

Executive Summary

Bioinformatics and computational biology (B&CB) refer to the investigation of phenomena in the life sciences through the application of statistics, biometrics, data storage, sharing, and analysis. The field arose in the latter half of the 20th century as the theoretical foundations of biology became more receptive to mathematical applications and as computer technology and biological instrumentation were introduced capable of handling the models, the volume of data and the analysis required of the newly developing discipline.

The first chapter of this report lays out the ground rules; bioinformatics and computational biology grew out of molecular biology, biophysics and computer science. We define Bioinformatics as the interdisciplinary science of the management and interpretation of quantitative biological information by which data acquired from biological phenomena are analyzed and modeled using numerical methodologies. Computational Biology is the application and development of computational methods and tools for modeling and analyzing complex biological systems.

We review the origins of these disciplines and how they fused together to form the present-day domain of investigation and study. The field today is massive, and this report sorts through the tremendous volume of material to produce a synthesis which connects it to the biotechnology industry.

The second chapter explores the issue of hardware and how it drives B&CB. The field is completely dependent on vast stores of information that can be analyzed using the modeling techniques of B&CB. Without extremely sophisticated instrumentation for gathering information on the properties of living systems, in particular protein sequencing, there could be no science of bioinformatics. This instrumentation includes biomedical imagery platforms, mass spectrometry, x-ray crystallography, NMR, high-throughput image analysis, and sequencing microarrays.

While hardware is essential for gathering data, computers capable of handling terabytes of data must be available; otherwise the data is of no value. This means that there must be capacity to store the data, but there must also be well-designed search engines that can effectively retrieve the information. From both sides of the issue, it appears that the industry is presently able to handle these demands, but major challenges may be ahead as the storehouses of data grow exponentially.

The third chapter discusses current applications of B&CB to the real world of pharma and biotechnology. The –omics fields are the subject of large-scale projects, as academics and commercial interests probe the possibilities of systems biology, *in silico* drug discovery, genomics, proteomics, metabolomics, and a wealth of subdisciplines. Initially, no new drugs originated from these approaches, and there was some skepticism concerning their relevance to drug development. As the field develops and as a more robust understanding of complex networks and their applications grows more integrated, these approaches have the potential to impact drug discovery.

The Sanger method, a laborious exercise for DNA sequencing executed by hand, in which truncated DNA molecules were separated and identified on large acrylamide gels, gave way to process automated capillary sequencing machines. A number of new platforms are now in use, including pyrosequencing, while some are still on the drawing boards, including nanoknife edge sequencing, sequencing-by-synthesis chemistry, and nanopore array-based systems.

As genomic information becomes more readily available, concerns over privacy and how the data are going to be protected are hotly discussed. DNA sequences are available for forensic analysis and tracing family lineages and, at present, privacy issues have not been thoroughly addressed.

Proteomics is another mega-science activity with important ramifications for health care and the application of B&CB technologies. Initially, it was thought that obvious health-related benefits would fall out from the sequencing of many, many proteins, but this turned out to a naïve proposition. The fact that more than 30 years after the discovery of oncogenes, so few therapeutics have made it to the marketplace is testimony to the extreme difficulty of developing effective drugs against even the most obvious targets. Proposals for moving the analysis of the proteome forward include a bioinformatics approach based on complex network analysis, including modeling of protein interaction and signaling.

Cancer is probably the most actively pursued indication of the major diseases. Recent improvements in survival statistics have encouraged researchers and clinicians to redouble their efforts, and B&CB figure prominently in these approaches. Because cancer is a complex and multifaceted disease, a systems biology approach to its investigation appears appropriate. Moreover, a largely untapped area of cancer research, biomarker characterization lends itself to a bioinformatics approach, focusing on the plasma proteome.

Despite the great appeal of oncogenes as targets for cancer therapeutics, few drugs have made their way through to FDA approval. Progress to develop inhibitors of protein-protein interactions has been beset with problems. The challenge of developing low-molecular-weight alternative molecules with sufficient potency, pharmacokinetic, and safety profiles to be considered suitable for testing in humans has often proved insurmountable.

There are a number of areas of investigation that are opening up, including the role of food constituents in protein expression. Proteomics will play an invaluable role in sorting out these complex and confusing contribution of nutritional factors to protein dynamics.

Infectious diseases, epidemiology, and immunology are all disciplines that are taking advantage of B&CB technologies and strategies in an effort to develop new insights which will lead to improved therapies. Using *in silico* modeling, numerous pathological conditions, including autoimmune and allergic diseases are under analysis.

B&CB also figures prominently for *in silico* drug development. Numerous large and small pharma companies and biotechs are vigorously pursuing this route in the hopes of shortcutting the tedious process of standard drug discovery.

Chapter 4 deals with the management of data and the challenges to workers in the field of B&CB brought on by the extraordinary growth in the amount of data pouring out of laboratories all over the world. Massive stores of data represent a huge challenge for the IT industry and, by extension, for B&CB. There is an increasing availability of well-designed storage area networks that are much easier to install and maintain.

A growing number of software firms are confronting the problem of condensing data. It currently costs roughly \$60,000 to sequence a human genome, and the cost will perhaps reach \$1,000 per genome within the next three years. However, companies are creating an approach that would sequence an individual's genome for a few dollars, and the technology could read the complete genome in a single workday. Notwithstanding, most workers in the field agree that the technology required to achieve a \$100 genome is still at least five years away.

According to popular media reports, the findings of personalized medicine will soon be translated into precise diagnostic tests and targeted therapies, yet this has not yet taken place. Precise studies are challenging for genetic markers with modest effects, and it is difficult to rule out environmental contributions. With the development of bioinformatics, many genetic determinants with minor effects on phenotype have been investigated as potential markers, but only with limited success.

One of the most significant tools for investigating multifactorial or polygenic inheritance is the HapMap, a large-scale science project in which the entire human genome is tracked with allelic variants or SNPs along its course. These serve as markers which can be used to tag or identify variants affecting a complex disease or multifactorially inherited condition. It is under intense consideration as a method for development of personalized medicine, yet it remains controversial and unproven at this time.

One solution to the problem of extremely demanding models based on B&CB construction, such as those derived from personalized medicine, is the possibility of adoption of the "grid network" approach, which would supply additional computing power for dealing with huge collections of data. However, in order to take advantage of such systems, it will be necessary for personalized medicine to be placed on a sounder theoretical footing than that on which it currently resides.

Chapter 5 is a compilation of interviews with a number of specialists in the field of B&CB who give their opinions of the state of the art and point out strategies, technologies, and companies that they feel are particularly noteworthy. Also considered are large projects such as systems biology and the challenges that they must address.

In the final chapter, the challenges and limitations of B&CB are addressed, including the applicability of B&CB to the search for cancer treatments. An urgent question is whether the elaborate technology that drives the advance of B&CB will move beyond the ability of the biotechnology community to build hypotheses that can be tested. In the coming years, B&CB will figure prominently in the formulation of models for various disease and conditions based on the host's cellular activities. In the final analysis, no technology is better than quality of data and the imagination of those who drive the project forward.

