Bioinformatics and information technology: reshaping the drug discovery process

Jeffrey Augen
IBM Life Sciences
Route 100
Somers
NY 10589
USA
tel: +1 914 766 3657
fax: +1 914 766 8370
e-mail: jaugen@us.ibm.com

During the past decade, advanced computing technologies have evolved into core components of both the drug discovery process and biotechnology in general. Although it is not unusual for advances in one industry to enable progress in another, it is unusual for such advances to become crucial differentiating factors and the drivers of growth – precisely the effect that information technology (IT) has had on the drug discovery process.

High-performance computing, data management software and the Internet are now facilitating the rapid industrialization of many of the processes that drive the drug discovery pipeline. The transformation is evident across all processes and includes, but is not limited to, gene discovery, structure and function determination of target proteins, combinatorial chemistry and the molecular dynamics of lead compound–target protein interaction and the selection of subjects for clinical trials. The latter is often based on computerized analysis of mRNA expression profiles and medical histories of the candidate patients.

Desktop computing

Many of these advances are related more to the availability of high-performance desktop computing than the availability of large-scale computing horsepower. The desktop is now home to machines that can execute over two billion instructions per second, store hundreds of billions of bytes of information, display structures with more resolution than the human eye can resolve and communicate across the Internet at speeds that facilitate real-time collaboration between researchers. As a result, the desktop computer has become the engine that drives modern scientific investigation.

The rapid march of desktop computing for biotechnology has its roots in the early 1980s when most instrument manufacturers began shipping their products with personal computers and associated software. Scientists quickly became comfortable using these machines to store output data from laboratory machines, and soon began to design algorithms for searching and analyzing the growing body of information related to DNA and protein sequence, structure and function. The efforts of these researchers eventually drove the launch of the public data infrastructure for biotechnology, which has grown to include hundreds of databases such as those of the European Bioinformatics Institute (Hinxton, Cambridge, UK) and the US National Center for Biotechnology Information (Bethesda, ML, USA). Today's scientists access these databases from their desktops for gene and protein sequences, 3D protein structures, medical and scientific literature and a variety of related medical and biological information. When combined with proprietary data obtained from internal company sources, this information becomes a crucial engine of discovery.

'Coopetition'

The process is remarkable in that it depends on the competitive sharing of publicly available information combined with the use of privately held data and tools. The dynamics of drug discovery, therefore, have taken on many of the characteristics of cooperative competition, sometimes referred to in the computer industry as 'coopetition'. Such strategies lend themselves to analysis through the mathematics of cooperative game theory, which has become the dominant model in economic theory and has made significant contributions to such diverse fields as political science and biology.

The trend is evident in the business models of a small number of leading edge drug discovery and biotechnology companies that have launched software and services businesses as an adjunct to their drug discovery business. Examples are diverse and include companies such as Celera (Rockville, ML, USA), Lion Biosciences (Heidelberg, Germany)

1359-6446/02/$ – see front matter PII: S1359-6446(02)02187-6

and Millennium Pharmaceuticals (Boston, MA, USA). Each of these organizations is directly involved in basic 'wet chemistry' research while also realizing significant revenue from operations directly related to software, algorithms, databases containing genomic or proteomic content, and integration services. These companies are willing to make core information technologies available to their competitors, despite the fact that these technologies are key differentiators for other parts of their business.

Such efforts represent the most advanced strategic use of these technologies and are paralleled in other industries such as banking and telecommunications, in which competitors share infrastructure, bandwidth, technology and even intellectual property portfolios while differentiating on other strengths such as business logistics and design skills. The remarkable conclusion is that advances in IT, specifically the availability of high-performance desktop computers, are driving both advances in drug discovery technology and changes to the basic business strategies of companies in the pharmaceutical industry.

Recent advances

One of the most startling changes has been in the area of structure prediction. Fifteen years ago, X-ray crystallography of proteins was a long, complex and tedious process that often took months, even years. Furthermore, it was almost impossible to solve structures of membrane-bound proteins and large protein complexes. Even when structures were solved they represented crystal structures, which often differ markedly from structures in solution. Diffraction data were collected on film, measurements were made by hand, analysis software was crude and models were built by hand using balls and sticks. Today's picture is far different. X-ray diffraction data are fed directly into computer systems in which electron density maps and candidate structures are rapidly calculated. These structures are displayed on advanced desktop systems and the software enables researchers to examine the data easily by rotating and modifying the subject molecule. Finally, other computationally intensive structure-prediction techniques, such as nuclear magnetic resonance (NMR), have advanced rapidly and researchers can now combine multiple sources of information to help drive more accurate structure determination than was previously possible. Computationally intensive chemical techniques, such as crystallography and NMR, are also being supplemented by a new generation of software designed around *ab initio* folding predictions from basic sequence information.

These technologies are driving the industrialization of structure prediction, and thousands of structures can now be determined in a timeframe that once allowed only a single solution. Ambitious efforts are already underway to determine the structure of every protein coded in the human genome. The result will be an explosion in the number of available target molecules and a dramatic increase in the amount of knowledge that can be gleaned through *in silico* studies of the interactions between targets and lead compounds – a technique that is also enabled by rapid advances in desktop systems and software.

The rapid march of IT is also enabling more precise modeling of the complex interactions between various metabolic systems and the effects of different pharmaceutical compounds on those systems. This new emerging field, commonly referred to as systems biology, has already begun to yield new insights into the ripple effects that occur when a complex metabolic system is perturbed at the molecular level. Such experiments are an important milestone in the history of drug discovery and will soon become a cornerstone of predictive medicine.

Target-rich, lead-poor

Finally, the industrialization of genome sequencing coupled with dramatic improvements in protein chemistry has left the pharmaceutical industry in a target-rich, lead-poor environment. Technical improvements in areas that drive target discovery – genome sequence analysis, protein structure determination and the study of protein–protein interactions – continue to widen the gap. However, information technologies have the potential to shift the balance by enhancing the effectiveness and speed of combinatorial chemistry approaches to lead identification and optimization.

Traditionally, combinatorial chemistry involved synthesizing and screening thousands of compounds, a tedious and expensive process even when accelerated by advanced robotics and micro-scale chemistry. Newer computational approaches enable researchers to identify the best lead compounds by modeling the environment around each atom in the binding domain of the target–lead complex. The need to use such tools becomes apparent when one considers that there are approximately 10^{20} possible molecules in the molecular weight range of a typical pharmaceutical compound (<500 Da). Lead identification must therefore be a directed process with combinatorial chemistry serving to validate structural assumptions. The effectiveness of the process can be enhanced by improving the quality of the assumptions, and a new generation of software tools is evolving to meet these requirements. These technologies are made available to researchers through advanced desktop workstations capable of generating and displaying molecular models and executing the enormous numbers of calculations required to build such models. Such decentralized computing models are essential in an environment in which flexibility and independence are a central theme, as they are in the modern pharmaceutical research lab.

The future

The rapid improvements in IT described here, coupled with explosive growth in the quantity of available bioinformatic data, are driving a migration from *in vivo* to *in vitro* to *in silico* research that promises to enable the greatest advance of our time – the launch of molecular-based personalized medicine and a molecular-level understanding of both health and sickness.