Bioinformatics: an overview for cancer research

M. Al-Rajab[1], **J. Lu**[1]
[1]School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom

Abstract - *Bioinformatics is a new science that is glowing out in the recent years. It is a multidisciplinary science that is made out of different kinds of other scientific fields like biology, computer science, chemistry, statistics, mathematics and others. It was a big challenge for researchers to describe this new field in a systematic scientific way and bring out the attention of its applications and services; one of these important services that Bioinformatics can be applied in, is the cancer studies, research and therapies for many beneficial reasons. This paper will give a clear glance overview of bioinformatics, its definition, aims, applications, technologies, the large amount of data produced in the biological field and how bioinformatics can organize, analyze and store them, discuss some algorithms that can be implemented over bioinformatics data, and how to apply bioinformatics to discover and diagnose diseases like cancer.*

Keywords: Bioinformatics, Applications, Technologies, Data, Algorithms, Cancer.

1 Introduction

Bioinformatics is a new multidisciplinary field that comes out from the combination of other sciences and fields like biology, computer science, statistics, chemistry, mathematics and even more [3, 6, 8, 9, 14, 15, 16, 17]. In recent years new sciences have risen up due to the demand in understanding more the world around us like Bioinformatics, Biotechnology, Computational Biology, Biochemistry and others. It was a big challenge for researchers and scientists to give an adequate definition for each of these newly emerged sciences [5, 9, 18]. One of these sciences that have a huge influence in the medical field is Bioinformatics but also can play a key role in other fields like agriculture, livestock and even space explorations [1, 19]. Bioinformatics which attracts people in the academic field in addition an interest to those in the medical industry [4, 15, 20, 21].

There were many contributions to define and explain Bioinformatics in scientific ways, but all researchers agree that it is a combination of Biology, Computer Science, Statistics and Mathematics. Each one of these disciplines is playing an important role for collecting, organizing, analyzing and digitizing the biological data and even classifying and storing it in an efficient manner [1, 3, 12, 16, 19].

The main purpose of this paper is explore and explain Bioinformatics in a more scientific way, the paper will try to define Bioinformatics scientifically and try highlight applications of bioinformatics in the medical sector specially, and in the diagnosis of critical diseases like cancer. The race of bioinformatics research is now passing long rounds in many areas in the Biological life, so; the goal of this paper is to provide an overview summary of bioinformatics definition from different articles written in this field, what are the main implementations and aims under the skin of this science, how to understand the data and what are the most important databases used, give a snapshot over the most common algorithms implemented in the field and how important to apply bioinformatics in the cancer research and study.

This paper will target four categories of readership who are interested in the field. (1) Students who are interested in studying this new field. (2) Instructors who would like to prepare a fundamental course to teach in bioinformatics. (3) Researchers who would like to understand more about Bioinformatics and the relationship with cancer. (4) Experts in the medical field who are interested in implementing the understanding of this field in the medical life.

The remainder of this paper will be structured as follows: Section 2 will discuss the background in methodologies applied in this paper; while Section 3 will focus on Bioinformatics definition, on the other hand section 4 will figure out the aim of studying the field. Moreover in section 5 data, data types and databases will be presented in Bioinformatics. On the other hand, section 6 will discuss the most common Algorithms implemented in Bioinformatics. Section 7 will discuss the role of Bioinformatics in cancer research and how important to be implemented in that field. In section 8 current problems in Bioinformatics are represented, and finally section 9 will conclude this paper.

2 Background in methodologies

As well as sufficient number of papers, articles, websites, and books are talking about Bioinformatics. It was clear to us that all have no unified definition for Bioinformatics as a science or a new born field emerging in the life of biology and technology, add to that there were rare papers systematically constructing and directing the road for all Bioinformatics basic knowledge. From this point an effort was implemented to conduct a deep search to collect as many papers and articles discussing the historical and

fundamentals of Bioinformatics in order to establish a unified basis form understanding the basics of Bioinformatics and links that with importance of applying the field in the cancer study, research and therapy. More than seventy papers, articles, websites and books that are talking about introduction in bioinformatics were collected. A profound reading took place to classify the papers. To write about the basics, we put out all the keywords (bioinformatics, database, algorithms, technologies, cancer, applications), then we started classifying the papers related to the collected data as in Table 1.

Table 1: Summery of Papers Number Read

Topics	No. of Papers
Bioinformatics Definition	49
Databases	12
Algorithms	6
Technologies and Tools	12
Applications	12
Cancer	12

To remark the numbers in the table, 49 references were introducing a definition to Bioinformatics, 12 of them talked about the databases in bioinformatics, 6 discussed the most important algorithms used in Bioinformatics, 12 mentioned out the most important technologies and tools used in the field, the same number discussed where Bioinformatics is applied, and also the same number introduced the relationship of the field with cancer. After that grouped out the data that are relevant together from the different resources and put them together for the literature review and the findings. It was noticed that the different resources collected were not focusing on a basic knowledge of Bioinformatics, they started by defining the field then highlighting one part of the field like databases, tools, applications, algorithms, etc…

Our contribution in this paper is to gather all the distributed fundamental information about Bioinformatics and summarize them in a systematic fundamental way. Jawdat [1] discussed that the storage and analysis of biological data using certain algorithms and computer software is called Bioinformatics, so it was defined as the design, construction and use of software tools to generate, store, annotate, access and analyze data and information related to molecular biology. The authors in [2] said that bioinformatics is basically a study to model, to organize, to understand and to discover interesting information associated with the large scale molecular biological databases. The term Bio (Molecular Biology) informatics (Information Technology) which encompasses tools and methods used to manage, analyze, and manipulate large set of biological data. In [3] the authors claimed that the use of bioinformatics to organize, manage, and analyze genomic data which is the genetic material of an organism, this new IT discipline fuses computing, mathematics, and biology to meet the many computational challenges in modern molecular biology and medical research. Chavan in [4] argued that biological data include extensive information regarding genomic sequences of different species, changes due to evolution, and changes in their protein sequences. Such a massive data cannot be handled with ease. This requires systematic sieving of data to categorize and catalogue them. Based on this need arose the field of Bioinformatics. So Bioinformatics can be defined as the discipline, which encompasses branches like biology, computer science, IT and mathematics. It is a science of managing and analyzing vast biological data using advanced computing techniques. On the other hand, in [5] the authors commented that defining the terms bioinformatics and computational biology in addition is not an easy task. They are both multidisciplinary fields, involving researchers from different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics. In [6] Zadeh defines bioinformatics as a new discipline that has emerged from the areas of biology, biochemistry, and computer science. Bioinformatics is an interdisciplinary and rapidly evolving field that has emerged from the fields of biology, chemistry and computer science. Add to that Kasabov in [7] said that bioinformatics is concerned with the application and the development of the methods of information sciences for the analysis, modeling and knowledge discovery of biological processes in living organisms. Furthermore in [8] the authors illustrate Bioinformatics as the combination of biology and information technology which focuses on cellular and molecular levels for application in modern biotechnology. So as a result they said that Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with computer based analysis of large biological data sets. Fenstermacher in [9] is defining Bioinformatics as a multifaceted discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology, and genetics. So Bioinformatics is conceptualizing biology in terms of macromolecules and then applying “informatics” techniques to understand and organize the information associated with these molecules, on a large scale. Moreover, Nair in [10], explained Bioinformatics to be the application of computer sciences and allied technologies to answer the questions of Biologists, about the mysteries of life. In addition the authors in [11] discussed that bioinformatics is a new and rapidly evolving discipline that has emerged from the fields of

experimental molecular biology and biochemistry, and from the artificial intelligence (AI), database, pattern recognition, and algorithms disciplines of computer science. Finally, in [12] the authors summarized the definition of bioinformatics as the application of computer technology to the management of biological information.

3 Bioinformatics Definition

The origin of bioinformatics goes back to Mendel's discovery of genetic inheritance in 1865. Since the 1953, big revolution achievements took place by James Watson and Francis Crick as they determined the structure of DNA [13]. Later in 1960s, the hard work of bioinformatics research started, symbolized by Dayhoff's atlas of protein sequences and the early modeling analysis of protein and RNA structures [12]. After a while, the term Bioinformatics came to sense and use in around 1990s and was described by the management and analysis of DNA, RNA, and protein sequence data. Later in 2000 a big achievement took place which is the announcement of the initial draft of the Human Genome Sequence. Later after 13 years of research and work from 1990 up to spring 2003, in which the official announcement of the Human Genome Sequence Project took place. In this project around 20,000 – 25,000 of human genes where discovered, so the access to this huge amount of gene data and its information was not an easy task for the biologists and for this it opened the doors for a new era in modern biology with an assistant to new computerized technology or in other words the marriage between Biology and Computer Science to bear a new baby known as Bioinformatics which will play a significant role in gathering, analyzing, classifying and storing genetic data collected from the human project or at biological points in a more efficient or powerful way. From here raised the question, what is the importance of Computers in Biology? The accurate answer of this question will be resulted out from the following formula: Biology + Computer Science = Bioinformatics. So what is Bioinformatics? What are the main problems that this field can help in?

As a result of the literature review, Bioinformatics can be defined from different perspectives, first from the English Oxford Dictionary, and then from the summary of all researchers' definitions.

Bioinformatics: (According to the Oxford English Dictionary) (Molecular) bio – informatics: bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied math, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale.

In short, bioinformatics is a management information system for molecular biology and has many practical applications. So, Bioinformatics can be defined as a new hybrid emerging field of science in which biology, computer science, mathematics, statistics and Information Technology merge and interact together to form a whole new discipline field. It is a science used to manage, analyze, organize, and classify the huge amount of biological data by using well developed algorithms, computational and statistical techniques, designing and construction of software tools and theories to solve different problems arising from biological data and help in generating, storing, accessing and analyzing data and information that are related to molecular biology. Noting that the suffix "informatics" is from European origin; "informatique" means and indicates computer science in French and Bio means Biology [13]. Figure 1 below illustrates all the sciences that make up the Bioinformatics field.

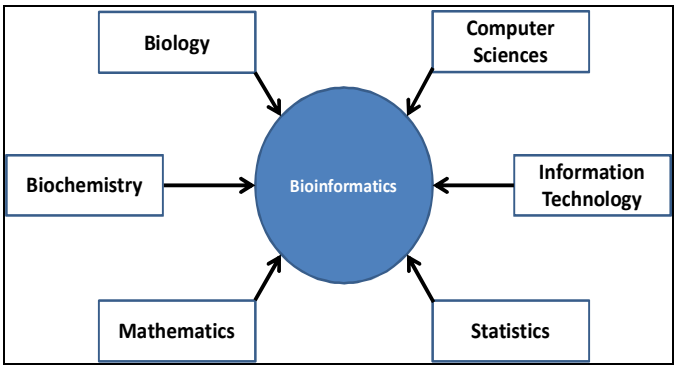

Figure 1: Bioinformatics multidisciplinary sciences

Bioinformatics has four main components: Databases, Computational Tools, Algorithms and Software. Biologists and other related people must be aware of the difference between Bioinformatics and Computational Biology and this is not an easy task, the latter is not a "field" like bioinformatics but it is "an approach" involved in using computers to study biology [9]. So, bioinformatics is concerned with information while computation biology is concerned with hypothesis [14].

4 Bioinformatics Aims

There are five main aims of Bioinformatics [12]:

1. To organize the biological data in an easy manner that helps biologists and researchers to store and access exiting information.

2. To develop and design software tools that help in the analysis and management of data.

3. To use these biological data in the analysis and interpretation of the results in a biological meaningful manner.

4. To assist researchers in the pharmaceutical industry to understand the protein structures that lead and help in the drugs industry development.

5. To help and assist physicians in the medical fields to understand gene structures that will help in detecting and diagnosing disease like cancer.

5 Biological Data, Data types and Databases

Biological Data is often characterized by huge size. There are four important data generated and collected at biological points [10]: DNA, RNA, Protein Sequences, and Micro Array images. The first 3 of them are text data and the last one is a digital image. As the different biological data generated, it can be noticed that these data is represented with different types. There are four types of the data structures [13]: String to represent DNA, RNA, and protein sequences; Trees to represent protein structures; Graphs to represent metabolic and signaling pathways; and Strings (like words and phrases) are also used to express comments that reflect meanings to researchers. Moreover, researchers and biologists are also interested in substrings, subtrees and subgraphs.

The large, huge and complex amount of biological data needed to be stored, accessed and manipulated in an efficient and powerful manner. So it was the need to build Bioinformatics databases which are classified into sequence databases, microarray databases, genome databases, protein structures databases and many more [2].

The sequence databases represent sequence information of all the organisms. GeneBank at the National Center for Biotechnology information, EMBL (European Molecular Biology Laboratory) DNA database, Bethesda and DNA Data Bank Japan (DDNJ), and Protein databases at SWISS-PORT (Protein sequence database at Swiss Institute of Bioinformatics, Geneva) all of them are the largest databanks of the sequence databases. Micro array databases include micro array gene expression under different biological conditions. Example databases of this category are Array Express, and Gene Expression mnibus. Genome databases collect organisms' gene (DNA) sequences. Example of this category databases are Xenbase, Corn, SEED, and RGD. There is another example of Bioinformatics databases that comes from the integration with cheminformatics which is the DrugBank database (http://redpoll.pharmacy.ualberta.ca/drugbank), this database contains 4300 drug entries for and more than 6000 protein sequences which are linked to these drug entries [1].

6 Common Bioinformatics Algorithms [12-13]

This section sheds the light on algorithms that are of interest to bioinformatitions and researchers. The following are some of the most important algorithmic trends in bioinformatics:

1. Finding similarities among strings (such as proteins of different organisms).

2. Detecting certain patterns within strings (such as genes).

3. Finding similarities among parts of spatial structures (such as motifs).

4. Constructing trees (called phylogenetic trees expressing the evolution of organisms whose DNA or proteins are currently known).

5. Classifying new data according to previously clustered sets of annotated data.

6. Reasoning about microarray data and the corresponding behavior of pathways.

7 Bioinformatics Applications in Cancer Research

Cancer is classified as a genetic disease in which the cells cannot follow the sequential phases of the cell cycle and divide in a normal manner. That is cells will lose the control in the cell cycle and starts to divide uncontrollably and the chromosomes of the cancer cells will be arranged incorrectly, or have large pieces missing.

Due to large and fast steps in the medical field research, a lot of efforts are extended in order to find a way to detect, diagnose and treat such hazardous disease. Also the raise of the Human Genome project discovery in 2003 had put more pressure on Bioinformatics to be applied in the cancer therapy. Bioinformatics is now being applied in the cancer research and therapy [21], and it is clear that experts and researchers have implemented rapid and expanded amount of research on the tools of bioinformatics that are considered necessary during the cancer therapies. One of these applications is to use the computerized models that represent biological data and information to know about the quantity of cancer cells in the body or about the biological state of the

patient [22]. Such way has a positive result after the cancer therapy in which experts are now being able to monitor the tumor growth that was not possible earlier during the absence of bioinformatics. In addition, many studies have indicated that gene expression of cancer cells is imperative and this will ensure efficient results after the treatment [9, 23]. Also bioinformatics can be applied to cancer by using the database among the cancer cells' expression and to study the drug response and tumor response also [23]. Until now bioinformatics studies show that it had succeeded in the cases of breast and ovarian cancer and future will insure the effectiveness of bioinformatics in the therapies of other cancer types [24]. Moreover, bioinformatics has made it possible for therapists to analyze immune responses that allow an understanding of the differences between controlled and uncontrolled tumors for better treatment of cancer patients. In other words bioinformatics succeeded in explaining out the effects of the chemotherapy and the radiation therapy with the help of the mathematical models that are part of the bioinformatics discipline. It was noticed that experts and physicians try to use the multiple databases available and the different search engines like Google in order to look for biological data and apply bioinformatics in cancer research and treatment, that due to some organizations and experts limit their work and information and do not allow other experts to benefit from the same work and information. In other words, integration of bioinformatics databases data types, and structures are an important factor to decide the future of Bioinformatics application the medical field science and especially in the cancer treatment and therapies.

The Human Genome Project has enriched the human research community with massive amount of huge biological data and information by the year 2003 [1]. In this case Bioinformatics has found its applications in many areas, and below is a list of some of the important problems where applications in Bioinformatics can be applied in[4, 10]:

• Analyzing DNA sequence data to locate genes.

• Analyzing RNA sequence data to predict their structures.

• Analyzing protein sequence data to predict their location inside the cell.

• Analyzing gene expression images.

• Understanding genetic diseases like cancer, cystic fibrosis, and sickle cell anemia.

• For gene therapy in general.

• In designing drugs for better treatment, and avoid drugs side effects and develop better drug delivery system.

Moreover, NASA's experts are using Bioinformatics in their operations to explore the space and study the universe. So, NASA is also interested in Bioinformatics in their researches and discoveries.

8 Conclusions

The paper tried to give an overview of this multidisciplinary field, by forming a unique clear definition that is introduced by the reaction of Biology and Computer Science in addition to some assessment factors like statistics and mathematics to result into the newly born field "Bioinformatics" after this strong reaction. At the end the paper highlighted the importance of applying bioinformatics in cancer research which will open the horizons for experts and researchers to continue in this specialized field. The future of Bioinformatics will be bright in many biological and life areas, but one of the important issues that must be worked in for this; is the integration of the wide and huge amount of data sources and databases to unify them for better life and for a huge revolution in the biological life as will reaching the moon.

9 References

[1] Jawdat, D.; , "The Era of Bioinformatics," Information and Communication Technologies, 2006. ICTTA '06. 2nd , vol.1, no., pp.1860-1865, 0-0 0

[2] Raut, S.A.; Sathe, S.R.; Raut, A.; , "Bioinformatics: Trends in gene expression analysis," Bioinformatics and Biomedical Technology (ICBBT), 2010 International Conference on , vol., no., pp.97-100, 16-18 April 2010

[3] See-Kiong Ng; Limsoon Wong; , "Accomplishments and challenges in bioinformatics," IT Professional , vol.6, no.1, pp. 44- 50, Jan.-Feb. 2004

[4] Dr.(Mrs.) Padma R. Chavan; , "Application of Bioinformatics in the Field of Cancer Research", 11th Workshop on Medical Informatics & CME on Biomedical Communication, vol., no., 20-22 November 2008.

[5] Ackovska, N.; Madevska-Bogdanova, A.; , "Teaching Bioinformatics to Computer Science Students," Computer as a Tool, 2005. EUROCON 2005.The International Conference on Computers as a Tool, vol.1, no., pp.811-814, 21-24 Nov. 2005

[6] Zadeh, J.; , "An undergraduate program in bioinformatics," Potentials, IEEE , vol.25, no.3, pp.43-46, July-Aug. 2006

[7] Kasabov, N.; , "Bioinformatics: a knowledge engineering approach," Intelligent Systems, 2004.

Proceedings. 2004 2nd International IEEE Conference , vol.1, no., pp. 19- 24 Vol.1, 22-24 June 2004

[8] Fulekar, M.H. and J. Sharma. 2008. “Bioinformatics Applied in Bioremediation”. Innovative Romanian Food Biotechnology. Vol. 2 No. 2. pp 28-36.

[9] David Fenstermacher, Introduction to bioinformatics: Research Articles, Journal of the American Society for Information Science and Technology, v.56 n.5, p.440-446, March 2005

[10] Achuthsankar S Nair, ; “Computational Biology & Bioinformatics – A gentle Overview”, Communications of Computer Scioety of India, January 2007.

[11] Doom, T.; Raymer, M.; Krane, D.; Garcia, O.; , "Crossing the interdisciplinary barrier: a baccalaureate computer science option in bioinformatics," Education, IEEE Transactions on , vol.46, no.3, pp. 387- 393, Aug. 2003

[12] Jana, R., Aqel, M., Srivastava, P., and Mahanti, P. K., Soft Computing Methodologies in Bioinformatics, European Journal of Scientific Research, Vol. 26, No. 2, pp. 189-203, 2009.

[13] Jacques Cohen, Computer science and bioinformatics, Communications of the ACM, v.48 n.3, p.72-78, March 2005

[14] DOMOKOS, A.. BIOINFORMATICS AND COMPUTATIONAL BIOLOGY. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture, North America, 6527, p. 571 – 574, 09 2008.

[15] Poe, D.; Venkatraman, N.; Hansen, C.; Singh, G.; , "Component-Based Approach for Educating Students in Bioinformatics," Education, IEEE Transactions on , vol.52, no.1, pp.1-9, Feb. 2009

[16] Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002;324:1018–1022.

[17] National Center for Biotechnology, "Bioinformatics Factsheet," http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html, last accessed June 13, 2012.

[18] Gavin J. Gordon, Bioinformatics in Cancer and Cancer Therapy (Cancer Drug Discovery and Development) [Kindle Edition] , ISBN: 978-1-58829-753-2 e-ISBN: 978-1-59745-576-3, Library of Congress Control Number: 2008931368

[19] Jacques Cohen, Computer science and bioinformatics, Communications of the ACM, v.48 n.3, p.72-78, March 2005

[20] Umarji, M.; Seaman, C.; Koru, A.G.; Hongfang Liu; , "Software Engineering Education for Bioinformatics," Software Engineering Education and Training, 2009. CSEET '09. 22nd Conference on , vol., no., pp.216-223, 17-20 Feb. 2009

[21] Simon R. Bioinformatics in cancer therapeutics-hype or hope?. Nat Clin Pract Oncol. 2005;2:223

[22] Goldin, L.; , "Bioinformatics Integration for Cancer Research-Goal Question analysis," Information Technology: Research and Education, 2006. ITRE '06. International Conference on , vol., no., pp.248-252, 16-19 Oct. 2006

[23] Kihara D, Yang YD, Hawkins T. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. Cancer Inform. 2007;2:25-35.

[24] Ardekani AM, Aslani F, Lakpour N. Application of genomics and proteomics technologies to early diagnosis of reproductive organ cancers. J Reprod Infertil. 2007;8(3):259-278.