PUBLIC HEALTH AND BIOINFORMATICS

Asangansi IE and Aiyetan PO

Final year medical students

University of Ibadan College of Medicine

Ibadan, Nigeria

"*Information is not knowledge*".

Albert Einstein (1879 – 1955)

Citation:

Asangansi I E, Aiyetan P O, Public Health and Bioinformatics. Dokita March 2005 Vol 30 (1) p24

ABSTRACT

Bioinformatics is simply *the application of information science to biology and health*. This paper briefly describes what bioinformatics is, its basic principles and discusses its relationship with public health.

DEFINITIONS

Bioinfomatics is the application of the science of informatics to biological problems. *Informatics* is the application of information science, computer science, mathematics and their associated technologies to a discipline or a domain[1]. Roughly, bioinformatics describes any use of a computer to handle biological information. In practice, however, the definition used by most people is narrower, thought to be a synonym to *computational molecular biology* - the use of computers to characterize the molecular components of living things.

According to Fredj Tekaia at the Institut Pasteur, bioinformatics is: "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."[2]

Most biologists talk about "doing bioinformatics" when they use *computers* to *store*, *retrieve*, *analyze* or *predict* the *composition* or the *structure* of biomolecules. *Simulate* could probably be added to this list of bioinformatics verbs. "Biomolecules" include genetic material - nucleic acids - and the products of your genes, proteins. Thus, the concern of "*classical*" bioinformatics is dealing primarily with sequence analysis.

Other Fields within Bioinformatics

There are other fields - for example medical imaging and image analysis - that are considered part of bioinformatics. There are also other disciplines of biologically inspired computation that includes genetic algorithms, artificial intelligence (AI) and neural networks. Often, these areas interact in strange ways. Neural networks, inspired by crude models of the functioning of nerve cells in the brain, are used to predict, surprisingly accurately, the secondary structure of proteins from their primary sequences.

What almost all fields of bioinformatics have in common is the processing of large amounts of biologically derived information, whether DNA sequences or breast X-rays. For example, computerization of the records at the University College Hospital (UCH) will be considered an endeavour in bioinformatics.

“Richard Durbin", Head of Informatics at the Wellcome Trust Sanger Institute, expressed an interesting opinion: "I do not think all biological computing is bioinformatics, e.g. mathematical modeling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, **particular genetic information**."[3]

In summary, the National Center for Biotechnology Information, USA (NCBI 2001)[4] defines bioinformatics as: *"… the field of science in which biology, computer science, and information technology merge into a single discipline. It is composed of three important sub-disciplines:*

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

Information dissemination, even in the form of myths and legends has been a vital tool in the survival of traditions from which present civilizations are born. This includes knowledge as a way of life. Present understandings of particular subjects are a culmination of many notable experiences that grew into a body of knowledge.

Over a century ago, bioinformatics history started with an Austrian monk named Gregor Mendel. He is known as the "Father of Genetics". He cross-fertilized different colors of the same species of flowers. He kept **careful records** of the colors of flowers that he cross-fertilized and the color(s) of flowers they produced. Mendel illustrated that the inheritance of traits could be more easily explained if it was controlled by factors passed down from generation to generation.

Since Mendel, bioinformatics has come a long way. The understanding of genetics and biology as a whole has advanced remarkably in the last thirty years. Computers have been introduced and data handling can be said to be revolutionized.

In 1988, the Human Genome Organization (HUGO), an international organization of scientists involved in the Human Genome Project, was founded. In 1989, the first complete genome map was published of the bacteria Haemophilus influenza. Today, the entire human genome (of about 3000 million base pairs) and numerous other genomes are available free of charge.[5]

Bioinformatics was fuelled by the need to create huge databases, such as **GenBank, EMBL** (the European Molecular Biology Laboratory bank) and the **DNA Database of Japan** where the human genome is stored and compared with other genome sequencing projects. Today, bioinformatics embraces protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways.[5]

PUBLIC DATABASES

A database (or data base) is a collection of data that is organized so that its contents can easily be accessed, managed, and modified by a computer. The most prevalent type of database is the relational database, which organizes the data in tables; multiple relations can be mathematically defined between

the rows and columns of each table to yield the desired information. An object-oriented database stores data in the form of objects, which are organized in hierarchical classes that may inherit properties from classes higher in the tree structure.

Biological raw data are stored in public databanks (such as Genbank or EMBL for primary DNA sequences). The data can be submitted and accessed via the World Wide Web. Protein sequence databanks like trEMBL provide the most likely translation of all coding sequences in the EMBL databank. Sequence data are prominent, but also other data are stored, e.g. yeast two-hybrid screens, expression arrays, systematic gene-knock-out experiments, and metabolic pathways. The stored data need to be accessed in a meaningful way, and often contents of several databanks or databases have to be accessed simultaneously and correlated with each other. This stems from the fact that ***genome sequencing risks becoming expensive molecular stamp collecting without the tools to mine the data and fuel hypothesis-driven laboratory-based research.***[6]

Special languages have been developed to facilitate this task (such as the Sequence Retrieval System **(SRS)** and the Entrez system). An unsolved problem is the optimal design of inter-operating database systems. Databases provide additional functionality such as access to sequence homology searches and links to other databases and analysis results. For example, **SWISSPROT** contains verified protein sequences and more annotations describing the function of a protein. Protein 3D structures are stored in specific databases e.g. the **Protein Data Bank**, now primarily curated and developed by the Research Collaboratory for Structural Bioinformatics. Organism-specific databases have been developed such as **ACEDB (**the **A C. Elegans DataBase)** for the *C. elegans* genome, **FLYBASE** for *D. melanogaster* and **PlasmoDB** for *Plasmodium*. [7]

Major problems are errors in databanks and databases (mostly errors in annotation), in particular since errors propagate easily through links. Also databases of scientific literature (such as **PUBMED, MEDLINE**) provide additional functionality, e.g. they can search for similar articles based on word-usage analysis. Text recognition systems are being developed that extract automatically knowledge about protein function from the abstracts of scientific articles, notably on protein-protein interactions.

BIOINFORMATICS IN PUBLIC HEALTH

The four major areas where bioinformatics finds application in public health issues include **drug design, vaccine creation, disease surveillance and national planning.**

Drug targets

Many aspects of bioinformatics are relevant for pharmacology. Drug targets in infectious organisms can be revealed by whole genome comparisons of infectious and non-infectious organisms. The analysis of single nucleotide polymorphisms reveals genes potentially responsible for genetic diseases. Prediction and analysis of protein 3D structure is used to develop drugs and understand drug resistance.[8]

Patient databases with genetic profiles, e.g. for cardiovascular diseases, diabetes, cancer, etc. may play an important role in the future for individual health care, by integrating personal genetic profile into diagnosis, despite obvious ethical problems. The goal is to analyze a patient's individual genetic profile and compare it with a collection of reference profiles and other related information. This may improve individual diagnosis, prophylaxis, and therapy.

Vaccine Development

The conventional approach to vaccine development is based on dissection of the pathogen using biochemical, immunological and microbiological methods. Although successful in several cases, this approach has failed to provide a solution to prevent several major bacterial infections. The availability of complete genome sequences in combination with novel advanced technologies, **microarrays** and **proteomics**, have revolutionized the approach to vaccine development and provided a new impulse to microbial research. The genomic revolution allows the design of vaccines starting from the prediction of all antigens *in silico*, independently of their abundance and without the need to grow the pathogen in vitro. This new genome-based approach, which is named **"Reverse Vaccinology",** has been successfully applied for Neisseria meningitidis serogroup B for which conventional strategies had failed to provide an efficacious vaccine. The concept of "Reverse Vaccinology" can be easily applied

to all the pathogens for which vaccines are not yet available and can be extended to parasites and viruses.[9, 10, 11, 12]

National health statistics and planning

Bioinformatics is involved in creating computer programs that can receive, process and interpret data. Sources of these data include hospitals, public surveys, and birth and death statistics. Accurate analysis of such a verse array of information no doubt derives better relationships between cause and effect. Bioinformatics makes available in shorter time adequate information to implement appropriate health intervention methods. Bioinformatics also is used to simulate and therefore predict the consequence of these interventions.

Gene therapy and Immunotherapy

Programs that involve interventions for people with genotypes that predispose to infectious diseases are already in existence: the best example is sickle cell disease. Genetically engineered (recombinant) therapies such as cytokines or chemokines, or agents targeted to their receptors, are mainly in the developmental stage. They hold great promise and may become standard parts of the treatment armamentarium, as is already the case with α-interferon and *hepatitis C*. Gene therapy (e.g., for immunodeficiencies) is becoming more common as technologies advance. New drugs that may be safer or more effective are being developed because of increased knowledge of the molecular structure of HLA and other molecules that influence host responses to pathogens. On the other hand, simple interventions such as reduction of iron-overload in persons with hereditary hemochromatosis or education of such persons about their increased risk to the severe consequences of Vibrio vulnificus infection (acquired through food, occupational, or recreational exposure) could have a great impact in reducing the burden of this and other infectious diseases. Future preventive opportunities will require the dissemination of genetic information to the public to ensure the translation of genetic knowledge into effective public health programs.[13]

THE ROLE OF PUBLIC HEALTH IN BIOINFORMATICS

Public health agencies will play an increasingly important role in determining risk factors for emerging or reemerging infectious diseases. One role is to continue to identify genes associated with risk of infectious diseases. The attributable fraction of each genotype and the interaction with other risk factors to these diseases can be determined. This can be accomplished through surveillance and applied research. Secondly, important ethical, legal and social aspects of genetic testing of persons or populations can be easily resolved when included in public health schemes. This also includes the quality of genetic testing. Through education and dissemination of information about genes and infectious diseases, persons with genes that increase the risk of certain infections may be encouraged to alter behaviors that increase their risk of infection. [13]

Public health agencies can play important roles in many aspects of the interaction between genetics and emerging infectious diseases. For example, through surveillance backed by genetic testing, public health agencies can monitor the safety of blood products (e.g., immunoglobulins or plasma derived products) used in treatment of persons with heritable immune-deficiencies.

CONCLUSION

The birth of bioinformatics, as it is, is a revolutionary adaptation that attempts to solve the problem of ***hyper-informosis*** - the presence of an overwhelming amount of information. Within a space of a few years, the human genome has been mapped but a high level of understanding would elude us for decades without bioinformatics.

The Nigerian case is peculiar. Inconstant power supply, inadequate infrastructure, insufficient manpower, lack of political will have militated against the development of bioinformatics in all its ramifications. However, bioinformatics offers opportunities in the face of these negative factors and it represents the most sensible doorway to the enormous amount of information being added every day.

ACKNOWLEDGEMENT

The authors thank Dr Happi and Mrs Folarin, both of the Institute of Advanced Medical Research and Training (IAMRAT), College of Medicine, University of Ibadan for their review.

REFERENCES

1. Elkin P L *Primer on Medical Genomics Part V: Bioinformatics Mayo Clin Proc*. 2003; 78:57-64.
1. Counsell D *UK Medical Research Council Human Genome Mapping Project Resource Centre Articles* http://www.bioinformatics.org/faq.
2. What is Computational Biology? http://www.cs.mcgill.ca/~kaleigh/compbio/whatis_compbio.html
3. NCBI Science Primer: Bioinformatics http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html
4. http://www.geocities.com/bioinformaticsweb/definition.html
5. Pallen M J *Microbial genomes. Mol. Microbiology.1999; 32:907-912*
6. Michael Nilges and Jens P. Linge, Unité de Bio–Informatique Structurale, Institut Pasteur, 25–28 rue du Docteur Roux, F–75015 Paris, France.
7. http://www.linge.de/science/reprints/bioinformatics_definition_2001.pdf
8. Capecchi B, Serruto D, Adu-Bobie J, Rappuoli R, Pizza M. *Curr Issues Mol Biol*. 2004 Jan;6(1):17-27.
9. Zagursky RJ, Olmsted SB, Russell DP, Wooters JL *Bioinformatics: how it is being used to identify bacterial vaccine candidates Expert Rev Vaccines.* 2003 Jun;2(3):417-36.
10. Mora M, Veggi D, Santini L, Pizza M, Rappuoli R *Reverse vaccinology* DDT Vol. 8, No. 10 May 2003.
11. Mäkelä P.H. *et al.* (1995) Vaccines against *Haemophilus influenzae* type B In *Molecular and Clinical Aspects of Bacterial Vaccine Development* (Ala'Aldeen, D.A.A. and Hormaeche, C.E., eds), pp. 41–91, John Wiley and Sons.
12. McNicholl J M, Hughes J M *Human Genetics and Emerging Infectious Diseases: Issues and Prevention Opportunities U.S. Medicine*, June 1998, p.4.