The Era of Bioinformatics

Dana Jawdat
The Atomic Energy Commission of Syria, Molecular Biology and Biotechnology Department,
P. O. Box 6091, Damascus, Syria
E-mail: djawdat@aec.org.sy

Abstract

The recent technological developments in the field of molecular biology have vastly contributed in feeding the scientific research community with massive amounts of biological data that will be eventually stored and analyzed. Scientists and researchers involved in the field of molecular biology are facing the challenge of analyzing and interpreting massive numbers of Cs, Gs, As and Ts sequences that encode functional biological processes guaranteeing an organism survival. Genomics and proteomics research aims at collecting various living organisms' genetic resources to study genes and their products that regulate biological processes granting food, fiber, energy and other compounds that are essential for human health and environment. The storage and analysis of biological data using certain algorithms and computer softwares is called BIOINFORMATICS. The present paper will point out bioinformatics elements and applications. It will portray the importance of bioinformatics for a better understanding of our environment. This paper will also make reference to some of the debatable issues related to bioinformatics and some of its applications.

1. Introduction

Bioinformatics holds the keys for a better understanding of structural, comparative and functional genomics and proteomics. It has been described that genomics is the study of the genome representing the whole set of genes, whereas, proteomics is the study of the proteome that describes the protein complement of the genome [1].

The integration of three scientific domains: molecular biology, mathematics and computer sciences have brought forth bioinformatics as an influential tool for accessing, comparing and analyzing the data stored. Bioinformatics common definition is the design, construction and use of software tools to generate, store, annotate, access and analyze data and information related to molecular biology.

If considering the human genome only which is around 3x10^{12} base pairs long, 3 GB of computer data storage space are needed to store the entire genome. This includes nucleotide sequence data only and does not include data annotations and other information that can be associated with sequence data (http://www.ornl.gov/sci/techresources/Human_Genome/faq/faqs1.shtml). Other genomes of other organisms are to be considered as well, which urges the efficient use of computational biology.

The present paper will state the history of bioinformatics and will underline bioinformatics tools and applications that are mostly used by related researchers. The paper will indicate current debates regarding bioinformatics ethics.

2. History of bioinformatics

Last century witnessed major advances in computer sciences and molecular biology that resulted in bioinformatics, an efficient tool in the hands of researchers. In 1933, Tiselius introduced electrophoresis for separating proteins in solution. Where in the 1950s, Watson and Crick proposed the double helix model for DNA and the sequence of the first protein to be analyzed, bovine insulin, was announced by Sanger. Meanwhile in the 1950s, the first integrated circuit was constructed by Jack Kilby. First use of molecular sequences for evolutionary studies by Zuckerkandl and Pauling was introduced in the 1960s. During the 1960s, ARPANET was created by linking computers at Stanford, UCSB, The University of Utah and UCLA. The 1970s witnessed a breakthrough in molecular biology and genetic engineering when the first recombinant DNA was created by Paul Berg and his group. Meanwhile, Southern published the experimental details for the Southern Blot technique of specific sequences of DNA [2]. Sequencing DNA was also first reported in the 1970s by Allan Maxam and Walter Gilbert (Harvard) and Frederick Sanger (U.K. Medical Research Council). The first Brookhaven Protein Data Bank was announced in 1977 [3]. The 1970s attended great breakthroughs in the field of connection and communication. The internet surfaced in the 1970s when Vint Cerf and Robert Kahn developed the concept of connecting networks of computers into an "internet" and developed the Transmission Control Protocol (TCP). The foundation of Microsoft Corporation by Bill Gates and Paul Allen has introduced the globe to the world of softwares and computer programmes. Algorithms that are essential for nucleic and protein sequences comparison started to emerge such as the Needleman-Wunsch algorithm. Moving to the 1980s, molecular biology took a different shape when Polymerase Chain Reaction (PCR) was used by Saiki et al. in 1988 [4] for DNA amplification with a thermostable DNA polymerase. In fact, the 1980s period was also famous for

0-7803-9521-2/06/$20.00 ©2006 IEEE.

publishing the physical map of *E. coli* [5]. Centers for biotechnology information were established such as The National Center for Biotechnology Information (NCBI) at the National Cancer Institute in the US. More databases revealed itself in order to manipulate the increasing amounts of data, such as The SWISS-PROT protein database (www.swissprot20.org). Molecular advances moved forward with parallel in computer sciences and mathematics applications. More algorithms for sequence alignments and comparison were published such as the Smith-Waterman algorithm and the FASTA algorithm. The Personal Computer was introduced to the market by IBM which had a great impact on facilitating researchers and others work in different domains of life. Marching towards the second millennium, researchers have witnessed the integration of softwares and computer programmes with crude bench work leading to the introduction of automated protocols and the production of high throughput techniques that accelerated the acquisition of biological data which in turn demanded the creation of an easy-access and analysis of data environment for molecular biologists. New computer operating systems were launched such as Linux by Linus Torvalds. UNIX and Linux operating systems allow a computer to be shared, providing secure and simultaneous access for multiple users at different times or at the same time. Such systems can carry out multiple tasks and allow sharing of resources such as files and hardware across a network. Recently, grid computing is used for massive data processing [6]. The progression of sequence alignments and sequence comparisons has resulted in the Basic Local Alignment Search Tool (BLAST) program [7] that is implemented in the majority of gene databases banks for search and fishing out related sequences from diverse organisms. The second millennium was crowned by publishing genomes of diverse organisms and most importantly the complete human genome in 2003. Reaching the golden era of molecular biology urges the integration of sciences for introducing the most comprehensive tools for analysis, besides the demand for data storage and maintenance. Bioinformaticians in the present time are busy creating suitable environments that are online and/or computer packages to facilitate molecular biologists work. Specialized computer packages are developed for a particular type of analysis and used intensively by researchers in the relevant area such as the Phred-Phrap package by Phil Green and co-workers at the University of Washington in Seattle (visit www.phrap.com for more information). Whereas, general packages offer comprehensive range of bioinformatics tools for sequence analysis for the majority of researchers. Such general packages are GCG Wisconsin package that runs on UNIX operating system and is available commercially. The free, open-source counterpart package is EMBOSS (emboss.sourceforge.net) that also needs UNIX operating system and have similar structure to the GCG package (www.accelrys.com/products/gcg/index.html).

3. Examples of international collaboration in the field of bioinformatics

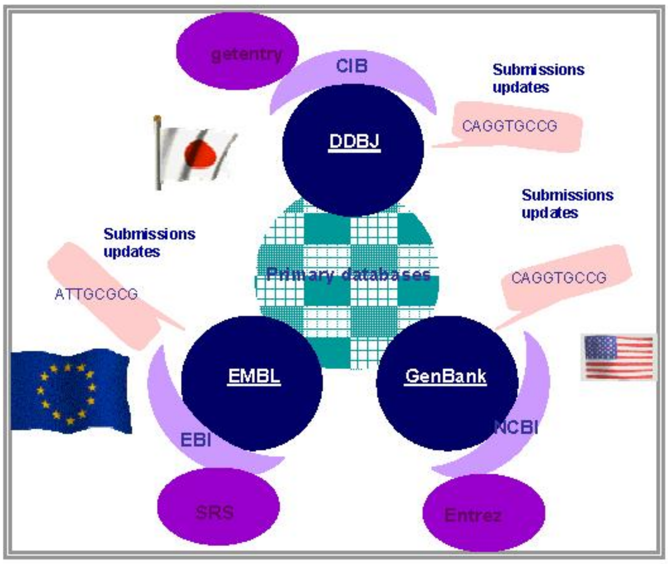
The international research community brought into being a non profitable collaboration project that enables researchers to submit data at one site and access same data at different sites. Three major bioinformatics institutes have joined their primary databases: The NCBI in the US (www.ncbi.nlm.nih.gov), The European Bioinformatics Institute (EBI; www.ebi.ac.uk) in Europe and the Center for Information Biology (CIB; www.cib.nig.ac.jp) in Japan. Researchers can submit their sequence data in primary databases in any of the following sites: GenBank at the NCBI, the European Molecular Biology Laboratory (EMBL; www.ebi.ac.uk/embl) at the EBI or the DNA Data Bank of Japan (DDBJ; www.ddbj.nig.ac.jp) at the CIB. Primary databases are concurrent for access at any of the databases mentioned earlier. This can assist in avoiding peak hours where servers in one continent get busy. Primary databases consists of original submissions by experimentalists and the content is controlled by the submitter. Whereas, derivative databases are built from primary databases and their content is controlled by a third party: Entrez at the NCBI, Sequence Retrieval System (SRS; http://srs.ebi.ac.uk) at EBI and Getentry at CIB. Figure (1) summarizes the joint collaboration among the three centers. Diverse tools and programmes for sequence search and analysis are available at each of the sites and a researcher is free to choose the most suitable tool for his/her research. Some of bioinformatics tools used by researchers are listed in table (1).

4. Bioinformatics applications

Bioinformatics tools proved to be the heaven of molecular biologists under this heavy bombardment of bio-data. The applications that bioinformatics offer to the civilized world are more than just being a researcher's tool for structural and functional analysis. Bioinformatics serves as a vehicle for a better understanding of the surrounding environment. Bioinformatics domain of science is trying to serve the scientific community and in consequences the lay communities with up-to-date bio-information. The Tree of Life web project (ToL; http://tolweb.org/tree) is a collaborative pioneer project of biologists from around the world. The site holds information about the diversity of organisms on earth, their evolutionary history (Phylogeny) and characteristics. The project employed morphological and genetical data for the analysis and determination of relationships among species. The project provides exciting information that can be easily obtained online by interested lay individuals besides related researchers.

The Human Genome Project (HGP; www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml) has enriched the human research community with massive amount of information by the year 2003 when collaborators have published the complete human genome.

Intensive efforts are conducted to identify new genes along the human DNA sequence. Bioinformatics tools are employed for structural, comparative and functional analysis and will increase the potential of curing inherited human diseases and producing new human medicines. Benefits to human mankind are vast with the help of bioinformatics.



The project aimed at identifying approximately 20.000 to 25.000 genes in human DNA, determining the sequences of $3x10^{12}$ chemical base pairs that make up the human DNA, storing information in databases and improving tools for data analysis besides other goals.

Figure 1. A schematic illustration of the international not-for-profit collaboration in the field of bioinformatics.

Table 1. Internet resources for the some bioinformatics tools used by researchers

Name	Objective	URL
Align	Pairwise alignment algorithms	www.ebi.ac.uk/emboss/align/index.html
ArrayExpress	Microarray data	www.ebi.ac.uk/arrayexpress
BLAST	Sequence search and analysis	www.ncbi.nlm.nih.gov/BLAST
Bookshelf	Books search	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books&itool=toolbar
CDART	Conserved domain architecture retrieval	www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps
Cn3D	Viewing 3 dimensional structures	www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
DaliLite	Comparisons of protein structures	www.ebi.ac.uk/DaliLite
EcoCyc	Encyclopedia of E. coli K12 genes and metabolisms	http://ecocyc.org
Entrez Gene	Database for gene search	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ExPASy	Proteomics server	www.expasy.org
FASTA	Protein database query	www.ebi.ac.uk/fasta33/index.html
GeneDB	Access to 32 genomes	www.genedb.org
GeneQuiz	Automated analysis of biological sequences	jura.ebi.ac.uk:8765/ext-genequiz
GeneWise	Comparison of a protein sequence to a genomic sequence	www.ebi.ac.uk/Wise2/index.html
GEO	Gene expression data browsing	www.ncbi.nlm.nih.gov/geo
Map Viewer	Integrated views of chromosome maps for many organisms	www.ncbi.nlm.nih.gov/mapview/static/MVstart.html
MEDLINE	Bibliographic databases	http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+MEDLINE
ORF Finder	Sequence analysis	www.ncbi.nlm.nih.gov/gorf/gorf.html
Patent abstracts	Search for a patent abstract	http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+query+-libList+PATABS
Pfam	Collection of protein domains and families	www.sanger.ac.uk/Software/Pfam
PubMed	Journals and references search	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
SMART	Simple Modular Architecture Tool	http://smart.embl-heidelberg.de/
Spidey	mRNA-to-single genomic sequence alignment	www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey
TAIR	The Arabidopsis information source	www.arabidopsis.org
Taxonomy	Taxonomy browser	www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html
UniProt	Universal Protein Resource	www.ebi.uniprot.org/index.shtml
VAST	Structure-structure similarity search	www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
VecScreen	Screening for contaminating vector sequences	www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html
WU-Blast2 ASD	Nucleotide database query	www.ebi.ac.uk/blast2/asd.html
WU-Blast2-Parasite	Parasite genomes database query	www.ebi.ac.uk/blast2/parasites.html

A significant application of bioinformatics is the discovery of mutation-induced diseases such as cancer diseases. Fishing out the mutant gene by means of bioinformatics tools will direct researchers to the gene of interest. Tumor protein P53 functions as a tumor suppressor and mutants of its gene occur in a number of human cancers where the mutation causes the loss of tumor suppressor activity. Information about the gene sequence and the protein sequence and all related published references, sequences, phenotypes, pathways and related links are available free of charge online at the NCBI databases (www.ncbi.nlm.nih.gov) for those interested. Another essential part of NCBI is the Online Mendelian Inheritance in Man (OMIM; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omim), which is a reference to human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins.

The Institute of Medical Genetics in Cardiff provides the human research community with its electronic gene mutation database; the Human Gene Mutation Database (HGMD; www.hgmd.cf.ac.uk). Another mutation database is the Nucleotide Variation and Mutation Database that is available online from www.mutationdiscovery.com. This site includes entries of thousand of genes specific to humans, mice, rats and other organisms. A researcher can view the sequence of a certain gene such as BRCA1 (breast cancer gene) and explore mutation sites along the sequence, which will facilitate disease diagnosis using molecular biology means. Variation among species that took place through evolution has emerged mainly due to Single Nucleotide Polymorphisms (SNPs). SNPs are inherited changes of one nucleotide to another. Genetic variation is of major interest to the research community due to its importance in studying speciation and biodiversity of organisms. Most SNPs are neutral and are used in DNA fingerprinting, forensic medicine and paternity tests. Recently, molecular systematic and phylogeny depend to a great deal on SNPs in DNA sequences to produce genetic trees that in turn support species trees and explore the relationship among species. Bioinformatics has demonstrated its capabilities in studying molecular biodiversity and vast number of computer packages are now available such as PHYLIP (available online at http://evolution.genetics.washington.edu/phylip.htmland free of charge) and PAUP (available on line at http://paup.csit.fsu.edu/but commercial but in a reasonable price).

Bioinformatics is going further and it is in the present time a major part in drug design. Bioinformatics and Drug Design Group (BIDD; http://bidd.nus.edu.sg/group/about.htm) offers the research community with Therapeutic Target Databases (TTD) that in turn provide information about known and explored therapeutic proteins and nuclear acid targets, the targeted disease, pathway information and corresponding drugs and ligands. Links to other related databases are also available making surfacing related databases an easy job for a specialist. Bridging the gap between bioinformatics and cheminformatics is an essential move for drug design. Protein structures can bridge the gap and make it possible to identify new drugs. Biopendium is a technology that compiles information about protein structure, sequences and ligands is available from Inpharmatica Ltd. (London). This technology can annotate sequences and relate them to a certain protein structure, even if it is distantly related. Bio-information obtained through Biopendium is integrated in Chematica (Inpharmatica's cheminformatics database) providing mechanisms for specifying new promising drugs. Another example of integrating bioinformatics and cheminformatics is the DrugBank Database (http://redpoll.pharmacy.ualberta.ca/drugbank), the database contains 4300 drug entries and more than 6000 protein sequences which are linked to these drug entries. Drugcard entries contain drug/chemical data and drug/target or protein data. The DrugBank database is a project that is supported by Genome Alberta and Genome Canada which is a not-for-profit organization.

Bioinformatics tools and databases are used in a condensed way to engineer plants and microbes. Recently, plants and microorganisms are used in treating soil and water pollution in what is called phytoremediation and bioremediation; for more information see [8] and [9]. The attainability of biological data will enable researchers in genetically modifying these organisms to increase their performance and competency in breaking or absorbing and breaking pollutants such as heavy metals or hydrocarbonic compounds. The storage and analysis of biological data will also enable the study of the biological diversity of these organisms and their evolution on the gene level which will make it an essential biological database for the protection of environment from pollution.

NASA's experts are developing bioinformatics environment in their mission to explore space and study origins of life in the universe. NASA is therefore focusing on information technology, biotechnology, nanotechnology and space biology. Ames research center was established by NASA's predecessor the National Advisory Committee of Aeronautics (NACA). The center is pursuing NASA's missions in science biology and the development of protein-based nanotubes, a crossover technology potentially capable of self-organization and replication. Ames conducts research and develops technologies biologically-inspired and employed within basic biological processes, including biomimetics, bioinformatics, and space genomic/protonomic systems that enable in situ character studies of genetic materials and proteins in space and extraterrestrial environments (www.sti.nasa.gov/tto/spinoff2001/ames.html).

Bioinformatics tools are available online and also as computer packages to navigate through published genomes such as the model plant *Arabidopsis thaliana* (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome prj&cmd=Retrieve&dopt=Overview&list_uids=9506), *Fugu rubripes* (http://fugu.biology.qmul.ac.uk/; http://genome.jgi-psf.org/fugu6/fugu6.home.html), *Drosophila melanogaster* (www.fruitfly.org/).

5. Bioinformatics education and ethics

Advances in different domains of science are accelerating and this will create a knowledge gap if no measures are taken in relation to education and awareness. The accelerated quantitative and qualitative biological data generated by the HGP and other genome projects has multiplied public awareness of genetics and biotechnology. An effective approach to reach out for the broader sector of society is through high schools students who have the maturity and scientific background to understand genomics and biotechnology [10]. Science teachers, genome researchers, ethicists, genetic counselors and business partners have to have joint efforts for improvement of genome education programmes. Bioinformatics, an essential part of any genome project, is urged to be included in any education programme and it can be an attractive choice for a large percentage of students that are already drawn into the field of informatics. An MSc Bioinformatics project is currently running in our lab aiming to establish a database that concerns plant physiologists (Al-Rawas and Jawdat, unpublished). This project is the first to be conducted in Syria and will open doors for applying bioinformatics.

Concerns about bioinformatics are heightened among the public in the wake of exploiting large genome projects particularly the HGP. The Ethical, Legal and Social issues research programme (ELSI) is considered the world's largest bioethics program. The programme is jointly funded by the US Department of Energy (DOE) and the National Institutes of Health (NIH). The programme is concerned about safeguarding the privacy of individuals and groups who contribute DNA samples for large-scale sequence-variation studies. The ELSI research goals are: examining issues surrounding the completion of the human DNA sequence and the study of human genetic variation, examining issues raised by the integration of genetic technologies and information into health care and public health activities, examine issues raised by the integration of knowledge about genomics and gene-environment interactions in non-clinical settings, exploring how new genetic knowledge may interact with a variety of philosophical, theological, and ethical perspectives and exploring how racial, ethnic, and socioeconomic factors affect the use, understanding, and interpretation of genetic information; the use of genetic services; and the development of policy [9].

Bioinformatics can be an extremely beneficial tool for discovering diseases and introducing new human medicines. However, with advances in the technology of information, guidelines must be addressed to strike a balance between the need for authorized access to personal information and the need to prevent unauthorized access that respects the privacy of personal patients data (www.amaassn.org/ama/pub/ category/4375.html).

6. References

[1] M. H. Maurer, The path to enlightment: Making sense of genomic and proteomics information, *Genomics and Proteomics Bioinformatics*, 2004, 2, 123-131.

[2] E. M. Southern, Detection of specific sequences among DNA fragments separated by gel electrophoresis, *Journal of Molecular Biology,* 1975, 98,503-517.

[3] F. C. Bernstein, T. F. Koetzle, G. J. P. Williams, E. F. Meyer, M. D. Brice, J. R, Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *Journal of Molecular Biology,* 1977, 112, 535-542.

[4] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science*, 1988, 239,487.

[5] Y. Kohara, K. Akiyama, and K. Isono, The physical map of the whole E. coli chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library, *Cell*, 1987, 50, 495-508.

[6] D. Sulakhe, A. Rodriguez, M. D'Souza, M. Wilde, V. Nefedova, I. Foster, and N. Maltsev, GNARE: automated system for high-throughput genome analysis with grid computational backend, J. Clin. Monit. Compute.,2005, 19, 361-369.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tools, *Journal of Molecular Biology*, 1990, 215, 403-410.

[8] D. Y. Kim, L. Bovet, S. Kushnir, E.W. Noh, E. Martinoia, and L. Lee, AtATM3 is involved in heavy metal resistance in Arabidopsis, *Plant Physiology*, 2006, E-print.

[9] D. R. Lovley, Cleaning up with genomics: applying molecular biology to bioremediation, *Nat. reviews in Microbiology*, 2003, 1, 35-44.

[10] M. Munn, P. O. Skinner, L. Conn, H. G. Horsma, and P. Gregory, The involvement of genome researchers in high school science education *Genome Research*, 1999, 9, 597-607.