Module 2 - Computational biology

Syllabus

Introduction to bioinformatics and data generation tools (FASTA, BLAST), Data base (GENBANK, PubMed, PDB), Genome sequencing, Protein sequencing, Overview of bioinformatics applications, Biosensors: Types and applications

Bioinformatics

Bioinformatics is science of information and information flow in biological systems, particularly the use of computational methods in genetics and genomics. From an information technology perspective, it is defined as a scientific discipline encompassing acquisition, storage, processing, analysis, interpretation and visualization of biological information.

Database

Database stores biological information/data like sequences, structures, alignments, annotations. The purpose of biological data is to store and manage biological information in computer readable form.

Classification of Database

- Primary database : They store primary sequences that are obtained from different molecular biology techniques.
- Secondary database: It stores annotated sequences. Primary sequences that are obtained from different molecular techniques are annotated and stored in secondary database.
- Specialized database: They store data of specialized organisms and diseased data.
- Other database: It can be classified on the basis of data stored,

- RNA database
- Nucleotide sequence database
- protein databases
- genomes databases
- gene expression database.

GENBANK

GenBank is the genetic sequence database maintained by the National Institute of Health (NIH). It is comprised of an annotated collection of all publicly available nucleotide and protein sequences. Generally, the records in the GenBank database are comprised of contiguous stretches of DNA or RNA that have been annotated to add information to the raw sequence segments within that record.
Presently, the records in GenBank are based upon the direct submission of DNA sequences by the original authors along with some level of data curation by the GenBank staff facilitating the search and retrieval of the submitted records.
To find the gene coding sequence, look at the Genomic regions, transcripts, and products section or the NCBI Reference Sequences (RefSeq) section of the Gene record: Clicking on the GenBank link displays the GenBank record in the Nucleotide database.The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data.
The sequence can be transformed into a variety of other formats, including those used in the EMBL database, as well as the FASTA format which is generally a minimalist format where most of all annotations have been stripped away.

PubMed

PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally. The PubMed database contains more than 35 million citations and abstracts of biomedical literature. PubMed was developed and is maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH).

Citations in PubMed primarily stem from the biomedicine and health fields, and related disciplines such as life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed facilitates searching across several NLM literature resources:

- MEDLINE : MEDLINE is the largest componenet of PubMed nad consists primarlily of citations from journals selected for MEDLINE; artilce indexed with MeSH(Medical Subject Headings) and curated with funding, genetic, chemical and other meta data.
- PubMed Central(PMC) : It is the second largest component of PubMed. PMC is a full text archive that includes articles from journals reviewed and selected by NLM for archiving, as well as individual articles collected for srchiving in compliance with funder policies.
- Bookshelf: The final component of PubMed is ciattions for books and some individual cahpters available on Bookshelf. Bookself is a full text archive of books, reports, databases and other documents related to biomedical, health and life science.

PDB - Protein Data Bank

Protein Data Bank (PDB) is the single global archive of 3D structural data of biological macromolecules(contains greater than 120,000 entries). It includes data obtained by X-ray crystallography and nuclear magnetic resonance (NMR) spectrometry submitted by biologists and biochemists from all over the world.
The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). PDB is a secondary database. Titin is the largest protein chain in our body, with more than 34,000 amino acids. This titanic protein acts like a big rubber band in our muscles. PDB file can be freely downloadable from pdb.org. endenumerate

Genome sequencing

Genome sequencing is an important step toward correlating genotypes with phenotypic characters. Sequencing technologies are important in many fields in the life sciences, including functional genomics, transcriptomics, oncology, evolutionary biology, forensic sciences, and many more. The era of sequencing has been divided

into three generations. First generation sequencing involved sequencing by synthesis (Sanger sequencing) and sequencing by cleavage (Maxam-Gilbert sequencing).
Sanger sequencing led to the completion of various genome sequences (including human) and provided the foundation for development of other sequencing technologies. Since then, various techniques have been developed which can overcome some of the limitations of Sanger sequencing. These techniques are collectively known as "Next-generation sequencing" (NGS), and are further classified into second and third generation technologies. Although NGS methods have many advantages in terms of speed, cost, and parallelism, the accuracy and read length of Sanger sequencing is still superior and has confined the use of NGS mainly to resequencing genomes. Consequently, there is a continuing need to develop improved real time sequencing techniques.

Protein sequencing

Protein sequencing denotes the process of finding the amino acid sequence, or primary structure of a protein. Sequencing plays a very vital role in Proteomics as the information obtained can be used to deduce function, structure, and location which in turn aids in identifying new or novel proteins as well as understanding of cellular processes. Better understanding of these processes allows for creation of drugs that target specific metabolic pathways among other things.
Though several methods exist to sequence proteins the two dominant methods are Mass Spectrometry and Edman Degradation. Other methods that are not as frequently used still can serve very specific roles, such as overcoming inadequacies or acting as a preliminary, that compliment the two predominant methods.

FASTA

FASTA stands for FAST-All or FAST A. It is software to align DNA and protein sequences. It is a sequence analysis tool similar to BLAST. It was developed by David J Lipman and William R Pearson in 1985. The algorithm can be accesses from EBI site. FAST A give better results for nucleotide sequences than protein.

Types

- FastA : Compares a DNA query sequence to a DNA database, or a protein query to a protein database, detecting the sequence type automatically.

- FASTX : Compares a DNA query to a protein database. It may introduce gaps only between codons.
- FASTY : Compares a DNA query to a protein database, optimizing gap location, even within codons.
- TFASTA : Compares a protein query to a DNA database.

Working

- Finds regions of similarity by first breaking the sequence into short sub-sequences, then searching for diagonals with highest density of words that match.
- The alignment in diagonals is then refined.
- Its fast but is not guaranteed to find the best alignment, it may miss match.

Steps involved

- First FASTA prepares a list of words from the pair of sequences to be matched. Words can be 3-6 nucleotides or 1 or 2 amino acids.
- It uses non-overlapping words, it matches the words and makes
- It creates the word diagonal and finds a high scoring match. The output is labeled as unit 1.
- Only if score is sizeable it proceeds to the second level.
- In the second level for every best hit of words, it looks for neighboring approximate hits.

BLAST

Basic Local Alignment Search Tool. It is developed by Stephen Altschul in 1990. It is sequence similarity search program that is used to compare query sequence with sequences databases and finds similarity between them, if any. It shows better results for protein sequences than nucleotide sequences.

Types of BLAST

- BLASTn: Compares nucleotide query sequences to a nucleotide sequence or a database of nucleotide sequences.
- BLASTp : Compares one or more protein query sequences to a subject protein sequence or a database of protein sequences.
- BlASTx : Compares a nucleotide query sequence (translated in protein sequence) against a database of protein sequences.
- tBLASTn : Compares a protein query sequence against the six frame translations of database of nucleotide sequence.
- tBLASTx : Compares the translated nucleotide query sequence against the translated nucleotide sequence database.

Other class include

- MegaBLAST : It is a program optimized for aligning long sequences. It can only work with DNA sequences.
- PSIBLAST : It stands for position specific iterated BLAST. It is useful for protein similarity
- PHIBLAST : Pattern hit initiated BLAST, it can be used to search for a specific pattern or motif.

Working of BLAST

It works through use of a Heuristic algorithm. BLAST searches in two phases

- It looks for short sequences that are likely to have significant matches.
- It tries to extend these matched regions on both sides in order to obtain maximum sequence similarity.

BLASTing a query sequence

Steps involved

- Go to NCBI homepage and click on BLAST
- Copy the sequence of interest in FASTA format from GenBank.
- Paste it in a space provided by BLAST(if nucleotide sequence is taken then use nucleotide BLAST and if it is protein then use Protein BLAST).
- Run BLAST. Analyze the result.

Applications of BLAST

1. Homology searching
2. Species identification
3. Establishing phytogenetic relationships
4. DNA mapping and comparison.

Applications of Bioinformatics

Bioinformatics is a multidisciplinary field used in many life sciences, it has numerous applications.

1. Medicine : It has led to advances in personalised medicine, adapting treatments to each person's genetics.
2. Pharmacology : It has had a fundamental role in pharmaceutical research, particularly in combatting infectious diseases, and in developing vaccines.
3. Genetics : It is applied to gene therapy, particularly in illnesses caused by individual genes that have been affected or inherited. Likewise, the comparison of genomic data adds to our knowledge of the evolutional history of life on Earth.
4. Agriculture : The use of proteomics, metabolomics and genetics allows stronger crops to be developed that are more resistant to drought and insect pests.

5. Livestock : It is used to sequence the genome of farm animals, prevent diseases and to give them more resistance and a better quality of life.

6. Wastes : It allows the DNA sequencing of bacteria and microbes to identify and evaluate their use in wastewater clean-up, radioactive waste disposal or plastics recycling.

Biosensor

It is a sensor, used to measure biological or chemical reactions by simply generating signals which are proportional to the absorption of an analyte in the response. This sensor is designed with a transducer and a biological element like an enzyme, a nucleic acid/ an antibody. In 1956, Leland C. Clark, Jr created the first 'real' biosensor for oxygen detection. He is known as "Father of Biosensors," The oxygen electrode is names after him as "Clark electrode" .

Components of Biosensors

- Analyte : An analyte is a substance whose concentration must be determined. Example : glucose, urea, a medicine, or a pesticide

- Bioreceptor : It is a molecule that recognizes the analyte specifically. It include enzymes, cells, aptamers, deoxyribonucleic acid (DNA), and antibodies.

- Transducer : It transforms one form of energy into another. Its job in a biosensor is to turn a bio-recognition event into a quantifiable signal. This energy conversion procedure is called signalization. Generally, transducers generate either electrical or optical signals that are generally proportional to the quantity of analyte–bioreceptor interactions.

- Electronics : It is the part that simply processes the transduced signal and it displays on the display. It includes complex electronic circuitry that performs signal conditioning like amplification and signal conversion from analog form to digital. After that, the processed signals are quantified through the sensor's display unit.

- Display : The display unit mainly includes a user interpretation system like the LCD of a computer otherwise a printer that produces curves or numbers

understandable by the consumer. This element frequently includes a combination of hardware and software that gives output in an accessible way. The output signal on the LCD can be graphic, numeric, tabular otherwise an image based on the necessities of the user.

Bio-recognition

It is the process of signal creation (in the form of light, heat, pH, charge, or mass shift, etc.) when a bioreceptor interacts with an analyte.

Working principle of biosensor

The working principle of biosensors is signal transduction, so it detects changes within Biological processes and changes them into an electrical signal. Therefore, a Biosensor is a combination of a transducer and Biological sensing element, which changes the data into electrical signals. In addition, there is an electronic circuit that includes a signal conditioning unit (SCU), a Microcontroller/ Processor and a display unit.
Biological process is, any biological material or element such as enzymes, cells, microorganisms, tissues, acids, etc.

Types of biosensors:

1. Electrochemical biosensors : Electrochemical biosensors are simple devices that use bio electrodes to measure electric current, ionic, and conductance changes. It works by using an enzymatic catalytic reaction to consume or create electrons. Redox enzymes are a type of enzyme that does just that. It is further classified as,

    - Amperometric biosensor : The movement of electrons (i.e., the determination of electric current) as a result of enzyme-catalyzed redox reactions is the basis for these biosensors. A constant voltage can be determined when a constant voltage is passed between the electrodes. This causes a change in current flow, which may be measured. The current is proportional to the concentration of the substrate. The simplest type of amperometric biosensor is the Clark oxygen electrode, which measures oxygen decrease. A notable example is the measurement of glucose using glucose oxidase.

- Potentiometric biosensor : Ion-selective electrodes are used in these biosensors to determine changes in ionic concentrations. Because many enzymatic activities involve the release or absorption of hydrogen ions, the pH electrode is the most often used ion-selective electrode. Ammonia-selective and CO_2-selective electrodes are two more significant electrodes.
- Conduct metric biosensor : Changes in ionic species occur as a result of numerous processes in biological systems. The electrical conductivity of these ionic species can be tested. The urea biosensor with immobilized urease is a nice example of a direct metric biosensor

2. Thermometric biosensor : Thermal biosensors or calorimetric biosensors are the terms used to describe them. The creation of heat is linked to a number of biological reactions, and thermometric biosensors are based on this. It is made up of a heat-insulated box with a heat exchanger (aluminum cylinder). In a tiny enzyme packed bed reactor, the reaction takes place. The substrate is transformed to a product and heat is created as it enters the bed. Thermistors measure the temperature differential between the substrate and the product.

   Thermal biosensors can detect even the tiniest changes in temperature. Thermometric biosensors are used to calculate serum cholesterol levels. When the enzyme cholesterol oxidase oxidizes cholesterol, heat is produced, which can be measured. These biosensors can also estimate glucose (enzyme-glucose oxidase), urea (enzyme-urease), uric acid (enzyme-uricase), and penicillin G (enzyme-P lactamase).

3. Piezoelectric biosensor : Piezoelectric biosensors are sometimes known as acoustic biosensors since they work on the principle of acoustics (sound vibrations). These biosensors are made up of piezoelectric crystals. Positive and negative-charged crystals have distinct vibrational frequencies. The adsorption of specific molecules on the crystal surface changes the resonance frequencies, which can be detected using electronic instruments. These crystals can also hold enzymes with gaseous substrates or inhibitors.

4. Optical biosensor : Optical biosensor is a device that works on the principle of optical measurement. Fiber optics and optoelectronic transducers are used. The word optrode is a combination of the words optical and electrode. Antibodies and enzymes, as well as traducing elements, are primarily used in these sensors. Optical biosensors provide non-electrical remote detection of materials in a safe manner.

5. Immuno biosensors : Immuno-biosensors, also known as immunochemical biosensors, work on the basis of immunological specificity, with measurement (mainly) using amperometric or potentiometric biosensors. There are several possible configurations for immuno-biosensors and some of them are depicted in image.

Applications of biosensors

The following are the applications,

1. Biosensors provide the following advantages over lab-based equipment
    - Size is small.
    - Cost-effective
    - Quick outcomes
    - Very simple to use
2. In food industry
    - For quality assurance and control.
    - Applications in the agricultural field, such as crop cultivation and food processing.
    - Quality control in food manufacturing, since it ensures that healthy food has a long shelf life and conforms with standards.
    - It monitors estimates cross-contamination of surfaces and food products.
3. Environment
    - Chemical agents, organic pollutants, potentially poisonous substances, and infections that may represent a health threat have all been detected using biosensors in environmental monitoring.
    - Pollutants are detected using biosensors that measure color, light, fluorescence or electric current.
4. Medical science Cancer diagnosis and monitoring, cardiovascular disease monitoring, and diabetes control are just a few of the applications that have profited from the development of biosensors.

- Biosensors can be used in medicine to monitor diabetic blood glucose levels, identify infections, and diagnose and track cancer growth. Early identification of cancer and effective therapy delivery could be aided by the use of developing biosensor technology.
- Biosensors can detect the presence of a tumor, whether benign or cancerous, by measuring the amounts of particular proteins expressed and/or secreted by tumor cells. They can also tell if treatment is effective in lowering or removing cancerous cells.
- Cardiovascular disorders, which are the leading cause of death, are still regarded as one of the world's most serious problems, with approximately one million individuals suffering from them.

Disadvantages of biosensors

Biosensors have relatively poor sensitivity for many of the clinically related targets and semi-quantitative or qualitative results. To enhance the detection limit, the latest efforts have focused on the amplification of the signal.

Abbreviations

- NCBI : National Center for Biotechnology Information
- EBI : European Bioinformatics Institute
- EMBL : European Molecular Biology Laboratory
- TrEMBL :Translated EMBL
- DDBJ : DNA Data Bank of Japan
- BLAST : Basic Local Alignment Search Tool

- FASTA : Fast All
- dbEST :Databases of Expresses Sequence Tag
- HMM : Hidden Markov Model
- CINEMA : Color Interactive Editor for Multiple Alignments
- ExPASy : Expert Protein Analysis System
- EcoGene : E.Coli Genome Database
- INSDC : International Nucleotide Sequence Database Collaboration
- PDB : Protein Data Bank
- BSML : Bioinformatic Sequence Markup Language
- CBCB : Center for Bioinformatics and Computational Biology
- BLOSUM : Blocks Substitution Matrix
- dbGaP : Databases of Genotypes and Phenotypes
- Pfam : Protein Family
- PSSM : Position Specific Scoring Matrix

- SMPDB : The small Molecule Pathway Database
- ENA : European Nucleotide Archive
- RefSeq : Reference Sequences
- UniProt : Universal Protein
- PIR : Protein Information Resources
- SCOP : Structural Classification of Proteins
- CATH : Class Architect Topology Homology
- OMIM : Online Mendelian Inheritance in Man
- dbSTS : database of Sequence Tagged Sites
- dbSNV : Database of Short Genetic Variation

Questions

1. Explain different types of data base.
2. What is BLAST and explain types of BLAST.
3. What is FASTA and explain types of FASTA.
4. What is Biosensors. Mention types of Biosensors.
5. Explain Biosensors.
6. Explain types of biosensors.

7. Explain applications of Biosensors.
8. What does EMBL, DDBJ and NCBI stands for.
9. Explain genome sequencing.
10. Explain protein sequencing.
11. Explain GenBank.
12. Explain PDB and PubMed.
13. Explain Thermometric Biosensor.
14. Explain applications of Bioinformatics.

Department of Physics, NIE Estd 1946