Pathways in Bioinformatics: A Window in Computer Science

C. P. E. Agbachi

Department of Mathematical Sciences, Kogi State University
Anyigba, Kogi State, Nigeria

***Abstract*—** *Bioinformatics is an emerging field galvanized by intense research in computational and molecular biology. It continues to grow with vast potential for the present and future. As a result, the role of computer science in driving research and results forward remains crucial. Yet this field is barely understood. It is not in the standard domain of course curriculum in many institutions, even at post graduate level. This paper looks at Bioinformatics from the perspectives of Computer Science. In the process, provides entry point and pathways for more active and productive participation.*

Keywords*—Genes, DNA, Sequence,Automata,PTA, SFA, ALERGIA.*

I. INTRODUCTION

Biology is as old as science. In recent decades, the field has gained strength in the subject of Biochemistry. Involving concepts in Molecular Biology, it provides a peep into a boundless ocean of exploration. The massive data sets that is often the case in laboratory research requires huge processing resources, givingrise to fields of Computational Biology and Bioinformatics.

A. Bioinformatics

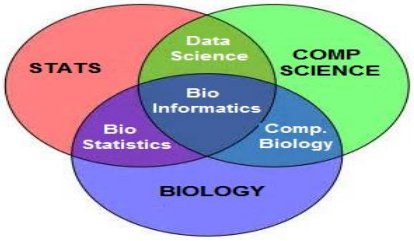


Fig. 1Bioinformatics

Informatics is a synonym for automatic information processing. It brings together the science of information and engineering of information systems [1]. Often understood as same to Computer Science, it howeverhas its emphasis and focus on information. Informatics therefore is broader, includes the study of biological and social mechanisms of information and also encompasses the study of communication using gesture, speech and language.

Given this vein, a new field emerges when data domain is in Biology. Bioinformatics is a crossdisciplineinvolving Biology, Computer Science and Mathematics (Statistics), Fig 1.Generally, it is the application of information technology to the storage, management and analysis of biological information [2].Fundamentally, this field of biology is based on the concept of molecules, andwherestudies are carried out through novelcomputational techniques.

1. Benefits:

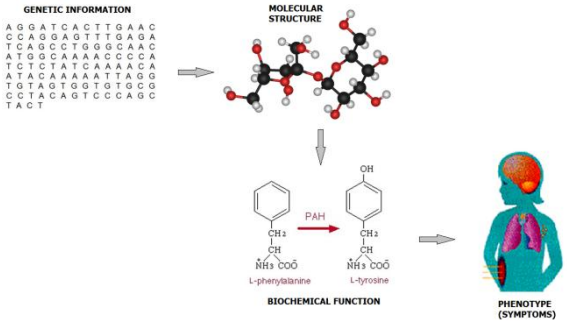


Fig 2

The main thrust in Bioinformatics can be described as developing techniques for analysing sequenced data and related structures [3]. And by so doing, endeavours to understand the molecular basis of life, Fig 2. There are broadly speaking four key areas of application [4, 5]:

Health Care: With massive acquisition of bio-medical data, Bioinformatics tools are increasingly able to provide requisite management and analysis. It leads to identification of disease susceptibility genes and the development ofmany new treatments. Furthermore, it assists the ability to predict those patients at risk for experiencing adverse reactions or patientswith a high probability of experiencing improved efficacy. It also provides the basis to accurately correlate clinical parameters of patient responsiveness toa particular therapy.

Drug Discovery:Traditionally, this process is very time consuming and expensive. However the trial and errors process is being replaced by a rational, structurebased drug designthat can reduce the time and cost of developinguseful pharmacological agents. The theme here is drug-likeness, meaning identification and elimination of candidate molecules that are unlikely to survive the

later stages of discoveryand development. Drug-likeness could be predicted by genetic algorithms. By so doing, with regularly updated public databases, bioinformatics contributes by providing functional information of target candidates and correlating this information to the biological pathways.

Forensic Analysis: Bioinformatics comes in strength with regards to personal identificationand relatedness to other individuals through forensic DNA analysis. Examples include mass disastercases that require managing, analysing, and comparing large numbers of biological samples and DNA profiles. Also whenever crime scene investigation needs identification of bacteria, insects, and plants, genomicsequences can be rearranged using microarray and analysed using bioinformatics standard techniques.

Agriculture: The backbone of biotechnology rests on bioinformatics and over the years have been a source to complement conventional breeding for crop and animalimprovement through targeted gene transfer. This biotechnological application is used to improve the yield ofcrop and animal species and their product quality such as nutritional value andshelf life. In addition to these benefits, this methodologyreduces the need for agrochemicals by creating disease and pest-resistant species,thereby reducing environmental pollution from chemical runoff. It is a fact that such increased yieldsand higher food quality can contribute to reducing world hunger and malnutrition.

B. Languages

A language provides the basic entry point in communications, and sois vital for the Computer Scientist in understanding Bioinformatics. In this respect, a review of components is as follows [6, 7]:

2. Alphabet:

An Alphabet comprises of a set of characters. This may be alpha, numeric or alphanumeric depending on application. In human languages, most often the alphabet comprises of the set, a – z. In machine language, it is only a two character set of 0 and 1 bits.

3. Strings:

Strings are formed out of the alphabet as finite sequences of characters. Examples include the words in a language dictionary. In computer electronics, 000,0101, 1100 etc. are strings of bits from the alphabet. Words are lexicons recognized in the language.

4. Regular Expressions:

Every language is governed by rules or syntax, which is often distinct. As an example, 0000, 0011, 0110 obeys the rules that stipulate a word length of 4 and prefix of 0. This rule can be stated formally as a regular expression, $w= 0.\Sigma^{*}|$ Length(w) = 4, making it possible for machines to generate or vet input strings. Often, search engines employ regular expressions in the algorithm.

5. Grammar:

Sentences are part of languages and conform to the rules of the grammar. In Computer Science, CFGs – context free grammar – are used to model natural languages.They also play important roles in compiler construction.

6. Semantics:

Every word or sentence has a meaning. There are two main techniques, namely Denotational and Axiomatic Semantics. Through these paths, a machine can interpret results and act accordingly.

This overview is based on classical concepts in Computer Science. With respect to Bioinformatics, it is a window for understanding and appreciation.

II. DATA CAPTURE

Data collection is a very important process in the input progression of any computer system. Inthis application however, the journey begins in the research laboratories of molecular biologyin assembly of DNA sequences.

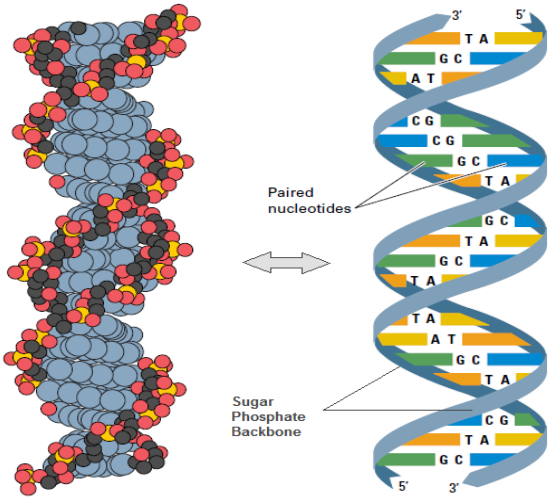


Fig.3DNA

DeoxyriboNucleicAcid (DNA),is a long polymer made from repeating units known as nucleotides. It is the domain of all biological information regarding

the cells inliving organisms and therefore the custodian of information about life.As illustrated in Fig. 3, there are base pairs of matching nucleotides forming a rung against the backbone of sugar phosphate. The structure is of the form of a double helix.

The key unit of hereditary is the Gene which broadly is a sequence of DNA encoding product [8].A Genome on the other hand is complete set of genes,genetic information that describes an organism.While Chromosomes, represent storage form of DNA withinthe cell. Conceptuallygenes are packed in these structureswith the largest human chromosome, 85nm long, comprising of about 220 million base pairs. The human genome comprises of estimated 3 billion base pairs arranged in 46 chromosomes, and out of which 23 is from each parent.

The DNA language centres on 4 alphabets derived from the nucleotides bases:Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). It is thus a language over four characters: L = {A, C, G, T}. In computer systems, a machine reads program code and performs actions in a fetch-and-execute cycle.

Similarly, ribosome-machines in living organisms read the genetic code and use this information as instructions to repair cells, and produce proteins etc. To this extent, the DNA can be likened to Turing Machine [9] where the infinite tape has analogy with a DNA molecule, Fig 4.

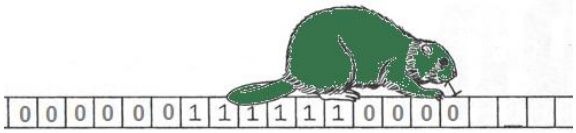


Fig. 4Busy-Beaver Turing Machine

Data capture starts in researchlaboratorywith analysis to determine sequences of the alphabet.

A. Gel Electrophoresis

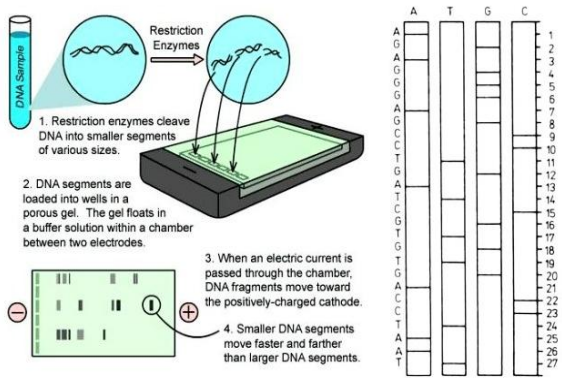


Fig.5

Gel electrophoresis is a laboratory method used to separate mixtures of DNA according to molecular size. In this process, the molecules to be separated are pushed by an electrical field through a gel that contains small pores. The molecules travel through the pores in the gel at a speed that is inversely related to their lengths. This means that a small DNA molecule will travel a greater distance through the gel than will a larger DNA molecule. It is followed by detection of bands of molecule, Fig 5.

1. Sequence Assembly:

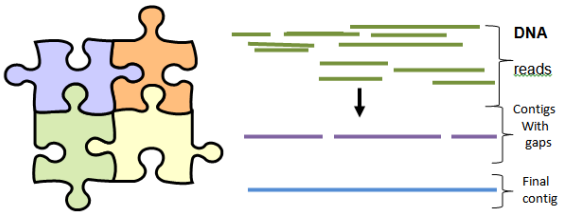


Fig 6a Jigsaw Puzzle and Sequence Assembly

The size of a human genome is put at about 10^{10} base pairs. However sequencing technology such as electrophoresis allows biologists to determine 10^3 base pairs at a time.So fora given genome, how the sequence would be determined becomes an issue. Asolution can be found in assembly of short fragments, as readsfrom a longer DNA sequence, Fig 6. It is then followed by reconstruction of the original sequence. This process involving aligning and merging of fragments is known as Shotgun Sequencing.

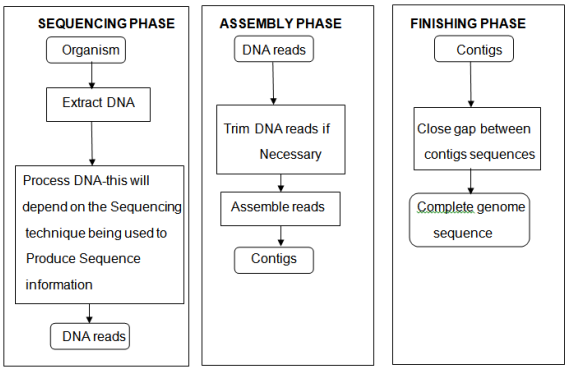


Fig 6b Phases in Shotgun sequence

Genome sequences from many research laboratories are deposited in GenBank [10] and other databases. From these sources, researchers can carry out other forms of sequence analysis.

2. Sequence Comparison:

The role of search engines is very much established in most computing environments. In the field of biology situations do arise where given a DNA sample, there is need to find out similarities with known sequences in the database. This process is performed through sequence comparisons.

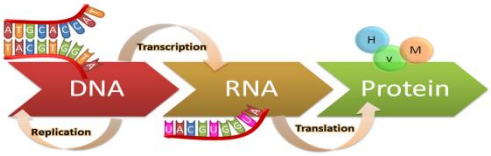
Fig. 7

A typical example is illustrated in Fig 7, between human DNA and chimpanzee. The similarities are indeed very obvious. However, searches in these databases do not necessarily yield exact matches. This is because genomes are dynamic with mutations, insertions and deletions. Furthermore, there could be human and machine errors in reading sequencing gels.

Nevertheless, the importance of sequence comparisons lies in providing clues about function and evolutionary relationships.

B. DNA Evolutions

DNA is not static but naturally evolves in the growth of organisms, along the Central Dogma of molecular genetics. Principally, there are three categories involved in this gene expression [11]:



1. Replication:

This is a copying process in which a single DNA molecule becomes two molecules.

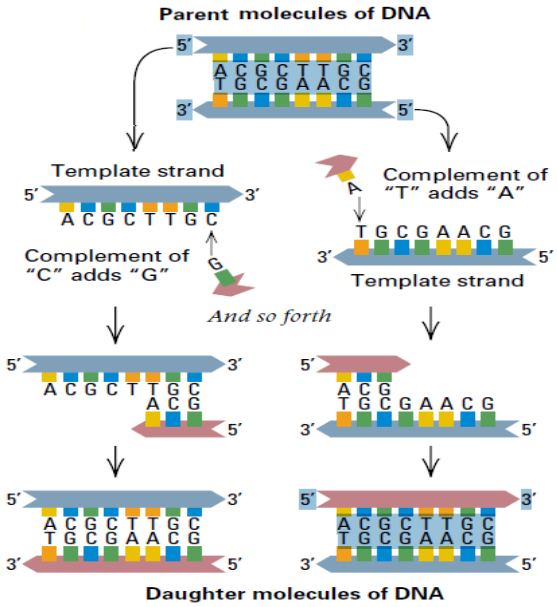


Fig. 8

It starts with two parental strands as templates for formation of a new child strand by means of A-T and G-C base pairing. The construction grows in length by the successive addition of single nucleotides tothe 3' end polarity, Fig 8.

2. Transcription:

The DNA as described earlier stores the genetic code of living organisms. However execution of these instructions involves another form, RNA – Ribonucleic Acid. Unlike the DNA, it is single strand. And where the DNA has a nucleoid T, in the RNA this replaced by U to form complementary base pair, A – U.

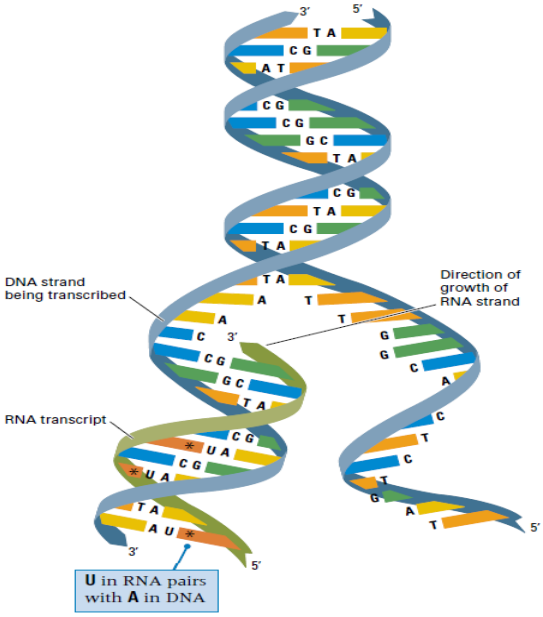


Fig. 9

Transcription is the process of makingan RNA strand by copying from a DNA template.The resulting RNA molecule is the transcript, Fig 9.It is vital in the functions provided by rRNA, mRNA and tRNA for synthesis of proteins.

3. Translation:

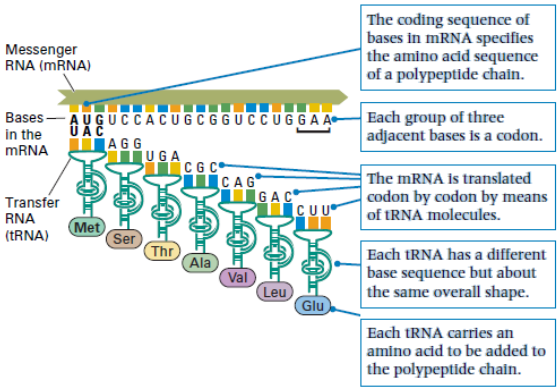


Fig. 10

The process of translation involves the messenger RNA. It delivers the information contained in asequence of DNA bases to a ribosome, where it is translated into a polypeptide chain. In the process, each transfer RNA (tRNA) moleculesbase-pair with a group of threeadjacent bases in the mRNA, Fig 10.

III.PROCESSING

Processing begins with a visit to genetic code generation and denotations. With respect to RNA, the alphabet comprises of Σ_R = {A, C, G, U}. The words known as codons are triplets from the alphabet. And by regular expression, the language is:

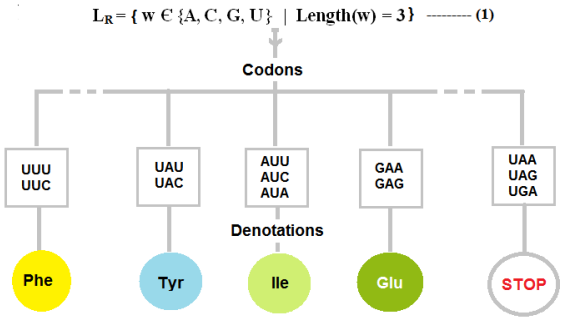


Fig.11Genetic Code

On the whole, there are 4^3 permutations, leading to 64 codons. However, two or more codons may stand for the same denotation, in20 amino acids.

A. Algorithms

Algorithms are keys in developing Bioinformatics. As they evolve, new techniques emerge with advances in methodology and solutions. Earlier discussions have highlighted sequence assembly and comparison. These are possible because of efficient algorithms.

1. Genome Sequencing:

One of the techniques in genome sequencing is the shot gun approach. There are a number of models in implementation, such as the Shortest Supersequence Problem [12, 13]. By definition, it means, given a set of sequences, find the shortest sequence *S*such that each of original sequences appears as subsequence of *S*.The algorithm may be described as follows:

- Create an overlap graphin which every node represents a fragment and edges indicate overlaps Fig 12.
- Determine which overlaps will be used in the final assembly: find an optimal spanning forestin overlap graph.

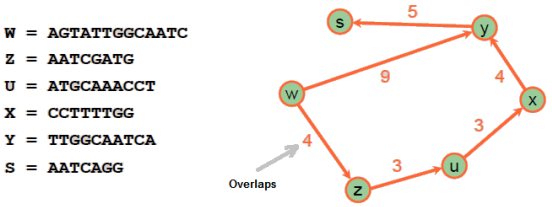


Fig. 12

- Look for paths of maximum weight: use greedy algorithm to select edge with highest weight at every step.
- Selected edge must connect nodes with in-and out-degrees <= 1.
- May end up with set of paths: each corresponds to a contig, Fig 13.

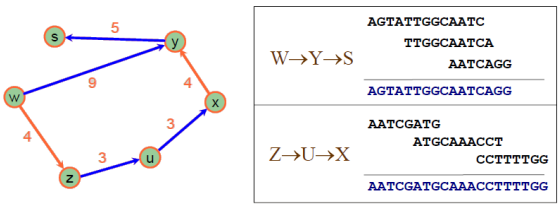


Fig.13

An alternative approach exists where analogy may be drawn with random sequences in survey data sets [14] in Geomatic Engineering. The solution here is based on Concurrent Doubly Key Data Structures. And by this application, a sorted data set translates into a genome sequence, Fig 14.

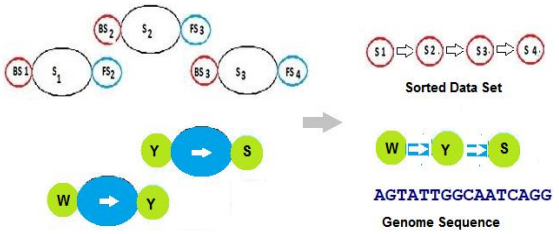


Fig.14

2. Sequence Comparison:

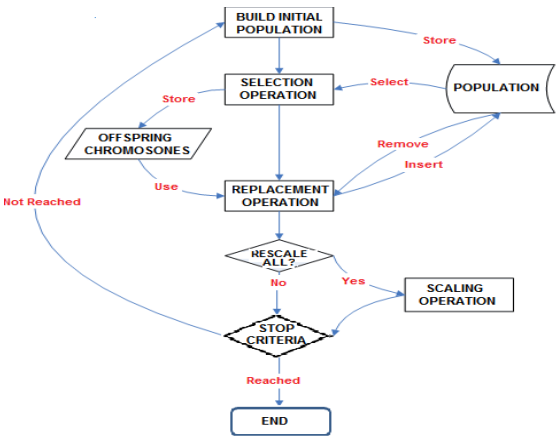


Fig 15

A sequence algorithm essentiallytakes one or more linearsequences as inputs. In a comparison based sequence algorithm, computation depends on comparisons between pair of values inthe sequence. It depends upon a comparison operator that is either previously defined or is passed to the algorithm [15]. Broadly the flowchart is of the form in Fig 15.

The sequence comparison problem is to quantify the degree of similarity or, equivalentlythe distance between the sequences. In this respect, an alignment may be constructed as an intermediatestep or as a goal in itself.

The exact definition of similarity or distance varies by application, but is usuallyformulated as a set of edit operations - mutations - that are then used to transform one sequence into the other. Often, scoring system exists in form of edit distance. These are operations comprising of single characterinsertions, deletions, and substitutions.

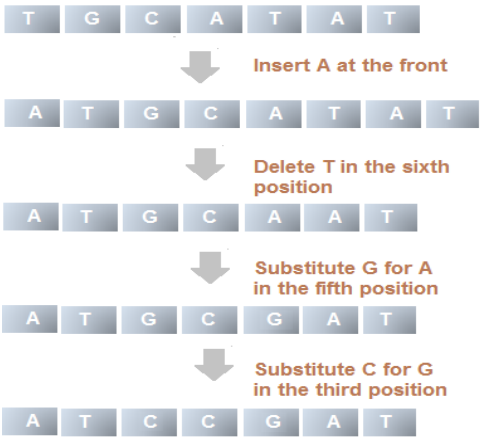


Fig. 16 Edit Operations

The edit distance is the smallest number of suchoperations required to transform one string into the other. In Fig 16, four operations are required to transform TGCATAT to ATCCGAT. Generally a cost is assigned to each operation to find the sequence of operations with the minimal cost. The setof possible operations and the cost of each operation therefore constitute a scoring system.

Mathematically, the sequence comparison problem may be described as follows [16]:

- Given an alphabet Σ, define a distance function $d:\Sigma^*\times\Sigma^*\rightarrow\mathbb{R}$
- Two strings on thealphabet $s\in\Sigma^*$ and $t\in\Sigma^*$
- Compute the value of $d(s,t)$

In addition, comparison enables similarity search. If t is large in comparison to s, we maywant to find all substrings $t_{i..j}$ of t such that $d(s,t_{i..j})\leqslant r$ where r is a threshold or cut-offvalue.

Early implementations were based on Dynamic Programming Algorithms such as Needleman-Wunsch in 1970 and Smith-Waterman in 1984. There were followed by Heuristic Algorithms typified by FASTA and BLAST in 1990.

Since then there have been further developments in research on improvements.One of such is ORIS (Ordered Index Seed Algorithm**)** Algorithm [17]. This algorithm addresses intensive or full genome comparisons with focus on fast execution times, and shows significant improvement in results.

IV. DNA AUTOMATA

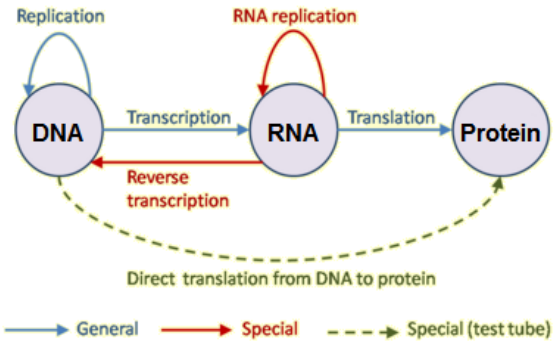


Fig. 18 Central Dogma Automata

A finite-state machine (FSM) or finite-state automaton (plural: *automata*), is a mathematical model of computation used to design both computer programs and sequential logic circuits. It is conceived as an abstract machine that can be in one of a finite number of states. These are the start state, the current state and the finalstate. A transition is then said to take place, when by a triggering event,a change occurs from one state to another,Fig 18.

Information theory, probability theory and randomness are key features of Kolmogorov theory. According to this theory, a string, which has patterns, can always be represented and written by some basic Turing machine. Though there is no practical way of expression,a finite automatoncan provide a good description, with implementation by Alergia algorithm.

Living organisms carry out complex physical processes dictated by molecular information. For example, biochemical reactions - and ultimately the entire organism's operations - are ruled by instructions stored in its genome, encoded in sequences of nucleic acids. An analogy can be construed therefore, Fig 4,between the intracellular processing of DNA and RNA, and the processing of information stored in the tape of the Turing machine.

Mathematically, a finite-state automaton M is a quintuple M = (∑, δ, Q, q_0, F), where ∑ is a finite input symbols,and δ is transition function δ:Q×M→Q. Q is a finite set of states, with $q_0$as the initial state (taken from the set Q) while F is a set of accepting states (subset of Q).

A common way to present finite automaton is a transition diagram composed of vertices and arrows connecting them. The vertices represent the states while each arrow describes a transition between vertices as a result of reading a specific symbol. The initial state is distinguished by an inward pointing arrow and accepting states usually by double circles.

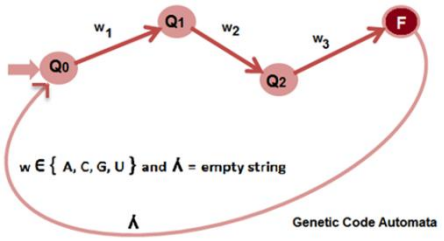


Fig. 19

A basic theory of automata is that a language is regular only and only if it is accepted or generated by FSA.Then considering Σ = {A, C, G, U}, and equation 1 in Fig 11, a genetic code automata would be of the form in Fig 19.

A. Data Mining

Data Mining in Bioinformatics is a quest for discovery of knowledge in vast fields of databases and is not unlike traditional exploration ventures for natural resources. As would be expected, the key approach is sequence mining [18]. The purpose is to discover useful sequential knowledge thattakes the form ofinsight into the structures of the data. It is this structurethat leads to expectations, a predictability thatcan be exploited.

1. Prefix Tree Acceptor(PTA):

DNA data sequence is a code in genetic language. As such, issues of grammatical inference are important in analysis, search for structures and pattern in a data sequence. While there are a number of techniques, such as Markov models, a starting point is PTA, the prefix tree acceptor [19].

Consider the set of strings P = {W, Z, U, X, Y,S}, Fig 12, where:

W = AGTATTGGCAATC
Z = AATCGATG
U = ATGCAAACCT
X = CCTTTTGG
Y = TTGGCAATCA
S = AATCAGG

For each string$P_i = a_1a_2\ldots a_i$, first, put a start node q_0. As the transition a_i, is followed itleads to next node q_i. Thisprocess continues until it reaches a nodethat accepts this string, Fig 20.

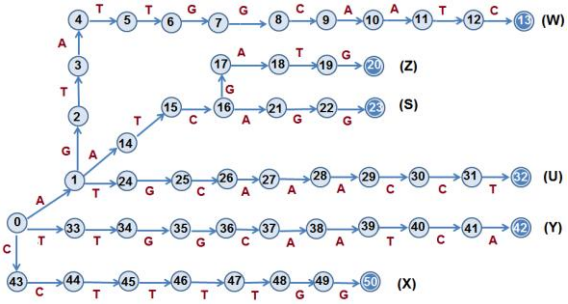


Fig. 20 PTA for {W, Z, U, X, Y, S}

This basic PTA offers the framework for optimization through merging of compatible nodes to increase flexibility and the range of sequence generation. It is achieved by introducing probabilities in transition process.

2. Statistical Finite Automata(SFA):

A Stochastic deterministic finite automata SFA [20]is defined as (Q, Σ, δ, q_0, F, P), consisting of the DFAand P, a probability function Q × ΣU{ε}→Q such that:

$$\forall\, q \in Q, \sum_{w \in \Sigma \cup \{\varepsilon\}} P(q,w) = 1 \qquad --(2)$$

P is a set of probability matrices of $p_{ij}(a)$, which states the probability of state *i* ending instate *j* with symbol $a \in \Sigma$. Let p_{if}be the probability of the string *w* ending in state *i*then the following applies:

$$p_{if} + \sum_{q_j \in Q} \sum_{a \in \Sigma} p_{ij}(a) = 1 \qquad --(3)$$

The probability of string w generated by Σ is defined by:

$$p(w) = \sum_{q_j \in Q} p_{ij}(w) p_{if} \qquad --(4)$$

The language generated here by the SFA, known as stochastic regular language, is givenas:

$$L = \{ w \in \Sigma^* : p(w) \neq 0 \} \qquad --(5)$$

Now for two languages to be equivalent, it needs to have the probability distribution to beidentical over Σ*, meaning, not only that the strings should be the same, but their probabilitiesshould be equal too.

$$\therefore L_1 \equiv L_2 \Leftrightarrow p_1(w) = p_2(w), \forall\, w \in \Sigma^* \quad -(6)$$

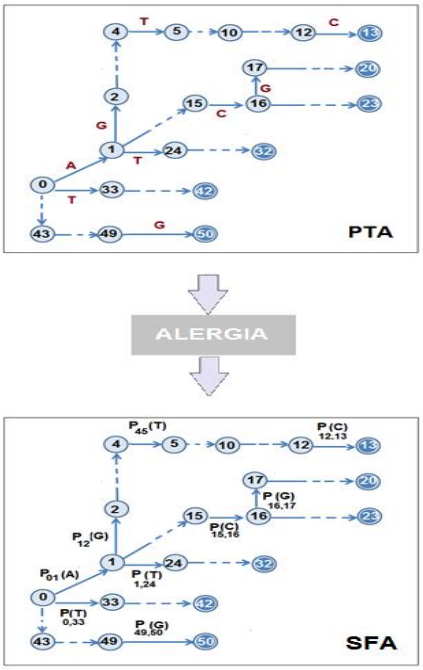


Fig. 21

The procedure for obtaining a Stochastic Finite Automata (SFA) starts with a PTA. This is fed into Alergiaalgorithm [21] where it is optimized through mergers. The output is SFA, Fig 21.

Having generated an SFA, evaluating equation 6 becomes the next step in determining findings or otherwise. Consider a known specimen, L_1 and field sample in L_2. If both the strings and probabilities are equal then L_2 belongs to the same species of L_1.

With further evolutions, inference from SFA provides theautomata's architecture, predictions and classifications,from a training sample [22, 23].

V. CONCLUSION

This paper provides an entry point to Bioinformatics for Computer Scientists. As such it starts with introduction to genetics and follows it up with data capture and processing, in a slant of language that most are accustomed to. What comes across then is the importance of adapting knowledge to solving problems in Biology. To this end the subjects of Automata, Algorithms, among others,are very crucial requiring further studies and applications.

Bioinformatics is indeed a very wide area, Fig 1. In this context, this paper can be seen as a snap shot. The detailthough is adequate, in providing requisite path ways into theemerging field. By so doing, it is hoped more computer scientists would lend support to active research,with benefits of good health to mankind.

REFERENCES

[1] Informatics,The University of Edinburgh, "What is Informatics",http://www.ed.ac.uk/files/atoms/files/-what20is20informatics.pdf.

[2] Bruno Gaeta, "BioInformatics - What and Why", © Copyright eBioInformatics Pty Ltd.

[3] Mark Gerstein, "BioInformatics - Intoduction", Yale University, bioinfo.mbb.yale.edu/mbb452a

[4] Himanshu Singh, "Bioinformatics: Benefits to Mankind", International Journal of PharmTech Research, 2016,Vol9 No 4,pp 242-248.

[5] Oliver Brandenberg et al, "Introduction to Molecular Biology and Genetic Engineering," FAO of the United Nations, Rome, 2011.

[6] SamarjitChakraborty, "Formal Languages and Automata Theory", Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology (ETH), Zurich, 2003.

[7] Jeffrey Shallit, "A Second Course in Formal Languages and Automata Theory", © Cambridge University Press.

[8] Genetics Home Reference, "Guide to Understanding Genetic Conditions", US National Library of Medicine, https://ghr.nlm.nih.gov/primer/basics/dna

[9] OrogunKehindeGbemisola, "Bioinformatics: Concepts, Design and Methodology (A Case Study of DNA Sequencing)", Bachelor's Thesis, Kogi State University, Anyigba, Nigeria, 2015.

[10] Dennis A. Benson et al, "GenBank", Nucleic Acids Research, 1998, Vol. 26, No. 1, © 1998 Oxford University Press.

[11] Introduction to Molecular Genetics and Genomics,http://www.bio-nica.info/Biblioteca/-AnonimoxxxIntroductionMolecularGenetics.pdf.

[12] Dan Lopresti, "Introduction to Bioinformatics",Computer Science & Engineering, Lehigh University, BioS, 2010.

[13] Neil C. Jones, Pavel A. Pevzner, "An Introduction to Bioinformatics Algorithms",© 2004 Massachusetts Institute of Technology.

[14] C. P. E. Agbachi, "Design and Application of Concurrent Double Key Survey Data Structures", IJCTT, Volume 36 Number 3 - June 2016.

[15] David R. Musser and Brian Osman, "Sequence Algorithm Concepts ", http://www.cs.rpi.edu/~musser/gp/algorithm-concepts/sequence-algorithms-screen.pdf

[16] J. Christopher Bare, "The Evolution of Sequence Comparison Algorithms", University of Washington.

[17] Dominique Lavenier, "Ordered Index Seed Algorithm for Intensive DNA Sequence Comparison",IRISA / CNRS Campus de Beaulieu 35042 Rennes.

[18] Philip Hingston, "Using Finite State Automata for Sequence Minining", School of Computer and Information Science, Edith Cowan University, Mt Lawley, WA 6050.

[19] HasanIbneAkram," Grammatical Inference",TechnischeUniversitätMünchen, 14 January, 2010.

[20] Asmi Shah, "DNA Sequence Representation by Use of Statistical Finite Automata",Master's Projects. Paper 40, San Jose State University, 2009.

[21] Rafael C. Carrasco, Jose Oncina, "Learning Stochastic Regular Grammars by Means of a State Merging Method",Departamento de TecnologiaInformatica y Computaci6n Universidad de Alicante, E-03071 Alicante.

[22] André Y. Kashiwabara and Alan M. Durham, "Biological Signal Prediction Using Stochastic Regular Grammars", Departamento de Ciência da Computação (IME), Universidade de Sao Paulo.

[23] F. Psomopoulos, S. Diplaris, P. A. Mitkas, "A Finite State Automata Based Technique for Protein Classification Rules Induction", Proceedings of the Second European Conference on Data Mining and Text Mining in Bioinformatics, Pisa, Italy. 24. September, 2004.