HSPI The Open Access Publisher

J F S R JOURNAL OF FORENSIC SCIENCE AND RESEARCH
ISSN 2575-0186

Research Article

Bioinformatics as a modern tool in forensic science for data understanding & investigation in research

Pranav Kumar Ray*

Guest Speaker, Jharkhand Police Academy, Forensic Investigator, In-Charge Forensic Science Laboratory, Jharkhand Raksha Shakti University, Ranchi, India

More Information

***Address for Correspondence:** Pranav Kumar Ray, Guest Speaker, Jharkhand Police Academy, Forensic Investigator, In-Charge Forensic Science Laboratory, Jharkhand Raksha Shakti University, Ranchi, India,
Email: Jharkhand.pk932026@gmail.com

Submitted: December 02, 2022
Approved: December 07, 2022
Published: December 08, 2022

How to cite this article: Ray PK, Bioinformatics as a modern tool in forensic science for data understanding & investigation in research. J Forensic Sci Res. 2022; 6: 083-087.

DOI: 10.29328/journal.jfsr.1001040

Abstract

Modern-day biology is witnessing a data explosion with a vast amount of information generated from ongoing genome and sequencing projects. The abundance of data from genome sequences, functional genomics and another high throughput (HTP) technique with the potential of computing has led to rising of a new discipline namely 'bioinformatics'.

Bioinformatics is a young but fast-growing field for biological data collection, organization, interpretation, and modeling. Tools and techniques for bioinformatics are derived from multidisciplinary combinations of varied disciplines from natural and physical sciences. Previously various disciplines were carved out as and when sufficient specialization was achieved. However, now bioinformatics is borne out of an alliance between existing disciplines from life and non-life. Bioinformatics encompasses new foundations for the collection, organization, and mining of gene/ protein sequences, three-dimensional structures, and biochemical functions, for modeling biological processes of functioning cells. DNA sequencing performed on an industrial scale has produced a vast amount of data to analyze. Although the Human Genome Project is officially over, improvements in DNA sequencing continue to be made. The field of forensic science is increasingly based on biomolecular data and many European countries are establishing forensic databases to store DNA profiles of crime scenes of known offenders and apply DNA testing.

Introduction

Paulien Hogeweg and Ben Hesper coined the term 'Bioinformatics' in 1978 referring to the study of information processes in biological systems. As an interdisciplinary field bioinformatics draws contributions from biology, chemistry, mathematics, statistics, and computer science; to understand life and its processes. With the emergence of disciplines such as genetics, biochemistry, molecular biology, and structural biology, the focus of the study of 'life' shifted from the 'macro' properties to 'micro' properties. Bioinformatics and forensic DNA are inherently interdisciplinary and draw their techniques from statistics and computer science bringing them to bear on problems in biology and law.

The National Centre for Biotechnology Information (NCBI 2001) defines bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics [1-7]: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."

History

Computers have become essential in molecular biology time since protein sequences have become available. The first bioinformatics databases were constructed a few years after the first protein sequence became available. The first protein sequence reported was bovine insulin after the groundbreaking work of Frederick Sanger in 1956. Early contributions to bioinformatics embrace comprehensive volumes of antibody sequences released in the works of Elvin A. Kabat in 1970. During the journey from the discovery of DNA to be the source of genetic information and elucidation of double-

helical arrangement of DNA molecules to the elucidation of human genome sequence and thereafter, bioinformatics has become an integral part of modern biology. Foundations of bioinformatics were laid in a breakthrough work by Margaret Oakley Day Hoff appropriately regarded as the 'father of bioinformatics'. A pioneer in the field of bioinformatics' Day Hoff assembled all sequence data information available to create the first bioinformatics database. Day Hoff compiled one of the first protein sequence databases initially published as 'Atlas of Protein Sequence and Structure in the year 1965. Margaret Oakley Day Hoff pioneered methods of sequence alignment and molecular evolution. Among the significant contributions of Day Hoff is the establishment of a one-letter code for the amino acids. Research in the 80s and early 90s focused primarily on the development of value-added derived databases to understand the 'sequence-structure-function relationship [6-10].

Chronological developments in bioinformatics

- **1902:** Emil Hermann Fischer wins the Nobel Prize for showing that amino acids are linked and form proteins.
- **1911:** Pheobus Aaron Theodore Levene discovers RNA.
- **1933:** Electrophoresis technique for separating proteins in solution introduced by Tiselius.
- **1941:** George Beadle and Edward Tatum identify that genes make proteins.
- **1943:** first true general-purpose electronic computer (ENIAC) was constructed at the University of Pennsylvania between 1943 and 1946.
- **1950:** Edwin Chargaff finds base pairing rule for cytosine with guanine and adenine with thymine.
- **1951:** First compiler developed by Grace Murray Hopper. Hopper developed the A-0 for the UNIVAC I. She also helped create the COBOL programming language.
- **1952:** Linus Pauling and Robert Corey propose α-helix and β-sheet protein structure.
- **1953:** Watson & Crick proposed the double helix structure for DNA based on X-ray crystallographic data obtained by Franklin & Wilkins.
- **1954:** Perutz's group develops heavy atom methods to solve the phase problem in protein crystallography.
- **1955:** Frederick Sanger analyzed the sequence of the first protein bovine insulin.
- **1958:** First integrated circuit constructed by Jack Kilby at Texas Instruments. Advanced Research Projects Agency (ARPA) was formed in the US.
- **1962:** Pauling gave the theory of molecular evolution.
- **1965:** Margaret Day Hoff's Atlas of Protein Sequences published.
- **1966:** First bioinformatics system: Margaret Oakley Day Hoff created the first protein sequence database and came up with the PAM model of protein evolution.
- **1968:** Packet-switching network protocols are presented to ARPA.
- **1970:** Details of Needleman-Wunsch algorithm for sequence comparison published.
- **1971:** E-mail program invented by Ray Tomlinson.
- **1972:** first recombinant DNA molecule was created by Paul Berg, Herbert Boyer, and Stanley N. Cohen.
- **1973:** Brookhaven Protein Data Bank announced. Robert Metcalfe from Harvard University describes 'Ethernet' in his Doctoral thesis.
- **1974:** Vinton Gray 'Vinton' Cerf and Robert Elliot Kahn developed the concept of connecting networks of computers into an 'internet' and develop Transmission Control Protocol/Internet protocol; TCP/IP. Specification of Internet Transmission Control Program by Vinton Cerf, Yogen Dalal, and Carl Sunshine, Network Working Group contains the first use of the term internet, as shorthand for internetworking.
- **1975:** Microsoft Corporation is founded by Bill Gates and Paul Allen. Two-dimensional electrophoresis for the separation of proteins on SDS -PAGE is combined with separation according to isoelectric points by P. H. O'Farrell.
- **1976:** Unix-to-Unix Copy Protocol developed at Bell Labs. E. M. Southern published details of the Southern Blot technique of specific sequences of DNA.
- **1977:** Allan Maxam and Walter Gilbert; Frederick Sanger reports methods for DNA sequencing.
- **1980:** Complete gene sequence of the first organism, a single-stranded bacteriophage φX174 published. Multi-dimensional NMR for protein structure determination described by Wuthrich et. al. Genetics Suite of programs for DNA and protein sequence analysis developed.
- **1981:** Smith-Waterman algorithm for sequence alignment is published. IBM introduces its Personal Computer.
- **1982:** Genetics Computer Group (GCG), created as a part of the University of Wisconsin, of Wisconsin Biotechnology Center. Gen Bank Released.
- **1983:** Production of DNA clone (cosmid) libraries by Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL).
- **1984:** Jon Postel's Domain Name System placed online. Macintosh was announced by Apple Computer.
- **1985:** FASTP/FASTN algorithm published. 'Genomics'

coined by Thomas Roderick appears for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes. SWISS-PROT database created by the Department of Medical Biochemistry, University of Geneva and European Molecular Biology Laboratory EMBL. PCR reaction is described by Kary Mullis and co-workers.

- **1986:** Automated sequencing technique by Leroy Hood.
- **1987:** Use of YAC's yeast artificial chromosomes described by David T. Burke and coworkers. A physical map of *E. coli* is published by Y. Kohara and coworkers. PERL - Practical Extraction Report Language released by Larry Wall.
- **1988:** National Centre for Biotechnology Information, NCBI created at NIH/NLM EMB net network for database distribution.
- **1989:** The FASTA algorithm for sequence comparison is published by Pearson and Lipman. Telomere sequence having implications for aging and cancer research is identified at LANL. Human Genome Initiative is started.
- **1990:** BLAST program is implemented. InforMax is founded with the company's products that address sequence analysis, database and data management, searching, publication graphics, clone construction, mapping, and primer design.
- **1991:** CERN research institute in Geneva announces the creation of the protocols which constitute the World Wide Web. Linus Torvalds announces a Unix-Like operating system which later becomes Linux creation. Use of expressed sequence tags ESTs described. Human chromosome mapping data repository, Genome Database GDB is established.
- **1992:** Low-resolution genetic linkage map of entire human genome published. Guidelines for data release and resource sharing announced by DOE and NIH.
- **1993:** International IMAGE Consortium was established to coordinate efficient mapping and sequencing of gene-representing cDNAs.
- **1994:** Netscape Communications Corporation founded; releases a commercial version of NCSA's Mozilla. PRINTS database of protein motifs is published by Attwood and Beck. EMBL-EBI European Bioinformatics Institute was established, in Hinxton, UK. Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.
- **1995:** Microsoft releases version 1.0 of Internet Explorer. Sun releases version 1.0 of Java. Sun and Netscape released version 1.0 of JavaScript. First non-viral whole genome sequenced for the bacterium Haemophilus influenzae. The sequence of the smallest bacterium, Mycoplasma genitalium, completed; provides a model of the minimum number of genes needed for independent existence. Physical map with over 15,000 STS markers published.
- **1996:** Saccharomyces cerevisiae genome sequence completed. PROSITE database is reported by Bairoch et.al. Affymetrix produces the first commercial DNA chips. The sequence of the human T-cell receptor region is completed. Archaebacteria- Methanococcus jannaschii genome sequenced; confirms the existence of the third major branch of life on earth.
- **1997:** Genome for E. coli published.
- **1998:** Genomes of Caenorhabitis elegans and baker's yeast are published. The Swiss Institute of Bioinformatics is established as a non-profit foundation. Craig Venter forms Celera Genomics in Rockville, Maryland.
- **1999:** First Human chromosome 22 completely sequenced.
- **2000:** Pseudomonas aeruginosa genome published. Arabidopsis thaliana genome sequenced. Drosophila melanogaster genome sequenced. International research consortium publishes chromosome 21 genome, the smallest human chromosome and the second to be completely sequenced.
- **2001:** Human genome published. Human Chromosome 20 completely sequenced.
- **2002:** genome sequence of the common house mouse 2.5 Gb published.
- **2003:** Human Genome Project completed.
- **2004:** Rattus norvegicus Brown Norway laboratory rat draft genome sequence completed.

Importance of bioinformatics

Understanding genetic diversity: Genetic diversity is the total number of genetic characteristics in the genetic makeup of a species, it ranges widely from the number of species to differences within species and can be attributed to the span of survival for a species.

Epidemiology: Epidemiology is the study (scientific, systematic, and data-driven) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health-related states and events (not just diseases) in specified populations (neighborhood, school, city, state, country, global) [11-17].

Vaccinology: Vaccinology is a field of microbiology and immunology covering vaccine development as well as the use of vaccines and their effects on animal health and public health. Developing vaccines is central to the control of infectious diseases in animals and new vaccines have the potential to reduce antibiotic use, prevent losses in livestock production and protect people from zoonotic infections.

Global health: Global health is the health of populations in the global context; it has been defined as "the area of study, research and practice that places a priority on improving health and achieving equity in health for all people worldwide".

Metabolic reconstruction: A metabolic reconstruction provides a highly mathematical, structured platform on which to understand the systems biology of metabolic pathways within an organism. The integration of biochemical metabolic pathways with rapidly available, annotated genome sequences has developed what are called genome-scale metabolic models.

Systems biology

Systems biology is the computational and mathematical analysis and modeling of complex biological systems.

- **Personalized medicine:** Personalized medicine, also referred to as precision medicine, is a medical model that separates people into different groups—with medical decisions, practices, interventions, and/or products being tailored to the individual patient based on their predicted response or risk of disease.

Fields related to bioinformatics

Computational biology: Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, ecological, behavioral, and social systems. The field is broadly defined and includes foundations in biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science, and evolution.

Genomics: Genomics is any attempt to analyze or compare the entire genetic complement of a species.

Proteomics: Proteomics is concerned with: "Qualitative and quantitative studies of gene expression at the level of the functional proteins themselves" that is: "an interface between protein biochemistry and molecular biology".

Pharmacogenomics: Pharmacogenomics is the application of genomic approaches and technologies to the identification of drug targets.

Pharmacogenetics: Pharmacogenetics is a subset of pharmacogenomics that uses genomic/bioinformatic methods to identify genomic correlates, for example, SNPs (Single Nucleotide Polymorphisms), characteristic of particular patient response profiles and use those markers to inform the administration and development of therapies.

Cheminformatics: "The combination of chemical synthesis, biological screening, and data-mining approaches used to guide drug discovery and development".

Medical informatics: "Biomedical Informatics is an emerging discipline that has been defined as the study, invention, and implementation of structures and algorithms to improve communication, understanding, and management of medical information." Medical informatics is more concerned with structures and algorithms for the manipulation of medical data, rather than with the data itself.

Uses of bioinformatics

- Store/retrieve biological information (databases)
- Retrieve/compare gene sequences
- Predict the function of unknown genes/proteins
- Search for previously known functions of a gene
- Compare data with other researchers
- Compile/distribute data for other researchers.

Bioinformatics in forensic science & from a research perspective is a very important and upgrowing field for the forensic division to understand biological things in a standard way with the help of technology and it is basically the implementation of the application of bioinformatics, which means biological science + computational knowledge, it may be utilized in biotechnology also for different purposes, the application of bioinformatics in forensic means the application of biological science & computational knowledge for the purpose of investigation or for investigative research.

DNA testing According to Forensic bioinformatics basic task is to make advancements in setting up forensic records which is useful to store a rough draft of the DNA of criminals that are taken from the crime scene and later presented for DNA testing (Ajay, et al. 2012). Statistical and technological progressions i.e., learning algorithms based on machine learning, DNA microarray sequencing, thin film transistor biosensors, etc. are used to improve the accuracy and authenticity of the results. Nowadays genetic tests have been extensively used for the detection of mass fatality and forensic evidence as well. A multidisciplinary panel including medical examiners, fingerprint professionals, and forensic pathologists gathers the data which is then incorporated with the results of genetic testing.

Result and discussion

Bioinformatics tools are very helpful in forensics but there is still a need to be more careful while generating results from computational tools because at times there are discrepancies arise between a set of statistical rules and biological reactions. As the most doubtful results were produced in phylogeny reconstructions and Cluster W reconstructed alignment. It is also observed that correct alignments are generated from those sequences which are very closely related with the help of the bootstrap method. At the same time, it is expected that the alignments which are produced from biological sequence sets produced inaccuracy in more than half of the alignments so such a method is used to determine the constancy of tree

topology but not give an accurate phylogenetic tree. But with the passage of time, there is an improvement in results and the computational programs are becoming more consistent progressively. Parentage testing and family reunification are also something that comes under the category of bioinformatics and forensics.

Though it's very useful many people condemn such test as it interrupts their privacy. In the last 20 years, the field of bioinformatics has become more advanced and the objective of production, as well as the assemblage of various documentation and investigative tools, has been accomplished. Worldwide, public realm assets such as Gen Bank have become a very crucial source for research purposes. Prasad (2008).

Currently, the lives of millions of people globally are influenced by forensic DNA technology. This approach is still getting a high rate of approval on a universal level. Forensics played well in major events like in 9/11 activist assault, and the victims were recognized through DNA profile analysis.

Nowadays Forensic DNA databases fast expansion put many questions on the standard of data related to placing and its maintenance, uncertainties related to its effectiveness and there are also chances of confidentiality violation of such huge private data Ge, et al. (2014). On the other hand, in earlier periods various types of transgressions were put under DNA investigation and as a result, numerous DNA profiles were produced which become helpful to generate novel measures i.e. in Familial DNA Database Searching, finding similarities between DNA profiles of executors' family member and the data collected from the crime scene and the first victorious familial search was carried out in 2004 in UK that confirm Craig Harman is responsible for assassination but many countries are against to use this type of facts i.e. according to Germany viewpoint, each autonomous society needs to enjoy the freedom and constitutional rights that's why the expansion of forensic database is discouraged Wallace, et al. (2014). There is a deficiency of funds, professionals, and data protection and also, and there is insufficient guidance as well as improper apparatus.

- University of Veterinary and Animal Sciences Lahore (UVAS),
- Government College University Lahore (GCU) and
- University of Punjab (PU)

Initiated DNA forensics research. Center of Excellence in Molecular Biology (CEMB) is a committed laboratory started in 2005 and it deals with cases including crimes, catastrophes, and paternity clashes. Higher Education Commission (HEC) should focus on this field also by acquiring advanced strategy in association with law enforcement institute facilitates forensics.

References

1. Christofidis G. Bioinformatics Role in Forensic DNA. AZO Life Sciences. https://www.azolifesciences.com/article/Bioinformatics-Role-in-Forensic-DNA.aspx#:~:text=The%20current%20state%20of%20bioinformatics,exome%20Aggregation%20Databases%20and%20ExAC
2. Buffalo V. Book "Bioinformatics Data Skills".
3. Choudhuri S. Book "Bioinformatics for Beginners".
4. Bianchi L, Liò P. Forensic DNA and bioinformatics. Brief Bioinform. 2007 Mar;8(2):117-28. doi: 10.1093/bib/bbm006. Epub 2007 Mar 24. PMID: 17384432.
5. Bioinformatics as forensic tool in coronavirus outbreak. https://genomealberta.ca/genomics/bioinformatics-as-forensic-tool-in-coronavirus-outbreak.aspx.
6. Forensic Bioinformatics. http://www.bioforensics.com/.
7. Neelakanta PS. Textbook of Bioinformatics.
8. Selzer PM, Koch RJM. Applied Bioinformatics: An Introduction.
9. An Introduction to Applied Bioinformatics (1st edition). https://github.com/applied-bioinformatics/An-Introduction-To-Applied-Bioinformatics
10. Bioinformatics, AI & Big Data. https://dbtindia.gov.in/schemes-programmes/research-development/knowledge-generation-discovery-research-new-tools-and-0.
11. https://www.ibioinformatics.org/.
12. http://www.scfbio-iitd.res.in/oldwebsite/scfbiogroup/bioinformaticsindia.htm.
13. https://www.ibab.ac.in/.
14. Handbook of mathematics of bioinformatics.
15. Lopes HS, Cruz LM. Computational Biology and Applied Bioinformatics.
16. Thampi SM. Introduction to Bioinformatics.
17. Gerstein M. Bioinformatics Introduction.