# Introduction to Sequences and Databases

Shifra Ben-Dor
Irit Orr

Bioinformatics Unit
Life Sciences Core Facilities

WEIZMANN INSTITUTE OF SCIENCE

# Lecture Outline:

- Technical Course Items
- Sequences
- Databases
  - This week and next week

# What “units of information” do we deal with in bioinformatics?

- DNA
- RNA
- Protein
- Sequence
- Structure
- Evolution
- Pathways
- Interactions
- Mutations

# Examples of biological data used in bioinformatics

- DNA (Genome)
- RNA (Transciptome)
- Protein (Proteome)

# DNA

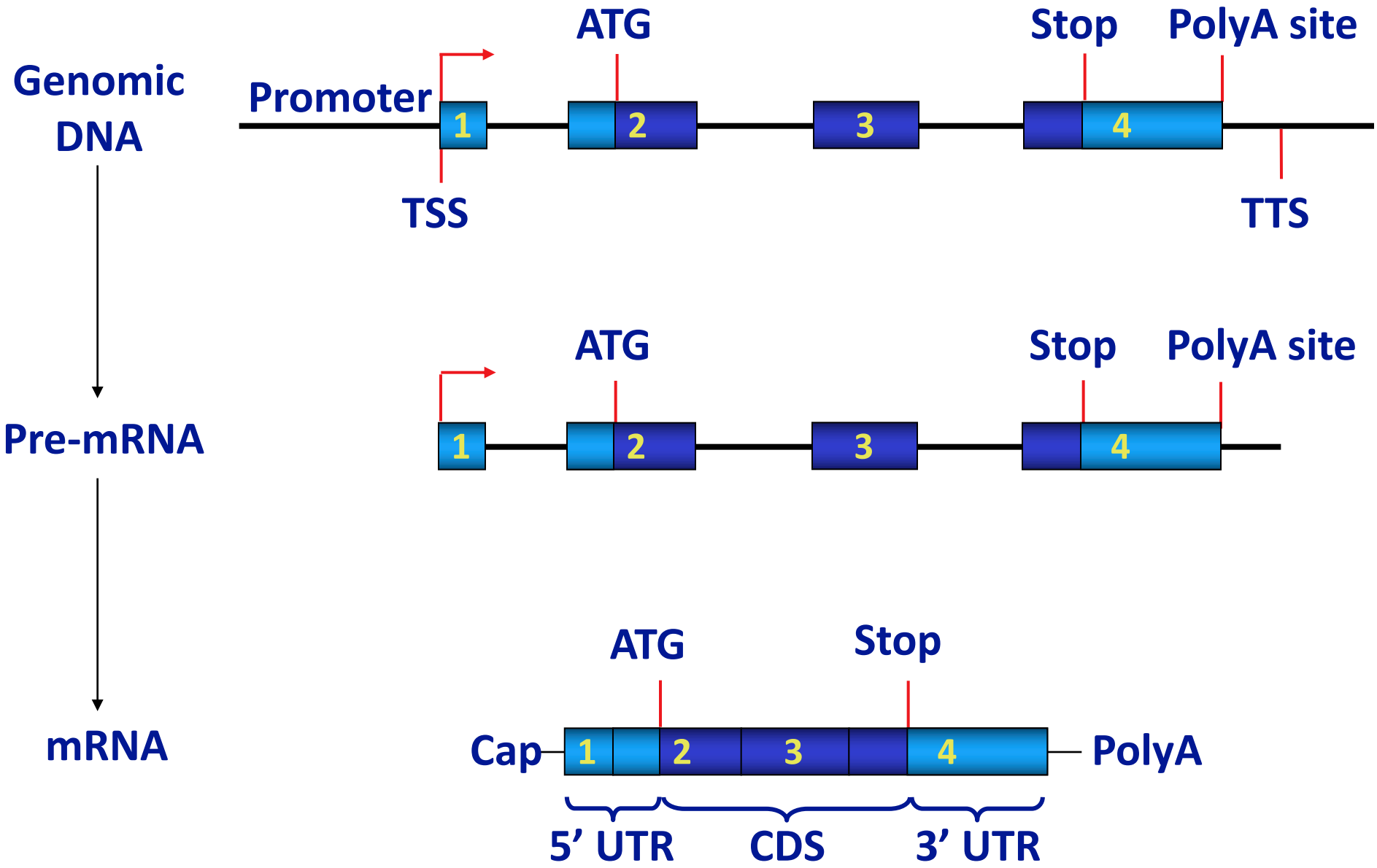
## Raw DNA Sequence

```
atggcaattaaaattggtatca
atggttttggtcgtatcggccg
tatcgtattccgtgcagcacaa
caccgtgatgacattgaagttg
taggtattaacgacttaatcga
cgttgaatacatggcttatatg
ttgaaatatgattcaactcacg
gtcgtttcgacggcactgttga
agtgaaagatggtaacttagtg
gttaatggtaaaactatccgtg
taactgcagaacgtgatccatc
```

- Coding or Not coding?
- Parse into genes?
- Other important genomic elements?
- 4 bases: ACGT

# DNA/RNA sequences

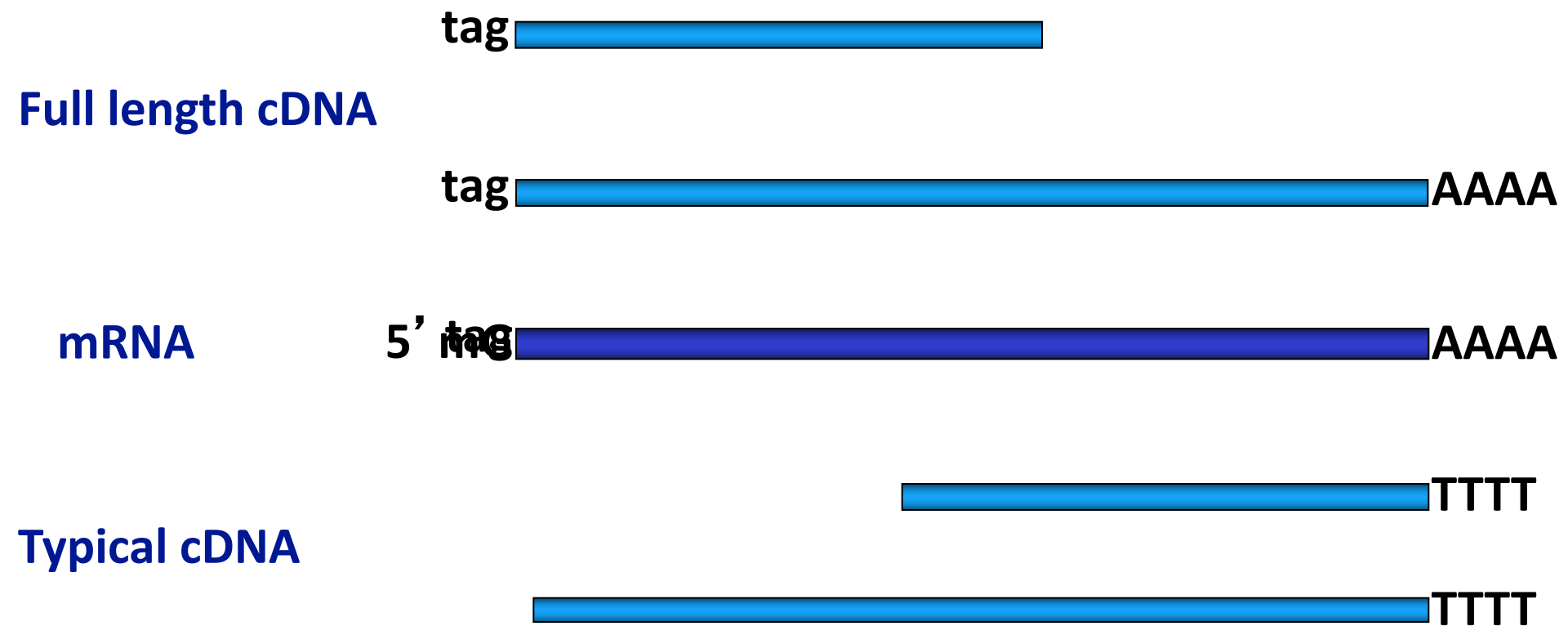
- Genes are encoded in genomic sequences.
- Genes are transcribed into pre-mRNAs (including coding, intronic, 5’ and 3’ untranslated regions).
- mRNAs are spliced (introns removed) and translated into proteins.
- mRNAs are copied to cDNAs (in the lab)




Modified from Zhang MQ Nat Rev Genet. 2002 Sep;3(9):698-709.

# Sources of mRNAs

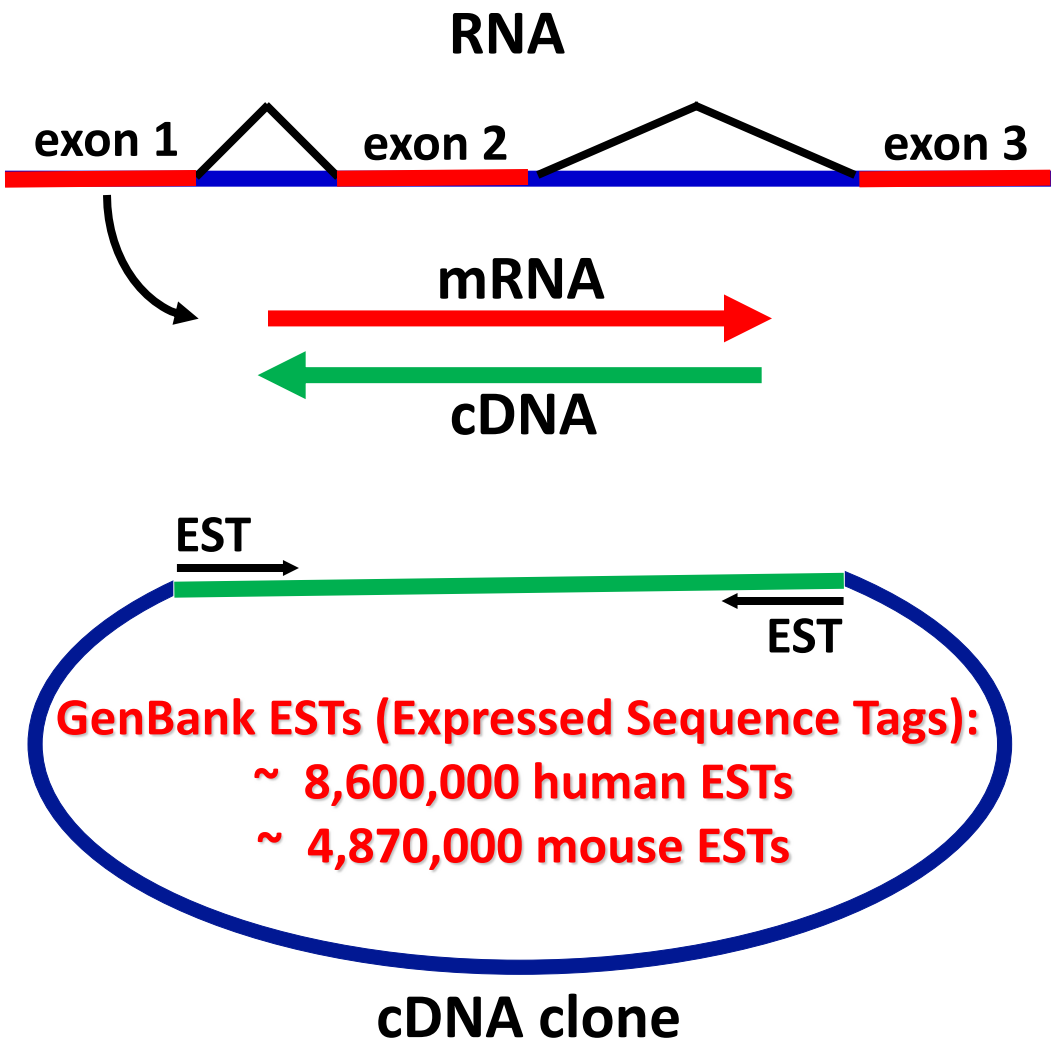
- Experimental
  - Clone new gene
  - “Clone” gene from database
  - RNA-Seq
- Database
  - “Typical” cDNA
  - Full length cDNA
  - EST (Expressed Sequence Tag)
  - Short read sequences
  - Long read sequences


tag
Full length cDNA
tag
AAAA
mRNA
5’ tag
AAAA
TTTT
Typical cDNA
TTTT

# Sources of mRNAs

- Experimental
  - Clone new gene
  - “Clone” gene from database
  - RNASeq (Short, Long)
- Database
  - “Typical” cDNA
  - Full length cDNA
  - EST (Expressed Sequence Tag)
  - Short read sequences
  - Long read sequences

# RNA, cDNA, and ESTs




Adapted with permission from Adam Sartiel

# Uses of ESTs

- prediction of coding regions
- detection of alternative splicing
- clustering to form “genes”

Problems with clustering:

- incomplete coverage breaks genes up
- gene families

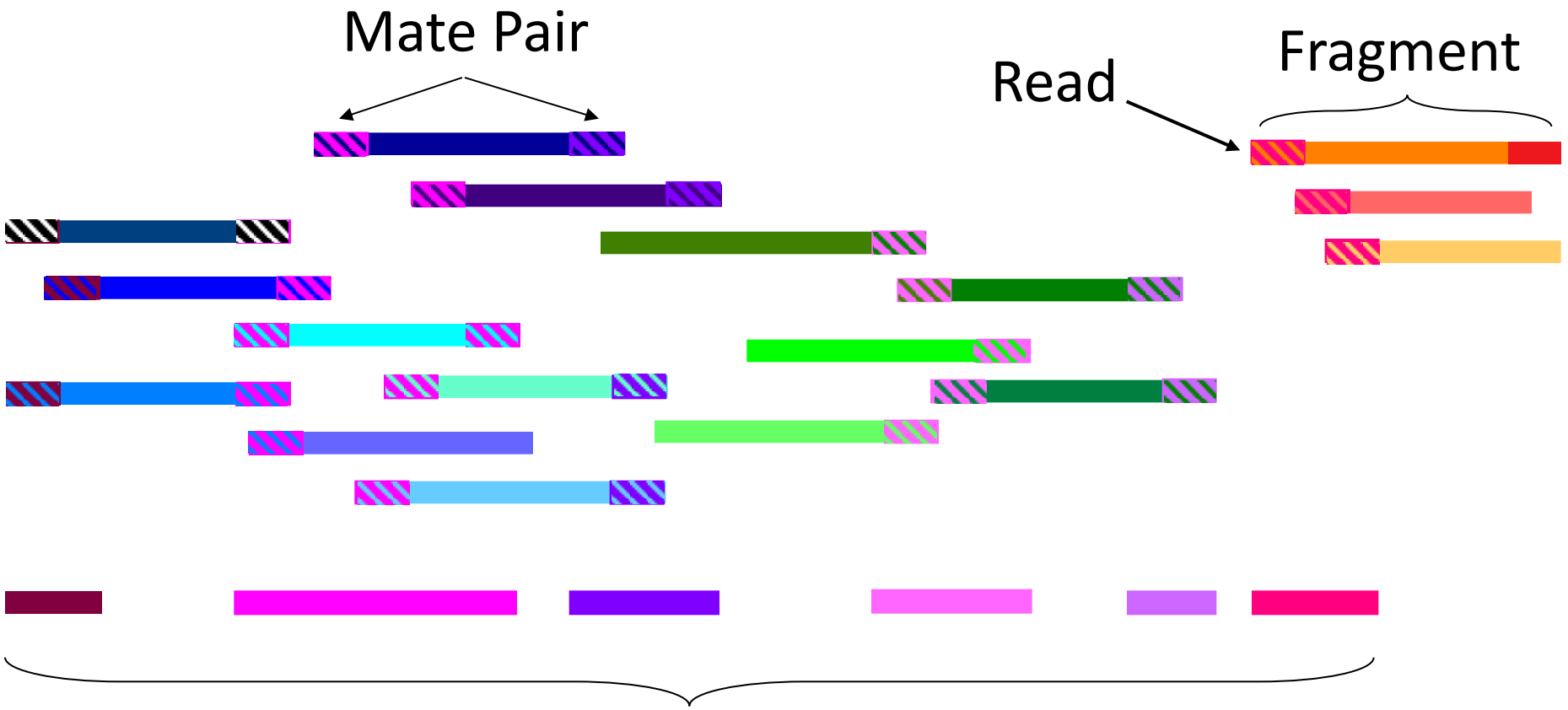
# Problems with ESTs

- low copy number genes
- rare tissues
- mistakes
- enrichment of 3’ ends of genes
- incomplete coverage of genes

# Short Read Sequencing

- Sequence lengths range from 20-25 bp to 75-100 to 150 bp reads
- Can be 3’ end only
- Can be paired or single read

# Paired end reads



# Problems with Short Reads

- have to be assembled to make transcripts
- incomplete coverage breaks genes up
- can’t tell which splice goes with which
- gene families
- rare tissues
- mistakes
- only sequence 3’ ends of genes

# EST vs Short Read

- ESTs have longer continuous sequence, so better to see gene structure (alternative splicing)
- Short reads generally have higher accuracy
- Both cannot give a picture of a whole gene

# Long reads

- Sequence lengths from several hundred to many thousand
- Have the potential to get full transcripts with splicing
- Technologies are more error prone than short read

# Long read problems

- Error rate
- Dependent on quality of RNA – may be 5’ end degraded
- Have to be clustered, and splicing defined
- Easier to do with known genomic sequence
- May be more expensive

# Protein

- 20 letter alphabet
  ACDEFGHIKLMNPQRSTVWY
  **But not** BJOUXZ
- Strings of ~300 aa in an average protein
  (e.g. bacteria)
- Protein are divided into domains

```
MLNCIVAVSQNMGIGKNG
DLPWPPLRNEFRYFQRMT
TTSSVEGKQNLVIMGKKT
WFSILNSIVAVCQNMGIG
KDGNLPWPPLRNEYKYFQ
RMTSTSHVEGKQNAVIMG
KKTWFSIISLIAALAVDR
VIGMENAMPWNLPADLAW
FKRNTLDKPVIMGRHTWE
SITAFLWAQDRNGLIGKD
GHLPWHLPDDLHYFRAQT
VGKIMVVGRRTYESF
```

# Protein

- Proteome of an Organism
- 2D gels
- Mass Spec
- 2D Structure
- 3D Structure
- 4D Structure (interactions)

# Lecture Outline:

- Technical Course Items
- Sequences
- Databases

# Databases: Outline

- Introduction
  - Data and Database types
  - Database components
- Data Formats
- Sample databases
- How to text search databases

# What “units of information” do we deal with in bioinformatics?

- DNA
- RNA
- Protein
- Sequence
- Structure
- Evolution
- Pathways
- Interactions
- Mutations

SNPs

Nucleotide sequence

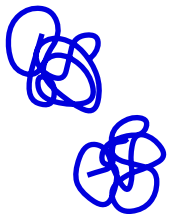
AAGTGCCACTGCATAAATGACCATGAGTGGGCACCGGTAAGGGAGGGTGATGCTATCTGGTCTGAAG

Genes

mRNA

Protein primary sequence





Protein 3D structure

Protein Function

Acts as a tumor suppressor in many tumor types. induces growth arrest or apoptosis depending on the physiological circumstances or cell type, but both activities are involved in tumor suppression.

Involved in the transport of chloride ions. Defects in CFTR are the cause of cystic fibrosis. It is the most common genetic disease in the caucasian population, with a prevalence of about 1 in 2000 live births. cf, an autosomal recessive disorder, is a common generalized disorder of exocrine gland function

Slide provided by Dr. Vered Caspi

All of these have databases and tools that were created to work with them

What do we want from databases?

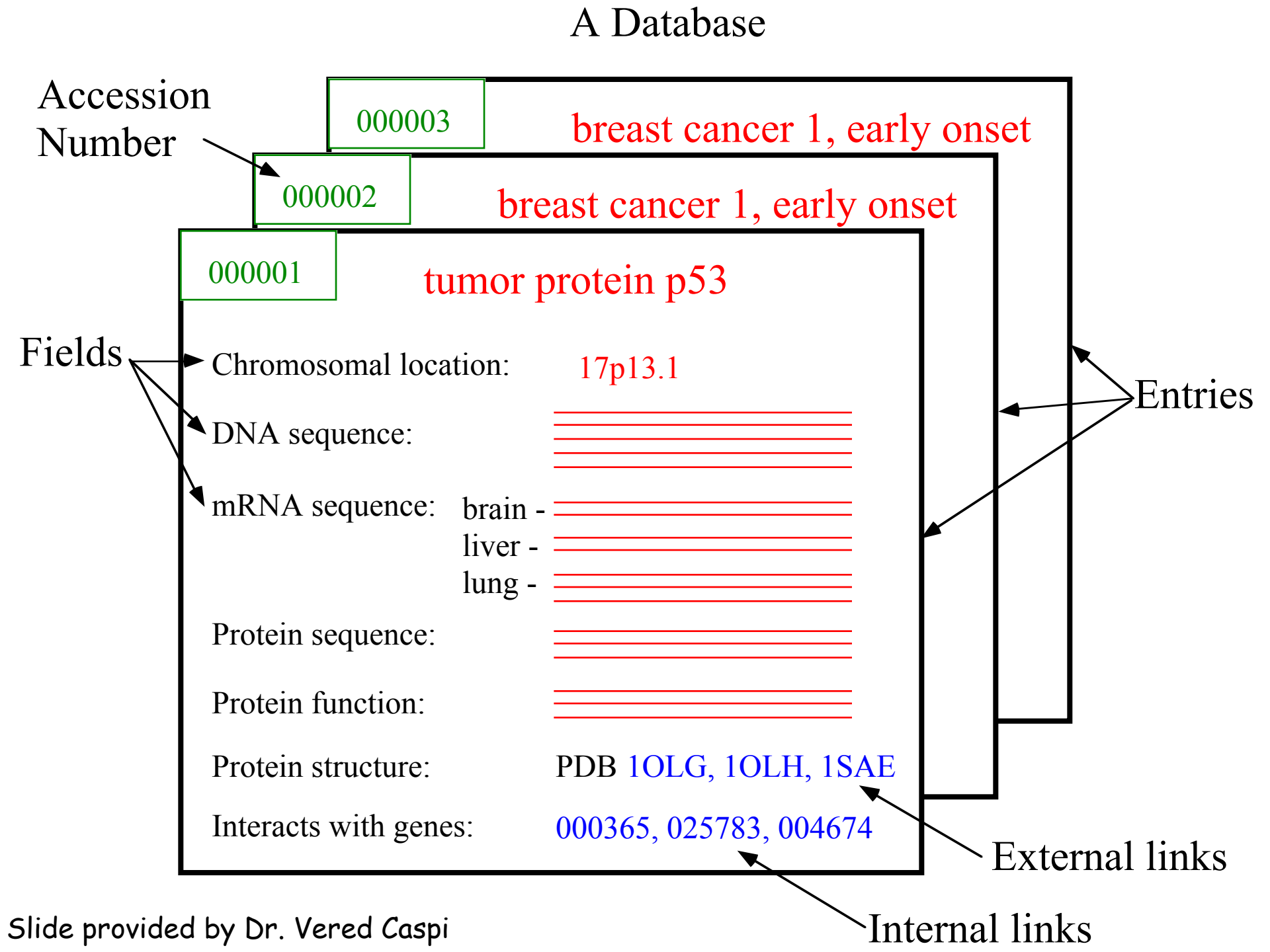
# Information retrieval from sequence databases

Biological databases contain enormous amounts of data.

- Databases need to be well annotated.
- Databases need to be easily searched.
- Data found in databases should be easily retrieved.
- Data in databases should be in standard formats.

# Integrated Information Retrieval

- Many databases contain logical relations between specific entries.
- One interface - connecting many biological databases.
- For example: a database that connects between protein sequence, protein domain, protein structure and reference databases. (Interpro)
- Another example: Connection between references, protein sequence, DNA sequence, and structure databases. (Entrez)


A Database
Accession Number
000003
breast cancer 1, early onset
000002
breast cancer 1, early onset
000001
tumor protein p53
Fields
Chromosomal location:
17p13.1
DNA sequence:
mRNA sequence:
brain -
liver -
lung -
Protein sequence:
Protein function:
Protein structure:
PDB 1OLG, 1OLH, 1SAE
Interacts with genes:
000365, 025783, 004674
Entries
External links
Internal links
Slide provided by Dr. Vered Caspi

# Core Data and Annotation

Databases generally have (at least) two types of data:

Core data: The data the database was generated to organize

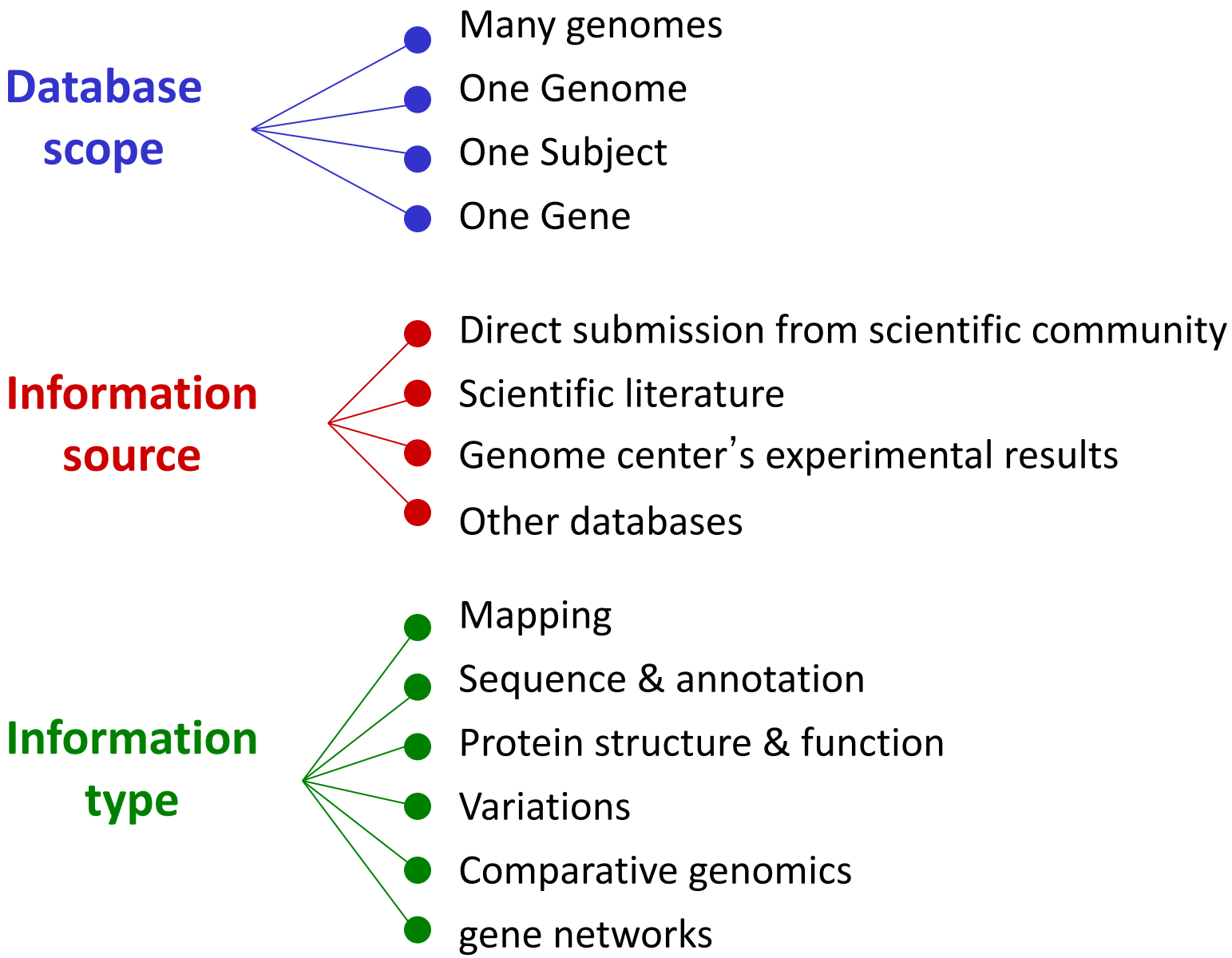
Annotation: Extra information that rounds out our picture of the core data

For example in a genome database, the sequence is the core data, and the location of genes is the annotation

# Database Issues

- Printed journals vs. databases
- Direct submission to databases (e.g. GenBank, PDB)
- Archival vs. curated databases
- Databases that publish experimental results of large genomic centers
- Public vs. private databases
- Raw data vs Processed Data

# For Example: Classification of Genomic Databases




Slide provided by Dr. Vered Caspi

# User Interface

- Database search
  - free text
  - field-specific
  - sequence-based
- Database output
  - text
  - graphics
  - dynamic

# Data Formats

There are many data formats used for sequences (both nucleic and amino acid)

- Fasta Format
- GenBank Format
- Fastq Format

- (EMBL Format)

# Fasta Format

- Simplest format
- Least information
- Starts with a > and sequence name on one line
- The sequence in plain text follows

>OB2T2

GTGACAACATGTACAGCTGTGAGCGGTGTAAGAAGCTGCGGAACGGAGTGAAGTACTGCA
AAGTCCTGCGGTTGCCCGAGATCCTGTGCATTCACCTAAAGCGCTTTCGGCACGAGGTGA
TGTACTCATTCAAGATCAACAGCCACGTCTCCTTGCCCTCGAGGGGCTCGACCTGCGCCC
CTTCCTTGCCAAGGAGTGCACATCCCAGATCACCACCTACGACCTCCTCTCGGTCATCTG
CCACCACGGCACGGCAGGCA

>TNRC_HUMAN P36941 (tumor necrosis factor c receptor)
MLLPWATSAPGLAWGPLVLGLFGLLAASQPQAVPPYASENQTCRDQEKEYYEPQHRICCS
RCPPGTYVSAKCSRIRDTVCATCAENSYNEHWNYLTICQLCRPCDPVMGLEEIAPCTSKR
KTQCRCQPGMFCAAWALECTHCELLSDCPPGTEAELKDEVGKGNNHCVPCKAGHFQNTSS
PSARCQPHTRCENQGLVEAAPGTAQSDTTCKNPLEPLPPEMSGTMLMLAVLLPLAFFLLL
ATVFSCIWKSHPSLCRKLGSLLKRRPQGEGPNPVAGSWEPPKAHPYFPDLVQPLLPISGD
VSPVSTGLPAAPVLEAGVPQQQSPLDLTREPQLEPGEQSQVAHGTNGIHVTGGSMTITGN
IYIYNGPVLGGPPGPGDLPATPEPPYPIPEEGDPGPPGLSTPHQEDGKAWHLAETEHCGA
TPSNRGPRNQFITHD
>TNRC_MOUSE P50284 lymphotoxin-beta receptor precursor
MRLPRASSPCGLAWGPLLLGLSGLLVASQPQLVPPYRIENQTCWDQDKEYYEPMHDVCCS
RCPPGEFVFAVCSRSQDTVCKTCPHNSYNEHWNHLSTCQLCRPCDIVLGFEEVAPCTSDR
KAECRCQPGMSCVYLDNECVHCEEERLVLCQPGTEAEVTDEIMDTDVNCVPCKPGHFQNT
SSPRARCQPHTRCEIQGLVEAAPGTSYSDTICKNPPEPGAMLLLAILLSLVLFLLFTTVL
ACAWMRHPSLCRKLGTLLKRHPEGEESPPCPAPRADPHFPDLAEPLLPMSGDLSPSPAGP
PTAPSLEEVVLQQQSPLVQARELEAEPGEHGQVAHGANGIHVTGGSVTVTGNIYIYNGPV
LGGTRGPGDPPAPPEPPYPTPEEGAPGPSELSTPYQEDGKAWHLAETETLGCQDL
>TNR1_RAT P22934 tumor necrosis factor receptor 1 precursor (p60)
MGLPIVPGLLLSLVLLALLMGIHPSGVTGLVPSLGDREKRDNLCPQGKYAHPKNNSICCT
KCHKGTYLVSDCPSPGQETVCEVCDKGTFTASQNHVRQCLSCKTCRKEMFQVEISPCKAD
MDTVCGCKKNQFQRYLSETHFQCVDCSPCFNGTVTIPCKEKQNTVCNCHAGFFLSGNECT
PCSHCKKNQECMKLCLPPVANVTNPQDSGTAVLLPLVIFLGLCLLFFICISLLCRYPQWR
PRVYSIICRDSAPVKEVEGEGIVTKPLTPASIPAFSPNPGFNPTLGFSTTPRFSHPVSST
PISPVFGPSNWHNFVPPVREVVPTQGADPLLYGSLNPVPIPAPVRKWEDVVAAQPQRLDT
ADPAMLYAVVDGVPPTRWKEFMRLLGLSEHEIERLELQNGRCLREAHYSMLEAWRRRTPR
HEATLDVVGRVLCDMNLRGCLENIRETLESPAHSSTTHLPR

# Known Issues with Fasta Format

- Different programs treat the header line differently:
  - Some read 10 characters, some 30
  - Some read until the first space
- Make sure you have unique names!!!
- Header lines should be under 80 characters
- Length of sequence line can differ

# Fastq Format

```
@SRR2976060.1 1 length=202
NAAGCTCTCACCCATGGAGACCAAGGCGATTAGGGTTTTTCTCTTCGCTCTCCTCCT
+SRR2976060.1 1 length=202
#1=DDFFFHHHHHJJJEIJJJJJIJJJJFHGJIIJ9DHIIIJJJJGIIJJJGIIIJJ
```

Four lines:

1 – starts with @ and is a unique identifier
2 – the actual sequence
3 – starts with a + and can have an identifier again
4 – the quality of the bases