What is Bioinformatics?

Greg Butler

June 2000

Abstract

Scientists working on the genome projects have been early adopters of internet technology. Faced with a huge amount of data, that is growing rapidly in terms of volume, location, and diversity, they have developed pragmatic approaches to resolve their problems of data access, data analysis, large-scale computation, and intelligent data mining to create scientific knowledge from the raw experimental data.

A major driving force in bioinformatics is the development of new experimental techniques. So one can not foresee which questions, which data, nor which paradigms may be relevant from one day to the next.

1 Introduction

Bioinformatics is the meeting point of algorithmics and information technology with the biological sciences, in particular, with the study of gene sequences and related information [1, 2, 3, 4, 5]. The demand for computational tools arises because the scientists are faced with a dataset that is growing rapidly in terms of volume, location, and diversity. This has lead to pragmatic approaches to their problems of data access, data analysis, large-scale computation, and intelligent data mining to create scientific knowledge from the raw experimental data.

We review the current state-of-the-art in bioinformatics for data management, large-scale computation, and for intelligent systems. While impressive advances have been made, they have been driven by pragmatics, and face severe limitations because of this. Extensibility and scaleability are particularly lacking in most bioinformatics systems.

Available technology in software, databases, and the internet can address some of the limitations of existing bioinformatics systems. Ongoing research in these areas of computer science offer even more potential to improve these systems. For still other issues, research into the technologies themselves is required.

The layout of the paper is to first present an overview of genomics and the typical problems faced by biologists. Then we discuss the state-of-the-art in terms of data infrastructure, computation infrastructure, and intelligent computing infrastructure available to the bioinformatics community.

2 Sample Biological Problems

First we provide a simple overview of genomics, and the areas such as proteomics and pharmacogenomics that will develop in the post-genomics era. More background material can be found in the textbooks [6, 7].

The decoding of a gene into a protein is the basic process of genomics. A *gene* is a region of DNA, which consists of the bases A, C, G, T. It is copied to produce a complementary messenger RNA (mRNA), which consists of the complementary bases U, G, C, A. This process is called *transcription*. The mRNA is a sequence of codons. Each codon is a sequence of three bases and it encodes for a specific one of the twenty amino-acids that are the building blocks of proteins. There are also special codons: a *start codon*, AUG, that marks where decoding is to commence; and *stop codons*, that cause the decoding process to stop. The sequence of codons in the mRNA is read to produce the protein as a sequence of amino-acids in one-to-one correspondence with the sequence of codons. This process is called *translation*.

In higher organisms, called eukaryotes, the decoding process is complicated by an intermediate stage where the mRNA is spliced before being decoded.

In the laboratory, the mRNA can be isolated from the cells. It can then be replicated to produce complementary DNA (cDNA).

The proteins are responsible for all the processes in the cell. Each protein has its own specific three-dimensional shape. The shape determines how proteins interact. The sites where interaction takes place are called *binding sites* or *active sites*.

It is the binding of proteins to DNA that controls the transcription. There are often a number of proteins, or a complex of proteins, that bind at adjacent sites in order to regulate transcription.

Some proteins need to be transported to the appropriate compartment of the cell so they may be where they are required for an interaction. Such proteins may reside in the cell membranes and may be part of the signalling mechanisms within a cell or between cells. Small molecules and hormones also play roles in signalling.

Most multi-cellular organisms differentiate their cells into *tissues*, such as nerves and blood vessels, which are organized into *organs*, such as kidneys and liver. The metabolism of a cell varies depending upon its tissue type. Hence, the subset of genes that are active in a cell also depend on its tissue type. Developmental biologists study the mechanisms that lead to differentiation. Drug developers need to be able to characterize the cells in different tissues, so they may target a drug at the site of the disease.

The Nature article [8] describes several levels of data and analysis of interest in genomics:

- the **genome** is the complete set of genes of an organism;
- the **transcriptome** is the complete set of mRNA molecules present in a cell, tissue, or organ;
- the **proteome** is the complete set of protein molecules present in a cell, tissue, or organ; and

- the **metabolome** is the complete set of metabolites — the low-molecular-weight intermediate compounds — present in a cell, tissue, or organ.

The last three datasets vary over time as the activity in a cell unfolds. They also vary with their context within a tissue or organ because differentiated cells have differentiated activities.

Automation of laboratory activities, either through the development of specialised instruments, or the use of robotics, or a combination of both, greatly speeds up the rate of data collection. It also allows parallel use of automation to further increase the rate.

Nanotechnology has shrunk the size of "test tubes" where reactions take place, so that thousands may fit upon a chip or gel or glass slide and take place simultaneously. These are called gene chips, or microarrays, or biochips. Advanced optics, often using combining the use of lasers to excite fluorescent probes, allows the automatic reading of the results of the reactions.

All this technology aids high-throughput production, screening, and collection of experimental data.

Common kinds of experiments that apply this technology to genomics are as follows.

Sequencing experiments use automatic sequencers to determine DNA or cDNA sequences. They may then be submitted to the public archival sequence databases, such as GenBank, or stored locally in proprietary databases. These sequences may be analysed by various computational tools in order to provide evidence of evolutionarily-related sequences, of structure or function for the corresponding protein, or of genes. The sequences may also be assembled into maps showing the layout of complete chromosomes.

Gene expression experiments attempt to determine the role of a gene. The organism may be genetically engineered to *knock-out* the gene, so its effect can be studied. Minor changes may be made to the gene, in a process called *mutagenesis*, to see which parts of the gene determine protein structure, particularly within the active sites. Microarray technology allows an experiment to be performed on thousands of sequences at once, so their behaviour in different contexts can be studied.

Drug discovery first screens for genes or proteins that appear linked to the disease. The properties of the protein are studied. In particular, its active sites are determined. Then a screen of known chemical compounds for ones that may *dock* with the site (and hence interfere with the gene's effect) is carried out. The selected compounds may be slightly altered to enhance docking, or stability of the compound, or lower toxicity, etc. The effect of the compound is then studied in vivo using *model organisms*, such as mouse, where the gene is also present. A model may be created by transferring the gene into a suitable organism: this is called *transgenics*. After these pre-clinical trials in the laboratory, clinical trials with human patients are carried out.

Now we outline some typical bioinformatics activities as they relate to these experiments. In general these involve (1) the keeping of lab notebooks describing experimental protocols, ingredients, reagents, instruments, and results; (2) the management of one or more projects, each of which involves a series of experimental investigations; and (3) the scientific analysis of experimental results. General tools that these use are word processing, document management; workflow, document management, databases; and databases, data mining, modeling; respectively.

2.1 Project and Laboratory Management

A lab notebook is a vital record of all experiments. It records the experimental protocol used. This details the steps in the experiment. The lab notebook also records which batches of reagents and ingredients are used in an experiment, which particular instruments are used, and any observations and results of the experiment.

Traditionally, a lab notebook is kept as a hand-written paper book. Today, word processing with digital images and other automatically collected datasets have either replaced or augmented the paper notebook. Barcodes and barcode readers are often used to track instrument usage, batches of reagents, and individual test tubes and dishes.

Increasingly, we expect to see handheld PDAs (Personal Data Assistants) being used as the lab notebook and data collector, with regularly synchronization of its data with the lab network. A future possibility is for infrared or wireless networks to connect all instruments in a lab for automatic tracking of experimental steps.

The details of experiments in the lab notebook feed into a larger document management system that tracks all projects. These project management tools, such as Documentum [www.documentum.com] or PharMatrix [www.base4.com], are web-based. They provide access to all data related to a particular project, so a manager knows its status. The prime purpose is often to support "stop-go" decision-making by the manager. XML [www.w3c.org] will become the standard for these document management tools.

2.2 Automated Sequence Analysis

Instrumentation has made the collection of sequence data a routine activity. However, sequencers have limits to the length of DNA or cDNA fragment they can process. So the result of sequencing is a collection of DNA fragments given by the sequence of their bases. It is not a collection of genes!

Assembly seeks to combine fragments into longer sequences; ultimately the complete chromosome or genome [10]. This is essentially a problem in combinatorics that is heavily computational. Laboratory experiments and multiple sequencings may be needed to supplement the computation and resolve ambiguities.

Analysis investigates the sequence fragments as they are. Problems include pattern matching to locate regions of interest within the sequence, or comparisons with known sequences

in databases. By finding a known sequence that is evolutionarily related to the sequence fragment, the researcher is in a position to infer by analogy from the properties of the known sequence to those of the fragment. These problems are combinatorial string matching [4]. They make heavy use of heuristics to achieve acceptable performance. The patterns, and their classifiers, are usually determined through machine learning.

Most biologists use, rather than develop, the variety of assembly and analysis algorithms available. The system, SEALS [www.ncbi.nlm.nih.gov], provides a set of Unix utilities for the algorithms so they may be pipelined to provide automatic sequence analysis. A fixed set of analyses, with intelligent assessment of confidence in the results of individual algorithms, is available in the GeneQuiz system [9].

2.3 Gathering a Dataset

The public archives, such as the GenBank sequence database at National Center for Biotechnology Information (NCBI), record all the sequence entries submitted by researchers. Similar databases are maintained in Europe and Japan. Structural information is available in the Protein Database (PDB). Functional information is available in the carefully curated database SwissProt. The medical and biological scientific literature is maintained online by Medline.

Most research projects rely on these archival databases, but need a more focussed dataset, and may need to pull together disparate information from several databases, collate and curate it, before beginning the actual research. Researchers may then add to the information, look for patterns or trends, and carry out laboratory experiments to confirm their hypotheses.

These specialized databases reference the archival data through links to integrated databases, in particular, to Entrez or SRS.

This activity is very common. Each January issue of the journal *Nucleic Acids Research* summarizes the most significant databases for bioinformatics. The online catalogue, DBCat [11], has links to over 400 bioinformatics databases.

2.4 Data Mining

Machine learning has been extensively applied in bioinformatics to construct classifiers from experimental data [3]. Almost all of these applications learn to classify sequences, or regions of sequences, as indicative of properties of gene structure, or protein structure or function [12, 13]. This is not surprising, since machine learning techniques need a body of examplars from which to learn, and to date the available data has primarily been sequence data.

The use of hidden Markov models (HMMs) has had particular success for the construction of classifiers [2].

Many databases include information on these patterns or classifiers as entries. For example, the Prosite database on protein families stores patterns as scoring matrices in order to classify a region of a sequence as indicative of a functional family of proteins.

Biologists generally access the classifiers, the results of machine learning, rather than apply the machine learning techniques themselves.

Recently, the need to populate databases with information about reactions and interactions has encouraged researchers to automatically extract such information from the scientific literature. This use of natural language processing (NLP) techniques dates back to mid 1990's work of Futrelle.

2.5 Brainstorming

The crux of scientific discovery is to complement the automation with insight. When the result of experiments is predictable, then automation will lead to the predicted information that can be analysed and documented as planned. However, one expects the unexpected, and is then required to carry out ad hoc, creative, original analyses.

Support for creative work must quickly respond to a scientist's intuition. It must be possible to harness available tools in novel ways so that datasets can be gathered, patterns determined, and analysis performed almost instantaneously. Any slower, and the scientist loses their train of thought.

Unfortunately, current tools are not this flexible, reusable, or interoperable.

3 Available Infrastructure for Bioinformatics

We provide an overview of the state-of-the-art in bioinformatics tools. We focus on the infrastructures available for data management, large-scale computation, and intelligent computing.

3.1 Data Infrastructure

Key to the success of genomics and bioinformatics has been the central data archives in USA, Europe, and Japan. They have provided tools as well as data archives, all of which are accessible over the web. An overview of available databases and their technology can be found in [14, 15].

At NCBI in Maryland, the GenBank sequence database can be queried using the BLAST alignment tool. The Entrez system integrates GenBank with the Protein Database (PDB) of protein structure and the Medline archive of the scientific literature. The three databases are cross-referenced, and the sequences are linked to the closely similar sequences (as determined by BLAST). These links and cross-references can be followed when browsing, and referenced in form-based queries of Entrez.

The quality of data is a major concern. While the archival sequence databases accept all submissions, there are several very carefully curated databases, such as SwissProt, too.

Curation is supported by annotation tools and platforms, such as Magpie [16]. Confidence in the database entries plays a role in automated sequence analysis.

The prevailing database technologies are flat files, relational databases, and AceDB (an object-based database system custom built for genome projects). For the databases listed in DBCat [11], about 80% are flat files, 5% use relational databases, and 12% use AceDB.

A number of companies provide corporate-wide data management solutions, such as Base4 [www.base4.com], Lion [www.lion-ag.de], Synomics [www.synomics.com], and Netgenics [www.netgenics.co
Most are CORBA-based document and project management systems with hooks to some bioinformatics analysis tools. Netgenics has a very open system that can be extended by registering a new application with the name server. This provides information about APIs, capabilities, and input/output formats. Once registered, the new service can be accessed across the system through the web browser.

Pragmatic solutions to the problem of data integration are provided by Entrez (NCBI), SRS (Lion), and OPM (GeneLogic). Stronger solutions are BioKris (previously called BioKleisli and based on the collection programming language, CPL) and TAMBIS (Manchester). BioKris provides a language for defining collections, translations of data types, and set-comprehension queries and filters. It integrates at the data level, whereas TAMBIS integrates at the conceptual level by providing an executable ontology.

An integrated Information retrieval system, ENTREZ, is maintained by the NCBI and can be accessed through the web. It provides links to the literature, weighted key term analysis to automatically retrieve related articles, following references ("hard links") to other data-sources, and online database searches with a submitted sequence. It has recently become a major source for projects extracting information by applying NLP techniques.

Another information retrieval, SRS, is maintained at the EBI and focuses on databases rather than literature. It provides its own query language including the usual set of boolean operators and provides links between the databases. As a consequence it is possible to retrieve subsets of databases with a certain feature annotated in another database while all databases can be stored locally by the user in their own formats.

TAMBIS is a system that provides the user integrated access to databases by first creating a homogenising layer on top of the different sources. The mediator between sources and the layer is an information broker based on a conceptual knowledgebase which is based on GRAIL, a description logic language. Wrappers access and retrieve information from the individual databases. TAMBIS represents an ontology of biological information as a hierarchy of logic expressions.

OPM (Object Protocol Model) is unique in that its data model includes the modeling of laboratory protocols.

3.2 Computation Infrastructure

The web provides a generic infrastructure for computation, in that researchers can access tools and services at a variety of web sites. The range of facilities used by researchers includes

individual PCs and workstations, clusters of workstations, beowulfs [beowulf.gsfc.nasa.gov] which are tightly coupled clusters, and supercomputers [17]. Distributed computing is supported by the CORBA-based systems for project management (see above). The Canadian Bioinformatics Resource in Halifax provides high-speed access over Canada's optical advanced network CA*net3 using GO-Joe [www.graphon.com] connectivity software.

MobiDick [18] is a platform for distributed computation that supports load balancing, process management, and scheduling of access to a collection of computing resources. The computing resources can be a heterogeneous mix of PCs, workstations, beowulfs, or supercomputers. MobiDick manages a task that is parameterised and where each parameter set defines a parallel subtask. Each subtask is identical, except for the parameters. MobiDick partitions the parameters, assigning a subset to each computing resource. Each computer has the code for the subtask already installed, receives the request for execution over `http` and `cgi`, and returns results to the MobiDick server.

At a simpler level, the SEALS package from NCBI provides a set of Unix utilities for computation and data translation that can be pipelined. The GeneQuiz system provides a fixed distributed set-up to perform its individual analyses.

The MOE system from Chemical Computing Group [www.chemcomp.com] is an open computing environment based on a parallel vector language called SVL. MOE contains implementations in SVL for many algorithms from bioinformatics and computational chemistry. These can be customized by the user, or used in new algorithms, by coding in SVL. While conceptually SVL supports parallel and distributed computation, the SVL implementations do not at present.

3.3 Intelligent Computing Infrastructure

There is no available infrastructure for intelligent computing in bioinformatics. Netgenics is working towards such facilities for its CORBA-based framework. They are incorporating IBM's data mining tools.

The GeneQuiz system has an architecture of interest for intelligent computing. It uses a simple relational database to collect the results of the individual analysis of each sequence. It uses an expert system with a small ruleset to interpret the results and infer a best guess for the function of the sequence. This expert system and its ruleset are actually hard-coded in C, so it is not as flexible as it should be. The summary of the analysis is provided as a web page, with links to those individual analysis results that contributed to the best guess. In this way, the researcher may "drill down" and see the actual data and an explanation for the reasoning.

4 Conclusion

The Internet is a significant technology for bioinformatics. Scientists working on the genome projects have been early adopters of the technology as they attempt to create scientific knowledge from the raw experimental data. Bioinformatics heavily uses software technology and

database technology too. Scientists need an infrastructure for data management, large-scale computation, and for intelligent systems. It should be easy-to-use, quickly customizable, extensible, and scaleable. They need the future promise of Internet computing today!

References

[1] Andreas Baxevanis and B.F. Francis Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* John Wiley and Sons, 1998.

[2] Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison. *Biological Sequence Analysis.* CUP, 1998.

[3] Pierre Baldi and Soren Brunak. *Bioinformatics: The Machine Learning Approach.* MIT Press, 1998.

[4] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences.* CUP, 1997.

[5] Steffen Schulze-Kremer. *Molecular Bioinformatics: Algorithms and Applications.* Walter de Gruyter, Berlin, 1996.

[6] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D. Watson. *Molecular Biology of the Cell.* Garland Publishing, New York, 3rd edition, 1994.

[7] Lubert Stryer. *Biochemistry.* W.H. Freeman and Co., New York, 3rd edition, 1988.

[8] Stephen Oliver. Guilt-by-association goes global. *Nature*, 403:601–603, 2000.

[9] M.A. Andrade, N.P. Brown, C. Leroy, S. Hoersch, A. de Daruver, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis, C. Sander. Automated genome sequence analysis and annotation. *Bioinformatics*, 15(5):391–412, 1999.

[10] Special Issue on Computational Biology. *IEEE Computing in Science and Engineering*, May-June 1999.

[11] Claude Discala, Marion Ninnin, Frederic Achard, Emmanuel Barillot, and Guy Vaysseix. DBcat: a catalog of biological databases. *Nucleic Acids Research*, 27(1):10–11, 1999.

[12] Simon Kasif. Datascope: Mining biological sequences. *IEEE Intelligent Systems*, 38–43, November-December 1999.

[13] Steven L. Salzberg. Gene discovery in DNA sequences. *IEEE Intelligent Systems*, 43–48, November-December 1999.

[14] Stanley I. Letovsky. *Bioinformatics: Databases and Systems.* Kluwer Academic Publishers, Boston, 1999.

[15] Dimitrij Frishman, Klaus Heumann, Arthur Lesk, Hans-Werner Mewes. Comprehensive, comprehensible, distributed and intelligent databases: Current status. *Bioinformatics*, 14(7):551–561, 1998.

[16] Terry Gaasterland and Christoph Sensen. Fully automated genome analysis that reflects user needs and preferences — a detailed introduction to the MAGPIE system architecture. *Biochemie*, 78:302–310, 1996.

[17] Carl J. Beckman, Donald D. McManus, George Cybenko. Horizons in scientific and distributed computing. *IEEE Computing in Science and Engineering*, 23–30, January-February 1999.

[18] Moez Dharsee and Christopher Hogue. MoBiDiCK: Modular bioinformatics distributed computing kernel. http://bioinfo.mshri.on.ca/projects/mobi/index.html

[19] P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, A. Brass. An ontology for bioinformatics applications. *Bioinformatics*, 15(6):510–520, 1999.