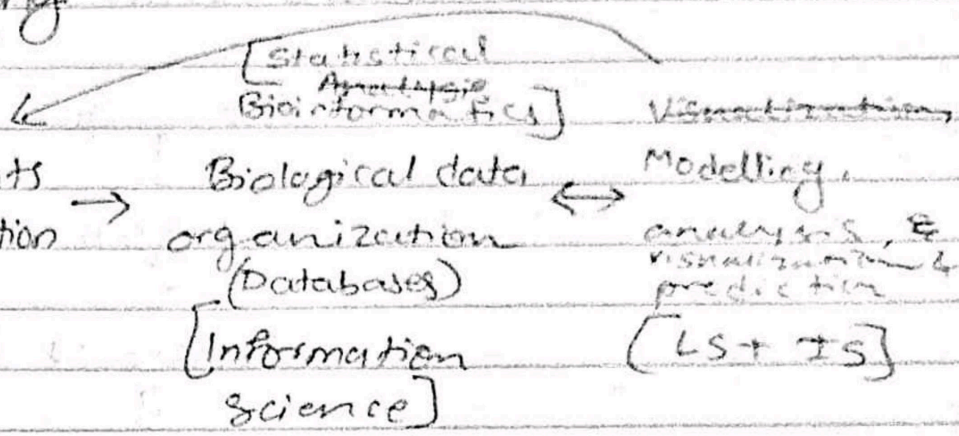
BOOK → Pavsner
→ Essentials of Bioinformatics
↳ by Zing Zong
→ David Mount



- Term Bioinformatics was invented by Paulien Hogeweg and Ben Hesper in 1970 as Study of informatic processes in biotic systems
  ↳ Paulien Hogeweg is a dutch- theoretical biologist & complex systems researchers study

- Bioinformatics is a science that is not only used to ~~algo~~ develop algorithms, store, retrie organize & analyze biological data but to curate data.

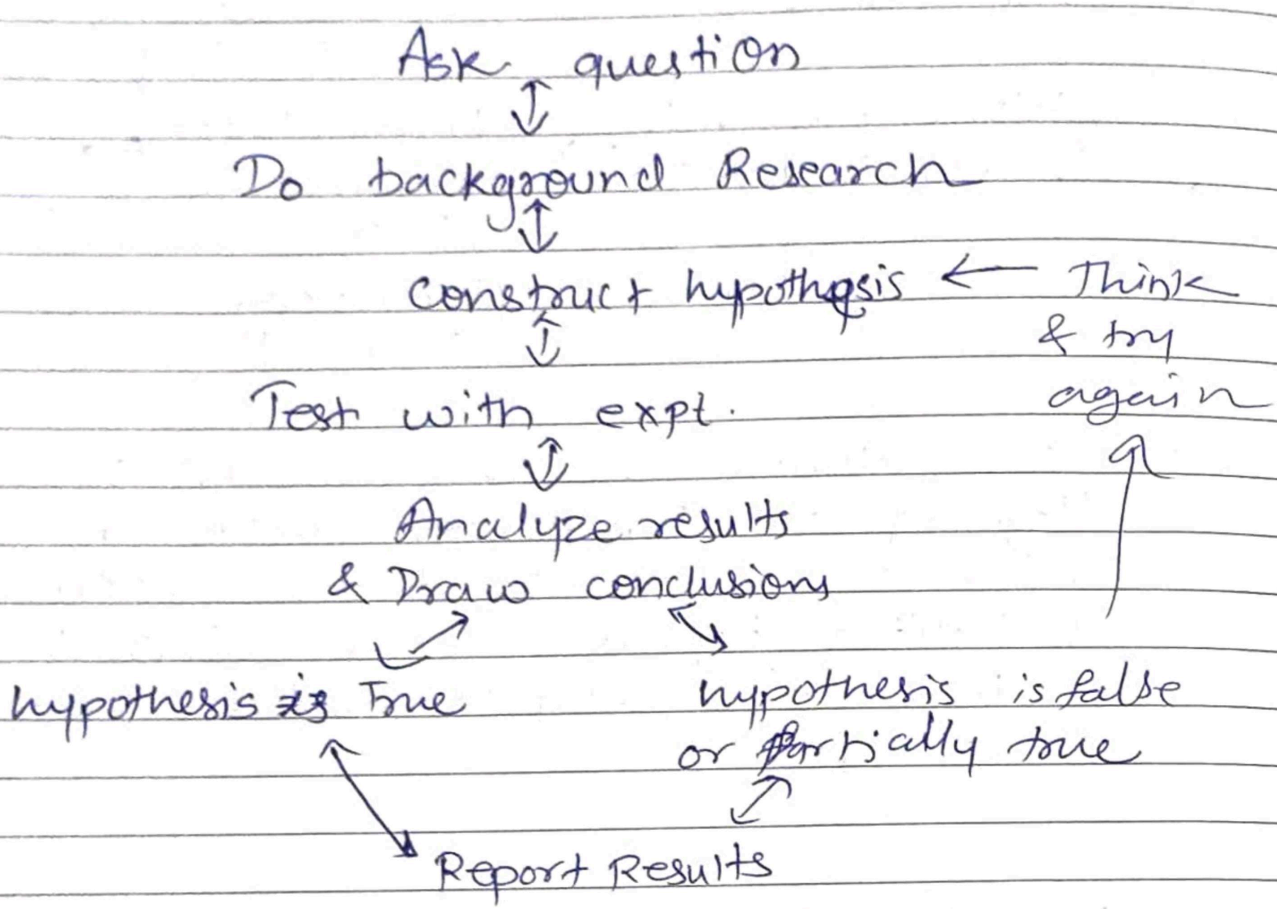
- Components of Bioinformatics
  - Biological data
  - Algorithms
  - Computers

- Speedups data production & data interpretation

- Bioinformatics = Biologist + CS + IT

- Human genome ⇒ 3.5 billion basepairs

⇒ How to work with bioinfo

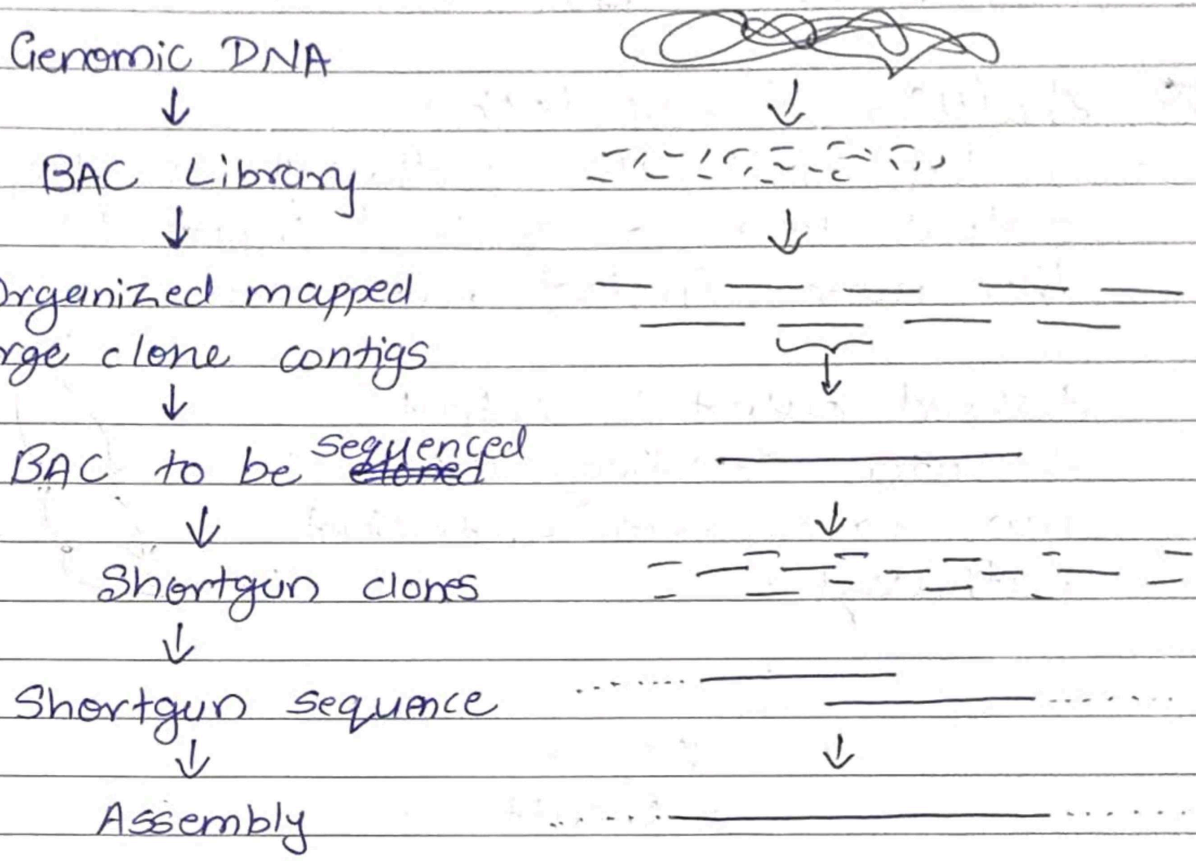


DATE 9 8 23 (3)

✱ Sequence Assembly

→ NGS [Shortgun]

Sequence Repository Database (SRA)



→ Illumina is the base platform for NGS

✱ Genome annotation

↳ naming/labeling of the genes/genome that comes from NGS.

- annotations can be done by researchers or NCBI. NCBI is preferred for cross-checking.
- Genome Browsers ⇒ UCSC, ENSEMBL, NCBI, etc.
- Genome tracks
- Data validation ✱ via. cross-checking & backtracing

✱ Molecular evolution

↳ Eg. Phylogenetic analysis

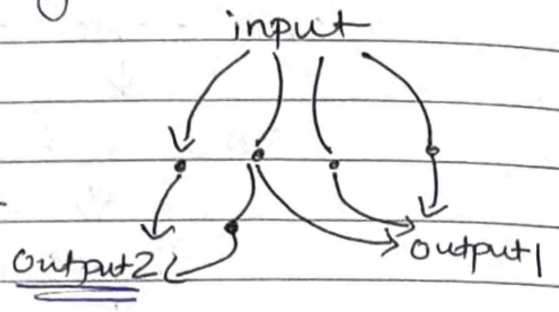
Analysis of gene expression.

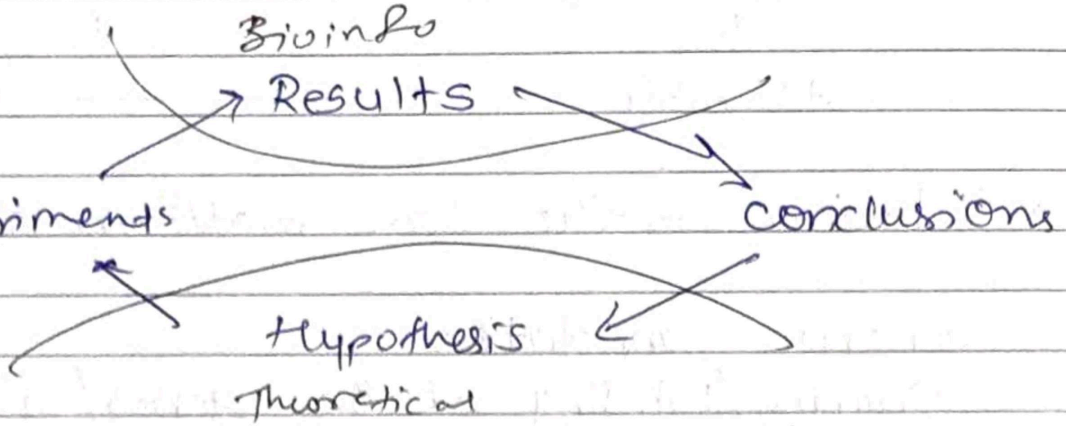
- to find out which changes in which gene or its expression causes which change.
- Database ⇒ Gene Expression (GEO)

Analysis of regulation.

- study all the pathways for the particular output from a specified input to select the most effecient pathway.

desired output is output 2 & later selecting the most favourable & efficient pathway.





DATE 10 8 23 (5)

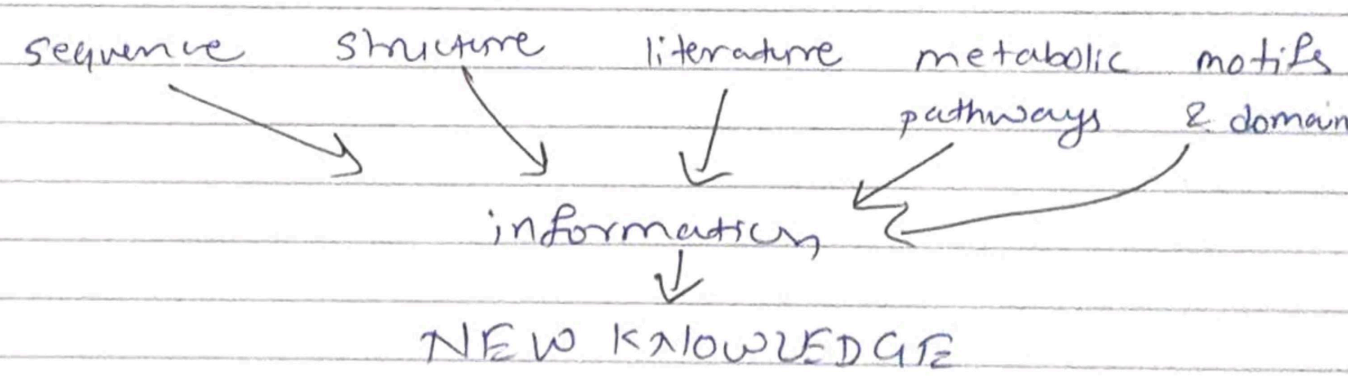
Pairwise Sequence Alignment

File formats for biomolecular sequences

⇒ Aims
- to understand the conventions regarding the presentation of DNA & Protein sequence info.
- to understand the logic underlying these conventions.
- to become familiar with commonly used sequence file formats
- to become familiar with READSEQ programme for the interconversion of file formats

- present a nucleotide or protein

⇒ where does data comes from



⇒ Database / Resource
→ Collection of data in related format
↳ structured
↳ updated periodically (releases)

Cross linked data within all dbs all over world → NCBI ← sequence
PDB ← structure

→ Includes associated tools/software necessary for accessing database, updating DB, DB information insertion, DB information deletion, etc.

→ Type & content of Data.
  - Sequence, structure, nucleic acid or Protein
  - imp. biological info such as about enzyme & their pathways, mutations, drugs, diseases, images, etc.

→ Based on source of data
  - primary db → raw info shared to portal by user
  - secondary db → curated info
  - knowledge bases → open access to all/specialized info.
  - integrated db → cross-linking of data.

Amino acid (one & 3 lettercode) | physicochem prop. | structure

DATE 17 8 23

⇒ DNA

File types

Myoglobin

H_2N- Met – Val – Tyr – Gly – Iso – Lys – COOH

MVYGIK

single letter protein

- many software packages have been developed for the analysis of DNA and protein sequences
  - ↳ these software stores & analyzes the DNA & protein sequence in different file formats.
- The variety of software packages will usually only accept a specific file format.
- Diff. databases holds info. in diff formats.
  - ↳ This makes the situation worse.
    - ↳ inorder to use this data, the info has to modified & desired file format must be created.
- It is an essential skill to be able to recognize the diff. file formats & to be able to interconvert files between formats.

⇒ Primary Biological Databases

→ Nucleic acid
1. EMBL
2. GenBank
3. DDBJB

→ Protein
1. PIR
2. MIPS
3. SWISS-PROT
4. TrEMBL
5. NRL-3D

→ Nucleotide DB
1. EMBL → Nucleotide seq. db
2. Ensembl → Automatics annotation of eukaryo genomes
3. Genome server → Overview of completed genomes at EBI
4. Genome MOT → Genome monitoring table
5. EMBL-Align → Multiple sequence alignment db
6. Parasites → Parasite genome db
7.
8.

→ EMBL/GenBank/DDBJB
- same info (diff syntax & format)
- non-confidential data is exchanged daily.

→ DB related to Genomics.
- info on genes, gene mapping (location), nomenclature & links to seq. db.
- exsists for most org. imp. for life sci research
- Eg. ① MIM
②
③
④
⑤

⇒ Plain Sequence Format.
- may contains only IUPAC characters & spaces but not numbers.
- file in (PSF) may contain one sequence, while

⇒ FASTA Format (FF)
- can contain several sequence
- begins with greater than sign (>) & single-line description, followed by the sequence on next line. & ends with 2 slashes (//)

⇒ EMBL Format (EF)
- contains several seq.

⇒ GenBank Formad (GF)

✗ ENSEMBL
- contains all human genome DNA seq. currently available in public domain.
- Automated annotation
  ↳ by using diff. software tools, features are identified in DNA sequence.
  ①
  ②
  ③
  ④
- Maintained by EMBL & Sanger's

DATE 18 8 23 B

Protein Databases

→ SWISS-PROT
- Annotated Sequence DB

→ TrEMBL
- DB of EMBL nucleotide translated sequences

→ InterPro
- Integrated res. for protein families, domains & functional sites.
- best for cross-referencing proteins

→ CluSTr
- offers automatic classification of SWISS-PROT & TrEMBL

→ IPI
- A non-redundant human proteome set constructed with SWISS-PROT, TrEMBL, EnsEMBL, RefSeq.
- no duplicated data

→ GO
- Provides assignments of gene products of Gene Ontology (GO) resource

→ Proteome
- Statistical & comparative analysis of predicted proteomes of fully

→ Protein profiles
- Tables of SWISSPROT & TrEMBL entries & alignment for protein families of protein profile

→ IntEnz
- Integrated relational Enzyme database (IntEnz) will contain enzyme data approved by Nomenclature Committee

✡ SWISS PROT
- Annotated protein sequence db, established in 1986
- maintained collaboratively by Dept. of Medical Biochemisty, University of Geneva & EBI, since 1987.
- complete, curated, non-redundant & cross-referenced with 34 other dbs
- Highly cross-referenced.
- more than 80000 species are available in DB.
- The first 20 results give about 42%.

⇒ SWISS PROT file format.

✡ TrEMBL [Translation of EMBL]
- computer-annotated supplement of SWISS-PROT, as it is impossible to cope up with the flow of data: ~~using only EMBL~~
- Well structure SWISS-PROT- like resource
- It contains the data that isn't yet available in SWISS-PROT

✡ Protein DB (PDB)
- imp in solving real-problem in molecular biology
- PDB, established in 1972 at Brookhaven National Laboratory (BNL)
-

Pubmed → 36876833

DATE 19/8/24 NO. 16

- PubMed contains summarized or brief information about the specific paper or topic. Whereas, PubMed Central (PMC) contains the whole scientific or research papers or articles.
- Amount of data in PMC is much more than it is in PubMed.