CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

INSTITUTE-UNIVERSITY INSTITUTE OF ENGINEERING

ACADEMIC UNIT-II

Computer Science Engineering

Subject Name-Biology For Engineers

Subject Code- 20SZT148

FASTA AND DATA MINING

DISCOVER . LEARN . EMPOWER

FASTA AND DATA MINING

Course Outcome

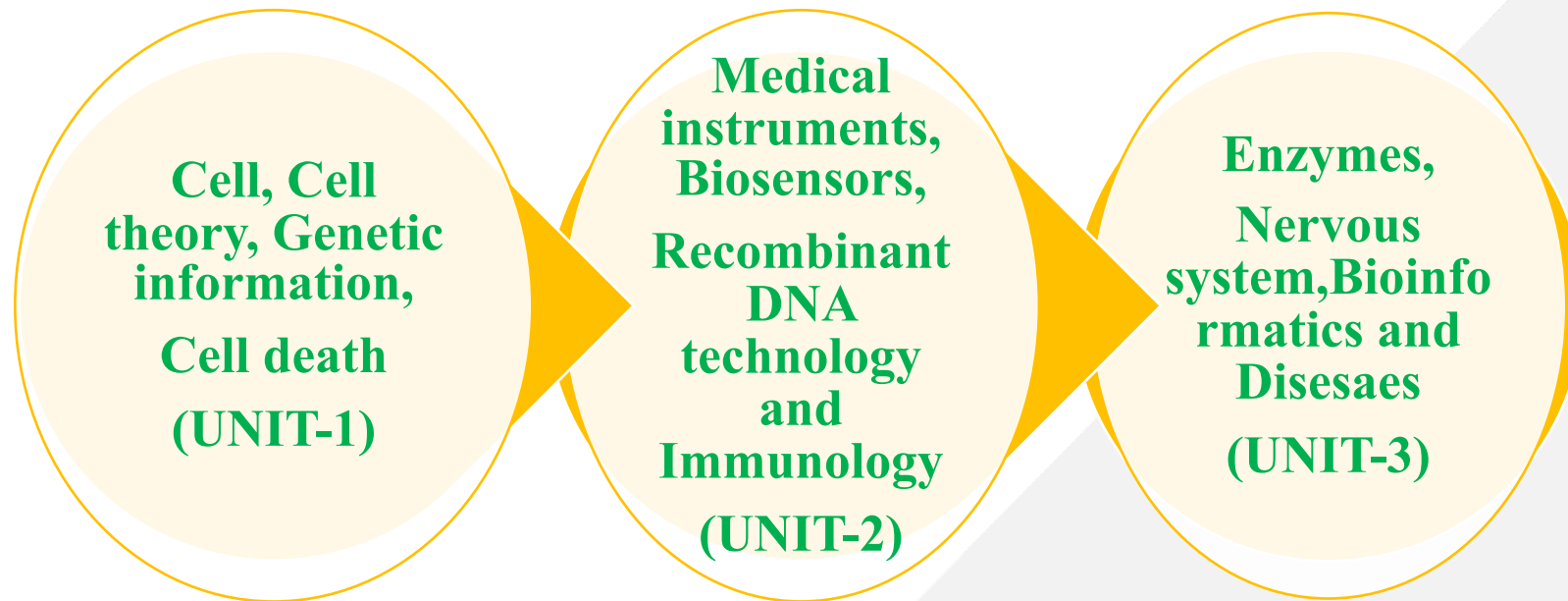| CO Number	Title	Level
CO1	It gives an idea about the about the basic cell biology.	Understanding
CO2	It deals with the idea of uses of biology in engineering.	Understanding
CO3	It provide knowledge about the uses of softwares in biology field.	Remembering

2015-03-23   47

Will be covered in this lecture

https://www.electronicshub.org/types-of-transducers/

BIOLOGY FOR ENGINEERS

Cell, Cell theory, Genetic information, Cell death (UNIT-1)

Medical instruments, Biosensors, Recombinant DNA technology and Immunology (UNIT-2)

Enzymes, Nervous system, Bioinformatics and Disesaes (UNIT-3)

FASTA

- FASTA package was 1st described as by Lipman & Pearson in 1985.

- FASTA is a DNA & protein sequence alignment software.

- FASTA is a fast HOMOLOGY search tool.

- Similar to BLAST , but this tool will speed up sequence comparison when compared with BLAST.

- Hence it is called FASTA.

FASTA

- FAST-P stands for protein , compare the amino acid sequence of proteins & FAST-N stands for nucleotide alignment , compare the nucleotide sequence of DNA.

- It is Bioinformatic tool.

- It is biological database.

- It compare the amino acid sequence of different proteins or the nucleotide sequence of different DNA.

FASTA

FASTA Format :

- Header line which begins with '>' gives a unique identifier for the sequence.

- The gap character '-' is found in locations corresponding to gaps in the sequence.

- E.g. >JC2395   NVSDVNLNK---

- YIWRTAEKMK---QKIQLLQWYQSHGKT

FASTA

Amino acid code	Meaning
A	Alanine
B	Asparagine
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
J	Leucine (L)
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
O	Pyrrolysine

https://image.slidesharecdn.com/fasta-170226094121/95/fasta-7-638.jpg?cb=1488102133

FASTA

Amino acid code	Meaning
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threosine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamic acid (E)

https://image.slidesharecdn.com/fasta-170226094121/95/fasta-7-638.jpg?cb=1488102133

FASTA

FASTA Programs:
- FASTA - compares a query protein sequence to a protein library or DNA sequence to a DNA library.
- TFASTA – compares a query sequence to a DNA library, after the DNA sequence has been translated in all 6 reading frames.

- FASTS – compares a set of ordered peptide fragments, obtained from mass-spectrometry analysis of a protein against a protein database.
- FASTX & Y – compares a query DNA sequence to a protein sequence database , translating the DNA sequence in all 6 reading frames.

FASTA

salient features of fasta

• FASTA find out patches of sequence SIMILARITY between the query and database.

• FASTA may contain GAPS.

• FASTA is RAPID.

• FASTA is working on the basis of FASTA algorithm developed by Fredi and Barton.

FASTA

- Programs can be run locally or by sending query sequence to the FASTA Email serer.

- Program achieve a high level of sensitivity for similarity search at high speed.

- FASTA is not guaranteed to find best alignment between the query and database.

USES OF FASTA

- Identification of species.
- Establishing phylogeny.
- DNA mapping.
- Understanding biochemical functions of protein.
- Studying evolution of species.
- Identification of mutations.
- Calculating molecular weight.
- Studying atomic composition.
- Knowing amino acid composition.
- Identification of specific function of sequence.
- Comparison of DNAs.

DATA MINING

- Obtaining information from several scientific resources becoming is more difficult as the volume of information grows.

- Number of different resources/databases is growing and simple search has to be repeated for each of them.

- Filtering relevant information is a big intellectual/computational burden.

DATA MINING

- Obtaining information from several scientific resources becoming is more difficult as the volume of information grows.

- Number of different resources/databases is growing and simple search has to be repeated for each of them.

- Filtering relevant information is a big intellectual/computational burden.

DATA MINING

- Retrieval, analysis and formatting (parsing) of information into searchable databases

- Recognition of patterns

- Recognition of natural language

- Extraction of semantic or grammatical relationships

- Coreference: terms that refer to the same object

DATA MINING

Text mining example

- Query: Find promoters known to work in E.coli with s70 holenzyme (Es70) aka sD
-
```
PREFIX sbol:http://sbols.org/sbol.owl#
PREFIX pr:http://partsregistry.org/#
SELECT DISTINCT ?name
WHERE {
?part a sbol:Part;
        sbol:status ?st;
        sbol:name ?name;
        sbol:dnaSequence ?seq;
        a pr:promoter;
        a ?cl.
FILTER (?cl =pr:sigma70_ecoli_prokaryote_rnap
        && ?st !='Deleted')}
```

DATA MINING

Applications of Data Mining

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

https://image.slidesharecdn.com/introductiontobioinformatics-150319175041-conversion-gate01/95/introduction-to-bioinformatics-51-638.jpg?cb=1427042135

CONCLUSION

- FASTA is a DNA & protein sequence alignment software.

- FASTA is a fast HOMOLOGY search tool.

- FASTA - compares a query protein sequence to a protein library or DNA sequence to a DNA library.

- TFASTA – compares a query sequence to a DNA library, after the DNA sequence has been translated in all 6 reading frames.

REFERENCES

- C.B.Powar, 2010.Cell Biology.5th Ed,Himalyan Publishing House.

- Leshie Cromwell, Fred.J. Weibell and Erich.A.Pfeiffer. 2003. Biomedical instrumentation and measurements. 2nd edition, PHI.

- John G. Webster 1998. Medical Instrumentation: Applications and Design, 3rd edition, Jon Wiley and Sons, New York.

- Jeremy M. Berg, John L. Tymoczko and Lubert Stryer. 2006. "Biochemistry," 6th Ed. W.H. Freeman and Co. Ltd.

- Robert Weaver. 2012 "Molecular Biology," 5th Edition, MCGraw-Hill.

- Jon Cooper, **, 2**004. "Biosensors A Practical Approach" Bellwether Books.

- Martin Alexander, 1994 "Biodegradation and Bioremediation," Academic Press.

THANK YOU

For queries
Email: subject_code_2020@gmail.com