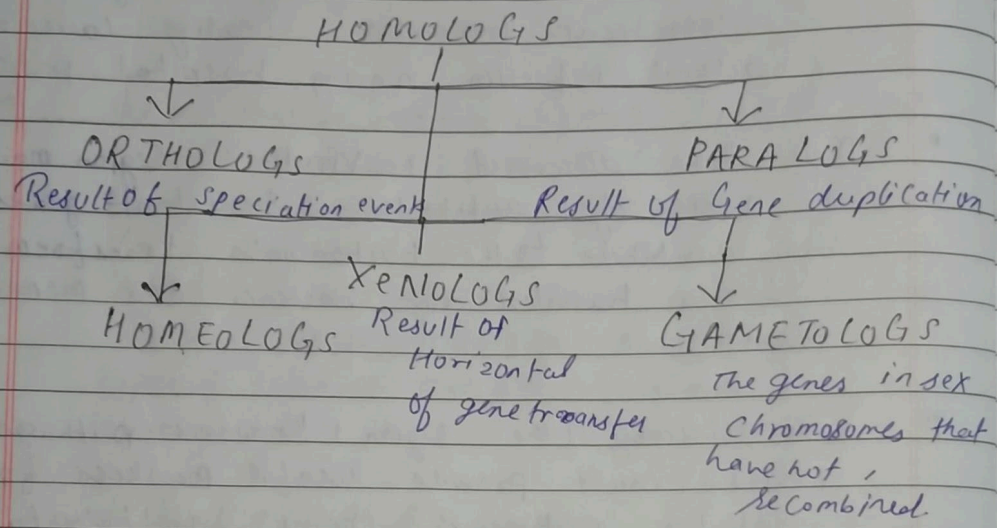
The Notion of Homology

Date .....
Page .....

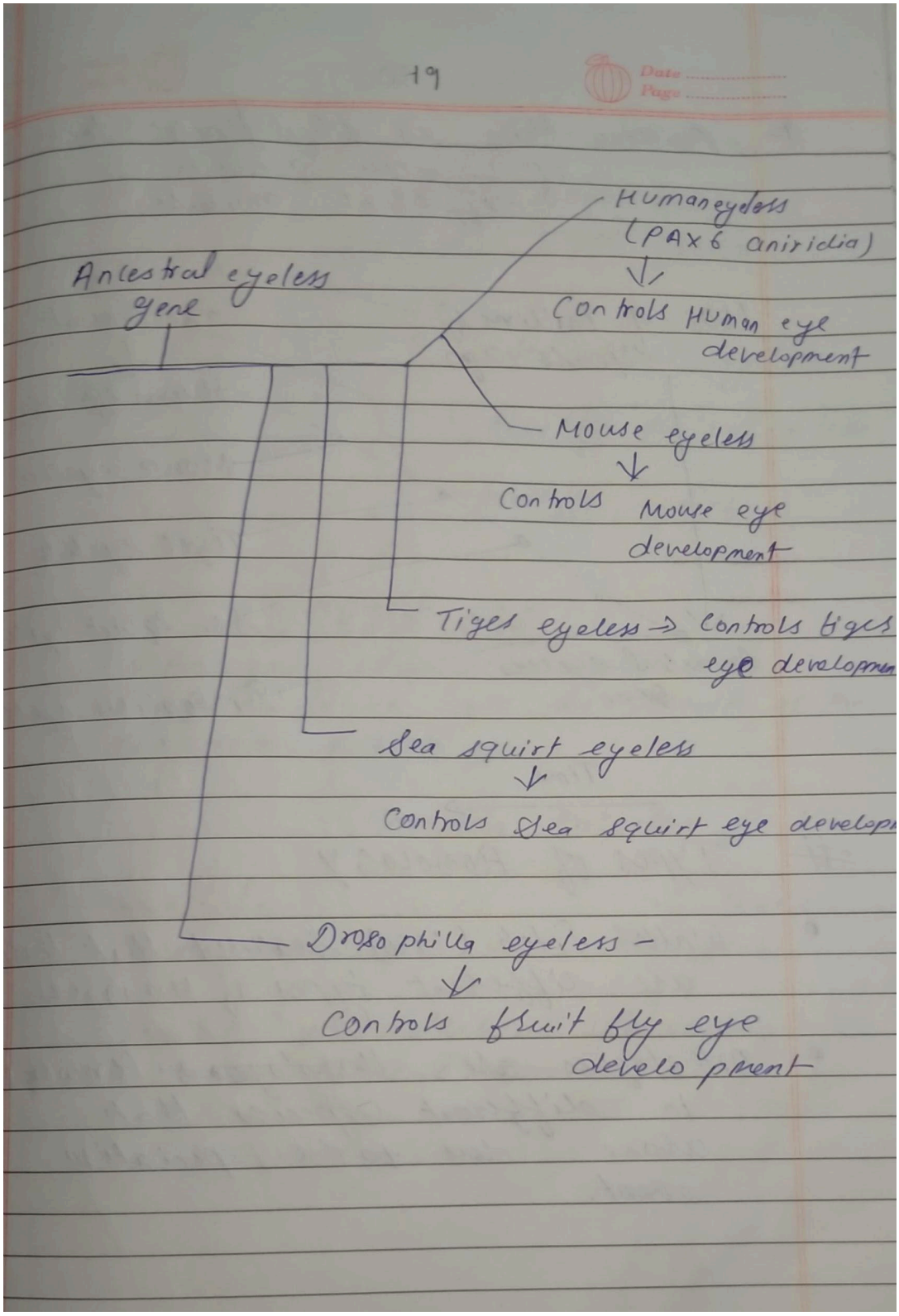
Introduction

Homologs are two or more sequence that descend from a common Ancestral sequence. Homologs are results of divergent evolution.



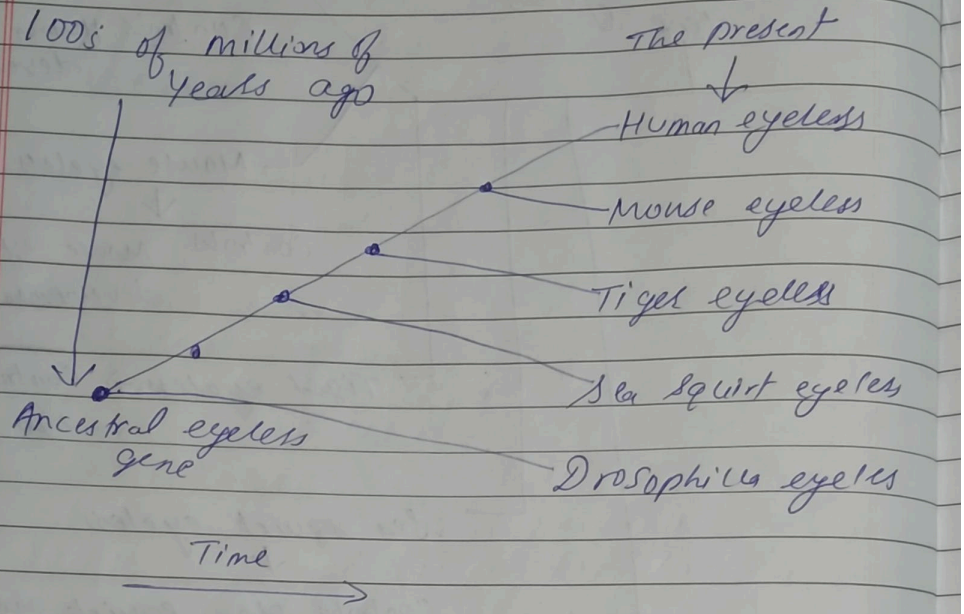
Slightly different versions of the eyeless gene control eye formation in many animals.

eyeless genes in different animals are Homologues.
i.e - they are Homologous (related) genes that descended from an ancestral gene in the ancestor of all these animals.


Ancestral eyeless gene
Human eyeless (PAX6 aniridia)
Controls Human eye development
Mouse eyeless
Controls Mouse eye development
Tiger eyeless → Controls tiger eye development
Sea squirt eyeless
Controls Sea squirt eye development
Drosophila eyeless
Controls fruit fly eye development

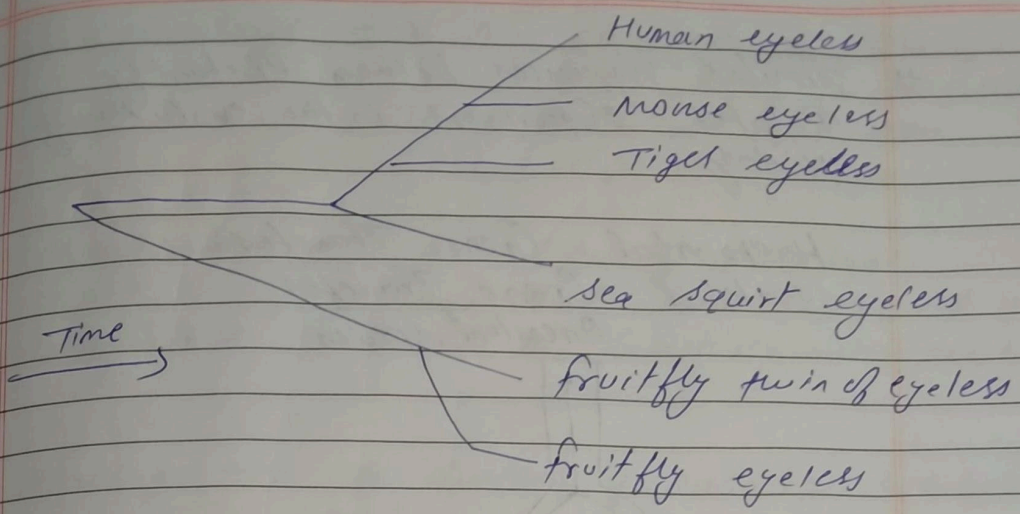
Date .....
Page .....

Aside: this is phylogenetic tree of eyeless genes in different animals.



Types of Homology

- Walter Fitch (1970) realised that there are different types of Homologues.
- Orthologues are ~~Homologeous~~ Homologues in different species that arose due to the speciation event.



ORTHOLOGS

orthologous sequences are those sequences which are the results of the speciation process.

Common Ancestor

(Human) ↓ Hemoglobin

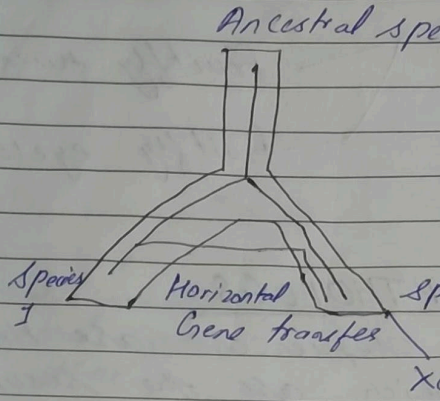
↓ (Rat) Hemoglobin

Similar or identical sequences in different species.

XENOLOGS

Similar Sequences between Distantly Related Organisms in the evolutionary History.

Horizontal Gene Transfer
Lateral Gene Transfer



Class — Description

Orthologs ⟶ Genes that diverged by speciation are Orthologous. They may or maynot have the same function.

Paralogs ⟶ Genes that originate from an ancestral duplication and reside within the same species are paralogous.

Date
Page

In-paralogs → A subtype of paralogs. Genes that have duplicated after the last speciation event are in-paralogs.

Out-paralogs or alloparalogs → A subtype of paralogs. Genes that have duplicated before the latest speciation event.

Ohnologs → A subtype of paralogs. A whole genome duplication results in ohnologous genes. They have a common origin and have diverged an identical length of time.

Gametologs → The genes in sex chromosomes that have not recombined.

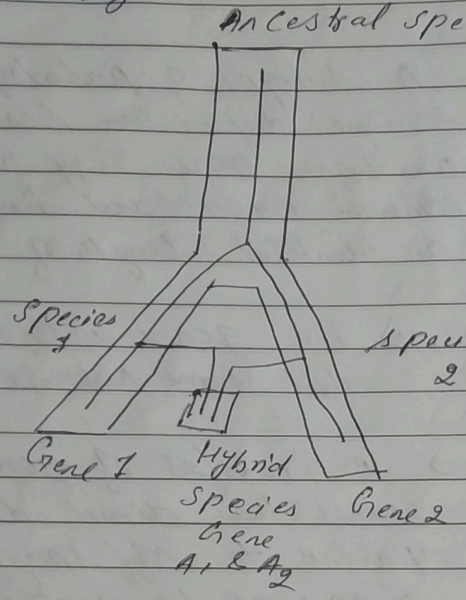
Xenologs → A transfer of genes, by means other than direct decent from parents to offsprings from one organisms to another, results in xeno-xenologous genes.

Homoeologs → A subtype of xenologs. The hybridization of genes from two separate species produces homoeologous species

HOMEOLOGS

Homeologs are special cases of Xenologs.

The genes which are separated by a speciation event when Hybridized together via Lateral gene transfer the gene sequences are Known as Homeologs.



Homeologs

Date
Page

SUMMARY

HOMOLOGS :- Two or more sequence that descend from a common ancestral sequence.

ORTHOLOGS :- Sequence which are the results of the speciation process.

paralogs :- Sequence which are the results of the gene duplication process.

Xenologs :- Similar sequence b/n Distantly related organisms in the evolutionary history.

Analogs :- Sequence that show similar structure or function but don't share any common ancestral sequence.

SYMA PARALOGS

Genes that have duplicated after the latest speciation event are known as In-Paralogs or Syma paralogs.

Date
Page

Alloparalogs

Genes that have duplicated before the latest speciation event are known as out-paralogs or alloparalogs.

OHNOLOGS

These are genes that are a result of a duplication of a whole Genome

ANALOGS

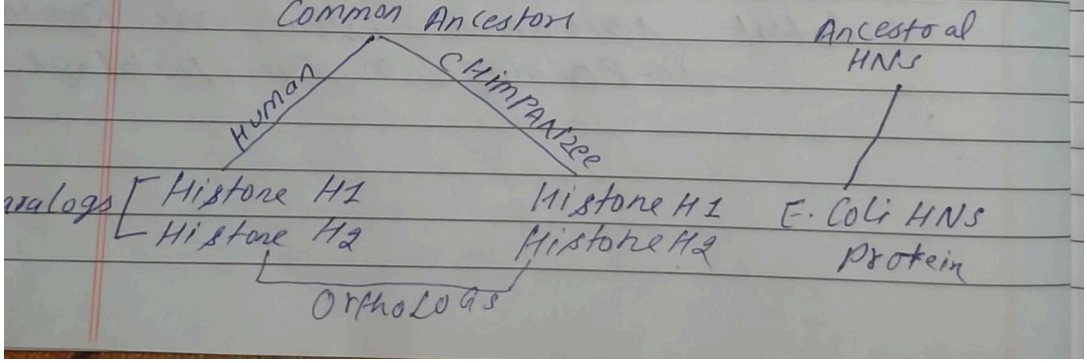
Analogeous genes are genes that have identical or similar functions but don't share a common ancestor.

Convergent evolution

(I) Squid eye (II) Human eye

Analogs have Homologous activity but heterologous origin

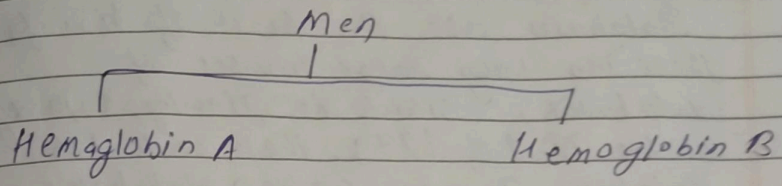
HOMOLOGS VS. ANALOGS



Date
Page

PARALOGUES

paralogous Sequence are Homologeous that have diverged within one species i.e arising during gene



Similar Sequences within SAME Species.

The End

Sequence Information Sources

Databases :—

A databases is a vast collection of data pertaining to a specific topic e.g— nucleotide sequence, protein sequence etc.— in an electronic environment.
Databases are the heart of bioinformatics There is a very large number of databases, which is growing rapidly. at the end of 1999, there were 226 databases. In 2000, 55 new databases were created, raising the total no to 281.

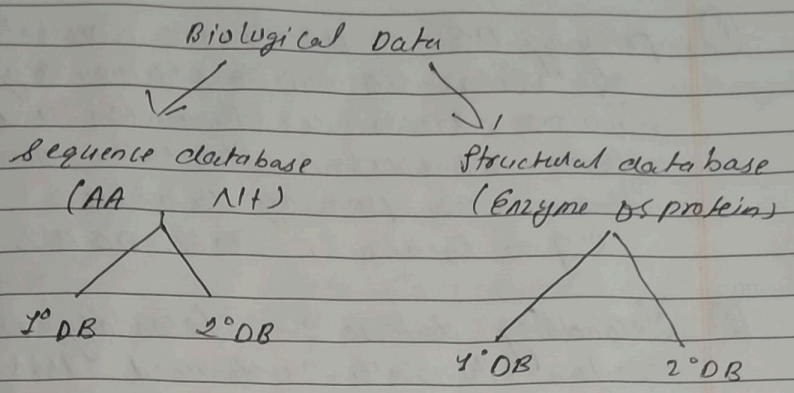
Nucleotide Sequence databases :—

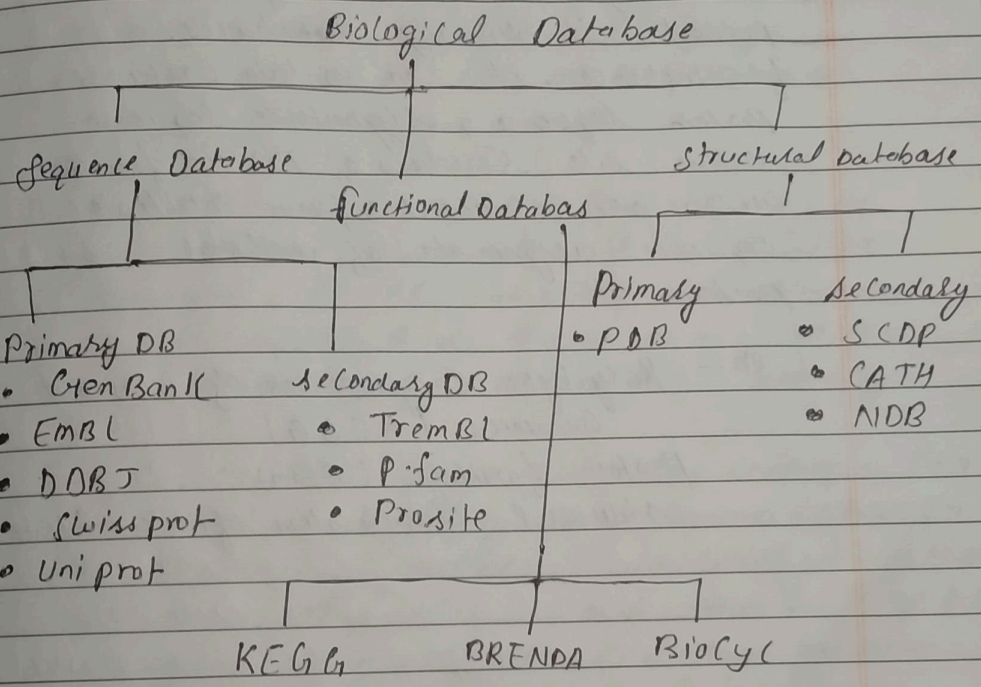
The major nucleotide sequence databases are Gene-bank held by NCBI, U.S.A DNA Data bank of Japan (DDBJ) and the Nucleotide sequence database maintained by EMBL.

In addition to other nucleotide sequence databases have been created.

Date
Page

Classification of Databases.





Types of Databases

(1) primary D.B :— Databases consisting of data desired experimentally such as nucleotide sequences and three dimensional structures are known as primary database.
e.g — Genbank, EMBL, DDBJ, NDB

(2) Secondary database :— Secondary databases contain the data is obtained through the analysis of treatment of data present in primary databases. for instance, it can contain conserved protein sequence, signature sequence active site resiclues of protein families which are obtained from multiple sequence alignment of related protein etc.

eg — Kaisyotype Encyclopedia of Genes and Genomics (KEGG)
Protein families (Pfam)
structural classification of protein (SCOP)

Date
Page

functional Database

Development of metabolic databases derived from the comparative study of metabolic pathways cater the industrial needs in more efficient manner to further the growth of systems biotechnology.

KEGG :— The Kyoto Encyclopedia of Genes and Genomics (KEGG) is the primary resource for the Japanese Genome Net service that attempts to define the relationships b/n the functional meanings and utilities of the cell or the organisms and its genome information.

BRENDA :— It is main collection of enzyme functional data available to the scientific community. It is maintained and developed at Biochemistry and Bioinformatics at the technical University of Braunschweig, Germany.

BioCyC :— The BioCyc database collection is a compilation of pathway and genome information for different organisms. It includes two other databases, EcoCyc, which describe E.coli K-12 metaCyc, which describe pathways for more than 3000 organisms.

Date
Page

Sequence Database

Nucleotide and protein sequence database represent the most widely established biological databases.

Serve as repositories for wet lab result and primary source for exponential result.

Genbank :— The Genbank nucleotide database is maintained by the NCBI which is part of the National Institute of Health (NIH), a federal agency of the us agency.

DDBJ :— DNA Data protein bank of Japan is a biological database that is run by the National Institute of Genetics, Japan.

PIR :— The protein information Resource an integrated public bioinformatics that support genomic and proteomic Research and scientific. PIR has provided many protein databases and analysis tool including PSD of the functional annotated protein sequences.

Date
Page

Swiss Port :- Swiss port is a protein database sequence and Knowledge database. It is well known for its minimal redundacy high quality of annotation, use of standard homenclature. It contains all amino acid sequence database taxonomic group data and citation information.

TrEmbl :- This data consist of computes annoted entries derived from the translation of all coding sequence in the nucleotide databases.

Structure Database

Protein Data Bank :-
The PDB is the single world wide archieve of structural data of Biological Macromolecules established in the Germany. National lab in 1971.

It Contains structural info. of the macromolecules determined by X-ray Crystallographic NMR methods. PDB is maintained by the Research Collaboratory for structural Bioinformatics (RCSB).

Secondary :-

SCOP - structural classification of protein)
CATH - Class, Architecture Topology Homologous family
NDB - Nucleic Acid data bank.

Date
Page

Entrez :—

Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence, gene-centered and genomic-mapping information. 3D structure data, PUB med medline.

Ton
The system is produced by NCBI and available via Internet.

Entrez covers over 20 databases including the complete protein sequence data from PIR-International, PRF, SWISS, PROT, and PDB and nucleotide sequence data.

The Entrez retrival system uses an intuitive users interface for rapidly searching sequence and bibliographic data.

UNIT-2

Date
Page

PROTEIN INFORMATION SOURCES

protein database can be sequence database or structural database.

Protein sequence database:-

The protein sequence database was developed at National biomedical research foundation by ~~margaret~~ Margaret dayoff in 1960's.

The protein sequence database was collaborately maintained by PIR.

The Protein Information Resource (PIR) Database:-

It is main protein sequence database.

This database is classified into 4 classes:-
PIR1:- Classified and annotated entries.

PIR2:- Prilimihary entries

PIR3 — Unvesified entries

PIR4 — Conceptual translation of the sequence that are not transcribed, that are genetically engineered etc.

Date
Page

SWISS PROT :—

It is a protein sequence database maintained collaborative by medical biochemistry at the university of Geneva Germany.

The d.B endevours to provide high level annotation description of the function of the protein and structure of the domains post translational modifications Vasients and so on.

They are interlinked to many source and have minimal redundancy.

TREMBL :—

It was created in 1966 as a computer annotated suppliment to swiss prot.

The d.B contains translation of all coding sequences.

Two main sections :—

SP – TREMBL – Contain entries that not been annotated but they are eventually incorporated into swiss prot.

Date
Page

REM - Trembl :— Contains entries that are not included into swiss prot
e.g — Eg seq - synthetic seq.

NRL - 3D :—
This d.B is produced by PIR from sequences extracted from PDB.
It is used both for similarity searches and keyword introgation.

Structural Database

They store a collection of 3 dimensional biological macromolecules structures of proteins.

The last established database for protein structures is protein data bank (PDB).

PDB : It contains following information :—

Name of the protein
The species
Structural determination
Amino Acid sequence
Additional Information.

Date
Page

Introduction to BLAST

An important goal of genomics and proteomics is to determine if a particular sequence is like another sequence. This is accomplished by comparing the new sequence with sequence that have already been reported and stored in database.

This process is principally one that using alignment procedures to uncovers the "like" sequence in the database.

The alignment process will uncovers those regions that are identical or close similar and those regions with little similarity.

Two alignments types are used Global and Local.

BLAST

BLAST - stands for BASic Local Alignment Search tool.

It is local alignment tool.

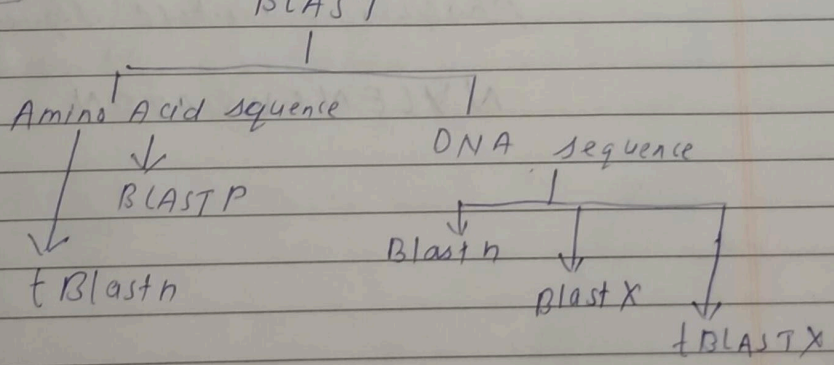
It helps to find regions of local similarity between sequences.

It is a program Compares nucleotide or Protein sequence to sequence databases and Calculates the statistical significance of matches.

BLAST Can be used to infer functional and evolutionary relationships b/n sequence as well as help identify members of gene families.

Google $\xrightarrow{\text{search}}$ NCBI HomPAge $\longrightarrow$ Sequences

Types



Steps

Specifying A sequence Of Interest

↓

Selecting BLAST program

↓

Selecting Database

↓

Selecting Optional Parameters

↓

Selecting formatting Parameters

PROCESS

The first step of the BLAST algorithm is to break the Query into short words of specific length.

e.g — 12 amino acid near the terminal of the Arbidopsis thaliana protein phosphoglucomutase sequence are!-

NYLENFVQATFN

Date
Page

This sequence is broken down into three character words by selecting the first Amino Acid characters.

NIYL YLE LEN ENF NFV FVQ
VQA QAT ATF TFN

These words are then compared against a sequence in a database.

for ex — word match with rabbit ~~against a sequence in a database.~~ muscle phosphoglucomutase

Query ENF
SSTNYAEMTIQSRMSTVEPAQR

Application of BLAST

BLAST can be used for several purposes these include.

* Identifying species
* Establishing phylogeny
* DNA mapping
* Locating Domains.

Theory

Date
Page

Multiple Sequence Alignment

A multiple Sequence Alignment (MSA) is a basic tool for the sequence alignment of two or more biological sequences.

Generally protein, DNA or RNA.

In many cases, the input set of query sequences are assumed to have an evolutionary relationship.

By which they share a lineage and are descended from a common ancestor.

Compare all sequence pairwise

Perform cluster analysis on the pairwise data to generate a hierarchy for alignment.

This may be in the form of a binary tree or a simple ordering.

Build the multiple Alignment by first aligning the most similar pair of sequence.

Then the next most similar pair and so on

Date
Page

Once an alignment of two sequence has been made, then this is fixed.

Thus for a set of sequences A, B, C, D having aligned.

A with C and B with D alignment of A, B, C, D is obtained by comparing the alignments of A and C with that of B and D using averaged scores at each aligned position.

eg— VTISCTGSSNAIG-NHVKW YQQLPL
VTISCTGSSINGS-ITVNW YQLPG

Application of MSA :—

Detecting similarities b/n sequences (closely or distantly related.)

Detecting conserved regions or motifs in sequences.

Detection of structural Homologies.

Thus, assisting the improved prediction of secondary and tertiary structures of proteins.

Making pattern used to predict new pattern sequences falling in a given family.

Inferring evolutionary trees or linkages.

Progressive Alignment Method

The most widely used approach to multiple sequence alignments.

Also known as the Hierarchical or Tree method.

Progressive Alignment Algorithms.

(i) Clustal W

(ii) T-Coffee

Iterative Refinment method! —

A set of methods to produce MSA while reducing the errors inherent in progressive methods.

TREE Alignment! —

In computational phylogenetic, it is used to analyse a set of sequences with evolutionary relationship using a fixed tree.

Essentially, Tree alignment is an algorithm for optimizing phylogenetic Tree.

STAR Alignment

Another form of Tree alignment.

Thank you

FASTA

Date .....
Page .....

FASTA formats allows the sequence naming and comments to introduce the sequences. The use of FASTA format has become a standard for biologist to analyze the sequencing.

The formats of FASTA codes is no longer than 120 characters.

Features of FASTA

Rather than trying to find out the best alignment between your data, it finds the patches of regional similarity.

It is rapid program - You can run the program locally or you can also send queries to an email server.

The alignments of FASTA can contain gaps. The sequence which contain the gap FASTA highlight these codes with red color.

Another features of FASTA it ignores the complete sensitivity and provide information about the expected matched alignments.

Date
Page

USES OF FASTA

FASTA can be used for the Alignment of all types of proteins and DNA.

It can also use for the translation of algorithms which handle frame shift errors.

It can used for calculating the similarity which can help biologists to decide whether the alignment is occured by chance or it is due to infer homology.

You can also use FASTA for calculating the optimal score for alignment.

Help in identify the members of gene family.

DATA STRUCTURE

Data in FASTA is presented in a singal code sequences. It has got a different search methods which help in searching the proteins.
eg — Smith - Waterman type of algorithm FASTA help you to find out the potential matches and save your time as well.

Date
Page

While the results of FASTA are reported in the form of histogram where the expected value are compared to random search set. While the lower part of the histogram contain information about the matches of intrest.

DATA SUBMISSION

NCBI, Bethesda Maryland.

Info

Houses series of databases relevant to biotechnology and biomedicine.

Mainly genbank for DNA and pubmed, a bibliographic database for biomedical literature epigenomics database.

Director: David Lipman

founded in 1988 by claude peppes.

NCBI

To submit a sequence in NCBI we need certain tools, which are easily found in NCBI page itself.

Database

Genbank

Sequence Read Archive

dbSNP (single nucleotide polymorphism)

dbVar (genome variant)

Geo (Gene expression omnibus)

BankIt

We have a single sequence, a simple set of sequences. (e.g - 16S rRNA, matK, ITS/tRNA) or a small batch of different sequences.

We prefer to use a web-based submission tool.

Sequin

We prefer to work on our submission off-line.

We have a sequence or sequence that are complex.

We would like graphical viewing and editing options, including an alignment editor.

Network access analysis tool.

Date .....
Page .....

Submission follow :—

Contact Information

Release date information.

Reference information

Submission category and type

Nucleotide sequence

Organisms name

Feature of the sequence

Sequin (Registration)

(I) Author form
(II) submission page
(III) Contact page
(IV) Authors page
(V) Affiliation Page
(VI) Sequence format form
(VII) submission type
(VIII) Seq. Data format
(IX) Submission category
(X) Organisms and sequence form
(XI) Nucleotide page.

The end

Genome Annotation
pattern and Repeat finding :-

Also called as DNA Annotation.

After DNA sequence DNA annotation is done. The Dna sequence will not make any sense without Annotation in this :-

- # Location of genes are identified.
- # All coding Regcoding Regions of the Gene are identified.
- # Start and stop point of the gene are identified.
- # Function of the Gene is identified.

Annotation Means :-

Extra information associated with any document is called as it annotation.

Includes comment or explanation.

Highlighting of key words, Underlining certain lines in your answers are examples of annotation.

Date .....
Page .....

Types of Annotation

Structural Annotation

↓

Identification of Genomic Elements

↓

(I) ORF's

(II) Gene Structur

(III) Location of Regulatory Regions / Motifs

(IV) Coding Regions

Functional Annotation

↓

Attaching Biological Information to the Genomic Elements

↓

(I) Biochemical function

(II) Biological function

(III) Expression.

Gene Structural Annotation tool:—

Repeat Masker is a program that screens DNA sequence for interspersed repeats and low complexity DNA sequences.

The output of the program is a detailed annotation of the repeats that are present in the query sequences as well as a annotated repeats have been masked.

Identification of Non-coding Regions :—

Identify the start and stop codon, We know that start codon is present at the start of the gene's coding region. This codon is generally 'AUG' (Methionine). Other start codons are also known. Generally the start region is 20 bp in length. When we compare various start regions, we can easily identify the region b/n coding (Exon) and non-coding region (intron).

Same is for stop codon. The stop codons are present at the end of the gene. This Regions is rich in thymine and is about 20-30 bp in length.

Gene finding :—

(1) Done with the help of Gene Knockout experiments. The gene of interests is knocked out from the animal body. Then the animal is made to grow in the laboratory, phenotypic changes are observed.

Date
Page

BLAST, FASTA

Find out Regions similarity between the given of the sequences. If a Regions has very high similarity to mRNA or Protein product, then it means that this Regions is a protein coding Gene.

Functional Annotation:—

Similarity functional Annotation has been useful in various files of Life science includes:

Cancer cell profiling.
Study of Complex disorder like Alzeimes disease, Schizophernia etc.

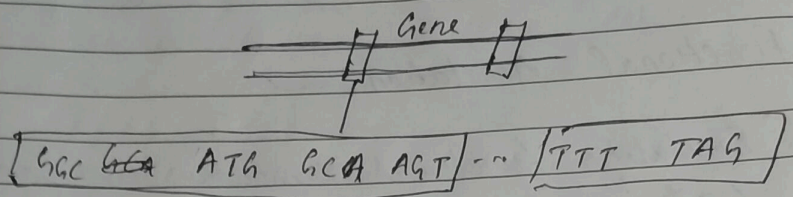
Gynaecological studies like preterm birth.

Genome wide association studies
Brain related disorderss like depression, Anxiety etc.

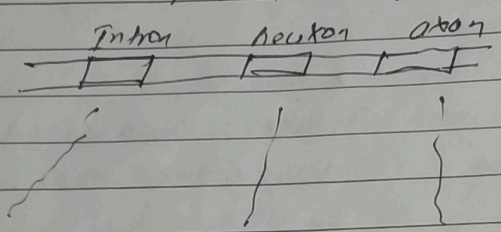
It is important in area of study genomics & proteomics.

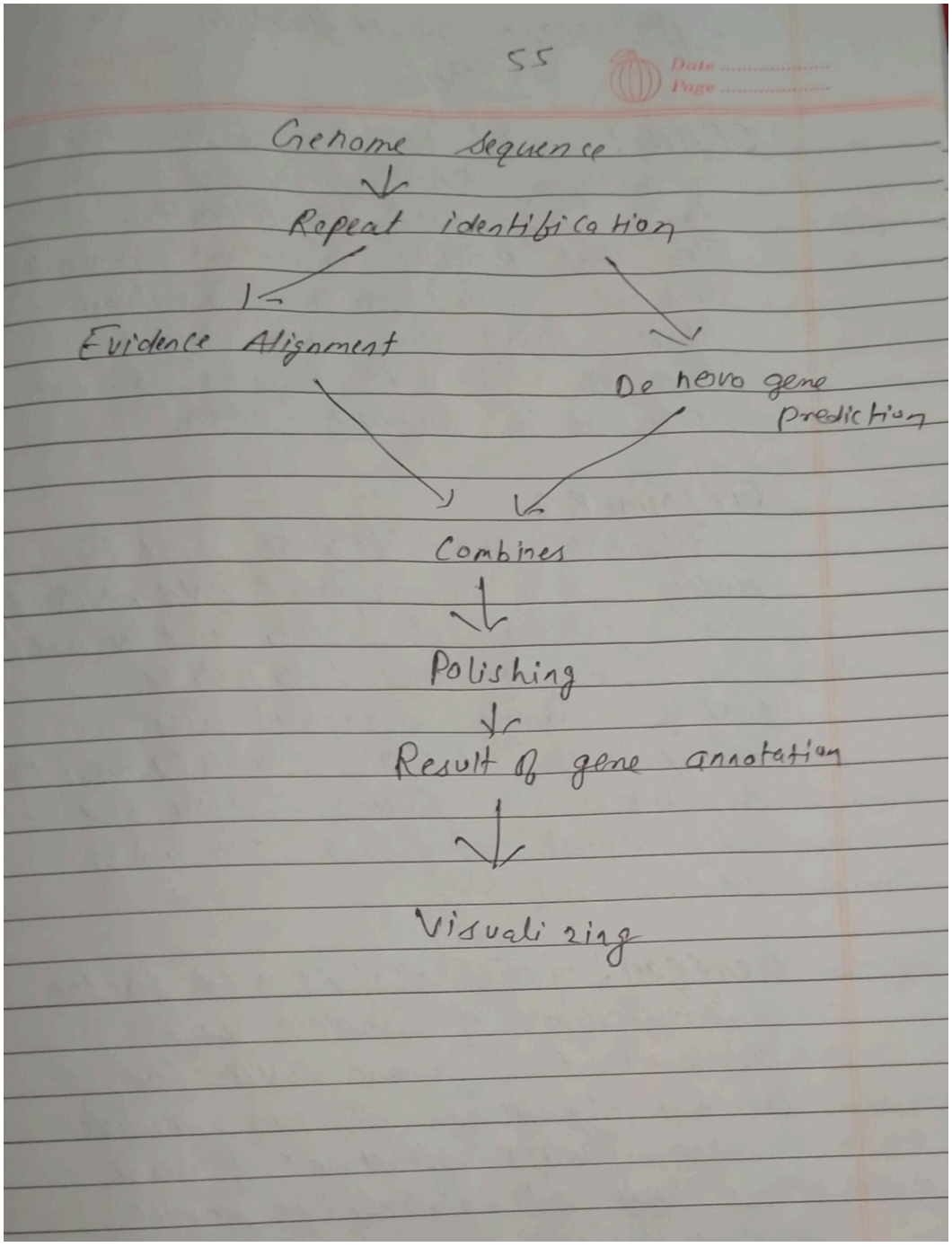
Useful annotation

(I) Automated annotation pipelines
(II) Human Genome project
(III) TIGR



Open Reading frame




Genome sequence
Repeat identification
Evidence Alignment
De novo gene prediction
Combines
Polishing
Result of gene annotation
Visualizing

Gene identification Tools

CRAIL :- It is one of the most Commonly Known Computational tools for ORF identification. This tool provides important information such as splice junctions, translation start points and non-coding scores of 60 base regions on both sides of the putative exon.

GLIMMER :-
Glimmer is a software used for finding genes in microbial DNA, especially the genomes of bacteria and archaea. Gene Locator and Interpolated markov modeler (Glimmer) uses interpolated markov models (IMMs) to recognize the coding regions and differentiate them from non-coding DNA.

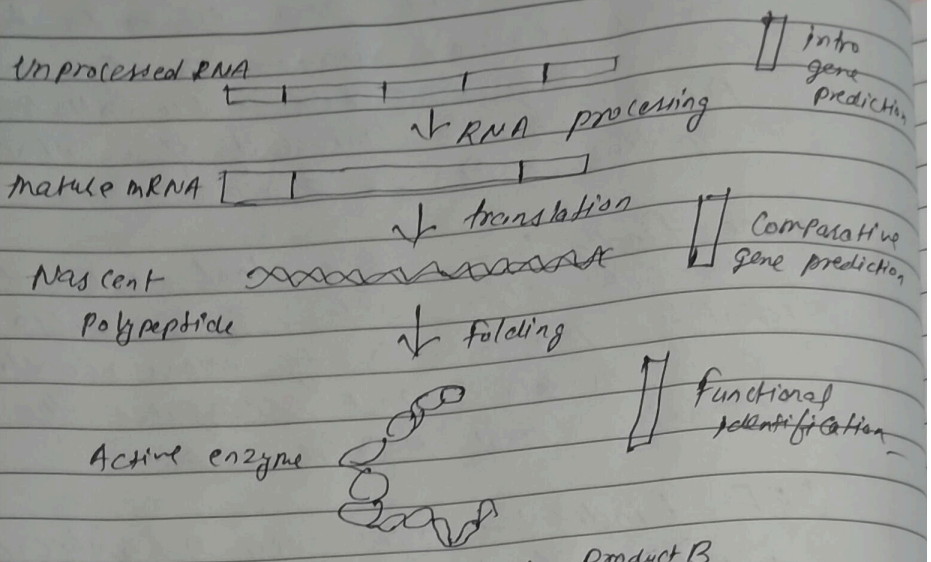
Genscan :- This tool is used for the identification of Complete gene structures in genomic DNA for various organisms. It can predict exon - intron structures of genes as well as locations in genomic sequences.

Genie :— This gene finder is based on generalized hidden Markov models. Genie was developed as a collaborative project by the University of California. Computional Biology group.

Gene finder :— This tool is used to predict slice sites. It can also identify protein-coding exons, construct gene models, and recognize the promoter and poly-A region.

ORF finder :— This is a graphical analysis tool that can detect open reading frames along with their protein translation from sequence in already in the database. This program is used to search new DNA sequences for potential protein-encoding segments.

Easy Gene :— This tool is used to identify genes in prokaryotes, the current version of which includes 128 different organisms. Each gene identified by easy gene is attributed with a significant score (P-value), which reveals the probability of a sequence to be a non-coding open reading frame rather than a real gene.


Unprocessed RNA
RNA processing
Mature mRNA
translation
Nascent Polypeptide
Folding
Active enzyme
Function
Reactant A → Product B
intro gene prediction
Comparative gene prediction
Functional identification

Thank you

- Author - Amit Kumar Raaj
- Published - Biotechnologyme.blogspt.com
- Subject - Bioinformatics Notes