Bioinformatics and Computational Biology: The Interface Between Computing and the Biosciences

John C. Wooley, Ph.D.

The Field of Bioinformatics and Interest in its Application

What is bioinformatics? What is the “buzz” all about, and why has the field of bioinformatics become so visible, popular, useful and even apparently essential? What might bioinformatics provide for the readers of this journal, what are some of the challenges for the science, and does it have a future? These are among the obvious questions when anyone is asked to describe bioinformatics and place the technical term, scientific field and methodological approaches in the context of interest for the audience. Every reader of Asia Pacific Biotech News will be familiar with the “hot job” market in the field for almost a decade; that the job market has continued to grow; and that the private sector sustained bioinformatics before the educational sector came on board. But now, programs and departments of bioinformatics are being created in every kind of university and technology institute around the world. Yet, simple definitions and perspectives on bioinformatics remain elusive, like the proverbial collection of seven blind men examining an elephant. This “elephant” is very complex; indeed, it might seem it is more like analogous examinations around an entire zoo. With that caveat in mind, let us consider the answers to the questions.

Bioinformatics concerns catalyzing scientific inquiry at the interface between computing and biology, in order to achieve a level of knowledge, and even wisdom, about an integrated view of biology and in turn, to serve society through specific applications to human wellbeing i.e. to health and the environment. Generally speaking, computational biology is a synonym for bioinformatics, and increasingly, the entire frontier at the interface—the use of computer-driven methods to enable biological knowledge—is considered bioinformatics, which clearly is the central growth area for all of biology. The essence is the use of technology (hardware and software) and increasingly, the philosophy or methodologies of scientific
computing to facilitate our understanding of the data acquired through the experimental biosciences. In turn, as experimentalists have accepted and adopted the use of computer tools, bioinformatics has moved from the periphery to the forefront of attention and application. Federal agencies have held numerous workshops and developed new funding programs for basic and applied academic science support, while the pharmaceutical and biotechnology industries had already begun more than a decade ago to depend on bioinformatics. Overall, the stimulation of research exploiting computing and information technology in the service of the biosciences portends nothing short of a sea change in the very nature of biological inquiry and in its application.

### The Changing Nature of the Biological Sciences

Over the past decade, the biosciences have been transformed into a field characterized by data-intensive research obligations. The rapid growth, both extraordinary and unanticipated, in very complex data, and the difficulties which have become “impossibilities” associated with manual analysis resulting from the molecular biology revolution, with its exclamation mark around the completion of the complete sequence of the human genome, has changed forever how biological research is conducted. All fields of pure and applied life sciences or biosciences including each discipline in biology, from molecular to ecological, biomedicine, translational medicine, and clinical medicine — are characterized by a rapidly growing collection of highly complex data. Ever larger data sets are acquired both through high throughput experimental methods (such as sequencing and screening) and by high information content methods (such as electron microscopy and magnetic resonance imaging), while some approaches (such as microarray analysis) have both features.

The key word is complexity; there are other scientific fields that are far richer in data volume, but the data can often be reduced through mathematical operations or is highly redundant. In contrast, the data in biology is highly heterogeneous and hierarchical in nature, ranging over vast orders of magnitude in space and time and over extraordinary levels in organization (molecules to patients to ecosystems), and is acquired by a wide range of means with different standards, yielding highly disparate types of data. The biosciences correspondingly have ever expanding requirements for the establishment and maintenance of well-organized, integrated, and stored data, as well as of well-understood data, i.e. information synthesized into knowledge.

### The Origins and Early Contributions of Bioinformatics

Bioinformatics began as a combination of the activities associated with biological databases, on the one hand, and the use of computer-performed comparisons of biological information, such as sequence and structural
data, and particularly of pattern analysis and the search for clues to provide an explicit knowledge of the implicit information content of biological sequences and structure, and the evolutionary relationships of macromolecules. Bioinformatics expanded to include the analysis of functional information including genetic expression, inference of evolutionary or phylogenetic relationships for species, and the study of cellular relationships, including the details of genetic circuits, metabolic pathways, signaling cascades, and even of general physiology in health and disease.

Subsequent to the pioneering and essential establishment of computer tools for searching and comparing macromolecular sequence information, a contribution that is often considered as the earliest exemplar — demonstration to experimental biologists of the wide ranging, deep importance of bioinformatics as a powerful vehicle for discovery and innovation was the computer-driven finding of the existence of cellular cancer genes or oncogenes. Specifically, bioinformatics research using algorithm-enabled searching and comparison of the sequence information for all genes then known revealed that a newly found cancer inducing gene was the same as a normal gene in healthy cells, where it is engaged in routine cellular development and growth. Consequently, our understanding of cancer and our approaches to studying cancer changed, in that bioinformatics opened our eyes to the possibility that cancer might be caused by a failure in genetic control, that is, by the expression at the wrong time, place or level of a perfectly normal gene involved in the processes of cell growth.

A recent overview article has described one possible list of the 10 top advances (out of a very large set) contributed by bioinformatics in the past decade, including those in sequence data integration, multiscale biological modeling, and networking of scientific inquiry and research collaboration. The list, largely in chronological order, includes the introduction of sequence alignment tools, the implementation of reliable, high throughput sequencing and its broad extension into systems biomedicine; the identification of genes for disease susceptibility; the development of an accurate model for HIV infection; the establishment of tomographic reconstruction; the advanced design of new prosthetic devices; the provision of detailed, realistic models of the electrical behavior of excitable tissues including neurons; the dissemination of biological knowledge via the internet; the extension of expertise and instrument access to distant locations through telescience; telemedicine and other forms of remote collaborations, and the establishment of information resources for the surveillance of infectious disease on a worldwide basis.

For molecular bioinformatics approaches per se (i.e., setting aside neuroinformatics, environment and ecosystem informatics, and medical informatics, all of which have parallel and additional extensions), the field today includes considerations of biological databases and information retrieval; the analysis of genetic information at the level of nucleotide and amino acid sequences and of protein structure and macromolecular interactions, recognition, complexes and pathways; the inference of relationships (and implications for experimental validation) for assembling of DNA sequence data; and the analysis
of sequence relationships, gene identification and genome annotation, phylogenetics, comparative genomics, gene expression and protein science (or proteomics).

**The Wide Content and Implications of Bioinformatics**

The intersections of experimental studies from all domains of biology with bioinformatics provide in toto the basis for establishing a systems biology, or a high information content biology, and the full integration of expertise and innovation from both sides of the frontier—computing and biology—will provide the basis for the ultimate expectations of research in the life sciences for the indefinite future. Establishing this basis, or infrastructure, will include, for example, research to establish how regulatory elements and protein factors guide the implementation of the developmental pattern specified in the genome; to identify and characterize the multiprotein complexes that execute cellular functions and govern cell form; to characterize gene regulatory networks; to establish (increasingly more accurate representations of) the tree of life; to understand the origins of multicellularity; and to characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level.

Bioinformatics provides the essential glue in linking the results of biological experiments done on individual scales, such as genome sequences, macromolecular structures, regulatory networks (from metabolic pathways, genome circuits to signaling cascades), to develop a comprehensive understanding of biology. A step beyond establishing the components and processes of living cells is to establish tools to predict the behavior of cells to external stimuli, and thus, for example, how plant cells respond to environmental changes. Any such mechanistic description will require a new generation of computational methods and capabilities to advance understanding of any complex biological system before cellular or organismic responses to environment can be predicted.

All of those components—assembly and annotation of genomes; analysis of proteomics; modeling of pathways and networks; modeling of cells; and the modeling and simulation of microbial community actions; computing the tree of life; modeling ecosystems and community ecology; looking at the capture and analysis of distributed biological data; federating biological databases; and providing data resources that range across the multiple scales characteristic of living organisms—are components of bioinformatics. More generally, research objectives of bioinformatics include analyzing the vast scales of time, space and organizational complexity that characterize biology; understanding the relationship of structure and function, and providing the basis for understanding the complexity of living systems, as well as more narrow considerations, such as network analysis and simulation, new databases, data integration, microbial
ecology support, modeling and simulations, and visualization) lie entirely within
the purview of bioinformatics.

The Challenges and Opportunities from the Biosciences

Traditional approaches in informatics for business and science, which work
so well for airline schedules or the physical and chemical properties of materials,
are inadequate for application in biology; among other reasons, this is due to
the complexity of biosystems, the lack of common vocabularies and standards
for structuring data and its representation even within individual biological
subdisciplines, the wide range of methods and instruments used, and the production
of biological data in a large collection of laboratories around the world. The
opportunities for bioinformatics to transform biology are widespread, central to
temporary bioscience research, and have often been presented as a list of
grand challenges. An extensive list of such grand challenges (from numerous
other, earlier reports and presentations) is given in Appendix B in a recent study
of the USA National Research Council. As a consequence of the extent of such
difficult and intellectually compelling challenge problems, an intense, interactive
and innovative partnership between biologists and computer scientists is beginning
in order to establish the information platforms managing the breadth of biological
data, the extensive range of modalities of observation and the even more distributed
nature of data acquisition.

With the acceleration of bioinformatics as a research method and discipline,
the community recognizes ever more challenge problems that will require further
adventures at the frontier, at the interface between computing and biology. Two
notable, more recent lists of top or grand challenges in bioinformatics identify
those in the life sciences as a whole and in biomedicine. Among the challenges
arising as bioinformatics and biosciences mature together are extracting regulatory
networks from DNA sequence information and large scale comparative genomics,
the synthesis of data obtained from multiple, high throughput experiments with
different methods; the visualization and further exploration of large volume;
multiscale data; the conversion of static cartoons of macromolecular networks
into dynamic models, the accurate prediction of structure and function from
sequence; finding molecular signatures for cellular states such as health and
disease; the creation of hierarchical models across vast scales of time and space;
and the reduction of detailed, sophisticated multiscale models to discover the
fundamental or underlying principles— i.e., the synthesis of robust theories for
the biosciences.

For more applied biosciences including biomedicine, translational research
and preclinical medicine, the challenges include computational screening of drug
compounds. Predicting function from complex macromolecular architecture,
developing efficient, comprehensive and accurate models of the dynamics of
traditional and emerging infectious diseases and their spread; automated or smart,
computer-driven methods for powerful mining of the biological, biomedical,
pharmaceutical and clinical literature; improving the accuracy of annotation
and implementing fully automatic methods to provide explicit information about
completed genomes and manage the continuing exponential growth in sequence information; innovations to create a computer-based healthcare delivery system (including definitive electronic medical records, remote, efficient 24 by 7 patient monitoring, and on demand, on location access to essential patient information for urgent and distributed care); enabling the interplay of computing, theory and experimentation required for systems biology; optimizing biomedical software for each new generation of computer hardware; and involving the tools of bioinformatics in education at all levels in order to ensure that an adequate cadre of a new generation of individuals trained in the quantitative sciences will be available to drive all of the above challenges for the pure and applied life sciences.

Frontier Science at the Interface between Computing and Biology

Bioinformatics, today, is pervasive across the bioscience fields, and it is not possible to cover in a brief article even a hint of the intellectual richness and potential of the field.

There are even specialties, in that many scientists distinguish traditional or classical bioinformatics, meaning especially, the use of pattern matching tools to study and apply genome information and data collected in molecular cell biology and structural biology, from other specialized extant domains that are advancing as rapidly, such as ecological informatics, neuroinformatics, and medical informatics. An extensive consideration of bioinformatics, which would be comprehensive but not truly inclusive, would include:

I. Computing and information technology in building and implementing tools to acquire, store, manage, query, and analyze biological data in a myriad of forms and in enormous volume for its complexity.

II. Technologies for computing that will allow biological scientists to understand many types of biological data in context, and even at very large data volumes, and to make model-based predictions that can then be tested empirically.

III. Information and computing abstractions that can be used to interpret and understand biological mechanisms and function.

IV. The role of cyberinfrastructure in supporting 21st century biology. Cyberinfrastructure is a term coined to refer to distributed computer, information, and communication technologies and the associated organizational facilities to support modern scientific and engineering research conducted on a global scale.\(^5\)

The essential, biologically oriented computing tools, introduced through computing and information technology, will allow biologists to move from the study of individual phenomena to the study of phenomena in a biological context, to move across vast scales of time, space and organizational complexity, to understand the relationship of structure to function, and to utilize properties such as evolutionary conservation to
ascertain biochemical and biological functional details. The use of computing models and methods will allow biological scientists to tackle harder problems that could not readily be posed without visualization tools, rich databases and query tools, and the development and implementation of new methods for making quantitative predictions. Biological modeling itself has become possible because data is available in unprecedented richness and because computing itself has matured enough to support the analysis of such complexity. Because both computing and biology are concerned with function, information and computing abstractions can provide well-understood constructs that can be used to characterize the biological function of interest. Further, such abstractions may well provide an alternative and more appropriate language and set of abstractions for representing biological interactions, describing biological phenomena, or conceptualizing some characteristics of biological systems. For the biological sciences, the provision of bioinformatics tools and access to all as well as a commitment to education is at the heart of cyberinfrastructure. Because 21st century biology seeks to integrate scientific understanding at multiple levels of biological abstraction and in sets of interactions between components of biological systems, cyberinfrastructure will be an enabler for 21st century biology in integrating enormous data-intensive efforts distributed over multiple laboratories and investigators.

The impact on biology of these four aspects within scientific computing and advanced information technology can accurately be described as a paradigm change happening as the life sciences research endeavor enters the 21st century. To foresee what lies ahead due to this paradigm shift, consider an earlier, now “classic” and famous case that has many parallels; namely, the impact of molecular biology on the life sciences as a whole over the past few decades. Molecular biology became ever more essential for the success of all disciplines of biology over this period, and then extended to biomedicine and today, to clinical medicine. Originally, multiple disciplines from the physical and biological sciences were integrated and applied in new ways to understand the mechanisms by which simple bacteria and viruses function. The impact of the early efforts was so significant that a new discipline, molecular biology, emerged, and many biologists, including those working at the level of tissues or systems and whole organisms, came to adopt the approaches and even often the techniques. Molecular biology has had such inclusive, demonstrable success and such overarching, fully pervasive impacts that it is no longer a discipline but rather, simply an integral part of bioscience research itself. Today, it is fair to say that among other traits, all biologists are molecular biologists.

The Past, Present and Future of Bioinformatics

In an exceptionally prescient article,10 Walter Gilbert noted what was to occur in the next Century; namely, a paradigm shift in biology in moving to genome enabled science:

“The new paradigm, now emerging, is that all of the genes will be known in the sense of being resident in databases available electronically, and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture only then turning to experimentation to follow or test that hypothesis.”
This vision was well presented more than a decade ago. During this period, entire genome sequences have been completed, numerous advances in high throughput biology have been instituted, and both the computer sciences and the biological sciences have matured in extraordinary, “exponential” ways. Information technology and scientific computing provides an infrastructure for the integration of broad disciplines in developing a quantitative systems approach, or, in other words, an integrative or synthetic approach, to understanding the interplay of complex biological entities as biological research moves up in scale. Bioinformatics provides the glue for systems biology; computational biology provides the methods for obtaining new insights into how best to approach or interpret key experimental approaches and how to tackle the challenges of nature.

The entire biosciences today represent an endless frontier, thanks to molecular biology and the introduction of bioinformatics. The interplay between molecular biology and bioinformatics led the basic and applied life sciences research to become rich and they will soon become exceptionally rich, in diverse, complex data. Increasingly, research will combine computational and experimental approaches to enable an integrative or synthetic understanding, and a predictive capacity for fundamental biology, biomedicine and clinical medicine. The language for describing, studying and understanding biology at a systems level will be that of bioinformatics (IT), just as calculus (math) has been the language for understanding the physical sciences.

Computing and information technology applied to biological problems, in sum, is likely to play a role for the next quarter century that is in many ways analogous to the role that molecular biology has played in biological research across all fields for the last quarter century and that computing and information technology will become embedded with biological research itself. Many observers of scientific progress and policy believe that bioinformatics will extend its reach even faster than molecular biology did and will serve to integrate our understanding into a biological synthesis frequently called 21st Century Biology. As such, for sure, the challenges and excitement of bioinformatics will continue to grow; knowing bioinformatics will be important for doing cutting edge bioscience and biomedical research; bioinformatics will be part of the biosciences of the future, for all of the future. A notational summary of this article is provided in Fig. 1.
Bioinformatics Challenges to Enable a Biological Synthesis

Dynamic Form and Function: Characterizing the Biological Machinery Across Multiple Scales

- Scientific Challenges
- Algorithmic Challenges
- Data Integration Challenges
- Computational Challenges

Adapted from Summary Slide from the NSF, NCI, DOE 1998 Workshop on “Next Generation Biology: The Challenges for Next Generation Computing”- Convened by J. Wooley, chaired by R. Subramaniam and J. Wooley

Fig 1: A summary of the challenges in bioinformatics and computational biology that must be addressed to implement a synthesis of biology, a robust theoretical biology perspective around an integrated understanding of biological systems.

References