Nanoinformatics:
Emerging Computational Tools in Nano-scale Research

K. Ruping* and B.W. Sherman**

* Massachusetts Institute of Technology, Cambridge, MA, USA, ruping@mit.edu
** Massachusetts Institute of Technology, Cambridge, MA, USA, woody420@mit.edu

ABSTRACT

Over the past several years basic research across several disciplines has revealed the promise of nanotechnology as an applied science in a wide range of industries. This progress has attracted the attention of government granting authorities, commercial researchers, and venture capital investors. Increased funding for nano-scale research initiatives has led to a growing demand for specialized human resource skills, research tools, and laboratory infrastructure.

This paper looks to a new area of computational tools that has emerged to meet the needs of the nano-scale research lab, and predicts the development of the field of nanoinformatics that will enable new advances and new applications in nanotechnology. We start with the historical roots of computer science and informatics. Next we look to the development of informatics into particular areas of science and commercial industries. Finally we document the growing use of computer science, database management tools, and information technology architectures in nano-scale research.

Keywords: informatics, computer science, nanoinformatics.

1. THE ORIGINS OF INFORMATICS

The term "informatics" has taken on various meanings, depending on the time period, geographical location, and field of use. For example, in Russia and parts of Europe the term informatics is closely related with computer science as a whole, and computer programming in particular. [8] Here we adopt a narrower definition of informatics to mean the science of applying computational theories and tools to the gathering, storage, manipulation, and interpretation of digital information. Such a definition still closely couples computer science with informatics, and envisions various applications of computational tools across different industries.

The history of computer science and informatics dates back to the 1830s when Charles Babbage, a Cambridge mathematician, conceived the first mechanical computing machine called a Differential Engine. Having found part of this never-completed device, Howard Aiken built the first electronic computer, the Harvard Mark I in 1944. However, applications for this new class of "calculators" were limited by instruction sets that were either hard-wired or manually inputted into the computational device through an I/O peripheral such as punch cards. This meant that any reprogramming required substantial human effort to re-wire the device or manually input new instructions, both of which were fraught with human error as well as time delays.

The advent of informatics came one year later when John von Neumann proposed that the programming code be incorporated into the electronic data resident on memory.[1, 3] The von Neumann architecture provided a level of abstraction away from the physical hardware, transforming the mundane task of programming into a field of engineering. With computer instructions liberated from the unchanging hardware elements and the continually changing data inputs, informatics emerged as an exciting area of research across university classrooms and industry labs. Government grant organizations, however, did not quickly recognize the strategic importance that informatics would have on other technical fields.

Informatics applications started with the first commercial computer. In 1951 the Census Bureau secured a UNIVAC from the Eckert-Mauchley Division of Remington Rand to tabulate census data. Three years later General Electric and the Metropolitan Life Company were using their UNIVACs to process employee payroll checks and perform other business processing needs. [6, 7] Soon IBM was offering a competing stored memory computers for computational science applications as well as for data management functions. Suddenly a growing number companies such as General Electric, RCA and Honeywell were entering the computer market with large mainframe computers, but with limited software flexibility.

Moving into the Cold War period, government support for emerging technologies became a strategic interest of several granting agencies. The momentum behind informatics as a discipline was not to be overlooked by the early 1960s, and computer science as a whole was increasingly viewed as a strategic differentiator in both economic and military competitive analysis. The first grant institution that

included informatics research in its administrative scope was the Information Processing Techniques Office (IPTO), established in 1962 under the Defense Advanced Research Projects Agency (DARPA) and headed by J.C.R. Licklider. In 1967 the National Science Foundation (NSF) united its disparate computer science activities into a single office, the Office of Computing Activities (OCA). The funding strategy was focused on facilities, however, with $11.3 million of the office's $12.8 million total budget going toward institutional support in its first year of operation. [5]

In these early years computing power was seen to come from hardware advances alone. Software typically came bundled with the hardware, with little chance for third-party vendors to develop and sell independent products or extensions to existing software applications. That changed in 1969 when IBM decided to unbundled the pricing of its software in light of antitrust pressures. Into the 1970s a new generation of independent software emerged which enabled the extension of informatics to fields beyond the traditional mathematical, scientific and business processing applications.

2. INFORMATICS GROWS UP

As memory capacity expanded and processor speeds raced forward in what was to become Moore's Law, computer science engineers applied these capabilities to the sophisticated management and manipulation of large databases. These tools became attractive to a group of biologist working on genomic research. The foundation of bioinformatics was set.

2.1 Bioinformatics

Bioinformatics was the result of both the growing capabilities of the hardware systems supporting computational tasks, the increasingly sophisticated software that enabled complex data management, and the needs of research scientists who had growing economic resources with which to either design or purchase high-end informatics software. This new field evolved from the biochemical advances fueling the molecular revolution in biology across the 1970s and 1980s, such as gel electrophoresis, amino acid sequencing, polymerase chain reaction, and gene mapping techniques. [12] Bioinformatics became an industry when government research grants coincided with corporate R&D and venture capital investment into biotechnology. These origins closely map the current state of computational science in nanotechnology, which we will return to shortly.

Throughout the 1960s government support of basic research was broken down into the traditional academic categories, including biology but excluding computer science. In the 1970s new NSF funding was introduced to molecular biology projects. As research programs were formalized, it soon became apparent that new tools were needed to handle data and to provide high-throughput computation of biological data.

Bioinformatics started with government interest in sequencing the human genome, first articulated in 1984 and institutionalized with the Human Genome Project in 1990. The objective was an informatics challenge: to identify the approximately 30,000 genes in human DNA, determine the sequences of the 3 billion chemical base pairs that comprise human DNA, and manage this information in publicly accessible databases. Since the inception of this program, the amount of DNA sequence data has grown exponentially to nearly cover the complete DNA sequence of the human genome. The funding of this project and its related computational demands on the researchers facilitated the development of specialized informatics tools.

As a result, academic research facilities attracted a new generation of multidisciplinary researchers familiar with biology, statistical modeling, and computer science. Advances in statistical methods of data analysis and progress in computational technology enabled a growing community of specialists to effectively deal with the explosion of sequence data. They developed a multidisciplinary research area aimed at organizing, classifying, and parsing the immense richness of sequence data. Soon these tools were to become products for sale to others as the bioinformatics community turned into an industry.

2.2 Cheminformatics

Other fields have emerged from the interface between science and computational engineering. Similar to early genomics, chemistry researchers are now challenged by the complex data structures and the computation power needed to model reactions. Combinatorial chemistry is an increasingly common field of chemistry research where a set of different compounds are reacted in combination with each other to form libraries of resulting substances and their characteristics. [11] High throughput chemical analysis demands process control, data capture, and information processing. Chemical analysis, test equipment, and informational databases are leading the way to the new field of cheminformatics -- the organization of chemical data in a logical form to facilitate the process of understanding and making inferences. [9]

With the vast space of chemical compounds, it is essential to have computational methods and informatics tools to organize this information in a manner that makes research advances possible at an appreciable rate. The value of this organization can be seen in the pride and secrecy that pharmaceutical companies hold related to their chemical databases. These databases typically range between 100,000 and 1 million compounds and obviously require

sophisticated data management and computational support. Even with such volumes today, these databases must grow by orders of magnitude to begin to span the chemical specie needed to find optimal drug candidates.

3. NANOINFORMATICS

The early development of nanotechnology is even more controversial than the origins of computer science and bioinformatics, in part because nanotechnology cuts across multiple scientific disciplines. The origins of nanotechnology lie in Richard Feynman's 1959 speech entitled "There's Plenty of Room at the Bottom." [2] In this presentation to the annual meeting of the American Physical Society, Feynman proposed the possibility of manipulating matter at the atomic level. Scientists had few tools with which to enable nanotechnology research, and most of the work at this scale was either theoretical or limited to chemistry.

Some of the most important early advances in nanotechnology have focused on innovative research equipment and new experimental methods applying such equipment. For example, Gerd Binnig and Heinrich Rohrer invented the Scanning Tunneling Microscope (STM) at IBM in 1981, for which they were awarded the Nobel Prize in Physics eight years later. Due to the physical scale of nanotechnology, much of the research equipment performing nano-scale experimentation is computer-driven. These tools have internal operating systems, computational software applications, and data management tools. As such the increasingly complex computer-driven research equipment, producing an ever increasing volume of digital output, were prime candidates for the next generation of informatics, which we call nanoinformatics.

There are two basic characteristics of nanotechnology research, also found in the development of bioinformatics, that has led to the emerging field of nanoinformatics. First is the deluge of data that computer-driven tools generate, particularly in nano-scale experimentation that incorporates a growing set of variables from across an increasing base of scientific knowledge. Second, the need for control -- of the systems themselves and of elements that the researcher hopes to manipulate -- increases in complexity as does the refinement of the research equipment itself. Both data management and system control come together in experimental tasks or industrial processes at a size-scale and time-scale that requires delicate sensing, massive data, complex calculation, and precision movement to translate virtual modeling into actual mechanical movement.

We are at the early stages in the emergence of a nanoinformatics science, and still earlier in the development of a nanoinformatics industry. Progress has been rapid. Early research tools required physical media, such as floppy disks or CDs, to transfer data between equipment. Data capture and transfer was overcome by today's second generation of equipment networked across the nano-scale research lab. Equipment that was once reliant on an internal system for memory, computational power, and software tools now come with an Ethernet card and a communications interface. More problematic is the growing number of data formats and the lack of interoperability across the laboratory.

4. CONCLUSION

Similar to the liberation of software tools from hardware in the 1970's, the decoupling of laboratory equipment from informatics tools is enabling more dynamic research as well as increased knowledge of the results of such laboratory equipment. We are now at the inflection point of a new era of research management and information tools that will take advantage of computation power at a time when nano-scale data needs are expanding in both scale and complexity.

We believe progress in the application of informatics solutions to nano-scale research challenges will continue. The next generation nanoinformatics tools will focus on computational cooperation and systems integration, at which time the deluge of data will be translated into a windfall of knowledge. These advances will have to move beyond data sharing to task sharing, when a nanotechnology lab experiment touches on the computing power of a set of equipment behind a common user interface. Future progress in this field will move toward intelligent automation, complex data mining, and intuitive visualization of results.

REFERENCES

[1] J. von Neumann, "First Draft of a Report on the EDVAC," 1945.

[2] R. Feynman, "There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics," December 29, 1959, printed in Engineering and Science, February 23, 1960.

[3] W. F. Aspray, "Pioneer Day `82: History of the Stored Program Concept," Ann. Hist. Comp., Vol. 4, No. 4,1982.

[4] History of the DOE Human Genome Program, http://www.er.doe.gov/production/ober/history.html.

[5] A.L. Norberg and J.E. O'Neill, "A History of the Information Processing Techniques Office of the Defense Advanced Research Projects Agency," Charles Babbage Institute, 1992

[6] Paul Ceruzzi, "A History of Modern Computing," 1998.

[7] M. Campbell-Kelly & W. Aspray, "Computer: A History of the Information Machine," 1996.

[8] V. Kasyanov, "SIMICS: Information System on Informatics History. International Federation of Information Processing," ICEUT 2000

http://www.ifip.or.at/con2000/iceut2000/iceut05-05.pdf
[9] K. Watkins, "Bioinformatics," Chemical & Engineering News, Feb 19, 2001.
[10] D. Baird, A. Shew, "Probing the History of Scanning Tunneling Microscopy," SHOT October 2002, Society for the History of Technology, http://shot.press.jhu.edu
[11] K. Schwall, E. Shanbrom, "Narrowing the Boundaries," Bioinformatics World, Spring 2003
[12] Dibner Institute for the History of Science and Technology, "The History of Bioinformatics," http://hrst.mit.edu