

## The commercialization of bioinformatics

**Phillip B.C. Jones**

E-mail: pbcj@wolfenet.com

Senior Patent Attorney, Patent Department

ZymoGenetics Inc., 1201 Eastlake Ave., East, Seattle, WA 98102

Tel: (206) 442-6681

Fax: (206) 442-6678

**Keywords:** Computational biology, Nucleotide sequences, Pharmaceuticals.

### Table of Contents

#### **Key factors for the development of bioinformatics**

Circumstance of Excessive Data

(a) Human Genome Project

(b) EST approach: Focus on active sequences.

Computational Biology and Computer Technology

#### **Commercialization of bioinformatics**

Innovation at the Macro-Level

Innovation at the Micro-level

(a) Human Genome Sciences.

Company Strategies

(a) Transformation of sequence data to gene information.

(b) The drug target as a commodity

(c) Gene expression information: Value added to gene sequence.

(d) Transformation of sequence data into information on human gene variation.

#### **The Pharmaceutical Industry and its Innovation Deficit**

#### **Challenges for Implementation of Bioinformatics**

Standardization for Data Integration

Data Mining: Converting Data into Information

#### **References**

Biological research has experienced a paradigm shift from *in vivo* or *in vitro* experimentation to *in silico* experimentation, a development that relies upon bioinformatics. The beginning of bioinformatics stems from the fortuitous timing of the adoption of new DNA sequencing methods and the availability of mini- and bench-top computers, which became the tools to store and to analyze the sequence data. Another fortunate coincidence was the popularization of the Internet, which provided a means to exchange sequence data and sequence analysis software, and the establishment of the Human Genome Project, which stimulated the need for sophisticated data management and analysis tools. Market pull has rapidly stimulated bioinformatics commercialization as pharmaceutical companies discovered a potential means to cure their innovation deficit. One of the early models for commercializing bioinformatics was simply to sell access to databases of human nucleotide sequences. This strategy is heading toward obsolescence as the public consortium nears its goal of sequencing the human genome. The key to future commercialization of sequence data will be to

develop informatics technology that transforms this data into information that is useful for diagnosis and therapy. A competitive transformation of sequence data into information will require improvements in data integration and data mining.

---

The modern biotechnology industry began in the late 1970s and early 1980s. At this time, the industry relied upon the early technology of molecular biology, which enabled the cloning and isolation of genes. These isolated genes provided a way to mass-produce the gene-encoded proteins, which were typically produced in small amounts by normal tissues. Many of these efforts were motivated by a known or likely utility of the proteins for therapy.

In the early 1990s, the first wave products of modern biotechnology were described as natural proteins and monoclonal antibodies to natural proteins (Eisner, 1991). Commentators speculated that second wave products would include nucleic acid molecules, carbohydrates, protein-based synthetic molecules, and modified cells for gene

therapy (Eisner, 1991; Hamilton, 1992). The impending major revolution driven by genomics and bioinformatics was still unseen.

“Genomics” and “bioinformatics” are terms that are not typically included in current dictionaries. Thomas Roderick coined the word “genomics” in 1986 to describe the discipline of mapping, sequencing, and analyzing genomes (Hieter and Boguski, 1997). Not incidentally, Roderick also wanted to provide a name for the new journal *Genomics*. Even today, the meaning of genomics is not very clear, but there is agreement that the term generally refers to the systematic use of genome information, in conjunction with new experimental data, to answer biological, medical, or industrial questions (Jordan, 1999).

The field of genomics relies upon bioinformatics, which is the management and analysis of biological information stored in databases (Durso, 1997). Bioinformatics developed after automated protein and DNA sequencing technologies were introduced around the mid-1970s, and after researchers started to use computers as central sequence repositories that were accessible remotely, a development that occurred in the mid- to late 1980s (Persidis, 1999). In the early 1990s, genomics was transformed from an academic undertaking to a significant commercial endeavor, a course followed by bioinformatics a few years later (Gershon, 1997).

The commercial value of bioinformatics continues to increase. In 1998, the bioinformatics market was worth an estimated US \$290 million, and the market may surpass the \$2 billion mark by 2005 (“Bioinformatics emerges”, 1999; “Oscar Gruss”, 2000). During March 2000, a rush to invest in bioinformatics-based companies was derailed by a plea from President Bill Clinton and UK Prime Minister Tony Blair for free access to human genome information, which the press misinterpreted as an intent to ban patents on human genes (Lewis, 2000; Licking et al. 2000; Rosenberg, 2000; Wadman, 2000). Nevertheless, interest in bioinformatics began to recover by the end of the month. This enduring attraction is due, at least in part, to the belief that bioinformatics can profoundly alter the way that drugs are developed (O’Brien, 2000).

The word “bioinformatics” is derived by combining biology and informatics. The linchpin of bioinformatics is that biological polymers, such as nucleic acid molecules and proteins, can be transformed into sequences of digital symbols. Moreover, only a limited alphabet is required to represent the nucleotide and amino acid monomers. It is the digital nature of this data that differentiates genetic data from many other types of biological data, and has allowed bioinformatics to flourish (Baldi and Brunak, 1998). Another key point is that the use of sequence data relies upon an underlying reductionist approach: sequence implies structure implies function (Murray-Rust, 1994). Consequently, sequence data can be treated as context-free,

because a prediction of the biological significance of a sequence can often be understood in isolation.

The performance of bioinformatics relies upon developments in computer hardware and software. However, it is the excessive amount of sequence data that has driven the development of bioinformatics, a circumstance that can be traced to the establishment of the Human Genome Project.

## Key factors for the development of bioinformatics

### Circumstance of Excessive Data

#### (a) Human genome project

“The Human Genome Project has been a technology-driven quest.” Dr. Michael W. Hunkapiller, senior vice president of Perkin-Elmer (“Perkin-Elmer”, 1998).

The Human Genome Project was conceived in the mid-1980s, and was widely discussed in the press and scientific community through the end of the 1980s (National Human Genome Research Institute, 1999). In the United States, the Human Genome Project officially started on October 1, 1990, as a 15-year program to map and sequence the complete set of human chromosomes, as well as those of several model organisms (Venter et al. 1998). The goal of sequencing an estimated three billion base pairs of the human genome was ambitious, considering that few laboratories in 1990 had sequenced even 100,000 nucleotides (National Human Genome Research Institute, 1998). Sequencing of the human genome began in earnest in 1996.

The U. S. Department of Energy and the National Institutes of Health were the main research agencies responsible for developing and planning the Human Genome Project (National Human Genome Research Institute, 1999). Other centers around the world soon joined in the project, including the Wellcome Trust (United Kingdom) (Cook-Deegan, 1994). By 1993, the Human Genome Project had become an established international effort, which included nine countries and the European Community (Cook-Deegan, 1994). Although any genome project center could determine a preferred method for generating sequencing data, all centers had to follow certain rules (Pennisi, 1998). The most significant rule was that the nucleotide sequences must contain no more than one error in every 10,000 bases, which represents an accuracy of 99.99%.

The strategy of this international project was to make a series of maps of each human chromosome at increasingly finer resolutions (DOE Human Genome Program, 1992). According to this approach, chromosomes were divided into smaller fragments that could be cloned, and then, fragments were arranged to correspond to their locations on

a chromosome. After mapping, each of the ordered fragments would be sequenced.

A number of major technical innovations were considered essential for the success of the Human Genome Project (Venter et al. 1998). The mapping and sequencing components of the Human Genome Project relied upon advances in technologies for constructing recombinant DNA libraries. The introduction of yeast artificial chromosomes greatly facilitated the construction of complete physical maps of complex genomes (Touchman and Green, 1998). The development of bacterial artificial chromosomes also provided a means to clone large fragments (about 150,000 base pairs). Complementary DNA (cDNA) library construction was aided by refined preparations of enzymes, which have allowed the generation of longer and more authentic primary cDNA (Touchman and Green, 1998).

However, the polymerase chain reaction, due to its sensitivity, specificity, and potential for automation, is considered the front-line analytical method for analyzing genomic DNA samples and constructing genetic maps (Touchman and Green, 1998). This method was invented in 1985, after initial discussions about the Human Genome Project. Over the years, incremental improvements in basic PCR technology have enhanced the power and practice of the technique.

Between 1970 and 1990, the number of nucleotides sequenced per year per investigator increased dramatically, reflecting the significant advancements in DNA sequencing technology (Cantor and Smith, 1999). Since the introduction of the first-semi-automated sequencer in 1987, and the development of Taq cycle sequencing in 1990, fluorescent labeling of DNA fragments generated by the Sanger dideoxy chain termination method has been the foundation of large-scale sequencing projects (Venter et al. 1998).

Technologies for capturing sequence information have also advanced. In the early 1980s, researchers could use digitizer pens to manually read DNA sequences from gels (Lewis, 1996). Then came image-capture devices, which were cameras that digitized the information on gels (Lewis, 1996). In 1987, Steven Krawetz (Center for Molecular Medicine and Genetics at Wayne State University in Detroit) helped to develop the first DNA sequencing software for automated film readers.

### EST approach: focus on active sequences

The prevailing view is that the bulk of useful information about the human genome can be gained from the regions of DNA that encode proteins. Analysis of these nucleotide sequences allows elucidation of the corresponding amino acid sequences. Although this seems simple, a significant problem is that gene density in the human genome is exceptionally low, and only about 3% of the genome

encodes proteins (Slater, 1998).

In the early 1990s, J. Craig Venter, a researcher at the National Institutes of Health, and his colleagues devised a new way to find genes (Wickelgren, 1999). Rather than taking the Human Genome Project strategy of sequencing chromosomal DNA one base at a time, Venter's group isolated messenger RNA molecules, copied these RNA molecules into DNA molecules, and then sequenced a part of the DNA molecules to create expressed sequence tags, or "ESTs." These ESTs could be used as handles to isolate the entire gene. Venter's method, therefore, focused on the "active" portion of the genome, which was producing messenger RNA for protein synthesis.

The EST approach has generated enormous databases of nucleotide sequences, and facilitated the construction of a preliminary transcript map of the human genome (Touchman and Green, 1998). The development of the EST technique is considered to have demonstrated the feasibility of high-throughput gene discovery, as well as provided a key impetus for the growth of the genomics industry (Fields, 1996).

As a result of the Human Genome Project and the parallel EST-based sequencing approaches, sequence data began to appear at an extraordinary rate. By mid-1999, the amount of GenBank nucleotide sequence data was doubling every 14 months, and a genetics laboratory could easily produce 100 gigabytes of data each day (Cook, 1999; "Drowning in data", 1999). As one commentator observed, "Biology has belatedly realised that it is, itself, an information technology" ("Drowning in data", 1999).

### Computational Biology and Computer Technology

"We are now witnessing two technology-driven revolutions that will transform our world within the next 10 to 20 years. The explosive growth in biotechnology is being paralleled almost precisely by the expansion of information technology. These areas come together in the concept of bioinformatics." (Murray-Rust, 1994).

Bioinformatics has its roots in computational biology, a field that has been driven during the last 17 years or so largely by the vast amounts of nucleotide sequence data generated and deposited in the public-domain databases (Durso, 1997). At present, computational-based analyses, storage, and retrieval of mapping and sequencing data are considered among the most critical and rapidly evolving genomic-based technologies (Touchman and Green, 1998).

The connection between molecular biology and computer science can be viewed as the outcome of coincidental timing (Smith, 1990). Departmental mini- and bench-top computers began to appear in laboratories at the same time that researchers were adopting techniques of cloning and nucleic acid sequencing. Thus, the tools needed to store,

search, and analyze the new sequence data developed alongside the tools necessary to generate the data.

After the formation of DNA and protein databases, software slowly became available to search sequence databases (Gardner, 1999). The first methods were simple and involved hunting for keyword matches and short sequence words. These approaches were followed by sophisticated pattern matching and alignment-based software. Suites of analysis algorithms were written by leading academic researchers at Stanford, CA, Cambridge, UK, and Madison, WI for in-house projects, and then became more widely available (Gardner, 1999). For example, in the late 1980s, the PC/GENE software package of IntelliGenetics (Mountain View, CA) made its appearance, enabling the researcher to translate a nucleotide sequence into an amino acid sequence, and to obtain basic protein structure predictions (Persidis, 1999). Today, genomics researchers rely upon software for a variety of activities, such as reading nucleotide sequences from electrophoresis gels, predicting encoded protein sequences, identifying primers for gene amplification, sequence comparison or alignment, database searching, analyzing evolutionary relatedness, pattern recognition, and structure prediction (Lewis, 1996; Smith, 1999).

Although conventional algorithms have been useful for analyzing biological information, these approaches are inadequate for many sequence analysis problems (Baldi and Brunak, 1998; Šali, 1999). This is due to the inherent complexity of biological systems and a lack of a comprehensive theory about molecular organization (Baldi and Brunak, 1998). Competent comparison of sequence patterns across species must take into account that biological sequences are inherently noisy, which reflects variability arising from random events amplified by evolution (Baldi and Brunak, 1998).

Machine learning approaches, such as neural networks, hidden Markov models, and belief networks, are suited for characterizing large amounts of data and noisy patterns in the absence of general theories (Baldi and Brunak, 1998; Cantor and Smith, 1999). The idea behind these approaches is to learn the theory automatically from the data through a process of inference, model fitting, or learning from examples (Baldi and Brunak, 1998; Cantor and Smith, 1999).

The advancement of bioinformatics generally, and the machine-learning expansion of bioinformatics in particular, has benefited greatly from progress in computer speed. Coincidentally, computer speed and the amount of sequence data have been growing at roughly the same rate since the late 1980s, apparently doubling about every 15 to 18 months by 1998 (Baldi and Brunak, 1998). Some claim that it is becoming increasingly difficult to separate advances in biotechnology from advances in high-performance computing (Van Brunt, 1999).

Access to sequence data is critical, and much of the new sequence data are distributed over the Internet. The Internet also provides a means to distribute software, and enables researchers to perform sophisticated analyses on remote servers. It was fortuitous, but fortunate, that the Human Genome Project and a major growth increase in Internet occurred in parallel (Touchman and Green, 1998).

Until the late 1980s, there were mainly three ways of accessing databases over Internet: electronic mail servers, File Transfer Protocol, and TELNET servers (Appel, 1997). Electronic mail servers allowed researchers to retrieve individual entries from databases by sending an electronic mail query to the mail server's Internet address. The researcher's query was then processed by the server, and the result was sent back to the sender's mailbox. Due to requirements of precise syntax, formulating the query was cumbersome and subject to frequent errors (Appel, 1997). Moreover, the process was slow by Internet standards, taking from minutes to hours. With File Transfer Protocol, entire databases could be downloaded and searched locally (Appel, 1997). A drawback, here, was that a researcher would have to download the database after each update. TELNET allowed a user to remotely log onto a computer and access its facilities. This method was useful for occasional queries, but required extensive management of user identifications, and often overloaded the remote computer's processing power (Appel, 1997).

In the early 1990's, the introduction of GOPHER and WAIS (Wide Area Information Server) increased the selection of database accession schemes. However, both protocols have been widely replaced by the World Wide Web, which Tim Berners-Lee (CERN; Geneva, Switzerland) invented in 1990 (Berners-Lee, 1999). Shortly after the National Center for Supercomputer Applications (Urbana-Champaign, IL) released the user-friendly Mosaic™ browser, it became clear that the World Wide Web would greatly enhance the power of cross-references by providing active integration of databases over Internet, thus eliminating the need to download and maintain local copies of databases (Murray-Rust, 1994; Appel, 1997). Thus, a researcher could easily navigate across database entries through active hypertext cross-references with the guarantee that each retrieved piece of information was up to date. ExPASy (Expert Protein Analysis System), the first molecular biology Web server, was set up at Geneva University Hospital and University of Geneva in 1993 (Appel, 1997). During the following months, most major genome databases were made accessible on World Wide Web servers throughout the world. Currently, there are at least 400 Internet-accessible databases of biological data (Discala et al. 1999).

### Commercialization of bioinformatics

“Private information is practically the source of every large modern fortune.” Sir Robert Chiltern, *An Ideal Husband* (Wilde, 1994).

“This is a quick and dirty grab – like the wild West, where everyone was trying to stake a claim.” (Fisher, 1994).

The Human Genome Project relies upon international cooperation and the sharing of knowledge. In this way, the rapidly growing data set of human nucleotide sequences reflects a macro-level of innovation at the international level. However, the transformation of that data into information is taking place at the national level, where governments are supporting commercialization of genomics and bioinformatics, and at the company level.

#### Innovation at the Macro-level

In Japan, for example, the government introduced a new program to launch 1,000 new biotechnology-related companies within the next decade (Saegusa, 1999a). This announcement represented a policy shift from academic research to commerce, and was consistent with the government’s program to produce a 25-fold expansion in Japan’s biotechnology market by 2010 (“Japan aims”, 1999; Saegusa, 1999a). The country’s commitment to the effort is clearly reflected in the 2000 budget, in which the Japanese government allocated US \$3.4 billion for biotechnology among the five science-related ministries; US \$561 million is set aside for genomics research alone (Triendl, 2000).

The first phase of the government program, which is backed by the Ministry of International Trade and Industry, aims to increase Japan’s competitiveness in the genomics field and to strengthen Japan’s intellectual property position (Saegusa, 1999a). One specific objective is to create vast databases of genomic information to provide data to Japanese research institutions and biotechnology companies for development of products and technologies (Saegusa, 1999a). According to the Ministry of International Trade and Industry, another objective is to sequence 30,000 cDNA clones by 2001 (Saegusa, 1999b). This project is led by Tokyo University’s Institute of Medical Sciences and Japanese companies.

Meanwhile, the Genome Sciences Centre (Wako City) will take the central role in human and animal research for the genomic databases (Saegusa, 1999a). Genome Sciences Centre was set up last year by the Institute of Physical and Chemical Research, which has developed a high-speed DNA sequencer. It is anticipated that the new government program will allow biotechnology companies to combine forces with electronics and multimedia industries (Saegusa, 1999a).

The products of the new bioinformatics initiative are already apparent. In July 1999, Helix Research Institute Inc. (Kisarazu-shi, Chiba), a genomics company funded by the government and industry, filed patent applications for more than 6,000 full-length human cDNA clones (Saegusa, 1999c). Over the next several years, Helix hopes to produce

an additional 20,000 full-length human cDNA clones in collaboration with other companies and research institutes, possibly through a consortium (Saegusa, 1999c).

Helix, which was established in 1996 as a joint venture between the Ministry of International Trade and Industry and ten private companies, consists of three research departments (Noguchi, 1996; Saegusa 1999c). The First Research Department combines established methods for identifying gene function with new core technologies, including the use of electro-optical devices for measurement of expression profiles, and high-throughput cloning of full-length DNA. The Second Research Department is responsible for bioinformatics and intends to have a computer system equipped with high-speed parallel processors, database servers, and graphics workstations. This department will also develop new software for the analysis of sequence data. The Third Research Department aims to develop methods to evaluate gene function and will analyze biological mechanisms through expression profiles. Thus, Helix includes in-house capabilities for gene sequencing, gene analysis, as well as hardware and software for bioinformatics.

China has also identified genomics as a major funding priority for biological and biomedical research (Hui, 2000; Triendl, 2000). New genome centers in Shanghai and Beijing receive funding from multiple sources, and are staffed with scientists from local hospitals and various institutes of the Chinese Academy of Sciences (Beijing).

The Canadian government is implementing a national genomics initiative with the objective of propelling the country into the role of a major player in genomics (“Genome Canada”, 1999; Hoyle, 1999). Canadian genomics researchers in Canada now have access to the Canadian Bioinformatics Resource (Halifax, Nova Scotia), the world’s first gigabyte network (Hoyle, 1999). The Canadian Bioinformatics Resource utilizes the national high speed CA\*net II internet and includes over 60 high performance servers and workstations at six National Research Council of Canada centers nation wide (Hoyle, 1999). The plan is to replace CA\*net II links of 45 megabytes per second with CA\*net III links of 300 gigabytes per second (Hoyle, 1999). When the system is fully implemented, Canadian Bioinformatics Resource should allow users to access and decipher raw data from over 100 databases in real time (Hoyle, 1999). Once again, bioinformatics and Internet access are developing in parallel.

Genome Canada’s directive is to coordinate Canadian genomics programs into a network of centers to provide the platform technologies and knowledge required for further research (“Genome Canada”, 1999; Hoyle, 1999). The network’s infrastructure is based on a hub and spoke model with five centers: Halifax, Montreal, Toronto, Saskatoon, and Vancouver (“Genome Canada”, 1999; Hoyle, 1999). The Canadian model of networked, geographically distinct

centers builds on the successes of the U.K., European, and Swiss bioinformatics efforts (Hoyle, 1999).

The Canadian genomics program centers will probably spur the creation of biotechnology clusters. The clustering phenomenon in the biotechnology industry is well-established, and is explicitly supported by certain governments (Sainsbury et al. 1999). The United States, often held as a model for biotechnology cluster development, contains a number of genomics clusters, such as the genomics cluster in Cambridge, Massachusetts. The magnet for activity in this particular area is the Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, which is one of the leaders in the Human Genome Project (Blanton, 1999). Not coincidentally, Cambridge also provides universities, entrepreneurs, medical research schools, and supporting biotech companies, as well as the attention of industry analysts (Blanton, 1999).

#### Innovation at the Micro-level

In addition to the phenomenon of cluster development, the U.S. bioinformatics industry is characterized by a dominant micro-level of innovation, as opposed to the macro-level innovation illustrated above. Yet Harry Mangalam, CEO of tag Informatics (Irvine, CA), observed that, “what has made US biotech and bioinformatics so vibrant was the ability of small companies to start up very easily to exploit an idea, and then, either grow or go bust based on their potential” (Hoyle, 1999). This validity of this comment is best illustrated by Human Genome Sciences, Inc. (HGS; Rockville, MD), the first company to commercialize genomics.

#### (a) Human Genome Sciences

In 1992, venture capitalist Wallace Steinberg gave US \$70 million, to be paid over ten years, to J. Craig Venter with the objective of creating The Institute of Genomic Research (TIGR; Gaithersburg, MD), a nonprofit organization, which would conduct basic genetic research (Marshall, 1997; Wickelgren, 1999). At the same time, Steinberg asked William A. Haseltine to leave the Dana-Farber Cancer Institute and Harvard Medical School (Boston) to become the chairman and CEO of HGS. (Pickering, 1999c; Wickelgren, 1999). Haseltine was a cancer researcher and entrepreneur, who had worked on the Human Genome Project (“Human Genome Sciences”, 1999).

Venter and Haseltine soon formed an alliance with SmithKline Beecham, which would invest an initial US \$125 million (Marshall, 1997). According to the terms of the collaboration, HGS would perform some research, develop medical products, and manage finances (Marshall, 1997). HGS would also transfer US \$85 million from SmithKline in quarterly payments to TIGR. In exchange, HGS would have the right to preview TIGR’s findings and HGS would have commercial rights to all discoveries

(Marshall, 1997). However, HGS grew increasingly concerned about TIGR’s plans to publish sequence data, and set up a duplicate intramural human gene sequencing program by 1993 (Marshall, 1997).

During 1995, HGS and TIGR signed a deal with Takeda Chemical Industries, Ltd. (Chuo-ku, Osaka, Japan) to develop and commercialize human genome products in Japan (“Human Genome Sciences”, 1999). The number of alliances increased during the following year when HGS and SmithKline enlarged the scope of their agreement to include Schering Plough Corp. (Madison, NJ), Synthelabo (now: Sanofi-Synthelabo; Paris, France), and Merck KGaA (Darmstadt, Germany) (“Enough to go around”, 1996). Despite these successes, the business relationship between HGS and TIGR grew increasingly distant, and in June 1997, HGS and TIGR announced the end of their partnership (Wickelgren, 1999).

In the early years of operation, HGS used automatic gene sequencing technology to compile one of the largest databases of human and microbial genes in the world (“Human Genome Sciences”, 1999). Today, HGS lists technologies that include gene isolation, sequencing, bioinformatics, molecular biology, protein chemistry, cell biology, pharmacology, high-throughput biological screening, drug formulation, and manufacturing (“Pipeline”, 2000).

Although an early strategy of HGS was to collect royalties from companies that would develop products using HGS’s database, HGS has taken the position that their goal is to discover, develop, manufacture, and market new gene- and protein-based drugs (“Corporate profile”, 1999; “Human Genome Sciences”, 1999). That is, HGS intends to support activities encompassing the discovery of new human genes and extending through human trials. In December 1997, HGS took a step to effect their long-term goals by entering into a lease with the Maryland Economic Development Corporation for a manufacturing and process development plant to mass produce its new first drugs (“Manufacturing”, 2000; Wickelgren, 1999).

In sum, bioinformatics is one HGS’s key technology platforms, and it is a major component of strategic alliances. HGS is considered to be the world’s first fully integrated bioinformatics company, where sequence, biological assay, or expression data are captured directly from production machines, and manipulated electronically (Pickering, 1999c). By October 1999, HGS had produced three candidates in Phase II trials, backing its long-term goal of moving from a genome sequencing company to a pharmaceutical company (“HGS touts”, 1999).

An important aspect of HGS strategy is the acquisition of proprietary rights to new genes and proteins. As of February 2000, HGS had filed patent applications covering more than 7,500 newly discovered human genes (“Patents”, 2000). According to a recent estimate, the HGS patent

estate may be worth in the billions (Pickering, 1999c).

The HGS patent portfolio is not appreciated by all, and highlights the continuing tension in the field between business and academic views of sequence information. In 1994, HGS and SmithKline made their database available to university researchers, in exchange for certain patent rights (“Enough to go around”, 1996). The terms of access excluded pharmaceutical and other industry-affiliated scientists, which many companies did not welcome (Hoke, 1994). Merck and Co., Inc. (Whitehouse Station, NJ) reacted by sponsoring a project to deposit sequence data into a publicly accessible database (Hoke, 1994). The competing project, conducted at Washington University School of Medicine (St. Louis, MO), had the aim of sequencing about 200,000 cDNA segments in about 18 months, and depositing the results immediately into GenBank, a publicly accessible database managed by the National Center for Biotechnology Information at the National Library of Medicine (Bethesda, MD) (Hoke, 1994). This duplicative effort created a more data to fuel the bioinformatics industry.

### Company Strategies

Before looking at bioinformatics-based strategies, some redefinition is in order. Reports on the genomics industry often refer to “bioinformatics technology”. However, bioinformatics is a discipline that encompasses the management and analysis of digitalized biological information. Bioinformatics is not one type of technology. Rather, bioinformatics includes any number of technologies. When considering commercialization strategies, therefore, it is more useful to identify configurations of bioinformatics technologies that comprise a company’s production system.

#### (a) Transformation of sequence data to gene information

For example, a company could use the following group of technologies, which provide a basic bioinformatics configuration: analysis of raw nucleotide sequence data to assemble small sets of sequences into large contiguous nucleotide sequences, structural analysis of the large sequences to identify the presence of a gene, analysis of the putative amino acid sequence of the gene to provide protein structure predictions, and analysis of the amino acid sequence and predicted protein structure to provide protein function predications. The net result of this set of unit processes is the productive transformation of raw sequence data to a gene sequence combined with annotation on protein predictions.

As HGS illustrates, one way that companies have used such a “basic bioinformatics configuration” was to sell data, in the form of complete genes or gene fragments, which others could use to identify potential drugs or drug targets. Incyte Pharmaceuticals, Inc. (Palo Alto, CA) was another of the

early biotechnology companies to engage in high-throughput computer-aided nucleotide sequencing to identify new genes and their corresponding proteins with potential therapeutic applications (“The history of Incyte”, 2000). The company’s basic approach was to compare partial sequences with known sequences to predict biological function, and to offer companies a non-exclusive access to its genomic databases (“Corporate backgrounder”, 2000; “The history of Incyte”, 2000).

Often, deals based upon the selling of nucleotide sequence information are structured like leases, in which the collaborator uses the technology and services for about three to five years in exchange for subscription fees and payments for services (Pickering, 1998). The most common element of such an arrangement is that equipment, software, and data are generally reclaimed by the owner upon termination of the agreement (Pickering, 1998). Bioinformatics services also usually came with milestones and royalties attached to projects that made it into actual drug programs (Longman, 1999).

This bioinformatics service model of the mid-1990s was attractive to investors, who viewed it as an alternative to the high cost and high risk of individual therapeutic programs (Longman, 1999). However, there are drawbacks to marketing only such services. Due to rapid innovation and rapid obsolescence, companies that rely solely upon selling nucleotide information must consistently upgrade their technologies to stay competitive (Ratner, 1999). In fact, the problems facing such companies more closely resemble those of the high-technology industry, because the end products of a bioinformatics service company are tools and information, and product development cycles are so compressed (Ratner, 1999).

There are additional drawbacks to the basic bioinformatics configuration. GenBank’s collection of human nucleotide sequences is currently experiencing an exponential growth. The sale of new nucleotide sequences is inherently limited by the number of human genes, and this natural limit constrains the time for the viability of the basic bioinformatics service model.

Considering the advancements in sequencing technology and the increasing numbers of contributions to public databases, it is clear that time is running out for the early strategy of selling sequence information. These sequences are generated by the publicly-financed consortium, which is financed principally by the NIH and the Wellcome Trust of London, and the duplicative efforts in industry. A major factor in the private sector is Celera Genomics Group (Rockville, MD). In 1998, instrument maker Perkin-Elmer (now, PE Corporation; Norwalk, CT), and J. Craig Venter of TIGR formed Celera to combine Perkin-Elmer’s DNA analysis technology with TIGR’s sequencing strategies (“Perkin-Elmer”, 1998; Wade, 1998). Venter, Celera’s head, had the goal to finish a complete sequence of a human genome within three years, a boast based upon

Celera's new sequencing strategy (Wade, 1998).

In contrast to the methodical, piecemeal approach of the Human Genome Project, Venter devised the "whole-genome shotgun strategy," which involves randomly breaking DNA into segments of various sizes and cloning the fragments into vectors (Marshall and Pennisi, 1998; Smaglik, 1998). Since the fragments are randomly cleaved from the genome, they tend to overlap, and a genome assembly program is used to fit contiguous pieces by matching overlapping ends (Wade, 1999). At a certain point, however, the assembly program cannot unambiguously match new ends, because the human genome contains numerous regions of repetitive DNA. As a result, it is difficult to determine how many repeated units reside within gaps between neighboring contiguous pieces. To bridge the gaps, Celera generates pieces of very large DNA fragments of known length with sequenced ends (Wade, 1999). The assembly program positions neighboring contiguous pieces by looking for a bridging link that has one end matching a DNA sequence in one contiguous fragment and the other end matching DNA in the other (Wade, 1999). A sufficient number of links is generated to ensure that each neighboring pair of contiguous fragments is straddled on average by at least two bridging links. Once the size of the gap is established, the assembly program can thread its way through the repetitive DNA between the two contiguous fragments (Wade, 1999).

In collaboration with Compaq Computer Corporation (Houston, Texas), Celera has developed what company officials refer to as the world's second largest supercomputer facility (Fickel, 2000). Celera has already installed more than 200 Compaq AlphaServer ES40 systems running 500MHz Alpha processors, 11 GS140 servers, and 50 terabytes of StorageWorks storage; this system runs on a network that supports throughput of 500GB per second (Fickel, 2000). The company has also leased 300 3700 DNA sequencers from PE Biosystems, and is reportedly working 24 hours a day and seven days a week (Fickel, 2000; Wade, 1999).

The effect of this effort was obvious in October 1999, when Celera announced that it had filed patent applications on 6,500 ESTs within one month, and in February 2000, when Celera announced that it had filed patent applications covering 10,000 genes (Gillis, 1999; "Company Says", 2000). Celera expects to have a rough draft of the entire human genome by midsummer of this year, a prediction that Celera backed up with the announcement that the company had completed the sequencing phase of one person's genome ("Celera Genomics", 2000; Licking et al. 2000).

Not to be outdone, the public consortium followed Celera's October announcement by publicizing the sequencing of its

one-billionth nucleotide, and the complete sequencing of human chromosome 22 (Butler, 1999; Dickson and Macilwain, 1999). In April 2000, the U.S. Department of Energy's Joint Genome Institute (Walnut Creek, CA) announced the decoding of human chromosomes 5, 16, and 19 in draft form (Osborne, 2000). More than ever, it is clear that the endgame is in sight. However, the collection and analysis of sequences to determine structure and predict function is just one part of bioinformatics. As one observer noted, "What started as a grab for gene sequences, of course, has turned into a race to find out what the genes do and which are the best targets for new drugs" ("Data mining", 1999).

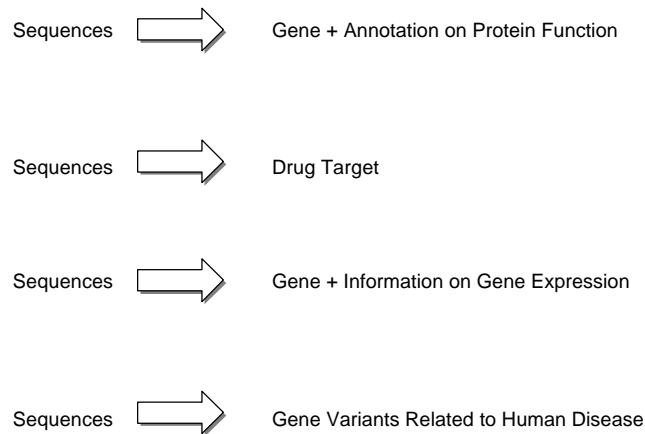
One sign of a maturing core technology is that the transformed products of that technology become relatively undifferentiated. In the case of the bioinformatics service model, the ready availability of sequences in the public domain has accelerated this maturation process. Recently, Hyseq Inc. (Sunnyvale, CA) created a subsidiary called "GeneSolutions Inc.," to sell Hyseq's genetic information over the Web (Henig, 1999). Lewis Gruber, Hyseq's CEO, has explained that the new company will allow researchers to purchase genes on a point-and-click basis (Regalado, 2000). GeneSolutions will charge from 50 cents for small bits of genetic data to US \$10,000 for exclusive one-year rights to patented genes (Henig, 1999). The price of nucleotide sequence information appears to be reduced to clear.

Pickering (1999a) observed that, in 1998, the bioinformatics industry passed beyond the proof-of-concept stage into a more mature business, which provides valuable products and services. However, the basic bioinformatics service model has reached its advanced years. As William Haseltine recently opined, "the bloom is already off the rose for database companies," (BioWorld®, 1999).

#### (b) The Drug Target as a Commodity

In light of the maturing bioinformatics service model, there has been an overhaul of the platform strategy of the basic bioinformatics service business. Figure 1 illustrates various tactics that companies use to commercialize bioinformatics. One approach is to modify the basic bioinformatics configuration to enhance the value-added transformation of sequence data by creating high-value intellectual property like validated targets (Longman, 1999). This strategy is illustrated by Millennium Pharmaceuticals Inc. (Cambridge, MA), which struck a US \$465 million deal to provide Bayer AG (Leverskusen, Germany) with 225 drug targets relevant to cardiovascular disease, cancer, osteoporosis, pain, liver fibrosis, hematology and viral infections (BioWorld®, 1999). Despite this lucrative arrangement, some believe that validated drug targeted discoverers have also entered into a race against commoditization (Longman, 1999).

### Bioinformatics Configurations



**Figure 1.** The figure illustrates various strategies that companies use to commercialize bioinformatics.

#### (c) Gene Expression Information: Value added to Gene Sequence

The availability of massive amounts of nucleotide sequences has motivated the development of various ways to examine this data, as reflected in the creation of functional genomics and pharmacogenomics technologies in the mid-1990s (Boguski, 1999). A systematic extrapolation from gene sequence to function is considered as the major challenge facing industry and academia (Rastan and Beeley, 1997). As a result of this shift in emphasis from sequencing and mapping genes to gene function, genome analysis is now considered to be divided into “structural genomics” and “functional genomics” (Hieter and Boguski, 1997). “Structural genomics,” a term coined in 1997, typically refers to the initial phase of genome analysis with the endpoint of constructing high-resolution genetic, physical, and transcript maps of an organism. (Hieter and Boguski, 1997; Gaasterland, 1998).

In contrast, functional genomics encompasses the development and application of genome-wide experimental approaches to assess gene function by using the information and reagents provided by structural genomics (Hieter and Boguski, 1997; Strausberg and Austin, 1999). The field is characterized by high throughput or large scale experimental methodologies combined with statistical and computer analysis of the results (Hieter and Boguski, 1997). The fundamental strategy in a functional genomics approach is to expand the scope of investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion. The emphasis is on gene function as gleaned from gene expression observations. Figure 2 presents the details of CuraGen’s (New Haven, CT) bioinformatics configuration for selling information that relates to gene expression. Gene Logic

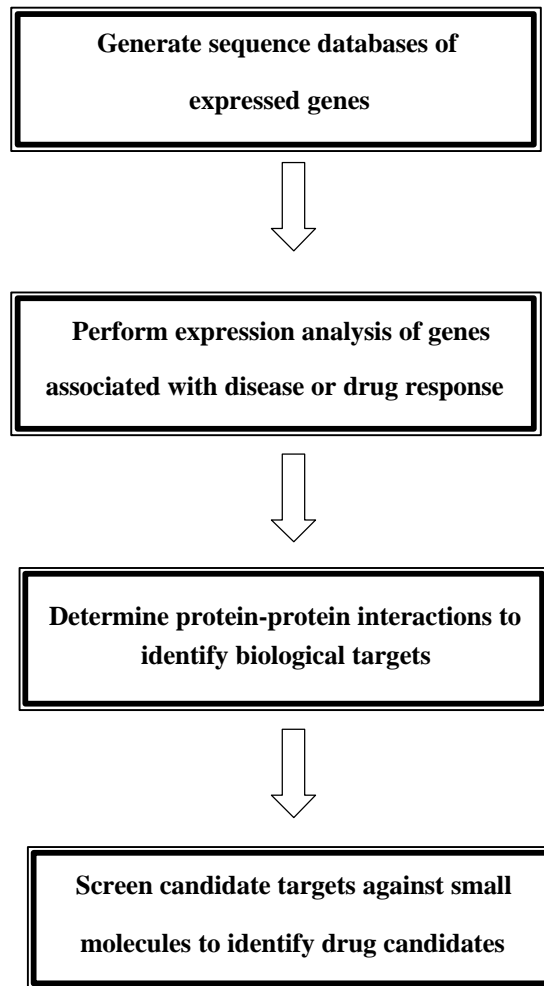
(Gaithersburg, MD) uses a similar approach to commercializing bioinformatics.

#### (d) Transformation of Sequence Data into Information on Human Gene Variation

“Pharmacogenomics” generally refers to a particular application of genomic technologies in drug discovery and development (Jain, 1999). Proponents of pharmacogenomics argue that knowledge about genetic differences, or polymorphisms, that contribute to variability in drug responses can be used to accelerate drug development and improve both the safety and efficacy of some currently available drugs (Schachter, 1998). This stand is based on the recognition that inherited differences in the metabolism and disposition of drugs, and genetic polymorphisms in the targets of drug therapy (such as receptors) can have an even greater influence on the efficacy and toxicity of medications than factors such as severity of disease being treated, drug interactions, patient’s age, nutritional status, renal and liver function (Evans and Relling, 1999). These polymorphisms can be characterized as base insertions, deletions, or substitutions (Czarnik, 1998). Single base differences between copies of the same gene are termed single nucleotide polymorphisms (“SNPs”), and these SNPs are the greatest source of variability within the population (Czarnik, 1998).

The field of pharmacogenomics has roots in the 1982 founding of the Human Polymorphism Study Center (Paris, France), which was the first academic institution to propose the systematic study of polymorphisms (“The SNP timeline”, 1999). The Center’s primary objective was to make an inventory of all possible human polymorphisms and associate them with diseases. However, the technology to support such a project would not appear for more than a decade

## CuraGen Strategy



**Figure 2.** The figure presents details of CuraGen’s (New Haven, CT) bioinformatics configuration for selling information that relates to gene expression (“Technology,” 1999).

In 1996, feasibility studies indicated that that mapping SNPs might be feasible and that the resulting data should be useful (“The SNP timeline”, 1999). During the following year, Abbot Laboratories (Abbott Park, IL) contracted with Genset (Paris, France) to begin a targeted search for 60,000 targeted SNPs that will be potentially useful in pharmacogenomics. When Celera formed in 1998, it included, as part of its business model, the selling of access to a human SNP database. In the same year, the National Institutes of Health began to fund research projects for searching and mapping 50,000 targeted SNPs and for the mapping of 50,000 randomly generated SNPs. The information will be entered into a public database (“The SNP timeline”, 1999).

Once again, patents have become an issue. In April 1999, Glaxo Wellcome, SmithKline Beecham, Pfizer, AstraZeneca, Bayer, Bristol-Meyers Squibb, F. Hoffman-La Roche, Hoechst Marion Roussel, Novartis, and Searle pooled their money to form a consortium with the objective of discovering 300,00 SNPs (Roberts, 1999; Russo and Smaglik, 1999). The consortium members will release the data to the public as a preemptive strike against patenting the SNPs.

According to simulations reported by Kruglyak (1999), one SNP per 6,000 base pairs may be needed to detect association between a marker and disease, which amounts to 500,000 for whole genome studies. A later study,

however, indicates that a SNP map with a density of one per 10,000 to 30,000 base pairs could be used to identify areas of DNA thought to be linked to disease (Loder, 1999). Glaxo Wellcome scientists have already used SNPs to locate areas of DNA for migraine, type II diabetes, and psoriasis. Therefore, researchers showed for the first time that it is possible to narrow the search for human genes using small variations in the genome. It is still uncertain, however, that SNP mapping can be applied to a wide range of diseases (Roberts, 2000).

Additional methods of commercializing bioinformatics are apparent. For example, certain bioinformatics-based companies are leveraging their technologies to become fully integrated drug discovery operations (Frew, 1998; Madden, 1998; Fisher, 1999; Ratner, 1999). This approach was illustrated by the evolution of Human Genome Sciences. Other bioinformatics companies are merging with drug discovery companies, resulting in a substitute technological approach to drug development.

Companies are also commercializing bioinformatics by using a hybrid database/application service provider model. Application service providers, or “ASPs,” can offer subscribers access to software programs, data storage, and computing facilities via the Internet. DoubleTwist.com (Oakland, CA), for example, provides access to clustering and alignment tools, as well as access to its own annotated genomic data. Similarly, LabOnWeb, a web site operated by Compugen, Inc. (Tel-Aviv, Israel), allows clients to analyze sequences using the company’s algorithms and propriety database. Both of these ASPs protect transferred information with encryption (“Data security”, 2000; “DoubleTwist Security Statement”, 2000).

### **The pharmaceutical industry and its innovation deficit**

“There’s very often a certain leap of faith, especially in the pharmaceutical industry, that if one can construct a database . . . it can be interrogated in a useful and productive way.” John N. Weinstein, senior research investigator in the Laboratory of Molecular Pharmacology at the National Cancer Institute (Durso, 1997).

“An in silico revolution is emerging that will alter the conduct of early drug development in the future.” Dale Johnson, Chiron Corporation (Johnson, 1999).

As discussed above, key factors that led to the development of bioinformatics included the excessive amounts of data generated by sequencing efforts, as well as advances in computational molecular biology, computer technologies, and the Internet. Another key factor was the availability of a market for bioinformatics information – the traditional

pharmaceutical industry. As Gardner (1999) observed, “in an industry worth comfortably over \$150 billion a year, any innovations that promise not only to find a new drug candidate more rapidly, but to revolutionise the way a pharmaceutical company fills its product pipeline for years to come, are compelling” (paragraph 2).

In the mid-1990s, pharmaceutical companies were primed for the new approaches of bioinformatics due to a lack of innovative new products in the traditional drug pipeline (Gershon, 1995; Gardner and Flores, 1999). This impending profit-gap is a particularly significant problem in view of the industry’s annual growth rates. Within the next decade, the leading pharmaceutical companies may need to bring to market ten times as many compounds per year as they currently manage just to maintain growth levels of 10 to 15%, a rate anticipated by investors (Purcell, 1998; Gardner, 1999).

Another disquieting development for the pharmaceutical industry is that a large number of blockbuster drugs will lose patent protection within the next few years (Purcell, 1998; Waldholtz et al. 1999). According to one estimate, drugs with sales approaching US \$25 billion in revenues will come off-patent by the year 2002 (Purcell, 1998). Consider Merck and Co. (Whitehouse Station, NJ) as an example. Within the next two years, Merck will lose U.S. patent protection for five major products, which brought the company US \$4.38 billion in U.S. sales and royalties during 1999 alone (Harris, 2000).

In light of these trends, the pharmaceutical industry needs an infusion of new blood to sustain earnings. Consequently, the industry is turning to bioinformatics-based approaches to shore up drug discovery programs. The adoption of bioinformatics may lead to the most comprehensive revolution of pharmaceutical research and development since the late 1930s (Gardner, 1999).

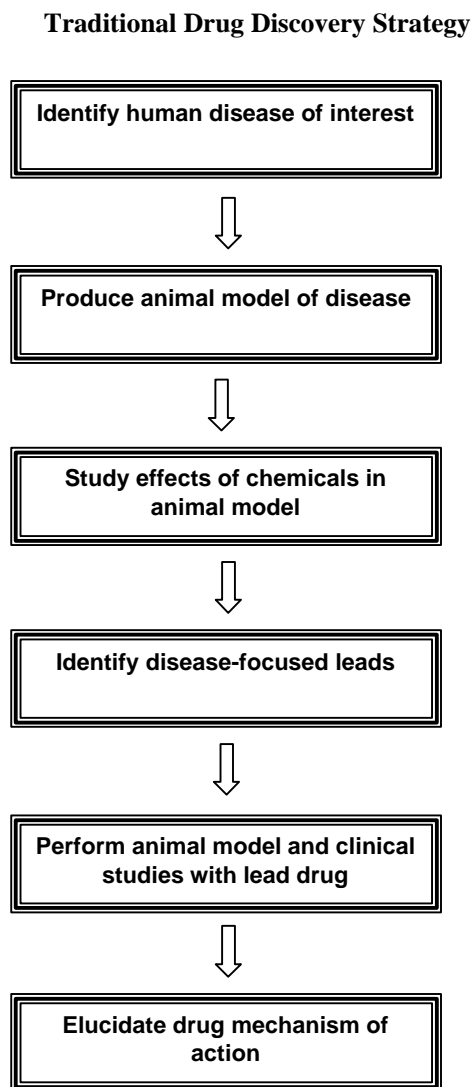
A successful integration of bioinformatics methods into drug discovery programs can improve target discovery and validation, and accelerate drug development by focusing research efforts on novel genomic targets (Bellavance et al. 1999). However, this does not represent a simple modification of traditional drug discovery.

There are two basic drug development approaches: a traditional drug screening approach and the newer “rational methods” (Gardner, 1999). Traditional drug development programs, which are organized around defined diseases, create an animal model of a disease as the first step in the drug discovery process (Crooke, 1998). This is illustrated in Figure 3. In this approach, the mechanism of drug action may not be elucidated until after a chemical had demonstrated therapeutic value in the clinic. In other words, the molecular target of the therapeutically useful drug may be identified last. In contrast, current drug discovery often focuses around molecular targets thought to be relevant to particular diseases, and mechanism of action

is evaluated very early in the discovery process (Crooke, 1998).

The adoption of a bioinformatics-based approach to drug discovery provides a further paradigm shift. With

bioinformatics, genotypes associated with pathophysiologic conditions may be defined first, and this will lead to the identification of potential molecular targets (Crooke, 1998). As shown in Figure 2, CuraGen is following this strategy.



**Figure 3.** The figure shows the strategy of a traditional drug development programs, as modified from Crooke (1998).

There is a widespread recognition by pharmaceutical companies of the need to better leverage information and informatics, because the path from gene to drug is neither simple nor quick (Larvol and Wilkerson, 1998). Some pharmaceutical companies have been expanding their bioinformatics groups, and adopting computer science technologies proven in other industries, such as financial trading (Gardner, 1999). However, the establishment of a new informatics framework for handling information requires both technology and a change in management's views. Some argue that the effective integration and use of

information will be the single biggest differentiator of pharmaceutical R&D competitive advantage in the next decade (Gardner, 1999).

One consequence of the rapid developments in drug discovery and development technologies has been a closing of the traditional technological advantages that one pharmaceutical company holds over another (Gardner, 1999). Pharmaceutical companies have made large investments in technologies such as genomics, microarrays, high throughput sequencing and combinatorial chemistry,

which has led to a fairly level playing field (Gardner, 1999). As a result, the pharmaceutical R&D bottleneck has moved from the creation of data about the activity of a small number of lead compounds to the manipulation and analysis of data to identify new targets.

Another consequence of the adoption of new technologies for discovery research is that, instead of working up to two to three targets slowly and in detail, a department may have as many as 20 to 30 targets to work on at once (Gardner, 1999). This increases the amount of communication required to perform effective research. Project managers who previously devoted weeks to single projects at a time must now be able multi-task those projects and be in a position to make stop/go decisions about projects and target molecules perhaps ten times as frequently as before (Gardner, 1999). As project timescales condense, the need for interdisciplinary teams to interact and share data increases, and a corporate culture of standardized data collection with quality assurance must be established (Gardner, 1999). To meet the challenges posed by pharmaceutical informatics in the next decade will require a major alteration of infrastructure and investment in people, networks, computers, and storage facilities.

## Challenges for implementation of bioinformatics

### Standardization for Data Integration

The commercialization of bioinformatics has spawned a support industry that provides bioinformatics tools, a function previously provided by university researchers. These companies are providing new data analysis tools and software platforms for data management, expression profile analysis, links to sequence and annotation databases, function prediction based on pathway information, data mining, and data tracking of automated processes (Pickering, 1999b). Companies that implement these new bioinformatics tools and software platforms are faced with problems that arise when trying to compare, store, and analyze data produced from multiple platforms (Pickering, 1999b; Ladd, 2000). Since there are no clearly accepted bioinformatics industry leaders, biotechnology and pharmaceutical companies are operating one or more outside systems with their own proprietary system to produce expression data (Pickering, 1999b). This situation leads to a practice of capturing only the lowest common denominator of data.

This problem is compounded in traditional pharmaceutical companies (Murray-Rust, 1994). The pharmaceutical industry has built large systems for corporate computing and traditional data processing, which reflect requirements imposed by the regulatory authorities. Security and auditing are two important features of these systems. In contrast, bioinformatics has been developed in an academic environment where the priorities are flexibility and responsiveness (Murray-Rust, 1994).

The development of individual molecular biology databases has generated a large variety of formats in their implementations, resulting in a situation compared with the tower of Babel (Benton, 1996; Frishman et al. 1998; Pennisi, 1999). The various Internet-based molecular biology databases have their own unique navigation tools and data storage formats, which make global searching difficult (Pennisi, 1999). Intercommunication between databases of different structure and format requires common semantic standards and controlled vocabulary in annotations that describe the sequences (Frishman et al. 1998). One of the key challenges in bioinformatics is to move through the transitional period of collections of incompatible components to integrated systems (Sobral, 1997; Jones and Franklin, 1998).

There are efforts to make the tools of bioinformatics as standardized as possible, similar to the development of standardized computer operating systems (Murray-Rust, 1994). One of the best-known examples of this approach is the BioStandards project of the European Bioinformatics Institute (Cambridge, UK). Funded jointly by the European Bioinformatics Institute, the European Commission, and several pharmaceutical companies, the project includes the development and adaptation of databases and software tools in terms of existing and emerging standards (Murray-Rust, 1994). As another example, Molecular Dynamics and Affymetrix have formed the Genetic Analysis Technology Consortium to attempt to standardize the growing field of microarray-based genetic analysis ("About the consortium", 1998). The consortium was created to provide a unified technology platform to design, process, read and analyze DNA arrays.

Thus, the integration of biological data will require some form of standardization. This is particularly important in view of the wealth of data available from Internet resources. There is a need for integrated access of information both within a company, and between companies and the Internet. Without a standard set by an innovative product or competitor, these standards will have to be set by cooperation between the industry, academia, and government agencies.

### Data Mining: Converting Data into Information

"Data is king. The long-term winners will generate it, interpret it, and apply it efficiently."  
Robert J. Olan, Chase Hambrecht and Quist analyst (Licking et al. 2000).

Regardless of who wins the race to sequence the human genome, the consequence of the race is more data. Sequence data combined with expression data of functional genomics is creating a bottleneck in the drug discovery process: data analysis. To place the problem in perspective, one study indicates that a typical high-throughput screening program in a large pharmaceutical company could have generated 200,000 data points per year in the early 1990s

(Drews, 1999). Five years later, a similar screening program generated five million to ten million data points, which may grow to 50 million by the end of the year (Drews, 1999). According to a recent estimate, one scientist can generate data in a few hours that might take months to analyze using conventional software (Brocklehurst et al. 1999).

Data mining encompasses the use of pattern recognition technologies and statistical techniques to examine large amounts of data (Wedin, 1999). The objective of data mining is to discover meaningful new correlations, patterns, and trends. Considering the bottleneck, data mining technology appears to represent the pacing technology of a company that uses bioinformatics for drug discovery.

The tools used for storage, retrieval, analysis, and dissemination of biological data are yet very similar to the original systems gathered together by researchers 15-20 years ago, and many of these are simple extensions of the original academic systems (Gardner, 1999). According to Gardner (1999), relational databases are still rare, and object-relational or fully object oriented systems are even rarer in mainstream applications.

This need is being addressed by the adoption of knowledge discovery approaches used for the business community, and by the development of new technology (Klevecz, 1999; Wedin, 1999; Persidis, 2000). An example of the latter is the development of visual data mining technology, which was strongly motivated by genome sequencing projects (Frishman et al. 1998). The development of pattern recognition tools is considered one of the fastest moving areas in bioinformatics, an opportunity reflected in the marketing of data mining software (Regalado, 1999). Pharmaceutical companies, which lack the necessary in-house expertise in informatics, are outsourcing informatics work as a means to speed genomics-based drug discovery and development (George, 1999; O'Neill, 1999).

According to Gardner and Flores (1999), the key differentiator of competitive advantage will shift from innovation in in vitro or in vivo biology to the exploitation of available information for rapid and accurate decision making. If so, then why would the pharmaceutical industry decide to buy informatics systems and services, rather than develop the technology in-house? Because now, there is a race to discover diagnostic and therapeutic uses for new nucleic acid molecules and proteins, and to acquire an intellectual property position on those new uses.

## References

About the consortium. (1998). [Online]. Available: <http://www.gatconsortium.org>.

Appel, R. D. (1997). *Interfacing and Integrating Databases*. In: *Proteome Research: New Frontiers in Functional*

*Genomics*, N. R. Wilkins (ed.), pp. 149-175. Springer-Verlag, Berlin.

Baldi, P. and Brunak, S. (1998). *Bioinformatics – the machine learning approach*. Cambridge, MA, The MIT Press.

Bellavance, L. L., Donlan, M.E. and Sharp, S. (1999). A bioinformatics primer. *Genetic Engineering News* 19:32-33.

Benton, D. (1996). Bioinformatics – principles and potential of a new multidisciplinary tool. *TIBTECH* 14:261-272.

Berners-Lee, T. (1999). *Weaving the web.*, HarperCollins Publishers Inc., New York, USA, 226 pp.

Bioinformatics emerges as a key technology for developing new drugs. (1999). *Chemical Market Reporter* 255:22.

BioWorld® Publishing Group. (1999). *BioWorld® 1999 genomics review: Companies leading the revolution.*, BioWorld® Publishing Group, Atlanta, GA, USA 110 pp.

Blanton, K. (1999, October 15). Gene pool. *The Boston Globe* H01.

Boguski, M. S. (1999). Biosequence exegesis. *Science* 286:453-455.

Brocklehurst, S. M., Hardman, C. M. and Johnston, S. J. T. (1999). Creating integrated computer systems for target discovery and drug discovery. In: *Pharmainformatics: A Trends Guide*, M. Owen (ed.), pp. 12-15. Elsevier Science Ltd., New York.

Butler, D. (1999). “Finishing” success marks major genome sequencing milestone . . . *Nature* 402:447-448.

Cantor, C. R. and Smith, C. L. (1999). *Genomics: The science and technology behind the Human Genome Project*. John Wiley and Sons, Inc., New York, USA, 596 pp.

Celera Genomics completes sequencing phase of the genome from one human being. (2000). [Online]. Available: <http://www.pecorporation.com> .

Company says it filed 10,000 gene patents. (2000). [Online]. Available: <http://dailynews.yahoo.com>.

Cook, J. L. (1999). Internet biomolecular resources. *Analytical Biochemistry* 268:165-172.

Cook-Deegan, R. (1994). *The Gene Wars*. W. W. Norton and Company, Inc., New York, USA, 416 pp.

Corporate backgrounder. (2000). [Online]. Available: <http://www.incyte.com>.

**Jones, P. B.**

- Corporate profile. (1999). [Online]. Available: <http://www.hgsi.com/cprofile/index.html>.
- Crooke, S. T. (1998). Optimizing the impact of genomics on drug discovery and development. *Nature Biotechnology* 16:29-30.
- Czarnik, A. W. (1998). Illuminating the SNP genomic code. *Modern Drug Discovery* 1:49-55.
- Data mining. (1999). Start-up 4:7-13.
- Data security. (2000). [Online]. Available: <http://www.labonweb.com>.
- Department of Energy (DOE) Human Genome Program. (1992). Primer on Molecular Genetics. Oakridge TN, Human Genome Management Information System.
- Dickson, D. and Macilwain, C. (1999). "It's a G": The one-billionth nucleotide. *Nature* 402:331.
- Discala, C., Ninnin, M., Achard, F., Barillot, E. and Vaysseix, G. (1999). DBcat: A catalog of biological databases. *Nucleic Acids Research* 27:10-11.
- Drews, J. (1999). Informatics: Coming to grips with complexity. In: *Pharmainformatics: A Trends Guide*, M. Owen (ed.), pp.1-2. Elsevier Science Ltd. New York.
- DoubleTwist Security Statement. (2000). [Online]. Available: <http://www.doubletwist.com>.
- Drowning in data. (1999). *The Economist* 351:93-94.
- Durso, T. W. (1997). As genomics grows, future for bioinformatics is bright. *The Scientist* 11:13.
- Eisner, R. (1991). Second wave of biotechnology revolution to crest with a new generation of drugs. *The Scientist* 5:1,8-9.
- Enough to go around. (1996). *The Scientist* 10:30.
- Evans, W. E. and Relling, M. V. (1999). Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* 286:487-491.
- Fickel, L. (2000). Writing the book of life. *CIO Magazine*. [Online]. Available: <http://www2.cio.com/> Corresponding author e-mail: RiceKid@ix.netcom.com
- Fields, C. (1996). Informatics for ubiquitous sequencing. *Trends in Biotechnology* 14:286-289.
- Fisher, L. M. (1999, August 26). The race to cash in on the genetic code. [Online]. Available: <http://www.nytimes.com>.
- Fisher, L. M. (1994). Profits and ethics clash in research on genetic coding. [Online]. Available: <http://www.nytimes.com>.
- Frew, J. (1998). Biotechnology's metamorphosis into a drug discovery industry. *Nature Biotechnology* 16:22-24.
- Frishman, D., Heumann, K., Lesk, A. and Mewes, H.W. (1998). Comprehensive, comprehensible, distributed and intelligent databases: Current status. *Bioinformatics* 14:551-561.
- Gaasterland, T. (1998). Structural genomics: Bioinformatics in the driver's seat. *Nature Biotechnology* 16:625-627.
- Gardner, S. (1999). The evolution of bioinformatics. *BITS Journal*. [Online]. Available: <http://www.bitsjournal.com>. Corresponding author e-mail: [stevegardner@csi.com](mailto:stevegardner@csi.com).
- Gardner, S. P. and Flores, T. P. (1999). Integrating information technology with pharmaceutical discovery and development. In: *Pharmainformatics: A Trends Guide*. M. Owen (ed.), pp.2-5. Elsevier Science Ltd., New York.
- Genome Canada: Planning session. (1999). [Online]. Available: <http://www.genomecanada.ca>.
- George, D. (1999). Bioinformatics outsourcing: The gains of letting go. *Drug Discovery Online*. [Online]. Available: <http://www.drugdiscoveryonline.com>.
- Gershon, D. (1997). Bioinformatics in a post-genomics age. *Nature* 389:417-418.
- Gershon, D. (1995). The boom in bioinformatics. *Nature* 375:262.
- Gillis, J. (1999, October 26). Md. gene researcher draws fire on filings. *The Washington Post*, E01.
- Hamilton, J. O'C. (1992, March 2). Biotech: America's dream machine. *Business Week* 3254:66-69, 73-74.
- Harris, G. (2000, February 9). How Merck plans to cope with patent expirations. *The Wall Street Journal*, Eastern Edition A1.
- Henig, P. D. (1999). Rent-A-Gene.com. *Red Herring* 73:156-168.
- HGS touts July 2001 Independence Day: All collaborations will expire. (1999). *The Pink Sheet* 61:24.
- Hieter, P. and Boguski, M. (1997). Functional Genomics: It's all how you read it. *Science* 278:601-602.
- Hoke, F. (1994). Limited access to cDNA database has drug manufacturer up in arms. *The Scientist* 8:1.

- Hoyle, B. (1999). Giga-speed bioinformatics to power Genome Canada. *Nature Biotechnology* 17:950. <http://www.redherring.com>. Corresponding author e-mail: [edit@herring.com](mailto:edit@herring.com)
- Hui, L. (2000). Money and machines fuel China's push in sequencing. *Science* 288:795-798. Manufacturing. (2000). [Online]. Available: <http://www.hgsi.com>.
- Human Genome Sciences, Inc. (1999). [Online]. Available: <http://www.hoovers.com>
- Jain, K. K. (1999). Strategies and technologies in functional genomics. *Drug Discovery Today* 4:50-53. Marshall E. and Pennisi, E. (1998). Hubris and the human genome. *Science* 280:994-995.
- Japan aims to launch 1,000 biotech companies in 10 years. (1999). *Nature* 397:554. Marshall, E. (1997). Genomic's odd couple. *Science* 275:778-780.
- Johnson, D. (1999). The discovery-development interface has become the new interfacial phenomenon. *Drug Discovery Today* 4:535-536. Murray-Rust, P. (1994). Bioinformatics and drug discovery. *Current Opinion in Biotechnology* 5:648-653.
- Jones, S. B. and Franklin, J. (1998). A strategy and demonstration for integrated biotechnology information. *Disease Markers* 13:237-243. National Human Genome Research Institute. (1999). [Online]. Available: <http://www.nhgri.nih.gov>.
- Jordan, B. R. (1999). "Genomics": Buzzword or reality? *Journal of Biomedical Science* 6:145-150. National Human Genome Research Institute. (1998). Reading the sequence of human DNA. [Online]. Available: <http://www.nhgri.nih.gov:80/NEWS>.
- Klevecz, R. (1999). The whole EST catalog. *The Scientist* 13:22-23. Noguchi, T. (1996). Turning towards applied research in genetic information: Helix Research Institute begins research. *Kazusa Special Edition*. [Online]. Available: <http://www.pref.vhiba.jp/Business/Kazusa/Hotline/Vol.16/special-e.html>.
- Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 22:139-144. O'Brien, S. (2000). DNA adds drama to biotech industry. *CBS MarketWatch*. [Online]. Available: <http://app.marketwatch.com>.
- Ladd, B. (2000). Intuitive data analysis: The next generation. *Modern Drug Discovery* 3:46-52. O'Neill, M. D. (1999). BioLIMS™ Information management system is crucial to target ID at Chugai Biopharmaceuticals. *BioBeatSM*. [Online]. Available: <http://www.biobeat.com>. Corresponding author e-mail: [oneillmd@pebio.com](mailto:oneillmd@pebio.com).
- Larvol, B. L. and Wilkerson, L. J. (1998). In silico drug discovery: Tools for bridging the NCE gap. *Nature Biotechnology* 16:33-34. Osborne, R. (2000). Researchers decode 3 more chromosomes: 5, 16 and 19. *BioWorld@Today* 11:1-2.
- Lewis, R. (2000). Clinton, Blair stoke debate on gene data. *The Scientist* 14:1,18,21. Oscar Gruss issues bioinformatics report. (2000). [Online]. Available: <http://www.prnewswire.com>.
- Lewis, R. (1996). Software helps researchers in sorting through the human genome. *The Scientist* 10:18-19. Patents. (2000). [Online]. Available: <http://www.hgsi.com>.
- Licking, E., Barret, A., Rae-Dupree, J., Carey, J. and Capell, K. (2000). Move over, dot-coms. Biotech is back. *Business Week* 3671:116-122. Pennisi, E. (1999). Seeking common language in a tower of Babel. *Science* 286:449.
- Loder, N. (1999). Genetic variations can point the way to disease genes. *Nature* 401:734. Pennisi, E. (1998). DNA sequencers' trial by fire. *Science* 280:814-817.
- Longman, R. (1999). Racing biotech's commoditization clock. *In Vivo* 17:46-53. Perkin-Elmer, Dr. J. Craig Venter, and TIGR announce formation of new genomics company. (1998). [Online]. Available: <http://www.pe-corp.com/press/prc5447.html>
- Madden, A. P. (1998). I am pharma, hear me roar. *Red Herring Online*. [Online]. Available: Persidis, A. (2000). Data mining in biotechnology. *Nature Biotechnology* 18:237.

**Jones, P. B.**

- Persidis, A. (1999). Bioinformatics. *Nature Biotechnology* 17:828-830.
- Pickering, L. (1999a). Bioinformatics comes of age. *Genetic Engineering News* 19:10, 30, 40.
- Pickering, L., (1999b). Gene expression analysis. *Genetic Engineering News* 19:1, 18, 54, 64.
- Pickering, L. (1999c). Fully integrating bioinformatics. *Genetic Engineering News* 19:21, 42, 52-54, 56.
- Pickering, L. (1998). Anatomy of a bioinformatics deal. *Genetic Engineering News* 18:12, 37, 44, 48.
- Pipeline: Functional Genomics. (2000). [Online]. Available: <http://www.hgsi.com/pipeline/index.html> .
- Purcell, D. J. (1998). Navigating biotechnology's new fiscal opportunities. *Nature Biotechnology* 16:51-53.
- Rastan, S. and Beeley, L. J. (1997). Functional genomics: Going forwards from the databases. *Current Opinion in Genetics and Development* 7:777-783.
- Ratner, M. (1999). Should platform companies move toward products? *Nature Biotechnology* 17:16-17.
- Regalado, A. (2000). G-Commerce. [Online]. Available: <http://www.techreview.com> Corresponding author: [regalado@mit.edu](mailto:regalado@mit.edu) .
- Regalado, A. (1999). Mining the Genome. *Technology Review*. [Online]. Available: <http://www.techreview.com/articles/oct99/regalado.htm>. Corresponding author e-mail: [regalado@mit.edu](mailto:regalado@mit.edu).
- Roberts, L. (2000). SNP mappers confront reality and find it daunting. *Science* 287:1898-1899.
- Roberts, L. (1999, October 18). A short cut to the gene pool. [Online]. Available: <http://www.usnews.com/usnews/home.htm>
- Rosenberg, R. (2000, February 18). On Wall Street, a buzz over biotech. *Boston Globe* C01.
- Russo, E. and Smaglik, P. (1999). Big Pharma hedges its bets. *The Scientist* 13:1,8,32.
- Saegusa, A. (1999a). Japan pushes to capitalize on biotechnology. *Nature Biotechnology* 17:320-321.
- Saegusa, A. (1999b). Japan declares five-year plan to double genome research funds. *Nature* 400:389.
- Saegusa, A. (1999c). Japan joins effort to patent cDNA clones. *Nature* 401:520.
- Sainsbury, D. (1999). *Biotechnology clusters*. London, Department of Trade and Industry.
- Šali, A. (1999). Functional links between proteins. *Nature* 402:23-26.
- Schachter, B. (1998). *Pharming the genome*. H.M.S. Beagle. Issue 41. [Online]. Available: <http://www.biomednet.com/hmsbeagle>. Corresponding author e-mail: [drbethie@walrus.com](mailto:drbethie@walrus.com).
- Slater, G. St. C. (1998). Human EST sequences. In: *Guide to Human Genome Computing*, M. J. Bishop (ed.), pp. 205-214. Academic Press, San Diego, CA.
- Smaglik, P. (1998). Private genome sequencing effort may hasten separate public venture. *The Scientist* 12:1-5.
- Smith, C. (1999). Computational gold. *The Scientist* 13:21-23.
- Smith, T. F. (1990). The history of the genetic sequence databases. *Genomics* 6:701-707.
- Sobral, B. W. S. (1997). Common language of bioinformatics. *Nature* 389:418.
- Strausberg, R. L. and Austin, M. J. F. (1999). Functional genomics: Technological challenges and opportunities. *Physiological Genomics* 1:25-32.
- Technology. (1999). [Online]. Available: <http://www.curagen.com>.
- The history of Incyte. (2000). [Online]. Available: <http://www.incyte.com>.
- The SNP timeline. (1999). *The Scientist* 13:9.
- Touchman, J. W. and Green, E. D. (1998). The genomic revolution: Enabling technologies of the Human Genome Project. *Focus* 20:58-61.
- Triendl, R. (2000). Genomic forges ahead in East Asia. *Nature Biotechnology* 18:278-279.
- Van Brunt, J. (1999). Restless in Seattle. *Signals*. [Online]. Available: <http://www.signalsmag.com> . Corresponding author e-mail: [signals\\_edit@recap.com](mailto:signals_edit@recap.com)
- Venter, J. C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* 280:1540-1542.
- Wade, N. (1999, October 26). Rivals reach milestones in genome race. [Online]. Available: <http://www.nytimes.com>.
- Wade, N. (1998, August 18). New company joins race to

sequence human genome. [Online]. Available: <http://www.nytimes.com>.

Wadman, M. (2000). Balance is returning after US biotechnology shares scare. *Nature* 404:424.

Waldholtz, M., Tanouye, E. and Harris, G. (1999 November 4). With executives aging and patents expiring, industry is ripe for megamergers. *The Wall Street Journal* B1, B4.

Wedin, R. (1999). Visual data mining speeds drug discovery. *Modern Drug Discovery* 2:39-47.

Wickelgren, I. (1999). Mining the genome for drugs. *Science* 285:998-1001.

Wilde, O. (1994). An ideal husband. In: *The Complete Works of Oscar Wilde*. New York, Barnes and Noble Books. pp. 482-551