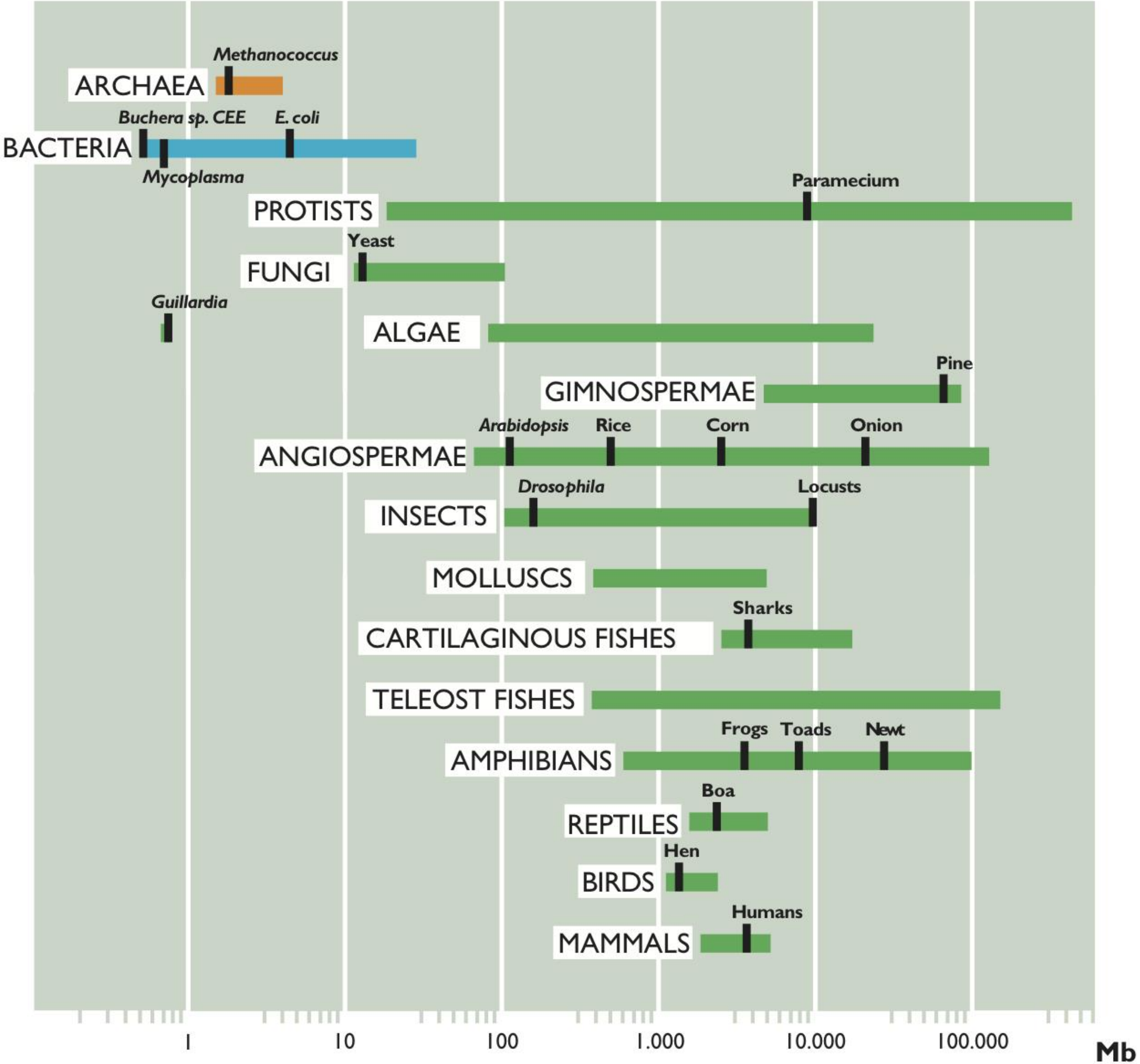
Genome organization

Subha Narayan Rath

Genome, transcriptome, proteomes

- Genome of a typical bacterium like E. coli: 4.6* 10e6 bp with a single DNA molecule
- If extended it would be 2 mm long and it fits into 0.001 mm of diameter of a cell
- Human cells contain 23 pairs of chromosomes and size is 3223*10e6 bp
- Transcriptome: for all the RNA content of the cells
- Proteome: for all the protein content of the cells….in humans only 20000 protein coding genes, but number of proteins are very huge


Methanococcus
ARCHAEA
Buchera sp. CEE
E. coli
BACTERIA
Mycoplasma
Paramecium
PROTISTS
Yeast
FUNGI
Guillardia
ALGAE
Pine
GIMNOSPERMAE
Arabidopsis
Rice
Corn
Onion
ANGIOSPERMAE
Drosophila
Locusts
INSECTS
MOLLUSCS
Sharks
CARTILAGINOUS FISHES
TELEOST FISHES
Frogs
Toads
Newt
AMPHIBIANS
Boa
REPTILES
Hen
BIRDS
Humans
MAMMALS
1
10
100
1.000
10.000
100.000
Mb

Twenty-One Amino Acids

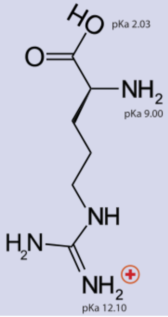
⊕ Positive ⊖ Negative

• Side chain charge at physiological pH 7.4

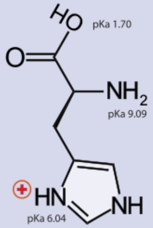
A. *Amino Acids with Electrically Charged Side Chains*

Positive

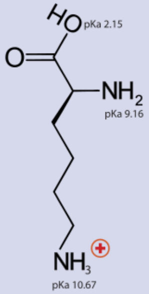
Arginine (Arg) R



Histidine (His) H

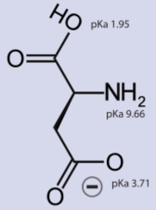


Lysine (Lys) K

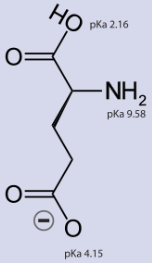


Negative

Aspartic Acid (Asp) D

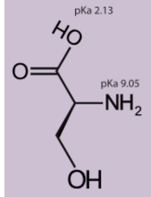


Glutamic Acid (Glu) E

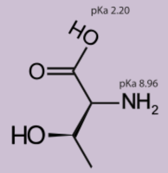


B. *Amino Acids with Polar Uncharged Side Chains*

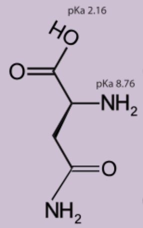
Serine (Ser) S



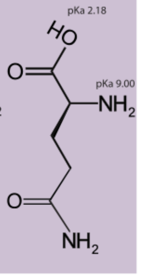
Threonine (Thr) T



Asparagine (Asn) N

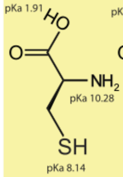


Glutamine (Gln) Q

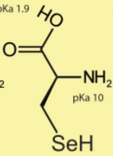


C. *Special Cases*

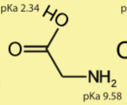
Cysteine (Cys) C



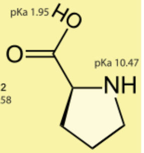
Selenocysteine (Sec) U



Glycine (Gly) G

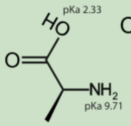


Proline (Pro) P

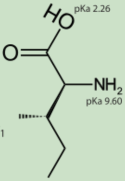


D. *Amino Acids with Hydrophobic Side Chain*

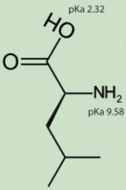
Alanine (Ala) A



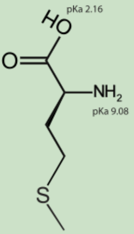
Isoleucine (Ile) I



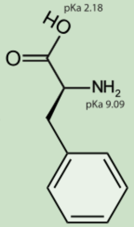
Leucine (Leu) L



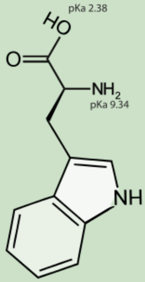
Methionine (Met) M



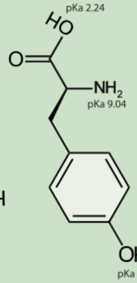
Phenylalanine (Phe) F



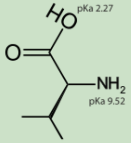
Tryptophan (Trp) W



Tyrosine (Tyr) Y



Valine (Val) V



pKa Data: CRC Handbook of Chemistry, v.2010

Dan Cojocari, Department of Medical Biophysics, **University of Toronto**, 2010

Problem 3.1

- The overall base composition of E.COLI genome is A=T=49.2%. In a random sequence of 4 639 221 (normal bp of E. coli) with these proportions, what is the expected number of occurrences of the sequence CTAG?

Genes

- They may appear in either strand of DNA
- In bacteria: functional unit of genetic sequence are
  - 3N nucleotides presenting
  - N amino acids of a protein
- In eukaryotes: one gene is split into separate segments in genetic DNA
  - EXON: Expressed region
  - INTRON: Intervening region
  - Cellular machinery splices together the initial mRNA to make the product
- Control mechanism may turn genes ON or OFF
  - Or regulate gene expression more finely
  - Cascade of controls respond to conc. of nutrients, to stress, or to control cell cycle

Control regions of DNA lie near the segments coding for proteins

- They contain signal sequences that serve as binding sites for molecules that causes transcription like TFs
- Or, bind regulatory molecules that block transcription
- Bacterial genes: there are OPERONS in line with the genes
- In Eukaryotes: epigenetic signals like
  - DNA methylation
  - Histone modification
- …they direct tissue specific expression of developmentally regulated genes
- DNA methylation is stable during tissue differentiation surviving cell division
- Reversible chemical modification of histones render the transcription sites more or less accessible
- ….so it is like 3.2 Gb of data in a mass storage device..

Proteomics

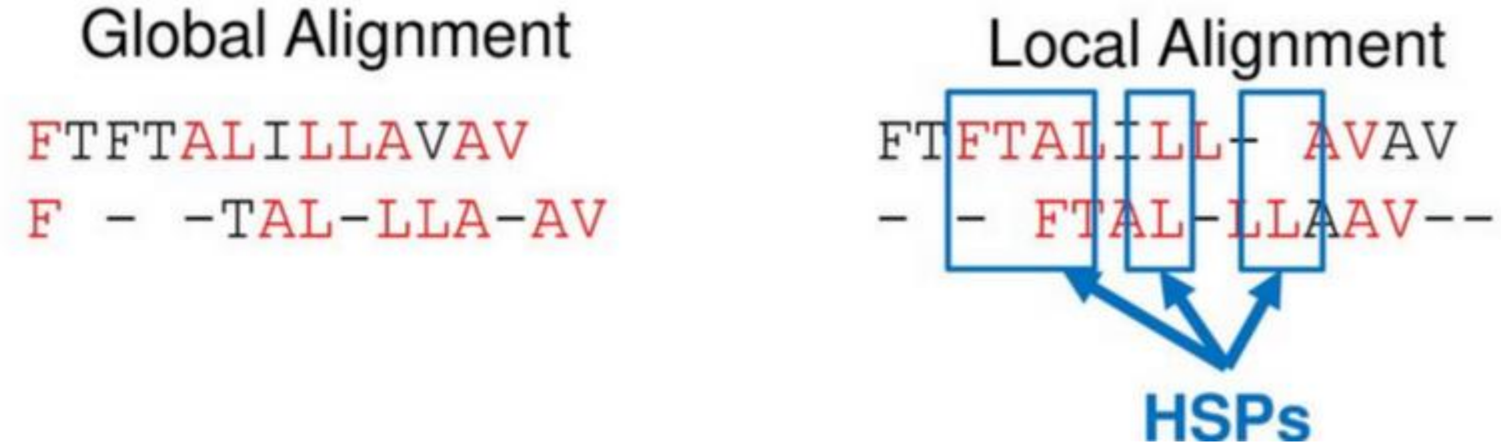
- 2 methods can measure:
  - High resolution 2D PAGE
  - Mass spectrometric techniques
- Previously by direct sequencing of proteins; but now by translation of DNA sequences…new protein sequence data are determined; but the later is always hypothetical unless experimentally proven
- The pattern recognition programs that can do it will be subject to 3 types of errors:
- 1. protein sequences might be missed entirely, may incorrectly spliced, might be from exons in different ways of combination and in different tissues can't be predicted. If mRNA is edited before translation, it can't be known.
- 2. No clue for quarternary structure, prosthetic group binding, patterns of disulphide bridges
- 3. Post translational modifications: covalent alterations within a cell, addition of a ligand, or cleavage of a protein to an active form
- Inteins are proteins that have self splicing activity compared to done by proteases

Transcriptomics

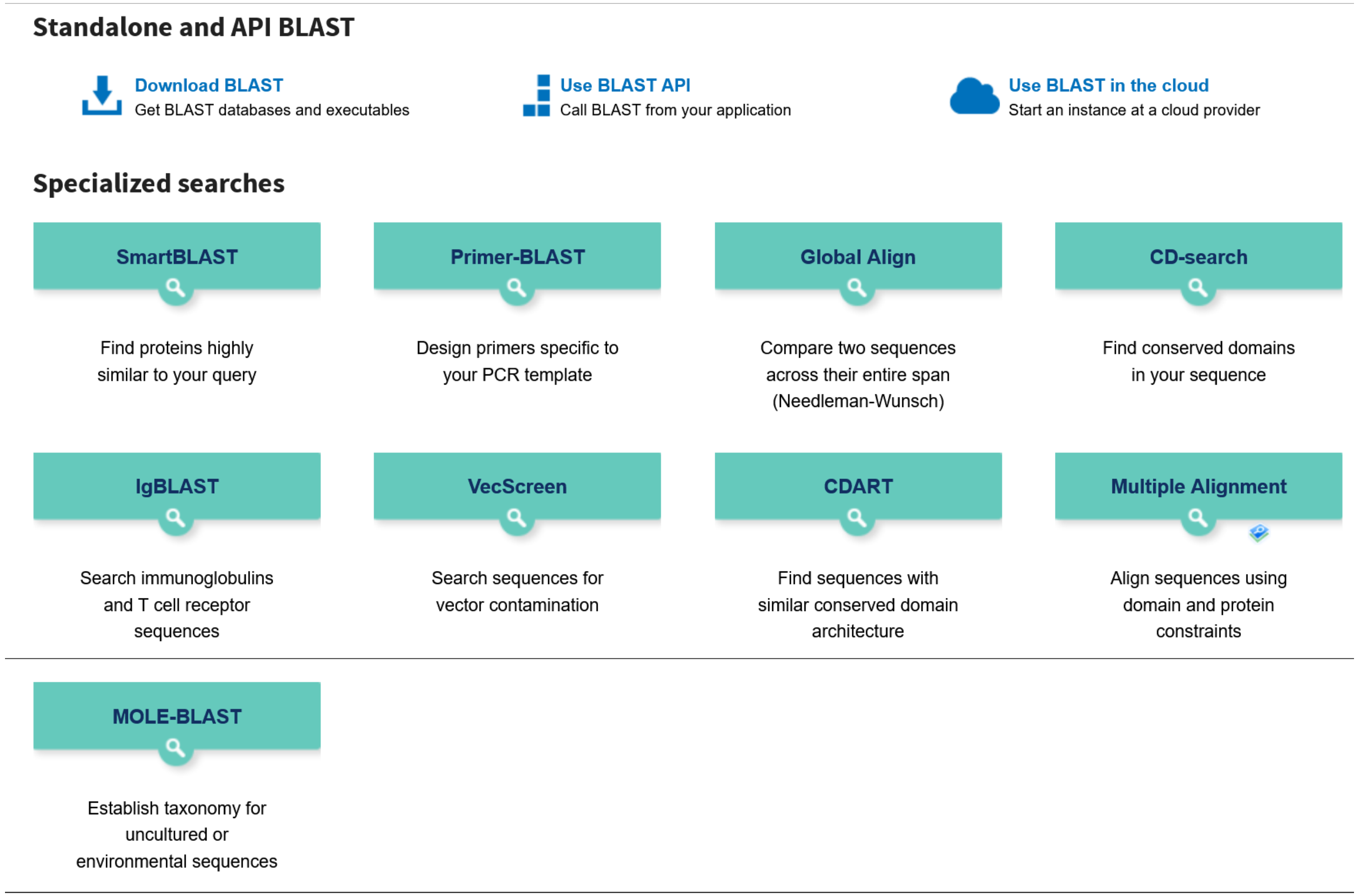
- Many RNA transcripts are not protein coding
- Transient: mRNA
- Stable: rRNA
- 1. RNA seq methods by RNA to cDNA and sequencing methods or real time PCR
- 2. RNA can be sequenced directly

BLAST = Basic local alignment search tool

- Rapidly compare a query sequence to a database of subject sequences
- Generate alignments between them= the quality of which is by ALIGNMENT SCORE
- Return alignments that pass user defined score and statistical significance thresholds
- BLAST uses local alignment to find high scoring segment pairs (HSP) between two sequences
- BLAST HIT: A Subject sequence that is aligned to the query



https://blast.ncbi.nlm.nih.gov/Blast.cgi



Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

IgBLAST
Search immunoglobulins and T cell receptor sequences

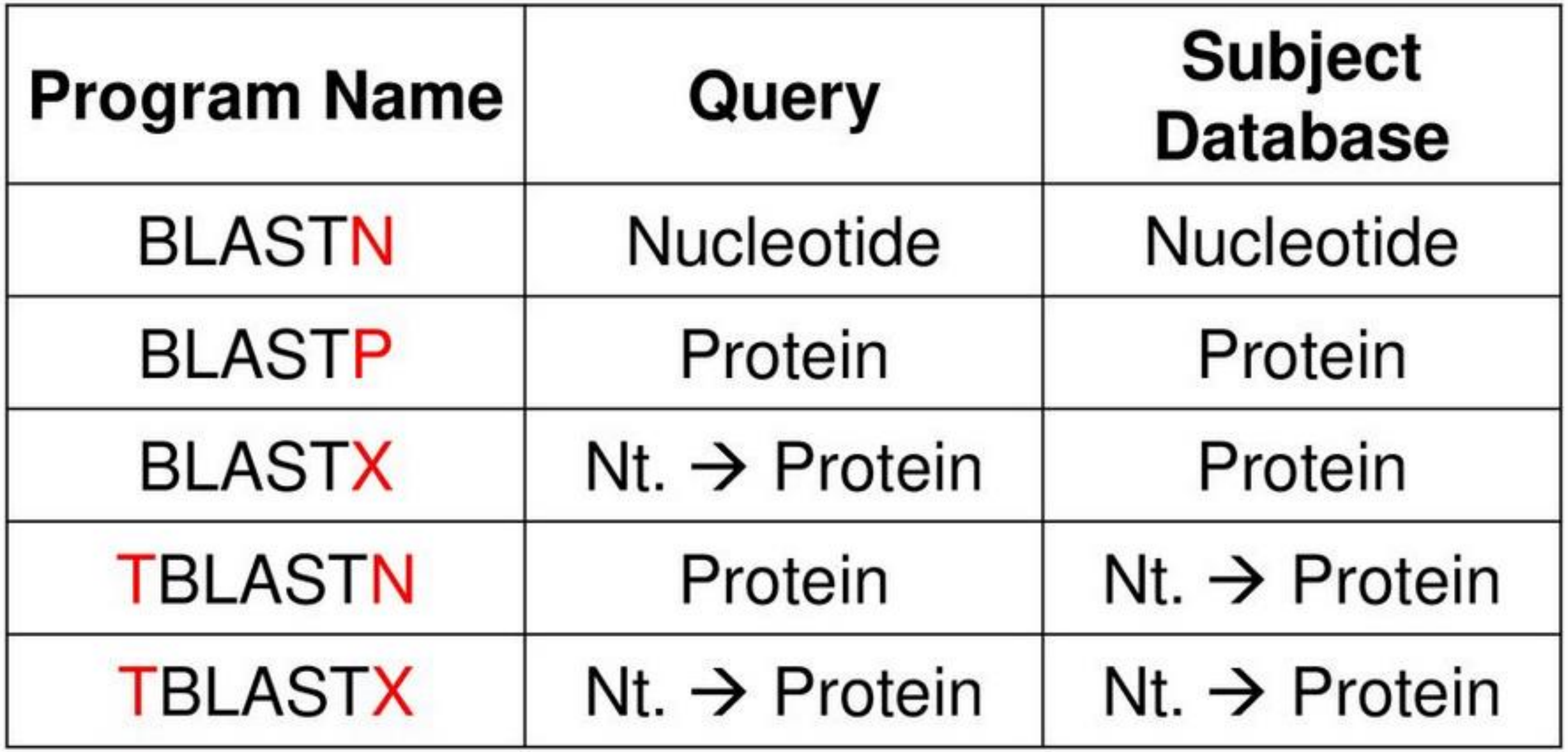
VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

Multiple Alignment
Align sequences using domain and protein constraints

MOLE-BLAST
Establish taxonomy for uncultured or environmental sequences

Common blast programs



Program Name	Query	Subject Database
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nt. → Protein	Protein
TBLASTN	Protein	Nt. → Protein
TBLASTX	Nt. → Protein	Nt. → Protein

BLASTN is for:

- Mapping oligonucleotides to genome
- Comparing DNA from closely related species
- Aligning expressed sequence tags to a genome

BLASTP is for:

- Exploring protein function
- Initial discovery for conserved domains

BLASTX

- Nucleotide query is translated into all 6 reading frames
  - 3 reading frames in + strand
  - 3 reading frames in – strand
- Each reading frame is compared to a protein database
- It is used for:
  - Gene finding in genomic DNA (Annotations)
  - Annotating ESTs

TBLASTN

- Query is a protein sequence
- Nucleotide database is translated into 6 RFs
- The query is then compared to each RF
- It is used for
  - Mapping a protein to genome database
  - Finding ESTs that map to a protein sequence
  - Finding RNA Seq reads that map to a protein sequence

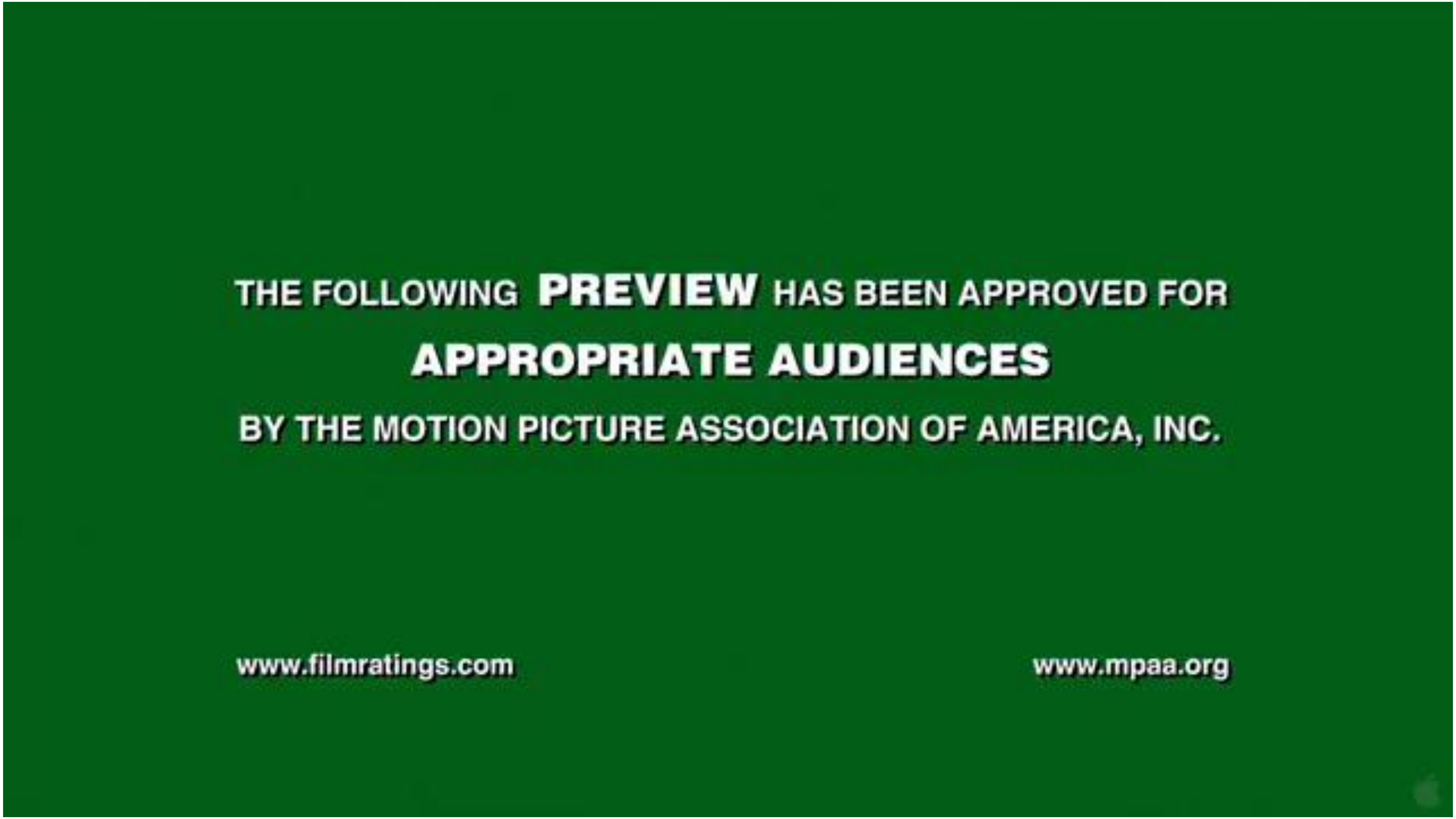
TBLASTX

- BOTH query and subject database (both are nucleotides) are converted to 6 RFs and compared
- It is best used for:
  - Comparing the nucleotide sequence from distantly related species
  - Identify coding regions in ESTs
  - Sensitive but expensive

Problem 4.1

- What are characteristic features of a good primer?
- Make primers for :Protein A, B, C, D as per groups by real time PCR BLAST it to human genome (nr for different animal) databases.
- Which chromosome it lies? What is the function of it and write in a paragraph?
- What are self and self 3` complementarity and how to read the scores? Which primers you would select based on this?
- Find the unknown values of BLAST OUTPUT page and describe them and interpret the results.
- If you are working in rat, mice and human stem cells, is there a problem of rat/mice cell contamination?
- ********************
- ….adenyl cyclase, beta actin, collagen IV, desmin.

Thanks!! Any questions!!



THE FOLLOWING **PREVIEW** HAS BEEN APPROVED FOR

APPROPRIATE AUDIENCES

BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.

www.filmratings.com

www.mpaa.org

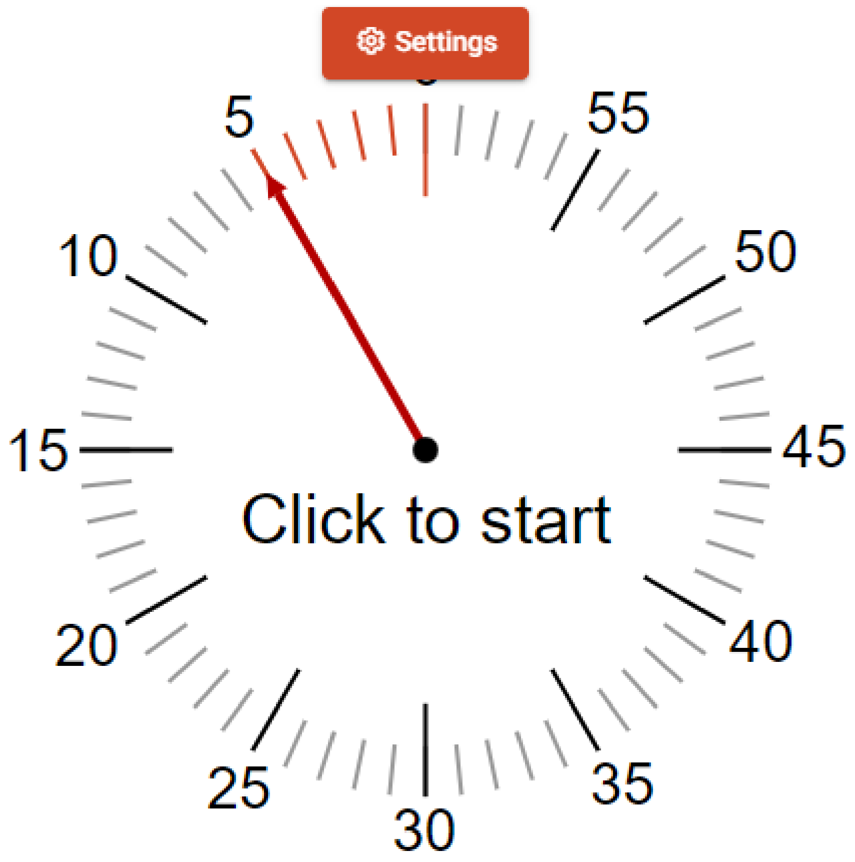
High resolution maps

- Previously, genes are only visible portion of genome
- Now any feature of DNA that vary among individuals can serve as markers
- 1. VNTRs: variable-number tandem repeats also called mini-satellites, 10-100 bp long same sequence repeated
- They can be mapped to disease phenotypes, and used for genetic fingerprinting for criminal cases
- 2. STRPs: short tandem-repeat polymorphisms also called micro-satellites2-5bp repeated for 10-30 times
- They are more uniformly distributed over the genome
- Panels of microsatellite markers: for identification of genes
- 3. Contig maps: series of overlapping DNA clones stored in yeast or bacterial cells as YACs or BACs

4. Sequence tagged site (STS)

YACs	10e6 bp	Human genome in 10,000 YACs
BACs	250,000 bp (1/4th of above)	Greater stability and ease of handling, so preferred

- Short sequenced region of DNA, 200-600 bp that appears in a unique location of genome
- STS can be mapped into the genome by PCR to test its presence in a contig map to the gene of interest
- One type of STSs arise from expressed sequence tag (EST) from cDNA from mRNA of expressed gene
- This is only the exons of the gene, spliced together to form the sequence which encode the protein



Click to start

Other stuff

- For bioperl program:
- https://bioperl.org/
- https://global.oup.com/uk/orc/biosciences/bioinf/leskbioinf4e/