Bioinformatics

Introduction to genomics and proteomics I

ulf.schmitz@informatik.uni-rostock.de

Bioinformatics and Systems Biology Group

www.sbi.informatik.uni-rostock.de

Outline

Genomics/Genetics

1. The tree of life
    - Prokaryotic Genomes
        - Bacteria
        - Archaea
    - Eukaryotic Genomes
        - Homo sapiens
2. Genes
    - Expression Data

Genomics - Definitions

✍ **Genetics:** is the science of *genes*, *heredity*, and the *variation* of organisms.

- Humans began applying knowledge of genetics in prehistory with the domestication and breeding of plants and animals.
- In modern research, genetics provides *tools* in the investigation of the *function* of a particular gene, e.g. analysis of *genetic interactions*.

✍ **Genomics:** attempts the study of large-scale genetic patterns across the genome for a given species. It deals with the systematic use of genome information to provide answers in biology, medicine, and industry.

- *Genomics* has the potential of offering new therapeutic methods for the treatment of some diseases, as well as new diagnostic methods.
- Major tools and methods related to genomics are bioinformatics, genetic analysis, measurement of gene expression, and determination of gene function.

Genes

- a *gene* coding for a *protein* corresponds to a sequence of *nucleotides* along one or more regions of a *molecule* of DNA
- in species with double stranded DNA (dsDNA), genes may appear on either strand
- bacterial genes are continuous regions of DNA

bacterium:

- a string of 3*N* nucleotides encodes a string of *N* amino acids
- or a string of *N* nucleotides encodes a structural RNA molecule of *N* residues

eukaryote:

- a gene may appear split into separated segments in the DNA
- an *exon* is a stretch of DNA retained in mRNA that the *ribosomes* translate into protein

Genomics

Genome size comparison

	Species	Chrom.	Genes	Base pairs
	Human (Homo sapiens)	46 (23 pairs)	28-35,000	3.1 billion
	Mouse (Mus musculus)	40	22.5-30,000	2.7 billion
	Puffer fish (Fugu rubripes)	44	31,000	365 million
	Malaria mosquito (Anopheles gambiae)	6	14,000	289 million
	Fruit Fly (Drosophila melanogaster)	8	14,000	137 million
	Roundworm (C. elegans)	12	19,000	97 million
	Bacterium (E. coli)	1	5,000	4.1 million

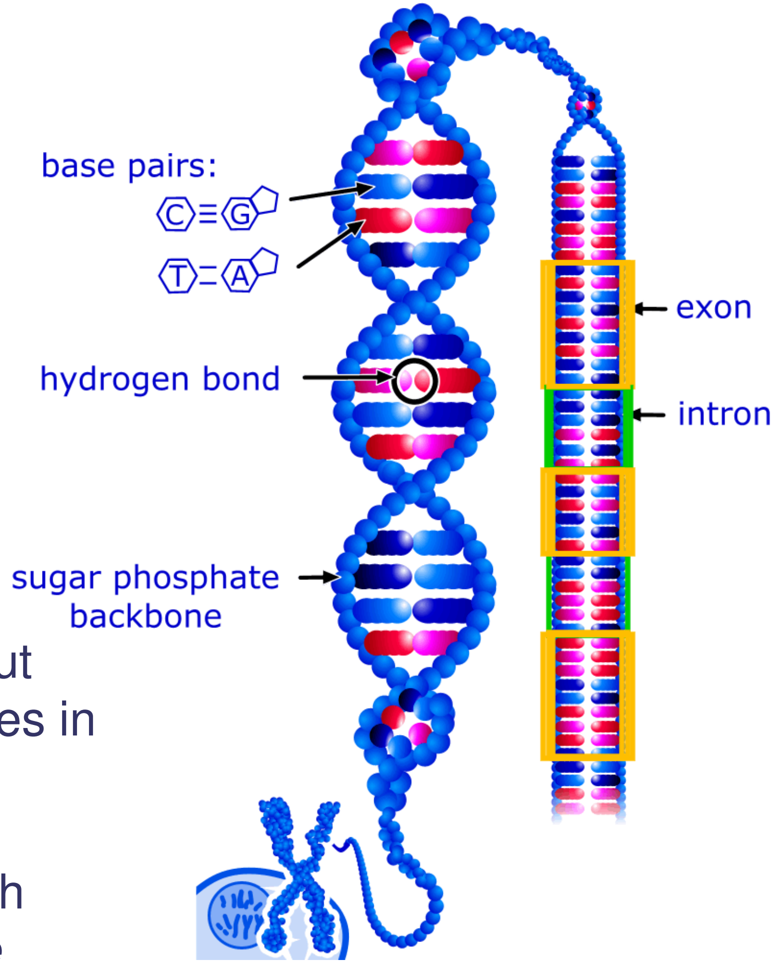
Genes

exon:

A section of DNA which carries the *coding sequence* for a protein or part of it. Exons are separated by intervening, non-coding sequences (called *introns*). In eukaryotes most genes consist of a number of exons.

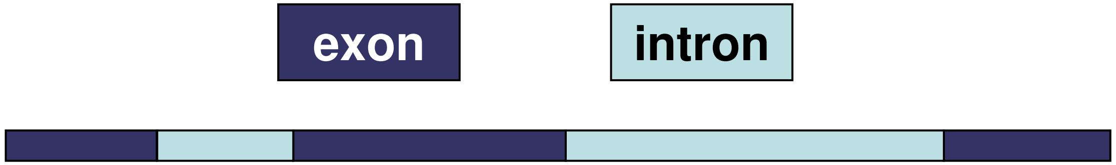
intron:

An intervening section of DNA which occurs almost exclusively within a *eukaryotic* gene, but which is not translated to amino-acid sequences in the gene product.
The introns are removed from the pre-mature mRNA through a process called *splicing*, which leaves the *exons* untouched, to form an *active* mRNA.

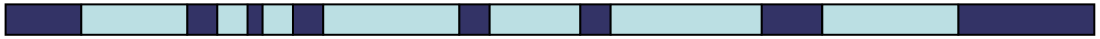


Genes

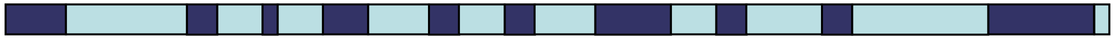
Examples of the exon:intron mosaic of genes



Globin gene – 1525 bp: 622 in exons, 893 in introns



Ovalbumin gene - ~ 7500 bp: 8 short exons comprising 1859 bp



Conalbumin gene - ~ 10,000 bp: 17 short exons comprising ~ 2,200 bp

Picking out genes in genomes

- Computer programs for genome analysis identify *ORFs* (*open reading frames*)
- An *ORF* begins with an initiation codon ATG (AUG)
- An *ORF* is a potential protein-coding region
- There are two approaches to identify protein coding regions…

Picking out genes in genomes

1. **Detection of regions similar to known coding regions from other organisms**

- Regions may encode amino acid sequences similar to known proteins
- Or may be similar to *ESTs* (correspond to genes known to be expressed)
- Few hundred initial bases of *cDNA* are sequenced to identify a gene

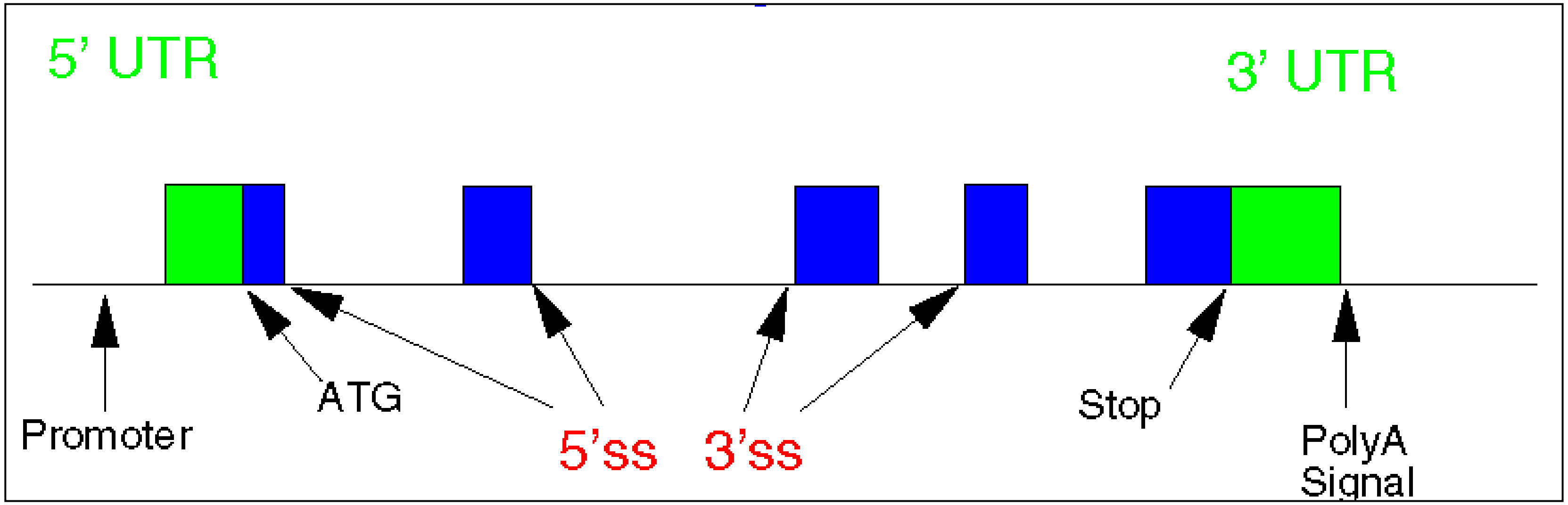
2. **Ab initio methods, seek to identify genes from the properties of the DNA sequence itself**

- Bacterial genes are easy to identify, because they are *contiguous*
- They have no *introns* and the space between genes is small
- Identification of *exons* in higher organisms is a problem, assembling them another…

Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

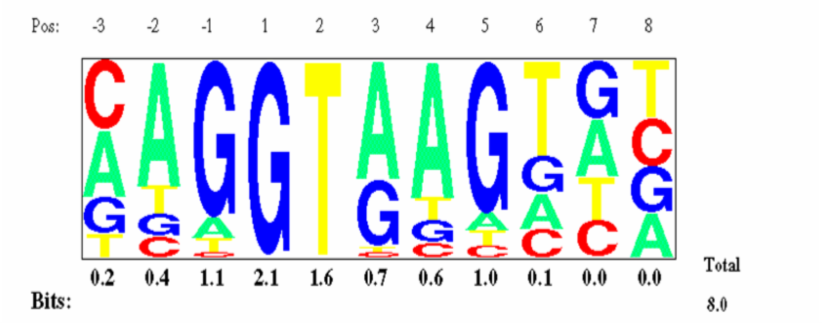
- The **initial (5´) exon** starts with a transcription start point, preceded by a core *promoter* site such as the `TATA` box (~30bp upstream)
  - Free of stop codons
  - End immediately before a `GT` splice-signal



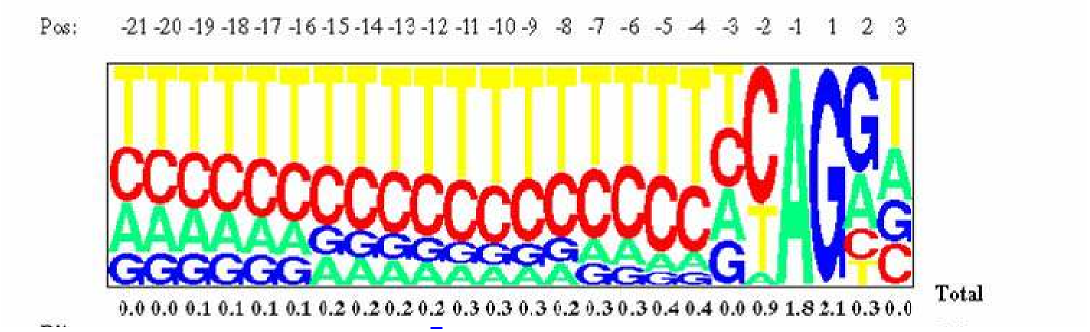
binds and directs RNA polymerase
to the correct transcriptional start site

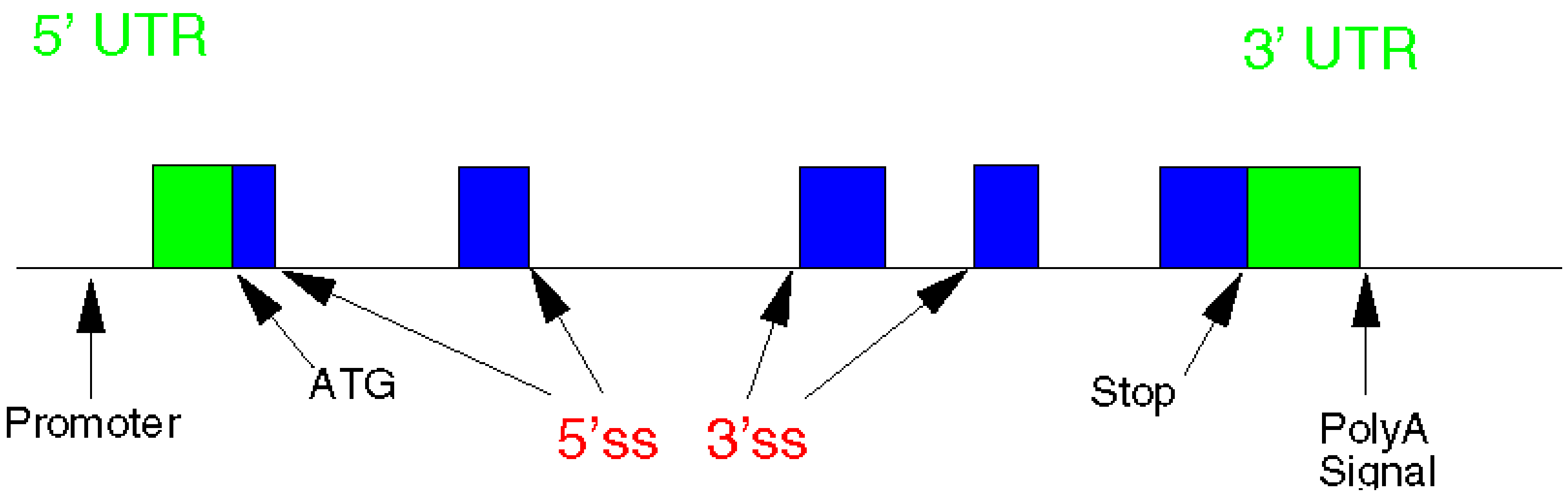
Picking out genes in genomes

5' splice signal



3' splice signal





5' UTR

3' UTR

Promoter

ATG

5'ss

3'ss

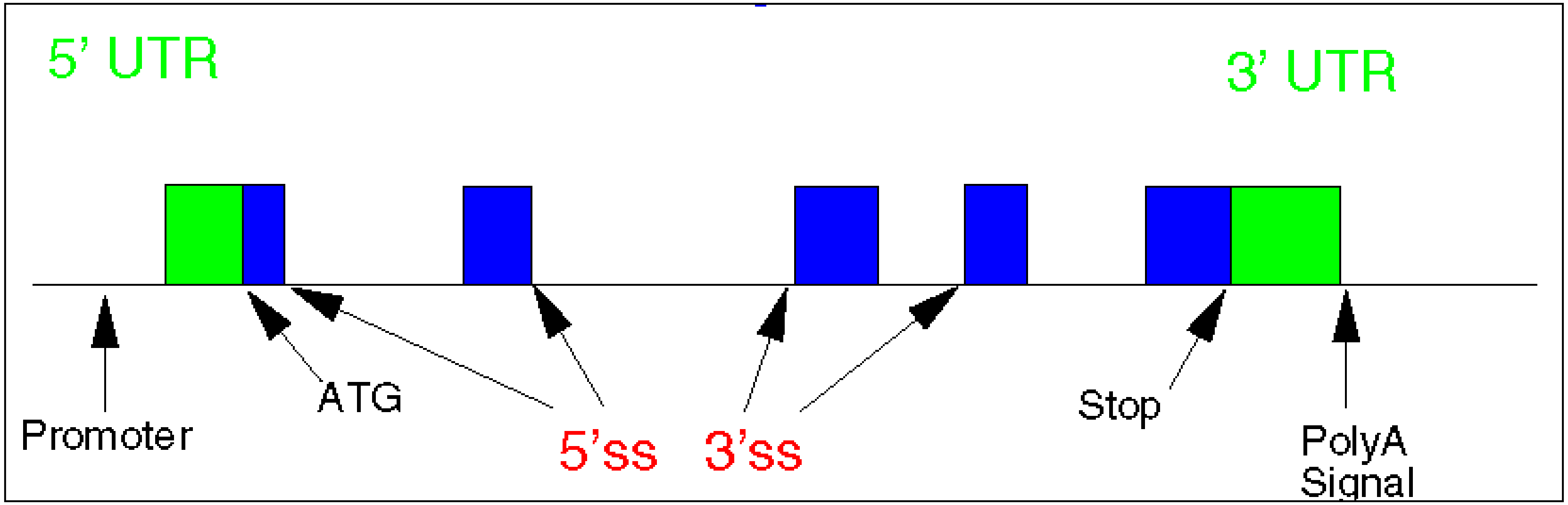
Stop

PolyA Signal

Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

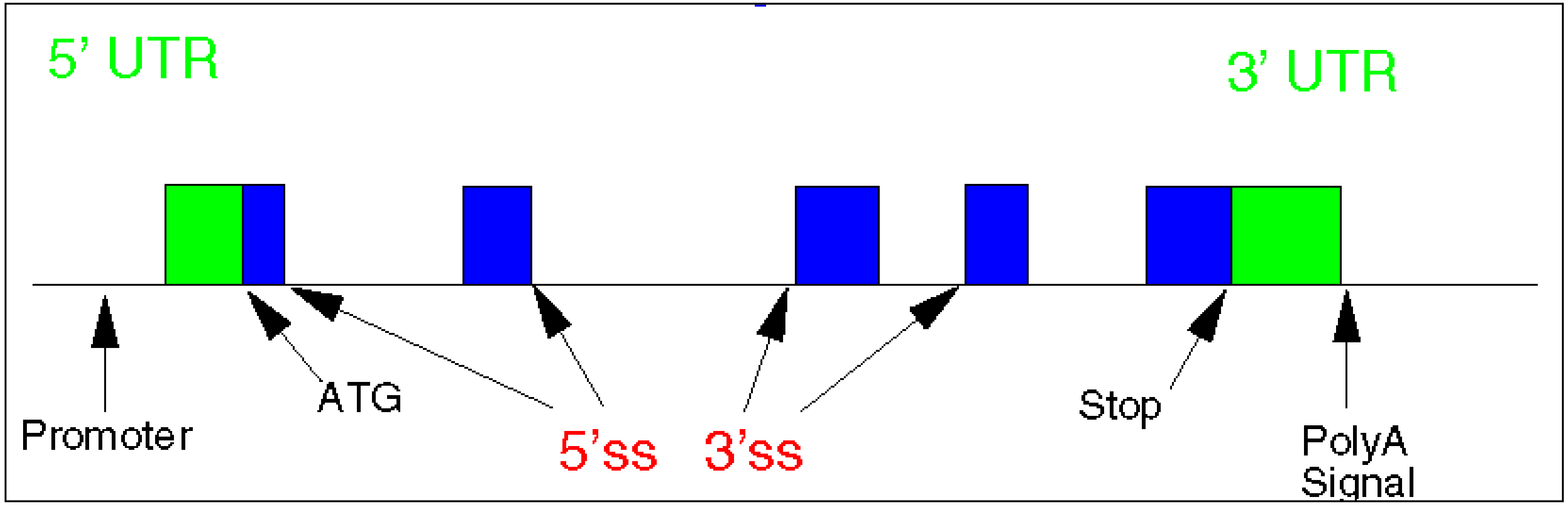
- **Internal exons** are free of stop codons too
  - Begin after an `AG` splice signal
  - End before a `GT` splice signal



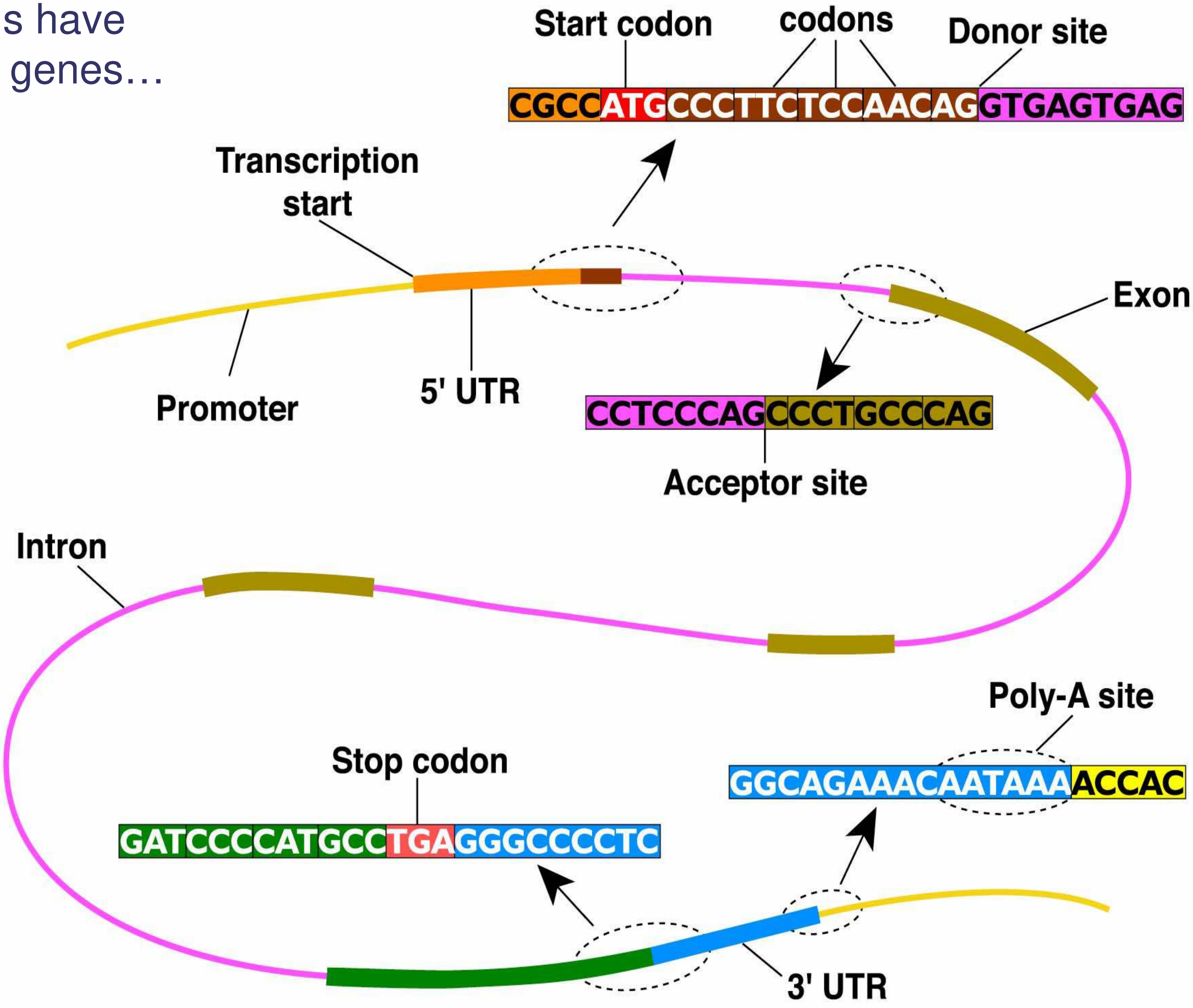
Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

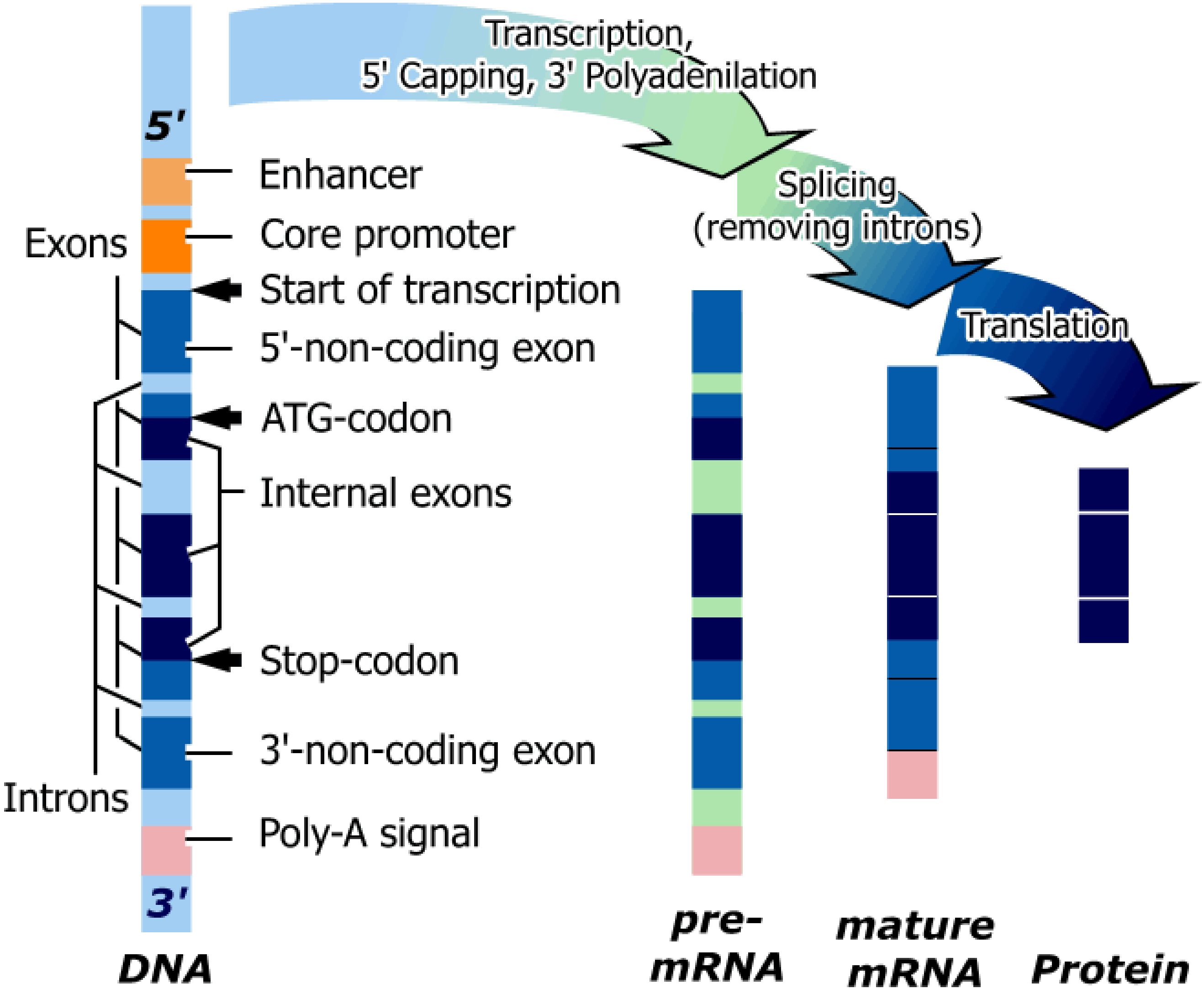
- The **final (3´) exon** starts after a an `AG` splice signal
  - Ends with a stop codon (`TAA`,`TAG`,`TGA`)
  - Followed by a polyadenylation signal sequence



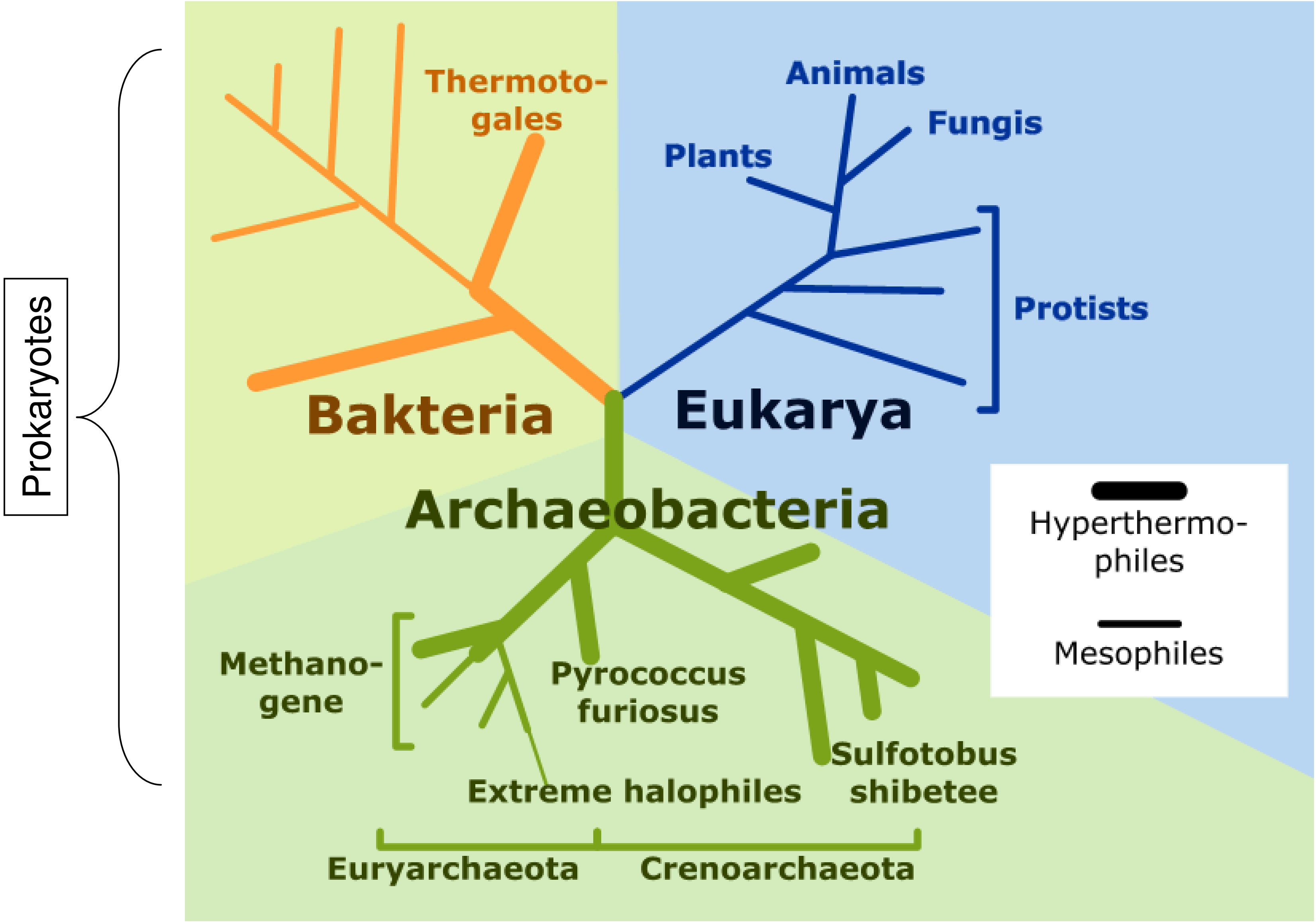
Humans have spliced genes…



DNA makes RNA makes Protein

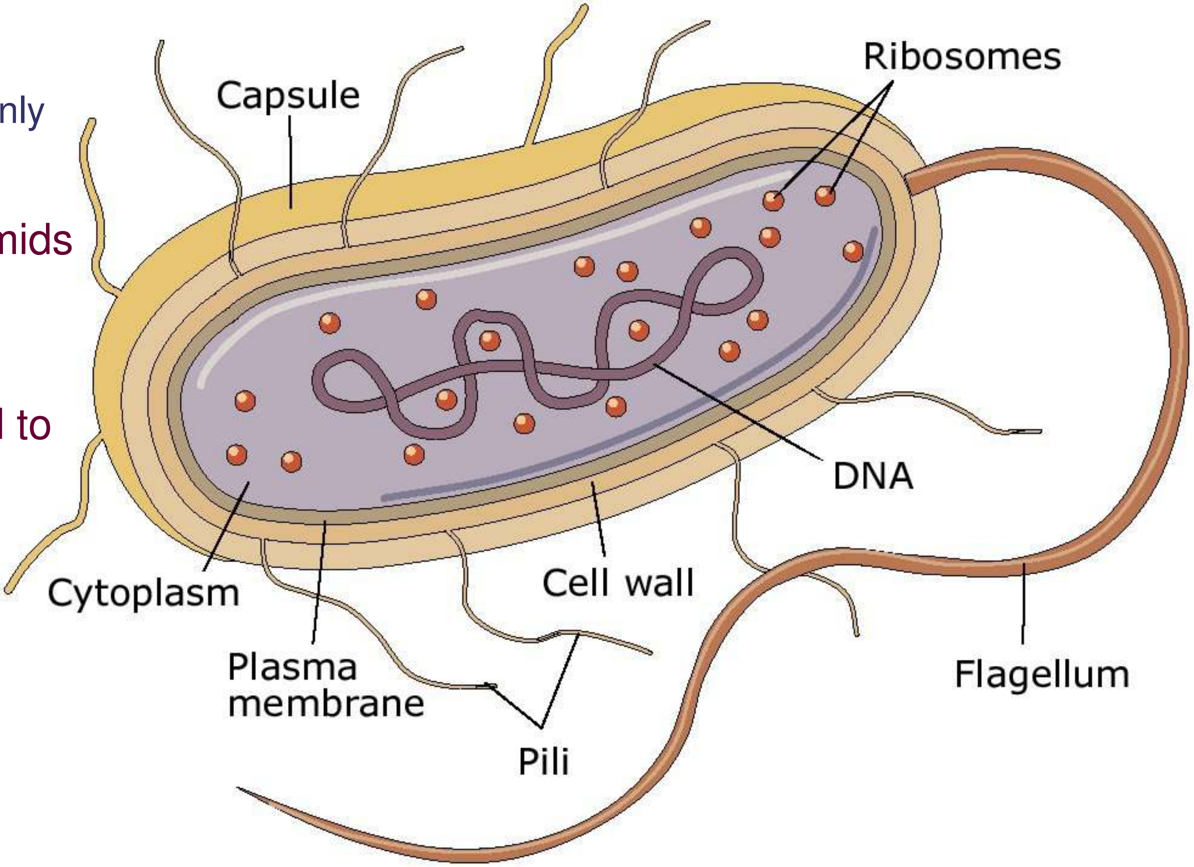


Tree of life


Prokaryotes
Thermoto-gales
Bakteria
Animals
Fungis
Plants
Protists
Eukarya
Archaeobacteria
Hyperthermo-philes
Mesophiles
Methano-gene
Pyrococcus furiosus
Sulfotobus shibetee
Extreme halophiles
Euryarchaeota
Crenoarchaeota

Genomics – Prokaryotes

- the genome of a *prokaryote* comes as a single double-stranded DNA molecule in ring-form
  - in average 2mm long
  - whereas the cells diameter is only 0.001mm
  - < 5 Mb
- *prokaryotic* cells can have plasmids as well (see next slide)
- protein coding regions have no *introns*
- little non-coding DNA compared to eukaryotes
  - in E.coli only 11%

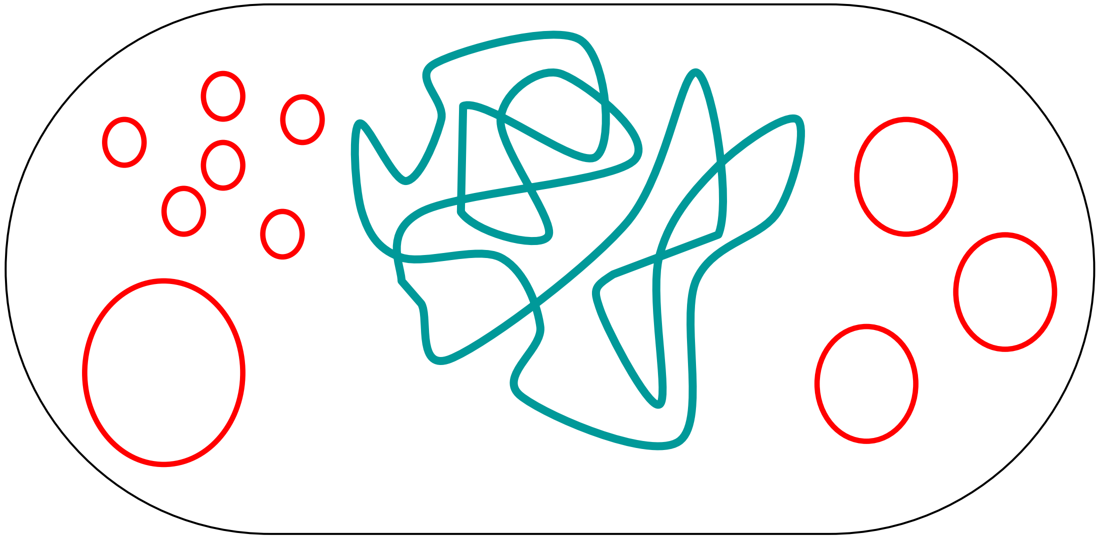


Genomics - Plasmids

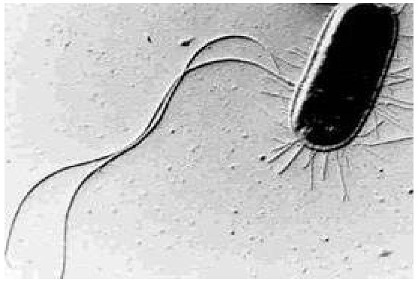
- *Plasmids* are circular double stranded DNA molecules that are separate from the *chromosomal* DNA.

- They usually occur in *bacteria*, sometimes in *eukaryotic* organisms

- Their size varies from 1 to 250 kilo base pairs (kbp). There are from one copy, for large plasmids, to hundreds of copies of the same plasmid present in a single cell.

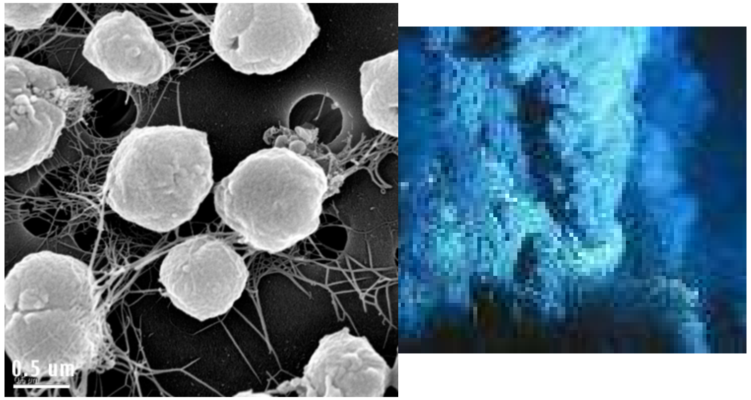


Prokaryotic model organisms



E.coli (Escherichia coli)

Methanococcus jannaschii (archaeon)



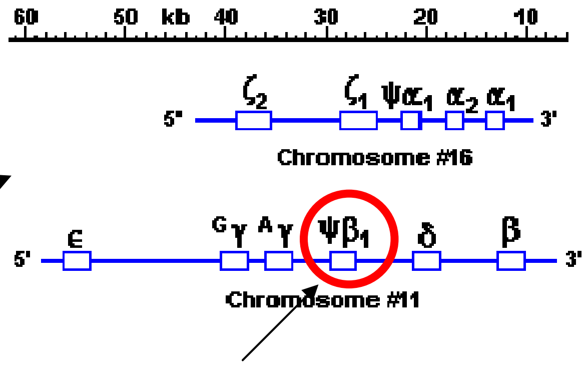
Mycoplasma genitalium
(simplest organism known)

Genomics

- DNA of higher organisms is organized into *chromosomes* (human – 23 chromosome pairs)
- not all DNA codes for proteins
- on the other hand some genes exist in multiple copies
- that’s why from the genome size you can’t easily estimate the amount of protein sequence information

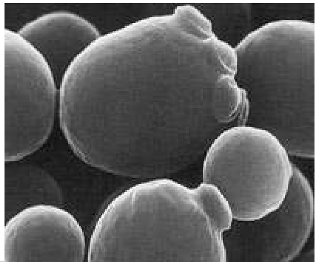
Genomes of eukaryotes

- majority of the DNA is in the nucleus, separated into bundles (*chromosomes*)
  - small amounts of DNA appear in organelles (mitochondria and chloroplasts)
- within single chromosomes gene families are common
  - some family members are *paralogues* (related)
    - they have duplicated within the same genome
    - often diverged to provide separate functions in descendants (Nachkommen)
    - e.g. human *α* and *β* globin
  - *orthologues* genes
    - are homologues in different species
    - often perform the same function
    - e.g. human and horse *myoglobin*
  - *pseudogenes*
    - lost their function
    - e.g. human *globin* gene cluster

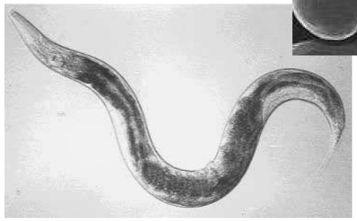


pseudogene

Eukaryotic model organisms



- *Saccharomyces cerevisiae* (baker’s yeast)



- *Caenorhabditis elegans* (C.elegans)

- *Drosophila melanogaster* (fruit fly)

- *Arabidopsis thaliana* (flower)
- *Homo sapiens* (human)

The human genome

- ~3.2 x 10^9 bp (thirty time larger than *C.elegans* or *D.melongaster*)
- coding sequences form only 5% of the human genome
- Repeat sequences over 50%
- Only ~32.000 genes
- Human genome is distributed over *22 chromosome pairs* plus *X* and *Y* chromosomes
- *Exons* of protein-coding genes are relatively small compared to other known eukaryotic genomes
- *Introns* are relatively long
- Protein-coding genes span long stretches of DNA (dystrophin, coding a 3.685 amino acid protein, is >2.4Mbp long)

- Average gene length: ~ 8,000 bp
- Average of 5-6 exons/gene
- Average exon length: ~200 bp
- Average intron length: ~2,000 bp
- ~8% genes have a single exon
- Some exons can be as small as 1 or 3 bp.

The human genome

Top categories in a function classification:

Function	Number	%
Nucleic acid binding	2207	14.0
DNA binding	1656	10.5
DNA repair protein	45	0.2
DNA replication factor	7	0.0
Transcription factor	986	6.2
RNA binding	380	2.4
Structural protein of ribosome	137	0.8
Translation factor	44	0.2
Transcription factor binding	6	0.0
Cell Cycle regulator	75	0.4
Chaperone	154	0.9
Motor	85	0.5
Actin binding	129	0.8
Defense/immunity protein	603	3.8
Enzyme	3242	20.6
Peptidase	457	2.9
Endopeptidase	403	2.5
Protein kinase	839	5.3
Protein phosphatase	295	1.8
Enzyme activator	3	0.0

Function	Number	%
Apoptosis inhibitor	132	0.8
Signal transduction	1790	11.4
Receptor	1318	8.4
Transmembrane receptor	1202	7.6
G-protein link receptor	489	3.1
Olfactory receptor	71	0.0
Storage protein	7	0.0
Cell adhesion	189	1.2
Structural protein	714	4.5
Cytoskeletal structural protein	145	0.9
Transporter	682	4.3
Ion channel	269	1.7
Neurotransmitter transporter	19	0.1
Ligand binding or carrier	1536	9.7
Electron transfer	33	0.2
Cytochrome P450	50	0.3
Tumor suppressor	5	0.0
Unclassified	4813	30.6
Total	**15683**	**100.0**

The human genome

- Repeated sequences comprise over 50% of the genome:
  - *Transposable* elements, or *interspersed* repeats include *LINEs* and *SINEs* (almost 50%)
  - Retroposed *pseudogenes*
  - Simple '*stutters*' - repeats of short oligomers (*minisatellites* and *microsatellites*)
  - *Segment duplication*, of blocks of ~10 - 300kb
  - Blocks of *tandem repeats,* including gene families

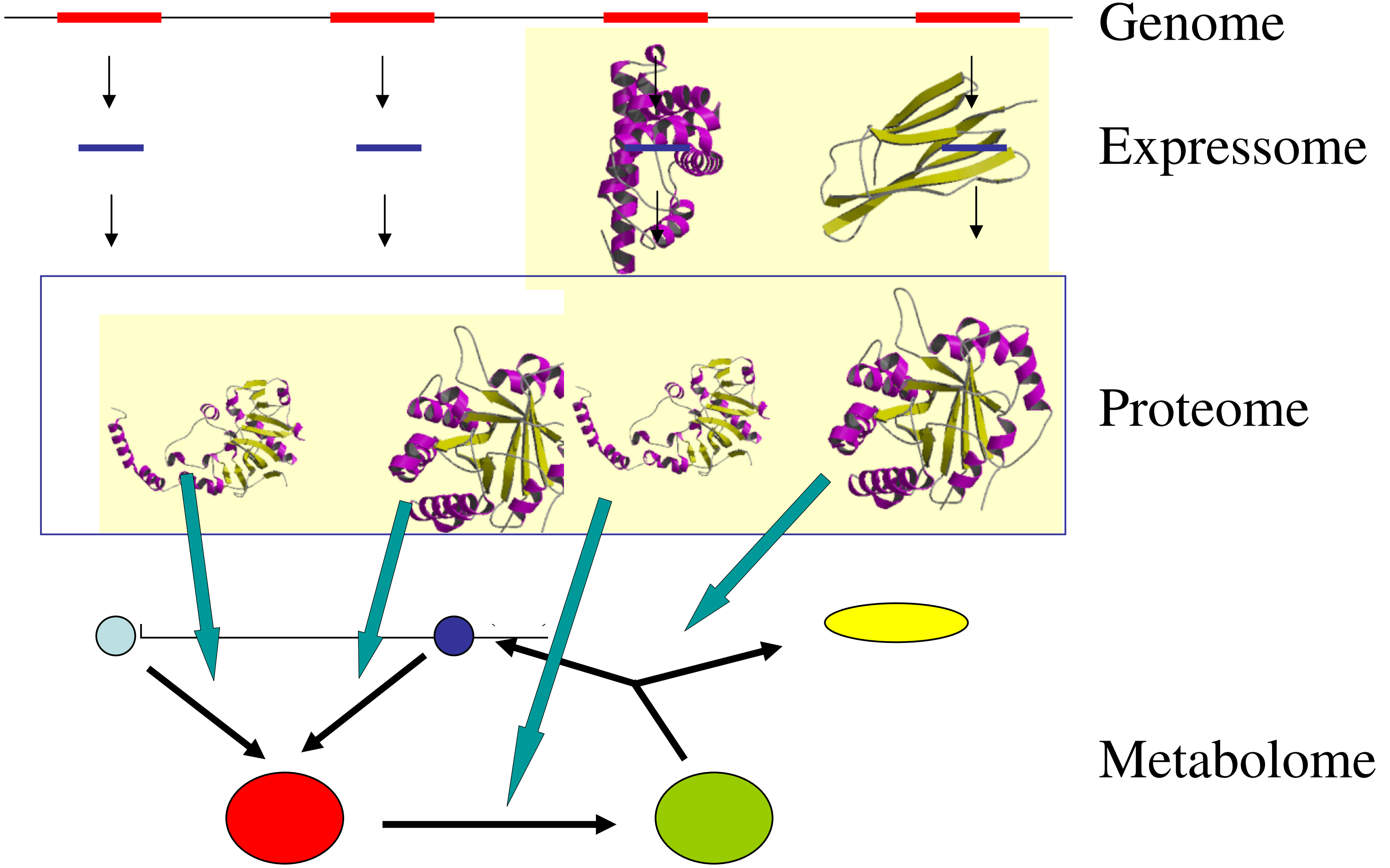
Element	Size (bp)	Copy number	Fraction of genome %
Short Interspersed Nuclear Elements (SINEs)	100-300	1.500.000	13
Long Interspersed Nuclear Elements (LINEs)	6000-8000	850.000	21
Long Terminal Repeats	15.000 -110.000	450.000	8
DNA Transposon fossils	80-3000	300.000	3

The human genome

- All people are different, but the DNA of different people only varies for 0.2% or less.
- So, only up to 2 letters in 1000 are expected to be different.
- Evidence in current genomics studies (Single Nucleotide Polymorphisms or SNPs) imply that on average only 1 letter out of 1400 is different between individuals.
- means that 2 to 3 million letters would differ between individuals.

Functional Genomics

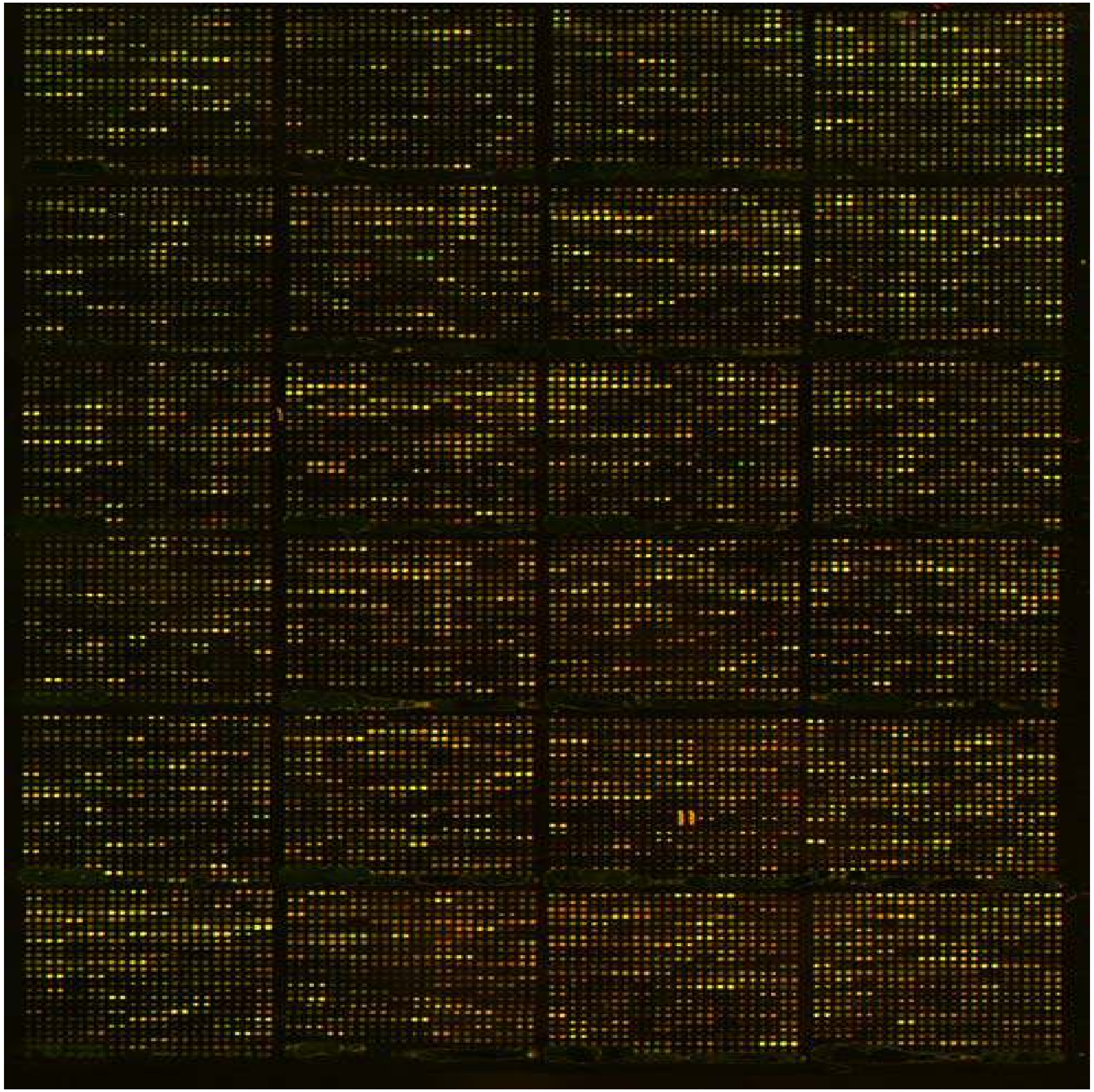
From gene to function



DNA makes RNA makes Protein:

Expression data

- More copies of mRNA for a gene leads to more protein
- mRNA can now be measured for all the genes in a cell at ones through microarray technology
- Can have 60,000 spots (genes) on a single gene chip
- Color change gives intensity of gene expression (over- or under-expression)

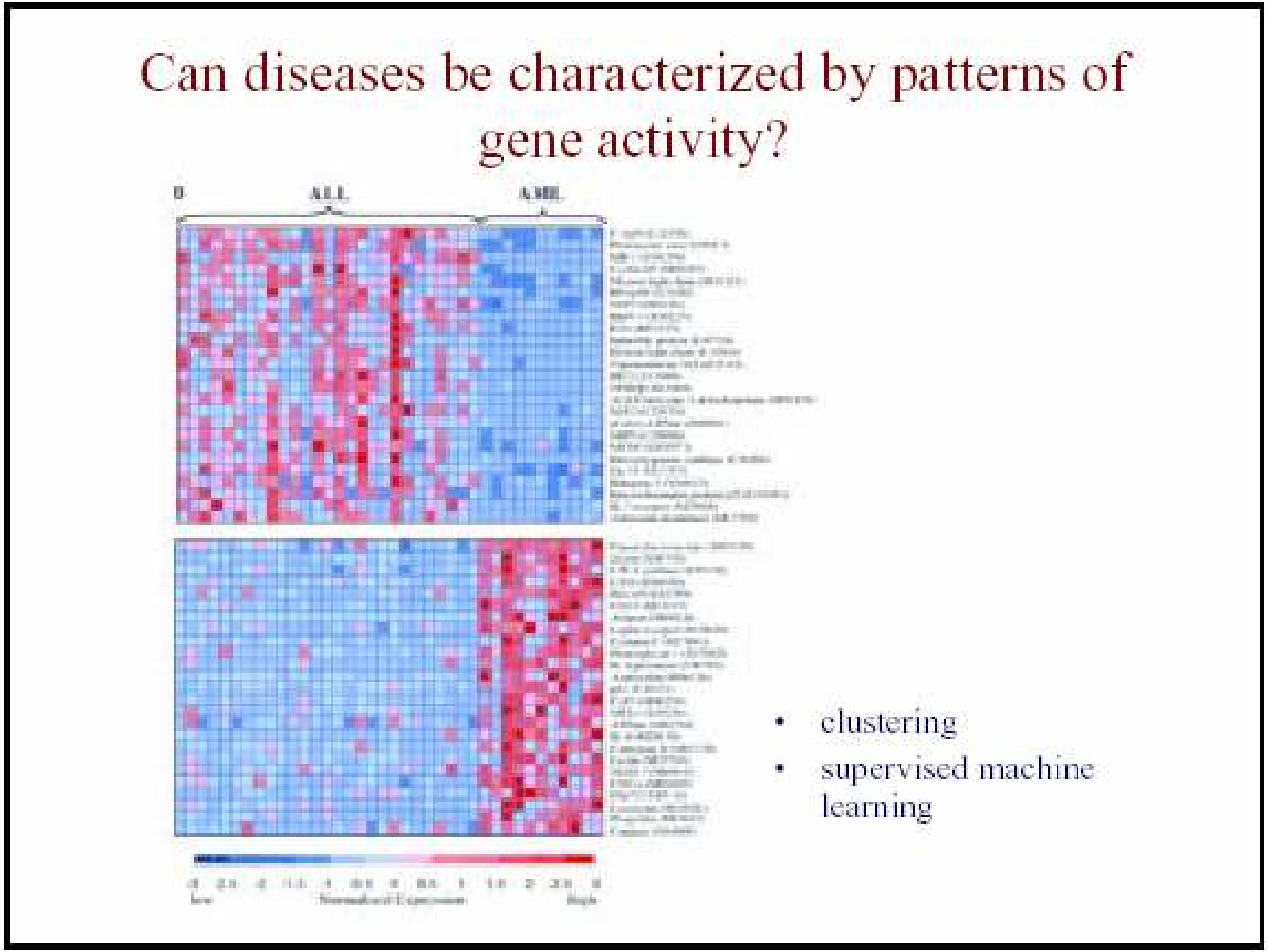


Genes and regulatory regions

regulatory mechanisms organize the expression of genes

- genes may be turned *on* or *off* in response to concentrations of *nutrients* or to *stress*
- control regions often lie near the segments coding for proteins
- they can serve as binding sites for molecules that transcribe the DNA
- or they bind regulatory molecules that can *block* transcription

Expression data



Can diseases be characterized by patterns of gene activity?

- clustering
- supervised machine learning

Outlook – coming lecture

Proteomics

- Proteins
- post-translational modification
- Key technologies

- Maps of hereditary information
- SNPs (Single nucleotide polymorphisms)
- Genetic diseases

Thanks for your attention!