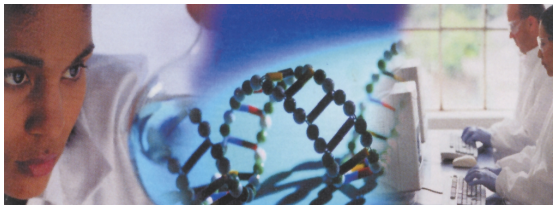
Information technology to the rescue!

Data management solutions are being tailored to the specific demands of large and small companies alike, says Jeffrey Augen.

It is unusual for a technology revolution in one industry to be enabled by advances in another. However, this is exactly the situation in the world of biotechnology, where several information technology (IT) platforms are coming together to enable the rapid advance of the drug discovery and target identification processes.

For the first time, IT has become the driver of experimental biology, a trend that is evident in genomics, proteomics, and the emerging field of metabolomics. High-performance computers, data-management software, and the Internet have helped to "industrialize" many aspects of biomedical research. For example, the basic steps of identifying, purifying, and cloning a gene, followed by characterization of the encoded proteins, have been automated and streamlined to a degree that no one would have predicted 10 years ago. IT has also maximized productivity within life sciences by enhancing collaboration. Life scientists now share data—both public and private—and collaborate in virtual environments.



However, genomics and proteomics generate staggering amounts of data, and the ability to process, analyze, and manage those data and the resulting information has become one of biotechnology's greatest challenges. The need for computational horsepower is driving emerging biotechnology companies to seek novel high-performance solutions (e.g., Grids and Linux clusters) from commodity-priced machines.

Byte-sized problems

It will be impossible for biotechnology companies to successfully leverage the massive amount of information available if they do not work across boundaries both within and outside their own organizations. Inefficiencies have already arisen from the inability to freely access and share information, and the inadequacy of conventional tools to explore relationships between pieces of data. For example, it is difficult to correlate messenger RNA expression patterns with previous research available through the public data sources. And pharmaceutical companies need to combine data housed in internal corporate databases with information held outside the IT firewall, such as subscription databases, public sources of information, and collaborations with partners.

Conservative estimates suggest that the study of the structure and function of proteins (proteomics) will involve 1,000 times more data and more computational power than that required for the Human Genome Project. More than one million human proteins regulate the structure and function of cells, tissues and organs, playing a key role in triggering and controlling most diseases.

Biotechnology and pharmaceutical researchers are amassing data at nearly one terabyte (10^{12} bytes) per week—the equivalent of 50 million printed pages of information. In just six weeks of work, these researchers will accumulate the equivalent of almost all the printed material in three academic libraries. Many biotechnology companies report that their IT needs are doubling every 6 months. For the first time in recent history, the IT demands of an industry are outpacing Moore's Law, which predicted that the performance of computers would double every 18 months.

Jeffrey Augen is director of business strategy, IBM Life Sciences Solutions, Armonk, NY.

Superpower solutions

To meet the ever-growing demands on their IT infrastructure, biotechnology companies will need to seek a variety of solutions: storage devices with file systems optimized for specific purposes, network hardware and software, large "back-end" computational engines and application servers, and client interfaces and devices with appropriate application software. To help reduce costs, smaller companies will look to IT companies as hosts of services, thereby sharing the costs with other companies.

Life sciences researchers will also have to grapple with the problem of data arising from highly heterogeneous sources. Biological data are geographically dispersed, and to uncover patterns and associations among genes and proteins, researchers need to organize and integrate this information. However, there is often little consistency across formats, access methods, and data models. Data sources must often be joined across the Internet through firewalls or across geographical boundaries on private data networks. Hardware platforms, operating environments, and application environments introduce additional heterogeneity into the equation.

Integrating IT

The key to continued growth, therefore, involves biotechnology companies carefully planning their computer infrastructure and striving to agree on data representation standards among industry participants. Efforts are underway to create standards both for medical informatics and bioinformatics: There are currently 25 different XML vocabularies focused on various life sciences segments, and draft standards have been proposed for chemical structures and molecular models. In addition, several organizations from the worlds of IT, drug development, biotechnology, and government research have recently teamed up to launch a standards organization known as I3C, which will develop a set of data interchange standards.

Data-warehousing solutions typically do not address the dynamic nature of genomic or proteomic data, because their construction involves collecting and combining disparate data sources. In the life sciences, data are constantly being refreshed, and conversion to a single format would be prohibitively expensive for most pharmaceutical operations—especially when dealing with some of the world's largest and fastest-growing databases. For example, the growth of GenBank, one of the most popular data sources, has skyrocketed from 1,700 bases per hour in 1990 to 833,000 bases per hour today.

Data-integration software also enables researchers to access heterogeneous repositories of data. For example, in the case of IBM DiscoveryLink, the use of "wrappers" allows users to access information housed in multiple databases regardless of the underlying formats, database structure, or query API (a programming interface by which an application program accesses operating system and other services). Other data-integration solutions allow users to connect to and view data sources through a single interface. Lion Biosciences' (Heidelberg, Germany) SRS solution, for example, provides access to over 400 different biological databases. Finally, in the absence of comprehensive turnkey solutions, users have often developed custom solutions that allow access to the specific data sources that they use, limiting users to a selected set of queries chosen from a menu.

Small companies should not underestimate their IT needs and will need to seek novel technical solutions and partnerships to provide the computational horsepower and storage needed to drive their organizations forward.

Undoubtedly, enormous computing power will be essential in the future. Supercomputing systems created today enable researchers to process trillions of calculations per second (so-called teraops). IBM's $100 million supercomputer research project, Blue Gene, will be capable of more than one quadrillion operations per second (one petaflop), 500 times faster than the fastest supercomputer today. Blue Gene's massive computing power will initially be used to study the protein folding process. Such efforts will someday enable researchers to gain a molecular-level understanding of disease states (*Nat. Biotechnol.* 18, 1024, 2000). IBM is also building the world's largest commercial supercomputer and will collaborate with NuTec Sciences, an informatics company (Atlanta, GA), to investigate how genes interact in the human body to cause life-threatening diseases.

Looking ahead

So what do small companies need to think about when planning IT infrastructure? For emerging biotechnology companies, competitive advantage hinges on their ability to obtain competitive computing technology, including large-scale storage, high-performance processing, and distribution of content. For these reasons, the biotechnology industry is increasingly turning to partnerships with IT companies to manage data centers, corporate intranets, and business IT functions.

Smaller life science companies may choose alternative means of meeting their IT requirements. Collaboration between IT and biotechnology companies, fostered through such arrangements as Internet portals, will become available as they have in other industries. The use of server farms, or "leased" supercomputing power, will help host larger amounts of biological data than could be afforded by any single company's computing budget. Finally, new technologies for sharing computing power have already made an impact on the physics community and will likely to surface in the biological world as well.

In addition, application service providers (ASPs) will provide biotechnology companies with remote sourcing and help managing technically complex applications using the Internet. Application of the ASP model to life sciences R&D efforts will enable small- and mid-size biotechnology and pharmaceutical companies to have the same access to state-of-the-art technology as their larger competitors. Internet-based management of clinical trials is one area where the ASP model is likely to emerge, bringing increased efficiency.

Overcoming these technological challenges is key to accelerating the drug discovery process, and in the future biotechnology companies will need to ensure that IT is an integral part of their early business planning. Small companies should not underestimate their IT needs and will need to seek novel technical solutions and partnerships to provide the computational horsepower and storage needed to drive their organizations forward.ons and partnerships to provide the computational horsepower and storage needed to drive their organizations forward.