently issued a series of reports on patient safety in which bioinformatics occupies a central role in developing processes to ensure appropriate patient care and to interpret medical records.</td>
</tr>
<tr>
<td>Centers for Disease Control</td>
<td>The CDC, through its Public Health Information Network, is developing a shared vocabulary and technical infrastructure for the rapid and accurate worldwide exchange of information for the detection and monitoring of incipient epidemics.</td>
</tr>
</tbody>
</table>

Even on campuses lacking large interdisciplinary centers, bioinformatics programs present significant challenges. Bioinformatics is itself perhaps without peer in the extent to which it is a multidisciplinary field. Bioinformaticians require training in computer and information science, statistics, mathematics, and the life sciences. Thus, bioinformatics does not gracefully fit into traditional departmental and college structures. Development of educational programs requires unequaled levels of coordination across typical academic disciplines. This is especially true on campuses that include medical schools, where contemporary genetic approaches in diagnostic and therapeutic clinical care and research increasingly depend on bioinformatics and where there is a desire to coordinate bioinformatics activities across the full range of the life sciences.  

The Biomedical Informatics Research Network (BIRN) provides excellent examples of the types of information technology required for today’s bioinformatics research environments. The BIRN is an initiative sponsored by the National Institutes of Health and the National Center for Research Resources that fosters large-scale biomedical science collaborations. The BIRN uses emerging cyberinfrastructure—high-speed networks, distributed high-performance computing, and software and data integration capabilities—to support a consortial effort among 12 universities and 16 research groups engaged in investigations of human neurological disease and associated animal models. Data derived from individual subgroups are being used to drive the definition, construction, and daily use of a federated data system, collected and stored across
geographically separated sites but presented as a unified data archive that can be securely accessed across institutional boundaries to address issues of data privacy and automatic translation of data formats. The BIRN is providing the means to test new hypotheses in the neurosciences through the integrated use of these shared data sets and site-independent resources.6

Bioinformatics activities also represent opportunities for invoking information technology best practices that might meet more significant resistance in other campus segments. Reasonable, appropriate management of rare information technology resources will certainly justify full invocation of security policies (both physical and cyber), data backup and restoration procedures, and disaster recovery processes. These can serve as examples to the general computing environment of the main campus. In environments in which data are generated from humans or human-derived samples, or when research is funded by corporate contracts, privacy and data confidentiality issues must be addressed.

Inasmuch as bioinformatics is itself interdisciplinary and its research inputs emerge from multiple academic segments, its information technology support should foster collaborative approaches. While bioinformatics research and education centers frequently have stand-alone information technology support operations, linkages between these centers and central campus information technology support groups can be leveraged to capitalize on investments in equipment and staff expertise. These arrangements represent a unique opportunity for campus information technology operational centers to directly develop rich, mutually beneficial working relationships with faculty researchers and educators. Particular advantages can accrue when bioinformatics activities act as a catalyst to drive cross-campus partnerships among life scientists, the central information technology shop, and diverse academic units. The benefits will extend far past the tightly coupled working environment required to best support bioinformatics research.

Key Questions to Ask

- What life sciences initiatives currently exist on campus?
- To what degree does the institution consider advanced life sciences studies a strategic priority?
- How can we better enable life sciences research and education by applying information technologies?
- Which elements of the information technology infrastructure are well positioned to support advanced life sciences initiatives? Which elements must be improved?
Where to Learn More


Endnotes


About the Author

Gary K. Allen (allengk@missouri.edu) is Associate Professor of Veterinary Pathobiology and Director of Advanced Computing Environments at the University of Missouri–Columbia and Executive Director of the University of Missouri Bioinformatics Consortium for the University of Missouri System.