NATIONAL OPEN UNIVERSITY OF NIGERIA

FACULTY OF SCIENCES

COURSE CODE: BIO316

COURSE TITLE: **INTRODUCTION TO BIOINFORMATICS**

BIO316: INTRODUCTION TO BIOINFORMATICS

Course Team

COURSE REVIEWER: Dr. Kabir Mohammed Adamu
IBB, Lapai. Niger State
Nigeria.

COURSE COORDINATOR: Dr. Uduak Aletan
National Open University of Nigeria
Abuja, Nigeria

HEAD OF DEPARTMENT: Dr. Maureen N. Chukwu
National Open University of Nigeria
Abuja, Nigeria

Reviewed: 2023

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters University Village
Plot 91, Cadastral Zone Nnamdi Azikiwe
Expressway Jabi, Abuja

Lagos Office
National Open University of
Nigeria Headquarters
14/16 Ahmadu Bello
Way Victoria Island
Lagos

e-mail:
centralinfo@nou.edu.ng
URL:
www.nou.edu.n
g

Published By:
National Open University of

Nigeria

Printed 2012,

Reviewed 2023

Course Guide: BIO 316: INTRODUCTION TO BIOINFORMATICS

Introduction

Bioinformatics (BIO 316) is a second semester year three course. It is a compulsory two-credit unit course hosted in the Department of Biological Sciences.

Bioinformatics also referred to as computational molecular biology is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing macromolecular data. Bioinformatics knowledge covers many branches such as biology, mathematics, computer science, laws of physics & chemistry, and sound knowledge of IT to analyze biotechnology data.

The course shall be looking at use of overview of bioinformatics; scripting language; biological databases; basic tasks and processes in bioinformatics; database search and sequence retrieval techniques; database searching algorithms (BLAST, FASTA); pairwise and multiple sequence alignments; phylogenetic analysis and data mining; the use of bioinformatics tools in biotechnology/biopharma and current topics in bioinformatics and use of perl to facilitate biological analysis.

Course Competencies

The course will provide general overview of the course synopsis; this course material shall be divided into appropriate sections to help the learners understand and assimilate the contents of the course. The course guide will help students to understand how to go about Tutor- Marked-Assignment which will form part of the overall assessment at the end of the course.

Similarly, structured on-line facilitation classes in this course shall increase the comprehension of the course thus students are encouraged to activity participate. This course exposes students to data collection, management and analysis, the knowledge will be helpful during your project data collection and analysis, it is indeed very interesting field of Biology.

Course Objectives

This course is aimed at providing students the knowledge of bioinformatics and its application. The course objectives are to;

i. Discuss the overview of bioinformatics;
ii. Justify the use of scripting language in bioinformatics
iii. Explain the different biological databases
iv. Enumerate the different basic tasks and processes in bioinformatics
v. Explain the various techniques in database search and sequence retrieval
vi. Justify the use of database searching algorithms (BLAST, FASTA)
vii. Calculate and extrapolate the pairwise and multiple sequence alignments
viii. Explain the phylogenetic analysis and data mining
ix. Use bioinformatics tools in biotechnology/biopharma
x. Justify the current topics in bioinformatics
xi. Explain the use of perl to facilitate biological analysis.

Working Through this Course

The successful completion of this course entails the studying of the course guide and the reference textbooks/materials as well as other materials provided by the National Open University of Nigeria. The course guide is divided into sections, each section has self-assessment exercise. The practice of the assessment will positively influence your academic performance in the course. The course is expected to cover a minimal period of 8 weeks to complete.

Study Units

The Modules of this course shall be in accordance with the course objectives thus;

Module 1
Unit 1: An Overview of Bioinformatics
Unit 2: Scripting Language
Unit 3: Biological Databases
Unit 4: Basic Tasks and Processes tools in Bioinformatics
Unit 5: Database Search and sequence Retrieval Techniques

Module 2
Unit 1: Database searching algorithms (BLAST, FASTA)
Unit 2: Pairwise and multiple sequence alignments
Unit 3: Phylogenetic analysis and Data mining
Unit 4: Use of Bioinformatics tools in Biotechnology/Biopharma
Unit 5: Current topics in bioinformatics and use of perl to facilitate biological analysis.

References and Further Readings

In every section or Module, Reference materials shall be provided for further reading.

Presentation Schedule

Assignment	Marks
TMA 1-4	Four T M A s , best three marks of the four count at 10% each – 30% of course marks.
End of course examination	70% of overall course marks
Total	100% of course materials

Assessment

In every section or Module, self-assessment questions shall be provided for further practice.

How to get the Most from the Course

The course guide is designed in a simplified form to assist self comprehension. In addition, further references with web links are provided in each sectio/module or unit. Similarly, the course has facilitation session that will provide information on any grey areas.

Online Facilitation

Eight weeks is scheduled for online facilitation. This facilitation is divided into two session (synchronous and asynchronmous). The synchronous session is a live session that is provided by a facilitator through University approved source (Zoom) for 1 hour. While the asynchronous session is an alternative interaction session that may not be live. In the facilitator dashboard, students have access to the course materials, recorded online facilitation, weblinks, virtual library and host of others that would improve the course comprehension.

Course Information

Course Code:	BIO 316
Course Title:	Bioinformatics
Credit Unit:	2
Course Status:	Core
Course Blub:	http://elearn.nouedu2.net
Semester:	Second
Course Duration:	2 hours per week (16 hours per semester)
Required Hours for Study:	3 x 2hours x 8 weeks (48hrs)

Ice Breaker

Dr. Kabir Mohammed Adamu, is an Associate Professor of Hydrobiology & Fisheries Biotechnology; Fish Nutrition and Physiology, Department of Biological Sciences, Ibrahim Badamasi Babangida University, Lapai, (IBBUL) Niger State, Nigeria. He is an external Facilitator with the Department of Biological Sciences, National Open University of Nigeria (NOUN), where he facilitates the BIO 316 (Bioinformatics) amongst other courses. He has been teaching/lecturing Bioinformatics for the past six (6) years at various level (undergraduate and postgraduate) students in different tertiary institutions (IBBUL, NOUN, Nasarawa State University, Keffi, Nile University of Nigeria, Abuja). Dr. Kabir's research interest is in circular economy by understanding the interaction of freshwater fisheries with the environment, using both phenotypic and genotypic techniques in characterization of fisheries resources and their roles in healthy aquatic ecosystem. Understanding the protein requirement of fish and seeking for protein (especially insect protein) resource fish growth, nutrition and physiology.

Module 1

Unit 1: An Overview of Bioinformatics

Unit Structure

1.1: Introduction
1.2: Intended Learning Outcomes
1.3: Main Body
    1.3.1: Introduction to Bioinformatics
    1.3.2: Aims of Studying Bioinformatics
    1.3.3: Rationales of Studying Bioinformatics
    1.3.4: Goals of Studying Bioinformatics
    1.3.5: Applications of Bioinformatics
    1.3.6: Fields of Bioinformatics
    1.3.7: Subfields of Bioinformatics
    1.3.8: Limitations of Bioinformatics
    1.3.9: Historical Development of Bioinformatics
1.4: Summary
1.5: References/Further Readings/Web Sources
1.6: Possible Answers to Self-Assessment Exercises

 **1.1 Introduction**

This section of the course material provides the overview of the course, Bioinformatics. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. It was primary used since late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing. This section also provides insights into the antecedent and the numerous scientists that contributed to the field of Bioinformatics.

 **1.2 Intended Learning Outcomes (ILOs)**

At the end of the section, the students should be able to;

a. Define Bioinformatics
b. State the application of Bioinformatics
c. Explain the various subfields of Bioinformatics
d. Give brief historical antecedent of Bioinformatics.

 **1.3: Main Body**

1.3.1: Introduction to Bioinformatics

The mathematical, statistical and computing methods that aid to solve biological problems using DNA and amino acid sequences and related information is called Bioinformatics. It is the use of internet technology (IT) in biotechnology for data storage, data warehousing and analyzing.

According to National Center for Biotechnology Information (NCBI), Bioinformatics is a research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data. In Bioinformatics, knowledge of many branches is required like **biology**, **mathematics**, **computer science**, **laws of physics & chemistry**, and sound knowledge of IT to analyze biotech data. **Therefore**, it is an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing macromolecular data.

A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. It therefore, develops algorithms and biological software to analyze and record the data related to biology. Bioinformatics differs from a related field known as computational biology. Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered **computational molecular biology**. However, computational biology encompasses all biological areas that involve computation. For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

Beside this distinction, it is worth noting that there are other views of how the two terms relate. For example, one version defines bioinformatics as the development and application of computational tools in managing all kinds of biological data, whereas computational biology is more confined to the theoretical development of algorithms used for bioinformatics. The confusion at present over definition may partly reflect the nature of this vibrant and quickly evolving new field.

1.3.2: Aims of Studying Bioinformatics

Bioinformatics is aimed at

- the development of powerful software for data analysis, and
- benefit the researchers through disseminating the scientifically investigated knowledge, etc.

1.3.3: Rationales of Studying Bioinformatics

This is done in order to;

i. provide large storage of data.
ii. provide sequencing, crystallography and DNA chips.
iii. enable fast retrieval of data and database searching.
iv. enable data mining and analysis from integrate diverse sources.

1.3.4: Goals of Studying Bioinformatics

The ultimate goal of bioinformatics is to;

i. better understand a living cell and how it functions at the molecular level. This is achieved by analyzing raw molecular sequence and structural data.
ii. generate new insights and provide a “global” perspective of the cell. The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the “central dogma” of biology (Plate 1) in which DNA is transcribed to RNA, which is translated to proteins. Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches has proved to be a fruitful endeavor.

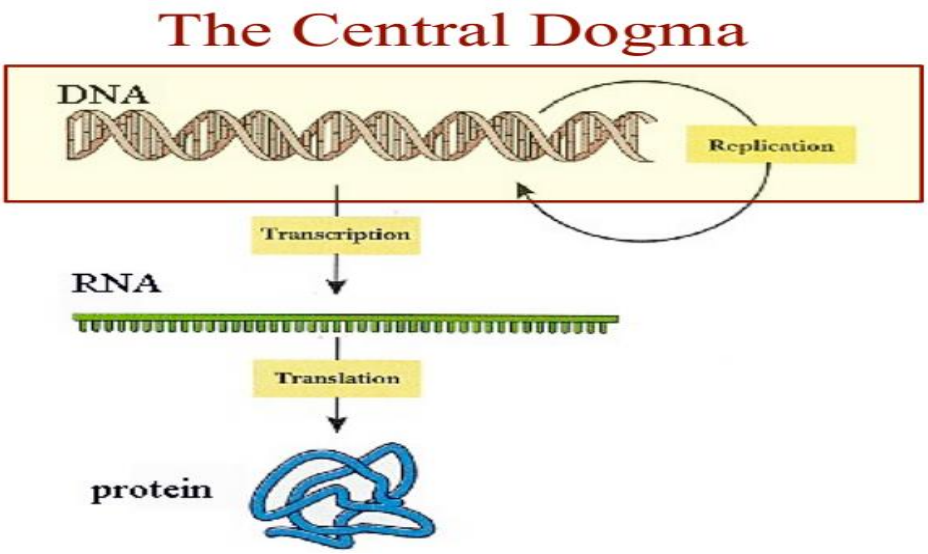


Plate 1: Schematic representation of Central Dogma theory

1.3.5: Applications of Bioinformatics

i. It has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences.

ii. It has applications, in
   a. knowledge-based drug design,
   b. forensic DNA analysis, and
   c. agricultural biotechnology.

iii. Computational studies of protein–ligand interactions provide a rational basis for the rapid identification of novel leads for synthetic drugs.

iv. Knowledge of the three-dimensional structures of proteins allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity. This informatics-based approach significantly reduces the time and cost necessary to develop drugs with higher potency, fewer side effects, and less toxicity than using the traditional trial-and-error approach.

v. In forensics, results from molecular phylogenetic analysis have been accepted as evidence in criminal courts. Some sophisticated Bayesian statistics and likelihood-based methods for analysis of DNA have been applied in the analysis of forensic identity.

vi. It is worth mentioning that genomics and bioinformatics are now poised to revolutionize our healthcare system by developing personalized and customized medicine. The high-speed genomic sequencing coupled with sophisticated informatics technology will allow a doctor in a clinic to quickly sequence a patient's genome and easily detect potential harmful mutations and to engage in early diagnosis and effective treatment of diseases.

vii. Bioinformatics tools are being used in agriculture as well. Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

1.3.6: Fields of Bioinformatics

- Microbial genome applications
- Molecular medicine
- Personalized medicine
- Preventative medicine

- Gene therapy
- Drug development
- Antibiotic resistance
- Evolutionary studies
- Waste cleanup
- Biotechnology
- Climate change Studies
- Alternative energy sources
- Crop improvement
- Forensic analysis
- Bio-weapon creation
- Insect resistance
- Improve nutritional quality
- Development of Drought resistant varieties

1.3.7: Subfields of Bioinformatics

Bioinformatics consists of two subfields:

i. the development of computational tools and databases and

ii. the application of these tools and databases in generating biological knowledge to better understand living systems. These two subfields are complementary to each other. The tool development in these subfields includes writing software for sequence, structural, and functional analysis, as well as the construction and curating of biological databases. These tools are used in three areas of genomic and molecular biological research.

   a. Molecular sequence analysis: these areas include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison.,

   b. Molecular structural analysis: this area includes protein and nucleic acid structure analysis, comparison, classification, and prediction.

   c. Molecular functional analysis: this area include gene expression profiling, protein–protein interaction prediction, protein subcellular localization prediction, metabolic pathway reconstruction, and simulation.

These aspects are not isolated but often interact to produce integrated results. For example, protein structure prediction depends on sequence alignment data; clustering of gene expression profiles requires the use of phylogenetic tree construction methods derived in sequence analysis. Sequence-based promoter prediction is related to functional analysis of co expressed genes. Gene annotation involves a number of activities, which include distinction between coding and noncoding sequences, identification of translated protein sequences, and determination of the gene's evolutionary relationship with other known genes; prediction of its cellular functions employs tools from all three groups of the analyses.

1.3.8: Limitations of Bioinformatics

Having recognized the power of bioinformatics, it is also important to realize its limitations and avoid over-reliance on and over-expectation of bioinformatics output. In fact, bioinformatics has a number of inherent limitations. In many ways, the role of bioinformatics in genomics and molecular biology research can be likened to the role of intelligence gathering in battlefields.

Intelligence is clearly very important in leading to victory in a battlefield. Fighting a battle without intelligence is inefficient and dangerous. Having superior information and correct intelligence helps to identify the enemy's weaknesses and reveal the enemy's strategy and intentions. The gathered information can then be used in directing the forces to engage the enemy and win the battle. However, completely relying on intelligence can also be dangerous if the intelligence is of limited accuracy. Over reliance on poor-quality intelligence can yield costly mistakes if not complete failures. It is no stretch in analogy that fighting diseases or other biological problems using bioinformatics is like fighting battles with intelligence. Bioinformatics and experimental biology are independent, but complementary, activities. Bioinformatics depends on experimental science to produce raw data for analysis. It, in turn, provides useful interpretation of experimental data and important leads for further experimental research. Bioinformatics predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses. In addition, the quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms being used. Sequence data from high throughput analysis often contain errors. If the sequences are wrong or annotations incorrect, the results from the downstream analysis are misleading as well. That is why it is so important to maintain a realistic perspective of the role of bioinformatics.

Bioinformatics is by no means a mature field. Most algorithms lack the capability and sophistication to truly reflect reality. They often make incorrect predictions that make no sense when placed in a biological context. Errors in sequence alignment, for example, can affect the outcome of structural or phylogenetic analysis. The outcome of computation also depends on the computing power available. Many accurate but exhaustive algorithms cannot be used because of the slow rate of computation. Instead, less accurate but faster algorithms have to be used. This is a necessary trade-off between accuracy and computational feasibility. Therefore, it is important to keep in mind the potential for errors produced by bioinformatics programs. Caution should always be exercised when interpreting prediction results. It is a good practice to use multiple programs, if they are available, and perform multiple evaluations. A more accurate prediction can often be obtained if one draws a consensus by comparing results from different algorithms.

1.3.9: Historical Development of Bioinformatics

Bioinformatics began with protein sequences that was first collected by Fredrick Sanger that developed methods for determining amino acid sequences of protein molecules. In 1962, using sequence variability, Zuckerkandl and Pauling proposed a new strategy to study evolutionary relations between the organisms which is called molecular evolution. This theory was based on the facts that similarity exists among the functionally related (homologous) protein sequences.

Subsequently, Margaret Dayhoff (1972) and her colleagues became the first to assemble databases of these sequences into a protein sequence atlas in the 60s.; where collection center eventually became known as Protein Information Resource (PIR). Proteins were organized into families and super families based on sequence similarity. Tables were built that reflected frequency of changes in sequences of closely related proteins. Phylogenetic trees were constructed showing graphically which sequences were most related and therefore, share a common branch on

the tree. The development ofcomputer methods pioneered by Dayhoff and her research group is applicable:

- in comparing protein sequences,
- detecting distantly related sequences and duplication within sequences, and
- deducing the evolutionary histories from alignment of protein sequences.

Margaret O.Dayhoff found that during evolution protein sequences undergo changes according to certain patterns such as:

- preferential alteration (replacement) in amino acids with amino acids of similar physic-chemical characteristics (but not randomly),
- no replacement of some amino acids (e.g., tryptophan) by any other amino acids, and
- development of a point accepted mutation (PAM) on the basis of several homologous sequences.

The first sequence alignment algorithm was developed by Needleman and Wunsch in 1970. This was a fundamental step in the development of the field of bioinformatics, which paved the way for the routine sequence comparisons and database searching practiced by modern biologists. The first protein structure prediction algorithm was developed by Chou and Fasman in 1974. Though it is rather rudimentary by today's standard, it pioneered a series of developments in protein structure prediction.

The 1980s saw the establishment of GenBank and the development of fast database searching algorithms such as FASTA by William Pearson and BLAST by Stephen Altschul and coworkers. The start of the human genome project in the late 1980s provided a major boost for the development of bioinformatics.

In 1980, the advent of the DNA sequence database led to the next phase in database sequence information through establishment of a data library by the European Molecular Biology Laboratory (EMBL). The purpose of establishing data library was to collect, organize and distribute data on nucleotide sequence and other information related to them. The European Bioinformatics Institute (EBI) is its successor that is situated at Hinxton, Cambridge, United Kingdom.

In 1984, the National Biomedical Research Foundation (NBRF) established the protein information resource (PIR). The NBRF helps the scientists in identifying and interpreting the information of protein sequences.

In 1988, the National Institute of Health (NIH), U.S.A. developed the National Centre for Biotechnology Information (NCBI) as a division of the National Library of Medicine (NLM) to develop information system in molecular biology. The DNA Databank of Japan (DDBJ) at Mishima joined the data collecting collaboration a few years later. The NCBI built the GenBank, the National Institute of Health (NIH) genetic sequence database GenBank is an annotated collection of all publicly available nucleotide and protein sequences. The record within GenBank represents single contig (contiguous) selection of DNA or RNA with annotations.

In 1988, the three partners (DDBJ, EMBL and GenBank) of the International Nucleotide Sequence Database Collaboration had a meeting and agreed to use a common format. All the three centres provide separate points of data submission, yet exchange this information daily making the same database available at large. All the three centres are collecting, direct submitting and distributing them so that each centre has copies of all the sequences. Hence, they can act as a primary distribution centre for these sequences. Moreover, all the databases have collaboration with each other. They regularly exchange their data.

New sequence data are accumulating day-by-day. Therefore, there is a need for powerful software so that sequences can be analyzed. For the development of algorithms [any sequence of actions (e.g. computational steps) that perform a particular task] firm basis of mathematics is needed. Now mathematicians, biologists and computer scientists are taking much interest in bioinformatics. Moreover, biologists are curious to ask reservoir of all such information because they are widely interconnected through network.

In a parallel track, the foundations for the Swiss-Prot protein sequence database also were laid in the early 1980s, when Amos Bairoch at the University of Geneva converted PIR's *Atlas* to a format similar to that used by EMBL for its nucleotide database. In this initial release, called PIR+, additional information about each of the proteins was added, increasing its value as a curated, well-annotated source of information on proteins. In this initial release, called PIR+, additional information about each of the proteins was added, increasing its value as a curated, well-annotated source of information on proteins. In the summer of 1986, Bairoch began distributing PIR+ on the US BIONET (a precursor to the Internet), renaming it Swiss-Prot. At that time, it contained the grand sum of 3900 protein sequences; this was seen as an over-whelming amount of data, in stark contrast to today's standards. Because Swiss-Prot and EMBL followed similar formats, a natural collaboration developed between these two European groups; these collaborative efforts strengthened when both EMBL and Swiss- Prot's operations were moved to EMBL‟s EBI in Hinxton, UK. One of the first collaborative projects undertaken was to create a new supplement to Swiss-Prot. Maintaining the high-quality y of Swiss-Prot entries is a time-consuming process involving extensive sequence analysis and detailed curation by expert annotators.So as to allow the quick release of protein sequence data not yet annotated to Swiss-Prot‟s stringent standards, a new database called TrEMBL (for "translation of EMBL nucleotide sequences") was created. This supplement to Swiss-Prot initially consisted of computationally annotated entries derived from the translation of all coding sequences (CDS) found in DDBJ/EMBL/GenBank.

The development and the increasingly widespread use of the Internet in the 1990s made instant access to, and exchange and dissemination of, biological data possible. **The fundamental reason that bioinformatics** gained prominence as a discipline was the advancement of genome studies that produced unprecedented amounts of biological data. The explosion of genomic sequence information generated a sudden demand for efficient computational tools to manage and analyze the data. The term Bioinformatics was coined by?

In what year did the National Biomedical Research Foundation (NBRF) established the protein information resource (PIR)

Self-Assessment Exercises

1. Enumerate the rationales of studying bioinformatics.
2. State the application of the developed computer methods pioneered by Dayhoff *et al.*

1.4: Summary

Historically, protein databases were prepared first.; then. Nucleotide databases. Various bodies were involved in the formation of various databases. Dayhoff and others prepared Atlas of Protein Sequence. EMBL succeeded by EBI, developed a DNA sequence database. NBRF established PIR while NCBI built the GenBank. DDBJ later joined the data collection collaboration. The three

main partners of the International Nucleotide Sequence Database Collaboration are Genbank, EMBL and DDBJ.

1.5: References/Further Reading/Web Sources

- Apweiler R. (2005). *Sequence Databases. In: Bioinformatics* – A practical guide tothe analysis of genes and proteins. 3rd Ed. John Wiley & Sons Inc. Publication.

- Jin X. (2006) *Essential Bioinformatics*, Cambridge University Press. The Edinburgh Building, Cambridge 362pp

- Attwood, T. K., & Miller, C. J. (2002). Progress in bioinformatics and the importance of being earnest. *Biotechnol*. *Annu*. *Rev*. 8:1–54.

Bioinformatics
https://www.bing.com/ck/a?!&&p=4b9bffa3334058f8JmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTQ3OA&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=overview+of+bioinformatics+textbook&u=a1aHR0cHM6Ly9saW5rLnNwcmluZ2VyLmNvbS9ib29rLzEwLjEwMDcvOTc4LTEtNDQ3MS02NzAyLTA&ntb=1

Introduction to Bioinformatics
https://www.bing.com/ck/a?!&&p=26cdb2ff40d63451JmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTQ4OQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=overview+of+bioinformatics+textbook&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8yMzY2MzAyODNfSW50cm9kdWN0aW9uX1RvX0Jpb2luZm9ybWF0aWNz&ntb=1

https://www.bing.com/ck/a?!&&p=959ea7fc12b8db21JmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTE0NQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&u=a1L3ZpZGVvcy9zZWFyY2g_cT1vdmVydmlldytvZitiaW9pbmZvcm1hdGljcyZxcHZ0PW92ZXJ2aWV3K29mK2Jpb2luZm9ybWF0aWNzJkZPUk09VkRSRQ&ntb=1

https://www.bing.com/videos/search?q=overview+of+bioinformatics&&view=detail&mid=D6940E6C9A6AAD1BE4AED6940E6C9A6AAD1BE4AE&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3Doverview%2Bof%2Bbioinformatics%26FORM%3DHDRSC6

1.6: Possible Answers to Self-Assessment Exercises

1.

1. provide large storage of data.
2. provide sequencing, crystallography and DNA chips.
3. enable fast retrieval of data and database searching.

4. enable data mining and analysis from integrate diverse sources.

2.

a. in comparing protein sequences,
b. detecting distantly related sequences and duplication within sequences, and
c. deducing the evolutionary histories from alignment of protein sequences.

Unit 2: Scripting Languages

Unit Structure

2.1: Introduction
2.2: Intended Learning Outcomes
2.3: Main Body
    2.3.1: Introduction to Scripting Languages
    2.3.2: Origin of Scripting
    2.3.3: Features and Limitations of Scripting Languages
    2.3.4: General Advantages of scripting languages
    2.3.5: Application of Scripting Languages
2.4: Summary
2.5: References/Further Readings/Web Sources
2.6: Possible Answers to Self-Assessment Exercises

2.1 Introduction

This section shall discuss different scripting languages. All scripting languages are programming languages. The scripting language is basically a language where instructions are written for a run time environment. They do not require the compilation step and are rather interpreted. It brings new functions to applications and glue complex system together. A scripting language is a programming language designed for integrating and communicating with other programming languages.

2.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. Define scripting languages
b. List and explain the different types of Scripting languages.
c. Enumerate the advantages, limitations of each of the scripting languages
d. State the general advantages of scripting languages.

2.3: Main Body

2.3.1: Introduction to Scripting Languages

Scripting is the action of writing scripts using a scripting language. It is a new style of programming which allows applications to be developed much faster than traditional methods allow, and makes it possible for applications to evolve rapidly to meet changing user requirements. This style of programming frequently uses a scripting language to interconnect ‘off the shelf ‘components that are themselves written in conventional language. Applications built in this way are called ‘glue applications and the language are called a ‘glue language’. A **glue language** is a programming language (usually an interpreted scripting language) that is designed or suited for writing glue code – code to connect software components. They are especially useful for writing and maintaining: Custom commands for a command shell smaller programs than those that are better implemented in a compiled language "Wrapper" programs for executable, like a batch file that moves or manipulates files and does other things with the operating system before or after running an application like a word processor, spreadsheet, data base, assembler, compiler, etc. Scripts that may change Rapid prototypes of a solution eventually implemented in another, usually compiled, language. It distinguishes neatly between programs, which are written in conventional programming language such as CC++, java and scripts, which are written using a different kind of language.

Scripting languages are an important tool in present day applied computing research. There are several reasons why scripting languages are popular especially in applied computing research. Scripting languages are object-oriented in nature, easy to learn and apply, they have flexible syntax, and powerful string-handling abilities, portable, embeddable, extensible, rich sets of libraries and some of them also provide support for concurrent programming.

Scripting languages find applications in different applied computing areas such as software engineering, bioinformatics and computational biology. **In bioinformatics**, it is important and

involves researching, developing and applying computational tools and approaches in order to expand the use of biological, medical, behavioral or health data. This also includes acquiring, storing, organizing, archiving, analyzing and visualizing such data.

2.3.2: Origin of Scripting

The use of the word 'script' in a computing context dates back to the early 1970s, when the originators of the UNIX operating system create the term 'shell script' for sequence of commands that were to be read from a file and follow in sequence as if they had been typed in at the keyword.

For instance, an 'AWKscript', a 'perl script' etc. the name 'script 'being used for a text file that was intended to be executed directly rather than being compiled to a different form of file prior to execution. Other early occurrences of the term 'script' can be found. For example, in a DOS-based system, use of a dial-up connection to a remote system required a communication package that used proprietary language to write scripts to automate the sequence of operations required to establish a connection to a remote system. Note that if we regard a script as a sequence of commands to control an application or a device, a configuration file such as a UNIX 'make file' could be regard as a script. However, scripts only become interesting when they have the added value that comes from using programming concepts such as loops and branches.

2.3.3: Features and Limitations of Scripting Languages

The popular scripting languages are: Python, Haskell, Lua, Perl, Scala, PHP, JavaScript, Erlang, R and Ruby. They are popular in the sense that they rank higher in comparison to other known scripting languages in the TIOBE programming community index. The features, advantages and limitations are;

a. Python is a general-purpose, high-level programming language that also provides scripting capability. It first appeared in 1991 and was designed by Guido van Rossum. The language was influenced by ABC, ALGOL, C, Haskell, Lisp, Modula-3, Perl, and Java. It has also influenced the design of other languages namely: Boo, Cobra, D, Falcon, Groovy, Ruby, and JavaScript.
b. *Haskell*: It is an advanced, purely-functional programming language that supports scripting capabilities. It first appeared in 1990 and is an open-source product of more than twenty years of cutting-edge research which allows rapid development of robust, concise, correct software. The language was influenced by languages like: Standard ML, Lisp, and Scheme. It has in turn also influenced several other languages like: Python and Scala
c. Lua is a powerful, fast, lightweight, embeddable language that first appeared in 1993. "Lua" (pronounced LOO-ah) means "Moon" in Portuguese, by designed Roberto group. The language was inspired by C++, CLU, Modula, Scheme and SNOBOL. It has in turn inspired languages like: Io, GameMonkey, Squirrel, Falcon and MiniD.
d. Perl is a highly capable, feature-rich programming language that first appeared in 1987. It was developed by Larry Wall and can be used in mission critical projects. The language was influenced by languages like: AWK, Smalltalk 80, Lisp, C, C++, sed,

UNIX shell, and Pascal. It has in turn influenced the creation of Python, PHP, Ruby, JavaScript, and Falcon.

e. *Scala*: It is a general-purpose programming language designed to express common programming patterns in a concise, elegant, and type-safe way. It was designed by Martin Odersky and first appeared in 2003. The language was inspired by languages like Eiffel, Erlang, Haskell, Java, Lisp, Pizza, Standard ML, OCaml, Scheme and Smalltalk. It has in turn influenced the following languages namely: Fantom, Ceylon, and Kotlin.
f. *PHP*: It is a widely used general purpose scripting language that is especially suited for Web development and can be embedded into HTML. It was designed by Rasmus Lerdorf using the C programming language and first appeared in 1995. PHP was influenced by Perl, C, C++, Java and Tcl.
g. JavaScript is a lightweight programming language that first appeared in 1994 and was designed by Brendan Eich. The language was influenced by C, Java, Perl, Python, Scheme, Self. It has in turn influenced ActionScript, CoffeeScript, Dart, Jscript .NET, Objective-J, QML, TIScript, and TypeScript.
h. Erlang is a programming language designed at the Ericsson Computer Science Laboratory. It first appeared in 1986. The language was influenced by Prolog and ML. It has in turn influenced F#, Clojure, Rust, Scala, Opa and Reia.
i. R is a language and environment for statistical computing and graphics. It was designed by Ihaka and Gentleman and first appeared in 1993. It was influenced by S, Scheme, and XLispStat.
j. *Ruby*: It is a dynamic open-source programming language with a focus on simplicity and productivity. It was designed by Yukihiro Matsumoto and first appeared in 1995. The language was influenced by Ada, C++, CLU, Dylan, Eiffel, Lisp, Perl, Python, and Smalltalk. It has also in turn influenced Falcon, Fancy, Groovy, loke, Mirah, Nu, and Reia.

2.3.4: General Advantages of scripting languages

- **Easy learning:** The user can learn to code in scripting languages quickly, not much knowledge of web technology is required.
- **Fast editing:** It is highly efficient with the limited number of data structures and variables to use.
- **Interactivity:** It helps in adding visualization interfaces and combinations in web pages. Modern web pages demand the use of scripting languages. To create enhanced web pages, fascinated visual description which includes background and foreground colors and so on.
- **Functionality:** There are different libraries which are part of different scripting languages. They help in creating new applications in web browsers and are different from normal programming languages.

2.3.5: Application of Scripting Languages

Scripting languages are used in many areas:

- Scripting languages are used in web applications. It is used in server side as well as client side. Server-side scripting languages are: JavaScript, PHP, Perl etc. and client-side scripting languages are: JavaScript, etc.
- Scripting languages are used in system administration. For example: Shell, Perl, Python scripts etc.
- It is used in Games application and Multimedia.
- It is used to create plugins and extensions for existing applications.
  Define Scripting Language? What is a glue language?

Self-Assessment Exercises

1. List any five (5) scripting language.
2. Enumerate four (4) applications of scripting language.

2.4: Summary

Scripting languages are **a type of programming language.** They are interpreted rather than requiring compilation. These are languages designed for specific runtime environments to provide additional functions, integrate complex systems, and communicate with other programming languages

2.5: References/Further Reading/Web Sources

- Apweiler R. (2005). *Sequence Databases. In: Bioinformatics* – A practical guide tothe analysis of genes and proteins. 3rd Ed. John Wiley & Sons Inc. Publication.

- Jin X. (2006) *Essential Bioinformatics,* Cambridge University Press. The Edinburgh Building, Cambridge 362pp

- Attwood, T. K., & Miller, C. J. (2002). Progress in bioinformatics and the importance of being earnest. *Biotechnol*. *Annu*. *Rev*. 8:1–54.

https://www.bing.com/ck/a?!&&p=f6b29433800faadeJmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTE5OQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Scripting+Languages+textbook%2fjournals&u=a1aHR0cHM6Ly9kbC5hY20ub3JnL2RvaS8xMC41NTU1LzU1NjUzNQ&ntb=1

https://www.bing.com/ck/a?!&&p=97ee5fe33879ca29JmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTM4MQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Scripting+Languages+textbook%2fjournals&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNjU0MDI4NjlfQV9Nb2RlbF9mb3JfU2NyaXB0aW5nX2FuZF9EZXNpZ25pbmdfYV9EaWdpdGFsX1RleHRib29r&ntb=1

https://www.bing.com/videos/search?q=Scripting+Languages&&view=detail&mid=085E7D5D45D9226FB1F4085E7D5D45D9226FB1F4&&FORM=VRDGAR&ru=%2Fvideos%2Fsea

rch%3Fq%3DScripting%2520Languages%26qs%3Dn%26form%3DQBVR%26%3D%2525eManage%2520Your%2520Search%2520History%2525E%26sp%3D-1%26lq%3D0%26pq%3Dscripting%2520languages%26sc%3D10-19%26sk%3D%26cvid%3D63F020205E95479A9DB2EA1F4C8460D6%26ghsh%3D0%26ghacc%3D0%26ghpl%3D

https://www.bing.com/videos/search?q=Scripting+Languages&&view=detail&mid=18199B51B362D2384D2918199B51B362D2384D29&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DScripting%2520Languages%26qs%3Dn%26form%3DQBVR%26%3D%2525eManage%2520Your%2520Search%2520History%2525E%26sp%3D-1%26lq%3D0%26pq%3Dscripting%2520languages%26sc%3D10-19%26sk%3D%26cvid%3D63F020205E95479A9DB2EA1F4C8460D6%26ghsh%3D0%26ghacc%3D0%26ghpl%3D

2.6: Possible Answers to Self-Assessment Exercises

1. a. python
   b. Haskell
   c. Lua
   d. Perl
   e. Scala
   f. PHP
   g. JavaScript
   h. Erlang
   i. R
   j. Ruby

2.

- Scripting languages are used in web applications.
- Scripting languages are used in system administration.
- It is used in Games application and Multimedia.
- It is used to create plugins and extensions for existing applications.

Unit 3: Biological Databases

Unit Structure

3.1: Introduction
3.2: Intended Learning Outcomes
3.3: Main Body
   3.3.1: Introduction to Biological databases
   3.3.2: Classification of Biological Databases
   3.3.3: Bioinformatics Databases

3.3.4: Pitfalls of Biological Databases
3.3.5: Database Format
3.4: Summary
3.5: References/Further Readings/Web Sources
3.6: Possible Answers to Self-Assessment Exercises

3.1 Introduction

This section of the course shall review the different types of biological databases and file formats in relation to bioinformatics. Definitions, explanations and examples in each scenario shall be provided.

3.2 Intended Learning Outcomes (ILOs)

At the end of this section, the student should be able to;

a. define biological databases
b. list and explain the different types of biological databases
c. apply the different biological databases in bioinformatics.

3.3 Main Body

3.3.1: Introduction to Biological databases

A *database* is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria. Databases are composed of computer hardware and software for data management.

The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information. Each record, also called an *entry*, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates. To retrieve a particular record from the database, a user can specify a particular piece of information, called *value*, to be found in a particular field and expect the computer to retrieve the whole data record.

This process is called *making a query*. Although data retrieval is the main purpose of all databases, biological databases often have a higher level of requirement, known as *knowledge discovery*, which refers to the identification of connections between pieces of information that were not known when the information was first entered. For example, databases containing raw sequence information can perform extra computational tasks to identify sequence homology or conserved motifs. These features facilitate the discovery of new biological insights from raw data.

3.3.2: Classification of Biological Databases

The databases are broadly classified according to the level of processing of information contained in it. In this respect databases can be classified into three categories

a. Primary database

   They contain information of the sequence or structure alone of either protein or nucleic acid e.g., PIR or protein sequences, GenBank and DDBJ for genome sequences. Primary database tools are effective for identifying the sequence similarities, but analysis of output is sometimes difficult and cannot always answer some of the more sophisticated questions of sequence analysis. In 1998, GenBank obtained more than a million of sequences from more than18,000 organisms.

   The primary databases contain, for the most part, experimental results (with some

interpretation), but are not a curated review. Curated reviews are found in what are called secondary databases. The nucleotide sequences in DDBJ/EMBL/GenBank are derived from the sequencing of a biological molecule that exists in a test tube, somewhere in a lab. They do not represent sequences that are a consensus of a population, not do they represent some other computer-generated string of letters. This framework has consequences in the interpretation of sequence analysis. Each such DNA and RNA sequence will be annotated to describe the analysis from experimental results that indicate why that sequence was determined in the first place. A great majority of the protein sequences available in public databases have not been determined experimentally, which may have downstream implications when analyses are performed. For example, the assignment of a product name or function qualifier based on a subjective interpretation of a similarity analysis (e.g., BLAST) can be very useful, but sometimes can be misleading. Therefore, the DNA, RNA, or protein sequences are the "computable" items to be analyzed, and they represent the most valuable component of primary databases.

b. Secondary database,

They contain derived-information from the primary databases. They contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR) (successor of Margaret Dayhoff's Atlas of Protein Sequence and Structure. Sequence annotation information in the primary database is often minimal. To turn the raw sequence information into more sophisticated biological knowledge, much post processing of the sequence information is needed. This begs the need for secondary databases, which contain computationally processed sequence information derived from the primary databases. The amount of computational processing work varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions. A prominent example of secondary databases is SWISS-PROT, which provides detailed sequence annotation that includes structure, function, and protein family assignment. The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database. The annotation of each entry is carefully curated by human experts and thus is of good quality. The protein annotation includes function, domain structure, catalytic sites, cofactor binding, post translational modification, metabolic pathway information, disease association, and similarity with other sequences. Much of this information is obtained from scientific literature and entered by database curators. The annotation provides significant added value to each original sequence record. The data record also provides cross referencing links to other online resources of interest. Other features such as very low redundancy and high level of integration with other primary and secondary databases make SWISS-PROT very popular among biologists. A recent effort to combine SWISS-PROT, TrEMBL, and PIR led to the creation of the UniProt database,

which has larger coverage than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation. There are also secondary databases that relate to protein family classification according to functions or structures. The Pfam and Blocks databases contain aligned protein sequence information as well as derived motifs and patterns, which can be used for classification of protein families and inference of protein functions. The DALI database is a protein secondary structure database that is vital for protein structure classification and threading analysis to identify distant evolutionary relationships among proteins

c. Specialized database.

These are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data. There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide: GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Data Bank of Japan (DDBJ), which are all freely available on the Internet. Most of the data in the databases are contributed directly by authors with a minimal level of annotation. A small number of sequences, especially those published in the 1980s, were entered manually from published literature by database management staff. Presently, sequence submission to either GenBank, EMBL, or DDBJ is a precondition for publication in most scientific journals to ensure the fundamental molecular data to be made freely available. These three public databases closely collaborate and exchange new data daily. They together constitute the International Nucleotide Sequence Database Collaboration. This means that by connecting to any one of the three databases, one should have access to the same nucleotide sequence data. Although the three databases all contain the same sets of raw data, each of the individual databases has a slightly different kind of format to represent the data. Fortunately, for the three-dimensional structures of biological macromolecules, there is only one centralized database, the PDB. This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR. It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates. The web interface of PDB also provides viewing tools for simple image manipulation.

Moreover, based on the nature of biological information concerning themolecules, data bases can be classified into two categories:

a. Sequence database this can be nucleotide or/and protein data bases

The major sources of nucleotide sequence data are the databases involved in the International Nucleotide Sequence Database Collaboration: DDBJ, EMBL, and GenBank: again, new or updated data are shared between these three entities once every 24 hours. DDBJ/EMBL/GenBank nucleotide records often are the primary source of sequence and biological information from which records in other databases are derived. Because so many other databases are dependent on the accuracy of DDBJ/EMBL/GenBank records, some important considerations

immediately come to the fore:

i. If a coding sequence is not indicated on a nucleic acid record, it will not lead to the creation of a record in the protein databases. Sequence similarity searches against the protein databases, which are the most sensitive way of doing sequence similarity searches, therefore may miss important biological relationships.

ii. If a coding feature in a DDBJ/EMBL/GenBank record contains incorrect information about the protein, this incorrect information will be passed onto other databases directly derived from the record: it could even be propagated to other nucleotide and protein records on the basis of sequence similarity.

iii. If important information about a protein is not entered in the appropriate place within a sequence record, any programs that are designed to extract information from these records more than likely will miss the information, meaning that the information will not filter down to other databases.

b. Structural databases that involve only protein databases.

Protein Sequence Databases: with the availability of hundreds of complete genome sequences from both prokaryotes and eukaryotes, efforts are now focused on the identification and functional analysis of the proteins encoded by these genomes. The large-scale analysis of these proteins has started to generate huge amounts of data, in large part because of a range of newly developed technologies in protein science. For example, mass spectrometry now is used widely in protein identification and in determining the nature of posttranslational modifications. These and other methods make it possible to identify large numbers of proteins quickly, to map their interactions.

3.3.3: Bioinformatics Databases

a. NCBI, National Center for Biotechnology Information, has a number of useful databases for bioinformatics such as;

i. BioCyc: Over 20,000 pathway/genome databases (PGDBs). BioCyc encyclopedias integrate a diverse range of data and provide a high level of curation for important microbes.

ii. BLAST: The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

iii. ClinVar: A resource to provide a public, tracked record of reported relationships between human variation and observed health status with supporting evidence.

iv. dbSNP (Database of Short Genetic Variations) includes single nucleotide variations, microsatellites, and small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping

information for both neutral variations and clinical mutations.

v. dbVar (Database of Genomic Structural Variation) has been developed to archive information associated with large scale genomic variation, including large insertions, deletions, translocations and inversions. In addition to archiving variation discovery, dbVar also stores associations of defined variants with phenotype information.

vi. Database of Genotypes and Phenotypes (dbGaP) is an archive and distribution center for the description and results of studies which investigate the interaction of genotype and phenotype. These studies include genome-wide association (GWAS), medical resequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.

vii. Gene: A searchable database of genes, focusing on genomes that have been completely sequenced and that have an active research community to contribute gene-specific data. Information includes nomenclature, chromosomal localization, gene products and their attributes (e.g., protein interactions), associated markers, phenotypes, interactions, and links to citations, sequences, variation details, maps, expression reports, homologs, protein domain content, and external databases.

viii. Genome: Contains sequence and map data from the whole genomes of over 1000 organisms. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.

ix. MedGen: Organizes information related to human medical genetics, such as attributes of conditions with a genetic contribution.

x. Nucleotide: A collection of nucleotide sequences from several sources, including GenBank, RefSeq, the Third Party Annotation (TPA) database, and PDB. Searching the Nucleotide Database will yield available results from each of its component databases.

xi. Protein: The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

b. ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

c. DNA Databank of Japan: DDBJ Center collects nucleotide sequence data as a member of INSDC (International Nucleotide Sequence Database Collaboration) and provides freely available nucleotide sequence data and supercomputer system, to support research activities in life science.

d. European Nucleotide Archives (EMBL-EBI): The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.

e. The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

   i. Protein Data Bank: Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

f. Reactome: this an open-source, open access, manually curated and peer-reviewed pathway database. Which goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic and clinical research, genome analysis, modeling, systems biology and education.

g. UniProt: The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

3.3.4: Pitfalls of Biological Databases

One of the problems associated with biological databases is overreliance on sequence information and related annotations, without understanding the reliability of the information. What is often ignored is the fact that there are many errors in sequence databases. There are also high levels of redundancy in the primary sequence databases. Annotations of genes can also occasionally be false or incomplete. All these types of errors can be passed on to other databases, causing propagation of errors. Most errors in nucleotide sequences are caused by sequencing errors. Some of these errors cause frame shifts that make whole gene identification difficult or protein translation impossible. Sometimes, gene sequences are contaminated with sequences from cloning vectors. Generally speaking, errors are more common for sequences produced before the 1990s; sequence quality has been greatly improved since. Therefore, exceptional care should be taken when dealing with more dated sequences. Redundancy is another major problem affecting primary databases. There is tremendous duplication of information in the databases, for various reasons. The causes of redundancy include repeated submission of identical or overlapping sequences by the same or different authors, revision of annotations, dumping of expressed sequence tags (EST) data, and poor database management that fails to detect the redundancy. This makes some primary databases excessively large and unwieldy for information retrieval. Steps have been taken to reduce the redundancy. The National Center for Biotechnology Information (NCBI) has now created a *non-redundant* database, called RefSeq, in which identical sequences from the same organism and associated sequence fragments are merged into a single entry. Proteins sequences derived from the same DNA sequences are explicitly linked as related entries. Sequence variants from the same organism with very minor differences, which may well be caused by sequencing errors, are treated as distinctly related entries. This carefully curated database can be considered a secondary database. As mentioned, the SWISS-PROT database also has minimal redundancy for protein sequences compared to most other databases. Another way to address the redundancy problem is to create sequence-cluster databases such as UniGene that coalesce EST sequences that are derived from the

same gene. The other common problem is erroneous annotations. Often, the same gene sequence is found under different names resulting in multiple entries and confusion about the data. Or conversely, unrelated genes bearing the same name are found in the databases. To alleviate the problem of naming genes, re-annotation of genes and proteins using a set of common, controlled vocabulary to describe a gene or protein is necessary. The goal is to provide a consistent and unambiguous naming system for all genes and proteins. A prominent example of such systems is *Gene Ontology*. Some of the inconsistencies in annotation could be caused by genuine disagreement between researchers in the field; others may result from imprudent assignment of protein functions by sequence submitters. There are also some errors that are simply caused by omissions or mistakes in typing. Errors in annotation can be particularly damaging because the large majority of new sequences are assigned functions based on similarity with sequences in the databases that are already annotated. Therefore, a wrong annotation can be easily transferred to all similar genes in the entire database. It is possible that some of these errors can be corrected at the informatics level by studying the protein domains and families. However, others eventually have to be corrected using experimental work.

3.3.5: Database Format

The elementary format underlying the information held in DDBJ/EMBL/GenBank is the *flatfile*. The correspondence between individual flatfile formats facilitates the exchange of data between each of these databases; in most cases, fieldscan be mapped on a one-to-one basis from one flatfile format to the other. Over time, various file formats have been adopted and have found continued, widespread use; others have fallen to the wayside for a variety of reasons. The success of a given format depends on its usefulness in a variety of contexts, as well as its power in effectively containing the types of biological information that need to be achieved and communicated to the community.

In its simplest form, a sequence record can be represented as a string of nucleotides with some basic tag or identifier. The most widely used of these simple formats is FASTA, which provides an easy way of handling primary data for both humans and computers. Thus, although information in DDBJ/EMBL/GenBank is basically in flatfile format, the nucleotide sequence information in the flatfile is in FASTA format. FASTA nucleotide see records takes the following form:

> U54469.1

CGGTTGCAACTTCCGGAATTCCGGCCAAGTCGTCAGTCACGTACTCGGTTGCAA
CTTCCGATTCCGGCCAAGTCGTCAGTCGGTTGCGTACTCGGTTGCAACTTCCGA
ACTTCCGGACAC

Note: The dotted lines represent several nucleotides usually there are 60 characters per line.

Only two lines out of many are shown for brevity. The greater than character (>) designate the beginning of a new sequencerecord; this line is called the definition line or def line. After „>“ is the accession version number (U54469.1). The accession version number is followed by the DNA sequence either in uppercase or lower-case letters.

More detail can sometimes be added to this format making it more complex. For

instance, one can add more information to the def line making it more informative:

> >gb$U54469.1 $/$DMU5469 *Drosophila melanogaster* eukaryotic initiation factor4E (eIF4E) gene, alternative splice products, complete cds

The above modified FASTA file now has information on the source database (gb, for GenBank), its accession version number (U54469.1), a LOCUS name identifier (in GenBank), or entry name identifier (in EMBL; DMU54469), and a short description of what biological entity the sequence represents.

A Dissection of Nucleotide Sequence Flatflles

Since flatfiles represent the elementary unit of information within DDBJ/EMBL/GenBank and facilitate the interchange of information between these databases, it is important to understand what each individual field within the flatfile represents and what kinds of information can be found in various parts of the record. At this time, the DDBJ and GenBank flatfile formats are nearly identical, whereas EMBL uses line-type prefixes; these prefixes indicate the type of information p resent within each line of the record. Flatfiles can be separated into three major parts:

A. **The Header**: which contains the information (descriptors) that apply to theentire record;
B. **The features:** which are the annotations on the record; and
C. **The Nucleotide Sequence:** All major nucleotide database flatfiles endwith ??on the last line of the record.

The three primary biological databases are?
What is DALI database?

Self-Assessment Exercises

1. List and explain the parts of flatfiles
2. What is Biological Database?

3.4: Summary

Biological databases are the stores of biological information such as nucleotides and proteins. The different types of biological databases were discussed in this section. Similarly, the ability of the databases to shake hand with each other using the programme called flat files was also examined in this section.

3.5: References/Further Reading/Web Sources

- Apweiler R. (2005). *Sequence Databases. In: Bioinformatics* – A practical guide tothe analysis of genes and proteins. 3rd Ed. John Wiley & Sons Inc. Publication.

- Jin X. (2006) *Essential Bioinformatics*, Cambridge University Press. The Edinburgh

Building, Cambridge 362pp

- Attwood, T. K., & Miller, C. J. (2002). Progress in bioinformatics and the importance of being earnest. *Biotechnol*. *Annu*. *Rev*. 8:1–54.

https://www.bing.com/ck/a?!&&p=ecab5b8120acbe79JmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTMyMQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=biological+databases+References&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNDQ3NTk3NjNfQmlvbG9naWNhbF9EYXRhYmFzZXM&ntb=1

https://www.bing.com/ck/a?!&&p=0e837a75e7d4187fJmltdHM9MTY4Njg3MzYwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTMzNQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=biological+databases+pdf&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNTc1NzU0ODdfQmlvaW5mb3JtYXRpY3NfQV9QcmFjdGljYWxfR3VpZGVfdG9fTkNSV9EYXRhYmFzZXNfYW5kX1NlcXVlbmNlX0FsaWdubWVudHM&ntb=1

https://www.bing.com/videos/search?q=biological+databases&&view=detail&mid=8F05D1F8DD75E54C3C2A8F05D1F8DD75E54C3C2A&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3Dbiological%2Bdatabases%26FORM%3DHDRSC6

https://www.bing.com/videos/search?q=biological+databases&&view=detail&mid=3AE83F536526B7EE26DE3AE83F536526B7EE26DE&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3Dbiological%2Bdatabases%26FORM%3DHDRSC6

 **3.6: Possible Answers to Self-Assessment Exercises**

1.

A. **The Header**: which contains the information (descriptors) that apply to theentire record;
B. **The features:** which are the annotations on the record; and
C. **The Nucleotide Sequence:** All major nucleotide database flatfiles endwith ??on the last line of the record.

2. Biological databases are the stores of biological information such as nucleotides and proteins.

Unit 4: Basic Tasks and Processes tools in Bioinformatics

Unit Structure

4.1: Introduction
4.2: Intended Learning Outcomes
4.3: Main Body
    4.3.1: Bioinformatics Tasks Tools
    4.3.2: Bioinformatics Processing Tools
    4.3.3: Application of Programmes in Bioinformatics
    4.3.4: Bioinformatics Projects
4.4: Summary
4.5: References/Further Readings/Web Sources
4.6: Possible Answers to Self-Assessment Exercises

4.1 Introduction

Bioinformatic tools are software programs that are designed for extracting the meaningful information from the mass of molecular biology / biological databases and to carry out sequence or structural analysis Factors that must be taken into consideration when designing bioinformatics tools, software and programmes are:

a. The end user (the biologist) may not be a frequent user of computer technology
b. These software tools must be made available over the internet given the global distribution of the scientific research community

4.2 Intended Learning Outcomes (ILOs)

At the need of this section, the students should be able to;

a. List and explain any three (3) types of tasks tools in bioinformatics.
b. List and explain any fifteen (15) types of processes tools in bioinformatics.
c. Explain the different application programmes in bioinformatics.

4.3: Main Body

4.3.1: Bioinformatics Tasks Tools

1. Homology and Similarity Tools: Homologous sequences are sequences that are related by divergence from a common ancestor. Thus, the degree of similarity between two sequences can be measured while their homology is a case of being either true of false. This set of tools can be used to identify similarities between novel query sequences of unknown structure and function and database sequences whose structure and function have been elucidated.
2. Protein Function Analysis: This group of programs allow you to compare your protein sequence to the secondary (or derived) protein databases that contain information on motifs,

signatures and protein domains. Highly significant hits against these different pattern databases allow you to approximate the biochemical function of your query protein.

3. Structural Analysis: This set of tools allow you to compare structures with the known structure databases. The function of a protein is more directly a consequence of its structure rather than its sequence with structural homologs tending to share functions. The determination of a protein's 2D/3D structure is crucial in the study of its function.
4. Sequence Analysis: This set of tools allows you to carry out further, more detailed analysis on your query sequence including evolutionary analysis, identification of mutations, hydropathy regions, CpG islands and compositional biases. The identification of these and other biological properties are all clues that aid the search to elucidate the specific function of your sequence.

4.3.2: Bioinformatics Processing Tools

1. BLAST (Basic Local Alignment Search Tool) comes under the category of homology and similarity tools. It is a set of search programs designed for the Windows platform and is used to perform fast similarity searches regardless of whether the query is for protein or DNA. Comparison of nucleotide sequences in a database can be performed. Also, a protein database can be searched to find a match against the queried protein sequence. NCBI has also introduced the new queuing system to BLAST (Q BLAST) that allows users to retrieve results at their convenience and format their results multiple times with different formatting options. Depending on the type of sequences to compare, there are different programs:

    a) blastp compares an amino acid query sequence against a protein sequence database
    b) blastn compares a nucleotide query sequence against a nucleotide sequence database
    c) blastx compares a nucleotide query sequence translated in all reading frames against a protein sequence database
    d) tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
    e) tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

2. **FASTA:** this is homology search A ll sequences. An alignment program for protein sequences created by Pearsin and Lipman in 1988. The program is one of the many heuristic algorithms proposed to speed up sequence comparison. The basic idea is to add a fast prescreen step to locate the highly matching segments between two sequences, and then extend these matching segments to local alignments using more rigorous algorithms such as Smith-Waterman.
3. EMBOSS: European Molecular Biology Open Software Suite is a software-analysis package. It can work with data in a range of formats and also retrieve sequence data transparently from the Web. Extensive libraries are also provided with this package, allowing other scientists to release their software as open source. It provides a set of sequence-analysis programs, and also supports all UNIX platforms.

4. **Clustalw:** It is a fully automated sequence alignment tool for DNA and protein sequences. It returns the best match over a total length of input sequences, be it a protein or a nucleic acid.
5. RasMol: It is a powerful research tool to display the structure of DNA, proteins, and smaller molecules. Protein Explorer, a derivative of RasMol, is an easier to use program.
6. **PROSPECT:** PROtein Structure Prediction and Evaluation Computer ToolKit is a protein-structure prediction system that employs a computational technique called protein threading to construct a protein's 3-D model.
7. PatternHunter: is based on Java, can identify all approximate repeats in a complete genome in a short time using little memory on a desktop computer. Its features are its advanced patented algorithm and data structures, and the java language used to create it. The Java language version of PatternHunter is just 40 KB, only 1% the size of Blast, while offering a large portion of its functionality.
8. COPIA: COnsensus Pattern Identification and Analysis) is a protein structure analysis tool for discovering motifs (conserved regions) in a family of protein sequences. Such motifs can be then used to determine membership to the family for new protein sequences, predict secondary and tertiary structure and function of proteins and study evolution history of the sequences.
9. AACOMP(Ident, Sim): These are search tools at the Expasy server that use the composition of anamino acid sequence, rather than the sequence itself, to find sequences in a database (SWISS-PROT) with similar composition. (The composition of a query is searched against a database of the compositions of sequences in SWISS- PROT).
10. AACompident inputs the composition of the query and is useful when the query sequence is not known. In fact, it may be used to identify the sequence from its composition. AACompSim inputs a query sequence and computes its composition internally.
11. BLOSUM: This is a family of substitution matrices for scoring alignments, for example when doing BLAST searches. BLOSUM62 is the most widely used member of this family.
12. BOOTSTRAP: This is a statistical method used to access the robustness of phylogenetic trees produced by various tree-building methods. The original multiple alignment from which a phylogenetic tree was first produced is used to generate numerous bootstrap alignments. Thus, a bootstrap alignment has the same number of columns as the original alignment. A phylogenetic tree is built from each bootstrap alignment. Once all the trees have been built, these trees may be compared to yield a consensus tree or consensus subtrees. Well-conserved branches (or more generally, subtrees) are often deemed as reliable.
13. CLUSTER ANALYSIS: This is a computational and statistical procedure for partitioning a data set into subsets of similar items. It is often used to group together genes with similar expression patterns (or experiments with similar response patterns of genes) from microarray gene expression data. It is also used to group together proteins with similar sequences or similar structure (many protein classification databasesare constructed this way).
14. CONSERVED DOMAIN DATABASE: This is a database at NCBI which may be used to locate domains in a protein sequence. Domain motifs, represented as position-specific weight matrix profiles, are scanned against sequence to find which motifs hit were.
15. DOT MATRIX: This is a nice way to visualize a local alignment, especially many alignments of pairs of possibly overlapping, fragments in the two sequences. The horizontal

andvertical axes correspond to the two sequences. Solid diagonal lines denote aligned fragments.

16. ENSEMBL: This is a web server which provides access to automatically generated annotations of the genomes of complex organisms such as humans. One can search the human genome, browse chromosomes, find genes, find genomicsequences similar to a given protein sequence, etc.
17. ENTREZ: This is a text-based search engine for bioinformatics databases, at NCBI. It provides access to a literature database (PubMed), nucleotide database (GenBank), protein sequence database, 3D Structure database (MMDB),Genome database (complete genome assemblies), population sets, and Online Mendelian Inheritance in Man (OMIM).
18. GENSCAN: This is a popular genefinding program. It uses hidden Markov models
19. GLOBAL ALIGNMENT: This is a full alignment of two nucleotide or protein sequences, with gaps insertedin one or both sequences, as needed.
20. GRAIL: This is a popular genefinding program. It uses neural networks to combine information about predicted local sites such as splice sites with predicted coding regions.
21. LOCAL ALIGNMENT: This is the process of finding and aligning highly similar regions of two DNA or protein sequences.
22. NEME: This is a tool for automatically discovering motifs (represented by ungapped position weight matrix profiles) from a set of related protein or DAN sequences. The motif discovery algorithm is based on fitting a two-component mixture model to the given set of sequences, using the EM algorithm. One component describes the motif by a fixed-width position-weight matrix profile. The other component models background, i.e. all other positions in the sequences.
23. MULTIPLE ALIGNMENT: This is a global alignment of more than two nucleotide or protein sequences. The alignment is typically scored by scoring each column of the alignment and adding up the column scores. Gaps costs may be incurred in the score of a column, or separately. One way to score a column is by its degree of conserveness (more conserved columns, indicate better scoring alignments). This can be done by computing the information content of the column. A more widely used method is called the sum-of-pairs method. In this method, the score of a column is the sum of the scores of all distinct pairs of letters in the column, where a pair of letters is scored via a substitution matrix (such as BLOSUM or PAM in the case of protein sequences). A multiple alignment is also a first step in phylogenetic tree-building.
24. NEIGHBOR-JOINING METHOD: This is a phylogenetic tree-building method that is “one notch above” the UPGMAtree-building method.
25. PARSIMONY: This is a character-based method for constructing a phylogenetic tree from a multiply aligned set of sequences. The parsimony of a tree is defined as the minimum number of substitutions required to produce a given set of sequences, placed a particular way the leaves of the tree. The parsimony of a tree may be computed efficiently by a clever method. It is much more time consuming to find the tree which has maximum parsimony (minimum number of substitutions).
26. PHI-BLAST: Pattern Hit Initiated BLAST. This is a version of BLAST that inputs a protein sequence and a regular expression pattern in it. It may be used to find other protein sequences that not only contain the same patter n but are also similar to the input protein sequence in the proximity of the occurrence of the pattern, in thetwo sequences.

27. PHYLOGENETIC ANALYSIS: This is the process of building a phylogenetic tree from a given set of sequences (or other data). The first step is to do a multiple alignment of the sequences, using CLUSTALW perhaps. The second step is to clean up the multiple alignment (remove outlying sequences, handle gaps, downweight overrepresented sequences, etc.). The third step is to build one or more trees, using a distance-based method, a parsimony method, or a likelihood-based method. The fourth step (which may interact with earlier steps) is to assess the quality of the built trees perhaps using bootstrap.
28. PREDICTPROTEIN: This web server features the class of PHD programs for predicting secondary structure, solvent accessibility, and transmembrane helices, as well as programs for prediction of tertiary structure, coiled-coil regions, etc. At the same site, in the Meta Predict Protein section is another secondary structure prediction program, JPRED, which takes a consensus between a number of methods. There arealso three other programs for predicting transmembrane helices, TMHMM, TOPRED, and DAS. The tertiary structure prediction programs include TOPITS, SWISS-MODEL, and CPHmodels.
29. PRINCIPAL COMPONENTS ANALYSIS: This is a method of visualizing high-dimensional data by transforming it into avery low-dimensional space (usually 1, 2 or 3D). The transformation is achieved (by rotating and translating the axes of the original space) in such a way that the first few axes describe the data as best as possible. The remaining axes may then be "thrown away", with possibly some (but hopefully not a lot) of loss of information.
30. RESTRICTION MAP CONSTRUCTION: Restriction enzymes cut foreign DAN at locations called restriction sites. Arestriction map is a map of these locations in the DNA. These locations are not determined experimentally. What is determined experimentally are some properties of fragments formed after the DNA has been cut by various restriction enzymes in various ways. From this data, a restriction map is then constructed by a nontrivial computer algorithm. Graph theory and algorithms are often used.
31. SWISS-PROT: This is a curated database of protein sequences with ahigh level of annotation (functions of a protein, domains in it, etc) and low redundancy.
32. TANDEM REPEAT FINDER: This is a tool that finds tandem repeats ---two or more exact or near-exact copies of the same sequence fragment that are adjacent--- in a nucleotide sequence.
33. THREADING: This is an alignment of a protein sequence to the 3D structure of another protein.
34. TOPITS: This is a program for predicting the tertiary structure of a protein from its sequence. It uses a database of secondary structure strings derived from tertiarystructures in PDB. (This database has one such string for each tertiary structure in PDB). First, this program uses the PHDsec program to predict the secondary structure string sse(x) of a protein sequence x. Next, it aligns, one by one, this predicted string sse( ) with each secondary structure string in the database (Alignments are x done via dynamic programming).
35. TREMBL: This is an automatically annotated adjunct to SWISS-PROT.
36. UPGMA: This is a distance-based method for building a phylogenetic tree from a multiply aligned set of sequences. It performs a hierarchical clustering of the given data set, which yields this tree.

4.3.3: Application of Programmes in Bioinformatics

a. JAVA in Bioinformatics: Since research centers are scattered all around the globe ranging from private to academic settings, and a range of hardware and OSs are being used, Java is emerging as a key player in bioinformatics. Physiome Sciences' computer-based biological simulation technologies and Bioinformatics Solutions' PatternHunter are two examples of the growing adoption of Java in bioinformatics.
b. Perl in Bioinformatics: String manipulation, regular expression matching, file parsing, data format interconversion etc are the common text-processing tasks performed in bioinformatics. Perl excels in such tasks and is being used by many developers. Yet, there are no standard modules designed in Perl specifically for the field of bioinformatics. However, developers have designed several of their own individual modules for the purpose, which have become quite popular and are coordinated by the BioPerl project.

4.3.4: Bioinformatics Projects

a. BioJava: The BioJava Project is dedicated to providing Java tools for processing biological data which includes objects for manipulating sequences, dynamic programming, file parsers, simple statistical routines, etc.
b. BioPerl: The BioPerl project is an international association of developers of Perl tools for bioinformatics and provides an online resource for modules, scripts and web links for developers of Perl-based software.
c. BioXML: A part of the BioPerl project, this is a resource to gather XML documentation, DTDs and XML aware tools for biology in one location.
d. Biocorba: Interface objects have facilitated interoperability between bioperl and other perl packages such as Ensembl and the Annotation Workbench. However, interoperability between bioperl and packages written in other languages requires additional support software. CORBA is one such framework for interlanguage support, and the biocorba project is currently implementing a CORBA interface for bioperl. With biocorba, objects written within bioperl will be able to communicate with objects written in biopython and biojava (see the next subsection). For more information, see the biocorba project website at http://biocorba.org/ . The Bioperl BioCORBA server and client bindings are available in the bioperl-corba-server and bioperl-corba-client bioperl CVS repositories respecitively. (see http://cvs.bioperl.org/ for more information).
e. Ensembl: Ensembl is an ambitious automated-genome-annotation project at EBI. Much of Ensembl\'s code is based on bioperl, and Ensembl developers, in turn, have contributed significant pieces of code to bioperl. In particular, the bioperl code for automated sequence annotation has been largely contributed by Ensembl developers. Describing Ensembl and its capabilities is far beyond the scope of this tutorial The interested reader is referred to the Ensembl website at http://www.ensembl.org/.
f. bioperl-db: Bioperl-db is a relatively new project intended to transfer some of Ensembl's capability of integrating bioperl syntax with a standalone Mysql database (http://www.mysql.com) to the bioperl code-base. More details on bioperl-db can be found in the bioperl-db CVS directory at http://cvs.bioperl.org/cgi-bin/viewcvs/viewcvs.cgi/bioperl-db/?cvsroot=bioperl. It is worth mentioning that most of the bioperl objects mentioned above map directly to tables in the bioperl-db schema.

Therefore object data such as sequences, their features, and annotations can be easily loaded into the databases, as in $loader->store($newid,$seqobj) Similarly one can query the database in a variety of ways and retrieve arrays of Seq objects. See biodatabases.pod, Bio::DB::SQL::SeqAdaptor, Bio::DB::SQL::QueryConstraint, and Bio::DB::SQL::BioQuery for examples.

g. Biopython and biojava: Biopython and biojava are open-source projects with very similar goals to bioperl. However, their code is implemented in python and java, respectively. With the development of interface objects and biocorba, it is possible to write java or python objects which can be accessed by a bioperl script, or to call bioperl objects from java or python code. Since biopython and biojava are more recent projects than bioperl, most effort to date has been to port bioperl functionality to biopython and biojava rather than the other way around. However, in the future, some bioinformatics tasks may prove to be more effectively implemented in java or python in which case being able to call them from within bioperl will become more important. For more information, go to the biojava http://biojava.org/ and biopython http://biopython.org/ websites.

The type of BLAST that compares an amino acid query sequence against a protein sequence database is? FASTA was created by?

Self-Assessment Exercises

1. What is Bioinformatic tools?
2. List the four bioinformatics tasks tools.

4.4: Summary

This section reviewed four different types of bioinformatics task tools and about thirty-two bioinformatics processing tools. Similarly, the two applications of programmes in bioinformatics were explained with seven bioinformatics projects.

4.5: References/Further Reading/Web Sources

- Apweiler R. (2005). *Sequence Databases. In: Bioinformatics* – A practical guide tothe analysis of genes and proteins. 3rd Ed. John Wiley & Sons Inc. Publication.

- Jin X. (2006) *Essential Bioinformatics,* Cambridge University Press. The Edinburgh Building, Cambridge 362pp

- Attwood, T. K., & Miller, C. J. (2002). Progress in bioinformatics and the importance of being earnest. *Biotechnol. Annu. Rev*. 8:1–54.

https://www.bing.com/ck/a?!&&p=866f0f07bb72ce46JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTIwNw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Basic+Tasks+and+processes+tools+in+Bioinformatics+pdf&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zMDU2NTgwMjhfQmlvaW5mb3JtYXRpY3NfQmFzaWNzX0RldmVsb3BtZW50X2FuZF9GdXR1cmU&ntb=1

https://www.bing.com/ck/a?!&&p=9e81b037f0f4c6afJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTIzMw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Basic+Tasks+and+processes+tools+in+Bioinformatics+pdf&u=a1aHR0cHM6Ly93d3cuaW1zYy5yZXMuaW4vfmthYnJ1L3BhcmFwcC9iaW9pbmZvcm1hdGljc19ub3Rlcy5wZGY&ntb=1

https://www.bing.com/videos/search?q=Basic+Tasks+and+processes+tools+in+Bioinformatics+pdf&&view=detail&mid=D148AC5199C8035C9525D148AC5199C8035C9525&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DBasic%2BTasks%2Band%2Bprocesses%2Btools%2Bin%2BBioinformatics%2Bpdf%26FORM%3DHDRSC6

https://www.bing.com/videos/search?q=Basic+Tasks+and+processes+tools+in+Bioinformatics+pdf&&view=detail&mid=2342ACF1849B41619F922342ACF1849B41619F92&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DBasic%2BTasks%2Band%2Bprocesses%2Btools%2Bin%2BBioinformatics%2Bpdf%26FORM%3DHDRSC6

4.6: Possible Answers to Self-Assessment Exercises

1. Bioinformatic tools are software programs that are designed for extracting the meaningful information from the mass of molecular biology / biological databases and to carry out sequence or structural analysis.

2.

1. Homology and Similarity Tools
2. Protein Function Analysis
3. Structural Analysis
4. Sequence Analysis:

Unit 5: Database Search and **S**equence Retrieval Techniques

Unit Structure

5.1: Introduction
5.2: Intended Learning Outcomes
5.3: Main Body
    5.3.1: Coding for Nucleotide Bases and Amino Acid Residues
    5.3.2: Database Analyses
    5.3.3: Database Organisation
    5.3.4: Search Engines
    5.3.5: Sequence Retrieval System (SRS)
    5.3.6: Types of Sequence Retrieval Databases
    5.3.7: Search Sites
    5.3.8: Sequence Retrieval Tools
5.4: Summary
5.5: References/Further Reading/Web Sources
5.6: Possible Answers to Self-Assessment Exercises

5.1 Introduction

The completion of sequencing of a number of model organisms, along with the continued sequencing of others, underscores the necessity for all biologists to learn how to make their way effectively through this sequence space. GenBank, or any other biological database for that matter, serves little purpose unless the database can be easily searched and entries can be retrieved in a usable,meaningful format. Otherwise, sequencing efforts have no useful end, because the biological community as a whole cannot make use of the information hidden within these millions of bases and amino acids. Much effort has gone into making such data accessible to the average user, and the programs and interfacesresulting from these efforts are the focus of this chapter. The discussion centers on querying databases at the National Center for Biotechnology Information (NCB) because these more "general" repositories are far and away the ones most often accessed by biologists, but attention is also given to a specialized databases that provide information not necessarily found through Entrez.

5.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. provide the codes for nucleotides bases
b. give the codes for amino acids
c. list and explain the different types of database search techniques
d. list and explain the different types of sequence retrieval techniques.

5.3: Main Body

5.3.1: Coding for Nucleotide Bases and Amino Acid Residues

	Nucleotide Bases	Codes
1.	Adenine	A
2.	Guanine	G
3.	Thymine	T
4.	Cytosine	C
5.	Uracil	U

s/n	Amino Acid	3-letter Code	One Letter Code		Amino Acid	3-letter Code	One Letter Code
1.	Alanine	Ala	A	11.	Leucine	Leu	L
2.	Arginine	Arg	R	12.	Lysine	Lys	K
3.	Asparagine	Ash	N	13.	Methionine	Met	M
4.	Aspartic Acid	Asp	D	14.	Phenylalanine	Phe	F
5.	Cysteine	Cys	C	15.	Proline	Pro	P
6.	Glutamine	Gln	Q	16.	Serine	Ser	S
7.	Glutamic Acid	Glu	E	17.	Threonine	Thr	T
8.	Glycine	Gly	H	18.	Tryptophan	Trp	W
9.	Histidine	His	H	19.	Tyrosine	Tyr	Y
10.	Isoleucine	Ile	I	20.	Valine	Val	V

5.3.2: Database Analyses

Database analysis are of two categories:

a. Genomic analysis: includes analysis of nucleic acid composition, restriction enzyme cleavage sites, transcriptional factors, promoter sites, secondary structure and sequence similarity searches.
b. Proteomics analysis includes determination of amino acid composition, sequence alignment, phylogenetic analysis, sequence similarity searches, prediction of secondary structure, motifs, profiles, domains and tertiary structure.

5.3.3: Database Organisation

Searching of sequence databases is one of the most common tasks with a newly discovered protein or nucleic acid. This is used to find if

a. the sequence is already in a database,
b. it is new, then to infer its structure (secondary and tertiary), and its function, and
c. presence of active sites, substrate-binding sites etc.

There is a vast amount of gene sequence data available (e.g. from genome sequence project). Two main databases that are widely used for novel gene discovery are;

a. high-throughput genomic databases, and

b. the expressed sequence tag (EST) databases. EST databases are single pass, partial sequences of 50-500 nucleotides from cDNA libraries. They provide direct window onto the expressed genome. EST sequences are generated by shotgun sequencing method. The sequencing is random and a sequence can be generated several times, and can be inaccurate.

5.3.4: Search Engines

- Altavista
- Google
- Yahoo
- Infoseek
- Medicine
- Research Index
- Pedro's Biomolecular Research Tools

5.3.5: Sequence Retrieval System (SRS)

The SRS has been created by Swiss Institute of Bioinformatics and the European Bioinformatic Institute, who have also created the Swiss-PROT database. SRS allows retrieval from an extensive catalogue of more than 75 public biological databases. The link button is SRS will allow you to get all the entries in one databank which are linked to an entry (or entries) in another database. Hyperlinks made links between the entries.

SRS is the operation of accessing the precise order of gene in a DNA, RNA and protein. Sequence information comes from many sources in which some are reliable than others in different aspects of sequence curation. Many methods have been employed in determining the genomic sequence including;

a. the Sangers,

b. manual and
c. automated methods

But the recent knowledge of bioinformatics proffered an efficient and effective way of accessing the sequence using computer. When the name of a gene or its ID number is given, it is possible to find and retrieve its DNA sequence. Sequence of genes are given an accession number as identification for database processing. Each sequence has its own unique accession number, but there may be some sequences that have more than one accession number. The most consistent source of sequence data comes from sequencing centres. The foundation of online computer database for storing and distributing sequence data has made bioinformatics an invaluable tool for sequence retrieval.

5.3.6: Types of Sequence Retrieval Databases

There are different types of sequence retrieval database. The main resources for sequence retrieval are three large databases called global nucleotide sequence storage. They include the following:

a. National Centre for Biotechnology Information (NCBI) database – (www.ncbi.nlm.nih.gov/)
b. European Molecular Biology Laboratory (EMBL) database – (www.ebi.ac.uk/embl/)
c. DNA Database of Japan (DDBJ) database – (www.ddbj.nig.ac.jp/) They collect all publicly available DNA, RNA and protein sequence data and make it available for free. Due to their daily exchange of data, they contain essentially the same data.
d. Other sequence database are genome centered database and protein database.
    i. Genome centered database includes the following.
        - NCBI genomes: Entrez Life Sciences Search Engine (US National Institutes of Health)- www.ncbi.nlm.nih.gov/sites/gquery
        - Ensemble genome browser (European Bioinformatics Institute)- www. ensemble.org iii. UCSC genome bioinformatics site (University of California at Santa Cruz) - www.genome.ucsc.edu.
    ii. Protein database includes
        - Swiss-Prot
        - TrEMBL
        - PDB

5.3.7: Search Sites

- DDBJ:(http://www.ddbj.nig.ac.jp/). (DNA Databank of Japan). Anucleic acid database.
- EBI: (http://www.ebi.ac.uk/). (European Bioinformatics Institute;UK, an outstation of the EMBL).
- EMBL: (http://www.ebi.ac.uk/). (European Molecular BiologyLaboratory; Germany).
- ExPASy: (http://www.expasy.ch/). Expert Protein Analysis System, aMolecular Biology Server, Switzerland, with SWISS-PROT, PROSITE, 2D-PAGE, an other proteomics tools. Key site forprotein sequence and structure information.
- GenBank: (http://www.ncbi.nlm.nih.gov/Web/GenBank/). GenBank of the National Institute of Health (NIH, USA genetic sequence database is an annotated collection of all publicly available DNA sequences. GenBank is a part of the international nucleotide sequence database, which is comprised of the DNA databank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL, Germany) and GenBank at NCBI, USA.
- GRAIL: (http://compbio.ornl.gov/Grail-1.3/). Gene Recognition and Assembly Internet Link software. A suite of tools designed toprovide analysis and putative annotation of

DNA sequences both interactively and through the use of automated computation.

- NCBI: (http://www.ncbi.nlm.nih.gov/). (National Center for Biotechnology Information; NIH, USA)
- OMIM: (http://www3.ncbi.nlm.nih-gov/Omim). On-line Mendelianinheritance in Man (for human genes and genomics at NCBI).
- Sanger Center: (http://www.sanger.ac.uk/DataSearch/). Genomicsequencing and genomics analysis server (UK).
- PUBMED: (http://www.ncbi.nlm.nih.gov/PubMed/). Covers mainlymedical literature.
- SWISS-PROT: (http://expasy.hcuge.ch/sprot/sprot-top.html). A proteinsequence database (Switzerland).

GenBank

It contains 7 million sequence record covering 9 million nucleotide bases. Unless the databases are easily searched and entries retrieved in a usable and meaningful format, the biological databases serve a little purpose. Moreover, efforts made on sequencing will not be meaningful if biological community as a whole cannot make use of the information hidden within millions of bases and amino acids. There are several database retrieval tools such as ENTREZ, LOCUSTLINK, TAXONOMY BROWSER, etc.

OMIM (Online Mendelian Inheritance in Man)

OMIM is a non-sequence-based information resource that is very much useful in genomics. It is a web-based electronic version of catalogue that contains thousands of entries for human genes and genetic disorders. It serves as a phenotypic companion to Human Genome Project. It was founded by Victor McKusick at the Hohns Hopkins University (McKusick, 1998). A concise textual information is provided by OMIM from the published literature on the conditionsof human having genetic disorders and full citation information. At the NCBI, the online version of OMIM is housed. Also, links are provided to Entrez from all references cited within each OMOM entry. Internet resource for OMIM is: http://www.ncbi.nlm.nih.gov/omim.

The OMIM cytogenetic and morbid maps present cytogenetic locations for those genes with published locations and provide an alphabetical list of all the diseasesdescribed in OMIM. Therefore, it is necessary to consider the results of web- based biology reported in the scientific literature in order to validate the findings generated through computer-based comparative analysis. Hence, integration of scientific data with the literature is an important step for creating a unified information resource in the life science. For this purpose, individuals are provided with a direct link from OMIM to PubMed, the NCBI literature system.

It is very easy to perform OMIM searches. A simple query is performed by search engine on the basis of one or more words typed into a search window. Consequently, a list of documents is returned containing the query words. The users can select one or more disorders from the list so as to see the full text of OMIM entry.

PubMed (Publishers on Medicine)

PubMed is a Web search interface that provides to over 11 million journal citation in MEDLINE and contains links to full text articles at participating publisher's web site.

PubMed provides web-based access to over 11 million citations, abstracts and indexing terms for journal articles in the biomedical sciences. It also includes links to full-text journals. At present about 20 million searches are conducted per month and over 1,40,000 users seek information daily through PubMed

5.3.8: Sequence Retrieval Tools

- BLAST: Basic Local Alignment and Search Tool (Home Page: NCBI, USA) sequence retrieval and sequence similarity search engine, which consists of a suite of programs — BLASTN (nucleotide BLAST), BLASTP (Protein BLAST), BLASTX (Translated BLAST), PhyloBLAST and PIR-BLAST.
- CLEVER: Command-line ENTREZ Version from NCBI. It is an interactive tool to browse ENTREZ database using only testinput/output.
- ENTREZ: (http://www3.ncbi.nlm.nih.gov/ENTREZ). ENTREZ is a powerful search engine, a part of NCBI server. The NCBI contains all the nucleotides and protein sequences in GenBank and Medicine. The program allows one to startwith only tentative set of keywords, or a sequence identifiedin the laboratory, and rapidly accesses a set of relevant list and list related database sequences.
- FASTA: (http://www2.igh.cnrs.fr/bin/fasta-guess.cgi). Sequence retrieval and similarity search database.
- FETCH: FETCH is sequence retrieval program that retrieves sequences from the GenBank and other databases. The program requires the exact locus name or accession numberof a sequence.
- LOOKUP: LOOKUP is a sequence retrieval program that uses SRS (Sequence Retrieval System) and is useful if the accession number is not known, but one wishes to download sequences of all proteins related to the query protein. LOOKUP identifies sequence by name, accession number,keyword, title, reference, feature or date. The output is a listof sequences.

ENTREZ

The integrated information database retrieval system of NCBI is called Entrez. It is most utilized of all biological database systems. Using Entrez system you can access literature, prepare genome map, sequences (both protein and nucleotides) and get structural data (3D). To be very clear, Entrez is not a database, but it is the interface through which all of its component databases can be accessed and traversed. Entrez has ability to retrieve the related sequence structures and references. The Entrez information space includes PubMed records, nucleotide and protein sequence data, three-dimensional structure information, and mapping information. The strength of Entrez lies in the fact that all of this information, across numerous component databases, can be accessed by issuing one and only one query. Entrez is able to offer integrated information retrieval through the use of two types of connection between database entries: neighboring and hard links.
List the codes for nucleotides bases.

Self-Assessment Exercises

1. Write on the two categories of database analysis.
2. Itemize the essence of database organization.

5.4: Summary

This section was able to provide the coding for all the nucleotides bases and amino acids. The two basic database analyses were discussed. The varying sequence retrieval systems were discussed with their web links. The basic sequence retrieval tools were listed and discussed.

5.5: References/Further Reading/Web Sources

- Apweiler R. (2005). *Sequence Databases. In: Bioinformatics* – A practical guide tothe analysis of genes and proteins. 3rd Ed. John Wiley & Sons Inc. Publication.

- Jin X. (2006) *Essential Bioinformatics*, Cambridge University Press. The Edinburgh Building, Cambridge 362pp

- Attwood, T. K., & Miller, C. J. (2002). Progress in bioinformatics and the importance of being earnest. *Biotechnol. Annu. Rev.* 8:1–54.

https://www.bing.com/ck/a?!&&p=913ab8bb4b496165JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTMxMw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Database+Search+and+sequence+Retrieval+Techniques+in+Bioinformatics&u=a1aHR0cHM6Ly9tZ2N1Yi5hYy5pbi9wZGYvbWF0ZXJpYWwvMjAyMDA0MDYwMTU2MzhlYzIyNzU5MWY5LnBkZg&ntb=1

https://www.bing.com/ck/a?!&&p=dbc4134485a24ae7JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTQxOQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Database+Search+and+sequence+Retrieval+Techniques+in+Bioinformatics&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8yOTEzNTU4ODFfU0VRVUVOQ0VfUkVUUklFVkFMX0FORF9BTkFMWVNJU19BX1VTRUZVTF9UT09MX0lOX0JJT0lORk9STUFUSUNTXy1fQV9SRVZJRVc&ntb=1

https://www.bing.com/videos/search?q=Database+Search+and+sequence+Retrieval+Techniques+in+Bioinformatics&&view=detail&mid=6140F51C0394D11C7F956140F51C0394D11C7F95&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DDatabase%2BSearch%2Band%2Bsequence%2BRetrieval%2BTechniques%2Bin%2BBioinformatics%26FORM%3DHDRSC6

https://www.bing.com/videos/search?q=Database+Search+and+sequence+Retrieval+Techniques+in+Bioinformatics&&view=detail&mid=FF9C3A2DC9453696DFDCFF9C3A2DC9453696DFDC&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DDatabase%2BSearch%2Band%2Bsequence%2BRetrieval%2BTechniques%2Bin%2BBioinformatics%26FORM%3DHDRSC6

5.6: Possible Answers to Self-Assessment Exercises

1.

a. Genomic analysis: includes analysis of nucleic acid composition, restriction enzyme cleavage sites, transcriptional factors, promoter sites, secondary structure and sequence similarity searches.
b. Proteomics analysis includes determination of amino acid composition, sequence alignment, phylogenetic analysis, sequence similarity searches, prediction of secondary structure, motifs, profiles, domains and tertiary structure.

2.

a) the sequence is already in a database,
b) it is new, then to infer its structure (secondary and tertiary), and its function, and
c) presence of active sites, substrate-binding sites etc.

Glossary

Accession number (in GenBank):
A unique identifier assigned to the entire sequence record when the record is submitted to GenBank. The GenBank accession number is a combination of letters and numbers that are usually in the format of one letter followed by five digits (e.g., M12345) or two letters followed by six digits (e.g., AC123456). The accession number for a particular record will not change even if the author submits a request to change some of the information in the record. Take note that an accession number is a unique identifier for a complete sequence record, while a sequence identifier, such as a Version, GI, or Protein ID, is an identification number assigned only to the sequence data. The NCBI Entrez System is searchable by accession number using the Accession [ACCN] search field.

Adenine:
One of the nitrogenous bases that has a double - ring structure, classified as a purine, found in DNA and RNA.

A DNA:
A more dehydrated form of DNA than the typical, B form. It is more compact, with 11 nitrogen bases per turn of the helix. RNA – DNA and RNA – RNA helices typically exist in this form.

BLAST (Basic Local Alignment Search Tool):
A fast technique for detecting ungapped subsequences that match a given query sequence.

Bootstrap test:
A test that allows for a rough quantification of confidence levels

Entrez:
An online resource provided by the National Center for Biotechnology Information (NCBI). It organizes GenBank sequences and links them to the literature sources in which they originally appeared.

FASTA format:
A sequence in FASTA format begins with a single - line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater - than symbol (>) in the first column.

GenBank:
A data bank of genetic sequences operated by a division of the National Institutes of Health

Homology:
(Strict): two or more biological species, systems, or molecules that share a common evolutionary ancestor;
(general): two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein) that they register along their lengths. Sequence homology searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage, and possible function of the query gene

MIM number (also known as MIM#, OMIM number, or McKusick code)
A unique six - digit number assigned to each entry listed in the catalog of human genes and genetic disorders, “Online Mendelian Inheritance in Man” (OMIM)

Motif:
A conserved element of a protein sequence alignment that usually correlates with a particular function. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be predictive of any subsequent occurrence of such a structural or functional region in any other novel protein sequence. A motif is built from particular combinations of secondary structures (typically, α - helices and β - sheets).

Query (sequence):
A DNA, RNA of protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced

Thymine:
One of the nitrogenous bases that has a single - ring structure, classified as a pyrimidine, found in DNA but not in RNA

End of the module Questions

1. In which year did the SWISSPROT protein sequence database begin?
(a) 1988
(b) 1985
(c) 1986
(d) 1987

2.Which of the following scientists created the first Bioinformatics database?
(a) Dayhoff
(b) Pearson
(c) Richard Durbin
(d) Michael.J.Dunn

3. The human genome contains approximately_____.
(a) 6 billion base pairs
(b) 5 billion base pairs
(c) 3 billion base pairs
(d) 4 billion base pairs

4. Which of the following tools is used for the identification of motifs?
(a) BLAST
(b) COPIA
(c) PROSPECT
(d) Pattern hunter

5. Proteomics refers to the study of _____.
(a) Set of proteins in a specific region of the cell
(b) Biomolecules
(c) Set of proteins
(d) The entire set of expressed proteins in the cell

6. Which of the following are not the application of bioinformatics?
(a) Drug designing
(b) Data storage and management
(c) Understand the relationships between organisms
(d) None of the above

7. The stepwise method for solving problems in computer science is called_____.
(a) Flowchart
(b) Algorithm
(c) Procedure
(d) Sequential design

8. The term Bioinformatics was coined by _____.
(a) J.D Watson
(b) Pauline Hogeweg

(c) Margaret Dayhoff
(d) Frederic Sanger

Answers
1. (d) 1987.
2. (a) Dayhoff.
3. (c) 3 billion base pairs.
4. (b) COPIA.
5. (d) The entire set of expressed proteins in the cell.
6. (d) None of the above.
7. (b) Algorithm.
8. (b) Pauline Hogeweg.

Module 2

Unit 1: Database searching algorithms (BLAST, FASTA)

Unit Structure

1.1: Introduction
1.2: Intended Learning Outcomes
1.3: Main Body
    1.3.1: Introduction to Database Searching algorithm
    1.3.2: Working of FASTA and BLAST
    1.3.3: BLAST (Basic Local Alignment Search Tool)
    1.3.4: FASTA
1.4: Summary
1.5: References/Further Readings/Web Sources
1.6: Possible Answers to Self-Assessment Exercises

1.1 Introduction

Currently, there are two major heuristic algorithms for performing database searches which are BLAST and FASTA. The number of DNA and protein sequences in public databases is very large. Searching a database involves aligning the query sequence to each sequence in the database, to find significant local alignment.

1.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. state four (4) similarities between BLAST and FASTA
b. state three (3) differences between BLAST and FASTA
c. Explain the procedure of using BLAST
d. Explain the steps in the use of FASTA

1.3: Main Body

1.3.1: Introduction to Database Searching algorithm

BLAST and FASTA are two similarity searching programs that identify homologous DNA sequences and proteins based on the excess sequence similarity. They provide facilities for comparing DNA and proteins sequences with the existing DNA and protein databases.

1.3.2: Working of FASTA and BLAST

- FASTA and BLAST are the software tools used in bioinformatics. Both BLAST and FASTA use a heuristic word method for fast pairwise sequence alignment.
- It works by finding short stretches of identical or nearly identical letters in two sequences. These short strings of characters are called words.
- The basic assumption is that two related sequences must have at least one word in common.
- By first identifying word matches, a longer alignment can be obtained by extending similarity regions from the words.
- Once regions of high sequence similarity are found, adjacent high-scoring regions can be joined into a full alignment.
- The main difference between BLAST and FASTA is that BLAST is mostly involved in finding of ungapped, locally optimal sequence alignments whereas FASTA is involved in finding similarities between less similar sequences.

1.3.3: BLAST (Basic Local Alignment Search Tool)

The BLAST program was developed by Stephen Altschul of NCBI in 1990 and has since become one of the most popular programs for sequence analysis. BLAST uses heuristics to align a query sequence with all sequences in a database. The objective is to find high-scoring ungapped segments among related sequences. The existence of such segments above a given threshold indicates pairwise similarity beyond random chance, which helps to discriminate related sequences from unrelated sequences in a database.

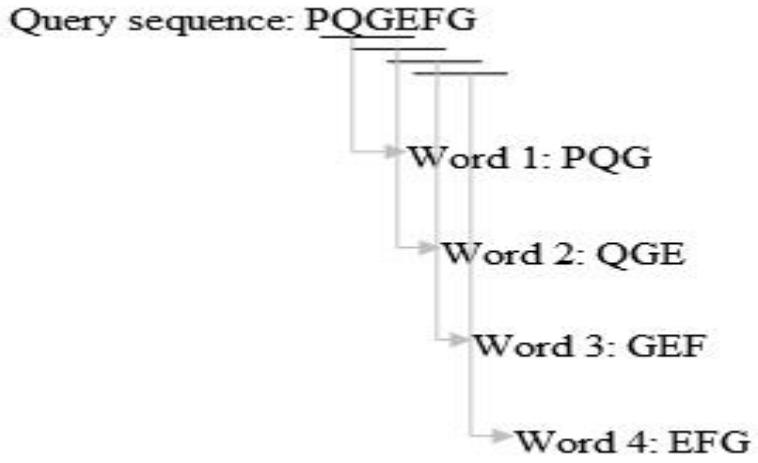
BLAST is popular as a bioinformatics tool due to its ability to identify regions of local similarity between two sequences quickly. BLAST calculates an expectation value, which estimates the number of matches between two sequences. It uses the local alignment of sequences.

The BLAST algorithm works by identifying short common regions of similarity (words) between a query sequence and database sequences. These words are of fixed lengths (4 amino acids for proteins and 11 base pairs for nucleotides) and are considered to be the minimum length needed to guarantee finding meaningful and significant patterns of similarity. Once a "word" is identified it is extended in either direction to searchfor extended regions of similarity between the query and the matched sequence, in order to determine the maximum level of identity between the two sequences being compared. This lining up of the two sequences is called an alignment and the solutions provided by BLAST are given in the form of alignments.

The variants of BLAST are;

a. **BLAST-N:** compares nucleotide sequence with nucleotide sequences
b. **BLAST-P:** compares protein sequences with protein sequences
c. **BLAST-X:** Compares nucleotide sequences against the protein sequences
d. **tBLAST-N**: compares the protein sequences against the six frame translations of nucleotide sequences
e. **tBLAST-X:** Compares the six frame translations of nucleotide sequence against the six frame translations of protein sequences.

BLAST Algorithm



To run the software, BLAST requires a query sequence to search for, and a sequence to search against (also called the target sequence) or a sequence database containing multiple such sequences. BLAST will find sub-sequences in the database which are similar to sub sequences in the query in typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides

Uses of BLAST

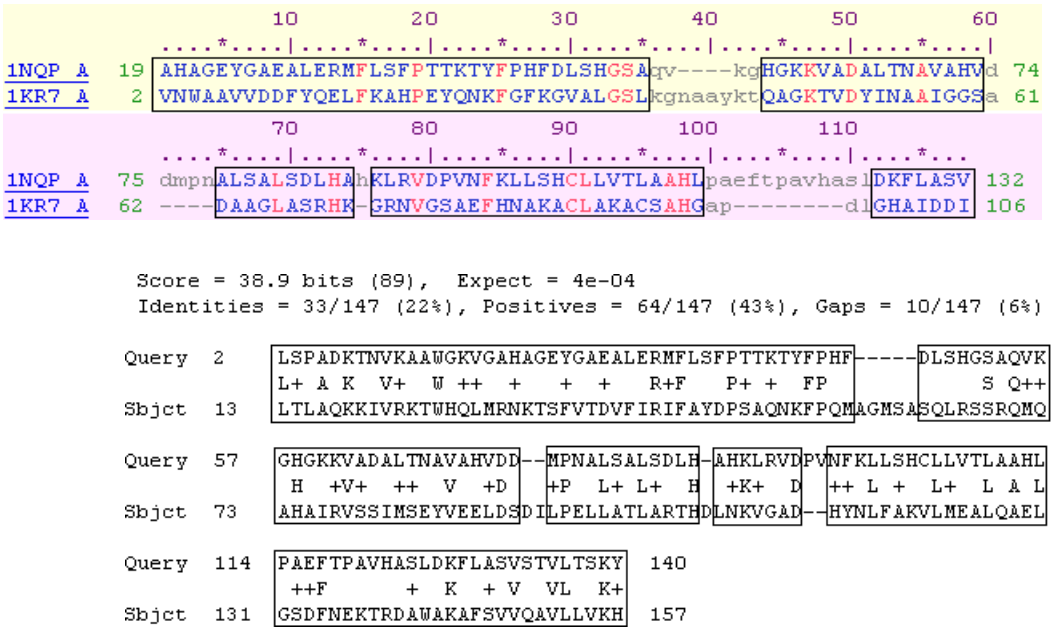
BLAST can be used for several purposes such as;

- Identifying species: Identifying species with the use of BLAST, can possibly correctly identify a species or find homologous species. This can be useful, for example, when researchers are working with a DNA sequence from an unknown species.

- Locating domains: when working with a protein sequence you can input it into BLAST, to locate known domain within the sequence of interest.
- Establishing phylogeny: u sing the results received through BLAST you can create a phylogenetic tree using the BLAST web-page. Phylogenies based on BLAST alone are less reliable than other purpose-built computational phylogenetic methods, so should only be relied upon for "first pass" phylogenetic analyses.
- DNA mapping: when working with a known species, and looking to sequence a gene at an unknown location, BLAST can compare the chromosomal position of the sequence of interest, to relevant sequences in the database(s). NCBI has a "Magic-BLAST" tool built around BLAST for this purpose.
- Comparison: when working with genes, BLAST can locate common genes in two related species, and can be used to map annotations from one organism to another.

It will be rewarding going through the following exercise using BLAST. Consider the DNA sequence below:
Note: Since it is DNA-DNA alignment, we use BLASTn

Sample DNA Sequence:
ATGGATTATCAAGTGTCAAGTCCAATCTATGACATCAATTATTATACATCGGA
GCCCTGCCAAAAAATCAATGTGAAGCAAATCGCAGCCCGCCTCCTGCCTCC
GCTCTACTCACTGGTGTTCATCTTTGGTTTTGTGGGCACATGCTGGTCATCC
TCATCCTGATAAACTGCAAAAGGCTGAAGAGCATGACTGACATCTACCTGCT
CAATGGCCATCTCTGACCTGTTTTTCCTTCTTACTGTCCCCTTCTGGGCTCA
CTATGCTGCCGCCCAGTGGGACTTTGGAAATACAATGTGTCAACTCTTGACA
GGGCTCTATTTTATAGGCTTCTTCTCTGGAATCTTCTTCACTCCTCCTGACAA
TCGATAGGTACCTGGCTGTCGTCCATGCTGTGTTTGCTTTAAAAGCCAGGAC
GGTCACCTTTGGGGTGGTGACAAGTGTGATCACTTGGGTGGTGGCTGTGTT
TGCGTCTCTCCCAGGAATCATCTTTACCAGATCTCAAAAAGAAGGTCTTCAT
TACACCTGCAGCTCTCATTTTCCATACAGTCAGTATCAATTCTGGAGAATTTC
CAGACATTAAAGATAGTCATCTTGGGGCTGGTCCTGCCGCTGCTTGTCATG
GTCATCTGCTACTCGGAATCCTAAAAACTCTGCTTCGGTGTCGAAATGAGAA
GAAGAGGCACAGGGCTGTGAGGCTTATCTTCACCATCATGATTGTTTATTTT
CTCTTCTGGGCTCCCTACAACATTGTCCTTCTCCTGAACACCTTCCAGGAAT
TCTTTGGCCTGAATAATTGCAGTAGCTCTAACAGGTTGGACCAAGCTATGCA
GGTGACAGAGACTCTTGGGATGACGCACTGCTGCATCAACCCCATCATCTA
TGCCTTTGTCGGGGAGAAGTTCAGAAACTACCTCTTAGTCTTCTTCCAAAAG
CACATTGCCAAACGCTTCTGCAAATGCTGTTCTATTTTCCAGCAAGAGGCTC
CCGAGCGAGCAAGCTCAGTTTACACCCGATCCACTGGGGAGCAGGAAATAT
CTGTGGGCTTGTGA
From the NCBI home page (http://www.ncbi.nlm.nih.gov) follow the “BLAST”link
Then from “Genomes” select “Human”
Paste the sequence above into the box (note that it is in FASTA format)
Click on “Begin Search”
Wait for a short while then try “Format”
Aligned (similar) regions are in rectangles



Comparing Amino Acid Sequences for Proteins in Different Organisms.

Part A. Hemoglobin Comparison

Open up your internet browser.

Go to the site: http://www.ncbi.nlm.nih.gov In the upper left corner of the website, search "all databases" for hemoglobin.You will get a page showing a variety of articles and genetic information e.g. "Pub Med" lists thousands of original papers related to hemoglobin, and "Nucleotide" reveals the sequence of the mRNA for hemoglobin in a variety of organisms.

HomoloGene. This will give you dozens of choices where you can compare the protein sequences of the various organisms. Just below the tab for "Limits" change the "display" to FASTA. You will then see the amino acid sequence of each organism (in a one letter code — for interpretation of the code go to http://www.chem.csustan.edu/chem4400/code.htm. Choose one of the comparisons (with at least five organisms) and using the human as a base, count the number of amino acid differences for the other organisms. Using this data, calculate the % of similarity of each organism to the human.

100 – (Number of differences/total amino acids X 100) = % similarity

Part B. Gene of Interest

Repeat the process above with any gene or protein that is of interest to you.

Part C. How many genes?

For an up-to-date report of the number of genes in various organisms, go back to the original NCBI site in step 2, and search "HomoloGene" and leave the space after the word "for" blank. You will see a recent count of identified genes for about 18 organisms.

Further Activity on BLASTn.

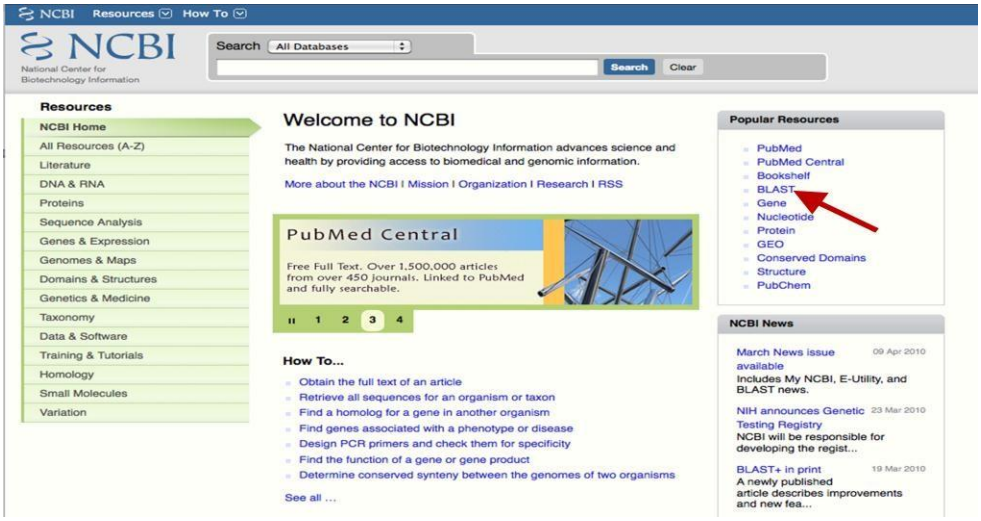
You will need to determine the source from which the following DNA fragmentwas obtained by comparing the "unknown 1" nucleotide sequence (query sequence) against a nucleotide sequence database using BLASTn.

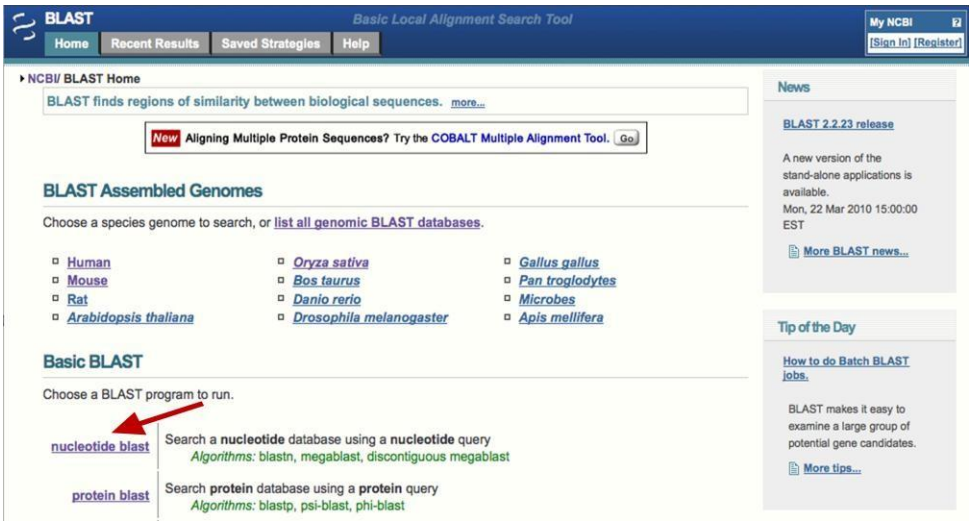
>Unknown 1 gagcaggtgcctcactatcgacaagccctagacatgatcttggacctggaacctgatgaagagctggaagaca
accccaaccagagtgacttgattgagcaggcggccgagatgctctatgggttgatccacgcccgctacatcctc
accaaccggggcattgcacaaatgttggaaaagtaccagcaaggagactttggctactgtcctcgagtatactg
tgagaaccagccgatgcttcccatcggcctttcggacatcccaggagaggccatggtgaagctctactgcccca
agtgcatggacgtgtacacacccaagtcctctaggcaccaccacacggatggcgcatacttcggcactggtttc
cctcacatgctcttcatggtgcatcccgagtaccggcccaagcggccggccaaccagtttgtgcccaggctctac
ggtttcaagatccatccaatggcctaccagctgcagctccaagccgccagcaacttcaagagcccagtcaaga
cgattcgctgagtgccctcccacctcctctgcctgtgacaccaccgtccctccgctgccacccctttcaggaagtctatggtttttagt

To perform BLAST on this sequence, follow the steps below:

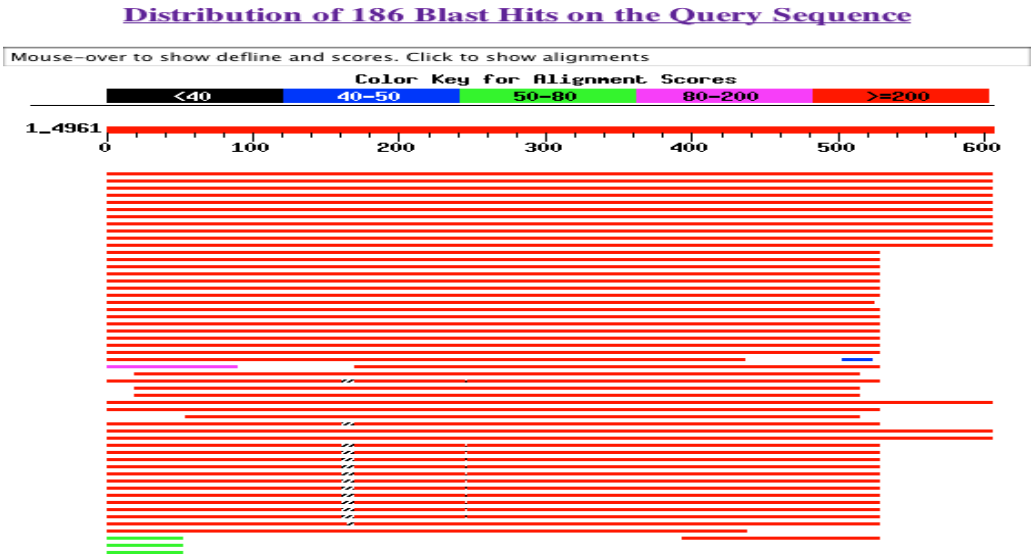
Go to the NCBI web page by typing the URL http://www.ncbi.nlm.nih.gov/
Go to BLAST.
Go to "**Nucleotide BLAST** under "**Basic BLAST.**"
Insert the query sequences (Unknown 1) in the window provided
Hit "BLAST."

You will be prompted with a window informing you that your request was successfully submitted to BLAST. It will assign you a "Request ID". **(*NOTE: The NCBI web page is changing constantly, so the figures presented in this example may differ from those you may encounter*).**





Get back to your BLAST result page. Look at the graphical view of the results. You will see a set of parallel horizontal bars of different colors and lengths. The color indicates the level of similarity between the query sequence and the matching sequence from the database. A red bar denotes a high similarity score while black denotes very low similarity between the two sequences.



Below the graph there is a list of sequences from the database producing significant alignments described by a one-line summary called “description.” The alignments are sorted by E-values (expected values) with the lowest score (0.0) presented at the top of the list. The E-values represent the probability of obtaining the particular alignment by chance rather than by real sequence similarity. Therefore the lower the value the more significant the alignment. E-values are very useful in helping one to decide what results are more meaningful. In addition to E-values, the description provides a score value, which is another statistical value to represent the alignment. The score of

sequence was obtained, and a description of the sequence. Following the description is the statistical information for the alignment, including the score, the E- value, the "identities" (the ratio of the number of nucleotides considered in the alignment to the number of well-matched nucleotides). In this case the identity is 100%, indicating that 605 nucleotides analyzed form the query sequence were identical to 605 nucleotides in the sequence of Rat casein II beta subunit. If you look at the alignment you will see that both sequences are identical. The low E-value (0.0) together with the high score (1199) and the 100% identity strongly suggest that the "Unknown 1" sequence comes from a rat and is a portion of the casein kinase II beta subunit (CK2) mRNA sequence.

You can obtain the complete record (the source for the matching sequence). Click on the NCBI-gi accession number in the sequence description that links to the full sequence of the gene. You will be prompted with a page containing a completedescription of the sequence, including the accession number under which the sequence is stored in the database, the name of the gene and of the organism from where the sequence was obtained, a complete taxonomic origin for the organism, a reference to the publication reporting the sequence with comments about the sequence, and the protein and nucleotide sequences. This page also includes "**Links**" to other relevant sections on NCBI like PubMed, Taxon Browser, Gene etc.

1.3.4: FASTA

FASTA stands for fast-all" or "FastA": FASTA (pronounced FAST-AYE) is a suite of programs for searching nucleotide or protein databases with a query sequence. FASTA itself performs a local heuristic search of a protein or nucleotide database for a query of the same type. FASTX and FASTY translate a nucleotide query for searching a protein database.

FASTA is one of the first widely-used database similarity search tools. FASTA (or FastA), an abbreviation for 'Fast-All', is a sequence alignment tool that takes nucleotide or protein sequences as input and compares it with existing databases. It was first developed by David J. Lipman and William R. Pearson in 1985 and has since been refined and adapted for various applications.

It was the first database similarity search tool developed, preceding the development of BLAST. FASTA is another sequence alignment tool which is used to search similarities between sequences of DNA and proteins. FASTA uses a "hashing" strategy to find matches for a short stretch of identical residues with a length of k. The string of residues is known as ktuples or ktups, which are equivalent to words in BLAST, but are normally shorter than the words.

Typically, a ktup is composed of two residues for protein sequences and six residues for DNA sequences. The query sequence is thus broken down into sequence patterns or words known as k-tuples and the target sequences are searched for these k-tuples in order to find the similarities between the two. FASTA is a fine tool for similarity searches. These methods are not guaranteed to find the optimal alignment or true homologs, but are 50–100 times faster than dynamic programming.

The text-based file format for representing nucleotide or protein sequences, which originates from the FASTA program, has now become a standard in bioinformatics. Many other sequence database search tools also use the FASTA file format.

FASTA Programs

FASTA was originally developed for comparing protein sequences. The original program was referred to as FASTP. It quickly became a popular tool for sequence alignment and database searching. The program has been continually updated and improved. There are now different FASTA programs available, each used for different types of sequence searches:

a. **FASTA** compares a DNA query sequence against a database of DNA sequences or a protein query sequence against a database of protein sequences using the FASTA algorithm.
b. **SSEARCH** performs protein-protein or DNA-DNA comparisons using the Smith-Waterman algorithm.
c. **GGSEARCH/ GLSEARCH** works using a global alignment algorithm (GGSEARCH) or a combination of global and local alignment algorithms (GLSEARCH) to compare protein and nucleotide sequences.
d. **FASTX/ FASTY** compares a DNA sequence and a database of protein sequences by translating the DNA sequence into three frames and allowing gaps and frameshifts.
e. **TFASTX/ TFASTY** compares a protein sequence and a database of DNA sequences. The DNA sequence is translated in six frames – three in the forward direction and three in the reverse direction.
f. **FASTF/ TFASTF** compares mixed peptide sequences against a protein (FASTF) or translated DNA (TFASTF) databases.
g. **FASTS/ TFASTS** compares a set of short peptide fragments against the protein (FASTS) or translated DNA (TFASTS) databases.

How FASTA Works

FASTA works by comparing a query sequence to a database of sequences to identify similar matches. The program uses a heuristic algorithm to quickly search the database and identify the most significant matches.

The working mechanism of FASTA is described in the following steps:

Step 1: Identifying Regions

The first step is identifying regions with high similarity by creating a lookup table for the query sequence. This step is also called hashing step. To create the lookup table, the query sequence is first broken down into smaller words known as k-tuples (ktup). When the ktup value is increased, the number of background word hits is reduced. By reducing the number of these background word hits, the algorithm can focus on the more relevant hits, enhancing the overall search speed. k-tuple is usually 2 for proteins and 6 for nucleotide sequences. Once the lookup table is created, it is used to identify matches between the k-tuples in the query sequence and the database sequences. Similar regions are represented as diagonals in a two-dimensional matrix. The ten regions with the highest density of word matches are the high-similarity regions, and these best ten diagonals are saved.

Step 2: Re-Scoring

In the second step, the ten best diagonals are rescored using suitable scoring matrices. For protein, BLOSUM50 or PAM matrix is used; for DNA sequences, the identity matrix is used. A subregion with the highest score is identified for each of the rescanned diagonal regions. These high-scoring subregions within the diagonals are called initial regions.

Step 3: Joining Threshold

Next, a score cutoff or the joining threshold is applied that excludes segments unlikely to be part of the final alignment. The library sequences are ranked based on their initial scores. The regions with initial scores above the pre-set threshold are selected and checked to see if they can be joined together. This step introduces gaps between the diagonals while applying gap penalties. The score of the gapped alignment is calculated by subtracting a penalty for each gap, which is used to rank the database sequences by similarity.

Step 4: Final Alignment

Finally, the gapped alignment is refined to produce the final alignment. This is done by using the banded Smith-Waterman algorithm, which is a dynamic programming algorithm that calculates the optimal score (opt) for alignment. This score is used for statistical calculations.

Statistical Significance and FASTA

- FASTA also provides an estimate of the statistical significance of each alignment found. It is evaluated using the E-value, which measures the likelihood of obtaining a sequence alignment score by chance. The smaller the E-value, the more significant the alignment.
- E-value is not the only statistical parameter. FASTA also uses other statistical measures, such as the bit score and the similarity score based on the scoring matrix and gap penalties, to evaluate the significance of sequence alignments.
- The FASTA output also includes an additional statistical parameter, the Z-score, which represents the number of standard deviations from the mean score of the database search. A higher Z-score value indicates a more significant match.

Applications of FASTA
FASTA has a wide range of applications. Some are:

- FASTA can be used in the sequence alignment to identify regions of similarity. This is useful for identifying conserved regions in DNA or protein sequences, which can help to identify functional domains or motifs. Identifying these functional domains or motifs can provide insights into the biological function of the sequence.
- FASTA can be used to search large databases of sequences to find matches to a given query sequence. This helps to identify homologous sequences, which can help to predict the function of a newly identified sequence.
- FASTA can construct phylogenetic trees by aligning sequences from different species and identifying evolutionary relationships between them.

Who developed BLAST in the year 1990? How many types of BLAST variants exist?

Self-Assessment Exercises

1. Explain four (4) purposes of BLAST
2. How many types of FASTA programmes exist?

1.4: Summary

This section examined the basic two (2) database searching algorithm the BLAST and FASTA. The different types and mode of action in each case.

1.5: References/Further Reading/Web Sources

- Barton, G. J. (1996). Protein sequence alignment and database scanning. Protein structure prediction: A practical approach, 31-63.
- Casey, R. M. (2005). *"BLAST Sequences Aid in Genomics and Proteomics"*. Business Intelligence Network.
- Lipman, DJ & Pearson, WR (1985). "Rapid and sensitive protein similarity searches". Science. 227 (4693): 1435 41. doi:10.1126/science.2983426. PMID 2983426.
- Oehmen, C. S & Baxter, D. J. (2013). *"ScalaBLAST 2.0: Rapid and robust BLAST calculations on multiprocssor_systems"*. Bioinformatics. 29 (6): Science Watch. July–August 2000.
- Lloyd, A. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43). Briefings in Bioinformatics. 2. 10.1093/bib/2.4.407.
- Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806087

https://www.bing.com/ck/a?!&&p=890ca54439a49f5fJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTM5Nw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Database+searching+algorithms+in+Bioinformatics&u=a1aHR0cHM6Ly9kbC5hY20ub3JnL2RvaS9ib29rLzEwLjU1NTUvMTU0MTkyNA&ntb=1

https://www.bing.com/ck/a?!&&p=e7627c8528e7b048JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTQ5Ng&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Database+searching+algorithms+in+Bioinformatics&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8yNzA1NDgyNjVfQmlvaW5mb3JtYXRpY3NfQWxnb3JpdGhtcw&ntb=1

https://www.bing.com/videos/search?q=Database+searching+algorithms+in+Bioinformatics+BLAST&&view=detail&mid=94BD97536FEBC904D63894BD97536FEBC904D638&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DDatabase%2Bsearching%2Balgorithms%2Bin%2BBioinformatics%2BBLAST%26FORM%3DHDRSC6

https://www.youtube.com/watch?v=jL_dbaGytNE

1.6: Possible Answers to Self-Assessment Exercises

1.

- Identifying species: Identifying species with the use of BLAST, can possibly correctly identify a species or find homologous species.
- Locating domains: when working with a protein sequence you can input it into BLAST, to locate known domain within the sequence of interest.
- Establishing phylogeny: u sing the results received through BLAST you can create a phylogenetic tree using the BLAST web-page.
- DNA mapping: when working with a known species, and looking to sequence a gene at an unknown location, BLAST can compare the chromosomal position of the sequence of interest, to relevant sequences in the database(s).
- Comparison: when working with genes, BLAST can locate common genes in two related species, and can be used to map annotations from one organism to another.

2.

7

Unit 2: Pairwise and Multiple Sequence Alignment

Unit Structure

2.1: Introduction
2.2: Intended Learning Outcomes
2.3: Main Body
    2.3.1: Types of sequence analysis
    2.3.2: Sequence Alignment Analysis
    2.3.3: Types of Sequence Alignment
2.4: Summary
2.5: References/Further Readings/Web Sources
2.6: Possible Answers to Self-Assessment Exercises

2.1 Introduction

Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping processes. The sequence analysis implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer.

2.2 Intended Learning Outcomes (ILOs)

At the end of this section, student should be able to;

a. distinguish between pairwise and multiple sequence alignment
b. distinguish between local and global sequence alignment

2.3 Main Body

2.3.1: Types of sequence analysis

a. Sequence alignment: is a way of arranging the sequence of DNA, RNA or protein to identify region of similarity that may be a consequence of functional, structural, or evolutional relationship between the sequence. it involves the identification of the correct location of deletion and insertions that have occurred of either of the two lineages since the divergence from a common ancestor.

b. Sequence assembly: refers to the reconstruction of a DNA sequence by aligning and merging small DNA fragments. It is an integral part of modern DNA sequencing. Since presently-available DNA sequencing technologies are ill-suited for reading long sequences, large pieces of DNA (such as genomes) are often sequenced by

i. cutting the DNA into small pieces,
ii. reading the small fragments, and
iii. reconstituting the original DNA by merging the information on various fragments.

2.3.2: **Sequence Alignment Analysis**

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. An alignment basically is used for building phylogenetic trees, looking for sites of interest/conservation within a gene (motifs, binding sites, etc., identifying positive/negative selection and references for short read analysis.

Sequence alignment is the process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology. Sequence similarity analysis is the single most powerful method for structural and functional inference of homology between proteins databases.

Sequence similarity analysis allows the inference of homology between proteins and homology can help one to infer whether the similarity in sequences would have similarity in function. Methods of analysis can be grouped into two categories

— sequence alignment-based search,
— profile- based search.

Fundamentally, sequence-based alignment searches are string-matching procedures. A sequence of interest (the query sequence) is compared with sequences (targets) in a databank-either pair-wise (two at a time) or with multiple target sequences, by searching for a series of individual characters. Two sequences are aligned by writing them across a page in two rows. Identical or similar characters are placed in the same column and non-identical characters can be placed opposite a gap in the other sequence. Gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. In optimal alignment, non-identical characters and gaps are placed to bring as many identical or similar characters as possible into vertical register.

The objective of sequence alignment analysis is to analyze sequence data to make reliable prediction on protein structure, function and evolution vis a vis the three-dimensional structure. Such studies include detection of orthologous (same function in different species), and paralogous (different but related functions within an organisms) features.

Sequence-similarity searches of unknown sequences to databases are commonly used in biological laboratories as the first approach to obtain clues about the function of a newly sequenced genes. The National Center for Biotechnology Information (NCBI) provides the Basic Local alignment Search Tool (BLAST) that allows for rapid comparison of nucleotide and protein query sequences to database sequences.

2.3.3: Types of Sequence Alignment

1. Pairwise sequence alignment

Pairwise alignment is a tool designed for performing sequence alignments in a wide variety of combinations. It implements sequence to sequence, sequence to profile and profile to profile alignments with optional support of secondary structure. Usually, a pairwise alignment is done and a clustering method is used to create a guide tree. The guide tree is used to create a succession of pairwise alignments starting with the two closest sequences and ending with the most distant from these. Pairwise alignment methods are important largely in the context of a database search but for the analysis of individual protein families, multiple alignment methods are critical.

Pair-wise alignment is a fundamental process in sequence comparison analysis. Pair-wise alignment of two sequences (DNA or protein) is relatively straightforward computational problem. In a pair-wise comparison, if gaps or local alignments are not considered (i.e., fixed-length sequences), the optimal alignment method can be tried and the number of computations required for two sequences is roughly proportional to the square of the average length. The problem becomes complicated, and not feasibleby optimal alignment method, when gaps and local alignments are considered.

That a program may align two sequences is not a proof that a relationship exists between them. Statistical values are used to indicate the level of confidence that should be attached to an alignment. A maximum match between two sequences is defined to be the largest number of amino acids from on protein that can be matched with those of another protein, while allowing for all possible deletions. Apenalty is introduced to provide a barrier to arbitrary gap insertion.

In summary, pairwise sequence alignment is

a. An alignment procedure comparing two biological sequences of either protein, DNA or RNA.
b. It can be generally categorized as global or local alignment methods.

c. Comparatively simple algorithm is used.
d. A general global alignment technique is the Needleman–Wunsch algorithm. A general local alignment method is Smith–Waterman algorithm.
e. Applications:
   - Primarily to find out conserved regions between the two sequences.
   - Similarity searches in a database.
f. Examples of pairwise alignment tools: LALIGN, BLAST, EMBOSS Needle; EMBOSS Water

Types of Pairwise Sequence Alignment

a. Global Alignment

Global alignment is an alignment of two nucleic acid or protein sequences over their entire length. The Needleman-Wunsch algorithm (GAP program) is one of the methods to carry out pair-wise global alignment of sequences by comparing a pair of residues at a time. Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array (one sequence along x-axis and the other along y-axis), and pathways through the array represent all possible comparisons (every possible combination of match, mismatch and insertion and deletion). Statistical significance is determined by employing a scoring system; for a match = 1 and mismatch = 0 (or any other relative scores) and penalty for a gap.

In summary, in global alignment,

a. an attempt is made to align the entire sequence (end to end alignment).
b. A global alignment contains all letters from both the query and target sequences.
c. If two sequences have approximately the same length and are quite similar, they are suitable for global alignment.

Examples of Global alignment tools include:

a. EMBOSS Needle

b. Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)

b. Local Alignment

Local alignment is an alignment of some portion of two nucleic acid or protein sequences. *Smith-Waterman* algorithm is a variation of the dynamic programming approach to generate local optimal alignments, best alignment method for sequences for which no evolutionary relatedness is known. The program finds the region or regions of highest similarity between two sequences, thus generating one or more islands of matches or sub-alignment in the aligned sequences. Local alignments are more suitable and meaningful for

a. aligning sequences that are similar along some of their lengths but dissimilar in others,
b. sequences that differ in length, or
c. sequences that share conserved regions or domains.

Table 2: Differences between Global and Local Sequence Alignment

BASIS OF COMPARISON	GLOBAL SEQUENCE ALIGNMENT	LOCAL SEQUENCE ALIGNMENT
Description	In global alignment, an attempt is made to align the entire sequence (end to end alignment).	Finds local regions with the highest level of similarity between the two sequences.
Examples of Tools	-EMBOSS Needle -Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)	-BLAST -EMBOSS Water -LALIGN
Function	A global alignment contains all letters from both the query and target sequences.	A local alignment aligns a substring of the query sequence to a substring of the target sequence.
Two Sequences	If two sequences have approximately the same length and are quite similar, they are suitable for global alignment.	Any two sequences can be locally aligned as local alignment finds stretches of sequences with high level of matches without considering the alignment of rest of the sequence regions.
Suitability	Suitable for aligning two closely related sequences.	Suitable for aligning more divergent sequences or distantly related sequences.
Use	Global alignments are usually done for comparing homologous genes like comparing two genes with same function (in human vs. mouse) or comparing two proteins with similar function.	Used for finding out conserved patterns in DNA sequences or conserved domains or motifs in two proteins.
General Technique	A general global alignment technique is the Needleman–Wunsch algorithm.	A general local alignment method is Smith–Waterman algorithm.

2. Multiple sequence alignment.

Multiple sequence alignment is an alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. The goal of multiple sequence alignment process is to generate a concise, information-rich table of sequence data to obtain relatedness of sequences to a gene family.

Homologous sites between sequences are aligned. This is achieved by inserting gaps. Alignments allow for identification of regions of similarity between sequences. They identify indels (insertion and deletions) caused by DNA/RNA replication. Multiple sequence alignment is progressive. The main principle underlying popular algorithms for multiple alignments is hierarchical clustering that roughly approximates the phylogenetic tree and guides the alignment.

The sequences are first compared using a fast method (e.g. FASTA) and clustered by similarity scores to produce a guide tree. Sequences are then aligned step-by-step in a bottom-up succession, starting from terminal clusters in the tree and proceeding to the internal nodes until the root is reached. Once two sequences are aligned, their alignment is fixed and treated essentially as a single sequence with a modification of dynamic programming. Thus, the hierarchical algorithms essentially reduce the O (n k) multiple alignment problem to a series of O (n 2) problems, which makes the algorithm feasible but potentially at the price of alignment quality. The hierarchical algorithms attempt to minimize this problem by starting with most similar sequences where the likelihood of incorrect alignment is minimal, in the hope that the increased weight of correctly aligned positions precludes errors even on the subsequent steps.

In summary, multiple sequence alignment is

a. An alignment procedure comparing three or more biological sequences of either protein, DNA or RNA.
b. MSA is generally a global multiple sequence alignment.
c. Complex sophisticated algorithm is used.
d. A technique called progressive alignment method is employed. In this approach, a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair, and so on.
e. Applications:
    - To detect regions of variability or conservation in a family of proteins;
    - Phylogenetic analysis (inferring a tree, estimating rates of substitution, etc.)
    - Detection of homology between a newly sequenced gene and an existing gene family prediction of protein structure;
    - Demonstration of homology in multigene families.
f. Examples of Multiple Sequence Alignment tools: MUSCLE; T-Coffee; MAFFT; CLUSTALW.

Table 3: Pairwise Alignment vs Multiple Sequence Alignment

Basis of Comparison	Pairwise Alignment	Multiple Sequence Alignment (MSA)
Description	An alignment procedure comparing two biological sequences of either protein, DNA or RNA	An alignment procedure comparing three or more biological sequences of either protein, DNA or RNA
Category	Pairwise alignments can be generally categorized as global or local alignment methods.	MSA is generally a global multiple sequence alignment
Algorithm	Comparatively simple algorithm is used	Complex sophisticated algorithm is used
Techniques	A general global alignment technique is the Needleman–Wunsch algorithm. A general local alignment method is Smith–Waterman algorithm.	A technique called progressive alignment method is employed. In this approach, a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair, and so on.
Application	Applications: a) Primarily to find out conserved regions between the two sequences b) Similarity searches in a database	Applications: a) To detect regions of variability or conservation in a family of proteins b) Phylogenetic analysis (inferring a tree, estimating rates of substitution, etc.) c) Detection of homology between a newly sequenced gene and an existing gene family prediction of protein structure d) Demonstration of homology in multigene families
Example of Tools	Examples of pairwise alignment tools: • LALIGN • BLAST • EMBOSS Needle • EMBOSS Water	Examples of Multiple Sequence Alignment tools: • MUSCLE • T-Coffee • MAFFT • CLUSTALW

The two methods of sequence alignment analysis are?
The objective of sequence alignment analysis is?

Self-Assessment Exercises

1. List the two types of sequence alignment.
2. List the two types of pairwise sequence alignment

2.4: Summary

This section was able to explain the pairwise and multiple sequence alignment. The varying difference of the two alignments were presented in tabular form.

2.5: References/Further Reading/Web Sources

- Oehmen, C. S & Baxter, D. J. (2013). *"ScalaBLAST 2.0: Rapid and robust BLAST calculations on multiprocssor_systems*". Bioinformatics. 29 (6): Science Watch. July–August 2000.

- Lloyd, A. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43). Briefings in Bioinformatics. 2. 10.1093/bib/2.4.407.

- Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806087

https://www.bing.com/ck/a?!&&p=6fe45a9dbbf0e9fcJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTE4OQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Pairwise+and+multiple+sequence+alignments+pdf&u=a1aHR0cHM6Ly93d3cubmNiaS5ubG0ubmloLmdvdi9DQkJyZXNlYXJjaC9Qcnp5dHlja2EvZG93bmxvYWQvbGVjdHVyZXMvUENCX0xlY3QwNV9NdWx0aXBfQWxpZ24ucGRm&ntb=1

https://www.bing.com/ck/a?!&&p=d0249d558199b632JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTI0OQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Pairwise+and+multiple+sequence+alignments+pdf&u=a1aHR0cHM6Ly9hcnhpdi5vcmcvcGRmLzA5MDEuMjc0Ny5wZGY&ntb=1

https://www.bing.com/ck/a?!&&p=7313c5a0f9796a48JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTU2Ng&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&u=a1L3ZpZGVvcy9zZWFyY2g_cT1QYWlyd2lzZSthbmQrbXVsdGlwbGUrc2VxdWVuY2UrYWxpZ25tZW50cyZkb2NpZD02MDM1NDE4NjU0NjExMjc1MTQmbWlkPTgzQ0FEQjk1RkI0Q0JDMjREQTk3ODNDQURCOTVGQjRDQkMyNERBOTcmdmlldz1kZXRhaWwmRk9STT1WSVJF&ntb=1

https://www.bing.com/ck/a?!&&p=d7afe0ab619cf384JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTU2Nw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&u=a1L3ZpZGVvcy9zZWFyY2g_cT1QYWlyd2lzZSthbmQrbXVsdGlwbGUrc2V

xdWVuY2UrYWxpZ25tZW50cyZkb2NpZD02MDM1MDE0MTk3NTQzNTg2NzkmbWlkPTk2RUExMkFDNDI0OUE1RkMyQkZFOTZFQTEyQUM0MjQ5QTVGQzJCRkUmdmlldz1kZXRhaWwmRk9STT1WSVJF&ntb=1

 **2.6: Possible Answers to Self-Assessment Exercises**

1. a. pairwise sequence alignment
b. multiple sequence alignment

2. a. local sequence alignment
b. global sequence alignment

Unit 3: Phylogenetic analysis & Data mining in novel genomes

Unit Structure

3.1: Introduction
3.2: Intended Learning Outcomes
3.3: Main Body
    3.3.1: Introduction
    3.3.2: Importance of Phylogenetics
    3.3.3: Importance of Phylogenetics
    3.3.4: Scope of Phylogenetic analyses
    3.3.5: Phylogeny
    3.3.6: Phylogenetic Tree
    3.3.7: History of Phylogenetic Tree
    3.3.8: Parts of a Phylogenetic Tree
    3.3.9: Approaches to make Phylogenetic Tree
    3.3.10: Steps for Phylogenetic Analysis
    3.3.11: Types of Phylogenetic Tree
    3.3.12: Significance of Phylogenetic tree
    3.3.13: Applications of the phylogenetic tree
    3.3.14: Limitations of Phylogenetic tree
    3.3.15: Phenetic and Cladistic Methods in Phylogenetic Analysis
    3.3.16: Cladogram
    3.3.17: Difference between a Cladogram and a Phylogenetic Tree
    3.3.18: Data Mining
    3.3.19: Application of Data Mining in Bioinformatics
3.4: Summary
3.5: References/Further Readings/Web Sources
3.6: Possible Answers to Self-Assessment Exercises

3.1 Introduction

Phylogenetic analysis is the analysis of evolutionary relationships. In sequence alignment, evolutionary theory was assumed as the basis. This assumption stems from the believe that similarity implies co-ancestry.

3.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. enumerate the importance of phylogenetic analysis
b. distinguish the various components of phylogenetic tree
c. explain the steps for phylogenetic analysis.
d. Define data mining,
e. List and explain the classes of data mining

3.3 Main Body

3.3.1: Introduction

An understanding of evolutionary theory, therefore, is critical to appropriate interpretation of bioinformatics results. However, it has been well documented that the closest BLAST hit is not often the nearest neighbor; that is, the sequence that is listed first in the BLAST output as being similar to the query sequence is not necessarily the closest according to phylogenetic analysis. The issue of convergent evolution, whereby evolutionary pressure forces sequences to be similar despite the fact that they had different ancestors, should be borne in mind. Knowing how to handle such issues and appreciating the need to perform robust evolutionary analyses is an important component of bioinformatics analysis.

3.3.2: Importance of Phylogenetics

Phylogenetics is important because it enriches our understanding of how genes, genomes, species (and molecular sequences more generally) evolve. Through phylogenetics, we learn not only how the sequences came to be the way they are today, but also general principles that enable us to predict how they will change in the future. This is not only of fundamental importance but also extremely useful for numerous applications.

3.3.3: Importance of Phylogenetics

a. **Classification:** Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. Phylogenetics now informs the Linnaean classification of new species.
b. **Forensics**: Phylogenetics is used to assess DNA evidence presented in court cases to inform situations, e.g. where someone has committed a crime, when food is contaminated, or where the father of a child is unknown.
c. **Identifying the origin of pathogens**: Molecular sequencing technologies and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to new recommendations for public health policy.

d. **Conservation**: Phylogenetics can help to inform conservation policy when conservation biologists have to make tough decisions about which species, they try to prevent from becoming extinct.
e. **Bioinformatics and computing**: Many of the algorithms developed for phylogenetics have been used to develop software in other fields.

3.3.4: Scope of Phylogenetic analyses

a. interpret bioinformatics analyses involving sequence alignment in an evolutionary context.
b. Differentiate between the common phylogenetic methods
c. Explain the common phylogenetic software and their uses.

3.3.5: Phylogeny

The field of phylogeny has the goals of working out the relationships among species, populations, individuals or genes. Relationship is considered in the sense of kinship or genealogy. This means assignment of a scheme of descendants of a common ancestor. Evolutionary relationship gives us a glimpse of historical development of life. Several characters can be used for phylogenetic studies. Indeed, many molecular properties have been used. Serological and cross-reactivity was used from the beginning of the last century until superseded by direct use of sequences. Today, DNA sequences provide the best measures of similarities among species for phylogenetic analysis. Phylogenetics is sometimes called cladistics, because the word clade a set of descendants from a single ancestor is derived from the greek word for branch.

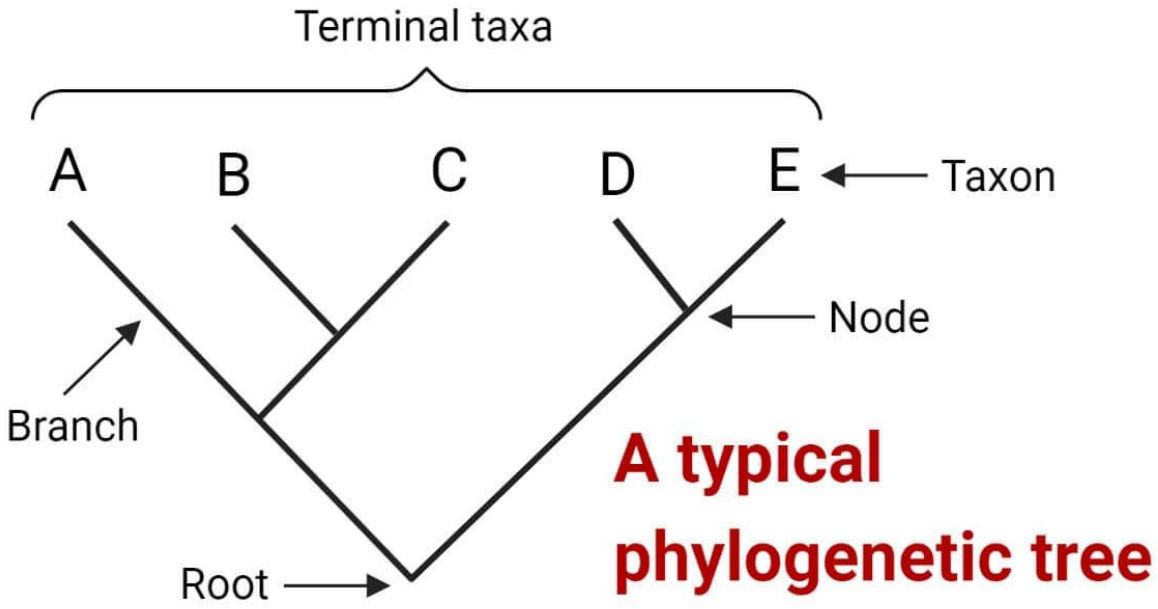
3.3.6: Phylogenetic Tree

Phylogenetic relationship is usually expressed as a tree called phylogenetic or evolutionary tree. In a tree, taxa are grouped into clades through a tree that comprises of a series of branches and nodes that mark bifurcating points in the branches. A characteristic phylogenetic tree is made up of a root, nodes, branches and leaves

3.3.7: **History** of Phylogenetic Tree

Ancient beliefs of a ladder-like evolution from lower to higher life forms gave rise to the concept of a "tree of life" (such as in the Great Chain of Being). A "paleontological chart" outlining the geological relationships between plants and animals can be found in Edward Hitchcock's book Elementary Geology as one of the earliest examples of "branching" phylogenetic trees (first edition: 1840). In his ground-breaking book The Origin of Species, Charles Darwin (1859) also created one of the first pictures and played a significant role in popularising the idea of an evolutionary "tree." The concept that speciation occurs through the adaptive and semi-random splitting of lineages is successfully communicated by tree diagrams, which are still used by evolutionary biologists to represent evolution more than a century after they were first used. The taxonomy of species has evolved to become more dynamic and less static.

An evolutionary or phylogenetic tree both have the same names. It is a branching diagram or tree that represents the relationships that have developed over time between different biological species or other entities based on the similarities and differences in their physical or genetic traits. One phylogenetic tree, which shows a common ancestor for all life on Earth, is present.



3.3.8: **Parts of a Phylogenetic Tree**

A phylogenetic tree consists of the following components:

a. Every branch denotes a lineage (single line of descent).
b. Each node on a branch (also known as a branch point) reflects the split in two or more evolutionary lineages from a common ancestor.
c. A taxon (plural: taxa), which might be a species or a group at any hierarchical level, is represented by each leaf, also known as a terminal node. Sister taxa are groups of related taxa that diverge from a single node. They stand for species that have a more recent common ancestor than other groups. Sister taxa have the closest relationships among their members. Taxa close to the root are called basal taxa. They are examples of species or groups that, early in the course of their evolutionary histories, diverge from the other members of the group.
d. The most recent common ancestor of all taxa is shown as the tree's root. Some phylogenetic trees do not have roots.

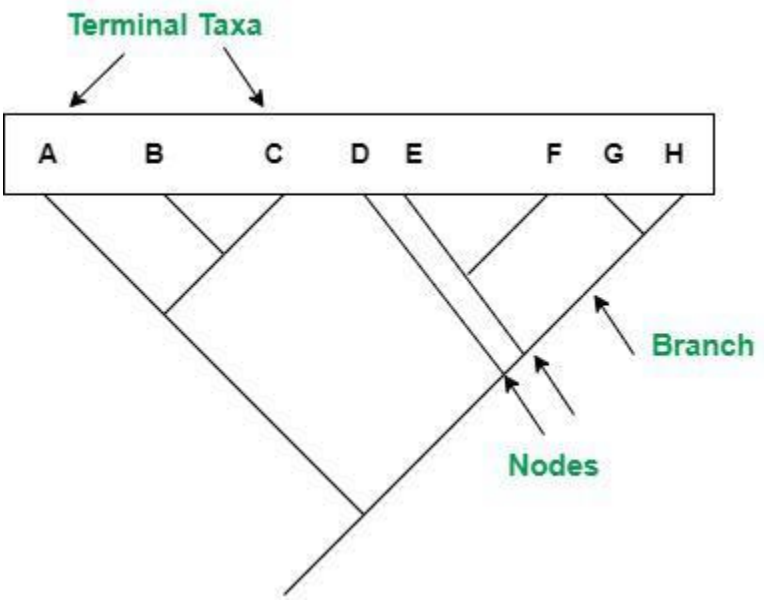


Fig 2: Parts of a phylogenetic tree

3.3.9: Approaches to make Phylogenetic Tree

The phylogenetic tree is created using one of two different approaches:

a. **Character-based Approach**: This method is also known as the discrete method because it is based solely on the sequence characters. Aligned characters are used in the character-based technique to build the phylogenetic tree. During the tree inference, these aligned characters either include DNA or protein sequences. Character-based methods involve analyzing sequence data by directly examining the sequence characters, rather than relying on pairwise distance comparisons. These methods evaluate all sequences at once by analyzing one character or site at a time. Character-based methods are generally considered more accurate than distance-based methods. However, character-based methods are more computationally intensive and require more sophisticated statistical models. maximum parsimony (MP) and maximum likelihood (ML) are the two most prevalent.

   i. Maximum parsimony (MP): Maximum parsimony method is a character-based method that selects the tree with the least number of evolutionary changes or the shortest total branch length. Initially, multiple sequence alignment is performed to identify potential positions in the sequences that correspond to each other. Each aligned position is analyzed to identify the trees that require the smallest number of evolutionary changes to produce the observed sequence changes. This process is repeated for all positions in the sequence alignment, and the trees that produce the lowest overall number of changes for all positions are selected. This method works best for relatively similar sequences and for small numbers of sequences.

   ii. **Maximum likelihood (ML)**: Maximum likelihood is a statistical method that uses probabilistic models to identify the most appropriate tree that has the maximum probability of generating the observed data. Similar, to the maximum parsimony method, this approach evaluates each column of a multiple sequence alignment during the analysis. However, unlike maximum parsimony, ML considers all possible trees that could explain the observed data. The likelihood of each possible

tree is calculated, and the tree with the highest probability is selected as the most likely evolutionary history of the sequences.

b. Distance-based Approach: This approach is based on how dissimilar or how far apart the two aligned sequences are from one another. The pairwise distances from the sequence data are then utilized to create a matrix, which is subsequently used to generate the phylogenetic tree in this method.

3.3.10: Steps for Phylogenetic Analysis

Any phylogenetic study starts with the following fundamental steps:

a. **Setup and alignment of a dataset**: Finding an interesting protein or DNA sequence is the initial stage, followed by compiling a dataset of related sequences. Using NCBI BLAST or other comparable search engines, DNA sequences of interest can be located. Multiple sequence alignment is produced after the selection and recovery of sequences. To find homology regions, a set of sequences must be arranged in a matrix. ClustalW, MSA, MAFFT, and T-Coffee are just a few of the websites and software tools available for doing multiple sequencing on a given set of molecular data.

b. **Create (estimate) phylogenetic trees**: From sequences using stochastic models and computational techniques. Statistical techniques are used to ascertain the tree topology and calculate the branch lengths that most accurately depict the phylogenetic relationships of the matched sequences in a dataset in order to construct phylogenetic trees. The most often used computational techniques are those that use distance matrices and discrete data, including maximum likelihood and parsimony. Many software programs, including Paup, PAML, and PHYLIP, use these most common techniques.

c. **Test and evaluate the estimated trees statistically:** One or more ideal trees are produced via tree estimation techniques. A number of statistical tests are run on this set of potential trees to see which is the best option and whether the suggested phylogeny makes sense. Jackknife Resampling techniques, as well as analytical techniques like parsimony, distance, and likelihood, are frequently used to evaluate trees.

3.3.11: Types of Phylogenetic Tree

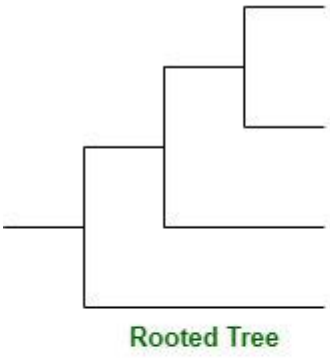
Distinct phylogenetic trees are divided into varied groups based o**n the basis of the presence or absence of a common root**

a. **Rooted tree:** A phylogenetic tree with a common ancestor on each node is referred to as a rooted tree. As a result, the categorization comes to a stop at one point, typically at the node that serves as the common ancestor of all the tree branches.

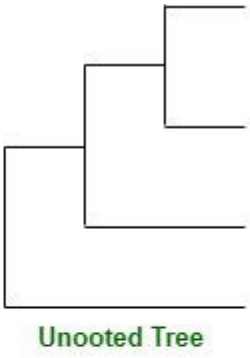
b. **Unrooted tree:** The non-rooted tree does not share a common ancestor with the rooted tree. The common ancestor or the tree node is always left out while creating the unrooted phylogenetic tree from the rooted tree.

c. **Bifurcating tree:** Phylogenetic trees that only have two branches or leaves are referred to as bifurcating trees. Additionally, it can be divided into rooted and unrooted bifurcating trees.

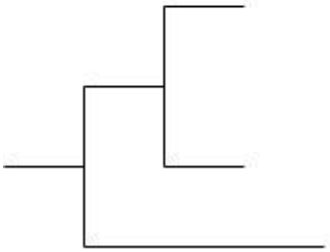
d. **Multifurcating tree:** Multiple branches can be found on a single node in a multifurcating tree, as the name suggests. Both a rooted multifurcating tree and an unrooted multifurcating tree are categories for it once more.



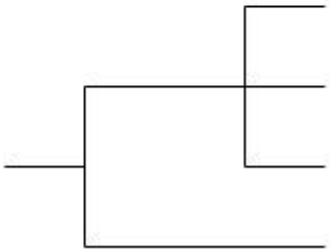
Rooted Tree



Unooted Tree



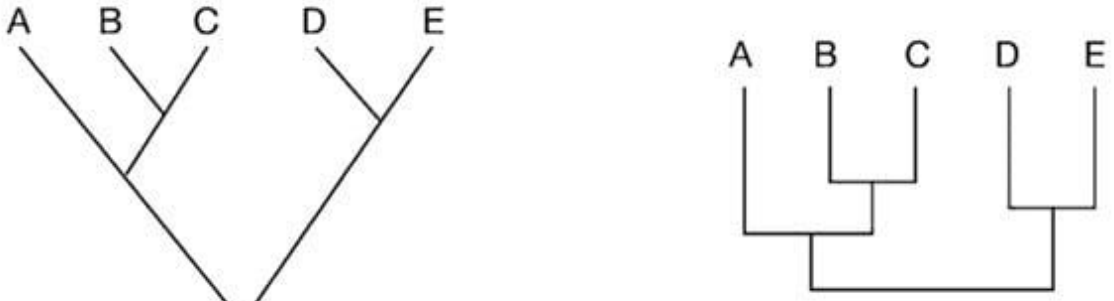
Bifurcating Tree



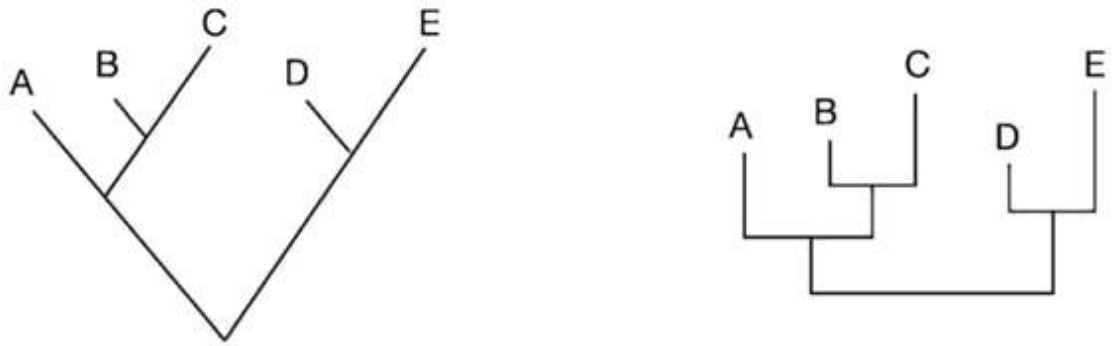
Multifurcating Tree

On the basis of topology

a. **Dendrogram-**A phylogenetic tree's diagrammatic representation is also known as a dendrogram because a dendrogram is a broad term for any tree, phylogenetic or not.
b. **Cladogram-**A cladogram solely depicts a branching pattern; as a result, its interior nodes do not represent ancestors and its branch lengths do not correspond to time or the relative degree of character change.
c. **Chronogram-**A Chronogram is a particular kind of Phylogenetic tree that uses the length of its branches to represent time.
d. **Phylogram-**A phylogenetic tree with branch lengths according to character change is called a phylogram. A phylogenetic tree called a chronogram explicitly displays time by the lengths of its branches.
e. **Dahlgrenogram-**A Dahlgrenogram is a diagram that shows a phylogenetic tree in cross-section.



Cladogram



Phylogram

3.3.12: Significance of Phylogenetic tree

3.3.13:

a. To illustrate the relationships between organisms thought to share some evolutionary origin.
b. Researching the shared ancestors of extinct and surviving species.
c. Employed to research the evolutionary past.
d. Employed in the hunt for new species.
e. The evolutionary histories of pathogenic bacteria can be tracked with the use of the phylogenetic tree.
f. Research the global dispersal of the species
g. It is used to determine the most recent shared ancestors and how closely related different species are to one another.
h. To connect the important turning points in the development of life to the tree of life.

3.3.13: Applications of the phylogenetic tree

a. Phylogenetic trees have various practical applications, including:
b. Phylogenetic trees can be used to study the evolutionary relationships between different species and to understand the evolutionary processes over time.
c. Phylogenetic trees can be used to study the diversity and distribution of species and to develop conservation strategies to protect endangered species and ecosystems.
d. Phylogenetic trees can be used to identify the origins of pathogens and to track the spread of diseases.
e. Phylogenetic trees can also be used in forensics to identify the origins of biological samples found at crime scenes and to link suspects to crimes.
f. Phylogenetic trees are useful for organizing and classifying organisms and species according to their DNA sequences and morphological similarities and differences.

3.3.14: **Limitations** of Phylogenetic tree

a. This evolutionary tree of craniates, which resembles a progressing ladder, evolved from an organism without a spinal column.
b. Therefore, depending only on the traits they share, various groupings of organisms, objects, or units are situated at the tips of each branch.
c. A phylogenetic tree illustrates the theories regarding the evolution and development of life.
d. They are only as accurate as the facts that they are based on and are supported by.
e. The information is derived from research and studies, which may contain some bias.
f. As a result, phylogenetic trees constructed using data from research and studies may always be erroneous, biased, or subject to manipulation.

3.3.15: Phenetic and Cladistic Methods in Phylogenetic Analysis

The two common phylogenetic analytical methods are phonetic or clustering approach and cladistic approach.

a. Phonetic (clustering) approach are capable of producing a tree even in the absence of evolutionary relationships. A simple clustering procedure e.g. UPGMA or unweighted Pair Group Method with Arithmetic Mean use pairwise dissimilarities. A modification of the UPGMA method called Neighbour-Joining is designed to correct the unequal rates of evolution in different branches of the tree.
b. Cladistic methods on the other hand deal with evolutionary patterns implied by the possible trees relating to a set of taxa. The most popular cladistic methods in molecular phylogeny are the maximum parsimony and maximum likelihood approaches. They are used for sequence data; they cannot be used for anatomical characters like height.

3.3.16: **Cladogram**

A phylogeny, or hypothetical link between groups of creatures, is depicted in a diagram known as a cladogram. A phylogenetic systematics researcher will use a cladogram to depict the groups of organisms being compared, their relationships, and their most recent shared ancestors. A cladogram might be quite complicated and compare every known form of life, or it can be extremely simple and compare just two or three groups of creatures.

3.3.17: Difference between a Cladogram and a Phylogenetic Tree

a. Phylogenetic trees and cladograms are frequently used interchangeably; however, they differ in some ways.
b. Cladograms demonstrate the evolutionary relationships between various organisms without demonstrating the course of evolution.
c. Through the evolutionary changes that have taken place between an ancestor and its descendant species or set of species, phylogenetic trees demonstrate how various organisms are connected.

3.3.18: Data Mining

Data Mining Data mining refers to extracting or "mining" knowledge from large amounts of data. Data Mining (DM) is the science of finding new interesting patterns and relationship in huge amount of data. It is defined as "the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses". Data mining is also sometimes called Knowledge Discovery in Databases (KDD). Data mining is not specific to any industry. It requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Data mining tasks The two "high-level" primary goals of data mining, in practice, are prediction and description. The main tasks wellsuited for data mining, all of which involves mining meaningful new patterns from the data, are:

a. Classification: Classification is learning a function that maps (classifies) a data item into one of several predefined classes.
b. Estimation: Given some input data, coming up with a value for some unknown continuous variable.
c. Prediction: Same as classification & estimation except that the records are classified according to some future behaviour or estimated future value).
d. Association rules: Determining which things go together, also called dependency modeling.
e. Clustering: Segmenting a population into a number of subgroups or clusters.
f. Description & visualization: Representing the data using visualization techniques.

Learning from data falls into two categories:

a. directed ("supervised") learning: The first three tasks – classification, estimation and prediction – are examples of supervised learning.
b. undirected ("unsupervised") learning. The next three tasks – association rules, clustering and description & visualization – are examples of unsupervised learning. In unsupervised learning, no variable is singled out as the target; the goal is to establish some relationship among all the variables. Unsupervised learning attempts to find patterns without the use of a particular target field. The development of new data mining and knowledge discovery tools is a subject of active research. One motivation behind the development of these tools is their potential application in modern biology.

3.3.19: Application of Data Mining in Bioinformatics

a. Gene finding,

b. protein function domain detection,
c. function motif detection,
d. protein function inference,
e. disease diagnosis,
f. disease prognosis,
g. disease treatment optimization,
h. protein and gene interaction network reconstruction,
i. data cleansing, and
j. protein sub-cellular location prediction.

Define a phylogenetic tree? Write the difference between a cladogram and a phylogenetic tree

Self-Assessment Exercises

1. What is a cladogram?
2. Write the importance of a phylogenetic tree

3.4: Summary

Database search are computer-based procedures which can be of two categories; genomic analysis and proteomic analysis. Once the database search is complete, the next course of action is data mining, analysis, and modeling procedures. The processes involve primary sequence alignment, secondary and tertiary structure prediction, homology modelling. Data mining is a follow-up to database search. Data mining gives biological meaning to a search.

3.5: References/Further Reading/Web Sources

- Oehmen, C. S & Baxter, D. J. (2013). *"ScalaBLAST 2.0: Rapid and robust BLAST calculations on multiprocssor_systems*". Bioinformatics. 29 (6): Science Watch. July–August 2000.

- Lloyd, A. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43). Briefings in Bioinformatics. 2. 10.1093/bib/2.4.407.

- Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806087

https://www.bing.com/ck/a?!&&p=f75fdf42244b01f2JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTI1NA&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Phylogenetic+analysis+in+Bioinformatics&u=a1aHR0cHM6Ly9iaXAud2Vpem1hbm4uYWMuaWwvZWR1Y2F0aW9uL2NvdXJzZS9pbnRyb2Jpb2luZm8vMDMvbGVjdDEyL3BoeWxvZ2VuZXRpY3MucGRm&ntb=1

https://www.bing.com/ck/a?!&&p=34dc0862beb1f907JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTIyOA&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=%2bData+mining+in+novel+genomes+in+Bionformatics&u=a1aHR0cHM6Ly9hcnhpdi5vcmcvcGRmLzEyMDUuMTEyNQ&ntb=1

https://www.bing.com/ck/a?!&&p=f6fb055c02cb7d6cJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTUwMw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&u=a1L3ZpZGVvcy9zZWFyY2g_cT1QaHlsb2dlbmV0aWMrYW5hbHlzaXMraW4rQmlvaW5mb3JtYXRpY3MmZG9jaWQ9NjAzNDg0MjU3MDYzODAwNjA1Jm1pZD0zMUZDN0QwQ0ZBMzVDQ0I0RjM5NTMxRkM3RDBDRkEzNUNDQjRGMzk1JnZpZXc9ZGV0YWlsJkZPUk09VklSRQ&ntb=1

https://www.youtube.com/watch?v=R9HNl2Pivx8

3.6: Possible Answers to Self-Assessment Exercises

1.

A cladogram solely depicts a branching pattern; as a result, its interior nodes do not represent ancestors and its branch lengths do not correspond to time or the relative degree of character change.

2.

- To illustrate the relationships between organisms thought to share some evolutionary origin.
- Researching the shared ancestors of extinct and surviving species.
- Employed to research the evolutionary past.
- Employed in the hunt for new species.
- The evolutionary histories of pathogenic bacteria can be tracked with the use of the phylogenetic tree.

Unit 4: Use of Bioinformatics tools in Biotechnology/Biopharma

Unit Structure

4.1: Introduction
4.2: Intended Learning Outcomes
4.3: Main Body
    4.3.1: Implementation of Bioinformatics in Biotechnology Research
    4.3.2: Bioinformatic Tools, Software and Database in Biotechnology
    4.3.3: Applications of bioinformatics tools in biotechnology
    4.3.4: Bioinformatics in Biopharmaceuticals
    4.3.5: List of Biopharmaceutical Companies that Uses Bioinformatics
4.4: Summary
4.5: References/Further Readings/Web Sources
4.6: Possible Answers to Self-Assessment Exercises

4.1 Introduction

Bioinformatics has provided computational ways for data analysis by employing informatics tools and software to determine protein/gene structure or sequence, homology, molecular modelling of biological system, molecular docking etc. to analyze and interpret data. Currently bioinformatics has become a principal technology in all life sciences research. Bioinformatics has been integrated into a number of different disciplines where it assists in better understanding of the data in a shorter time frame. With the massive advancement in biotechnology, bioinformatics is growing rapidly providing new ways and approaches for the assessment of valuable data.

4.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. Explain the implementation of bioinformatics in biotechnology research
b. List and explain five (5) bioinformatic tools, software and database in Biotechnology
c. List and explain three (3) biopharma companies using bioinformatics and their products.

4.3: Main Body

4.3.1: Implementation of Bioinformatics in Biotechnology Research

There is an unprecedented in infectious and transmitted disease problem in humans globally. The major diseases such as tuberculosis, dengue, malaria, influenza flu, cholera, hepatitis etc. in human are responsible for large number of losses among us. Therefore, an increasing the gene/genomes and proteomics data of microorganisms can help to better understanding and controlling of pathogenicity for suitable treatment.

Bioinformatics is a promising area which can accelerate wet laboratory research while avoid needless laboratory practices. That can also reduce the chemicals, enzymes and drugs, during experiment. It can help for in silico designing and in vitro validation of specific primers and probes for monitoring of pathogens.

Bioinformatics can also be used for analysis of evolutionary relationship of organisms using phylogenetic analysis. It can help to provide clue for existence of gene or protein present in other organisms. It can help to generate the RNA secondary structure for analysis of evolutionary stability. For the structural bioinformatics, we use unknown protein for modeling of 3-dimensional protein models for screening of drugs and antibiotics for better management of pathogen. That can help to superfluous use of drugs and antibiotics for treatment of disease.

Another approach like immunoinformatics is accelerating the development of antigen based diagnostic kits and vaccines. There are recent approaches in the field of metabolic engineering. That requires sometimes to replace the wild type promoter, transcription factor and coding gene by strong promoter or codon optimized synthetic gene for overexpression of metabolite, antibiotics, enzymes, drugs and fine chemicals.

Similarly, bioinformatics is also useful in systems and synthetic biology analysis of sequences of promoter, transcription factor and coding gene and their location. Promoter is considered as an engine of cell that play key role for gene expression thus, Bioinformatics can help for analysis of promoter and designing of new promoter library for gene regulation. In contrast, plasmid is also playing a key role in heterologous gene expression; thus, bioinformatics can help to in silico design and analyze the library of cloning and expression vectors. Before going to wet lab experiment, bioinformatics can be used for analysis of restriction enzymes sites, open reading frame and gene size. Genomics approaches use for analysis of number of genes in an organism based on the number of nucleotide base pairs. It uses a gene library that can be constructed or inserted into other organisms for improvement and silencing. In the areas of structural genomics, functional genomics and nutritional genomics, bioinformatics plays a vital role for better understanding of genomics complexity. In contrast, proteomics involves the sequences of amino acids in a protein that can be used for determination of three-dimensional structure and relating it to the function of the unknown protein.

4.3.2: Bioinformatic Tools, Software and Database in Biotechnology

Identification of Sequence Homology.

The Basic Local Alignment Search Tool (BLAST: http:// www.ncbi.nim.nih.gov/blast) is used for searching homology of the sequences either nucleotide or protein from existing databank. It compares query nucleotide or protein sequences to existing sequence in databases and calculates the statistical significance of matches. That can be useful to infer the functional and evolutionary relationships between sequences and also help to identify member of gene families

Sequences Alignment

A sequence alignment is a way of arranging the sequences of DNA or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues thus; the identical or similar characters are aligned in successive columns.

There are mainly two methods used in sequence alignment such as pairwise and multiple. In the pairwise sequences alignment is used to find the best-matching piecewise (local) or global alignments of two query sequences. It can be only used between the two sequences at a time. Whereas, multiple sequence alignment (MSA) can also use to align pairs of sequences. It is an extension of pair-wise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query. Multiple alignments are used for identification of conserved sequence regions across a group of sequences. Alignments are also used to help in the establishing evolutionary relationships by constructing phylogenetic tree. There

are a number of tools such as ClustalW, CodonCode Aligner, DNA Aligner, ClustalX and T-coffe used for sequences alignment.

Prediction of RNA Secondary Structure

There are mainly two types of RNA such as messenger RNA (mRNA) and non-coding RNA. The mRNA carries information from DNA to the ribosome, which provides a site for protein synthesis. While non-coding RNAs are transfer RNA (tRNA) and ribosomal RNA (rRNA), both are involved in the process of translation. Non-coding RNAs are also involved in gene regulation and processing. Thus, these RNA is a single stranded and forms a loop. That can have a diverse form of secondary structure and compared to base sequences and structural conservation. Base pairing is stabilized the structure whereas unpaired regions (loops) destabilized the secondary structure. A number of computational biology tools such as Mfold, RNAfold, Nupack and R-coffee are used for prediction of RNA secondary structure.

Construction of the Phylogenetic Tree

The nucleotide or protein sequences of any organisms are used to search the homology using BLAST and other homologous sequences that can be retrieved from NCBI-GenBank. All these sequences from various organisms are aligned using CLUSTALX. A numbers of tools such as MetaPIGA2, PAUP, PHYLIP, QuickTree, MEGA, SpllisTree and TreeAlign are available for construction of phylogenetic tree. That can be useful in the analysis of genetic relationship in any organisms. MEGA is a very popular tool for construction of phylogenetic tree. The aligned sequences are manually checked and verified subsequently using passion correction algorithm. While total 100-1000 bootstraps values are sampled to determine the measure support for each node on consensus tree.

Gene Designing and Codon Optimization

The DNA sequences are arranged into triplets (codons). It is replaced with new ones and generated with a given frequency distribution. In this process amino acid is same, but codon of low frequency of an amino acid is replaced with codon of high frequency. Gene designer (https:// www.dna20.com/ index.php?pageID=220) is used for designing and simulation of gene in a given expression hosts. Whereas, optimizer software is used for codon optimization and calculation of Codon adaptation index (CAI), G+C and A+T, CAIcal and Mr-Gene are also used for optimization of DNA sequences at maximum suitable threshold level. Codon optimization of gene is performed at 10-15% threshold level of host cellular codons. CAI is also calculated for each gene which is acceptable and effective measure of potential gene expression level. However, codon optimization is a technique to exploit the protein expression in living organism by increasing the translational efficiency of gene of interest. This gene is transformed into one species to another such as plant to human sequence, human sequence to bacteria or yeast.

Cell Designer

A living system can be viewed as a biochemical reaction network. A system biology approach is used to understand the living systems. Cell Designer is a structured diagram editor for drawing gene-regulatory and biochemical networks. Gene networks are drawn based on the process diagram, with graphical notation system proposed by Kitano, and are stored using the Systems Biology Markup Language (SBML). This is a standard for representing models of biochemical and gene regulatory networks. Networks are able to link with simulation and other analysis packages through Systems Biology Workbench (SBW).

RBS Calculator

A recently developed tools RBS calculator is available online for designing and calculation of RBS efficiency. But there is need to massively over-express a protein. Therefore, designing of

new RBS sequence are required for maximizing the translation initiation rate. The excess protein expression may form inclusion bodies or kill the cells.

APE- a Plasmid Editor APE- a plasmid editor software is used for designing of new plasmid or redesigning of existing plasmid. It contains the following features such as highlights restriction sites in the editing window, accurately reflects Dam/Dcm blocking of enzyme sites, highlights text using pre-defined and custom feature libraries, highlights text using user defined features, shows translation, Tm, %GC, ORF of selected DNA in real-time, reads DNA Strider, Fasta, it also have some more feature such as GenBank and EMBL files, saves files as DNA Strider-compatible or Genbank file format, highlights and draws graphic maps using feature annotations from genbank and embl files.

RNA Designer RNA designer takes as input a secondary structure description and outputs an RNA strand that is predicted to fold the RNA secondary structure. It is used to design the RNA molecules with certain structural properties, as part of the development of molecules with novel functional properties in order to understand the secondary structure. These elements are critical to specific function of cellular RNAs. RNA designer uses a stochastic local search algorithm, which decomposes the input structure in a hierarchical fashion, finds strands that fold to the resulting substructures. RNA designer uses the fold software from Vienna Package as part of its implementation.

Prediction of Epitope for Vaccine Designing Epitopes are also known as antigenic determinant which is a part of protein. It is recognized by immune system, specifically by antibodies. The part of an antibody that recognizes the epitope is called a paratope. Although epitopes are usually non-self-proteins, sequences derived from the host that can be recognized. The T-cell epitopes are presented on the surface of an antigen-presenting cell, where they are bound to MHC molecules. These T-cell epitopes (MHC class I) are small peptides between 8-11 amino acids in length, whereas MHC class II molecules present longer peptides, 13-17 amino acids. A number of tools such as HLArestrictor, ePitope, NetCTL, NetCTLpan, BIMAS, SYFPEITHI, MHCServer, Propred and Propred1 are used for prediction of epitopes from primary protein sequences. Propred and Propred1 tools cover maximum number of human leukocyte antigen (HLA) in comparison to other immunoinformatics tools. Thus; T-cell epitopes are considered the parameters during epitopes prediction such as 4% threshold with maximum binding score to HLA molecules. Whereas, IEDB is a collection of tools for prediction and analysis of epitope.

Structural Bioinformatics Approach - Prediction of protein structure

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein can be easily determined from gene sequence. Therefore, homology modeling is also known as comparative modeling of protein which refers to construct an atomic-resolution model of the target protein from its amino acid sequence. It uses experimentally generated 3-D structure of a related homologous protein. Homology modeling is used when the query protein sequences showed 25% homology with known crystal structure. A number of tools and softwares such as EsyPred3D, SWISS-MODEL, RaptorX, Geno3D and Modeller are available for generating 3-D model of protein. Modeller9v7 is more popular and frequently used in generating of the 3-D model and it is visualized by PYMOL. The evaluation of generated 3-D model is performed on the basis of free energy of model and template. Whereas, PROCHECK is used for validation of the 3-D structure. It generates Ramachandran plot and the amino acid residues in allowed, disallowed region and overall G-factor are considered.

Molecular Interaction

There is efficient software available for studying interactions among proteins, ligands and peptides. Whereas, types of interactions most often encountered in the field include–Protein–ligand (including drug), protein–protein and protein–peptide. Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational algorithms termed docking algorithms for studying molecular interactions

4.3.3: Applications of bioinformatics tools in biotechnology

GENOMICS

This field generates a vast amount of data from gene sequences, their interrelation and functions. To manage an escalating amount of genomic information, bioinformatics tools like Carrie, Clover, Cister, Possum etc are required to maintain and analyze the DNA sequences from different organism. Determination of sequence homology, gene finding, coding region identification, structural and functional analyses of genomic sequences etc.

COMPARATIVE GENOMICS

Bioinformatics plays an important role in comparative genomics by determine the genomic structural and functional relationship between different biological species. Tools like BLAST, HMMER, Clustal Omega, Sequerons all assist in DNA or protein sequence alignment, sequence profiling, multiple sequence alignment etc.

PROTEOMICS:

Advanced molecular based techniques led to the accumulation of huge proteomic data of protein activity patterns, interactions, profiling, composition, structural information, image analysis, peptide mass fingerprinting, peptide fragmentation fingerprinting etc. This enormous data could be managed by using different tools of bioinformatics such as SMM, ZDock, K2/FAST.

DRUG DISCOVERY

Bioinformatics is playing an increasingly important role in nearly all aspects of drug discovery, drug assessment and drug development. This growing importance is not because bioinformatics handles large volumes of data but also in the utility of bioinformatics tools to predict, analyze and help interprete clinical and preclinical findings.

EVOLUTIONARY STUDIES/PHYLOGENETICS

The study of evolutionary relationship among individuals or group of organisms is defined as phylogenetics. Taxonomists find the evolutionary relationship using various anatomical methods that takes too much time. Using Bioinformatics, phylogenetic trees are constructed based on the sequence alignment using various methods. Various algorithmic methods are developed for the construction of phylogenetic tree that are used depending on the various evolutionary lineages

CHEMINFORMATICS

Cheminformatics (chemical informatics) focuses on storing, indexing, searching, retrieving, and applying information about chemical compounds. It involves organization of chemical data in a logical form to facilitate the retrieval of chemical properties, structures and their relationships. Using bioinformatics, it is possible through computer algorithm to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically. Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.

PREDICTING PROTEIN STRUCTURE AND FUNCTIONS

Protein topology prediction is now so much easy thanks to bioinformatics which helps in the prediction of 3D structure of a protein to gain an insight into its function as well. E.g. PHD, Raptor X, Modeller etc.

MEDICINE

Doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning by employing bioinformatics tool.

MICROBIAL GENOME APPLICATIONS

Microbes have been studied at very basic level with the help of bioinformatics tools required to analyse their unique set of genes that enables them to survive under unfavourable conditions.

4.3.4: Bioinformatics in Biopharmaceuticals

A biopharmaceutical also known as a biological medical product, or is any pharmaceutical drug product manufactured in, extracted from, or semi synthesized from biological sources. Different from totally synthesized pharmaceuticals, they include vaccines, blood, blood components, allergenics, somatic cells, gene therapies, tissues, recombinant therapeutic protein, and living cells used in cell therapy. Biologics can be composed of sugars, proteins, or nucleic acids or complex combinations of these substances, or may be living cells or tissues. They (or their precursors or components) are isolated from living sources—human, animal, plant, fungal, or microbial.

4.3.5: List of Biopharmaceutical Companies that Uses Bioinformatics

1. **Anavex Life Sciences Corp:** is a pharmaceutical company that develops drug candidates. ANAVEX 2-73 has been shown to provide protection from oxidative stress, which damages and destroys neurons and is believed to be a primary cause of Alzheimer's disease.
2. **Bayhill Therapeutics Biopharmacy:** Bayhill Therapeutics is focused on the translation of research into therapeutics for the treatment of autoimmune diseases.
3. **Biocon Limited:** is an Indian biopharmaceutical company based in Bangalore, India. The Company manufactures generic active pharmaceutical ingredients that are sold in the developed markets of the United States and Europe. It also manufactures biosimilar Insulins, which are sold in India as branded formulations and in both bulk and formulation forms. In research services, Syngene International Limited is engaged in the business of custom research in drug discovery while the other fully owned subsidiary Clinigene International Limited is in the clinical development space.
4. **Halozyme Therapeutics:** Halozyme Therapeutics, based in San Diego, California, is a biopharmaceutical company developing and commercializing products targeting the extracellular matrix for the endocrinology, oncology, dermatology and drug delivery markets. What is the use of BLAST programme in identification of sequence homology?

Self-Assessment Exercises

1. Explain the application of genomics as bioinformatic tool in biotechnology

4.4: Summary

Bioinformatics holds significant importance in countless disciplines of biotechnology such as comparative genomics, drug designing, proteomics, molecular modelling, microbial genomics etc. And proper handling of this tools is very important to avoid misinterpretation of results.

4.5: References/Further Reading/Web Sources

- Oehmen, C. S & Baxter, D. J. (2013). *"ScalaBLAST 2.0: Rapid and robust BLAST calculations on multiprocssor_systems*". Bioinformatics. 29 (6): Science Watch. July–August 2000.

- Lloyd, A. (2001). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (Methods of Biochemical Analysis, 43). Briefings in Bioinformatics. 2. 10.1093/bib/2.4.407.

- Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806087

https://www.bing.com/ck/a?!&&p=4fe63c3582415db4JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTMxNA&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Use+of+Bioinformatics+tools+in+Biotechnology%2c+pdf&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNTExODIzNDVfQmlvaW5mb3JtYXRpY3MtX1Rvb2xzX2FuZF9BcHBsaWNhdGlvbnM&ntb=1

https://www.bing.com/ck/a?!&&p=0b236dbc64ee2abcJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTQwNw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Use+of+Bioinformatics+tools+in+Biopharma&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNTExODIzNDVfQmlvaW5mb3JtYXRpY3MtX1Rvb2xzX2FuZF9BcHBsaWNhdGlvbnM&ntb=1

https://www.bing.com/videos/search?q=Use+of+Bioinformatics+tools+in+Biotechnology%2c+pdf&&view=detail&mid=27ADC78891EB4A91F91227ADC78891EB4A91F912&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DUse%2Bof%2BBioinformatics%2Btools%2Bin%2BBiotechnology%252c%2Bpdf%26FORM%3DHDRSC6

https://www.bing.com/ck/a?!&&p=f79d8dace351bc36JmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTU4Mw&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&u=a1L3ZpZGVvcy9zZWFyY2g_cT1Vc2Urb2YrQmlvaW5mb3JtYXRpY3MrdG9vbHMraW4rQmlvcGhhcm1hJmRvY2lkPTYwMzUzNDAwNTY3MTQzNTM1NCZtaWQ9RDE0OEFDNTE5OUM4MDM1Qzk1MjVEMTQ4QUM1MTk5QzgwMzVDOTUyNSZ2aWV3PWRldGFpbCZGT1JNPVZJUkU&ntb=1

4.6: Possible Answers to Self-Assessment Exercises

- This field generates a vast amount of data from gene sequences, their interrelation and functions.
- To manage an escalating amount of genomic information, bioinformatics tools like Carrie, Clover, Cister, Possum etc are required to maintain and analyze the DNA sequences from different organism.
- Determination of sequence homology, gene finding, coding region identification, structural and functional analyses of genomic sequences etc.

Unit 5: Current topics in bioinformatics and use of perl to facilitate biological analysis.

Unit Structure

5.1: Introduction
5.2: Intended Learning Outcomes
5.3: Main Body
    5.3.1: Current topics in bioinformatics
    5.3.2: Bioperl
5.4: Summary
5.5: References/Further Readings/Web Sources
5.6: Possible Answers to Self-Assessment Exercises

5.1 Introduction

Bioinformatics has been one of the major focuses due to the rapid development and requirement of using bioinformatics approaches in biological data analysis, especially for omics large datasets.

5.2 Intended Learning Outcomes (ILOs)

At the end of this section, students should be able to;

a. identify the current areas of research in bioinformatics.
b. Explain the procedure of using Bioperl

5.3 Main Body

5.3.1: Current topics in bioinformatics

Researchers working in the scientific area always want to explore new and hot topics to make informed choices. "C*ancer, coronary artery disease, HIV, chronic infections*", and so on. *In silico drug designing* is always demanding in designing inhibitors or potential drugs for such diseases. Besides, a lot of scientists are working on *next-generation sequencing, big data,* and *cancer*. Cancer studies using bioinformatics is one of the leading current topics in bioinformatics. Cancer causes morbidity and mortality worldwide. There exists an urgent need to identify new biomarkers or signatures for early detection and prognosis.

The following topics are considered demanding in bioinformatics.

a. Cloud computing, big data, Hadoop
b. Machine learning
c. Artificial intelligence
d. Functional genomics
e. RNA-seq analysis (equally relevant along with next-generation sequencing techniques)
f. Data mining (including text search, data integration, database development, and management)
g. Neural networks
h. Mathematical modeling
i. Mirna function identification
j. Evolutionary studies
k. Genomics, transcriptomics, and proteomics
l. Metabolomics

5.3.2: Bioperl

BioPerl

Bioperl is a collection of perl modules that facilitate the development of perl scripts for bio-informatics applications. It provides reusable perl modules that facilitate writing perl scripts for sequence manipulation, accessing of databases using a range of data formats and execution and parsing of the results of various molecular biology programs including Blast, clustalw, TCoffee, Genscan, ESTscan and HMMER. Bioperl enables developing scripts that can analyse large quantities of sequence data in ways that are typically difficult or impossible with web based systems.

BioPerl is an open-source project that develops modules for biological data in Perl. A Perl module is a reusable package defined in a library file. BioPerl modules are stable and "easy" to use.

Modules include objects for sequence files, alignment files and database searching. These objects can interact: the objects provide a coordinated and extensible framework for computational biology

BioPerl module names minimize ‘namespace’ collisions by separating parts of a name by a double colon (::). For example: The module ‘Bio::DB::GenBank’; instructs Perl to go to the Database GenBank This module can automate retrieval of a set of sequences The ‘Bio::SearchIO’ module is used for parsing an input file and creating an output file with the specified information.

This module can be used to create tables that summarize results from BLAST searches Bioperl includes a wide array of utilities for biological and bioinformatics analysis. Utilities including databases used for bioinformatics information, analysis routines for genomics, proteomics, and evolutionary studies. It is capable of analysing the results from bioinformatics programs such as BLAST, Tcoffee, GenScan and ClustalW.

Using Bioperl

Bioperl provides software modules for many of the typical tasks of bioinformatics programming. These include:

a. Accessing sequence data from local and remote databases
b. Transforming formats of database/ file records
c. Manipulating individual sequences
d. Searching for "similar" sequences
e. Creating and manipulating sequence alignments
f. Searching for genes and other structures on genomic DNA
g. Developing machine readable sequence annotation Installation: The actual installation of the various system components is accomplished in the standard manner:
h. Locate the package on the network
i. Download
j. Decompress (with gunzip or a simliar utility)
k. Remove the file archive (eg with tar -xvf)
l. Create a ‘‘makefile’’ (with ‘‘perl Makefile.PL’’ for perl modules or a supplied ‘‘install’’ or
m. ‘‘configure’’ program for non-perl program
n. Run ‘‘make’’, ‘‘make test’’ and ‘‘make install’’ This procedure must be repeated for every CPAN.

Define BioPerl?

Self-Assessment Exercises

1. Enumerate three (3) usefulness of BioPerl.

5.4: Summary

In this section the current trend in the use of Bioinformatics were discussed. Explanations were provided for BioPerl usage and application.

5.5: References/Further Reading/Web Sources

- Xiong, J. (2006). Essential Bioinformatics. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511806087

- Hahn A., Mohanty S.D & Manda P. (2017) What's Hot and What's Not? – Exploring Trends in Bioinformatics Literature Using Topic Modeling and Keyword Analysis. In: Cai Z., Daescu O., Li M. (eds) Bioinformatics Research and Applications. ISBRA 2017. Lecture Notes in Computer Science, vol 10330. Springer, Cham. https://doi.org/10.1007/978-3-319-59575-7_25

https://www.bing.com/ck/a?!&&p=60f06fd1ee14e1bcJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTI5MQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=Current+topics+in+bioinformatics+pdf&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNDczMjMyMDNfQ3VycmVudF90cmVuZF9hbmRfZGV2ZWxvcG1lbnRfaW5fYmlvaW5mb3JtYXRpY3NfcmVzZWFyY2g&ntb=1

https://www.bing.com/ck/a?!&&p=5adcb746a7f3d45cJmltdHM9MTY4Njk2MDAwMCZpZ3VpZD0zNWJkZWU5MC0xOTQ1LTYyYjQtMzAzNi1mY2M1MTg1ODYzNGUmaW5zaWQ9NTM5OQ&ptn=3&hsh=3&fclid=35bdee90-1945-62b4-3036-fcc51858634e&psq=use+of+perl+to+facilitate+biological+analysis%2c+pdf&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi81MDIxMjM2OV9CSU9QaHlsby1waHlsb2luZm9ybWF0aWNfYW5hbHlzaXNfdXNpbmdfcGVybA&ntb=1

https://www.bing.com/videos/search?q=Current+topics+in+bioinformatics+pdf&&view=detail&mid=6FAFB1C4BD538602A8846FAFB1C4BD538602A884&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3DCurrent%2Btopics%2Bin%2Bbioinformatics%2Bpdf%26FORM%3DHDRSC6

https://www.bing.com/videos/search?q=use+of+perl+to+facilitate+biological+analysis%2c+pdf&&view=detail&mid=81069801161FBD6EABBB81069801161FBD6EABBB&&FORM=VRDGAR&ru=%2Fvideos%2Fsearch%3Fq%3Duse%2Bof%2Bperl%2Bto%2Bfacilitate%2Bbiological%2Banalysis%252c%2Bpdf%26FORM%3DHDRSC6

5.6: Possible Answers to Self-Assessment Exercises

- It provides reusable perl modules that facilitate writing perl scripts for sequence manipulation
- It provides reusable perl modules that facilitate writing perl scripts for accessing of databases using a range of data formats and execution and parsing of the results of various molecular biology programs including Blast, clustalw, TCoffee, Genscan, ESTscan and HMMER.

- Bioperl enables developing scripts that can analyse large quantities of sequence data in ways that are typically difficult or impossible with web-based systems.
- BioPerl is an open-source project that develops modules for biological data in Perl.
- BioPerl modules are stable and "easy" to use. Modules include objects for sequence files, alignment files and database searching. These objects can interact: the objects provide a coordinated and extensible framework for computational biology

Glossary

Algorithm: a series of steps defining a procedure or formula for solving a problem that can be coded into a programming language and executed. Bioinformatics algorithms typically are used to process, store, analyze, visualize, and make predictions from biological data. Alignment: the result of a comparison of two or more gene or protein sequences in order to determine their degree of nitrogen base or amino acid similarity or dissimilarity. Sequence alignments are used to determine the similarity, homology, function, or other degree of relatedness between two or more genes or gene products

Bifurcation: a point in a phylogenetic tree in which an ancestral taxon splits into two independent lineages

Data mining: the ability to query very large databases in order to satisfy a hypothesis (" top - down " data mining); or to interrogate a database in order to generate new hypotheses based on rigorous statistical correlations (" bottom - up " data mining)

Gap (affine gap): any maximal, consecutive run of spaces in a single string of a given alignment

Global alignment: two nucleic acid or amino acid sequences lined up along their entire length

In silico (in biology): the use of computers to simulate, process, or analyze a biological experiment

Modeling: (in bioinformatics) refers to molecular modeling, a process whereby the three - dimensional architecture of biological molecules is interpreted (or predicted), visually represented, and manipulated in order to determine their molecular properties. (general) a series of mathematical equations or procedures that simulate a real - life process given a set of assumptions, boundary parameters, and initial conditions

Proteomics: the study of a proteome. Typically, the cataloging of all the express- ed proteins in a particular cell or tissue type, obtained by identifying the proteins from cell extracts using a combination of two - dimensional gel electrophoresis and mass spectrometry. Proteomics includes the large - scale analysis of the amassed protein composition and function.

Similarity (homology) search: given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides, or short sequence patterns. The transfer of structure and function information to a potentially homologous protein is straightforward when the

sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity ca

Structure prediction: algorithms that predict the secondary, tertiary, and sometimes even quaternary structure of proteins from their sequences. Determining protein structure from a sequence has been dubbed "the second half of the genetic code" since it is the higher - level folded structure of a protein that governs how it functions as a gene product. As yet, most structure prediction methods have been only partially successful and typically work best for certain well - defined classes of proteins

End of the module Questions

1.The BLAST program was developed in the year?
a) 1992
b) 1995
c) 1990
d) 1991

2. In sequence alignment by BLAST, each word from query sequence is typically _____ residues for protein sequences and---------- residues for DNA sequences.
a) ten, eleven
b) three, three
c) three, eleven
d) three, ten

3. Which of the following is not a variant of BLAST?
a) BLASTN
b) BLASTP
c) BLASTX
d) TBLASTNX

4. Which of the following is not correct about BLAST?
a) The BLAST web server has been designed in such away as to simplify the task of program selection
b) The programs are organized based on the type of query sequences
c) The programs are organized based on the type of nucleotide sequences, or nucleotide sequence to be translated
d) BLAST is not based on heuristic searching methods

5. Which of the following is not correct about FASTA?
a) Its stands for FAST ALL
b) It was in fact the first database similarity search tool developed, preceding the development of BLAST
c) FASTA uses a 'hashing' strategy to find matches for a short stretch of identical residues with a length of k
d) The string of residues is known as blocks

6. The positional difference for each word between the two sequences is obtained by _____ the position of the----------sequence from that of the--------sequence and is expressed as the offset.
a) subtracting, second, first
b) adding, second, first
c) adding, first, second
d) subtracting, first, second

7. When did Smith–Waterman first describe the algorithm for local alignment?
a) 1950
b) 1970
c) 1981
d) 1925

8. Which of the following does not describe local alignment?
a) A local alignment aligns a substring of the query sequence to a substring of the target sequence
b) A local alignment is defined by maximizing the alignment score, so that deleting a column from either end would reduce the score, and adding further columns at either end would also reduce the score
c) Local alignments have terminal gaps
d) The substrings to be examined may be all of one or both sequences; if all of both are included then the local alignment is also global

9. Which of the following does not describe local alignment algorithm?
a) Score can be negative
b) Negative score is set to 0
c) First row and first column are set to 0 in initialization step
d) In traceback step, beginning is with the highest score, it ends when 0 is encountered

10. Local alignments are more used when _____
a) There are totally similar and equal length sequences
b) Dissimilar sequences are suspected to contain regions of similarity
c) Similar sequence motif with larger sequence context
d) Partially similar, different length and conserved region containing sequences

11. Which of the following does not describe BLOSUM matrices?
a) It stands for BLOcks SUbstitution Matrix
b) It was developed by Henikoff and Henikoff
c) The year it was developed was 1992
d) These matrices are logarithmic identity values

12. Which of the following is untrue regarding the gap penalty used in dynamic programming?
a) Gap penalty is subtracted for each gap that has been introduced
b) Gap penalty is added for each gap that has been introduced
c) The gap score defines a penalty given to alignment when we have insertion or deletion
d) Gap open and gap extension has been introduced when there are continuous gaps (five or more)

13. Among the following which one is not the approach to the local alignment?
a) Smith-Waterman algorithm
b) K-tuple method
c) Words method
d) Needleman-Wunsch algorithm

14. Which of the following does not describe k-tuple methods?
a) k-tuple methods are best known for their implementation in the database search tools FASTA and the BLAST family
b) They are also known as words methods
c) They are basically heuristic methods to find local alignment
d) They are useful in small scale databases

15. Which of the following does not describe BLAST?
a) It stands for Basic Local Alignment Search Tool
b) It uses word matching like FASTA
c) It is one of the tools of the NCBI
d) Even if no words are similar, there is an alignment to be considered

16. Which of the following is untrue regarding BLAST and FASTA?
a) FASTA is faster than BLAST
b) FASTA is the most accurate
c) BLAST has limited choices of databases
d) FASTA is more sensitive for DNA-DNA comparisons

Answers

1. Answer: c
   Explanation: The BLAST program was developed by Stephen Altschul of NCBI in 1990 and hassince become one of the most popular programs for sequence analysis. BLAST uses heuristics to align a query sequence with all sequences in a database.
2. Answer: c
   Explanation: The first step is to create a list of words from the query sequence. Each word is typically three residues for protein sequences and eleven residues for DNA sequences. The list includes every possible word extracted from the query sequence. This step is also called seeding.
3. Answer: d
   Explanation: BLAST is a family of programs that includes BLASTN, BLASTP, BLASTX TBLASTN, and TBLASTX. BLASTN queries nucleotide sequences with a nucleotide sequence database. The alignment scoring is based on the BLOSUM62 matrix.
4. Answer: d
   Explanation: BLAST and FASTA are based on heuristic searching methods. In addition, programs for special purposes are grouped separately; for example, bl2seq, immunoglobulin BLAST, and VecScreen, a program for removing contaminating vector sequences.

5. Answer: d
   Explanation: The string of residues is known as ktuples or ktups, which are equivalent to words inBLAST, but are normally shorter than the words. Typically, a ktup is composed of two residues for protein sequences and six residues for DNA sequences.
6. Answer: d
   Explanation: The positional difference for each word between the two sequences is obtained by subtracting the position of the first sequence from that of the second sequence and is expressed as the offset. The ktups that have the same offset values are then linked to reveal a contiguous identical sequence region that corresponds to a stretch of diagonal in a two-dimensional matrix.
7. Answer: c
   Explanation: The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences.
8. Answer: c
   Explanation: Local alignments never have terminal gaps, because a higher score could be obtained by deleting the gaps (which always have negative scores, i.e. penalties). In case of global alignment there are terminal gaps while analyzing.
9. Answer: a
   Explanation: Score can be negative. When any element has a score lower than zero, it means that the sequences up to this position have no similarities; this element will then be set to zero to eliminate influence from previous alignment. In this way, calculation can continue to find alignment in any position afterward.
10. Answer: a
    Explanation: The given description is suitable for global alignment. It attempts to align maximum of the entire sequence unlike local alignment where the partially similar sequences are analyzed.
11. Answer: d
    Explanation: These matrices are actual percentage identity values. Or simply, they depend on similarity. Blosum 62 means there is 62 % similarity.
12. Answer: b
    Explanation: Dynamic programming algorithms use gap penalties to maximize the biological meaning. The open penalty is always applied at the start of the gap, and then the other gaps following it is given with a gap extension penalty which will be less compared to the open penalty. Typical values are –12 for gap opening, and –4 for gap extension.
13. Answer: d
    Explanation: Local alignment can be distinguished on two broad approaches, Smith-Waterman algorithm and word methods, also known as k-tuple methods and they are implemented in the well-known families of programs FASTA and BLAST.
14. Answer: d
    Explanation: k-tuple or word methods are especially useful in large-scale database searches where a large proportion of stored sequences will have essentially no significant match with the query sequence. They are heuristic methods that are not guaranteed to find an optimal alignment solution but are significantly more efficient than Smith-Waterman algorithm.

15. Answer: d
    Explanation: If no words are similar, there is no alignment i. e. it will not find matches for very short sequences. But it is considerably accurate as compared to other tools and hence is quite popular.
16. Answer: a
    Explanation: BLAST is faster than FASTA and most other tools. The speed and relatively good accuracy of BLAST is the key why the tool is the most popular bioinformatics search tool.