
Nigerian Journal of Mathematics and Applications
Volume **23**, (2014), 70 – 76
©*Nig. J. Math. Appl.* *http* : *//www.njmaman.com*

BIOINFORMATICS: THE RELATIONSHIP BETWEEN INFORMATION TECHNOLOGY AND THE SCIENCE OF GENETIC ENGINEERING

Usman-Hamza F. E

Abstract

This study takes a look at the relationship between information technology and the science of genetic engineering. It explains how bioinformatics originated, explaining the various fields involved. This study briefly explains how information technology has helped solve complex biological problems by introducing learning conditional transducer algorithm which provides a framework for understanding DNA sequence comparison algorithms, many of which have been used by biologists to make important inferences about gene function and evolutionary history. We will also apply learning conditional transducer algorithm to gene finding and other bioinformatics problems,thus making it much simpler and more resourceful.

1. Introduction

Bioinformatics is one of the several branches of biotechnology. Biotechnology or biotech is the use of living organisms to develop or make useful products, it is the use of any technological application which involves biological systems, living organisms or derivatives thereof, to make or modify products (such as food, enzymes, drugs, vaccines etc) or processes for specific use. Biotechnology is the

Received January 22, 2015. * Corresponding author.
2010 *Mathematics Subject Classification.* 30C45.
Key words and phrases. Bioinformatics, Algorithm, DNA.
Department of Physical Sciences(Computer Science unit) Al- Hikma University, Ilorin, talk2tima@yahoo.co.uk

term used to describe genetically engineered food that contains genes modified by modern technologies [4]. Its applications are many, one of the major application of biotechnology is bioinformatics. This paper thus looks at bioinformatics, its origin and relationship with information technology. Bioinformatics is the use of problem solving tools such as algorithms and databases to solve biological problems. It uses computational techniques to organize and analyze biological data. Bioinformatics is majorly used to develop software tools to generate useful biological knowledge. The field may also be referred to as computational biology, and can be defined as, "conceptualizing biology in terms of molecules and then applying informatics techniques to understand and organize the information associated with these molecules, on a large scale [3].

2. History of Bioinformatics

Scientists were already thinking of establishing biological laws solely from data analysis by induction as far back 1920. It was the invention of powerful computers and the availability of data which can be treated by computation that launched bioinformatics as an independent field. One early contributor to bioinformatics was Elvin A. Kabat, who was one of the first person to work on biological sequence analysis with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991 Other people who played important roles in bioinformatics are Margeret Oakley Dayhoff and David Lipman, who is the director of National Centre for Biotechnology Information [2]. The term bioinformatics was formed because there is a relationship between computer science and biology. This relationship came into existence because of the phenomenal rate at which biological data are being produced and because these data have to be stored, analysed and made accessible, computer science had to come in. Three major types of datasets are being analyzed in bioinformatics: genome sequences, macromolecular structures, and functional genomics experiments (e.g. expression data, yeast twohybrid screens). However, bioinformatics analysis can also be applied to various types of other data e.g. taxonomy trees, relationship data from metabolic pathways, the text of scientific papers, and patient statistics [3]. There are two major ways of modelling a Biological system under bioinformatic approaches:
i. Static: this involves interaction of data and sequences of proteins,nucleic acides and peptides ii. Dynmic: this includes structures of protein, nucleic acids, ligands and peptides, systems bilogy and multi- agent based modelling approaches. In order to show and understand the relationship between computer science and biological science, this study explains the organization of biological knowledge in databases and the sequence in which the analysis takes place.

3. Specific Problem Statement

There are many issues in bioinformatics, one of it is the problem of not been able to analyze DNA sequence [5]. One of the main uses of the learning conditional transducer algorithm applied to bioinformatics is for DNA sequence analysis.Transducer learning algorithms are based on conditional probability computation. For example, based on a dataset with DNA sequences, classified following a criteria
(i.e.ACTGCTACTAGGGGCCTTTA Methanogenic;
CAGCTAAGAGCTTCTCTTA Protelitic;TAGACTACTAGGGGCCTTTA
..Cellulolitic;.TGCAGCTAAGAGCTTCTCTTA Protelitic; ..CAACTGCTACTAGGGGCCTTTA Methanogenic; etc.)
a transducer can be learned from this dataset to classify future sequences [5]. The classification is usually made by calculating the stochastic edit distance based on the learned transducers.

3.1. **Learning transducer algorithm.** This algorithm is described below: Learning a stochastic transducer Step 1: The dataset will be considered the Learning Set (Ls), i.e. each sequence is classified under some consideration (i.e. type 0 - methanogenics , type 1 - protelitics, type 2 - celulolitic, etc.). Step 2: from each sequence in Ls, a set of string pairs (pair dataset) Pi will be built in the form (x, y), where y = NN(x)=argminyLi-dE(x,y) ($y = NN(x)$ is the Nearest Neighbor of x, calculated as the minimum value (argmin) of the dE (the classic edit distance) between x and all other values in the same dataset (yiLi)). Step 3: a unique conditional transducer ct will be learned from all pair dataset (iPi) [6].

For a new given sequences, Step 4: s will be classified as the same type as y, where y Ls maximizing $p(y|s)$, where $p(y|s)$ is the conditional probability calculated using the learned transducer ct.

This algorithm is based on a dataset (Ls) of classified sequences under some conditions, e.g. the sequences are classified by their functionalities as type 0 - methanogenics, type 1 - protelitics, type 2 - celulolitic, etc. A new dataset, defined as pair-database (Pi), is created as $(x, NN(x))$, by calling the subalgorithm pair-database creation, where each sequence x is going to be paired with its nearest neighbor $y = NN(x)$, defined as the smallest edit distance between x and y ($dE(x, y)$). The conditional transducer (ct) is a matrix with the probabilities of the edit costs (insertions, deletion and substitutions) obtained by learning these values from the pair-database (Pi) [6]. The learning process is implemented using EM algorithm to estimate the parameters until a threshold precision is achieved [6]. Finally, a new sequence (z) can be classified comparing it with all elements in the original dataset (Ls) and calculating the stochastic distance using the transducer t. Thus, the sequence (y) where the stochastic distance is smallest (highest conditional probability - p(y—s)) will be chosen to classify s (i.e. s will

be classified as y is in Ls: type 0 - methanogenics, type 1 - protelitics, type 2 - celulolitic, etc. This algorithm is intrinsically high time-consuming. The pair-database creation algorithm has to create a new dataset using each element from the original dataset (Ls) and look for its nearest neighbor on the same dataset. Indeed, EM algorithm uses whole pair dataset (Ps) and calculates the probability to obtain a new transducer ct, as many times as threshold precision is achieved [6]. When this algorithm is used in bioinformatics dataset with several thousands of sequences, high execution times are expected. Based on our experiences, 95% of total execution time is related to pair-database creation and EM algorithms (approximately, pair-database creation 55% and EM 40%).

4. Organization of biological knowledge in databases

The database in biological science is called databank, it is the place in which biological raw data are stored in public (such as Genbank or EMBL for primary DNA sequences) [1]. In the biosciences, a databank (or data bank) is a structured set of raw data, most notably DNA sequences from sequencing projects (e.g. the EMBL and GenBank databases), the databank is used to store and organize biological data. Processes involve in analyzing biological data includes algorithms in artificial intelligence, soft computing, data mining, image processing and simulation.

The data is usually 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 twohybrid screens, expression arrays, systematic geneknockout experiments, and metabolic pathways [1]. The data stored is accessed in meaningful ways, the contents of various databanks or databases are accessed correlated and accessed simultaneously with each other. These simultaneous assessments are usually eveloped using special language such as the Sequence Retrieval System (SRS) and the Entrez system. Additionally, Databases provides access to sequence homology searches and links to other databases and analysis results. For example, a databank called SWISSPROT [1] contains verified protein sequences and more annotations describing the function of a protein. Databases which are of scientific literature such as PUBMED, MEDLINE search for similar articles based on wordusage analysis amongst other functions.

5. Analysing sequence data

The primary data of sequencing projects are DNA sequences. The DNA sequences of millions of organisms have been decoded and stored in databases. Analysis of the sequence of information is done to determine genes that encode polypeptides. These become only really valuable through their annotation. Annotation is the process of marking the genes and other biological features in a

DNA sequence. Several layers of analysis with bioinformatics tools are necessary to arrive from a raw DNA sequence at an annotated protein sequences. The ultimate goal of sequence annotation is to arrive at a complete functional description of all genes of an organism. Computer programs such as BLAST are used to search sequences from more than 400,000 organisms, containing over 200 billion nucleotides. It is to note that the simplified idea of one gene one protein one structure one function cannot take into account proteins that have multiple functions depending on context (e.g., subcellar location and the presence of cofactors). Well-known cases of moonlighting proteins are lens crystalline and phosphoglucose isomerase. Currently, work on ontologies is under way to explicitly define a vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. Families of similar sequences contain information on sequence evolution in the form of specific conservation patters at all sequence positions. Multiple sequence alignments are useful for many complete genomes of microorganisms and a few of eukaryotes are available [3]. By analysis of entire genome sequences a wealth of additional information can be obtained. The complete genomic sequence contains not only all protein sequences but also sequences regulating gene expression. A comparison of the genomes of genetically close organisms reveals genes responsible for specific properties of the organisms (e.g., infectivity). Protein interactions can be predicted from conservation of gene order or operon organisation in different genomes. Also the detection of gene fusion and gene fission (i.e, one protein is split into two in another genome) events helps to deduce protein interactions.

6. Conclusion

The relationship between computer science and biology is a natural one for several reasons. First, the phenomenal rate of biological data being produced provides challenges: massive amounts of data have to be stored, analysed, and made accessible. Second, the nature of the data is often such that a statistical method, and hence computation, is necessary. This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA. Third, there is a strong analogy between the DNA sequence and a computer program , this can be shown using the Turing Machine, this represents the DNA.
Analyses in bioinformatics focus on three types of datasets: genome sequences, macromolecular structures, and functional genomics experiments (e.g. expression data, yeast twohybrid screens). But bioinformatic analysis is also applied to various other data, e.g. taxonomy trees, relationship data from metabolic pathways, the text of scientific papers, and patient statistics. A large range of techniques are used, including primary sequence alignment, protein 3D structure alignment, phylogenetic tree construction, prediction and classification of protein structure,

prediction of RNA structure, prediction of protein function, and expression data clustering. Algorithmic development is an important part of bioinformatics, and techniques and algorithms were specifically developed for the analysis of biological data . Bioinformatics has a large impact on biological research. Giant research projects such as the human genome project [3] would be meaningless without the bioinformatics component. The goal of sequencing projects, for example, is not to corroborate a hypothesis, but to provide raw data for analysis later on. Once the raw data are available, hypotheses may be formed and tested. This has enabled computer experiments to answer biological questions which cannot be tackled by traditional approaches and has led to the production of dedicated bioinformatics. Three key areas are the organisation of knowledge in databases, sequence analysis, and structural bioinformatics.

Analyses in bioinformatics focus on three types of large datasets available in molecular biology: macromolecular structures, genome sequences, and the results of functional genomics. Bioinformatics employs a wide range of computational techniques including sequence and structural alignment, database design, macromolecular geometry. Bioinformatics integrates a variety of computational methods and heterogeneous data sources. Bioinformatics tools and services have important roles to play in drugs, they help design drugs, predict drug metabolism and toxicity.

Finally, in this paper, we have described the problem of DNA squencing approach and designed a conditional transducer learning algorithm applied to bioinformatics.Finite automata, in which each transition is augmented with an output label in addition to the familiar input label, are considered finite-state transducers. Transducers have been used to analyze some fundamental issues in bioinformatics. Weighted finite-state transducers have been proposed to pairwise alignments of DNA and protein sequences; as well as to develop kernels for computational biology. Machine learning algorithms for conditional transducers have been implemented and used for DNA sequence analysis.Futher studies can help implement and test this algorithm.

References

[1] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids *23, (1), Res. 25:33893402*

[2] Johnson, George:Tai Te Wua (2000), Kabat Database and its applications: 30 years after the first variability plot

[3] The Genome International Sequencing Consortium (2001), Initial sequencing and analysis of the human genome nature ,409:860921

[4] JC Venter et al. (2001) The sequence of the human genome science 291:13041351

[5] Marshall A. Beddoe (2007) Network Protocol Analysis using Bioinformatics Algorithms Springer Verlag. New York Inc.

[6] Abiel Roche-Lima1, Ruppa K. Thulasiram (2010). Bioinformatics algorithm based on a parallel implementationof a machine learning approach using transducers.