Elements of Bioinformatics and its applications in medicine

M. Kamran Azim, Ph.D.
H.E.J. Research Institute of Chemistry,
International Center for Chemical and Biological Sciences,
University of Karachi,
Karachi

1

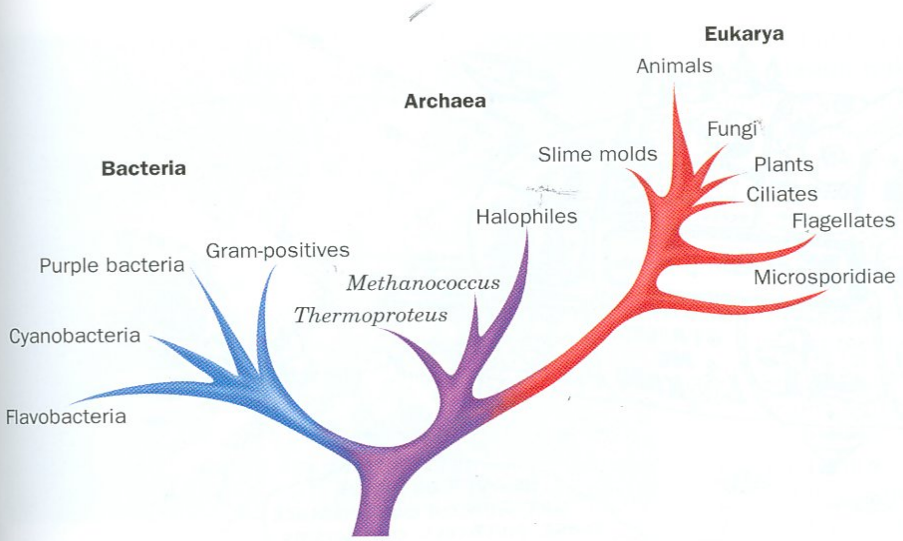
Outline

- Part-I: Overview of Molecular Biology
- Part-II: Biological Sequence Databases
- Part-III: Principles and Methods of Sequence Analyses
- Part-IV: Sequence Database Searching
- Part-V: Bioinformatics in Molecular Medicine

Part-I

Overview of molecular biology

Complexity and Diversity in Life



4

Molecular basis of life

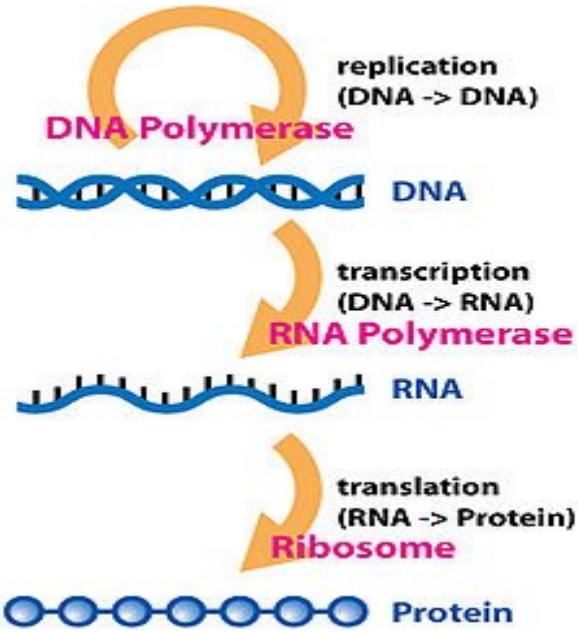
Living organisms-diversity

Molecular Biology brings Uniformity in life

From bacteria to human molecular biology is same

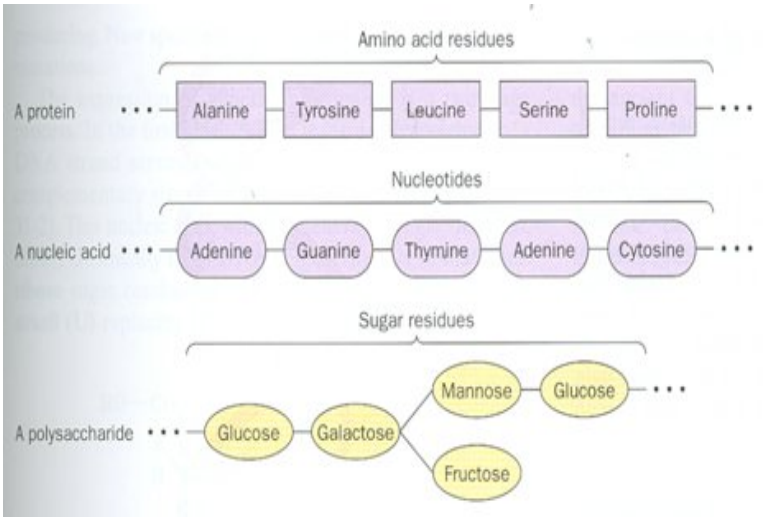
5

Central dogma of molecular biology



6

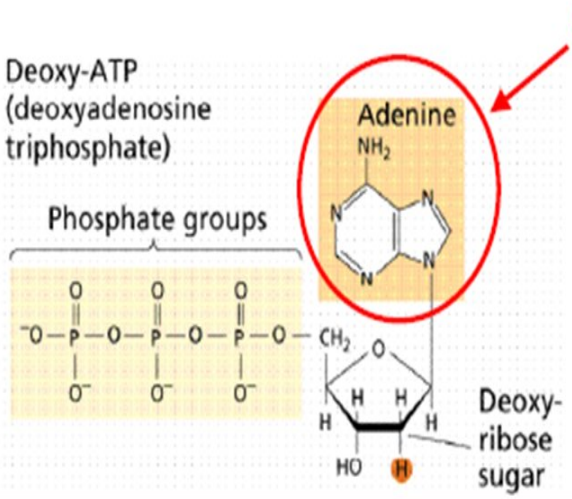
Polymeric organization in bio-macromolecules



7

Nucleic acids (DNA/RNA): Polymer of nucleotides

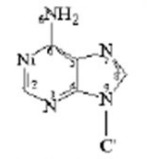
- **Nucleotide = deoxyribose sugar + phosphate group + base**



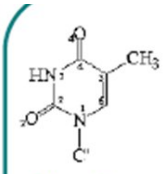
8

Four types of nucleotides due to difference in 'bases'

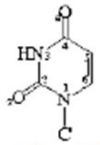
- **DNA**
  - A = Adenine
  - T = Thymine
  - C = Cytosine
  - G = Guanine
- **RNA**
  - A = Adenine
  - U = Uracil
  - C = Cytosine
  - G = Guanine



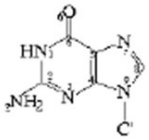
Adenine



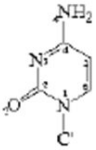
Thymine



Uracil



Guanine



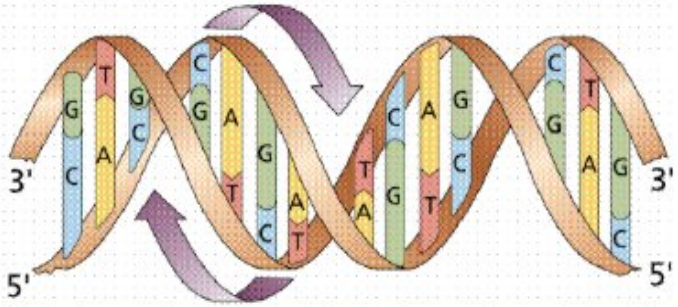
Cytosine

Purines

Pyrimidines

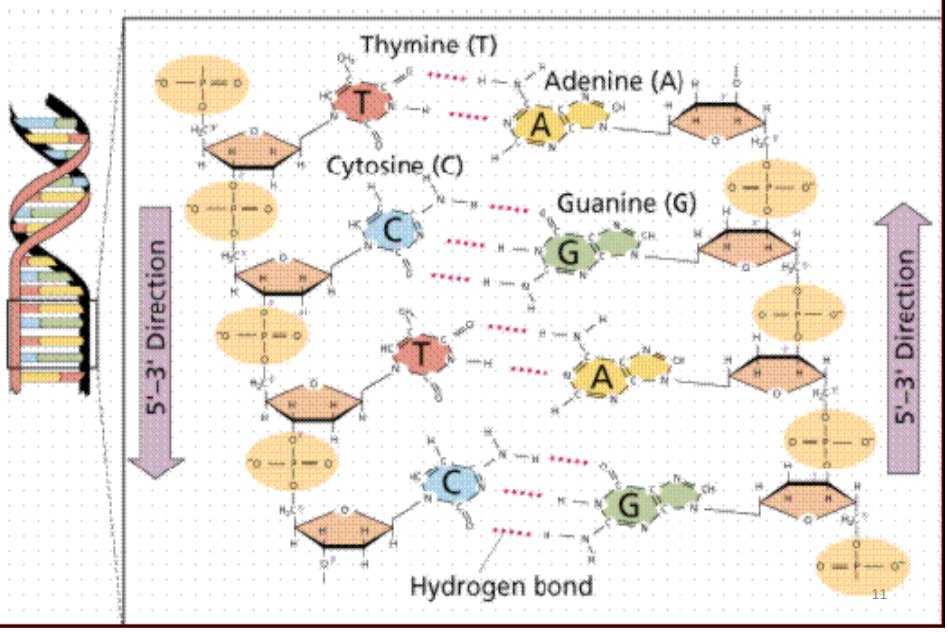
DNA

- **Physical structure**
  - Double (stranded) helix
  - Sugar & phosphate groups form backbone
  - Complementary bases (A-T, C-G) connected by hydrogen bond
  - 5' = end w/ free phosphate group
  - 3' = end w/ free oxygen group



10

Views of DNA Structure

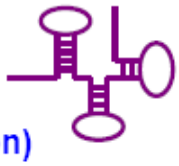


DNA

- **For bioinformatics**
  - DNA can be represented as a sequence of letters (A,C,G,T)
  - 5' A T A C G T A 3'
    3' T A T G C A T 5' (matching strand, redundant)
- **Terms**
  - Base pair (bp) – one pair of DNA bases (1 letter)
  - Gene – section of DNA that produces a functional product
  - Chromosome – physical linear sequence of DNA
  - Genome – entire collection of DNA for an organism
    - E Coli 1 chromosome 5×10^6 bases (5 Mbps)
    - Drosophila 8 chromosomes 2×10^8 bases (200 Mbps)
    - Human 46 chromosomes 3×10^9 bases (3 Gbps)

Ribonucleic acid (RNA)

- **Composition**
  - Sequence of nucleotides
  - Nucleotide = ribose sugar + phosphate group + base
  - Single stranded (but may form hairpin loops)
  - Uracil (U) instead of Thymine (T)



- **DNA → RNA (Transcription / Gene Expression)**
  - RNA polymerase (enzyme)
    1. Finds gene initiation marker (codon) on DNA strand
    2. Reads DNA strand containing marker
    3. Builds (complementary) strand of messenger RNA (mRNA)
    4. Stops when gene end marker (codon) found
  - Resulting RNA sequence = transcript

13

DNA for Information

Protein for Execution

14

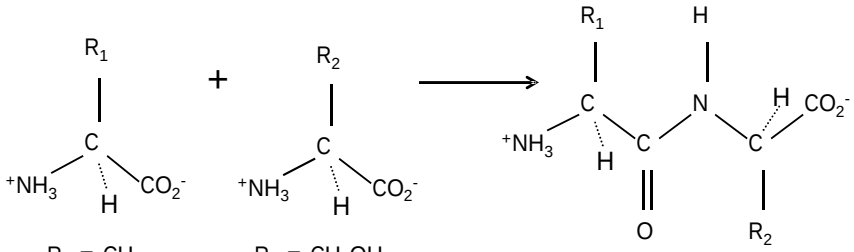
Proteins and the molecular basis of life

- Proteins have dynamic and diverse role(s)
  - Catalysing biochemical reactions (Enzymes; Regulators)
  - Formimg receptors and channels in cell membranes (Membrane proteins)
  - Providing intra/extracellular scaffolding support (Cytoskeleton protein)
  - Transporting molecules (Transport proteins)
  - Hormones
  - more

15

Proteins are the Polymers of amino acids

20 types amino acids in proteins; protein's length in the range of 50-2000 Amino acids; ~40,000 different proteins in human; millions of known proteins from other organisms



$R_1 = CH_3$
Alanine
Ala
A

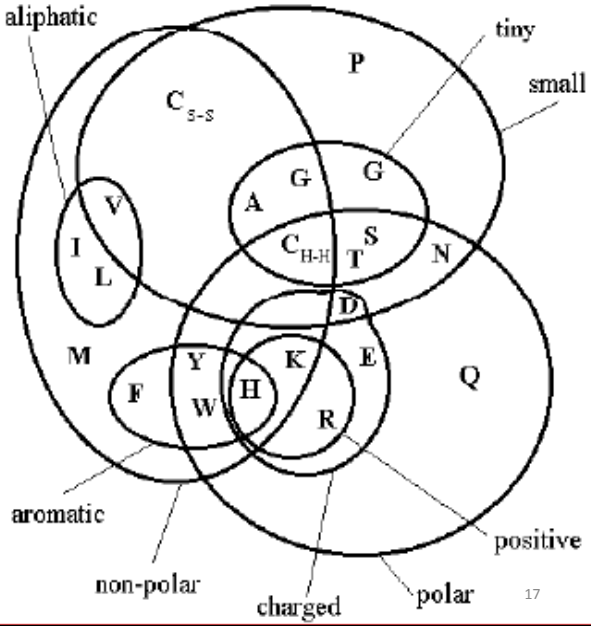
$R_2 = CH_3OH$
Serine
Ser
S

AlanylSerine
Ala-Ser
AS

16

Amino Acid Properties

- Amino acids have different physical & chemical properties
- Different amino acids are similar in different ways



17

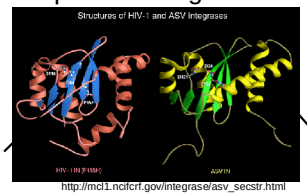
Protein Structure

primary structure – amino acid sequence

DRLEFIVTALLKPW

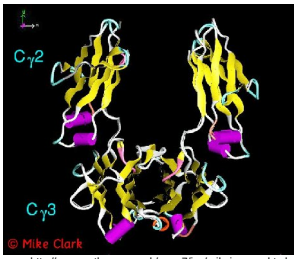
N-terminus C-terminus

secondary structure – local spatial arrangement



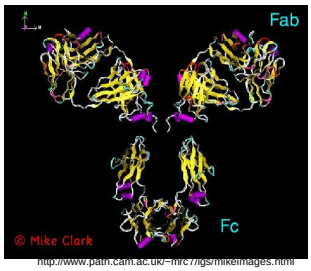
http://mcl1.ncifcrf.gov/integrase/asv_secstr.html

tertiary structure – protein folding



http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html

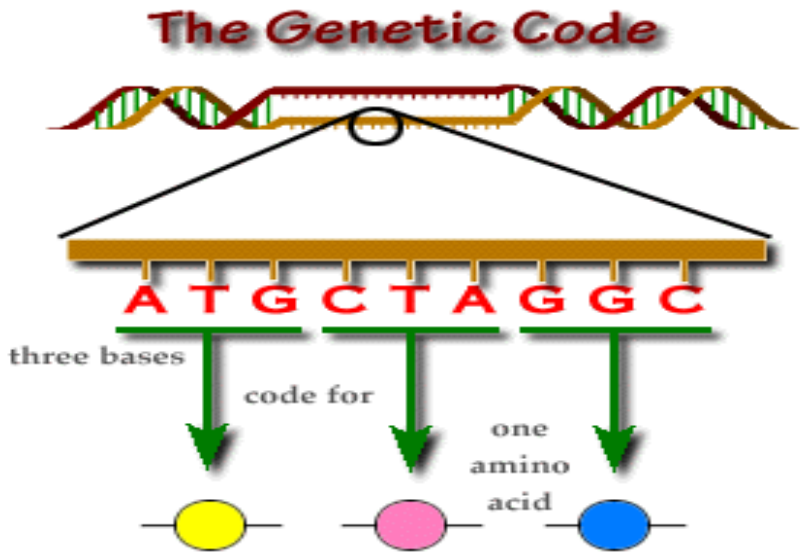
quaternary structure – multimeric complexes



http://www.path.cam.ac.uk/~mrc7/igs/mikeimages.html

18

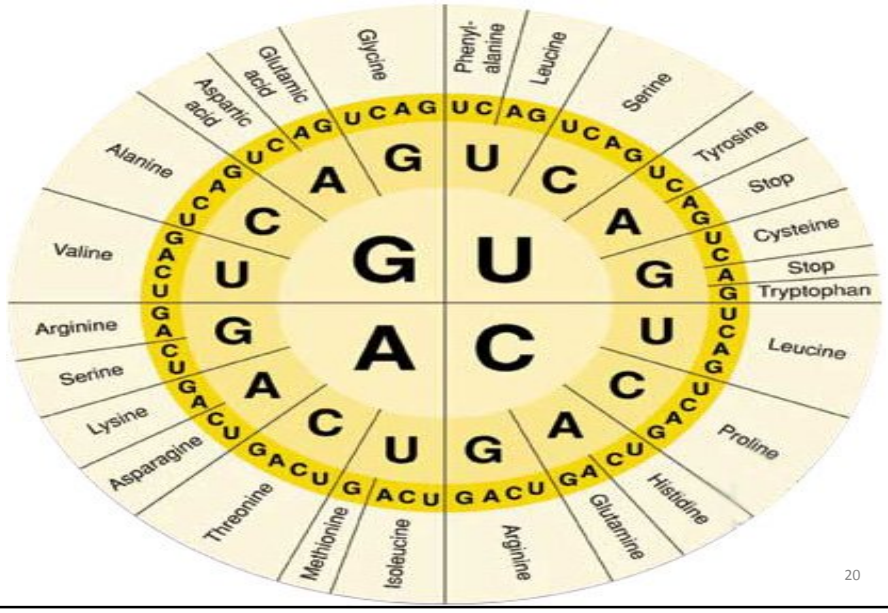
The correlation between Nucleic acids and protein sequences?



19

Genetic Code

Which nucleotide triplet codes for which amino acid



20

Coding and non-coding DNA

- DNA codes for proteins.
- Are all DNA sequences code for proteins?
- No. Only 2-3% of human DNA code for protein called as coding DNA or gene (coding regions).
- The coding region starts with the ATG (AUG) codon and ends with any of the stop codon.
- Rest is non-coding! Function(s) of major part of non-coding regions not known.

21

Mutations: Changes in DNA sequences

Important event during life processes
major cause of diversity; key reason for many diseases

- **Mutations**
  - Modifications during DNA replication
- **Possible changes**
  - Point mutation / single nucleotide polymorphism (SNP)
    - 5' A T A C G T A ...
    - 5' A T G C G T A ...
  - Duplicate sequence
  - Inverted sequence
  - Insert / delete sequence (indel)

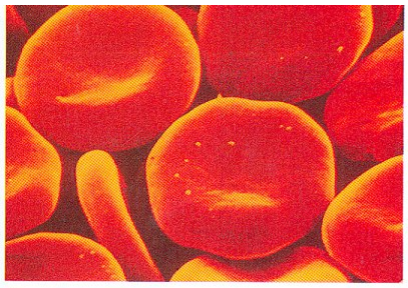
22

Case study: Disease due to mutation in the DNA
Sickle cell anemia

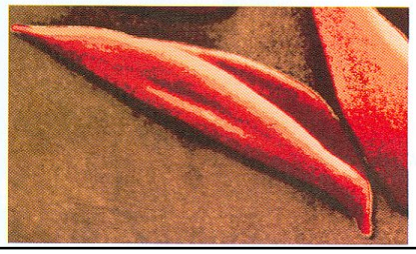
- In Sickle cell anemia, red blood cells assume an irregular crescent-like shape which increases the RBC's rigidity.
- Result in blockage of blood flow in capillaries and red cell destruction leading to anemia (deficiency of hemoglobin).
- Hemoglobin, the red blood pigment, is a protein responsible for transport oxygen throughout the body.
- A molecule of hemoglobin composed of four protein chains called globin.
- Main cause of Sickle cell anemia, is a mutation in the hemoglobin gene resulting in defective hemoglobin with single amino acid mutation (Glu6→Val).

23

Normal red cells versus Sickle cells



(a)



24

Part-II

Biological Sequence Databases

What is Bioinformatics?

Applying Mathematics, Statistics, Computer Science, Information Technology for solving Biological problems

Bioinformatics is the science of storing, retrieving and analyzing large amounts of biological information.

It cuts across many disciplines, including biology, computer science and mathematics.

26

Applications of Bioinformatics

Molecular basis of pathogenicity;
e.g. Amyloid protein in neurodegenerative diseases
Novel targets of therapeutic intervention;
e.g. Caspase inhibitors in diseases characterized by tissue degradation
Molecular Diagnostics;
e.g. Bird Flu
Host-pathogen interaction;
e.g. Bacterial adherence factors
Novel Research tools;
e.g. GFP-based techniques

27

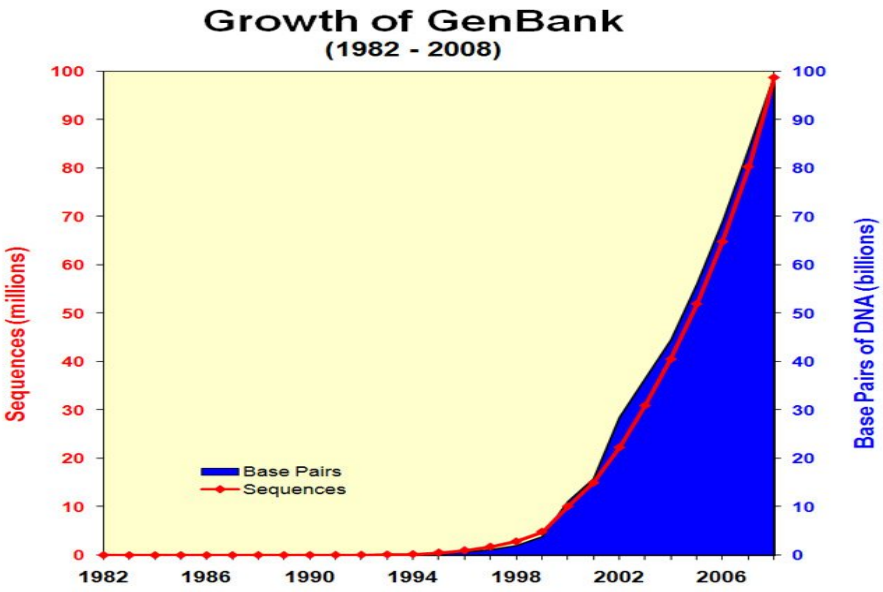
Why Bioinformatics?

- Large-scale DNA/genome sequencing projects (in particlular Human genome project) have led to an explosion of information concerning the DNA and protein sequence data.
- Development in the field of computer technology including the use of computerized databases for storing, retrieving and comparing sequences; computer graphics for displaying and manipulating three-dimensional structures.

28

The explosion in sequence information

billions of bases from over 100,000 species



Bioinformatics in Urdu poetry

زندگی کیا ہے! عناصر میں ظہورِ ترتیب

موت کیا ہے! انہی اجزاء کا پریشاں ہونا

Bioinformatics as the Science of Sequence

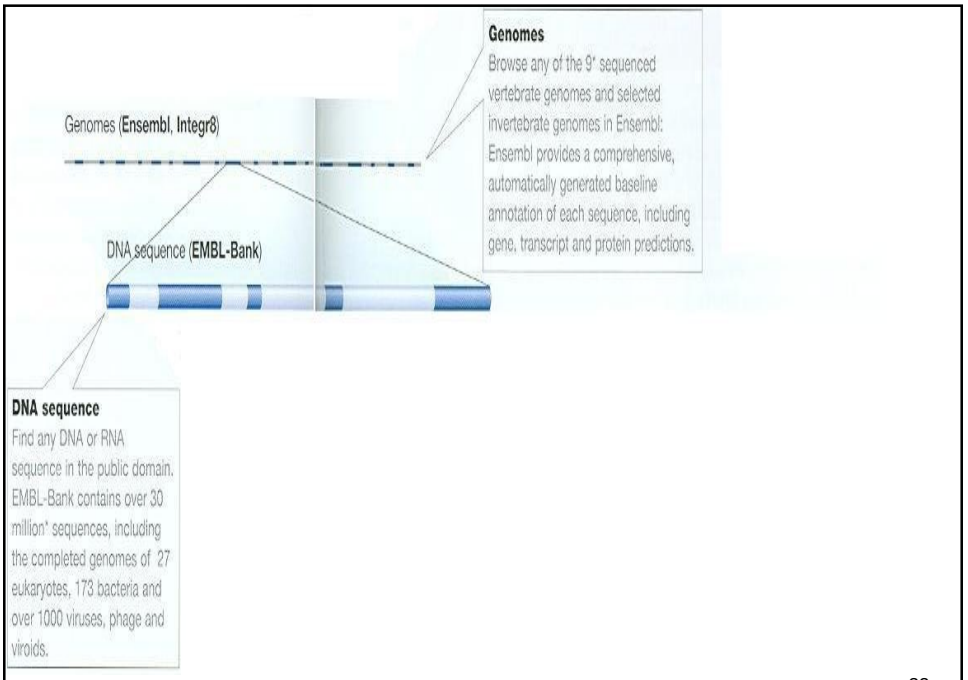
Sequence is the name of the game

30

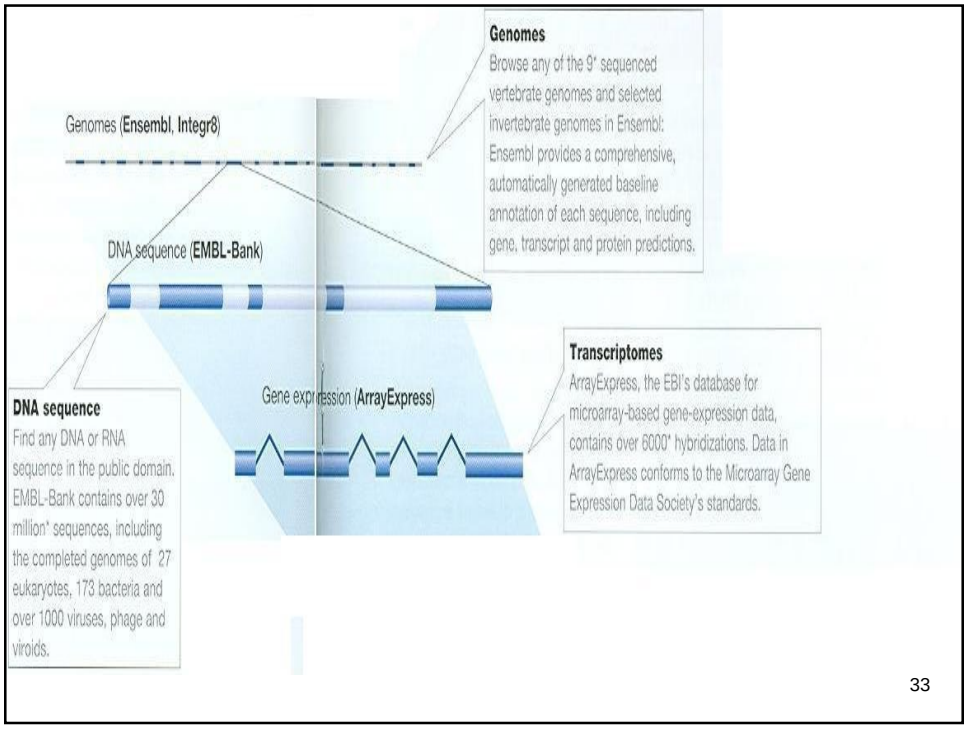
Prof. Zafar H. Zaidi and Bioinformatics at Karachi University

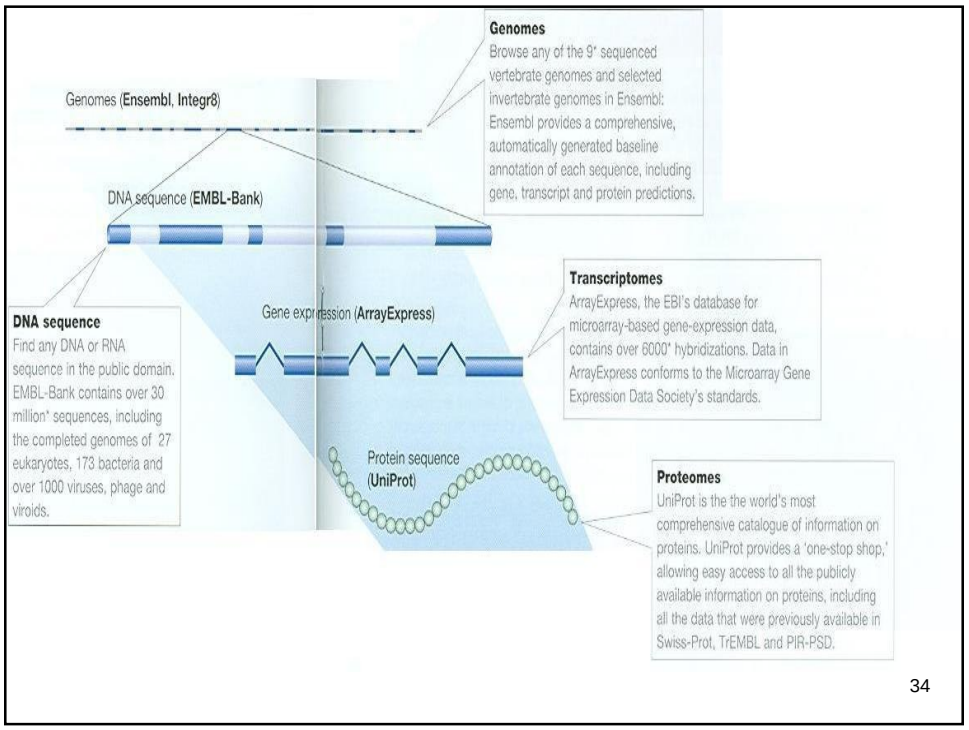
PROFESSOR SYED ZAFAR HASNAIN ZAIDI
(1939 - 2001)

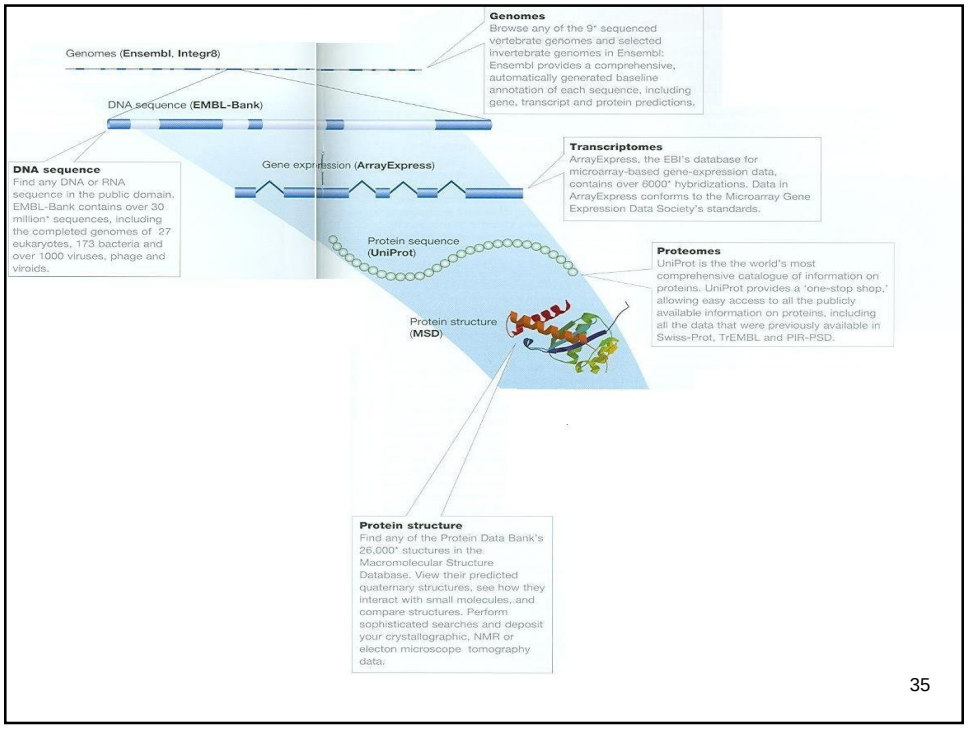
- Pioneered Protein Chemistry (Protein Sequencing; 1975-2001)
- Initiated Bioinformatics (Sequence Analysis; 1991-2001)
- First paper published structural Bioinformatics in 1994



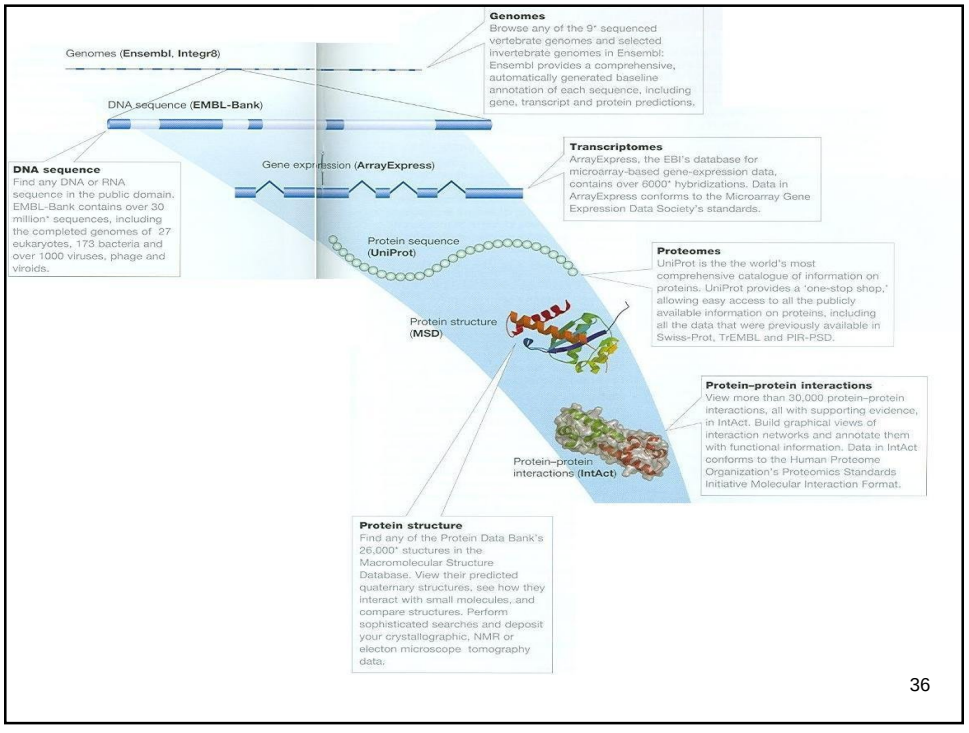
32


Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Gene expression (ArrayExpress)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
33

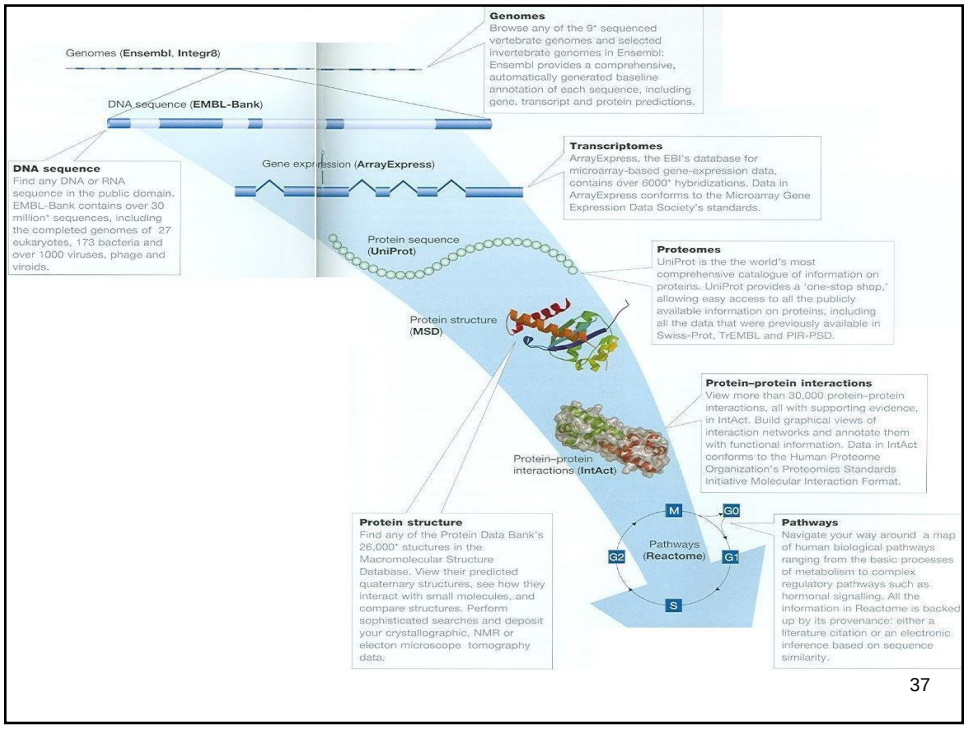

Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Gene expression (ArrayExpress)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
Protein sequence (UniProt)
Proteomes
UniProt is the the world's most comprehensive catalogue of information on proteins. UniProt provides a 'one-stop shop,' allowing easy access to all the publicly available information on proteins, including all the data that were previously available in Swiss-Prot, TrEMBL and PIR-PSD.
34


Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
Gene expression (ArrayExpress)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Protein sequence (UniProt)
Proteomes
UniProt is the the world's most comprehensive catalogue of information on proteins. UniProt provides a 'one-stop shop,' allowing easy access to all the publicly available information on proteins, including all the data that were previously available in Swiss-Prot, TrEMBL and PIR-PSD.
Protein structure (MSD)
Protein structure
Find any of the Protein Data Bank's 26,000* stuctures in the Macromolecular Structure Database. View their predicted quaternary structures, see how they interact with small molecules, and compare structures. Perform sophisticated searches and deposit your crystallographic, NMR or electon microscope tomography data.

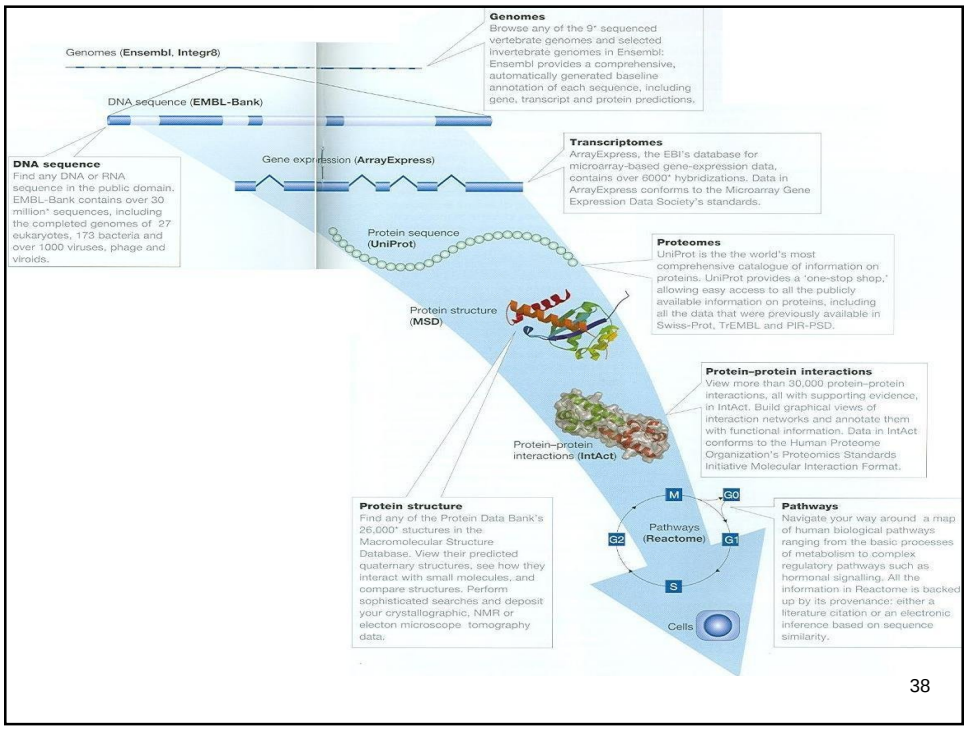
35


Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
Gene expression (ArrayExpress)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Protein sequence (UniProt)
Proteomes
UniProt is the the world's most comprehensive catalogue of information on proteins. UniProt provides a 'one-stop shop,' allowing easy access to all the publicly available information on proteins, including all the data that were previously available in Swiss-Prot, TrEMBL and PIR-PSD.
Protein structure (MSD)
Protein–protein interactions
View more than 30,000 protein–protein interactions, all with supporting evidence, in IntAct. Build graphical views of interaction networks and annotate them with functional information. Data in IntAct conforms to the Human Proteome Organization's Proteomics Standards Initiative Molecular Interaction Format.
Protein–protein interactions (IntAct)
Protein structure
Find any of the Protein Data Bank's 26,000* stuctures in the Macromolecular Structure Database. View their predicted quaternary structures, see how they interact with small molecules, and compare structures. Perform sophisticated searches and deposit your crystallographic, NMR or electon microscope tomography data.

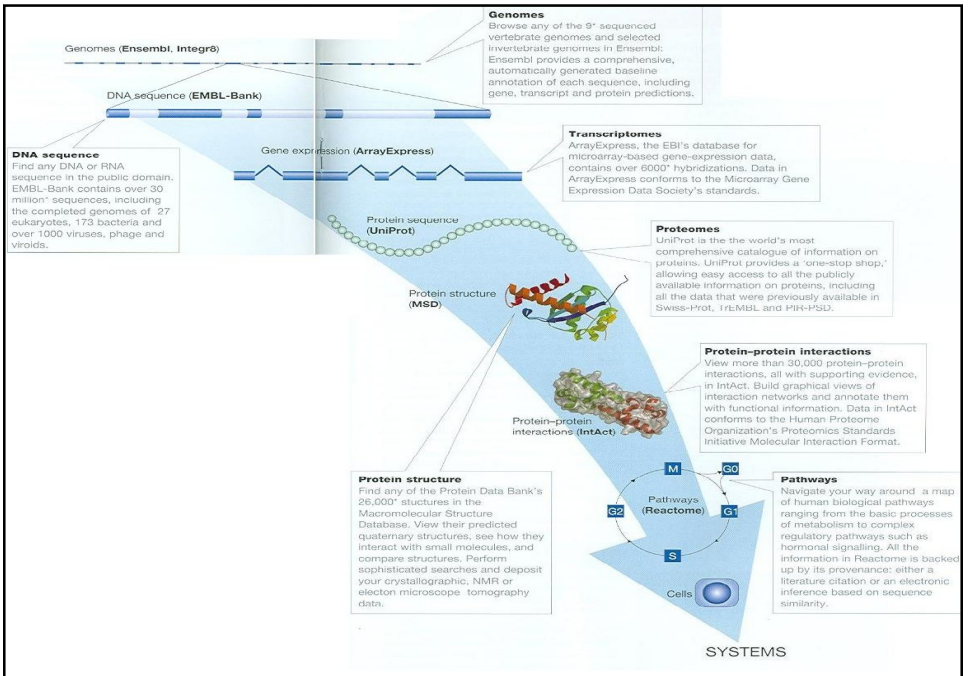
36


Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
Gene expression (ArrayExpress)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Protein sequence (UniProt)
Proteomes
UniProt is the the world's most comprehensive catalogue of information on proteins. UniProt provides a 'one-stop shop,' allowing easy access to all the publicly available information on proteins, including all the data that were previously available in Swiss-Prot, TrEMBL and PIR-PSD.
Protein structure (MSD)
Protein–protein interactions
View more than 30,000 protein–protein interactions, all with supporting evidence, in IntAct. Build graphical views of interaction networks and annotate them with functional information. Data in IntAct conforms to the Human Proteome Organization's Proteomics Standards Initiative Molecular Interaction Format.
Protein–protein interactions (IntAct)
Protein structure
Find any of the Protein Data Bank's 26,000* stuctures in the Macromolecular Structure Database. View their predicted quaternary structures, see how they interact with small molecules, and compare structures. Perform sophisticated searches and deposit your crystallographic, NMR or electon microscope tomography data.
M
G0
G2
Pathways (Reactome)
G1
S
Pathways
Navigate your way around a map of human biological pathways ranging from the basic processes of metabolism to complex regulatory pathways such as hormonal signalling. All the information in Reactome is backed up by its provenance: either a literature citation or an electronic inference based on sequence similarity.

37


Genomes
Browse any of the 9* sequenced vertebrate genomes and selected invertebrate genomes in Ensembl: Ensembl provides a comprehensive, automatically generated baseline annotation of each sequence, including gene, transcript and protein predictions.
Genomes (Ensembl, Integr8)
DNA sequence (EMBL-Bank)
DNA sequence
Find any DNA or RNA sequence in the public domain. EMBL-Bank contains over 30 million* sequences, including the completed genomes of 27 eukaryotes, 173 bacteria and over 1000 viruses, phage and viroids.
Gene expression (ArrayExpress)
Transcriptomes
ArrayExpress, the EBI's database for microarray-based gene-expression data, contains over 6000* hybridizations. Data in ArrayExpress conforms to the Microarray Gene Expression Data Society's standards.
Protein sequence (UniProt)
Proteomes
UniProt is the the world's most comprehensive catalogue of information on proteins. UniProt provides a 'one-stop shop,' allowing easy access to all the publicly available information on proteins, including all the data that were previously available in Swiss-Prot, TrEMBL and PIR-PSD.
Protein structure (MSD)
Protein–protein interactions
View more than 30,000 protein–protein interactions, all with supporting evidence, in IntAct. Build graphical views of interaction networks and annotate them with functional information. Data in IntAct conforms to the Human Proteome Organization's Proteomics Standards Initiative Molecular Interaction Format.
Protein–protein interactions (IntAct)
Protein structure
Find any of the Protein Data Bank's 26,000* stuctures in the Macromolecular Structure Database. View their predicted quaternary structures, see how they interact with small molecules, and compare structures. Perform sophisticated searches and deposit your crystallographic, NMR or electon microscope tomography data.
M
G0
G2
Pathways (Reactome)
G1
S
Cells
Pathways
Navigate your way around a map of human biological pathways ranging from the basic processes of metabolism to complex regulatory pathways such as hormonal signalling. All the information in Reactome is backed up by its provenance: either a literature citation or an electronic inference based on sequence similarity.

38



39

Scope of topics

- Biological databases (utilization, development and integration etc.)
- Analyses of nucleotide and protein sequence information
- Analyses of 3D structural data of macromolecules.
- Assessment of how small molecules interact with macromolecules in biological systems.
- Studies on networks of protein-protein interactions
- Simulation of biological processes
- More

40

Scope of topics

- Biological databases (utilization, development and integration etc.)
- Analyses of nucleotide and protein sequence information
- Analyses of 3D structural data of macromolecules.
- Assessment of how small molecules interact with macromolecules in biological systems.
- Studies on networks of protein-protein interactions
- Simulation of biological processes
- More

41

Bioinformatics Resources

Sequence Databases

- 1960s; The first sequences to be collected were those of proteins by Margaret Dayhoff at the NBRF, Washington, USA.

  [Protein sequence atlas; PIR]
- 1970s; First DNA sequences databases were

  (a) the GenBank at Los Alamos National Labotaroy, New Maxico, USA

  (b) EMBL at the European Molecular Biology Laboratory at Heidelberg, Germany.

42

Primary Bioinformatics Databases

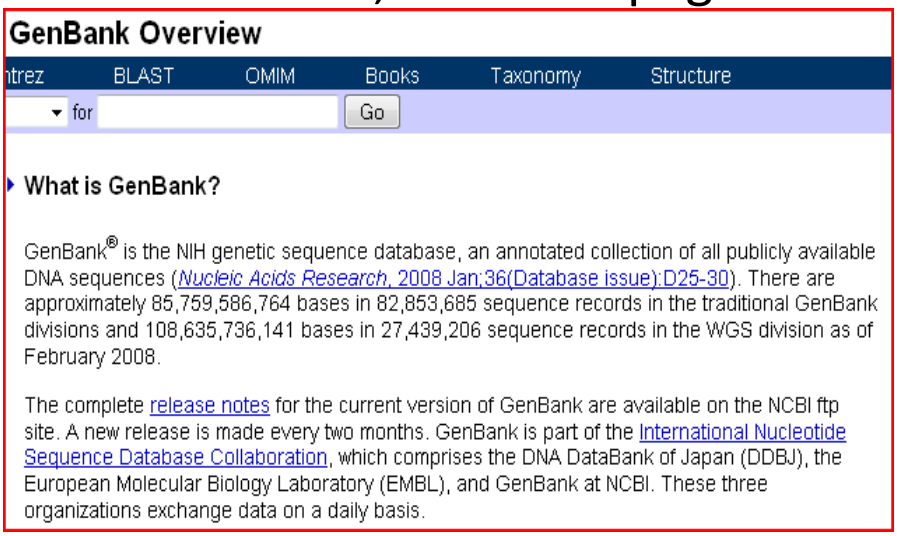
- Nucleotide (DNA) sequence databases
  GenBank, EMBL and DDBJ
- Protein sequence Databases
  SwissProt, PIR, UniProt
- Protein 3D structure databases
  PDB, SCOP, CATH
- Genome Centers databases
  Sanger Center, TIGR
- Specialized databases
  MEROPS, Protein Kinase Resource

43

Nucleotide Sequence Databases

- 3 major nucleotide sequence databases in INSC
  - Genbank at NCBI, Bethesda, Maryland, USA
  - EMBL at EBI, Hinxton, UK
  - DDBJ at NIG, Mishima, Japan
- Curators of these databases have made a consortium termed as International Nucleotide Sequence Collaboration (INSC)

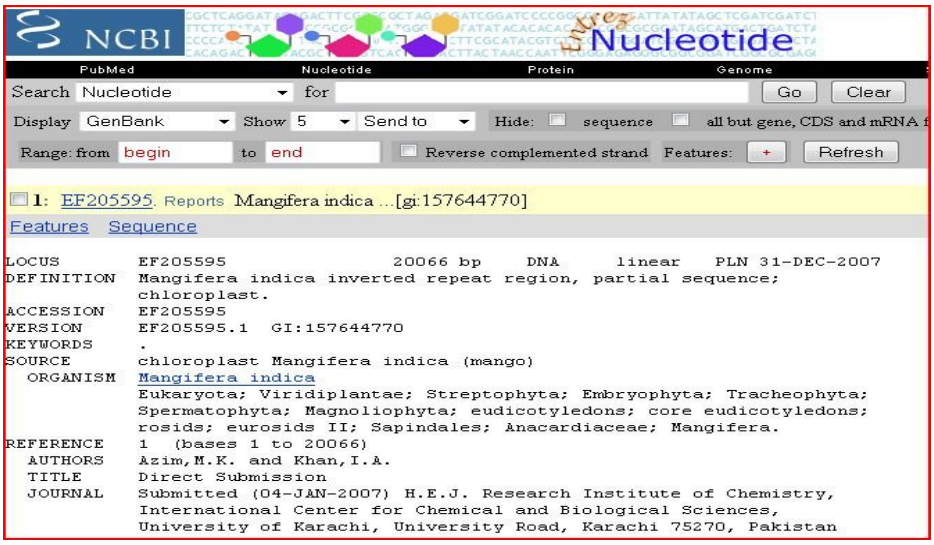
GenBank; the main page



Sequence Entry; The flatfile (3 major parts)

- These databases contain millions of 'entries'; each entry has nucleotide sequence of a portion of DNA.
- The sequence entry is a computerized text file called 'flatfile'.
- The flatfile comprises of 3 parts
  - The header (information that apply to the entire record)
  - The feature table (annotations on the record)
  - The nucleotide sequence

The Header



The Feature Table

```
FEATURES             Location/Qualifiers
     source          1..20066
                     /organism="Mangifera indica"
                     /organelle="plastid:chloroplast"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:29780"
     repeat_region   <1..>20066
                     /rpt_type=inverted
     gene            complement(55..1626)
                     /gene="rpl2"
     CDS             complement(join(55..489,1234..1626))
                     /gene="rpl2"
                     /codon_start=1
                     /transl_table=11
                     /product="ribosomal protein L2"
                     /protein_id="ABV59097.1"
                     /db_xref="GI:157644773"
```

The Feature Table (contd.)

```
CDS             complement(join(55..489,1234..1626))
                /gene="rpl2"
                /codon_start=1
                /transl_table=11
                /product="ribosomal protein L2"
                /protein_id="ABV59097.1"
                /db_xref="GI:157644773"
                /translation="MAIHLYKTSTPSTRNGAVDSQVKSNPRNNLIYGQHRCGKGRNAR
                GIITAGHRGGGHKRLYRKIDFRRNEKDIYGRIVTIEYDPNRNAYICLIHYGDGEKRYI
                LHPRGAIIGDTIVSGTEVPIKMGNALPLSTDMPLGTAIHNIEITLGKGGQLARAAGAV
                AKLIAKEGKSATLKLPSGEVRLISKNCSATVGQVGNVGVNQKSLGRAGSKCWLGKRPV
                VRGVVMNPVDHPHGGGEGRAPIGRKRPATPWGYPALGRRSRKRNKYSDNLILRRRSK"
gap             732..831
```

The Nucleotide Sequence

```
ORIGIN
        1 ttttttatc ttttgttttt tgtaaagacg aagaaaaaaa ttcgattttc tctcctattt
       61 actacggcgg cgaagaatca aattatcact atatttattc ctttttctac ttcttcttcc
      121 aagtgcagga taaccccaag gggttgcggg tctttttcta ccaattggag cccttccttc
      181 accaccccca tgggggtggt ctacagggtt cataactact cctcttacta caggacgctt
      241 acctagccaa catttcgatc cggctctacc caaacttttc tggttcaccc caacattccc
      301 cacttgtccg actgttgctg agcagttttt ggatatcaaa cggacctccc cagaaggtaa
      361 ttttaatgtg gccgatttcc cctcttttgc aatcagtttc gctacagcac ccgctgctct
      421 agctaattgt ccacccttte ctagtgtgat ttctatgtta tgtatggccg tgcctaaggg
      481 catatcggtt gaagtagatt cgtctttttg atcaatcaaa accccttccc aaaccgtaca
      541 agcttcttcc aaagcatacg gctttctggg tgtagatgat gatatctata cagatggatc
      601 ttatctatat catataatga agtaccacat gagtggatat ataggaatcc aaatctgccg
```

Submitting Sequences to the Databases

- Sequences can be submitted to the databases through the web or through computer programs (i.e. Sequin at GenBank)
- Web interfaces for sequence submission
  - at DDBJ is Sakura,
  - at EMBL is wEBIn and
  - at GenBank is BankIt

Nucleotide sequence databases

- RefSeq

  Many sequences are represented more than once in Genbank/EMBL/DDBJ (redundancy in sequence entries)

  RefSeq collection is a curated secondary DB

  provides a comprehensive, integrated, nonredundant set of sequences

- Organismal Divisions in databases

  BCT (Bacterial); FUN (fungal); HUM (human); PLN (plant)

- Functional Divisions in databases
  - EST (Expressed Sequence Tags)
  - HTG (High Throughput Genome sequences)
  - WGS (Whole Genome sequences)
  - CON (constructed 'contig' records of chromosomes and other long DNA sequences.

Protein Sequence Databases

Atlas of Protein Sequence and Structure

- The first comprehensive collection of macromolecular sequences in the *Atlas of Protein Sequence and Structure*, published from 1965-1978 under the editorship of Margaret O. Dayhoff.
- Dr. Dayhoff and her research group pioneered in the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences.

PIR-International

- The Protein Information Resource (PIR), located at Georgetown University Medical Center (GUMC), is an integrated public bioinformatics resource to support genomic and proteomic research, and scientific studies.
- PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information.
- For four decades, PIR has provided many protein databases and analysis tools freely accessible to the scientific community, including the Protein Sequence Database (PSD), the first international database (PIR-International), which grew out of *Atlas of Protein Sequence and Structure*.
- Today, PIR offers a wide variety of resources mainly oriented to assist the propagation and standarization of protein annotation:
  - PIRSF (Protein Family Classification System)
  - iProClass (Integrated Protein Knowledegebase)
  - iProlink (Literature, Information & Knowledege)

The SwissProt

- The Swiss-Prot is a manually annotated protein knowledgebase established in 1986 and maintained since 2003 by the UniProt Consortium, a collaboration between the Swiss Institute of Bioinformatics (SIB) and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown University Medical Center's Protein Information Resource (PIR).

UniProt Knowledgebase (UniProtKB)

- The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.
- This includes widely accepted biological information, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.
- The UniProt Knowledgebase consists of two sections:
  - "UniProtKB/Swiss-Prot" (reviewed, manually annotated): A section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis.
  - "UniProtKB/TrEMBL" (unreviewed, automatically annotated): A section with computationally analyzed records that await full manual annotation.

Accessing Bioinformatics Databases

Part-III

Principles and methods of sequence analyses

Scope of topics

- Biological databases (utilization, development and integration etc.)
- Analyses of nucleotide and protein sequence information
- Analyses of 3D structural data of macromolecules.
- Assessment of how small molecules interact with macromolecules in biological systems.
- Studies on networks of protein-protein interactions
- Simulation of biological processes
- More

59

Sequence Analysis and Sequence Analysis Programs

As more DNA sequences became available in the late 1970s, interest also increased in developing computer programs to analyze the sequences.

In early 1980s, the Genetics Computer Group (GCG) was started at the University of Wisconsin, USA, offering a set of programs for sequence analysis.

60

Sequence Analysis and Methods for Comparing Sequences

- The Dot Matrix method (DOTPLOT, COMPARE)
- Dynamic programming matrices
- Word or k-tuple methods (FASTA, BLAST)

61

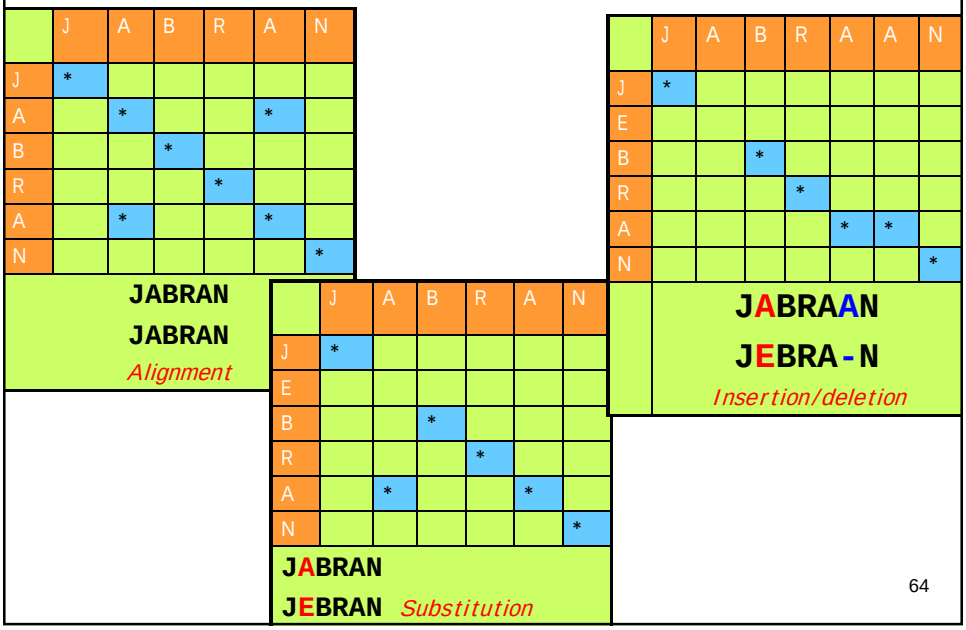
DotPlots

- One of the basic methods for comparing two sequences.
- Quickly identifies
  - regions of local alignment
  - substitutions/insertions/deletions
  - direct or inverted repeats
  - low complexity regions
- Provides easy to understand, intuitive representation of the similarity
- Do not provide statistical measure of the quality of alignment

DotPlots

- For constructing the DotPlots; letters in each sequence are written out across the top and down the side of a square matrix.
- Now, position by positon a dot is placed at each position where there is a match.
- Several publicly available tools for DotPlots;
  - Dotter
  - Dotlet
  - Dottup

Sequence analysis by DotPlots



64

DotDlot analysis; repetitive sequences

	J	A	B	R	A	N	J	A	B	R	A	N
J	*						*					
E												
B			*						*			
R				*						*		
A		*			*			*			*	
N						*						*
J	*						*					
E												
B			*						*			
R				*						*		
A		*			*			*			*	
N						*						*

Dynamic Programming for sequence alignment

identity and substitution scoring, gap penalty

Dynamic programming matrix:

		j → (sequence y)								
		0	1 T	2 G	3 C	4 T	5 C	6 G	7 T	8 = N A
i	0	0	←−6	←−12	←−18	←−24	←−30	←−36	←−42	←−48
	1 T	−6	5	←−1	←−7	←−13	←−19	←−25	←−31	←−37
	2 T	−12	−1	3	←−3	−2	←−8	←−14	←−20	←−26
	3 C	−18	−7	−3	8	← 2	3	←−3	←−9	←−15
	4 A	−24	−13	−9	2	6	← 0	1	←−5	−4
	5 T	−30	−19	−15	−4	7	4	←−2	6	← 0
(sequence x)	M = 6 A	−36	−25	−21	−10	1	5	2	0	11

Optimum alignment scores 11:

T	–	–	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	−6	−6	+5	+5	−2	+5	+5

66

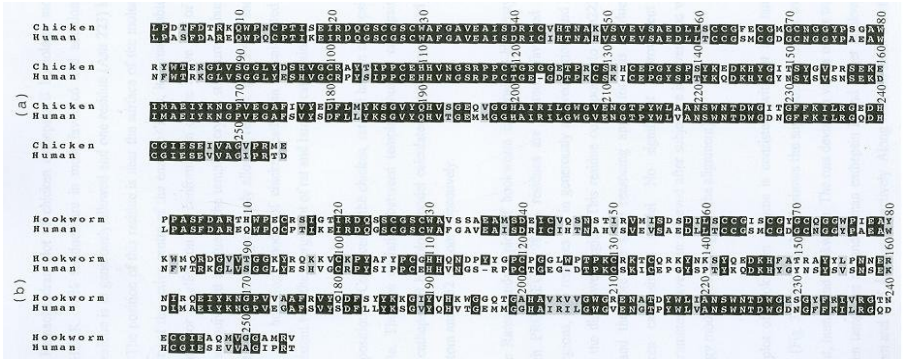
Sequence Analysis Programs

- Sequence comparison and alignment
  - Pairwise sequence alignment
    - FASTA; BLAST
  - Multiple sequence alignment
    - PILEUP; ClustalW
- Pattern search; PROSITE
- Phylogenetic analysis; Phylip
- Genome-level sequence analysis

67

Example: Pairwise sequence alignment of

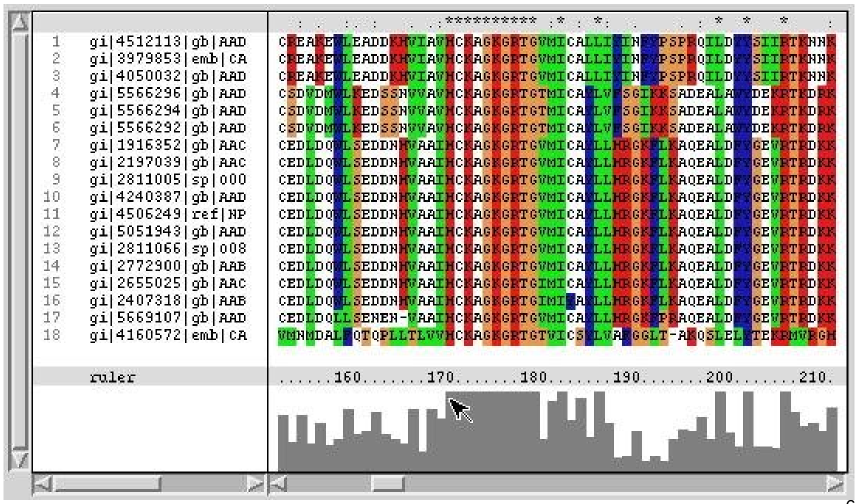
(a) human and chicken cathepsin B and
(b) human and hookworm cathepsin B.
Identical residues are indicated as dark blocks.



68

Multiple sequence alignment

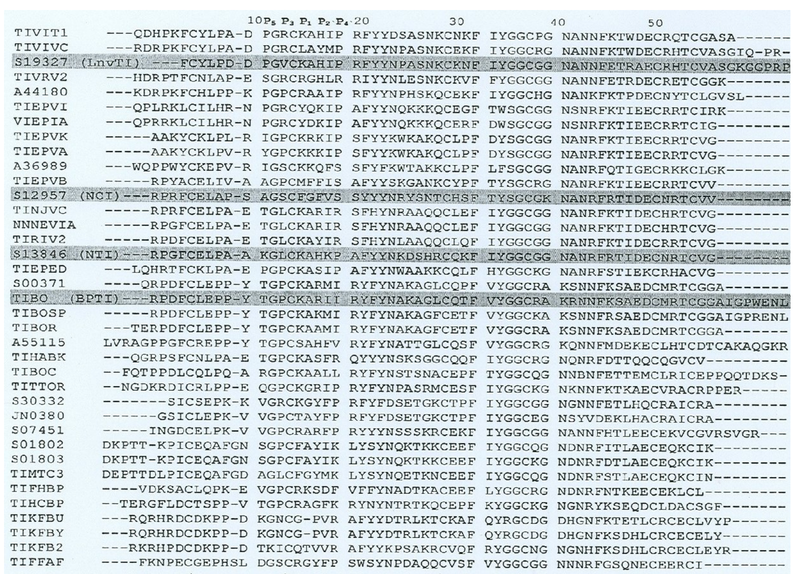
Alignment of 3 or more sequences.



69

Multiple sequence alignment

of the family of kunitz-type proteinase inhibitors

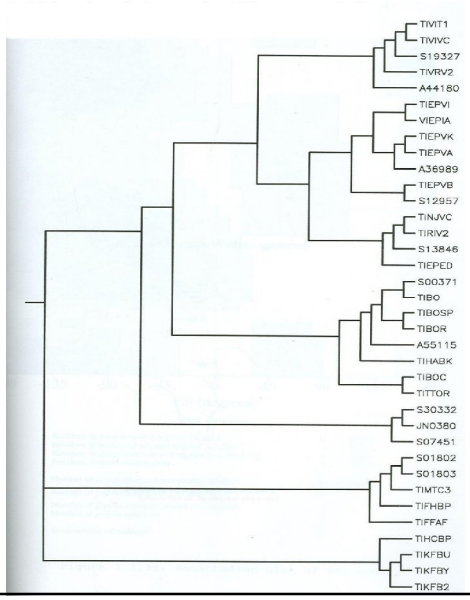


```
                          10P5 P3 P1 P2 P4 20       30            40           50
TIVIT1         ---QDHPKFCYLPA-D PGRCKAHIP RFYYDSASNKCNKF IYGGCPG NANNFKTWDECRQTCGASA------
TIVIVC         ---RDRPKFCYLPA-D PGRCLAYMP RFYYNPASNKCEKF IYGGCRG NANNFKTWDECRHTCVASGIQ-PR-
S19327 (LnvTI) ---FCYLPD-D PGVCKAHIP RFYYNPASNKCKNF IYGGCGG NANNFETRAECRHTCVASGKGGPRP
TIVRV2         ---HDRPTFCNLAP-E SGRCRGHLR RIYYNLESNKCKVF FYGGCGG NANNFETRDECRETCGGK-------
A44180         ---KDRPKFCHLPP-K PGPCRAAIP RFYYNPHSKQCEKF IYGGCHG NANKFKTPDECNYTCLGVSL-----
TIEPVI         ---QPLRKLCILHR-N PGRCYQKIP AFYYNQKKKQCEGF TWSGCGG NSNRFKTIEECRRTCIRK-------
VIEPIA         ---QPRRKLCILHR-N PGRCYDKIP AFYYNQKKKQCERF DWSGCGG NSNRFKTIEECRRTCIG--------
TIEPVK         -----AAKYCKLPL-R IGPCKRKIP SFYYKWKAKQCLPF DYSGCGG NANRFKTIEECRRTCVG--------
TIEPVA         -----AAKYCKLPV-R YGPCKKKIP SFYYKWKAKQCLPF DYSGCGG NANRFKTIEECRRTCVG--------
A36989         ---WQPPWYCKEPV-R IGSCKKQFS SFYFKWTAKKCLPF LFSGCGG NANRFQTIGECRKKCLGK-------
TIEPVB         -----RPYACELIV-A AGPCMFFIS AFYYSKGANKCYPF TYSGCRG NANRFKTIEECRRTCVV--------
S12957 (NCI)---RPRFCELAP-S AGSCFGFVS SYYYNRYSNTCHSF TYSGCGK NANRFRTIDECNRTCVV--------
TINJVC         -----RPRFCELPA-E TGLCKARIR SFHYNRAAQQCLEF IYGGCGG NANRFKTIDECHRTCVG--------
NNNEVIA        -----RPGFCELPA-E TGLCKARIR SFHYNRAAQQCLEF IYGGCGG NANRFKTIDECHRTCVG--------
TIRIV2         -----RPDFCELPA-E TGLCKAYIR SFHYNLAAQQCLQF IYGGCGG NANRFKTIDECRRTCVG--------
S13846 (NTI)---RPGFCELPA-A KGLCKAHKP AFYYNKDSHRCQKF IYGGCGG NANRFRTIDECNRTCVG--------
TIEPED         ---LQHRTFCKLPA-E PGPCKASIP AFYYNWAAKKCQLF HYGGCKG NANRFSTIEKCRHACVG--------
S00371         ----QRPDFCLEPP-Y TGPCKARMI RYFYNAKAGLCQPF VYGGCRA KSNNFKSAEDCMRTCGGA-------
TIBO   (BPTI)---RPDFCLEPP-Y TGPCKARII RYFYNAKAGLCQTF VYGGCRA KRNNFKSAEDCMRTCGGAIGPWENL
TIBOSP         -----RPDFCLEPP-Y TGPCKAKMI RYFYNAKAGFCETF VYGGCKA KSNNFRSAEDCMRTCGGAIGPRENL
TIBOR          ---TERPDFCLEPP-Y TGPCKAAMI RYFYNAKAGFCETF VYGGCRA KSNNFKSAEDCMRTCGGA-------
A55115         LVRAGPPGFCREPP-Y TGPCSAHFV RYFYNATTGLCQSF VYGGCRG KQNNFMDEKECLHTCDTCAKAQGKR
TIHABK         ---QGRPSFCNLPA-E TGPCKASFR QYYYNSKSGGCQQF IYGGCRG NQNRFDTTQQCQGVCV---------
TIBOC          --FQTPPDLCQLPQ-A RGPCKAALL RYFYNSTSNACEPF TYGGCQG NNBNFETTEMCLRICEPPQQTDKS-
TITTOR         --NGDKRDICRLPP-E QGPCKGRIP RYFYNPASRMCESF IYGGCKG NKNNFKTKAECVRACRPPER-----
S30332         -------SICSEPK-K VGRCKGYFP RFYFDSETGKCTPF IYGGCGG NGNNFETLHQCRAICRA--------
JN0380         ------GSICLEPK-V VGPCTAYFP RFYFDSETGKCTPF IYGGCEG NSYVDEKLHACRAICRA--------
S07451         -----INGDCELPK-V VGPCRARFP RYYYNSSSKRCEKF IYGGCGG NANNFHTLEECEKVCGVRSVGR---
S01802         DKPTT-KPICEQAFGN SGPCFAYIK LYSYNQKTKKCEEF IYGGCQG NDNRFITLAECEQKCIK--------
S01803         DKPTT-KPICEQAFGN SGPCFAYIK LYSYNQKTKKCEEF IYGGCKG NDNRFDTLAECEQKCIK--------
TIMTC3         DEPTTDLPICEQAFGD AGLCFGYMK LYSYNQETKNCEEF IYGGCQG NDNRFSTLAECEQKCIN--------
TIFHBP         ----VDKSACLQPK-E VGPCRKSDF VFFYNADTKACEEF LYGGCRG NDNRFNTKEECEKLCL---------
TIHCBP         --TERGFLDCTSPP-V TGPCRAGFK RYNYNTRTKQCEPF KYGGCKG NGNRYKSEQDCLDACSGF-------
TIKFBU         ---RQRHRDCDKPP-D KGNCG-PVR AFYYDTRLKTCKAF QYRGCDG DHGNFKTETLCRCECLVYP------
TIKFBY         ---RQRHRDCDKPP-D KGNCG-PVR AFYYDTRLKTCKAF QYRGCDG DHGNFKSDHLCRCECELY-------
TIKFB2         ---RKRHPDCDKPP-D TKICQTVVR AFYYKPSAKRCVQF RYGGCNG NGNHFKSDHLCRCECLEYR------
TIFFAF         ----FKNPECGEPHSL DGSCRGYFP SWSYNPDAQQCVSF VYGGCGG NNNRFGSQNECEERCI---------
                         *          *            :.    *  *  : ** : .            *    *
```

70

Phylogenetic Analysis

of kunitz-type proteinase inhibitors based on multiple sequence alignment



71

Part-IV

Sequence Database Searching

Database searching

- Searching for sequences in a database that are similar to a query sequence
- Often referred to as homology searching
- Currently the most commonly used method in bioinformatics
- Applications
  - Identifying functions
  - Retrieving homologues
  - Predicting structures
  - Characterizing domains
  - Identifying coding regions
- General principle:
  - Align a query sequence to every sequence in the database
  - Select the database sequences with the highest score

Similarity searching algorithms

- Smith Waterman
  - Slow-but exhautive
- FASTA
  - Heuristic approach (www.ebi.ac.uk)
- BLAST
  - Currently industry standard (www.ncbi.nlm.nih.gov/blast)

Database searching using BLAST

- Algorithm based on dynamic programming
- Different BLAST 'versions'
  - Nucteotide BLAST: Search a nucleotode DB using nucleotide query
    - Blastn, megablast
  - Protein BLAST: Search a protein DB using query protein sequence
  - BlastX: Search protein DB using a translated nucleotide query
  - tBlastn
  - tBlastx
  - PSI-Blast
  - PHI-Blast

Applications: Predicting functions

- For a protein sequence
  - Blastp and PSI-blast of the non-redundant DB
- For a DNA/RNA sequence
  - Blastn and/or Fasta of the non-redundant nucleotide DB
  - If coding, Blastx or Fasta of the non-redundant protein DB

Applications: predicting protein structure

- Blastp, psi-blast of the PDB
- If a good homologue is found, its structure can be used as template to model the query sequence

Applications: detecting genes

- Blastx and/or FastX of the non-redundant protein database
- Blastn and/or tBlastx of the EST database and of the well characterized mRNA database (e.g. RefSeq)

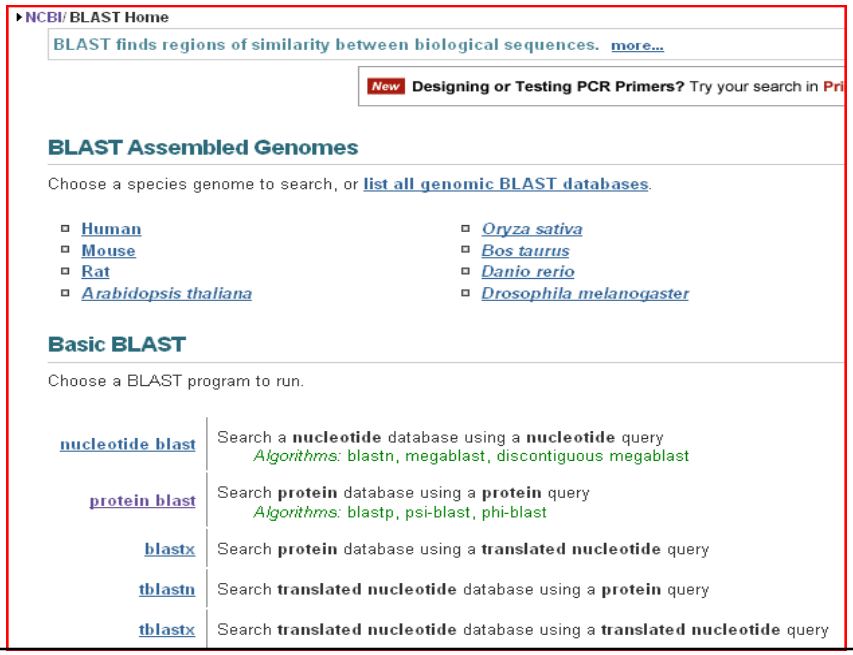
Database Searching Tips

- Use the latest version of the DB, preferably a non-redundant DB
- For sensitive searches, use more than one program
- Translate DNA query to search a protein DB
- When using a protein query sequence, also search a 6-frames translation of a DNA DB
- Split large query sequences into smaller segments (if > 1000 for DNA or > 200 for protein)
- If a sequence is found to contain a segment with a lot of hits in the DB, delete this segment and repeat the search

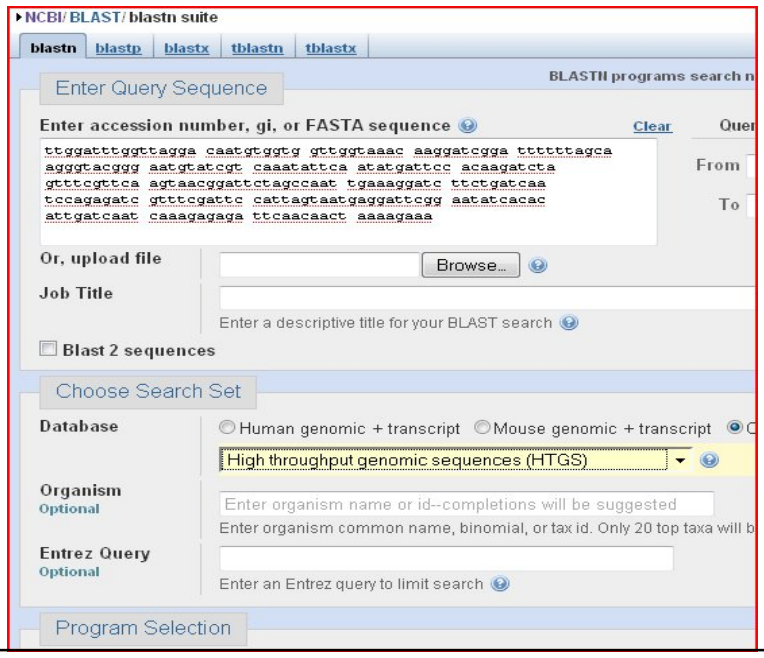
Database Searching Tips

- Sequences with an E value < 0.02-0.05 (protein) or < 0.001 (DNA) are statistically significant and almost always biologically significant
- What is biologically significant may not always be statistically significant: carefully examine hits with an expectation (E value) with 0.05-10.
- Use biological reasoning and common sense when interpreting the results
- Do not expect that computer will tell you the truth (always).

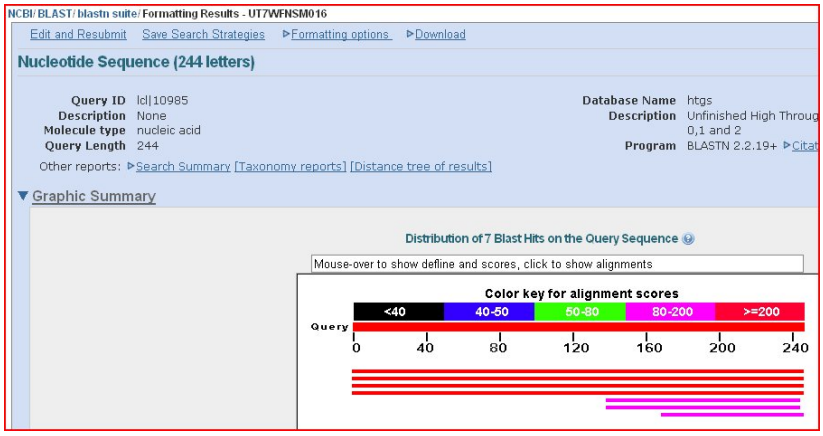
Blast page at the NCBI website



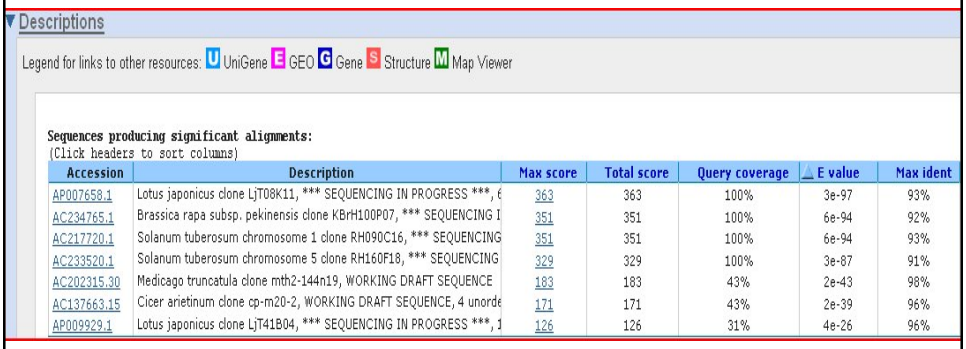
Blast sequence submission page



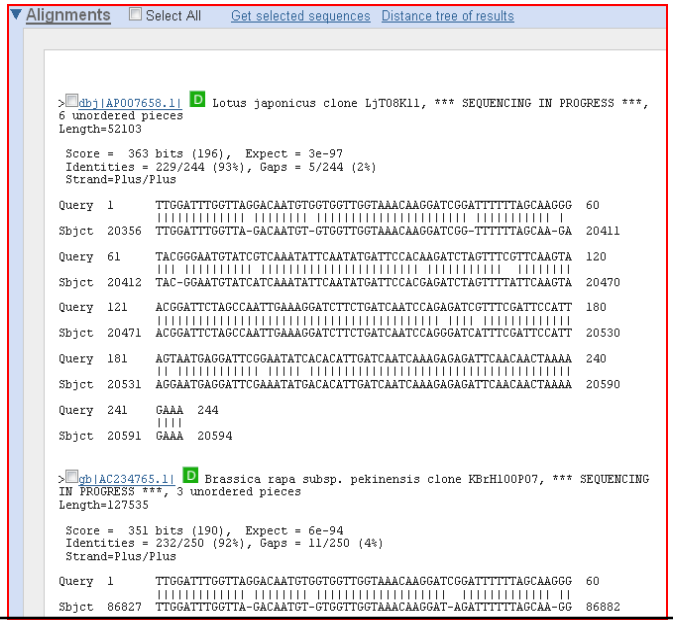
Blast search result-Header



Blast result page-Description



Blast result page-Alignment



Alignments Select All Get selected sequences Distance tree of results

>dbj|AP007658.1| Lotus japonicus clone LjT08K11, *** SEQUENCING IN PROGRESS ***, 6 unordered pieces
Length=52103

Score = 363 bits (196), Expect = 3e-97
Identities = 229/244 (93%), Gaps = 5/244 (2%)
Strand=Plus/Plus

```
Query  1      TTGGATTTGGTTAGGACAATGTGGTGGTTGGTAAACAAGGATCGGATTTTTTAGCAAGGG  60
              ||||||||||||| |||||||| ||||||||||||||||||||| |||||||||| |
Sbjct  20356  TTGGATTTGGTTA-GACAATGT-GTGGTTGGTAAACAAGGATCGG-TTTTTTAGCAA-GA  20411

Query  61     TACGGGAATGTATCGTCAAATATTCAATATGATTCCACAAGATCTAGTTTCGTTCAAGTA  120
              ||| |||||||||| |||||||||||||||||||||| ||||||||||  |||||||||
Sbjct  20412  TAC-GGAATGTATCATCAAATATTCAATATGATTCCACGAGATCTAGTTTTATTCAAGTA  20470

Query  121    ACGGATTCTAGCCAATTGAAAGGATCTTCTGATCAATCCAGAGATCGTTTCGATTCCATT  180
              ||||||||||||||||||||||||||||||||||||||||||||||| |||| ||||||||||||
Sbjct  20471  ACGGATTCTAGCCAATTGAAAGGATCTTCTGATCAATCCAGGGATCATTTCGATTCCATT  20530

Query  181    AGTAATGAGGATTCGGAAATATCACACATTGATCAATCAAAGAGAGATTCAACAACTAAAA  240
              || |||||||||||| ||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  20531  AGGAATGAGGATTCGAAATATGACACATTGATCAATCAAAGAGAGATTCAACAACTAAAA  20590

Query  241    GAAA  244
              ||||
Sbjct  20591  GAAA  20594
```

>gb|AC234765.1| Brassica rapa subsp. pekinensis clone KBrH100P07, *** SEQUENCING IN PROGRESS ***, 3 unordered pieces
Length=127535

Score = 351 bits (190), Expect = 6e-94
Identities = 232/250 (92%), Gaps = 11/250 (4%)
Strand=Plus/Plus

```
Query  1      TTGGATTTGGTTAGGACAATGTGGTGGTTGGTAAACAAGGATCGGATTTTTTAGCAAGGG  60
              ||||||||||||| |||||||| |||||||||||||||||  |||||||||||| ||
Sbjct  86827  TTGGATTTGGTTA-GACAATGT-GTGGTTGGTAAACAAGGAT-AGATTTTTTAGCAA-GG  86882
```

Part-V

Bioinformatics in molecular medicine

How Bioinformatics supports Molecular Medicine?

- Genome-level sequence analysis of medically important organisms in order to;

  gain comprehensive knowledge for their life cycle,
  characterization of disease causing factors,
  identify new targets for therapeutic intervention
- Novel drug targets

  Computational tools can identify and validate new drug targets that act on the cause, not merely the symptoms of the disease.
- Personalised medicine

  Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- Preventative medicine
- Gene therapy

87

Specialized databases useful in Molecular Medicine

- OMIM- Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders.
- ENSEMBL- is designed to allow free access to all the genetic information available about the Human Genome.
- Human Gene Mutation DB- contains sequences and phenotypes of human disease-causing mutations.
- KEGG- to computerize knowledge of molecular interactions namely metabolic pathways, regulatory pathways and molecular assemblies.
- dbSNP- Single Nucleotide Polymorphisms DB
- GeneCards- an integrated DB of human genes that includes automatically-mined genomic, proteomic and transcriptomic information, as well as orthologies, disease relationships, SNPs, gene expression, gene function etc.

88

A Case Study

To understand the capabilities of bioinformatics

- A patient told to the physician that she has “a genetic form of diabetes”.
- Physician can gather information in the following way.
  - OMIM→diabetes→Diabetes Mellitus, Autosomal Dominant, Type II→men ons defect in control of the Glucokinase gene
  - ENTREZ→Glucokinase→DNA/protein sequence as well as literature
  - Retrieve the sequence and perform sequence analysis
  - Motif search associated with various functions revealed binding sites of glucose and ATP
  - GENBANK→analysis of DNA bases that encode for glucokinase
  - PDB→overall structure, precise loca on of glucose and ATP binding

A Case Study

To understand the capabilities of bioinformatics

- The physician has used a number of databases
- On the basis of collected information about the molecular basis of disease, the physician can take following steps;
  - Confirm the diagnosis by sequencing of relevant part of DNA
  - Prescribe an appropriate treatment
  - Discuss the prognosis of disease
  - Genetic counseling
  - Might ask of the participation in the clinical study of new treatment(s)

The nature of biology in the "post-genome era":

"The new paradigm, now emerging, is that all genes will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical."

Walter Gilbert, 1993

End Note

- Bioinformatics is the body of Knowledge; A wealth of data on sequences and structures.
- Key Resource is KNOWLEDGE
- And the key technology is INFORMATION HANDLING

92

Leading Bioinformatics Institutions

European Bioinformatics Institute, Cambridge, UK
National Center for Biotechnology Information, USA
National Human Genome Research Institute, USA
EMBL, Heidelberg, Germany
J. Craig Ventor Institute, USA
The Sanger Institute, UK

Bioinformatics Journals and Books

Bioinformatics
Genome Research
Nucleic Acid Research
Bioinformatics by D.W. Mount
Introduction to Bioinformatics by Attwood
Structural Bioinformatics by P.E. Bourne
Bioinformatics; A beginner's Guide by Claverie
Bioinformatics Computing by B. Bergeron

Bioinformatics Societies

International Society for Computational Biology (ICSB)
Asia Pacific Bioinformatics Network (APBioNet)
European Conference on Computational Biology (ECCB)

94