The evolving role of information technology in the drug discovery process

Jeffrey Augen

Information technologies for chemical structure prediction, heterogeneous database access, pattern discovery, and systems and molecular modeling have evolved to become core components of the modern drug discovery process. As this evolution continues, the balance between *in silico* modeling and 'wet' chemistry will continue to shift and it might eventually be possible to step through the discovery pipeline without the aid of traditional laboratory techniques. Rapid advances in the industrialization of gene sequencing combined with databases of protein sequence and structure have created a target-rich but lead-poor environment. During the next decade, newer information technologies that facilitate the molecular modeling of drug–target interactions are likely to shift this balance towards molecular-based personalized medicine – the ultimate goal of the drug discovery process.

Jeffrey Augen
Director, Strategy
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 25 years, the drug discovery process and a variety of information technologies have co-evolved to the point where they have become inseparable components of a pipeline that begins with basic research and ends with disease specific pharmaceuticals. The impact of this trend has been tremendous with regard to acceleration of the drug discovery process. For example, using traditional drug development techniques it took nearly 40 years to capitalize on a basic understanding of the cholesterol biosynthesis pathway to develop statin drugs – those that inhibit the enzyme 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase, the rate limiting step in cholesterol biosynthesis [1,2]. Conversely, a molecular-level understanding of the role of the HER-2 receptor in breast cancer led to the development of the chemotherapeutic agent Herceptin® within only three years [3]. The developers of Herceptin® enjoyed the advantages of *in silico* molecular modeling, HTS and access to databases containing genomic and proteomic information.

The entanglement of information technology (IT) and the drug discovery process has its roots in the academic research laboratories of the mid- to late-1970s. One key enabler was the development of air-cooled, floor standing 'minicomputers'. Although high-end machines of this era had less computing power than a modern day wrist-watch, they were the platform that launched a new age of laboratory automation. Sophisticated research groups usually developed their own software. Applications included Fourier transform analysis for X-ray crystallography, enzyme and chemical kinetics, various types of spectroscopy, early statistical algorithms for protein structure prediction and simple ligand binding experiments. In the 1970s, genetics was exclusively a bacterial science, gene cloning was new, monoclonal antibodies were yet to be discovered and the techniques of DNA and protein sequencing were cumbersome and time consuming. Little structural data was generated and researchers had not yet begun to populate databases with the results of their sequencing and structure determination efforts.

During the early 1980s, chemists and biochemists began using computer technology as a core component of their research effort. This era saw the launch of the personal computer, starting with the Apple II, which became ubiquitous in the laboratory environment, and ending with the IBM PC. By 1985, most laboratory instruments were associated with a personal computer and the relevant software. These devices included liquid chromatography systems, various types of spectroscopy equipment, protein sequencers, DNA and protein

1359-6446/02/$ – see front matter PII: S1359-6446(02)02173-6

synthesizers, and many other devices used around the laboratory. Many applications that were previously run on larger systems could now be executed using personal computers. Scientists began to use desktop machines routinely for data analysis and PCs proliferated throughout the laboratory and research office environment. During the mid-1980s eukaryotic genetics became a mainstream science and scientists began storing DNA sequence information (along with protein sequence information) in large, publicly shared databases. As DNA sequencing techniques improved, leaders in the field began to articulate a plan for sequencing the entire human genome. These thoughts spawned the public infrastructure that now contains dozens of protein and DNA sequence-and-structure databases. The wide availability of computer horsepower also spawned the development of algorithms for pattern discovery and sequence homology testing. These algorithms became the foundation for today's science of bioinformatics – a crucial component of today's drug discovery process.

Emergence of a public infrastructure for molecular biology

These trends continued through the late 1980s with the emergence of the National Center for Biotechnology Information (NCBI; Bethesda, MD, USA) and other international centers for the management of biological data, such as the European Bioinformatics Institute (EBI; Hinxton, England, UK) (discussed later). The NCBI, which was established in 1988, is a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). NLM was chosen for its experience in creating and maintaining biomedical databases and because, as part of NIH, it could establish an intramural research program in computational molecular biology. The collective research components of NIH currently make up the largest biomedical research facility in the world.

The NCBI assumed responsibility for the GenBank DNA sequence database in October 1992. NCBI staff with advanced training in molecular biology built the database from sequences submitted by individual laboratories and by data exchange with the international nucleotide sequence databases, European Molecular Biology Laboratory (EMBL; Heidelberg, Germany) and the DNA Database of Japan (DDBJ). Arrangements with the US Patent and Trademark Office enable the incorporation of patent sequence data.

In addition to GenBank, NCBI supports and distributes a variety of databases for the medical and scientific communities. These include the Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) of three-dimensional (3D) protein structures, the Unique Human Gene Sequence Collection (UniGene), a Gene Map of the Human Genome, the Taxonomy Browser and the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute (Bethesda, MD, USA).

Another center for genomic information and biological research, the EMBL, was established in 1974. In 1980 the EMBL Data Library was founded – the first central repository of nucleotide sequence data in the world (precursor to EMBL's EBI outstation, which officially opened in Hinxton, Cambridge, UK in 1997). The EBI has been highly successful as a center for research and services in bioinformatics. The institute manages databases of biological information, including nucleic acid and protein sequences, as well as molecular structures (EMBL website; http://www.embl-heidelberg.de/).

The launch of a public infrastructure for aggregating, managing and disseminating biological data, combined with technical advances in databases and computer hardware, became the foundation for modern day, computer driven, rational drug design. During the past decade we have witnessed the industrialization of many aspects of biological research, most notably the sequencing of the human genome, HTS of drug targets and high-throughput protein-structure determination. This trend towards industrialization has served the pharmaceutical industry well because it drove the development of well-defined computer infrastructure, which also supports the drug discovery process.

During the past few years, the private sector has spawned dozens of companies, which supply various components of the information infrastructure that supports drug discovery. Unfortunately, a lack of standards prevents the seamless sharing of data between these environments and drug discovery companies are forced to integrate their own computing platforms despite the wide availability of tools. Additionally, scientists must be able to retrieve information from the public infrastructure to complement the work performed in their own laboratories. Such efforts are driving the development of a specific class of tools that enable scientists to query databases with heterogeneous data structures and naming conventions.

Finally, advances in computer technology are facilitating the development of a new class of applications for *in silico* biology. Included are applications for studying molecular dynamics, predicting tertiary protein structure, modeling binding kinetics, and a variety of applications designed to optimize the HTS process. As these tools mature, larger portions of the drug discovery process will make their way from the laboratory bench to the computer.

The ultimate goal is to be able to model a disease process at the molecular level, to predict which specific chemical

compounds are best suited to treating the disease for a genetically defined patient population, to perform all binding experiments *in silico*, and to accurately predict absorption, distribution, metabolism and excretion of the compound. Achieving these goals *in silico* will dramatically improve the drug discovery process and pave the way for personalized medicine based on a molecular level understanding of both the patient and the illness.

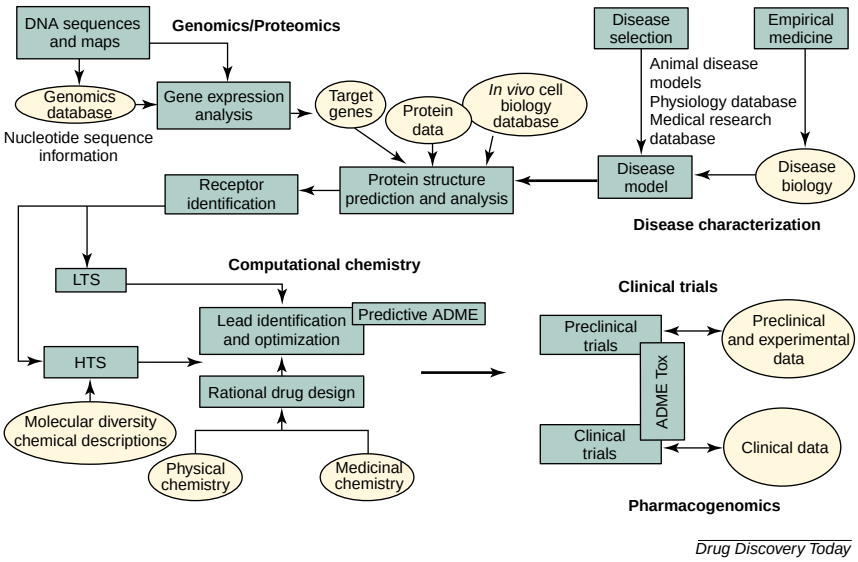

Figure 1. Flow of information through the drug discovery pipeline. Information from physiological databases, medical records and other sources drives disease selection. Target validation involves extensive analysis of protein structure and genomic sequence information. Target and lead compound interactions are the focus of computational chemistry techniques used to select final compounds for clinical trials. Finally, medical informatics techniques use databases containing information correlating clinical outcome with specific genetic makeup of the patient, gene expression data and disease states. Results from clinical trials are fed back to enhance the next round of target selection and lead compound optimization.

Dissecting and computerizing the drug discovery pipeline

The basic drug discovery pipeline is well known within the pharmaceutical industry. It consists of seven basic steps – disease selection, target hypothesis, lead compound identification, lead optimization, preclinical trial testing, clinical trial testing and pharmacogenomic optimization [4]. In actuality, each step of the pipeline involves a complex set of scientific interactions and each interaction has an information technology component that facilitates its execution. For example, the process of target validation requires access to data from a variety of sources with the goal of gaining a molecular level understanding of the potential role of specific protein molecules in modulating the progress of a disease state. Targets are identified using information gleaned from databases containing information about genomic sequences, protein sequence and structure, and mRNA expression profiles present in specific disease situations. The ultimate goal is to integrate these steps into a seamless process and to provide external interfaces to data sources regardless of differences in structure, design and data definition.

Although the simple pipeline model is both complete and correct with regard to logistics, the interactions between crucial components of the system must be described in detail to be useful to someone designing a computer infrastructure to support a specific environment. Various relationships between the pipeline's information components are outlined in Fig. 1. The flow of the diagram begins at the upper right with information about the basic biology of disease states. Information is combined from a variety of sources including physiological databases, medical records and other sources containing animal models for human diseases. The next section of the diagram represents the various activities that contribute to target selection and validation. Information from protein structure and/or sequence and genomic sequence databases is combined to help identify target molecules. Many of these targets are receptors that are displayed on the cell surface. An important part of target validation often involves purifying and studying a specific receptor to determine the parameters that molecules are likely to bind most efficiently. This binding of the target and lead compound is the subject of the next section of the diagram. Pharmaceutical companies often use combinatorial chemistry and HTS to predict these interactions. Additionally, during the past several years *in silico* techniques for predicting these molecular events have advanced to the point where biotech companies are beginning to skip much of the bench work involved in combinatorial chemistry and synthesize only the most promising compounds based on a structural understanding of the receptor and associated ligands. Finally, having identified the most relevant targets and selected the most promising lead compounds, the focus shifts to clinical trials. On the IT side, the emphasis becomes medical informatics and databases containing information correlating clinical outcome with the specific genetic makeup of the patient, gene expression data gathered from expression arrays and disease states. Results from clinical trials are fed

back to enhance the next round of target selection and lead identification and optimization.

Although each step in the process involves specific information tools, the tools are related and in some cases overlap. For example, much of today's animal model work involves comparative genomics including tools for multiple sequence homology and pattern matching. Many of these tools are also used to help find genes that code for target proteins. Likewise, both target validation and lead optimization are enhanced by the use of programs that facilitate predicting 3D structures of proteins and protein–ligand complexes.

Information technology in support of target identification

Target identification involves acquiring a molecular level understanding of a specific disease state and includes analysis of gene sequences, protein structures and metabolic pathways. The ultimate goal of the process is to discover macromolecules that can become binding targets for lead compounds, each one a potential drug.

One of the most important tools is the human genome itself and associated annotations. Sources of information include the public data infrastructure (NCBI and EBI), as well as private companies, such as Celera Genomics (Rockville, MD, USA) and Incyte Genomics (Palo Alto, CA, USA). The tools infrastructure is as important as the data and includes algorithms for sequence homology searching, transcription profiling software, and a new class of algorithms that aid in the process of structure prediction – one of the most computationally intensive exercises in the drug discovery process [5].

Computational structure prediction techniques are crucial because they can provide 3D information about the vast majority of proteins whose structures cannot be determined experimentally – membrane bound proteins, large complexes, certain glycoproteins and other molecules that cannot readily be crystallized [6]. IBM Research recently announced its intent to build a supercomputer, called Blue Gene/L, which will be optimized for protein structure prediction and other similar problems that lend themselves to parallelization. It was determined that the smallest machine that can be reasonably used to simulate the folding process will be capable of operating in the range of 1 petaflop – 1×10^{15} floating point operations per second. Furthermore, such a machine will require the order of one month to model the folding of even the smallest proteins. Box 1 depicts the complexity of the folding process in terms of time steps [7].

The calculation assumes that the protein being folded is very small, <100 residues, requiring an environment of only ~6000 water molecules, and that it folds quickly (<0.1 ms). With a time step of 5×10^{-15} sec, a petaflop size machine would still require 20 days to fold this small protein. By comparison, the largest commercial supercomputer in use today is approximately 7.5 teraflops (10^{12} floating operations per second) and the largest machine built to date, the ASCI White computer used by the US Government for nuclear simulations, is ~12 teraflops. These machines, despite their size and power, are two orders of magnitude too small to be useful for protein folding simulations. Therefore, it is necessary to restrict the size of the problem.

Box 1. Computational complexity of protein folding

Number of atoms	2×10^4
Number of instructions per atom per time step	5×10^6
Number of instructions per time step	1×10^{11}
Physical time for folding (sec)	1×10^{-4}
Physical time step (sec)	5×10^{-15}
Number of time steps needed	2×10^{10}
Number of instructions	2×10^{21}
Number of seconds in 20 days	1.7×10^6
Number of instructions per sec	1×10^{15}

Fortunately, many options are available to those using structure prediction as part of the target identification process. In many cases, for example, the molecule being modeled is membrane bound and the calculation is focused on the portion known, through chemical analysis, to be exposed outside the cell membrane. This approach has been helpful to researchers working to identify antibody targets on the surface of infected T-helper cells in HIV seropositive individuals. A recently created monoclonal antibody is believed to bind to a region of the gp41 transmembrane glycoprotein close to a transmembrane domain. This region is accessible to neutralizing antibodies and could form a useful target for vaccine design [8].

Another example involves using bioinformatic tools to predict which portion of a protein sequence is likely to be a biologically active binding site, and to model the specific structure of that site. An interesting example is the gp120 glycoprotein, a cell surface protein that, like gp41, mediates HIV entry into target cells initiating the replication cycle of the virus. The crystal structure of the core of gp120 has been solved recently [9,10]. It reveals the structure of the conserved HIV-1 receptor binding sites and some of the mechanisms evolved by HIV-1 to escape Ab response. The protein consists of three faces. One is largely inaccessible on the native trimer, and two faces are exposed but apparently

have low immunogenicity, particularly on primary viruses. The investigators modeled HIV-1 neutralization by a CD4 binding site monoclonal antibody, and propose that neutralization takes place by inhibition of the interaction between gp120 and the target receptors as a result of steric hindrance. Such structural knowledge could assist in the design of new AIDS drugs and vaccines [11].

Unfortunately, the chemical and structural experiments mentioned previously are not always feasible because many proteins involved in specific disease states are expressed as 'minor messages' and are difficult to purify and study. Furthermore, the single-protein–single-illness approach is very naïve and most medical situations can only be described in terms of the interactions of many metabolic pathways and protein–protein interactions [Institute for Systems Biology (Seattle, WA, USA); http://www.systemsbiology.org].

Gene versus transcript

In response, researchers have turned to mRNA expression profiling where the goal is to correlate the complex patterns of gene expression and medical history to treatment outcomes. The technique involves profiling the up- and downregulation of specific genes using microarray technology and analyzing the resulting data to help identify potential protein targets for drug therapy. Information technology is a crucial component at many stages of the process beginning with target identification, where a single microarray can produce hundreds of thousands of individual spots, each containing information about the expression of a single nucleotide sequence. Software products are often used to reorganize the information according to certain criteria, such as cell function–protein synthesis, carbohydrate metabolism, energy production, and so on. Clusters of regulated genes are ultimately identified, after which heterogeneous database access tools are often used to reach out across firewalls to the public infrastructure for information about the individual spots on the array. Once retrieved, these data are matched to information that is contained within proprietary databases inside the firewall. Additionally, large-scale database searches using pattern-matching algorithms are used to find expression profile matches in databases containing millions of such patterns. The task is similar to that of scanning a database containing millions of fingerprint records for patterns similar to a specific reference fingerprint. Finally, in the vast majority of situations, where thousands of genes are involved, it is necessary for a given set of expressed genes to correlate specific up- and downregulation patterns with treatment outcomes and phenotypic changes. Establishing the relevant correlations requires analyzing complex datasets using knowledge management tools and searching data sources with different schemas and data structures.

It is important to note that genomic information for target discovery can come from two distinct sources – the genome itself (DNA sequence information) and studies of mRNA expression patterns. Direct use of genomic sequence information can be a difficult endeavor because the information is in some sense encrypted. Individual genes code for multiple proteins, sometimes numbering in the hundreds [12]. Furthermore, the mechanism by which this complexity is accomplished frequently involves alternative splicing of mRNAs, and it is difficult to accurately locate splice sites and correctly piece together exons to assemble complete coding sequences [13]. Finally, the signaling system and controlling elements that regulate splicing decisions are poorly understood, making it extremely difficult to predict the sequences of mRNA or protein end product from chromosomal sequence information. Predicting the up- and downregulation of a given message is even more complex [14].

To grasp the complexity of these mechanisms one needs only to realize that the human genome contains approximately 10 times as many genes as the *Escherichia coli* bacterial genome, yet codes for an organism that is millions of times more complex [15–17]. The difference between an *E. coli* bacterium and a human is not so much related to the gene count as it is to the complexity of gene expression. Working with mRNA expression profiles eliminates some of the need to unravel this complexity, which in itself has become a challenging IT problem [18].

As data analysis tools become more sophisticated, researchers are beginning to take a 'systems' approach to understanding the complex web of interactions that define a cellular pathway. Recent approaches have included modeling a cellular pathway where components were analyzed using DNA microarrays, quantitative proteomics and databases of known physical interactions. In one specific project conducted at the Institute for Systems Biology, a global model was constructed, which was based on 20 systematic perturbations of a system containing 997 mRNAs from the galactose-utilization pathway in yeast. The experiments provided evidence that 15 of 289 detected proteins are regulated posttranscriptionally and the model that emerged identified explicit physical interactions governing the cellular response to each perturbation [19]. Such experiments are an important milestone in the history of drug discovery because they demonstrate that it is possible to develop and test complete systems models, which have the potential to advance the cause of predictive medicine rapidly. It is also important to note that the effort brings together contemporary tools of molecular biology with the latest in

IT – structure and pathway databases and expression array analysis–bioinformatics software.

Expression profiling has already yielded important results and academic research centers are beginning to populate large databases with specific patient information. One such project was recently announced between NuTec Sciences in Atlanta (GA, USA), the Winship Cancer Center associated with Emory University Medical School (Atlanta, GA, USA) and IBM (Armonk, NY, USA). Four specific areas of cancer were selected for study – breast, colorectal, lung and prostate. Information gathered from expression array profiling of various tissues and body fluids of affected patients is being analyzed in the context of medical history. The goal of the project is to make treatment decisions based on the newly discovered correlations. New algorithms are also being used to identify statistical correlations between genes and clusters of genes to help identify related traits and further refine the profiling technique.

Such projects can serve a multitude of purposes, including drug rescue for compounds that might fail clinical trials if tested on a patient population that is genetically too broad. Herceptin®, mentioned previously, is a perfect example of a compound that is only effective for certain patient populations and could not pass a clinical trials test in a patient population that was not genetically defined. The linkage between databases of expression data, target identification and clinical trials is important because it has the potential to revolutionize the drug discovery and delivery process. The same databases that support target identification, a pure research activity, will certainly support clinical trials, as well as medical diagnostics for clinical use. It is probable that diagnostics companies of the future will deploy the same databases as researchers and use them to make treatment recommendations. In the near future, treatment decisions will be made through a process that combines initial genetic data (DNA sequence information) with mRNA expression data and information from an individual's medical record – the exact same information that will be used by researchers to identify targets and test the efficacy of new therapeutics. The trend towards molecular-based personalized medicine will, therefore, drive closer ties between each step of the drug discovery pipeline and clinical medicine.

Information technology infrastructure – expression databases, links to medical record systems, data mining applications and algorithms for data analysis – will be central to these systems and their success will depend on the reliability and ease of use of the infrastructure. Furthermore, these systems will need to integrate medical informatics with bioinformatics. These activities are already driving the development of standards in the form of XML vocabularies for expression array data analysis, as well as other interfaces and standards for data exchange in the Life Sciences [20–22]. Over the next few years, these standards will evolve and mature in the same way that web-based standards have evolved. Those involved in the definition of these standards will have a decided business advantage and, for the first time ever, advantages in IT could translate into clear business advantages in the pharmaceutical and biotech industry.

Although, as mentioned previously, DNA sequence information is very dense and complex to interpret, an understanding of gene expression at the chromosomal level is often key to identifying targets for drug therapy [23]. This strategy is especially useful when the disease in question involves complex patterns of gene regulation and the targets sought are involved directly in modulating transcription. Appropriate therapies might directly block transcription of a key gene or interfere with the associated regulatory pathway.

The path from gene to target

An interesting and unique example that directly relates core genetic information to drug therapy involves polymorphisms in genes encoding drug metabolizing enzymes, drug transporters and/or drug targets. Recent experiments have taken a genome-wide approach to understanding the network of genes that govern an individual's response to drug therapy. For some genetic polymorphisms (e.g. thiopurine *S*-methyltransferase), monogenic traits have a dramatic effect on pharmacokinetics and ADME in general. Individuals with the enzyme deficiency must be treated with less than 10% of the normal thiopurine dose [24]. Likewise, polymorphisms in drug targets change the pharmacogenomics of the drug response for compounds specific for those targets – β-agonists are a notable example [24]. These pharmacokinetic effects are almost always determined by the interplay of several gene products and single nucleotide polymorphisms (SNPs). Unraveling the polygenic determinants that drive these effects is a complex IT problem, which includes kinetic modeling of metabolic pathways, algorithms for gene identification and gene structure prediction, and SNP databases.

It is also possible, using modern computational techniques, to conduct *in silico* binding experiments that model the interactions between small molecules and DNA [25]. This capability, which dates back to the late 1980s, is helping to launch a new generation of pharmaceuticals that bind directly to DNA to effect transcription. Although early experiments focused on using sequences that are known binding sites for small molecules as predictors for other DNA–small molecule interactions, more recent

studies have focused on more complex binding affinity calculations as the predictor [26]. Additionally, when the target is often a bacterial or fungal genomic site, an emerging area of focus is the regulation of transcription of disease-causing genes in humans.

Finally, tumors often express unusual combinations of genes. These genes typically represent a combination of mutant and normal phenotypes, and could be helpful with regard to target identification or tumor classification. Such is the case with breast cancers in individuals who exhibit *BRCA1* or *BRCA2* mutations and can be classified by their differential expression of a large number of related genes. One recent study included a statistical analysis of 5361 expressed mRNAs in breast cancer patients; a third exhibited *BRCA1* mutations, a third exhibited *BRCA2* mutations, and a third represented sporadic cases of the disease. An analysis of variance between the genotypes revealed 176 genes that were differentially expressed in the *BRCA1* and *BRCA2* cases suggesting that a heritable mutation influences the gene-expression profile of the tumor [27]. The ability to classify tumors according to their mRNA expression profiles is likely to become a cornerstone of diagnostic medicine in the immediate future. The tools that will enable this approach include databases containing medical histories, databases of mRNA expression profiles, pattern matching algorithms to compare mRNA profile images stored in the databases, and statistical analysis software. As the technique makes a transformation from research tool to clinical diagnostic procedure, large complex computer infrastructure will be required to link clinical physicians with appropriate databases and tools. This infrastructure will also provide the links to diagnostic companies or other entities that will be responsible for aggregating and managing the data.

Information technology in support of lead development and lead optimization

The next major stage in the drug discovery pipeline involves the identification of lead compounds as a complement to the target discovery process. During the past decade, industrialization of genome sequencing coupled with dramatic improvements in protein chemistry has left the pharmaceutical industry in a target-rich lead-poor environment [28]. Technical improvements in areas that drive target discovery have been dramatic over the past few years, and the pace is accelerating. For example, it took 22 years for Max Perutz and his colleagues at Cambridge University (Cambridge, UK) to complete the 3D structure of hemoglobin – an enormous accomplishment that was completed in the late 1950s [29]. The techniques of X-ray crystallography remained relatively unchanged for 30 years until, during the late 1980s, film was replaced by electronic devices that record X-ray diffraction patterns, and ball-and-stick models were replaced by computer programs for building and testing molecular models. These technical achievements have resulted in the emergence of a new field, known as high-throughput X-ray crystallography (HTX). Today, complete structures are routinely determined in days instead of years. Newer computational techniques, which will be discussed later, can determine the structure of crucial domains, where binding is likely to occur, at the rate of several structures per day.

Multiple order of magnitude improvements are also evident in DNA sequencing technology. Graduate students of the late 1980s often spent an entire graduate career sequencing a single gene. Conversely, modern day automated sequencers are capable of determining approximately one million base pairs per day, and new sequencing logistics combined with advanced algorithms have enabled a complete rendering of the human genome – three billion base pairs long – years ahead of schedule. (Sequencing of the human genome, completed early in 2001, was accomplished through the combined efforts of a publicly funded project and a private corporation. The public project was led by Francis Collins, Director of the National Human Genome Research Institute (Washington DC, USA); the private effort was led by Craig Venter, President of Celera Genomics.) One of the key enabling technologies was a technique called 'shotgun sequencing' [30,31]. A core component of the shotgun technique is the 'assembly' software used by Venter and his team at Celera. This software facilitates the positioning and alignment of millions of overlap sequences and obviates the need for the initial creation of detailed chromosomal maps. It has been calculated that the assembly of the human genome sequence is the most complex computer problem ever solved. To accomplish this feat, Celera built an enormously powerful IT infrastructure and hired an appropriately skilled team of programmers and computer scientists. Sequencing of the human genome was not so much a feat of wet chemistry or molecular biology as it was a feat of computer science.

For the pipeline to be efficient, improvements in lead identification and optimization must keep pace with the rapid improvements in the target identification technologies outlined previously. Historically, lead identification has involved wet chemical techniques – combinatorial chemistry combined with HTS to test for binding affinity. These techniques have not served the pharmaceutical industry well. At present, only 1% of all pharmaceutical research projects ever materialize into marketable drugs; the remainder are written off [32]. Worse still, many of the failures could potentially be saved through drug rescue

projects that focus on genetically defined targets and appropriately refined lead compounds.

The solution to increasing the efficiency of the lead identification and optimization process must involve structure-based drug design – a highly refined example of advanced computer modeling techniques. In structure-based drug design, various computer-generated 3D structures of a drug–target combination are used to guide lead discovery. This structural information can be obtained through X-ray crystallography or NMR, or a combination of both [33].

Early goals for structure-based drug design focused on *de novo* approaches where the outcome was a custom designed lead compound. The proposed binding site of a protein was space filled with a custom designed compound that had perfect charge complementarity and shape. Unfortunately the approach relied on molecular docking software and, more specifically, energy and thermodynamics calculations that were not accurate enough predictors of the potency of various lead compounds. The effect of these initial failures was to drive researchers to synthesize and test enormous numbers of compounds – essentially screening every combination possible for a given target. This approach launched the era of combinatorial chemistry and HTS, which we are just now beginning to exit.

In some cases it is possible to isolate interesting targets with attached ligands. Because these ligands are similar to the lead compounds being sought, X-ray structures of the protein–ligand complex can often reveal most of the information needed. It is now technically possible to generate accurate computer models of the portions of the molecules in question by analyzing electron density maps of the protein–ligand complex. In other situations, the structure of the protein is known and it is possible to predict and model, *in silico*, the part of the molecule where a ligand will probably bind. These techniques enable the rapid *in silico* screening of large numbers of compounds against a specific target and are fundamentally different from previous uses of X-ray crystallography and *in silico* modeling [34].

For a variety of reasons, structure determination as a basis for lead compound identification often proceeds in the complete absence of 3D structural information. Membrane proteins and large protein complexes, for example, are difficult and sometimes impossible to crystallize. Other proteins are difficult to purify because they represent minor messages in the cell. Finally, X-ray and NMR structures represent the behavior of a protein in specific environments and might not accurately model reality within the cell. For these and other reasons it is important to develop *in silico* techniques for protein structure prediction that rely only on base sequence information and can be adjusted to model folding in a variety of environments [35]. Other approaches combine information extracted from databases of known structures with experimentally obtained structural information and computational modeling to guide the lead compound screening process. Such approaches have driven active molecule hit rates of about 10%, as opposed to rates of only about 0.01% with conventional HTS techniques [36]. The output of this process includes new lead compounds as well as additional data that can be used to populate bioinformatics databases. As these databases grow in size and complexity, the pace of lead compound identification and the efficiency of lead optimization will increase until all stages of the pipeline operate at similar rates.

The future of *in silico* drug discovery

It is too early to predict if we will ever reach the ultimate endpoint of drug discovery – personalized molecular-based medicine. However, as the march towards molecular medicine continues, it is clear that IT will have a key role and that the drug discovery process will mature until *in silico* modeling is the central theme. Someday, in the near future, targets will be discovered by modeling genetic and metabolic processes and lead compounds will be synthesized and tested almost entirely *in silico*. Unfortunately, the systems that enable such advances will be enormously complex and it is unlikely that there will ever be a simple 'push button' solution that takes the process from disease identification all the way through to clinical trial.

Target identification will require access to heterogeneous databases containing genetic sequences, protein structures, metabolic pathways and mRNA profiles. These data sources will be used to create a molecular-level model of each specific disease, which includes genetic predictors and an understanding of the relationships between various mRNA and protein expression profiles and specific disease states. Once a target is identified its structure must be determined so that appropriate lead compounds can be synthesized. As mentioned previously, the most advanced computers available today are too small to model the folding of a single protein in any reasonable time. The next generation machine will make a small but noticeable dent in the problem by enabling us to model the folding of a single small protein in the timeframe of a month. The computer horsepower march will continue at the pace predicted by Moore's law [37] – performance doubling every 18 months – and by 2010 we will be able to predict the folding of simple proteins in only a few days.

Once a target is chosen and its structure is known, researchers will begin the process of identifying leads and modeling the interaction between target and lead – an

even more complex molecular dynamics experiment than generation of the original target structure. Finally, *in silico* ADME modeling will be used in advance of clinical trials to predict the metabolic behavior of the drug. Such modeling will require a detailed understanding of all related metabolic pathways and the effects the drug has on the expression of key genes whose protein products modulate those pathways.

Besides being computationally intensive, the process described previously requires access to many diverse data sources, some of which might not yet exist. At the time of writing, the life sciences community has not come to agreement on various standards for data interchange and data representation. A variety of standards exist for the representation of gene and protein sequences, as well as chemical structures, and software interfaces are defined at the time of development. Efforts are underway to create the required data standards and interfaces and these efforts will probably result in dramatic improvements in the ability of diverse organizations to pool their talents to drive the drug development process.

Finally, the pharmaceutical company of the future will be much different from that of today. Information technology infrastructure, efficient use of bioinformatics tools and the ability to leverage computer horsepower against biochemical and genetic insight are likely to become key business differentiators for the industry.

References

1 Corsini, A. *et al.* (1999) New insights into the pharmacodynamic and pharmacokinetic properties of statins. *Pharmacol. Ther.* 84, 413–428
2 Corsini, A. *et al.* (1995) Pharmacology of competitive inhibitors of HMG-CoA reductase. *Pharmacol. Res.* 31, 9–27
3 Slamon, D.J. *et al.* (1989) Studies of the HER-2/neuproto-oncogene in human breast and ovarian cancer. *Science* 244, 707–712
4 Clulow, M. (2001) *Life Science Informatics.* UBS Warburg LLC, 12 April 2001 report
5 Head-Gordon, T. and Wooley, J.C. (2001) Computational challenges in structural and functional genomics. *IBM Systems J.* 40, 265
6 Baker, D. and Andrej Sali, A. (2001) Protein structure prediction and structural genomics. *Science* 294, 93–96
7 IBM Blue Gene Team (2001) Blue Gene: a vision for protein science using a petaflop supercomputer. *IBM Systems J.* 40, 310
8 Zwick, M.B. *et al.* (2001) Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type-1 glycoprotein gp41. *J. Virol.* 75, 10892–10905
9 Malenbaum, S.E. *et al.* (2000) The N-terminal V3 loop glycan modulates the interaction of clade A and B human immunodeficiency virus type-1 envelopes with CD4 and chemokine receptors. *J. Virol.* 74, 11008–11016
10 Ye, Y.J. *et al.* (2000) Association of structural changes in the V2 and V3 loops of the gp120 envelope glycoprotein with acquisition of neutralization resistance in a simian-human immunodeficiency virus passaged *in vivo*. *J. Virol.* 74, 11955–11962
11 Poignard, P. *et al.* (2001) gp120: Biologic aspects of structural features. *Annu. Rev. Immunol.* 19, 253–274
12 Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* 32, 279–305
13 Li, Y. and Blencowe, B.J. (1999) Distinct factor requirements for exonic splicing enhancer function and binding of U2AF to the polypyrimidine tract. *J. Biol. Chem.* 274, 35074–35079
14 Sun, H. and Chasin, L.A. (2000) Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* 20, 6414–6425
15 Morton, N.E. (1991) Parameters of the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 88, 7474–7476
16 McClelland, M. *et al.* (2000) Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica serovars*, *typhimurium*, *typhi* and *paratyphi*. *Nucleic Acids Res.* 28, 4974–4986
17 Aparicio, S.A. (1998) How to count human genes. *Nat. Genet.* 2, 129–130
18 Sze, S.H. *et al.* (1998) Algorithms and software for support of gene identification experiments. *Bioinformatics* 14, 14–19
19 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* 292, 929–934
20 Achard, F. *et al.* (2001) XML, bioinformatics and data integration. *Bioinformatics* 17, 115–125
21 Gilmour, R. (2000) Taxonomic markup language: applying XML to systematic data. *Bioinformatics* 16, 406–407
22 Xie, G. (2000) Storing biological sequence databases in relational form. *Bioinformatics* 16, 288–289
23 Avise, J. (2001) Evolving genomic metaphors: a new look at the language of DNA. *Science* 294, 86–87
24 Evans, W. and Johnson, J. (2001) Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genetics.* 2, 9–39
25 Galat, A. (1989) Analysis of dynamics trajectories of DNA and DNA–drug complexes. *Comput. Appl. Biosci.* 5, 271–278
26 Ratilainen, T. *et al.* (2001) A simple model for gene targeting. *Biophys. J.* 81, 2876–2885
27 Hedenfalk, I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. Med.* 344, 539–548
28 Henry, C.M. (2001) Structure-based drug design. *Chem. Eng. News* 79, 69–74
29 Lehninger, A. (1975) *Proteins: three dimensional conformation*. In *Biochemistry* (2nd edn), pp. 145–146, Worth Publishers
30 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
31 Venter, J.C. *et. al.* (2001) The sequence of the Human Genome. *Science* 291, 1304–1351
32 Staff (2001) Drugs Ex Machina. *The Economist Technology Quarterly.* 22 September 2001, pp. 30–32
33 Lockwood, D. (2001) Structural proteomics: the upstream and downstream benefits of protein structure information. Cambridge Healthtech Institute articles, 1 June 2001
34 Stevens, R.C. *et al.* (2001) Global efforts in structural genomics. *Science* 294, 89–92
35 Duan, Y. and Kollman, P.A. (2001) Computational protein folding: From latics to all atom. *IBM Systems J.* 40, 297
36 Borman, S. (2000) Proteomics: taking off where genomics leaves off. *Chem. Eng. News* 78, 31–37
37 Moore, G. (1965) Cramming more components onto integrated circuits. *Electronics* 38