Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

1

Biomedical Data

2

Learning objectives

- After this lecture you should be able to:
  - understand fundamentals of molecular biology
  - define the types of molecular databases;
  - define accession numbers and the significance of RefSeq identifiers;
  - describe the main genome browsers and use them to study features of a genomic region;
  - use resources to study information about both individual genes (or proteins) and large sets of genes/proteins.

3

The Nature and Representation of Biomedical Data

- The first things to consider:
  - the various forms of biomedical data
  - how such data can be
    - represented in a computer
    - manipulated by a computer program
- In a typical biology/medical textbook
  - photographs, diagrams, drawings, chemical formulas, and lots of description
    - about the attributes of biological entities such as cells, organs, tissues, fluids, chemical compounds found in all those and about the relations between these entities and their properties.

4

The Nature and Representation of Biomedical Data

- Some of the properties of biological objects
  - numerical (quantities),
    - the concentration of certain chemicals in blood,
    - the size of a tumor,
    - the pH (degree of acidity) in a cell, tissue, organ, or body fluid.
  - qualities that can only be named but not quantified
    - the protein(s) produced by gene transcription,
    - the presence or absence of an organ in an organism,
    - the parts of an organ.
      - [These all have names, not numerical values]

5

Metadata

- A set of data that describes and gives information about other data.
  - Descriptive metadata
    - For finding or understanding a resource
  - Administrative metadata
    - Technical metadata
      - For decoding and rendering files
    - Preservation metadata
      - Long-term management of files
    - Rights metadata
      - Intellectual property rights attached to content
  - Structural metadata
    - Relationships of parts of resources to one another
  - Markup languages
    - Integrates metadata and flags for other structural or semantic features within content

6

Metadata

- Metadata is defined as the data providing information about one or more aspects of the data
- It is used to summarize basic information about data which can make tracking and working with specific data easier
- Some examples include:
  - Means of creation of the data
  - Purpose of the data
  - Time and date of creation
  - Creator or author of the data
  - Location on a computer network where the data was created
  - Standards used
  - File size

- http://www.ntsu.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

7

Overview of Molecular Biology

- **Molecular biology** is the study of biology at a molecular level.
- The field overlaps with other areas of biology, chemistry, genetics, and biochemistry.
- Molecular biology concerns itself with understanding the interactions between the various systems of a cell, including the interactions between DNA, RNA and protein biosynthesis and learning how these interactions are regulated.

8

Overview of Molecular Biology

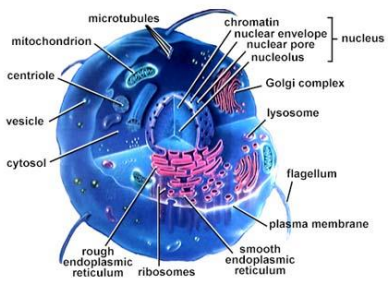
- Cells
- Chromosomes
- DNA
- RNA
- Amino Acids
- Proteins
- Genome/Transcriptome/Proteome

9

Cells

Example Animal Cell

www.ebi.ac.uk/microarray/ biology_intro.htm



10

Cells

- Complex system enclosed in a membrane
- The structural and functional unit of all known living organisms.
- The smallest unit of an organism that is classified as living, and is often called the building block of life.
- Some organisms, such as most bacteria, are unicellular
  - consist of a single cell
- Other organisms, such as humans, are multicellular.
  - Humans have an estimated 100 trillion or 10^{14} cells and 320 cell types
  - a typical cell size is 10 µm
  - a typical cell mass is 1 nanogram.

11

Cells

- Cell types
  - squamous cells,
  - epithelial cells,
  - muscle cells,
  - blood cells.
    - red cells, white cells
- These categorical attributes are also related back to numerical values.
  - Since we can count cells, it is possible to report for an individual the concentrations of different types of blood cells in the blood.

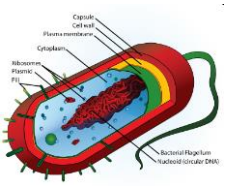
12

Organisms

- Classified into two types:
  - **Prokaryotes**:
    - lack a true membrane-bound nucleus and organelles (single-celled, includes bacteria)
  - **Eukaryotes**:
    - contain a membrane-bound nucleus and organelles (plants, animals, fungi,…)
- *Not all single celled organisms are prokaryotes!*

13

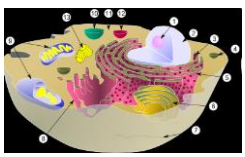
Prokaryotic cells



- Prokaryotes lack a nuclear envelope and a cell nucleus.
- Prokaryotes also lack most of the intracellular organelles and structures that are seen in eukaryotic cells.
- There are two kinds of prokaryotes:
  - bacteria and archaea,
    - but these are similar in the overall structures of their cells.
- Most functions of organelles, such as mitochondria, chloroplasts, and the Golgi apparatus, are taken over by the prokaryotic cell's plasma membrane.
- Prokaryotic cells have three architectural regions:
  - appendages called flagella and pili — proteins attached to the cell surface;
  - a cell envelope - consisting of a capsule, a cell wall, and a plasma membrane;
  - a cytoplasmic region that contains the cell genome (DNA) and ribosomes and various sorts of inclusions.

14

Eukaryotic cells



(1) nucleolus
(2) nucleus
(3) ribosome
(4) vesicle
(5) rough endoplasmic reticulum (ER)
(6) Golgi apparatus
(7) Cytoskeleton
(8) smooth endoplasmic reticulum
(9) mitochondria
(10) vacuole
(11) cytoplasm
(12) lysosome
(13) centrioles within centrosome

- Eukaryotic cells are about 10 times the size of a typical prokaryote and can be as much as 1000 times greater in volume.
- The major difference between prokaryotes and eukaryotes is that eukaryotic cells contain membrane-bound compartments in which specific metabolic activities take place.
- Most important among these is the presence of a cell nucleus, a membrane-delineated compartment that houses the eukaryotic cell's DNA.
- It is this nucleus that gives the eukaryote its name, which means "true nucleus.
- The eukaryotic DNA is organized in one or more linear molecules, called chromosomes, which are associated with histone proteins.
- All chromosomal DNA is stored in the *cell nucleus*, separated from the cytoplasm by a membrane.
- Some eukaryotic organelles such as mitochondria also contain some DNA.

15

Chromosomes

- **Chromosomes** are organized structures of DNA and proteins that are found in cells.
  - The word *chromosome* comes from the Greek χρῶμα (*chroma*, color) and σῶμα (*soma*, body) due to their property of being stained very strongly by some dyes.
- A chromosome is a singular piece of DNA, which contains many genes, regulatory elements and other nucleotide sequences.
- Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions.

16

Chromosomes

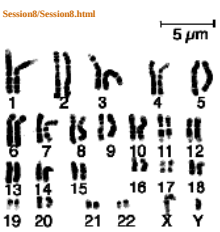
- Chromosomes vary extensively between different organisms.
  - The DNA molecule may be circular or linear, and can contain anything from tens of kilobase pairs to hundreds of megabase pairs.
  - Typically eukaryotic cells have large linear chromosomes and prokaryotic cells have smaller circular chromosomes.
    - In eukaryotes, nuclear chromosomes are packaged by proteins into a condensed structure called chromatin. This allows the very long DNA molecules to fit into the cell nucleus. The structure of chromosomes and chromatin varies through the cell cycle.
- Chromosomes may exist as either duplicated or unduplicated
  - unduplicated chromosomes are single linear strands, whereas duplicated chromosomes (copied during synthesis phase) contain two copies joined by a centromere.
    - Compaction of the duplicated chromosomes during mitosis and meiosis results in the classic four-arm structure.

17

Chromosomes

- In eukaryotes, nucleus contains one or several double stranded DNA molecules organized as chromosomes
- Humans:
  - 22 Pairs of autosomes
  - 1 pair sex chromosomes

Human Karyotype
(the **karyotype** is the characteristic chromosome complement of a eukaryote species)
http://avery.rutgers.edu/WSSP/StudentScholars/Session8/Session8.html



18

Chromosome numbers in some plants

Plant Species	#
Arabidopsis thaliana (diploid)	10
Rye (diploid)	14
Maize (diploid)	20
Einkorn wheat (diploid)	14
Durum wheat (tetraploid)	28
Bread wheat (hexaploid)	42
Potato (tetraploid)	48
Cultivated tobacco (diploid)	48
Adder's Tongue Fern (diploid)	approx 1,440

19

Chromosome numbers (2n) in some animals

Species	#	Species	#
Common fruit fly	8	Guinea Pig	64
Dove	16	Garden snail	54
Earthworm *Octodrilus complanatus*	36	Tibetan fox	36
Domestic cat	38	Domestic pig	38
Lab mouse	40	Lab rat	42
Rabbit	44	Syrian hamster	44
Hare	46	Human	46
Gorillas, Chimpanzees	48	Domestic sheep	54
Elephants	56	Cow	60
Donkey	62	Horse	64
Dog	78	Kingfisher	132
Goldfish	100-104	Silkworm	56

20

Chromosomes

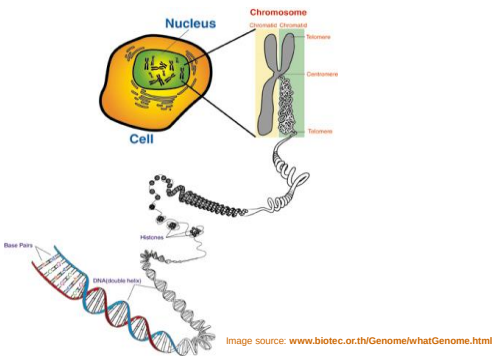


Image source: www.biotec.or.th/Genome/whatGenome.html

21

DNA

- **Deoxyribonucleic acid (DNA)** is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses.
- The main role of DNA molecules is the long-term storage of information.
- DNA is often compared to a set of blueprints or a recipe, or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules.
- The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

22

DNA

- Chemically, DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds.
  - These two strands run in opposite directions to each other and are therefore anti-parallel.
- Attached to each sugar is one of four types of molecules called bases.
  - It is the sequence of these four bases along the backbone that encodes information.
- This information is read using the genetic code, which specifies the sequence of the amino acids within proteins.
- The code is read by copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

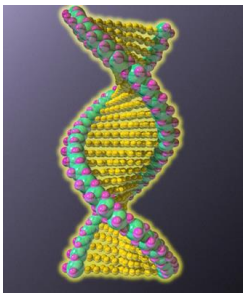
23

DNA

- Within cells, DNA is organized into structures called chromosomes.
- These chromosomes are duplicated before cells divide, in a process called DNA replication.
- Eukaryotic organisms (animals, plants, fungi, and protists) store their DNA inside the cell nucleus, while in prokaryotes (bacteria and archae) it is found in the cell's cytoplasm.
- Within the chromosomes, chromatin proteins such as histones compact and organize DNA.
  - These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

24

DNA is the blueprint for life



- DNA: Deoxyribonucleic Acid
- Every cell in your body has 23 *chromosomes* in the nucleus
- The *genes* in these chromosomes determine all of your physical attributes.
- Single stranded molecule (oligomer, polynucleotide) chain of nucleotides

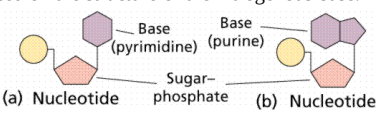
- 4 different nucleotides:
  - Adenine (A)
  - Cytosine (C)
  - Guanine (G)
  - Thymine (T)

25

Nucleotide Bases

- **Nucleotides** are organic compounds that consist of three joined structures:
  - a nitrogenous base,
  - a sugar,
  - a phosphate group.
- The most common nucleotides can be divided into two groups:
  - purines (A and G)
  - pyrimidines (C and T)

  based on the structure of the nitrogenous base.



- The joined sugar is either ribose or deoxyribose.
- Nucleotides are the structural units of RNA and DNA.

Image Source: www.ebi.ac.uk/microarray/ biology_intro.htm

26

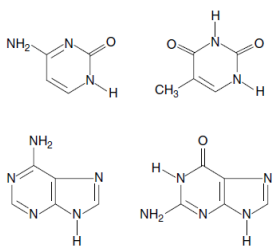
Genetic Code

- The **genetic code** is the set of rules by which information encoded within genetic material (DNA or mRNA sequences) is translated into proteins by living cells.
  - Translation is accomplished by the ribosome, which links amino acids in an order specified by messenger RNA (mRNA), using transfer RNA(tRNA) molecules to carry amino acids and to read the mRNA three nucleotides at a time.
- The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries.
- The code defines how sequences of nucleotide triplets (codons) specify which amino acid will be added next during protein synthesis.
  - With some exceptions, a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid.
  - The vast majority of genes are encoded with a single scheme (see the RNA codon table).
    - That scheme is often referred to as the canonical or standard genetic code, or simply *the genetic code*.

27

DNA and the Genetic Code

The chemical structures of the bases



- The pyrimidine bases,
  - cytosine and thymine,
    - two of the four bases that are constituents of the nucleotide units of DNA.
- The purine bases,
  - adenine and guanine,
    - the other two bases that are constituents of the nucleotide units of DNA.

28

DNA and the Genetic Code

- In most data sources, a typical representation of a DNA molecule or sequence would consist of a sequence of letters, G, C, A, and T, to represent each of the possible four nucleotide (also called "base") pairs that could appear in a double helix DNA strand.
  - Although the DNA molecule is double-stranded, the bases are paired uniquely,
    - A with T and G with C, so that only the bases on one strand need to be represented.

```
CACTGGCATGATCAGGACTCACTGCAGCCTTGACTCCCAGGCTCAGTAGATCCTCCTACCTCAGCCTCTC
GAGTAACTGGGACCACAGGCGAGCATCACCATGCTCAGCTAGTTTTTGTATTTGTAGAGATGAGGTTTCA
CCATATTGCCCAGGCTGGTCTTGAACTCCTGGGCTCAAGCAAGCCACCCACCTTGGCCACCCAAAGTGCT
```

  - a small fragment of the region around a well-known gene associated with breast cancer called BRCA ((the "BReast CAncer gene").

29

Genetic Code

- 4 possible bases (A, C, G, U)
- Each combination of three nucleotides is called a codon
- $4^3=4\times4\times4 = 64$ possible codon sequences
- Not all codons correspond to amino acids;
  - Start codon:
    - AUG
  - Stop codons:
    - UAA, UAG, UGA
- 61 codons to code for amino acids (AUG as well)
  - 20 amino acids – redundancy in genetic code
  - For most of the amino acids, there are several codons that represent the same amino acid.
    - For example, the amino acid lysine is represented by two codons, AAA and AAG, and leucine is represented by any one of six.

30

20 Amino Acids

- Glycine (G, GLY)
- Alanine (A, ALA)
- Valine (V, VAL)
- Leucine (L, LEU)
- Isoleucine (I, ILE)
- Phenylalanine (F, PHE)
- Proline (P, PRO)
- Serine (S, SER)
- Threonine (T, THR)
- Cysteine (C, CYS)
- Methionine (M, MET)
- Tryptophan (W, TRP)
- Tyrosine (T, TYR)
- Asparagine (N, ASN)
- Glutamine (Q, GLN)
- Aspartic acid (D, ASP)
- Glutamic Acid (E, GLU)
- Lysine (K, LYS)
- Arginine (R, ARG)
- Histidine (H, HIS)
- START: AUG
- STOP: UAA, UAG, UGA

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

31

Amino Acids

Letter	Abbreviation	Full name	Codons
A	Ala	Alanine	GCA GCC GCG GCT
C	Cys	Cysteine	TGC TGT
D	Asp	Aspartate	GAC GAT
E	Glu	Glutamate	GAA GAG
F	Phe	Phenylalanine	TTC TTT
G	Gly	Glycine	GGA GGC GGG GGT
H	His	Histidine	CAC CAT
I	Ile	Isoleucine	ATA ATC ATT
K	Lys	Lysine	AAA AAG
L	Leu	Leucine	TTA TTG CTA CTC CTG CTT

32

Amino Acids

M	Met	Methionine	ATG
N	Asn	Asparagine	AAC AAT
P	Pro	Proline	CCA CCC CCG CCT
Q	Gln	Glutamine	CAA CAG
R	Arg	Arginine	AGA AGG CGA CGC CGG CGT
S	Ser	Serine	AGC AGT TCA TCC TCG TCT
T	Thr	Threonine	ACA ACC ACG ACT
V	Val	Valine	GTA GTC GTG GTT
W	Trp	Tryptophan	TGG
Y	Tyr	Tyrosine	TAC TAT

33

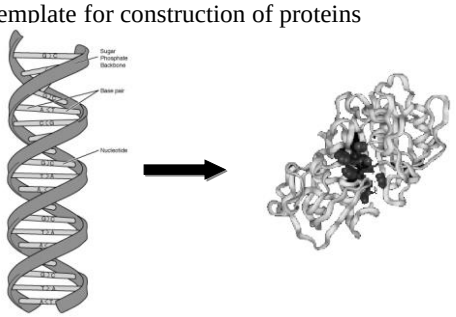
Amino Acids

- Amino acids are the basic structural building units of proteins.
  - **Hydrophilic** amino acids are water soluble
  - **Hydrophobic** are not water soluble
- They form short polymer chains called peptides or longer chains called either polypeptides or proteins.
- The process of such formation from an mRNA template is known as translation, which is part of protein biosynthesis.
- Twenty amino acids are encoded by the standard genetic code and are called proteinogenic or standard amino acids.
- Other amino acids contained in proteins are usually formed by post-translational modification, which is modification after translation in protein synthesis.
  - These modifications are often essential for the function or regulation of a protein.

34

How does the code work?

- Template for construction of proteins

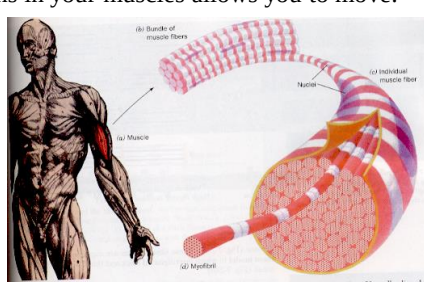


35

Proteins: Molecular machinery

- Proteins in your muscles allows you to move:

myosin and actin



36

Proteins: Molecular machinery

- Protein
  - One of the most important classes of constituents of living organisms
- Many hundreds of thousands of proteins found in living organisms have been named, catalogued, and their properties recorded.
  - These properties include
    - the function(s) of the protein,
    - the gene that codes for it,
    - diseases that may relate to its absence or mutation.

37

Proteins: Molecular machinery

- Proteins have many different roles in cells and tissues.
  - They can serve as enzymes to facilitate biochemical reactions,
    - such as metabolism of glucose.
  - They can regulate other processes by serving as signals.
- Proteins also are components of cell and tissue structure.

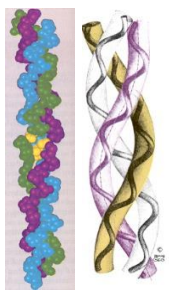
38

Proteins: Molecular machinery

- **Proteins** are large organic compounds made of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues.
  - The word *protein* comes from the Greek word *πρώτειος* (*proteios*) "primary".
- The sequence of amino acids in a protein is defined by a gene and encoded in the genetic code.
- Proteins are essential parts of organisms and participate in every process within cells.

39

Proteins: Molecular machinery



- Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism (digestion, catalysis)
- Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape

40

Proteins: Molecular machinery

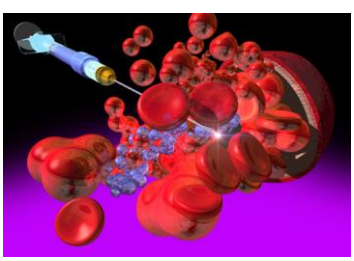


Image source: Crane digital, http://www.cranedigital.com/

- Other proteins are important in cell signaling (hormones, kinases), immune responses, cell adhesion, and the cell cycle.
- Transport (energy, oxygen)

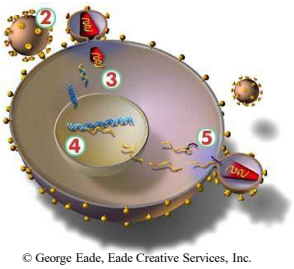
41

Proteins: Molecular machinery

- Proteins are also necessary in animals' diets, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food.
- Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism.

42

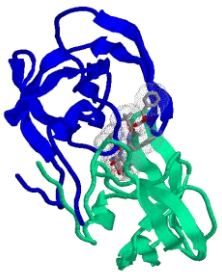
Example Case: HIV Protease



© George Eade, Eade Creative Services, Inc.
http://whyfiles.org/035aids/index.html

1. Exposure & infection
2. HIV enters your cell
3. Your own cell reads the HIV "code" and creates the HIV proteins.
4. New viral proteins prepare HIV for infection of other cells.

HIV Protease as a drug target



HIV Protease + Peptidyl inhibitor (1A8G.PDB)

- Many drugs bind to protein active sites.
- This HIV protease can *no longer* prepare HIV proteins for infection, because *an inhibitor is already bound* in its active site.

Protease: an enzyme which breaks down proteins and peptides.

Drug Discovery

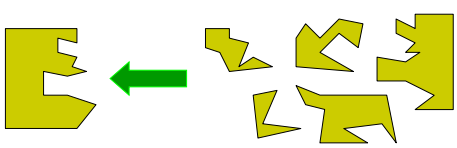
- Target Identification
  - What protein can we attack to stop the disease from progressing?
- Lead discovery & optimization
  - What sort of molecule will bind to this protein?
- Toxicology
  - Does it kill the patient?
  - Does it have side effects?
  - Does it get to the problem spots?

Drug discovery: past & present

- Put some of the infectious agent into thousands of tiny wells
- Add a known drug *lead compound* into each well.
  - Try nearly *every drug lead* known.
- See which ones kill the agent…
  - Too small to see, so we have to use chemical tests called *assays*

Finding drug leads

- Once we have a *target*, how do we find some compounds that might bind to it?
- The old way: exhaustive screening
- The new way: computational screening!



Problems in Bioinformatics

- Genomics
  - Gene finding
  - Annotation
    - Sequence alignment and database search
  - Functional genomics
    - Microarray expression, "gene chips"
- Proteomics
  - Structure prediction
    - Comparative modeling
  - Function prediction
- Structural bioinformatics
  - Molecular docking, screening, etc.

Genomics

- **Genomics** is the study of an organism's entire genome.
- The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts.
- For the United States Environmental Protection Agency,
  - "the term genomics encompasses a broader scope of scientific inquiry and associated technologies than when genomics was initially considered.
    - A genome is the total of all an individual organism's genes.
  - Thus, genomics is the study of all the genes of a cell, or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) levels."

Proteomics

- **Proteomics** is the large-scale study of proteins, particularly their structures and functions.
- Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells.
- The term "proteomics" was coined to make an analogy with genomics, the study of the genes.
  - The word "proteome" is a blend of "*prote*in" and "*genome*".
- The proteome is the entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or system.
  - This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes.

Structural bioinformatics

- **Structural bioinformatics** refers to the branch of bioinformatics which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as proteins.
- The term *structural* has the same meaning as in structural biology, and structural bioinformatics can be seen as computational structural biology.

RNA

- **RiboNucleic Acid** (**RNA**) is a nucleic acid, that is,
  - it consists of a long chain of nucleotide units.
  - Each nucleotide consists of a nitrogenous base, a ribose sugar, and a phosphate.
- RNA is very similar to DNA, but differs in a few important structural details:
  - in the cell, RNA is usually single-stranded, while DNA is usually double-stranded;
  - RNA nucleotides contain ribose while DNA contains deoxyribose (a type of ribose that lacks one oxygen atom);
  - RNA has the base uracil rather than thymine that is present in DNA.
- RNA is transcribed from DNA by enzymes called RNA polymerases and is generally further processed by other enzymes.
- RNA is central to the synthesis of proteins.
- There are RNAs with other roles – in particular regulating which genes are expressed, but also as the genome of most viruses.

mRNA

- A type of RNA called messenger RNA carries information from DNA to structures called ribosomes.
  - These ribosomes are made from proteins and ribosomal RNAs, which come together to form a molecular machine that can read messenger RNAs and translate the information they carry into proteins.
- Linear molecule encoding genetic information copied from DNA molecules
- **Transcription**: process in which DNA is copied into an RNA molecule

mRNA processing

- Eukaryotic genes can be pieced together
  - Exons: coding regions
  - Introns: non-coding regions
- mRNA processing removes introns, splices exons together
- Processed mRNA can be translated into a protein sequence

mRNA Processing

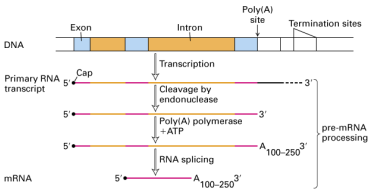


Image source: http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

tRNA

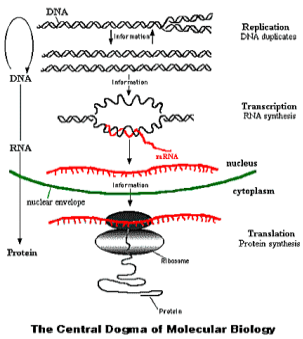
- **Transfer RNA** (**tRNA**) is a small RNA (usually about 74-95 nucleotides)
  - It transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation.
  - It has a 3' terminal site for amino acid attachment. This covalent linkage is catalyzed by an aminoacyl tRNA synthetase.
  - It also contains a three base region called the anticodon that can base pair to the corresponding three base codon region on mRNA.
  - Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.
- Well-defined three-dimensional structure
- Critical for creation of proteins

tRNA

- Amino acid attached to each tRNA
- Determined by 3 base anticodon sequence (complementary to mRNA)

- **Translation**:
  - process in which the nucleotide sequence of the processed mRNA is used in order to join amino acids together into a protein with the help of ribosomes and tRNA

Central Dogma



The Central Dogma of Molecular Biology

Gene and Genome

- **Gene** is the physical and functional unit of heredity that carries information from one generation to the next
- A **genome** is an organism's complete set of DNA, including all of its genes.
  - Each **genome** contains all of the information needed to build and maintain that organism.
- Genome size and number of genes does not necessarily determine organism complexity

Genome Comparison

ORGANISM	CHROMOSOMES	GENOME SIZE	GENES
Homo sapiens (Humans)	23	3,200,000,000	~ 30,000
Mus musculus (Mouse)	20	2,600,000,000	~30,000
Drosophila melanogaster (Fruit Fly)	4	180,000,000	~18,000
Saccharomyces cerevisiae (Yeast)	16	14,000,000	~6,000
Zea mays (Corn)	10	2,400,000,000	???

Transcriptome and Proteome

- Transcriptome
  - complete collection of all possible mRNAs (including splice variants) of an organism.
  - regions of an organism's genome that get *transcribed* into messenger RNA.
  - can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes.
- Proteome
  - the complete collection of proteins that can be produced by an organism.
  - can be studied either as static (sum of all proteins possible) or dynamic (all proteins found at a specific time point) entity

61

What Can Be Represented in a Computer?

- In order for computer programs to deal with biomedical data,
  - the data must be encoded according to some plan so that
    - the binary numbers in the computer's memory represent numeric data or represent text or possibly more abstract kinds of data.
- Many encoding schemes have been used for biomedical data.

62

What Can Be Represented in a Computer?

- Examples include biomolecular sequence data (DNA and proteins), laboratory data from tests done on blood samples and other substances obtained from patients in a clinic or hospital, and medical image data.
- In addition, everyone needs methods for searching and sorting through the vast collections of bibliographic reference data now available such as the MEDLINE database of journal articles and other informational items.
- The entries in such bibliographic databases are definitely not numeric but are indexed by keywords, and the search methods used depend on being able to manipulate lists and complex structures.

63

Representing DNA in Computer Programs

- In a typical computerized encoding, each letter is represented by its ASCII code,
  - each occupies 8 bits in a text file
    - although ASCII is a 7-bit code, it is usual to use 8 bit "bytes".
- We can represent base sequences as letter sequences or long strings.
- Another way to represent such sequences is to represent each nucleotide (or base) as a symbol in a list.

64

Representing DNA in Computer Programs

- The entire sequence then becomes a list of symbols and looks like this:
  - (C A C T G G C A T G A T C A G G A C T C A C T G C A G C C T T G A C T C C C A G G C T C A G T A G A T C C T C C T A C C T C A G C C T C T C G A G T A A C T G G G A C C A C A G G C G A G C A T C A C C A T G C T C A G C T A G T T T T T G T A T T T G T A G A G A T G A G G T T T C A C C A T A T T G C C C A G G C T G G T C T T G A A C T C C T G G G C T C A A G C A A G C C A C C C A C C T T G G C C A C C C A A A G T G C T)

65

The Nature and Representation of Biomedical Data

- Large digital data repositories are available containing information about proteins,
  - the UniProt / Swiss-Prot Knowledge Base,
    - a project of the Swiss Institute for Bioinformatics.
  - The UniProt data are downloadable in a variety of formats from the ExPASy web site http://www.expasy.ch,
    - maintained by the Swiss Institute for Bioinformatics.
- Next slide shows an abbreviated excerpt from a Swiss-Prot entry for the precursor protein from which human insulin is formed.

66

The Nature and Representation of Biomedical Data

```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DE   Insulin precursor [Contains: Insulin B chain; Insulin A chain].
GN   Name=INS;
OS   Homo sapiens (Human).
DR   GO; GO:0006953; P:acute-phase response; NAS:UniProtKB.
DR   GO; GO:0046631; P:alpha-beta T cell activation; IDA:UniProtKB.
DR   GO; GO:0008219; P:cell death; NAS:UniProtKB.
DR   GO; GO:0007267; P:cell-cell signaling; IC:UniProtKB.
DR   GO; GO:0006006; P:glucose metabolic process; TAS:ProtInc.
DR   GO; GO:0015758; P:glucose transport; IDA:UniProtKB.
FT   SIGNAL      1     24
FT   PEPTIDE     25    54       Insulin B chain.
FT                              /FTId=PRO_0000015819.
FT   PROPEP      57    87       C peptide.
FT                              /FTId=PRO_0000015820.
FT   PEPTIDE     90    110      Insulin A chain.
FT                              /FTId=PRO_0000015821.
FT   DISULFID    31    96       Interchain (between B and A chains).
FT   DISULFID    43    109      Interchain (between B and A chains).
FT   DISULFID    95    100
FT   VARIANT     34    34       H -> D (in familial hyperproinsulinemia;
FT                              Providence).
FT                              /FTId=VAR_003971.
FT   HELIX       35    43
FT   HELIX       44    46
FT   STRAND      48    50
FT   STRAND      56    58
FT   STRAND      74    76
FT   HELIX       79    81
FT   TURN        84    86
FT   HELIX       91    95
FT   HELIX       105   108
SQ   SEQUENCE   110 AA;  11981 MW;  C2C3B23B85E520E5 CRC64;
     MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY TPKTRREAED
     LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC SLYQLENYCN
```

- For a complete explanation of the individual field items, consult the documentation available at the ExPASy web site

The Nature and Representation of Biomedical Data

- The DR records are cross-references,
  - in this case to the Gene Ontology (GO).
    - It is useful to be able to have a computer program look up these cross-references so that information can then be used in combination with the data shown here.
- The FT records are feature descriptions.
  - Some of the features shown are the places in the amino acid sequence where different types of structures occur, such as α-helix structures, β strands, and turns.
    - In this record, the locations of disulfide linkages are also reported.

The Nature and Representation of Biomedical Data

- The names following the FT tags are the feature types.
- The numbers are the sequence start and end points for each feature.
  - This particular entry describes a polypeptide that is a precursor for the insulin protein.
  - The molecule folds up, forming the disulfide bonds indicated, and the section marked PROPEP in the FT records is spliced out, leaving the two PEPTIDE sections, linked by the disulfide bridges.
- Finally, the SQ section contains the actual sequence of amino acids, one letter for each.

The Nature and Representation of Biomedical Data

- Many proteins function as enzymes,
  - chemical compounds that facilitate chemical reactions
    - Many biologically important reactions do not proceed without the presence of the corresponding enzymes.
  - So, an important piece of information about a protein is its function.
    - Is it an enzyme, and what type of enzyme is it?
    - If it is an enzyme, what reaction(s) does it facilitate?
- Next slide shows an example, a small excerpt from the Enzyme database, at the ExPASy web site.

The Nature and Representation of Biomedical Data

```
ID   1.1.1.39
DE   Malate dehydrogenase (decarboxylating).
AN   Malic enzyme.
AN   Pyruvic-malic carboxylase.
CA   (S)-malate + NAD(+) = pyruvate + CO(2) + NADH.
CC   -!- Does not decarboxylates added oxaloacetate.
PR   PROSITE; PDOC00294;
DR   P37224, MAOM_AMAHP;  P37221, MAOM_SOLTU;  P37225, MAOM_SOLTU;
```

- The line beginning with ID is the Enzyme Commission number, unique to each entry.
- The DE line is the official name,
- The AN lines are alternate names or synonyms.
  - This enzyme catalyzes the reaction that removes a carboxyl group from the malate molecule, leaving a pyruvate molecule, and in the process also converting an NAD+ molecule to NADH.

The Nature and Representation of Biomedical Data

- The NAD+ and NADH molecules are coenzymes,
  - molecules that participate in the reaction
- The reaction catalyzed by malate dehydrogenase is described in a kind of stylized symbolic form on the line beginning with CA.
- The CC line is a comment,
  - not meant for use by a computer program.
- The PR line is a cross-reference to the Prosite database,
  - which has other information about proteins
- The DR line is a set of cross-references to entries in the Swiss-Prot database
  - where the sequences corresponding to various versions of this protein may be found.

Biological databases: two perspectives

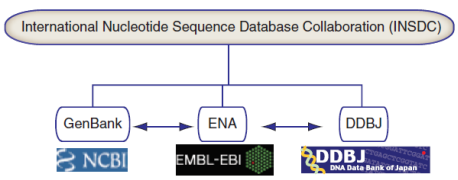
- We might want to study one gene, protein, DNA molecule, or other type of object in a database.
  - For example, for human beta globin there is a gene (HBB), a protein sequence, a protein structure, and entries for various kinds of variation.
- We can think about large groups, such as all the globin genes in the human genome, or all the known HBB variants.
  - Or we might want to study a set of 100 genes previously implicated in a disease (e.g. autism) to assess their variation in patient samples.
- These are different ways of thinking about searching databases.

globin: a colorless protein obtained by removal of heme from a conjugated protein and especially hemoglobin.

Biological databases

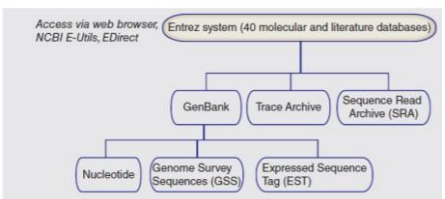
  - How much DNA sequence is stored in public databases?
  - Where are the data stored?
- Three main sites that have been responsible for storing nucleotide sequence data from 1982 to the present (next slide).
  - GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda
  - The European Molecular Biology Laboratory (EMBL)-Bank Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) in Hinxton, England
  - The DNA Database of Japan (DDBJ) at the National Institute of Genetics in Mishima

INSDC coordinates sequence data



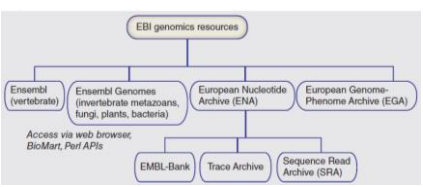
- The nucleotide collections of GenBank at NCBI, EMBL-Bank at the European Bioinformatics Institute, and DDBJ at the DNA Data Bank of Japan are all coordinated by
  - the International Nucleotide Sequence Database Collaboration (INSDC).

National Center for Biotechnology Information (NCBI): organization



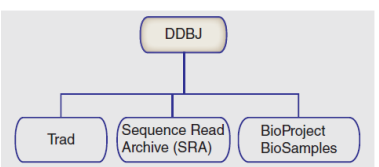
- houses GenBank as part of its Entrez system of 40 moleuclar and literature databases.
- The Trace Archive stores sequence traces, and the Sequence Read Archive (SRA) stores next-generation sequence data.
- GenBank includes separate divisions for nucleotides, genome survey sequences, and expressed sequence tags.

European Bioinformatics Institute (EBI): organization



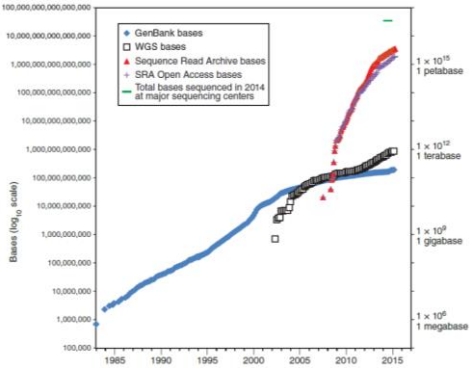
- EBI resources include Ensembl (with a focus on vertebrate genomes), Ensembl Genomes (centralizing data on broader groups of species), the European Nucleotide Archive (ENA), and the European Genome-Phenome Archive (EGA).
- Within ENA, EMBL-Bank includes the same raw sequence data as GenBank at NCBI.
- Similar data are also housed in the Trace Archive and SRA

DNA Database of Japan (DDBJ): organization



- The DDBJ also includes a SRA.
- Its traditional (Trad) division shares the same raw sequence data with GenBank and EMBL-Bank on a daily basis.
- All these various databases can be accessed by web browsing or via programs such as EDirect (for command-line access to Entrez databases).

Growth of DNA sequence in repositories



79

Scales of DNA base pairs

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	Size of a typical coding region of a gene
1,000,000	1 megabase pair	1 Mb	Size of a typical bacterial genome
10^9	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

80

Scales of file sizes

Size	Abbreviation	# bytes	Example
Bytes	--	1	Single text character
Kilobytes	1 kb	10^3	Text file, 1000 characters
Megabytes	1 MB	10^6	Text file, 1m characters
Gigabytes	1 GB	10^9	Size of GenBank: 600 GB
Terabytes	1 TB	10^{12}	Size of 1000 Genomes Project: <500 TB
Petabytes	1 PB	10^{15}	Size of SRA at NCBI: 5 PB
Exabytes	1 EB	10^{18}	Annual worldwide output: >2 EB

81

Taxa represented in GenBank (at NCBI)

Ranks	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	143	140	525	0	808
Bacteria	1,370	2,611	13,331	819	18,131
Eukaryota	20,443	67,606	297,207	22,608	407,864
Fungi	1,550	4,620	29,450	1,128	36,748
Metazoa	14,670	45,517	145,044	11,428	216,659
Viridiplantae	2,622	14,680	113,529	9,789	140,620
Viruses	618	442	2,349	0	3,409
All taxa	22,603	70,806	313,443	23,427	430,279

Taxa is plural of taxon,

– which is any group or rank in a biological classification into which related organisms are classified.

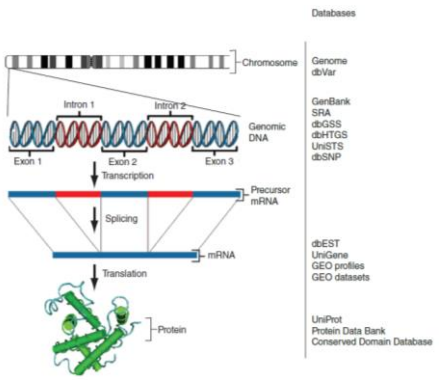
82

Types of data and examples of databases

- The types of data stored in various databases can be conceptualized in terms of the central dogma of biology in which genomic DNA includes protein-coding genes that are transcribed to precursor messenger RNA (mRNA), processed to mature mRNA, and translated to protein (next slide)
- The protein structure is from accession 1HBS.

83

Types of data and examples of databases



84

Top ten organisms for which expressed sequence tags (ESTs) have been sequenced

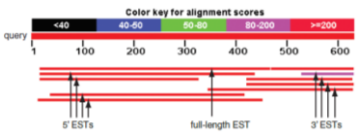
Organism	Common name	Number of ESTs
Homo sapiens	Human	8,704,790
Mus musculus + *domesticus*	Mouse	4,853,570
Zea mays	Maize	2,019,137
Sus scrofa	Pig	1,669,337
Bos taurus	Cattle	1,559,495
Arabidopsis thaliana	Thale Cress	1,529,700
Danio rerio	Zebrafish	1,488,275
Glycine max	Soybean	1,461,722
Triticum aestivum	Wheat	1,286,372
Xenopus (Silurana) tropicalis	Western clawed frog	1,271,480

85

UniGene database: clusters of EST sequences

86

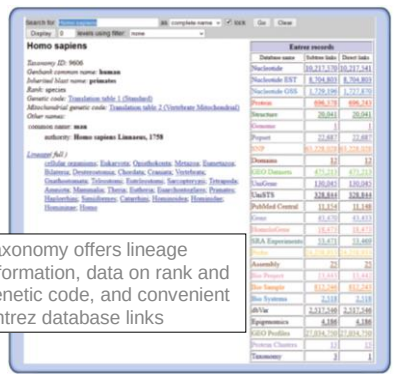
UniGene database: clusters of EST sequences



- ESTs are mapped to a particular gene and to each other.
- The number of ESTs that constitute a UniGene cluster ranges from 1 to over 1000;
  - on average there are 100 ESTs per cluster.
- Sometimes, separate UniGene clusters correspond to distinct regions of a gene (particularly for large genes).
- Here human beta globin (HBB) mRNA (NM_000518.4) was used as a query with BLAST and searched against nine ESTs selected from among >2000 available ESTs.
- Four of them are 5' ESTs, four are 3' ESTs (including a poly(A)+ tail), and one is a full length EST.
- The accession numbers are AA985606.1, AA910627.1, AI089557.1, AI150946.1, R25417.1, R27238.1, R27242.1, R27252.1, R31622.1, R32259.1.

87

Access to NCBI databases via Taxonomy Browser



88

Accession numbers are labels for sequences

- NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.
  - You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.
- DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

89

What is an accession number?

- a label used to identify a sequence.
- a string of letters and/or numbers that corresponds to a molecular sequence.
  - Examples:

DNA	• CH471100.2	GenBank genomic DNA sequence
	• NC_000001.10	Genomic contig
	• rs121434231	dbSNP (single nucleotide polymorphism)
RNA	• AI687828.1	An expressed sequence tag (1 of 184)
	• NM_001206696	RefSeq DNA sequence (from a transcript)
protein	• NP_006138.1	RefSeq protein
	• CAA18545.1	GenBank protein
	• O14896	SwissProt protein
	• 1KT7	Protein Data Bank structure record

A contig is a chromosome map showing the locations of those regions of a chromosome where contiguous DNA segments overlap.

90

NCBI's important RefSeq project: best representative sequences

- RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.
  - RefSeq identifiers include the following formats:
    - Complete genome NC_######
    - Complete chromosome NC_######
    - Genomic contig NT_######
    - mRNA (DNA format) NM_###### e.g. NM_006744
    - Protein NP_###### e.g. NP_006735

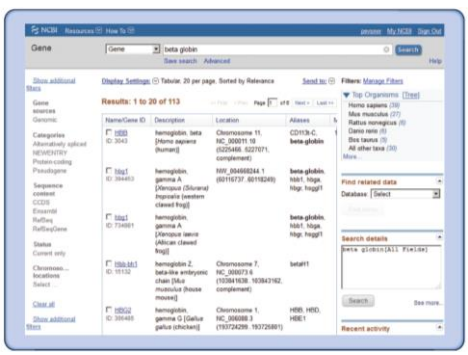
91

Access to sequences: Gene resource at NCBI

- NCBI Gene is a great starting point:
  - It collects key information on each gene/protein from major databases.
  - It covers all major organisms.
- RefSeq provides a curated, optimal accession number for each DNA
  - NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

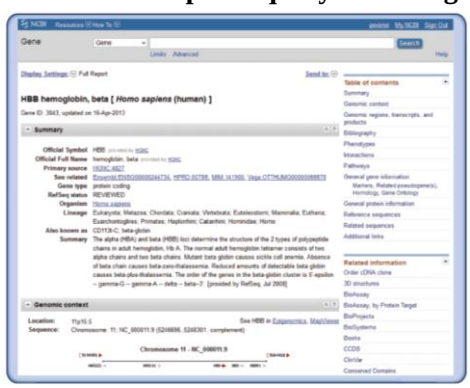
92

NCBI Gene: example of query for beta globin



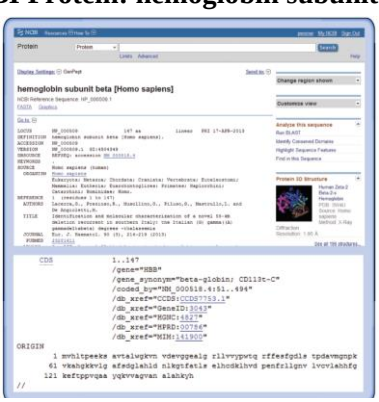
93

NCBI Gene: example of query for beta globin



94

NCBI Protein: hemoglobin subunit beta



95

NCBI Protein: hemoglobin subunit beta in the FASTA format

96

Genome Browsers

- Versatile tools to visualize chromosomal positions (typically on x-axis) with annotation tracks (typically on y-axis).
- Useful to explore data related to some chromosomal feature of interest such as a gene.
- Prominent browsers are at Ensembl, UCSC, and NCBI.
- Many hundreds of specialized genome browsers are available, some for particular organisms or molecule types

97

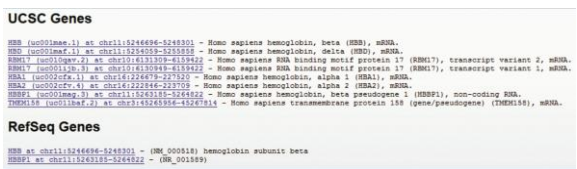
Genome Browsers: UCSC

- https://genome.ucsc.edu/
- Choose the group (e.g. mammal), genome (e.g. human), assembly (e.g. GRCh37 or GRCh38), position and/or search term (e.g. hbb).

- A genome build or assembly (e.g. GRCh37 or GRCh38) refers to a fixed, agreed-upon version of a reference genome.
- Assemblies are typically updated every few years

98

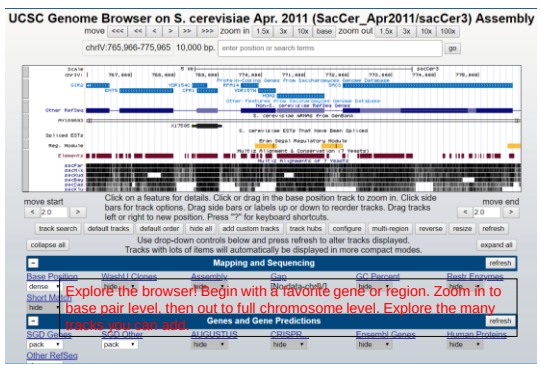
Genome Browsers: UCSC



- When you enter a query such as “hbb” you may have to specify which entry you want, such as the RefSeq version having accession NM_000518.

99

Genome Browsers: UCSC



Explore the browser! Begin with a favorite gene or region. Zoom in to base pair level, then out to full chromosome level. Explore the many tracks you can add.

100

Accessing sequence data for individual genes

- When you search for information about a particular gene, make sure you know the official gene symbol and choose the appropriate species.
  - visit http://www.genenames.org
- Some searches are particularly challenging.
  - For example, there are thousands of histones.
  - Use Boolean operators to limit the search results.
- Searching for HIV-1 proteins, note that there are vast numbers of protein and DNA results (approaching 1 million entries!) but there is only one RefSeq accession.
  - This highlights the usefulness of the RefSeq project.

101

Perspective

- The field of bioinformatics is growing quickly,
  - in part because of the introduction of vast amounts of sequence data.
- There are many databases that store genomic data, and many approaches to extracting information

102