Lessons 8-9. Contents

1. An overview of genomes and genes organization

2. Gene finding in prokaryote genomes

3. Gene finding in eukaryote genomes

BIOinformatics = genes + proteins + informatics

(part of computational biology, biocomputing)

GENE: DNA segment which codes for a specific protein and determines an hereditary feature

PROTEIN: expression product of a gene and ed EFFECTOR of the biochemical function whose information is stored in the gene

RNA: fundamental role in the gene expression regulation

The DNA length in the human genome is approximately **3.2 billion nucleotides**

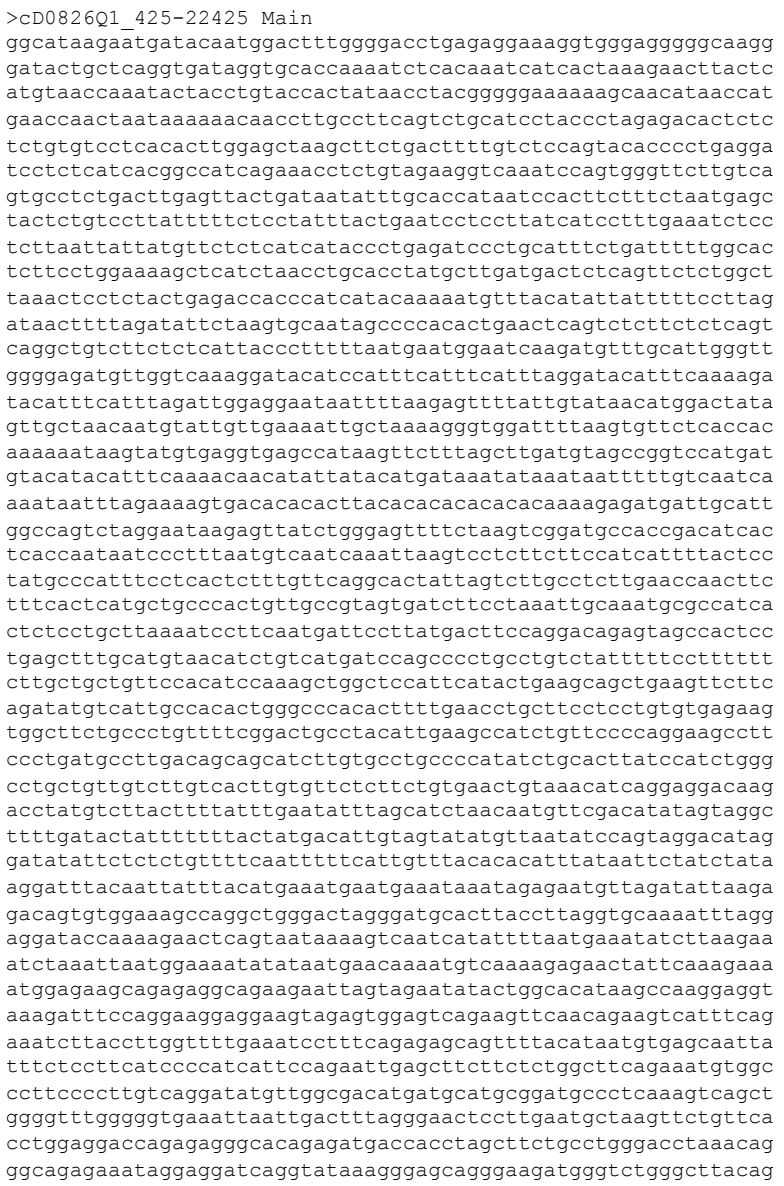
Such a nucleotide sequence (combination of A, C, G, T) is not random !!!

The large molecules in living organisms offer the most striking example of information density in the universe


Where is the gene?

Where is the gene?

Hay in a haystack
(A. Tramontano)



The cell…

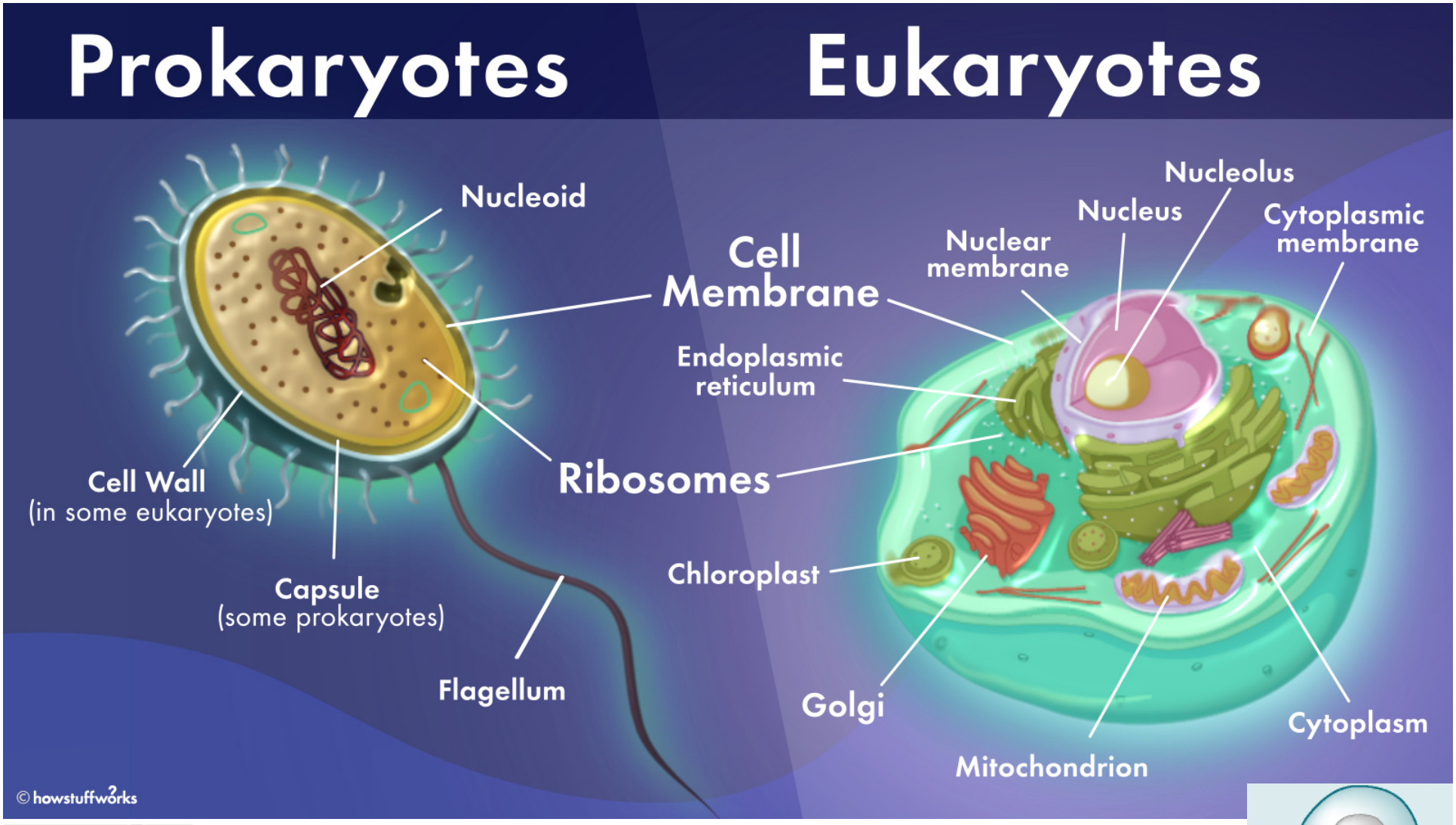
B

All living organisms are made of cells

Cell is the fundamental unit of life

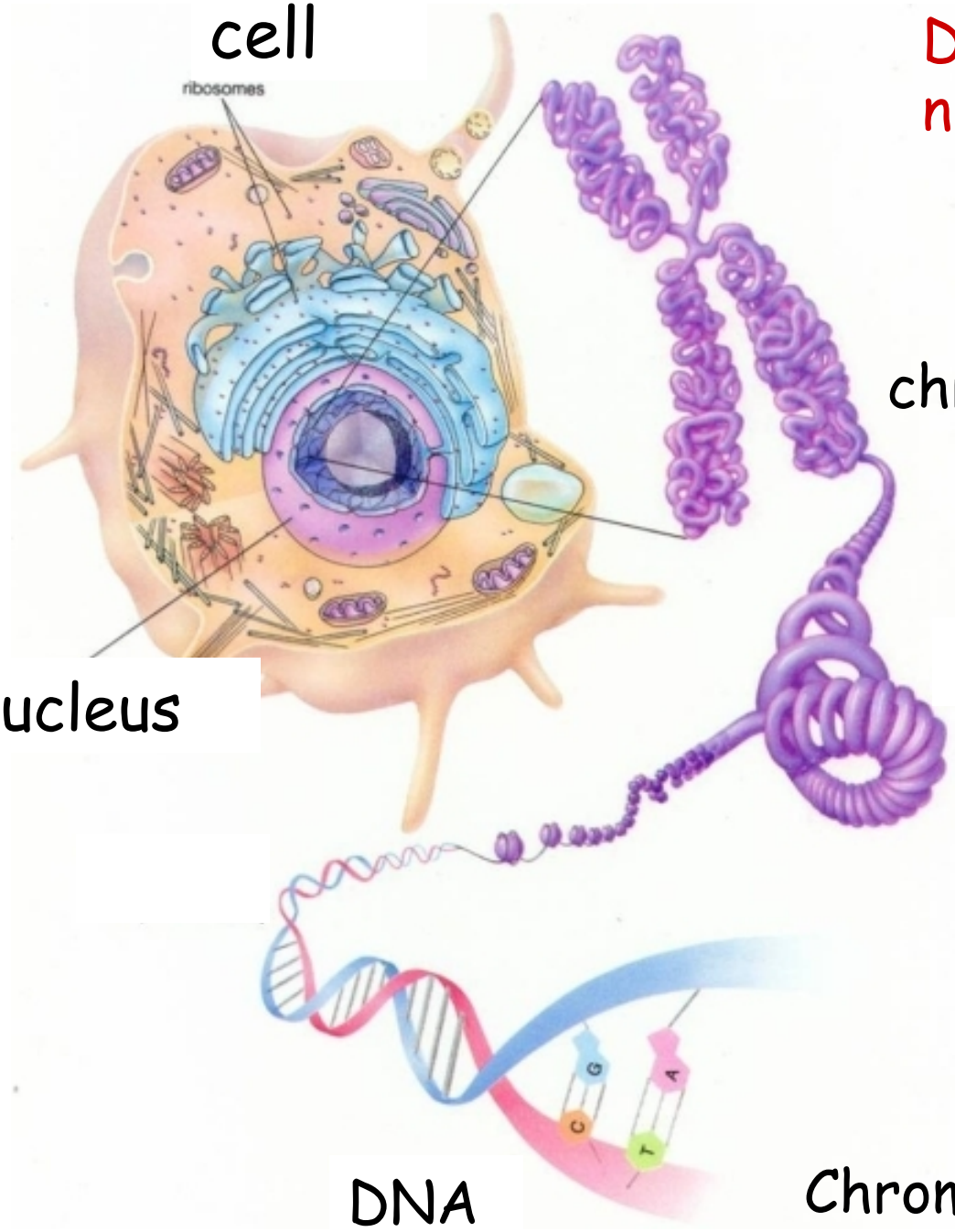
Organisms are classified as prokaryotes or eukaryotes based on the complexity of cells they are made of

Organisms are classified in prokaryotes and eukaryotes based on the complexity of their cells



Translation and transcription processes need to be considered separately in prokaryotes and eukaryotes

DNA is located in the cell nucleus (or nucleoid)



Chromosomes at the microscope

Genes complexity

- *Coding genes* in 100Kb (Kb=10^3 nucleotide bases)
  - *E. Coli* 87 } ***prokaryotes***
  - *S. Cerevisiae* 52 }
  - *C. elegans* 22 } ***eukaryotes***
  - *H. sapiens* 5 }

The “gene density” is much lower in eukaryotes and especially in complex organisms

This reflects the different complexity of the genomes (and genes) themselves

Prokaryotic genomes

- Small dimensions (< 10 Mb)
- High gene density (>85% of the genome is coding)
- Low nucleotide redundancy
- Introns (quasi) absent
- Low complexity of gene and genome architecture
- Easy to be sequenced

Gene finding approaches can be intrinsic or extrinsic

Gene finding methods which attempt to identify genes without reference to known sequences are also known as **intrinsic** methods

Gene finding methods which take advantage of similarity to sequences in databases are also known as **extrinsic** methods

The two approaches can of course be **integrated** to achieve an efficient prediction

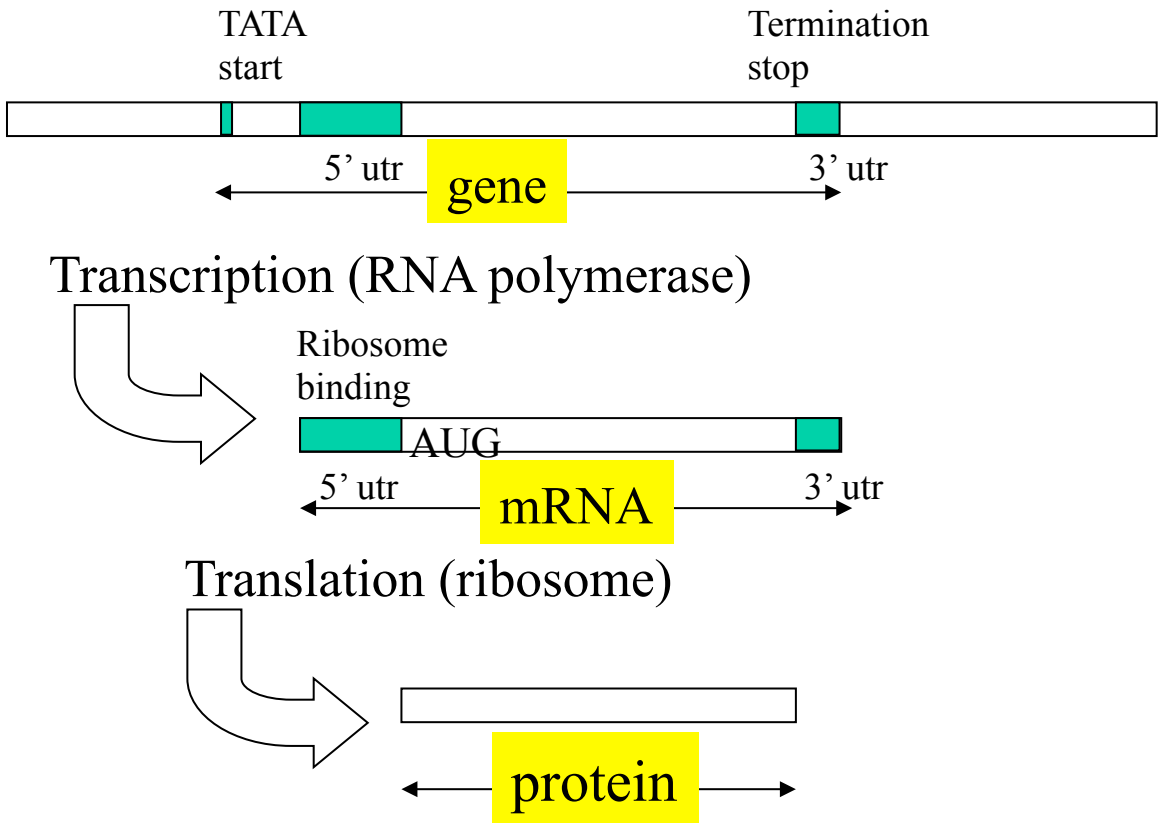
Methods for gene search

The main aim of the gene search is being able to distinguish between coding and non coding DNA and locate genes

1. Identifying ORFs (Open Reading Frames)
2. *Signals* search (e.g. promoters)
3. Difference in the *nucleotide composition*
4. Homology search (sequence similarity)

The first three approaches are intrinsic, the last one is extrinsic

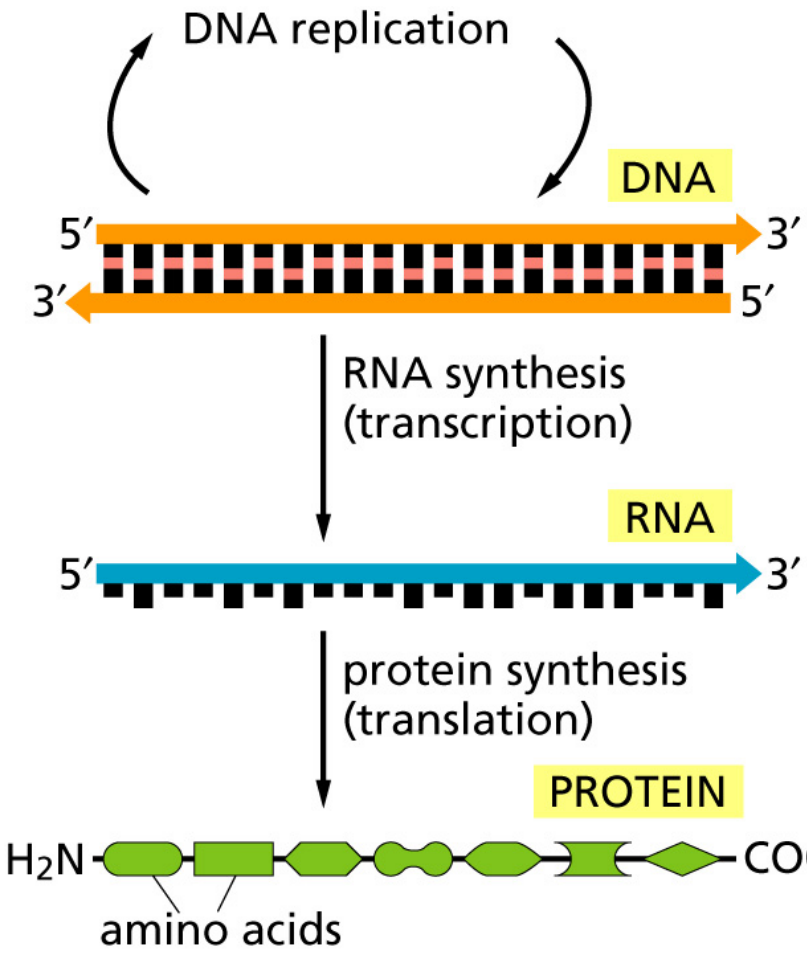
Structure of a prokaryotic gene

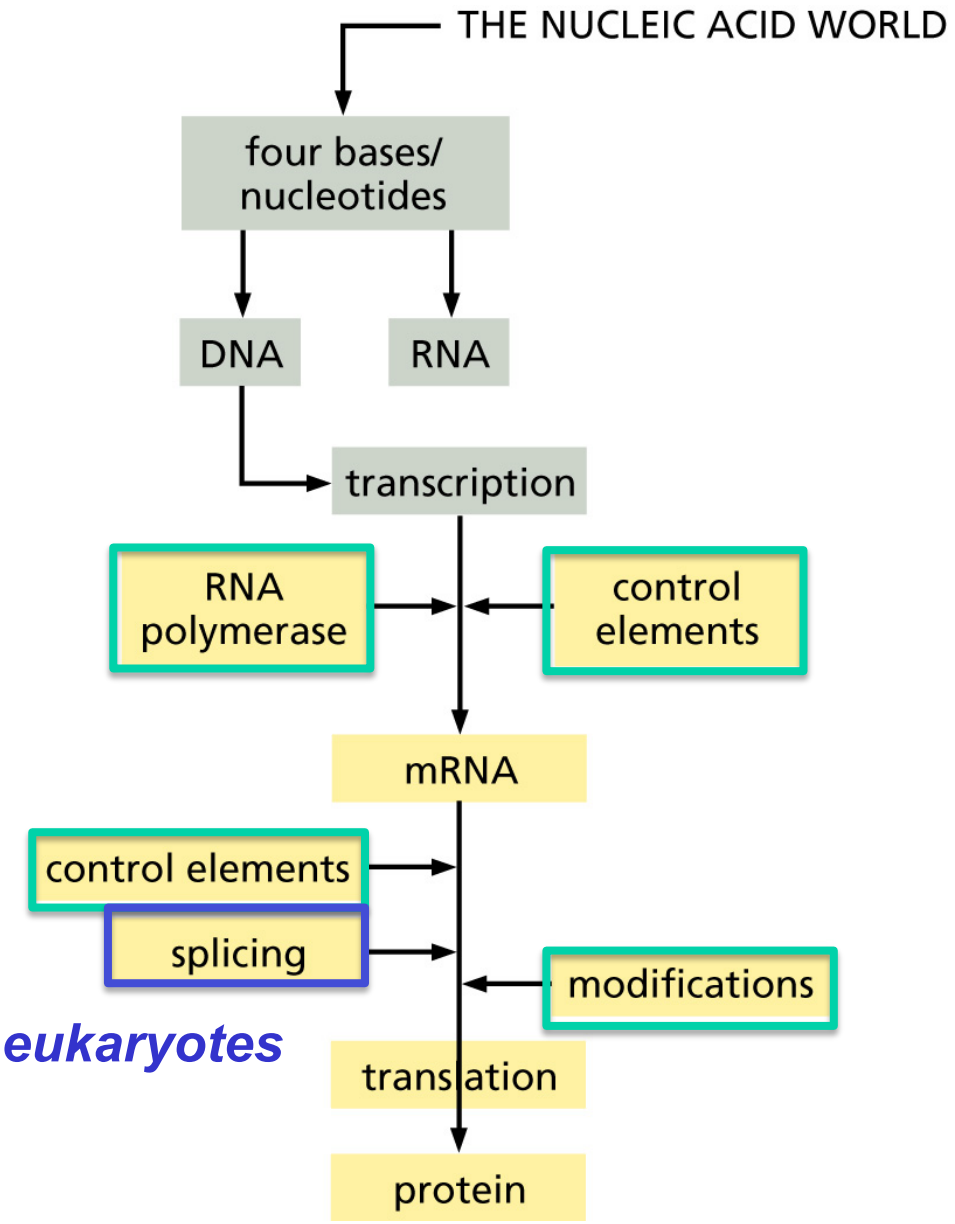


DNA —transcription→ RNA —translation→ protein

Molecular biology central dogma: a more realistic view

B





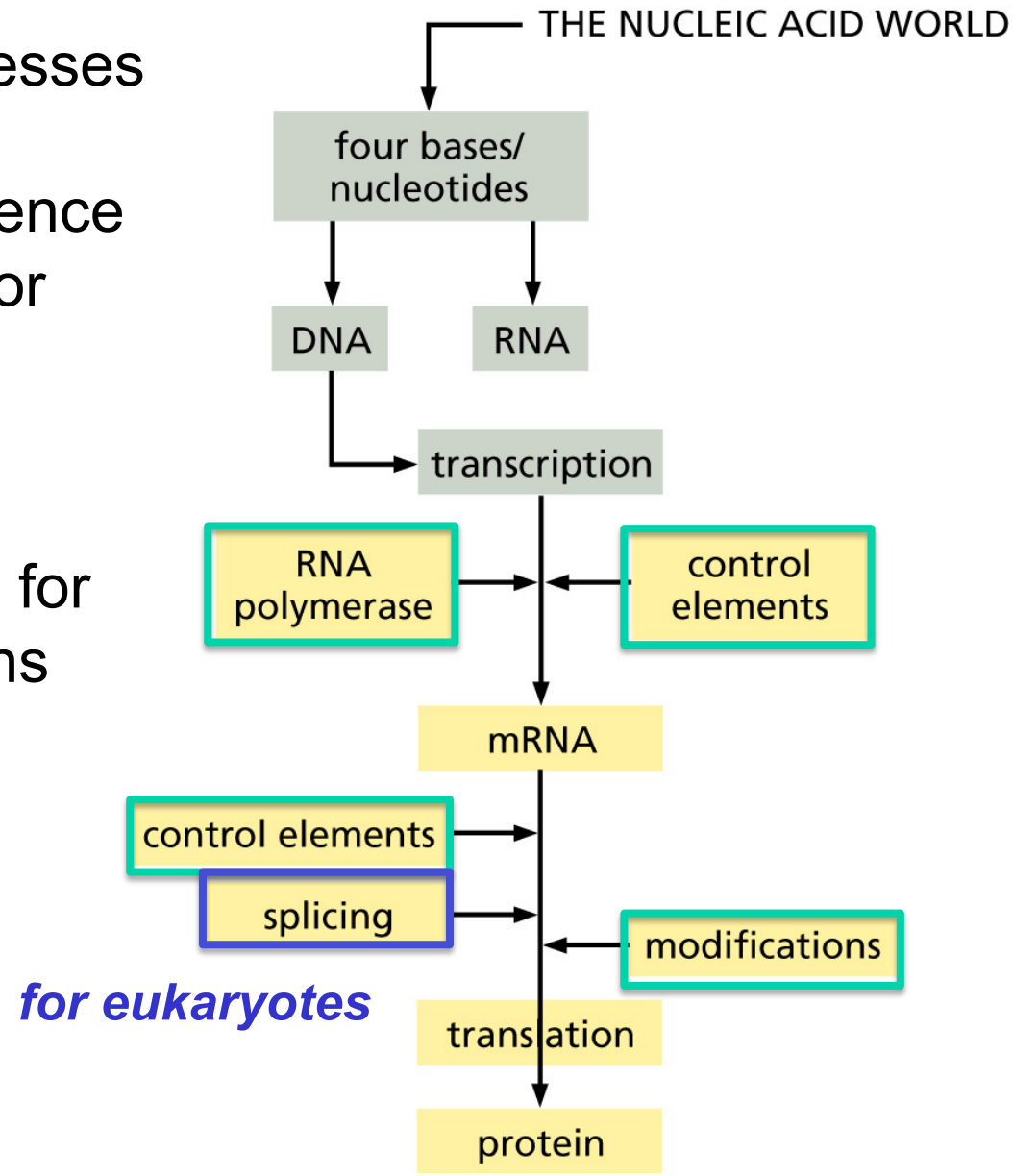
for eukaryotes

Molecular biology central dogma: a more realistic view

B

The regulation of many processes involved in transcription and translation relies on the presence of short **signal sequences** (or regulatory elements)

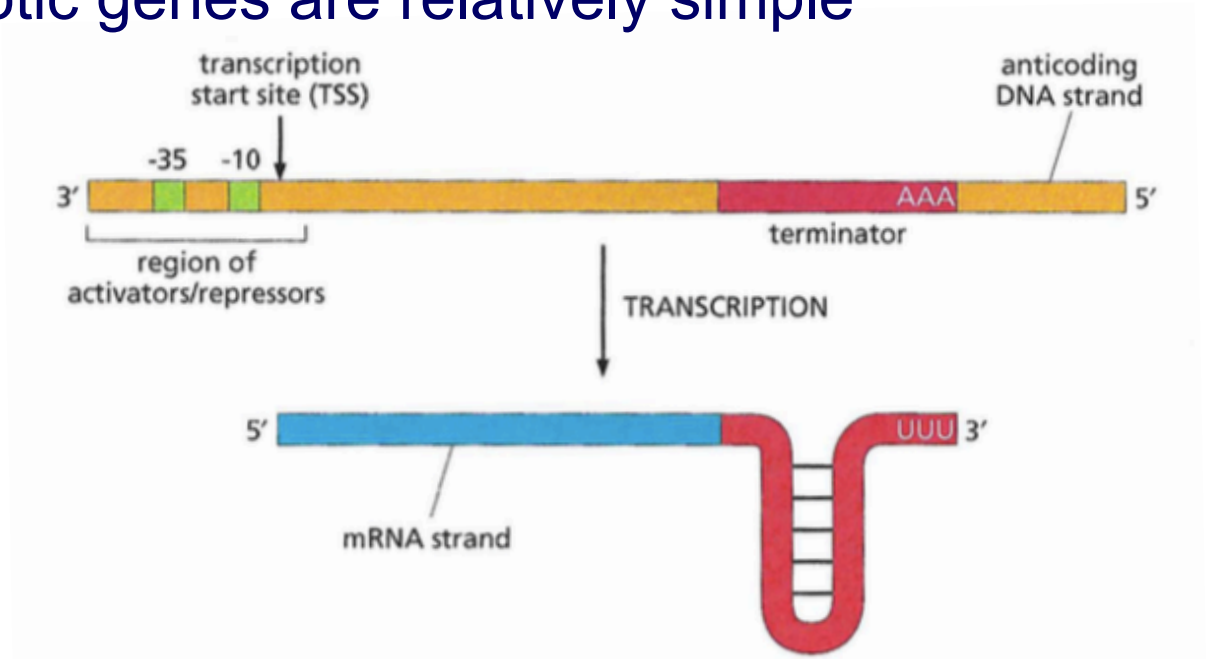
The majority of these **signal sequences** are binding sites for specialized regulatory proteins that interact with DNA to regulate transcription, RNA maturation and translation



for eukaryotes

Prokaryotic protein-coding genes

Prokaryotic genes are relatively simple



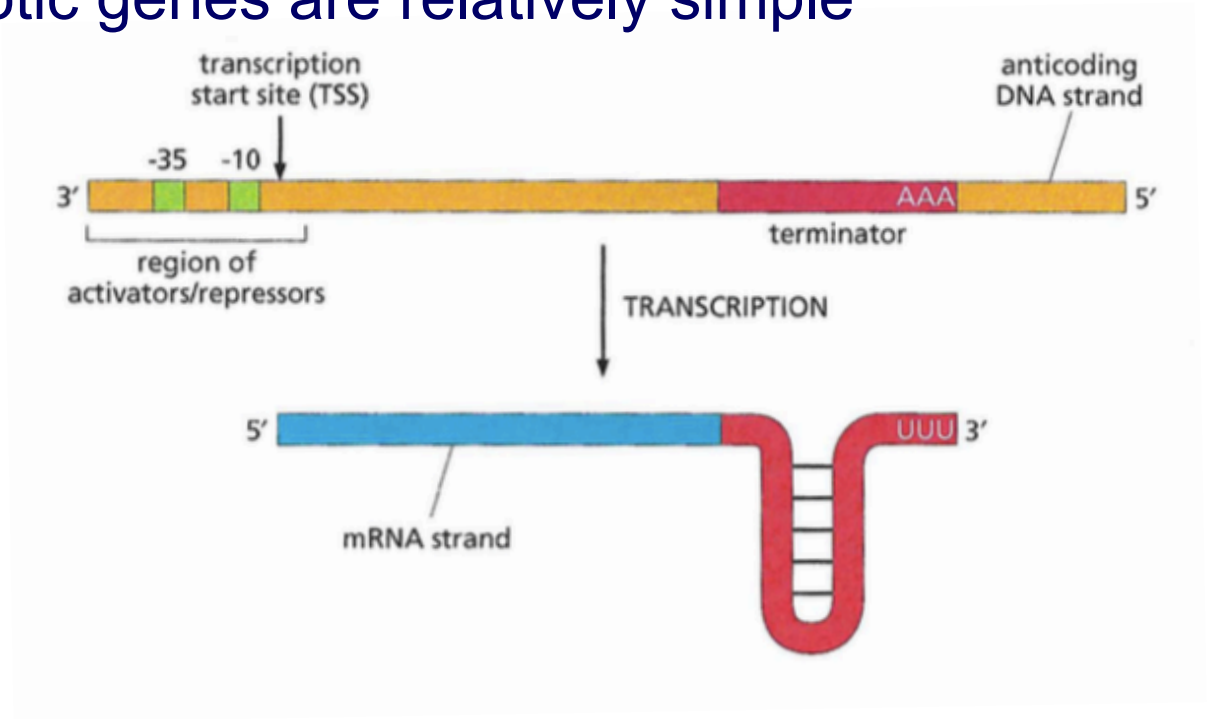
Control regions in DNA at which RNA polymerase binds to initiate transcription are named **promoters** (or start signals)

Prokaryotic genes typically feature two characteristic short **promoters** which are quite conserved and are centered around position -10 and -35 before the transcription start site

The signal to stop transcription (stop signal) is a short sequence forming a loop that prevents the transcription machine from continuing

Prokaryotic protein-coding genes

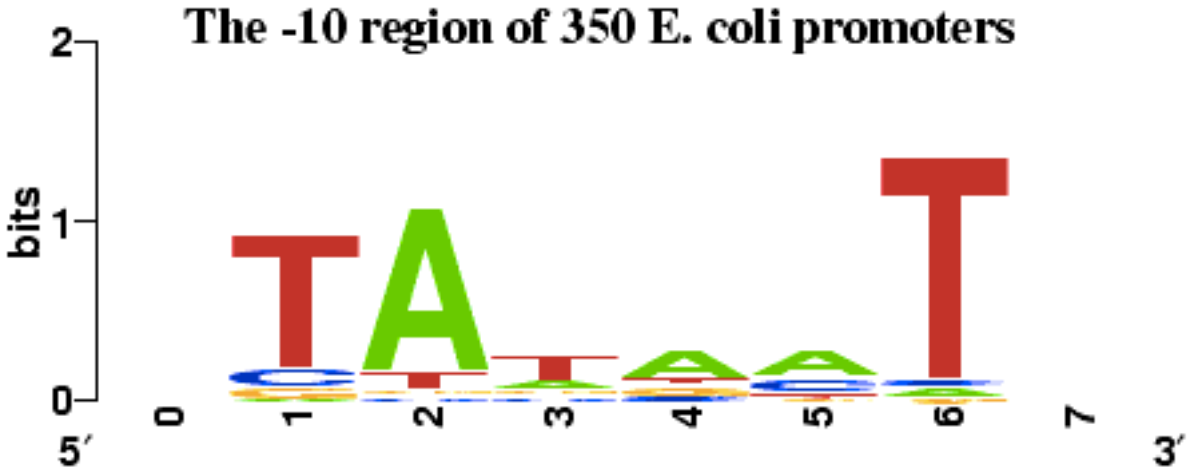
Prokaryotic genes are relatively simple



In addition, prokaryotes feature a **Shine-Delgado sequence** corresponding to the *consensus* AGGAGGU for ribosome binding, a few bases upstream of the start codon

Signals have a conserved sequence

Example: Sequence logo of the *-10 sequence (Pribnow box), recognized and bound by a subunit of RNA polymerase during initiation of transcription* in E. coli genes



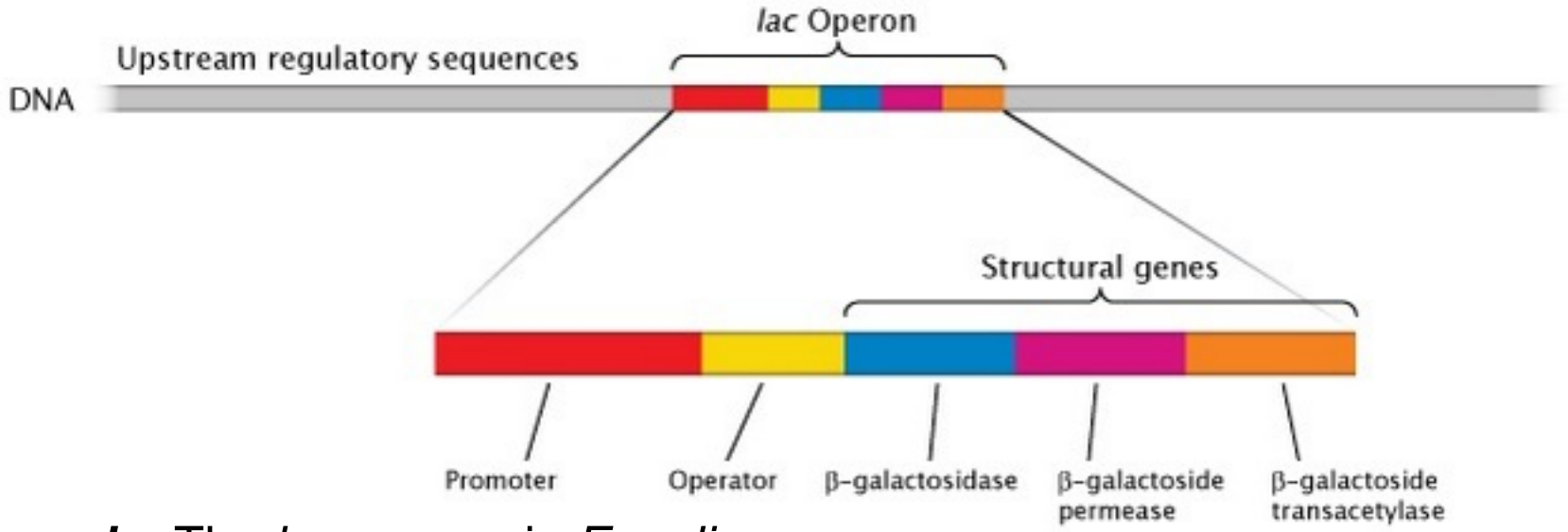
from *https://weblogo.berkeley.edu/examples.html*

A ‘TATA’ box with similar function is also found in archaea & eukarya

The AT-richness is important to allow the base pairs separation during transcription, since A and T are easier to break apart, both due to fewer hydrogen bonds and to weaker base stacking

Prokaryotic genes

More genes of prokaryotes (bacteria) are usually organized in operons, that are transcribed together to give a single messenger RNA (mRNA) molecule, which therefore encodes multiple proteins



Example: The *lac* operon in *E. coli*.

Three genes code for proteins involved in lactose import and metabolism in bacteria. The genes are organized in a cluster called the *lac* operon

A promoter is a sequence of DNA needed to turn a gene on or off

An operator is a genetic sequence which allows proteins responsible for transcription to attach to the DNA sequence

Gene *signals* search

Gene ***signals*** are discrete units with a recognizable *consensus* sequence (promotores, polyadenilation sites, donor and acceptor splice sites)

Algorithms used to search for gene signals are known as ***signal sensors***

Signal sensors can search a signal ***by itself or in a contextual way*** (i.e. relatively to other local signals)

Promoters search

As said before, coding sequences are flanked by promoter regions with a variable although quite conserved sequence

It is possible to define ***consensus sequences*** - featuring the most conserved nucleotide at each position - that all promoters of that type resemble to

For example, in *E. Coli* from the analysis of 263 promoters the following *consensus* has been derived:

```
     -35                    -10
5'---TTGACA--17bp--TATAAT--10bp--transcription
```

Generally **statistical measures** are used:

ex. Position-Specific Matrix

	1	2	3	4	5	6
A	0.1	0.1	0.1	**0.6**	0.1	**0.5**
T	**0.7**	**0.7**	0.1	0.1	0.1	0.3
G	0.1	0.1	**0.6**	0.1	0.1	0.1
C	0.1	0.1	0.2	0.2	**0.7**	0.1
	T	**T**	**G**	**A**	**C**	**A**

Consensus sequence

Frequency of occurrence of an event

For the T nucleotide at position 1:

```
   ...123456...
1  ...ATTACT...
2  ...TAGACA...
3  ...TTGTCT...
4  ...GTACAT...
5  ...TGGATA...
6  ...CCGAGG...
7  ...TTCGCC...
8  ...TTGACA...
9  ...TTGACA...
10 ...TTCCCA...
```

$$F_{(T,pos)} = \frac{N_T}{N_A + N_G + N_T + N_C}$$

$$F_{(T,1)} = \frac{7}{1 + 1 + 7 + 1} = 0.7$$

How can we use a *consensus* derived from known cases to predict novel cases?

Still about signals search, given this Scoring Matrix…

	1	2	3	4	5	6	7	8	9
A	0.30	0.50	0.10	0.45	0.65	0.32	0.12	0.20	0.65
T	0.20	0.15	0.05	0.20	0.05	0.25	0.30	0.44	0.05
C	0.45	0.20	0.80	0.30	0.25	0.40	0.20	0.30	0.10
G	0.05	0.15	0.05	0.05	0.05	0.03	0.58	0.26	0.20

Probability that the GGTCACAACGTTAGG sequence belongs to the region used to derive the above matrix is given by the product of each probability, that is:

$0.05 * 0.15 * 0.05 * 0.30 * 0.65 * 0.40 * 0.12 * 0.20 * 0.10 = 7.02 * 10^{-8}$

in case we assume the first G to correspond to the first position of the signal (table), if we start instead from the second G, the probability is:

$0.05 * 0.15 * 0.80 * 0.45 * 0.25 * 0.32 * 0.12 * 0.30 * 0.20 = 1.56 * 10^{-6}$

and so on…

The AACCCACTA gets a probability of:

$0.30 * 0.50 * 0.80 * 0.30 * 0.25 * 0.32 * 0.20 * 0.44 * 0.65 = 1.65 * 10^{-4}$

To be compared to a random probability (what we'd get by chance)

Probability of independent events

What is the probability to obtain “**6**” with dice-1 & “**5**” with dice-2 ?

DICE-1

DICE-2

It is two **independent events**:
the first event happening does not modify the probability of the second to occur;
The outcome of dice-2 does not depend on the outcome of dice-1, and viceversa.

Probability of independent events

What is the probability to obtain “**6**” with dice-1 & “**5**” with dice-2 ?

DICE-1

DICE-2

The probability of two **indipendent events** is given by the product of the probabilities of the two separate events

Probability of independent events

What is the probability to obtain "**6**" with dice-1 & "**5**" with dice-2 ?

$$probability = \frac{favorable\ cases}{total\ cases}$$

Favorable cases: **1**

Total cases: **6**

Probability = 1/6

DICE-1

DICE-2

The probability of two **indipendent events** is given by the product of the probabilities of the two separate events

Probability of independent events

What is the probability to obtain "**6**" with dice-1 & "**5**" with dice-2 ?

$$probability = \frac{favorable\ cases}{total\ cases}$$

Favorable cases: **1**

Total cases: **6**

Probability = 1/6

DICE-1 DICE-2

The probability of two **indipendent events** is given by the product of the probabilities of the two separate events

Probability of independent events

What is the probability to obtain “**6**” with dice-1 & “**5**” with dice-2 ?

DADO-1

DADO-2

The probability of two **indipendent events** is given by the product of the probabilities of the two separate events

P = 1/6 x 1/6 = 1/36 = 0.02$\overline{7}$

Position-Specific Scoring Matrix (PSSM) describing the *consensus* for a signal

	1	2	3	4	5	6	7	8	9
A	0.30	0.50	0.10	0.45	0.65	0.32	0.12	0.20	0.65
T	0.20	0.15	0.05	0.20	0.05	0.25	0.30	0.44	0.05
C	0.45	0.20	0.80	0.30	0.25	0.40	0.20	0.30	0.10
G	0.05	0.15	0.05	0.05	0.05	0.03	0.58	0.26	0.20
Caso 1 = $7.02 * 10^{-8}$	G 0.05 *	G 0.15 *	T 0.05 *	C 0.30 *	A 0.65 *	C 0.40 *	A 0.12 *	A 0.20 *	C 0.10
Caso 2 = $1.56 * 10^{-6}$	G 0.05 *	T 0.15 *	C 0.80 *	A 0.45 *	C 0.25 *	A 0.32 *	A 0.12 *	C 0.30 *	G 0.20
Caso 3 = $1.65 * 10^{-4}$	A 0.30 *	A 0.50 *	C 0.80 *	C 0.30 *	C 0.25 *	A 0.32 *	C 0.20 *	T 0.44 *	A 0.65

Case 1 GGTCACAACGTTAGG

Case 2 GGTCACAACGTTAGG

……

Case 3 AACCCACTA

Position-Specific Scoring Matrix (PSSM) describing the *consensus* for a signal

	1	2	3	4	5	6	7	8	9
A	0.30	0.50	0.10	0.45	0.65	0.32	0.12	0.20	0.65
T	0.20	0.15	0.05	0.20	0.05	0.25	0.30	0.44	0.05
C	0.45	0.20	0.80	0.30	0.25	0.40	0.20	0.30	0.10
G	0.05	0.15	0.05	0.05	0.05	0.03	0.58	0.26	0.20
Caso 1 = $7.02 * 10^{-8}$	G 0.05 *	G 0.15 *	T 0.05 *	C 0.30 *	A 0.65 *	C 0.40 *	A 0.12 *	A 0.20 *	C 0.10
Caso 2 = $1.56 * 10^{-6}$	G 0.05 *	T 0.15 *	C 0.80 *	A 0.45 *	C 0.25 *	A 0.32 *	A 0.12 *	C 0.30 *	G 0.20
Caso 3 = $1.65 * 10^{-4}$	A 0.30 *	A 0.50 *	C 0.80 *	C 0.30 *	C 0.25 *	A 0.32 *	C 0.20 *	T 0.44 *	A 0.65

To be compared with the probability that sequences belong to a random distribution:

1 2 3 4 5 6 7 8 9

$0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25 * 0.25 = 3.81 * 10^{-6}$

Generally, each value of previous PSSM is divided by 0.25 (the expected random value) and the logarithm base 2 ($\log_2 n$) of their ratio is calculated thus obtaining:

	1	2	3	4	5	6	7	8	9
A	0.26	1.00	-1.32	0.85	1.38	0.36	-1.06	-0.32	1.38
T	-0.32	-0.74	-2.32	-0.32	-2.32	0.00	0.26	0.82	-2.32
C	0.85	-0.32	1.68	0.26	0.00	0.68	-0.32	0.26	-1.32
G	-2.32	-0.74	-2.32	-2.32	-2.32	-3.06	1.21	0.06	-0.32
Caso 1 = -5.76	G - 2.32	G - 0.74	T - 2.32	C + 0.26	A + 1.38	C + 0.68	A - 1.06	A - 0.32	C - 1.32
Caso 2 = -1.29	G - 2.32	T - 0.74	C + 1.68	A + 0.85	C + 0.00	A + 0.36	A - 1.06	C + 0.26	G - 0.32
Caso 3 = 5.44	A + 0.26	A + 1.00	C + 1.68	C + 0.26	C + 0.00	A + 0.36	C - 0.32	T + 0.82	A + 1.38

Problem of 0 count!!! PSEUDOCOUNTS

Is it a biologically meaningful probability?

Se $F=0.25 \rightarrow F/0.25=1 \rightarrow \ln(F/0.25) = 0$

Se $F<0.25 \rightarrow F/0.25<1 \rightarrow \ln(F/0.25) < 0$ neg

Se $F>0.25 \rightarrow F/0.25>1 \rightarrow \ln(F/0.25) > 0$ pos

where F is the frequency of an event, e.g. of observing a given nucleotide at a given position

& 0.25 is the random probability of observing that event

Is it a biologically meaningful probability?

If F=Prandom → F/ Prandom =1 → ln(F/ Prandom) = 0

i.e. if the frequency with whom we observe something is the same that we expect by chance, the corresponding value in a log-odd matrix is “0”

If we draw a dice many many times and we observe “6”

1/6 of times

Is it a biologically meaningful probability?

If F=Prandom → F/ Prandom =1 → ln(F/ Prandom) = 0

i.e. if the frequency with whom we observe something is the same that we expect by chance, the correspondending value in a log-odd matrix is “0”

If we draw a dice many many times and we observe “6”

1/6 of times

This is what we expect by chance

Is it a biologically meaningful probability?

If F< Prandom → F/ Prandom < 1 → ln(F/ Prandom) < 0

i.e. if the frequency with whom we observe something is lower than what we expect by chance, the corresponding value in a log-odd matrix is negative

If we draw a dice many many times and we observe "6"

1/10 of times

The dice is modified

the possibility of obtaining "6" is disfavored as compared to what expected by chance

Is it a biologically meaningful probability?

If F> Prandom → F/ Prandom > 1 → ln(F/ Prandom) > 0

i.e. if the frequency with whom we observe something is higher than what we expect by chance, the corresponding value in a log-odd matrix is positive

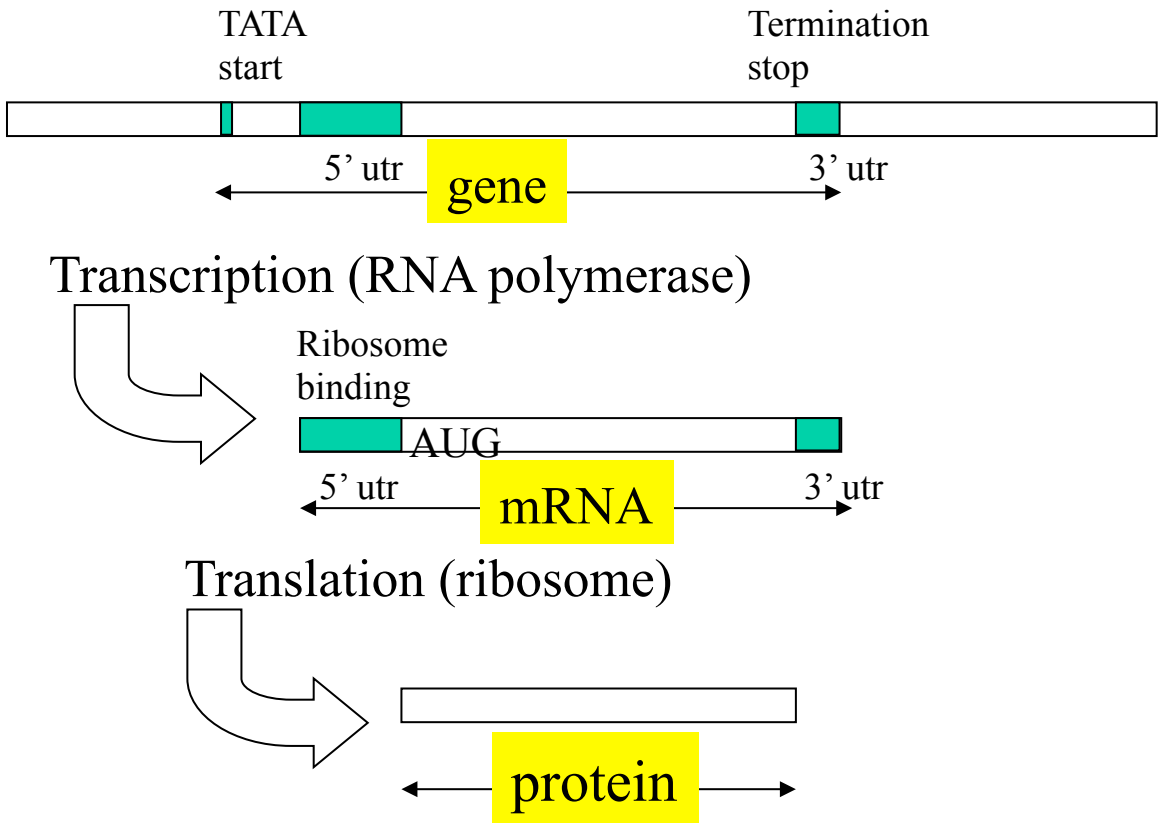
If we draw a dice many many times and we observe "6"

1/2 of times

The dice is modified

the possibility of obtaining "6" is favored as compared to what expected by chance

Structure of a prokaryotic gene



DNA —transcription→ RNA —translation→ protein

Prokaryotic genes

Prokaryotic genes are less complex:

Promoter - 5'utr – AUG – **ORF** – stop codon- 3'utr

Prokaryotic genes

Prokaryotic genes are less complex:

Promoter - 5'utr – AUG – **ORF** – stop codon- 3'utr

Recognition site for
RNA polymerase

Prokaryotic genes

Prokaryotic genes are less complex:

Promoter - 5'utr – AUG – **ORF** – stop codon- 3'utr

Utr: Untranslated Regions (for the mRNA-ribosome binding)

Prokaryotic genes

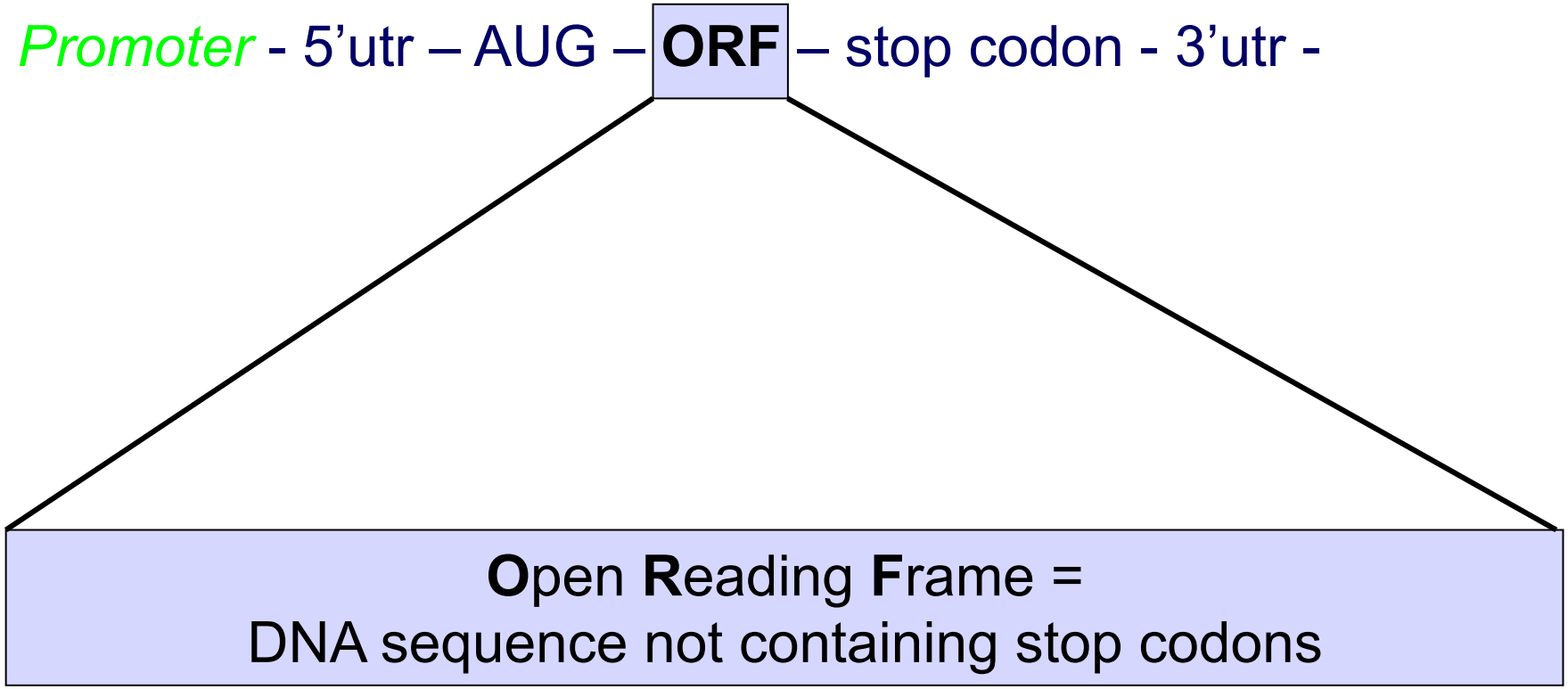
Prokaryotic genes are less complex:

Promoter - 5'utr – **AUG** – **ORF** – stop codon - 3'utr

Start codon

Prokaryotic genes

Prokaryotic genes are less complex:



Prokaryotic genes

Prokaryotic genes are less complex:

Promoter - 5'utr – AUG – **ORF** – stop codon- 3'utr

This region is translated into protein

Please note that occasionally translation does not start at the first encountered AUG sequence

Long ORFs in prokariotes

If we assume the codons to be uniformly (randomly) distributed, then we expect a stop codon every 3/64 codons (≈1/21 codons).

However, proteins correspond on avergae to roughly 1000 bp (≈330 codons/aa) and each coding region must contain only one stop codon (≈1/330 codons).

The **random probability** of finding a N-codon long ORF preceeded by ATG and followed by a stop codon can be calculated as:

1 start codon: *AUG*

61 continuing codons (ORF)

3 stop codons (termination)

Probability of independent events

What is the probability to obtain “**6**” with dice-1 & “**5**” with dice-2 ?

DICE-1

DICE-2

The probability of two **indipendent events** is given by the product of the probabilities of the two separate events

P = 1/6 x 1/6 = 1/36 = 0.027̄

Long ORFs in prokaryotes

If we assume the codons to be uniformly (randomly) distributed, then we expect a stop codon every 3/64 codons (≈1/21 codons).

However, proteins correspond on average to roughly 1000 bp (≈330 codons/aa) and each coding region must contain only one stop codon (≈1/330 codons).

The **random probability** of finding a N-codon long ORF preceeded by ATG and followed by a stop codon can be calculated as:

Long ORFs in prokaryotes

For N=10

$$P=(1/64) \times (61/64)^{10} \times (3/64) = 0.045$$

For N=100

$$P=(1/64) \times (61/64)^{100} \times (3/64) = 0.0006$$

Standard ORF length for searches ranges between 18 bp and 10^3 bp

Especially when combined with signals search the ORF prediction gives a low error rate in prokaryotes

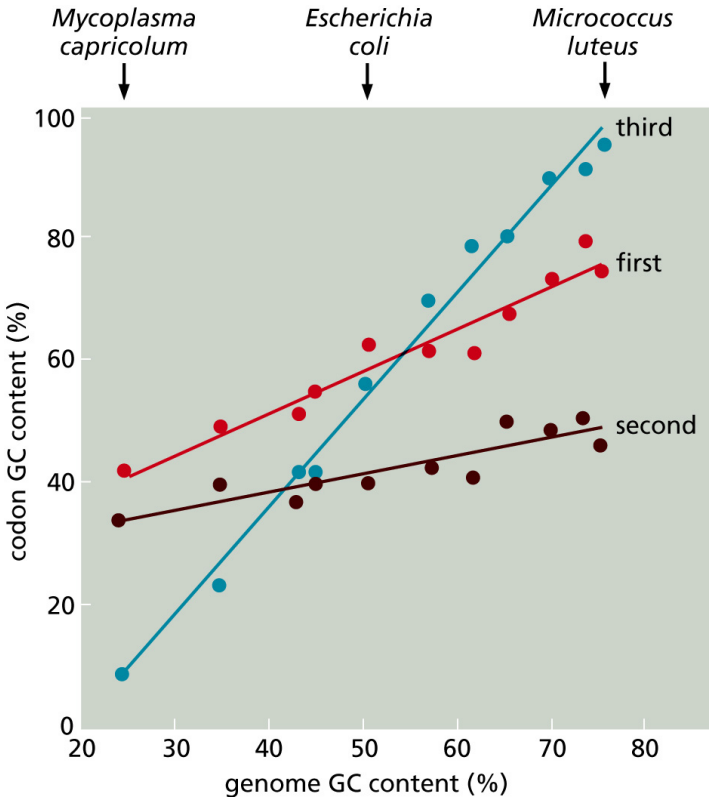
Differences in the nucleotide composition

Usually the GC-content (%) is biased in coding and non coding regions of the DNA of a given organism

The specific GC-content (%) is actually organism-specific

In addition, the GC-content (%) is biased at different codon positions (see Figure)

Knowing the specific GC-content (%) for an organism helps to locate missing genes



At the 1st & 2nd position of bacteria codons the GC-content is usually lower than that of the whole DNA

Difference in the nucleotide composition between coding & non coding DNA

Codon usage: preference for given codons, within the synonymous ones is specific to many organisms

Amino acid composition of the potential protein product…..

Differences in the nucleotide composition

Codon occurrences are distinct for coding (gene) and non coding (intergene) sequences and can be used as a gene-prediction feature

Codon frequencies are influenced by the amino acid composition of the encoded protein and by the codon bias imposed by the tRNAs availability

Differences in the nucleotide composition

Codon occurrences are distinct for coding and non coding sequences and can be used as a gene-prediction feature

Codon frequencies are influenced by the amino acid composition of the encoded protein and by the codon bias imposed by the tRNAs availability

In what frequency ratio do we find alanine and tryptophan in known protein sequences?

Differences in the nucleotide composition

Codon occurrences are distinct for coding and non coding sequences and can be used as a gene-prediction feature

Codon frequencies are influenced by the amino acid composition of the encoded protein and by the codon bias imposed by the tRNAs availability

Example 1: L, A e W (leucine, alanine & tryptophan) are coded by 6, 4 & 1 different codons respectively, therefore in a random DNA translation such amino acids should be in the 6:4:1 ratio. However, in a protein usually a 6:5:1 ratio is observed ➔ this implies that the coding DNA should contain the above codons in the 6:5:1 ratio

Relevance of non-coding RNAs

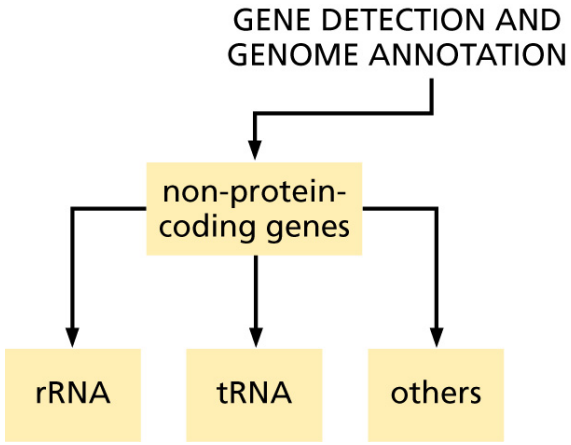
Genomes (especially eukaryotic ones) code for a large number of RNA molecules non coding for proteins (non-coding RNA, ncRNA)

Based on their dimension, ncRNAs are classified as:

sncRNA (short ncRNA, ≤ 200 bp), OR

lncRNA (long ncRNA, > 200 bp)

ncRNAs play a variety of functions, especially in regulating the gene expression at an *epigenetic and post-transcriptional level, affecting the cell development and differentiation*



Relevance of non-coding RNAs

Genomes (especially eukaryotic ones) code for a large number of RNA molecules non coding for proteins (non-coding RNA, ncRNA)

Based on their dimension, ncRNAs are classified as:

sncRNA (short ncRNA, ≤ 200 bp), OR

lncRNA (short ncRNA, > 200 bp)

ncRNAs play a variety of functions, especially in regulating the gene expression at an *epigenetic and post-transcriptional level, affecting the cell development and differentiation*

In the currently annotated version of the human genome:

≈ 22,000 genes "coding" for ncRNA (most probably an underestimate, up to 140,000 are hypothesized)

≈ 20,000 genes coding for proteins

rRNA and tRNA sequences are located before the protein-coding genes are

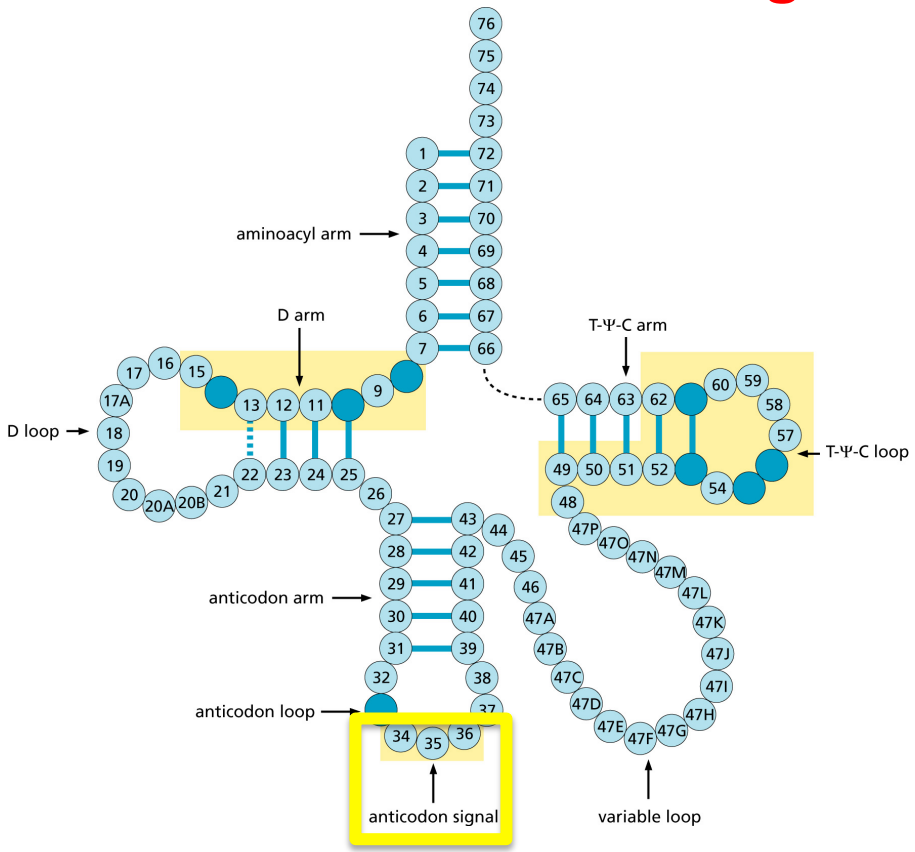
Genes not translated into proteins are generally located before the gene-detection methods discussed in the following are applied

rRNA (ribosomal RNA) sequences are well characterized and and well conserved, so they can be detected by sequence similarity quite easily

tRNA (transfer RNA) genes (important regulatory non-coding RNAs, over 50 in bacterial genomes) can also be located for instance by a decision tree

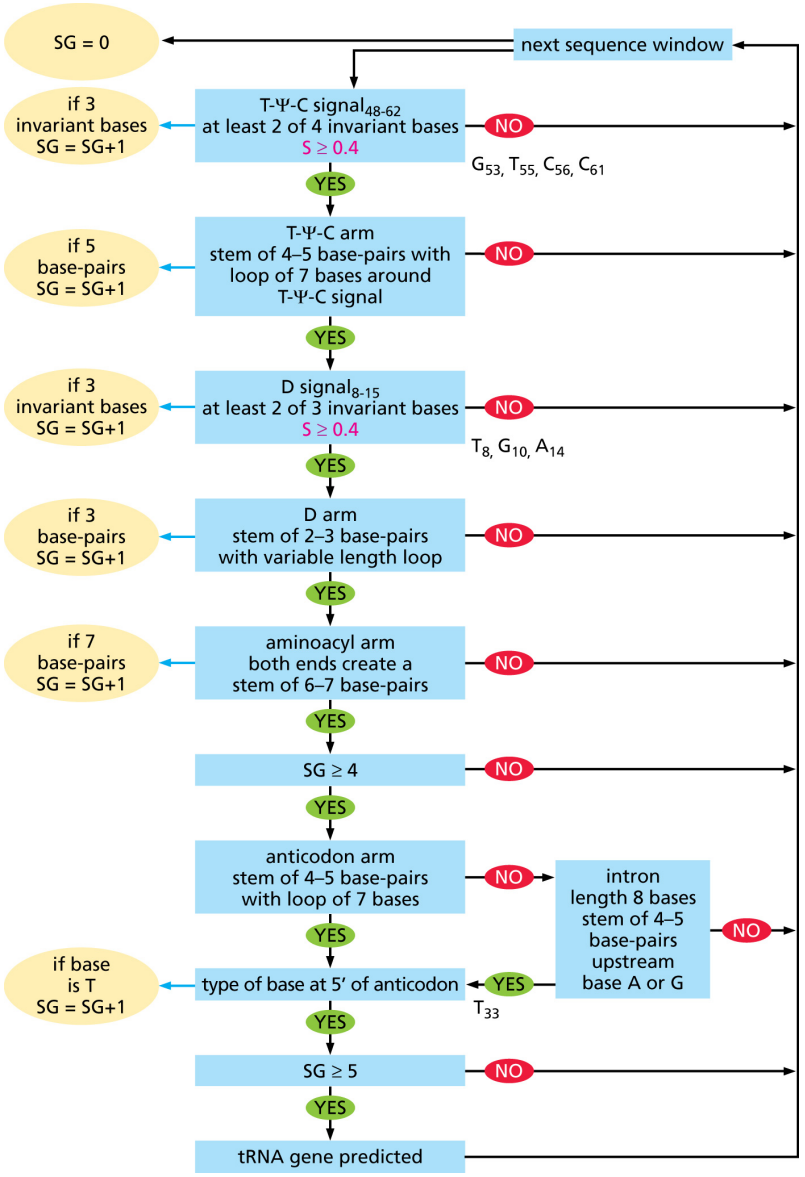
Knowing the set of tRNA molecules available to a genome is relevant as it is related to the specific codon usage

Finding tRNA genes



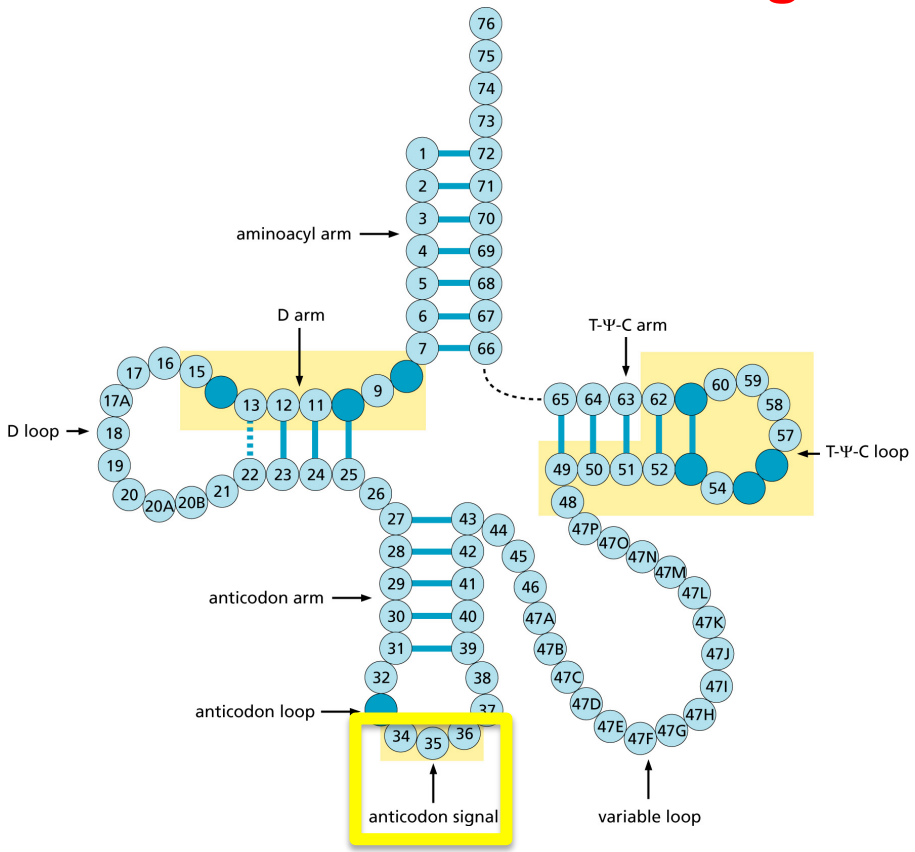
tRNA has a cloverleaf structure with several invariant or conserved elements

They are used in decision trees which predict or not tRNA genes based on their presence

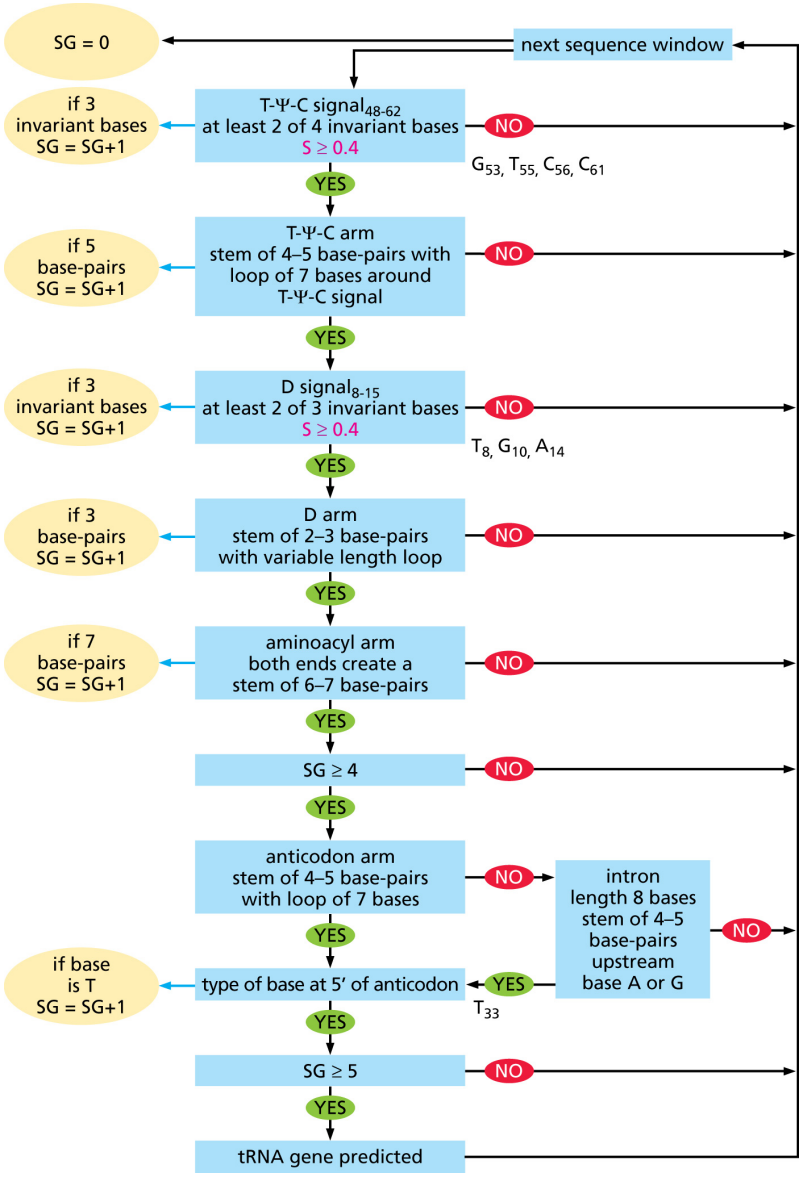


Example of decision tree for tRNA genes detection with added a general score (SG)

Finding tRNA genes



The prediction error rate is estimated as one base per several megabases, implying very few false positives for prokaryotes but not for eukaryotes (for which combined approaches are needed)



Example of decision tree for tRNA genes detection with added a general score (SG)

Extrinsic gene finding methods

The putative gene sequence (DNA) can be submitted for homology searches through databases of known sequences for DNA or proteins (gene products)

Search against **DNA** and **protein** databases, in case of close homologs present, can confirm the sequence to be a gene, and provide a clue about its structure (especially for eukaryotes) and its function

A certain fraction of genes will correspond to ORFans, i.e. they won’t have any homolog in the databases

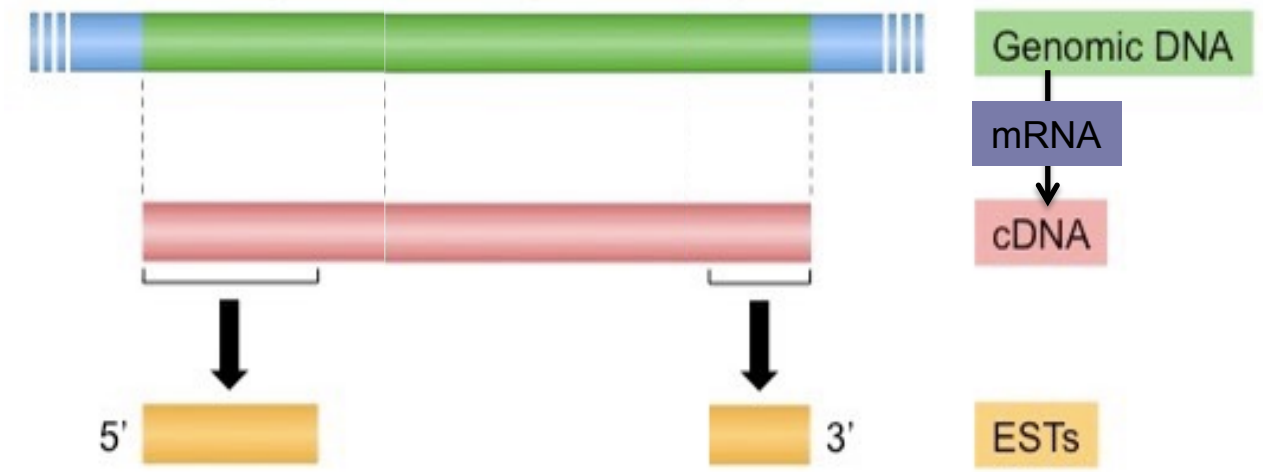
Extrinsic gene finding methods

The putative gene sequence (DNA) can also be submitted for searches through databases of **EST** sequences

ESTs, i.e. **Expressed Sequence Tags** are sub-sequences (usually 200 to 500 nts long) of complementary DNA (cDNA), usually at the 3' side

cDNA is DNA synthesized from a single-stranded mRNA (ready to be translated) by the enzyme reverse transcriptase, it has the sequence of (part of the) gene that will be translated

A match with an EST sequence thus implies that the query DNA sequence will be translated in the cell (i.e. it is a gene)



Extrinsic gene finding methods

The putative gene sequence (DNA) can also be submitted for searches through databases of **EST** sequences

ESTs, i.e. **Expressed Sequence Tags** are sub-sequences (usually 200 to 500 nts long) of complementary DNA (cDNA), usually at the 3' side

cDNA is DNA synthesized from a single-stranded mRNA (ready to be translated), it has the sequence of the (part of the) gene that will be translated

Searches against EST databases should be performed with the same organism of the target sequence(s); if a significant match is found, translated regions of genes can be identified

Genes complexity

- *Coding genes* in 100Kb (Kb=10^3 nucleotide bases)
  - *E. Coli* 87 } ***prokaryotes***
  - *S. Cerevisiae* 52 }
  - *C. elegans* 22 } ***eukaryotes***
  - *H. sapiens* 5 }

The "gene density" is much lower in eukaryotes and especially in complex organisms

This reflects the different complexity of the genomes (and genes) themselves

Prokaryotic genomes

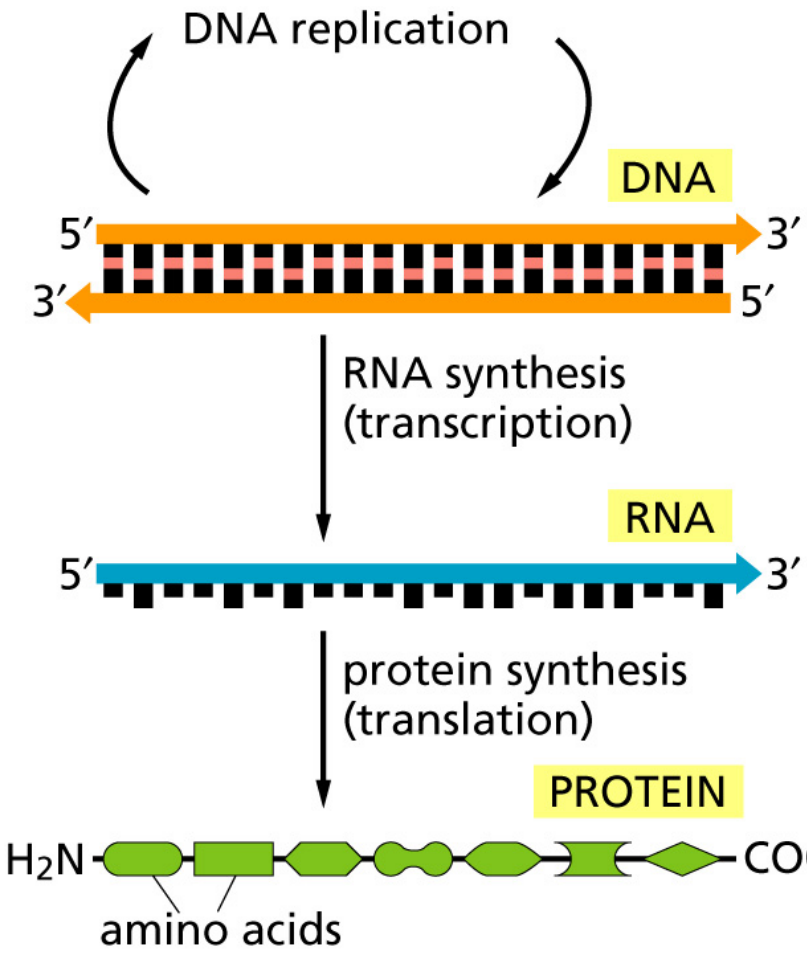
- Small dimensions (< 10 Mb)
- High gene density (>85% of the genome is coding)
- Low nucleotide redundancy
- Introns (quasi) absent
- Low complexity of gene and genome architecture
- Easy to be sequenced

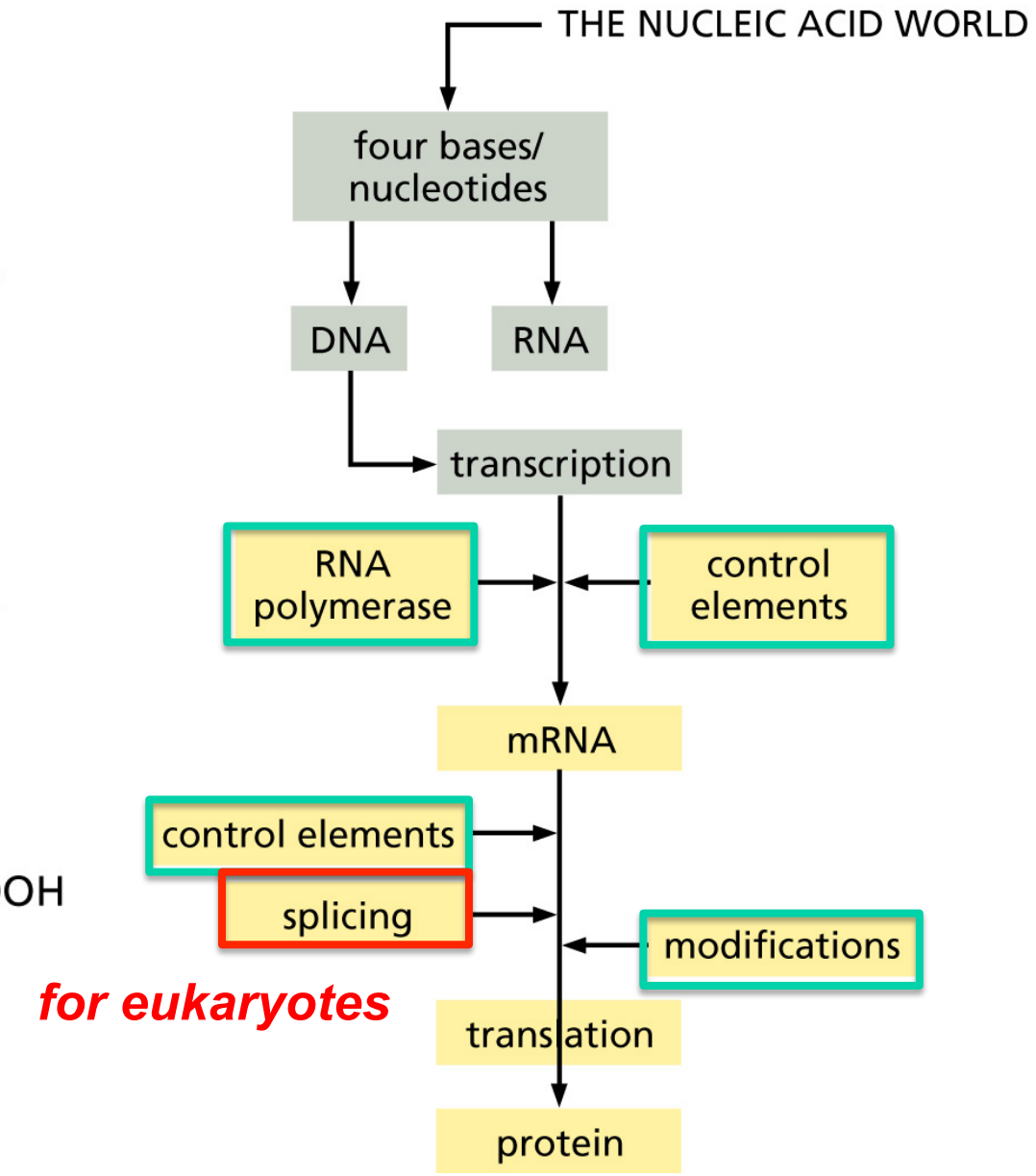
Eukaryotic genomes

- Large dimensions: from 13 Mb for the simplest fungi to 10.000 Mb for higher plants
- Low gene density (coding DNA: 70% in *S. Cerevisiae*, 25% in *D. Melanogaster*, 1-3% in vertebrates and higher plants
- High nucleotide redundancy
- High presence of **introns**
- High complexity of gene and genome architecture

Molecular biology central dogma: a more realistic view

B





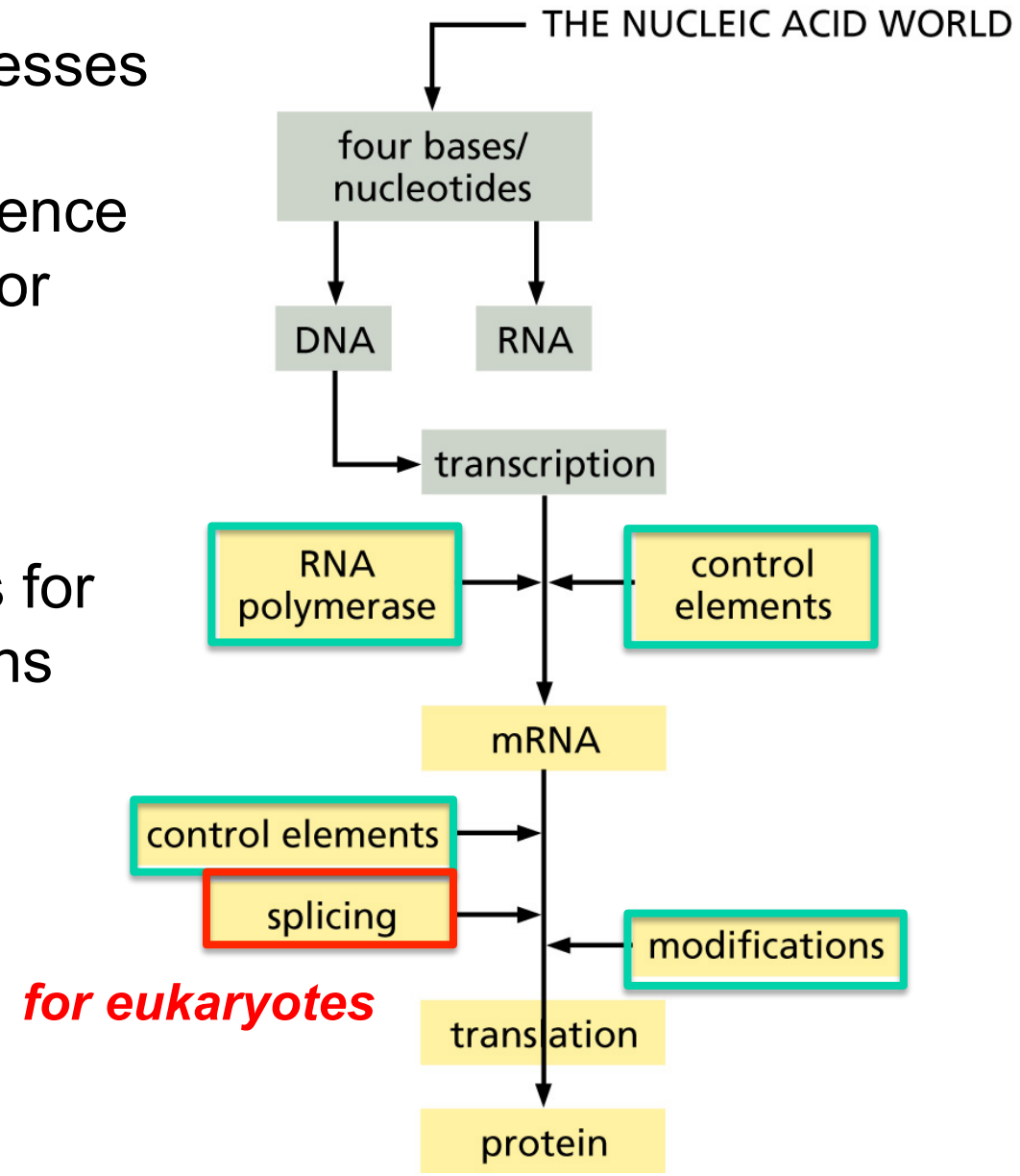
for eukaryotes

Molecular biology central dogma: a more realistic view

B

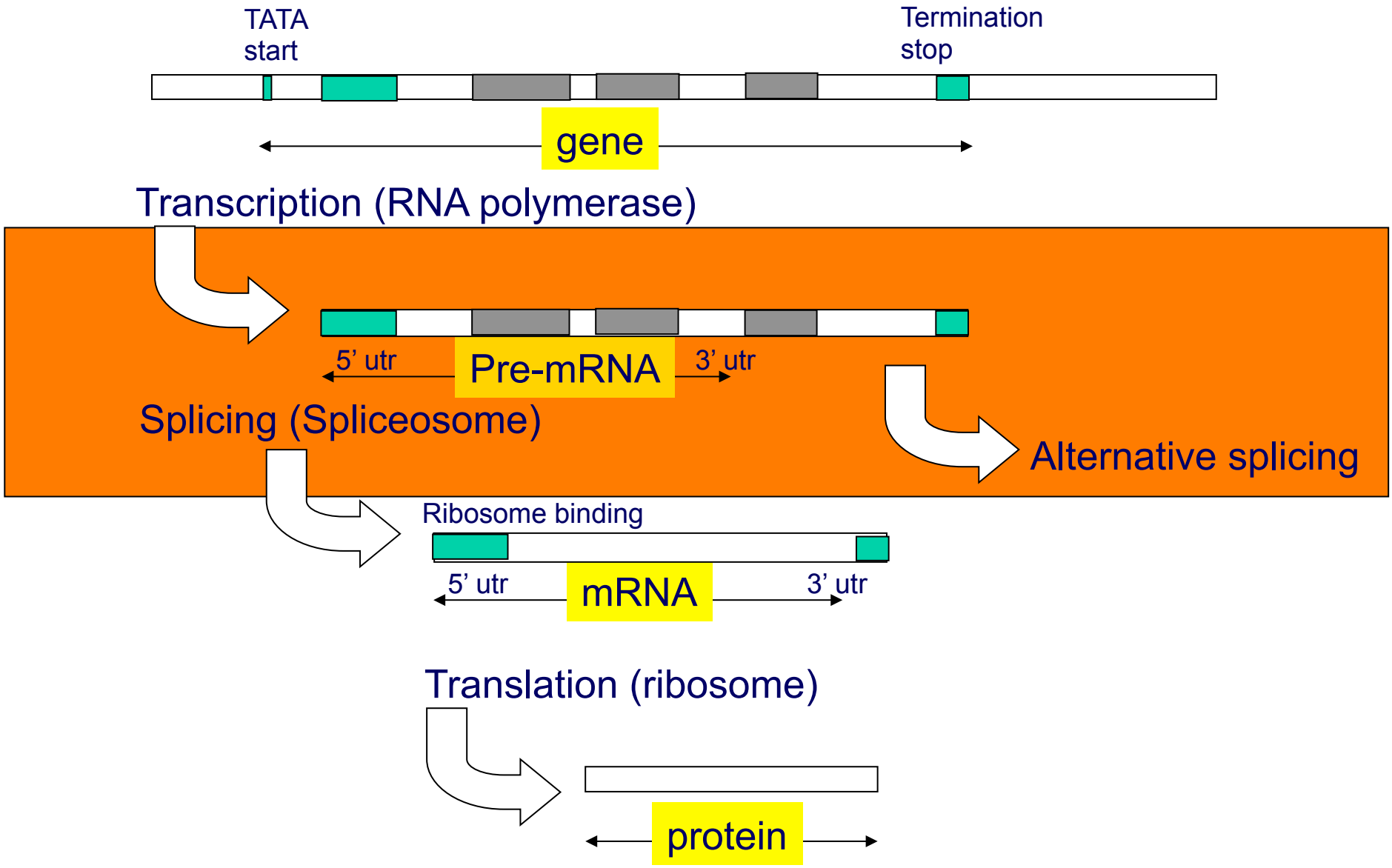
The regulation of many processes involved in transcription and translation relies on the presence of short **signal sequences** (or regulatory elements)

The majority of these **signal sequences** are binding sites for specialized regulatory proteins that interact with DNA to regulate transcription, RNA maturation and translation



for eukaryotes

Structure of an eukaryotic gene



Eukaryotic genes

Eukaryotic genes are more complex:

Promotor - 5'utr - Exon - Intron - Exon - … 3'utr

Eukaryotic genes

Eukaryotic genes are more complex:

Promotor - 5'utr - Exon - Intron - Exon - … 3'utr

Recognition site for
RNA polymerase

Eukaryotic genes

Eukaryotic genes are more complex:

Promotor - 5'utr - Exon - Intron - Exon - … 3'utr - $A(A)_nA$

Utr: Untranslated Regions (for the mRNA-ribosome binding)

Poly(A) tail

role in mRNA stabilization/ translation

Eukaryotic genes

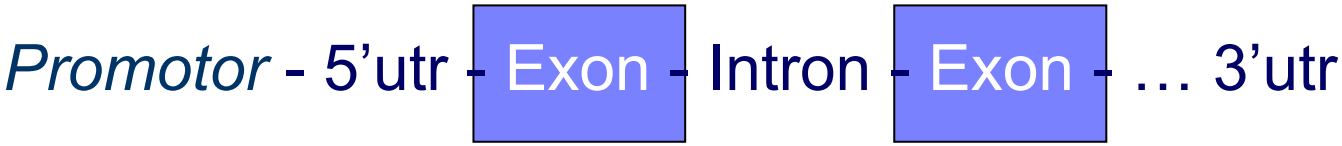
Eukaryotic genes are more complex:

Promotor - 5'utr - Exon - Intron - Exon - … 3'utr

these regions are
translated in proteins

Eukaryotic genes

Eukaryotic genes are more complex:



these regions are
translated in proteins

Eukariotic genes contain **esons** and **introns**, both of them are transcribed in RNA, but:

exons contain codons that are then translated into protein sequences

introns (intervening regions) are non coding sequences (thus not translated in proteins) 'sandwiched' between two exons

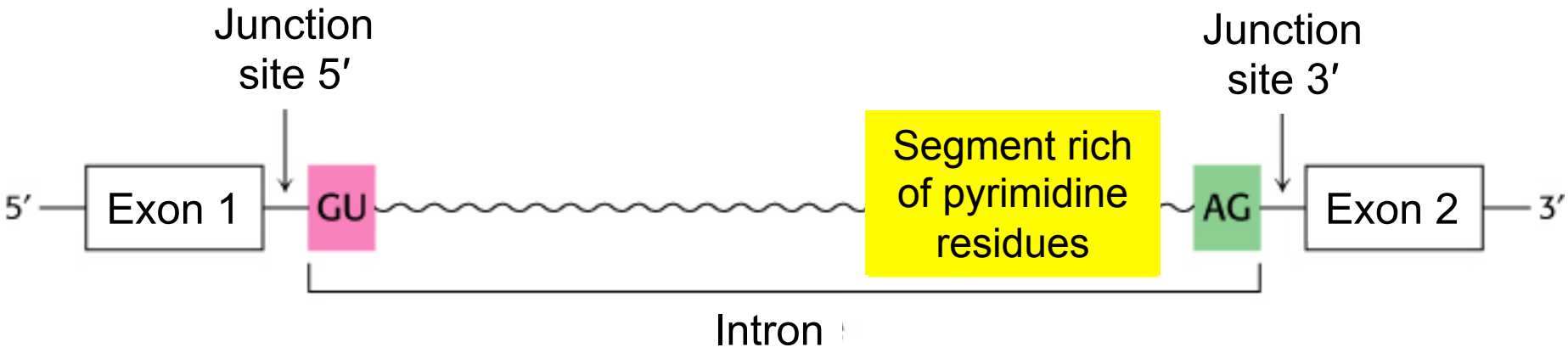
Eukaryotic genes

Eukaryotic genes are more complex:

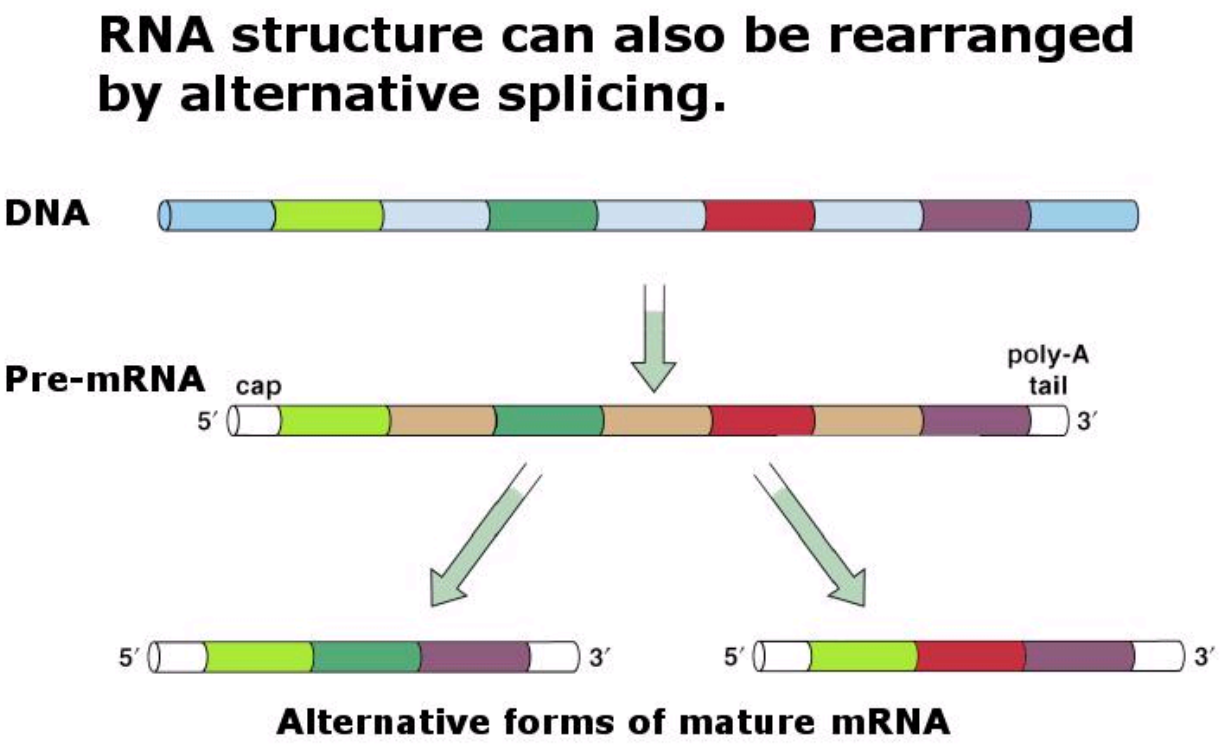
Promotor - 5'utr - Exon - Intron - Exon - … 3'utr

An **intron** can contain stop codons

An **intron** starts with a GU sequence and ends with a AG sequence (the intron signal is thus made of only 2 bp)



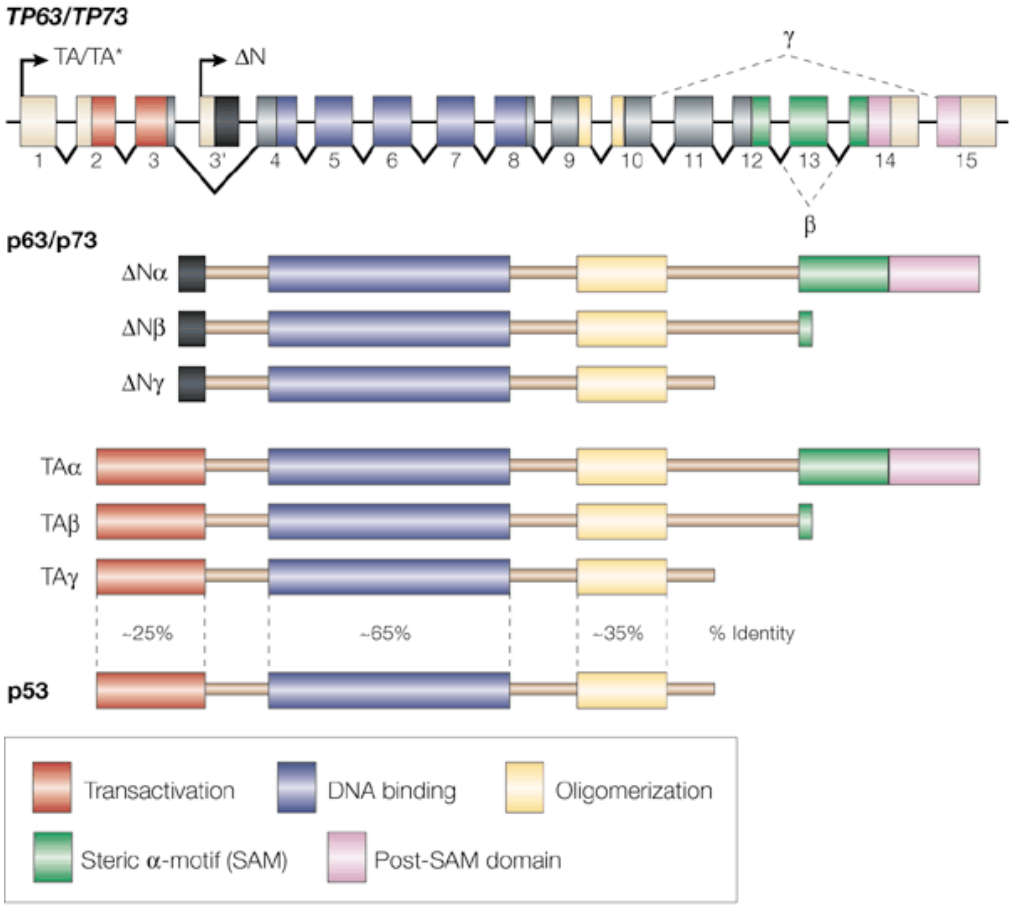
Alternative splicing



Multi-exon genes may undergo alternative splicing, i.e. they can combine differently their exons to encode proteins with different functions

Up to 95% of human multi-exon genes undergo alternative splicing. Moreover, around 15% of human hereditary diseases and cancers are associated with alternative splicing.

Alternative splicing



Nature Reviews | Molecular Cell Biology

An example: p53, p63 and p73 genes encode for multiple proteins containing different protein domains (isoforms) due to multiple splicing, alternative promoter and alternative initiation of translation. The different isoforms have a role in development and cancer

Alternative splicing is common in mammals; it is thought to allow a relatively small number of genes in the genomes to specify a much larger number of proteins

Genes complexity

- *Coding genes* per 100Kb
  - *E. coli* 87 } prokaryotes
  - *S. cerevisiae* 52 } eukaryotes
  - *C. elegans* 22
  - *H. sapiens* 5

Genes complexity

- *Coding genes* per 100Kb
  - *E. coli* 87 } prokaryotes
  - *S. cerevisiae* 52 } eukaryotes
  - *C. elegans* 22
  - *H. sapiens* 5

- Average N of introns | gene *l*. (Kb) | mRNA *l*. (Kb)
  - S. Cerevisiae 0 | 1.6 | 1.6
  - C. elegans 3 | 4.0 | 3.0
  - Drosophila 3 | 11.3 | 2.7
  - Gallus 8 | 13.9 | 2.4
  - H. sapiens 6 | 16.6 | 2.2

} eukaryotes

Eukaryotic genes: exceptions

Promotor - 5'UTR - Exon - Intron - Exon - … 3'UTR

Dystrophin gene is the largest known human gene = 2.4 Mb

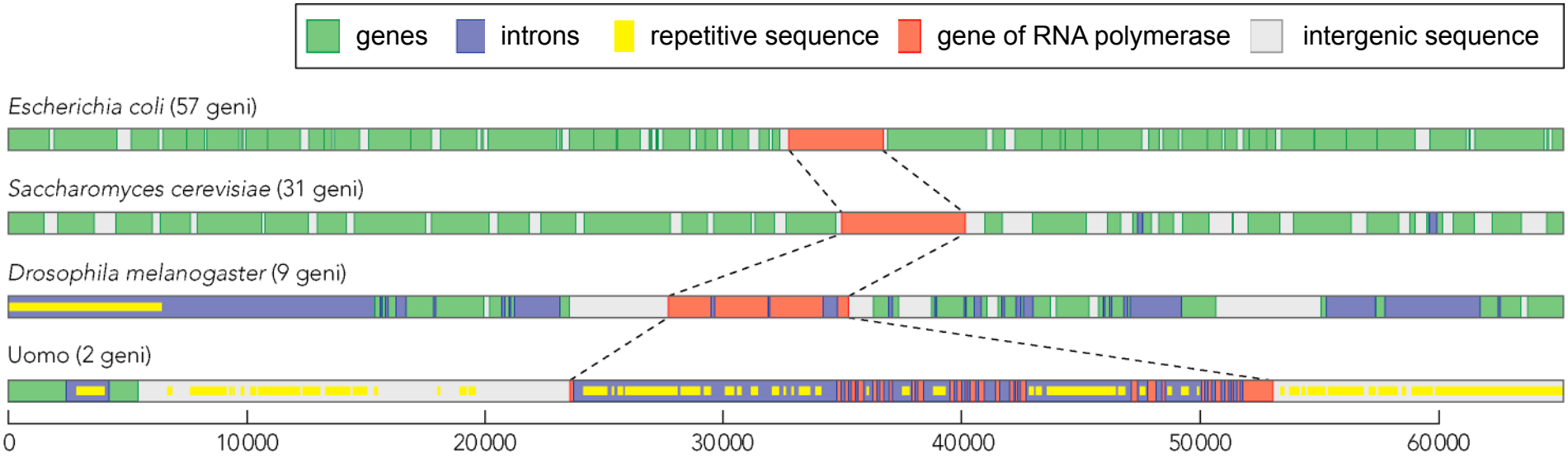
Blood clotting factor VIII = 26 exons from 69bp to 3106bp for a total of 186Kb, plus 32.4 Kb of introns.

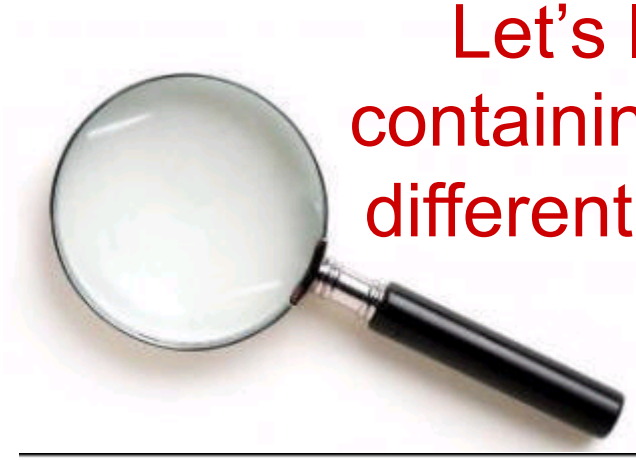
It is like a file made of 26 parts would be interrupted by long pages of irrelevant information

The 5' utr sequence is long on average 750bp, but can be much longer.

The 3' utr sequence is long on average ≈450 bp, but can be over 5Kb, like the *Kallman's syndrome* gene.

Let's look in detail a region of **~ 65 Kb** containing the gene of RNA polymerase in 4 different genomes: *E. coli, S. cerevisiae, D. melanogaster e H. sapiens*





Let's look in detail a region of **~ 45 Kb** containing the gene of RNA polymerase in 4 different genomes: *E. coli, S. cerevisiae, D. melanogaster e H. sapiens*

introns | repetitive sequence | gene of RNA polymerase | intergenic sequence

E. coli

S. cerevisiae

D. melanogaster

H. sapiens

20000 30000 40000 50000 60000

Methods for gene search in eukaryotes

The main aim of the gene search is being able to distinguish between coding and non coding DNA and locate genes

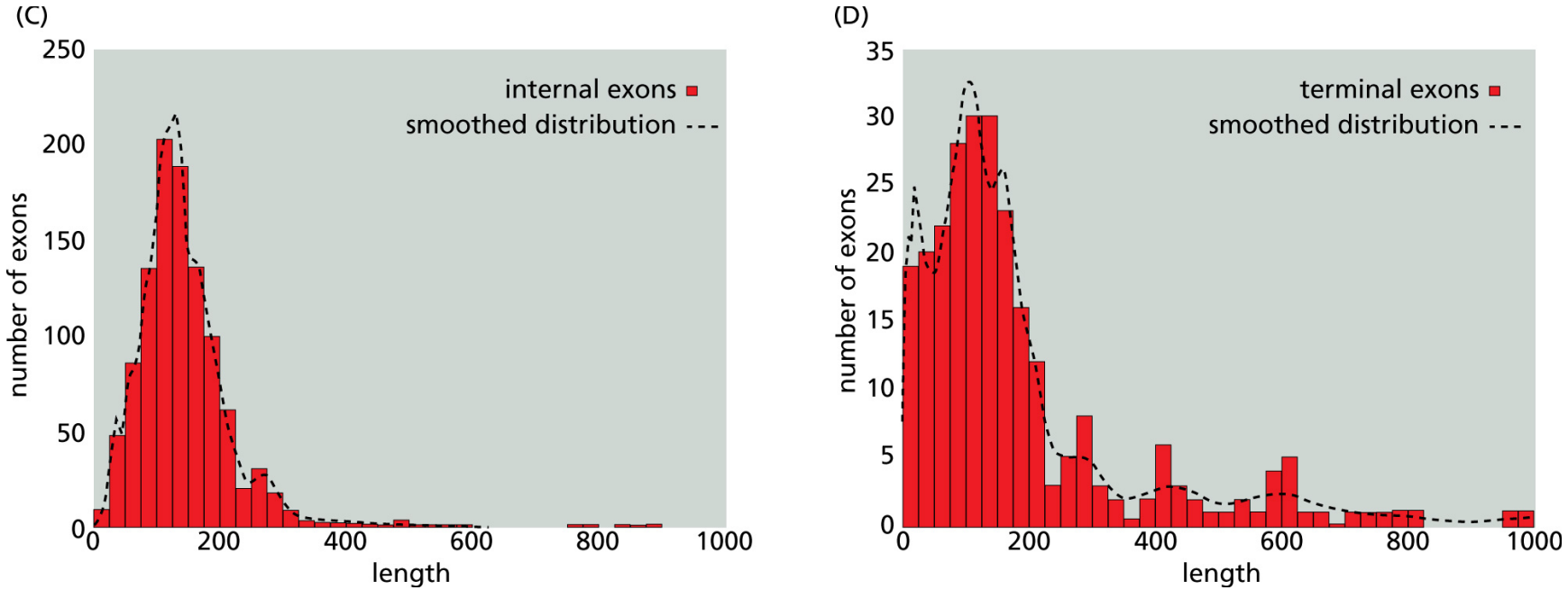
1. Identifying ORFs (Open Reading Frames) – not feasible
2. *Signals* search (e.g. promoters and splicing sites)
3. Difference in the *nucleotide composition*
4. Homology search (sequence similarity)

The first three approaches are intrinsic, the last one is extrinsic

Methods for gene search in eukaryotes

1. Identifying ORFs (Open Reading Frames) – not feasible

Exons are on average 100-200 (some even ≈12) nts long, thus considerably shorter than prokaryotic genes



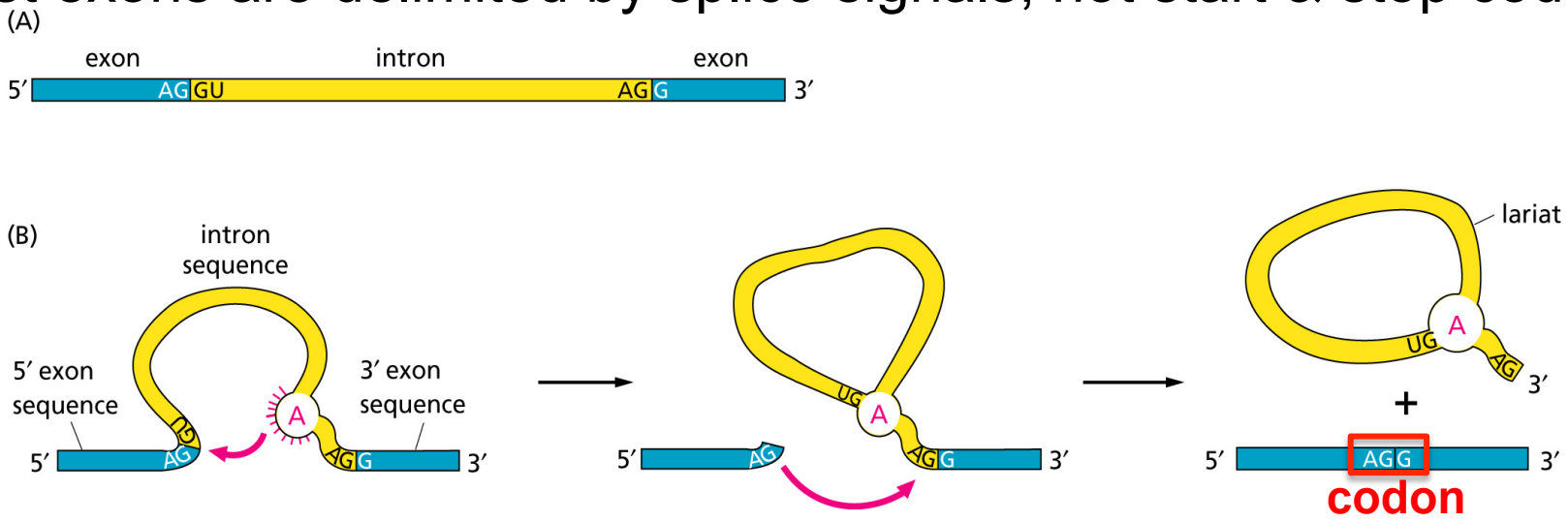
for vertebrates

Methods for gene search in eukaryotes

1. Identifying ORFs (Open Reading Frames) – not feasible

Exons are on average 100-200 nts (some just ≈12 nts) long, thus considerably shorter than prokariotic genes

Most exons are delimited by splice signals, not start & stop codons



Moreover, the initial and final exons will contain some untranslated (5' and 3' utr) regions

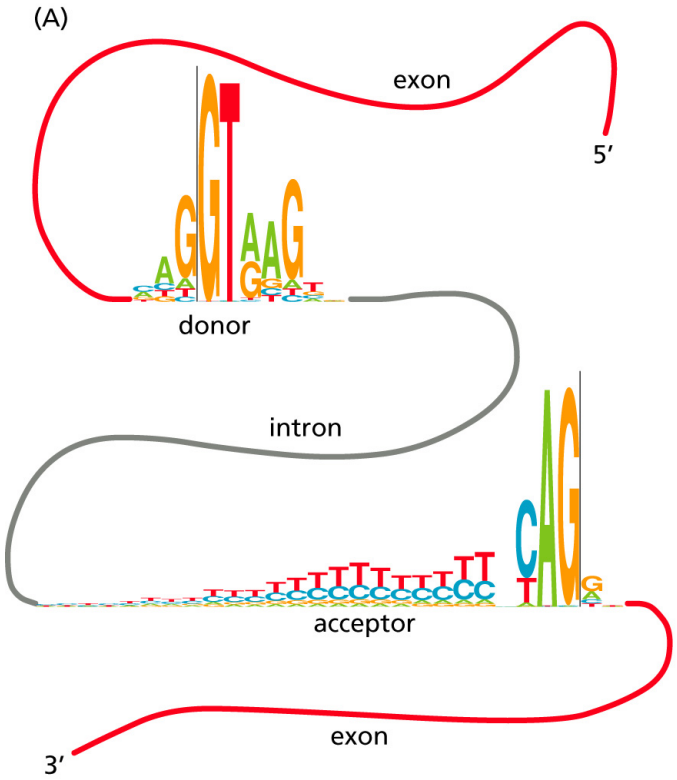
Therefore ORF prediction in eukaryotes can result in an extremely high false prediction rate

Methods for gene search in eukaryotes

2. *Signals* search (e.g. promoters and splicing sites)

Characterization of eukaryotic genes requires a prediction of their splicing structure

The signals for exon-intron (donor) and intron-exon (acceptor) junction sites are dominated by dinucleotides (GT & AG), with weak signals extending mostly on the intron side, which can be species-specific



Trying to identify splice sites solely on the basis of the GT/AT dinucleotides would lead to many false positive predictions

Methods for gene search in eukaryotes

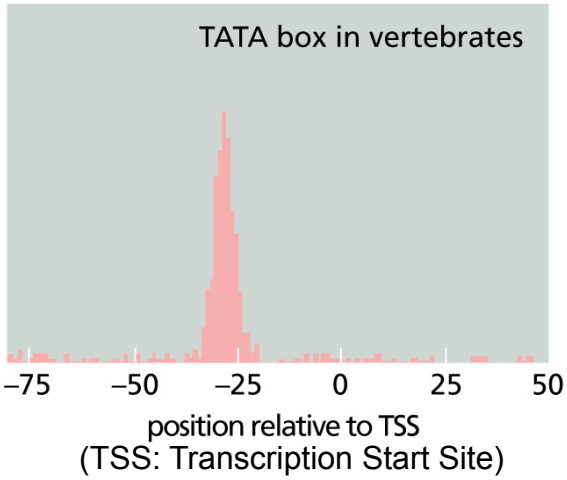
2. *Signals* search (e.g. promoters and splicing sites)

In eukaryotes, there are (core) promoter regions for the RNA polymerase binding, which are typically at a short distance from the transcription start site (e.g. the TATA box) and are quite well-defined

(A)

		sequence position														
		-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
observed bases	A	61	16	352	3	354	268	360	222	155	56	83	82	82	68	77
	C	145	46	0	10	0	0	3	2	44	135	147	127	118	107	101
	G	152	18	2	2	5	0	20	44	157	150	128	128	128	139	140
	T	31	309	35	374	30	121	6	121	33	48	31	52	61	75	71
weight matrix	A	-1.02	-3.05	0.00	-4.61	0.00	0.00	0.00	0.00	-0.01	-0.94	-0.54	-0.48	-0.48	-0.74	-0.62
	C	-0.28	-2.06	-5.22	-3.49	-5.17	-4.63	-4.12	-3.74	-1.13	-0.05	0.00	-0.05	-0.11	-0.28	-0.40
	G	0.00	-2.74	-4.38	-4.61	-3.77	-4.73	-2.65	-1.50	0.00	0.00	-0.09	0.00	0.00	0.00	0.00
	T	-1.68	0.00	-2.28	0.00	-2.34	-0.52	-3.65	-0.37	-1.40	-0.97	-1.40	-0.82	-0.66	-0.54	-0.61
consensus		G/C	T	A	T	A	A/T	A	A/T	G/A	G/C	C/G	G/C	G/C	G/C	G/C

(B)



Methods for gene search in eukaryotes

2. *Signals* search (e.g. promoters and splicing sites)

Furthermore, there are additional extensive promoter regions that may cover thousands of nts and contain binding sites for other gene-regulatory proteins that determine when and where the gene will be switched on in the organism

Control regions can be found either upstream or downstream of genes

The translation stop signal is one of the 3 stop codons, while the transcription stop signal is the polyadenylation signal (tail), with the consensus AATAAA

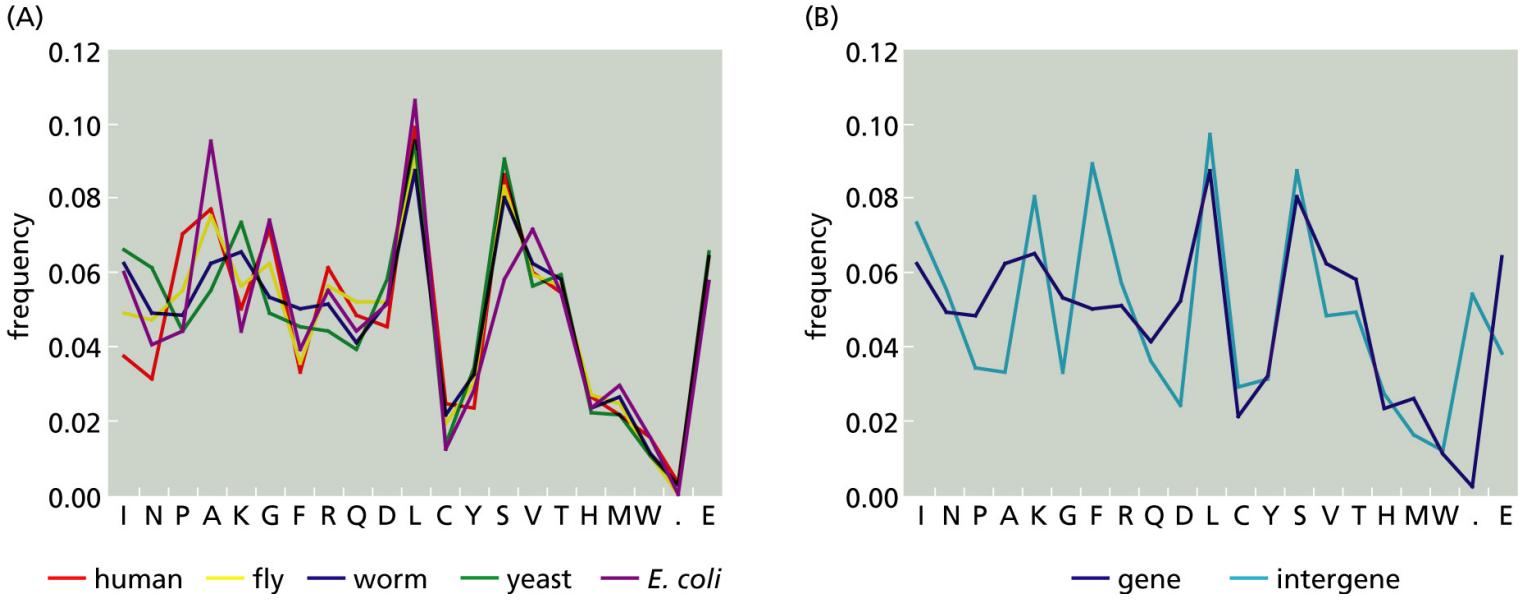
As prediction of signals in eukaryotes is more difficult, they are usually searched for not separately but contextually (relatively to each other)

Methods for gene search in eukaryotes

3. Difference in the *nucleotide composition*

The relative frequency of codon occurrence are different for coding and non coding sequences and can be used as a gene-prediction feature

Applications for gene finding focus on the frequency distribution of six-nucleotide sequences – hexamers or dicodons-, as it is more efficient



Difference in the nucleotide composition

3. Difference in the *nucleotide composition*

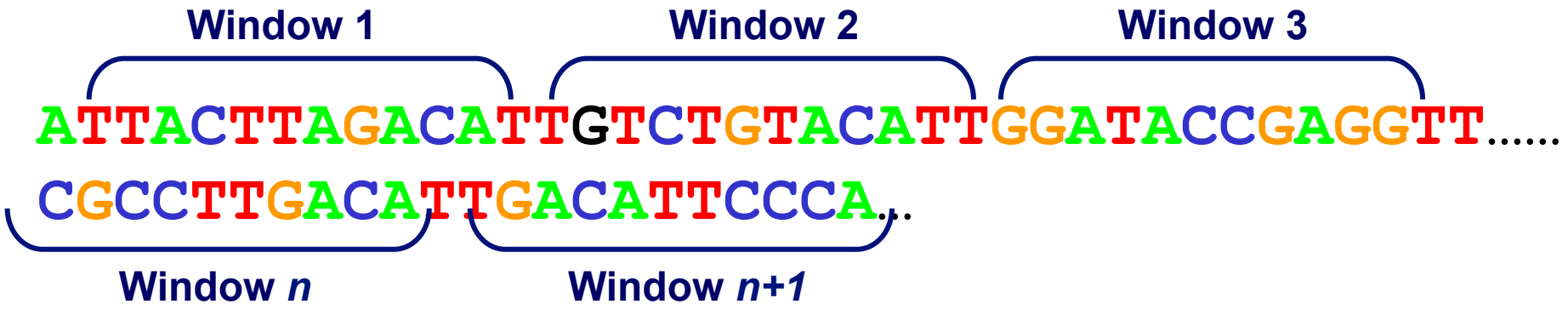
Coding and non-coding DNA significantly differ in their *hexameric frequency and dinucleotide dependence*

Hexameric frequency = frequency with whom a specific stretch of 6 nts is present in a DNA sequence

Nucleotide dependence = conservation of the distance between 2 given nucleotides in a DNA sequence (e.g. AXXXXT)

The hexameric frequency (EF) and nucleotide dependence (ND) are calculated for windows of a specific size and compared to each other.

Difference in the nucleotide composition



We obtain n+1 pairs of values:

EF-1	***EF-2***	***EF-3***	***EF-n***	***EF-n+1***
DN-1	***DN-2***	***DN-3***	***DN-n***	***DN-n+1***

We localize the gene:

FE-1 ~ FE-2<< FE-3 ~ FE-n >> FE-n+1
DN-1 ~ DN-2 << DN-3 ~ DN-n >> DN-n+1

The gene is identified because values in windows ***3*** to ***n*** are typical of coding regions

Methods for gene search in eukaryotes

4. Homology search (sequence similarity)

Exons (and promoters) are expected to have a greater similarity to the genome of related organisms as compared to introns

Furthermore, predicted exons can be translated to segments of protein sequences and used for an homology search *vs* a database of protein sequences: this will reliably confirm the prediction

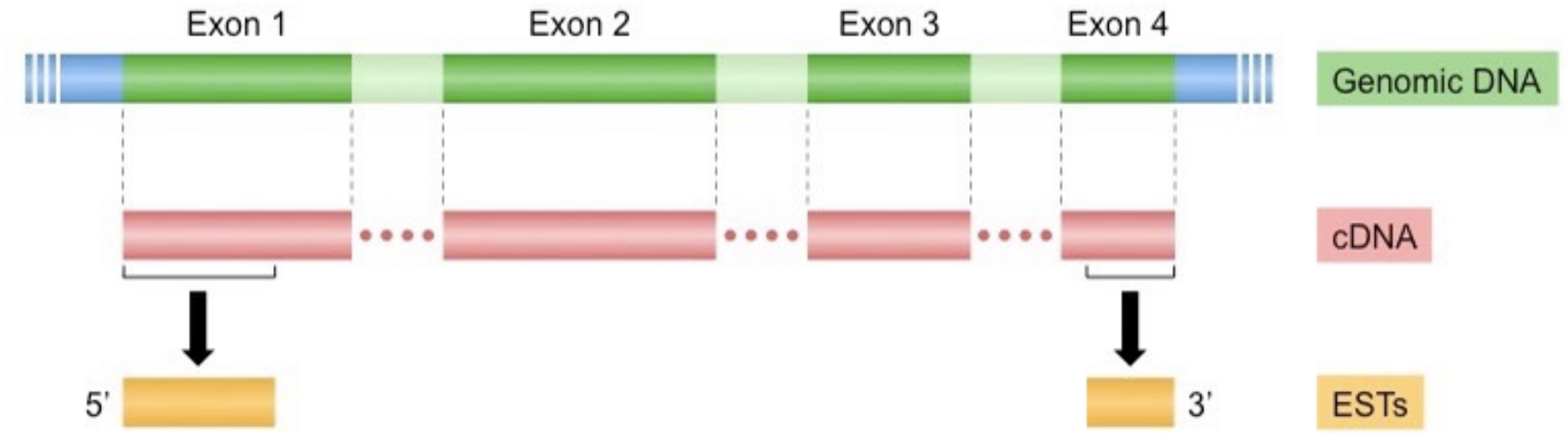
Homology to ESTs and cDNAs can also be used to confirm exon prediction

Methods for gene search in eukaryotes

4. Homology search (sequence similarity)

When the ESTs and cDNAs are from the same organism they can help identify a gene and its exon structure

EST databases can also be useful in providing evidence of alternative splicing



Eukaryotic genes: similarity-based (extrinsic) methods

- Comparison with EST and cDNA databases
- Comparison of the genomic sequence with related (homologous) genomic sequences
- Comparison of the translated genomic sequences *vs.* protein databases, also *via* **spliced alignment**

Given a genomic sequence and an ensemble of putative exons (all the possible fragments between donator and acceptor splicing sites):

1. Building all the possible exon chains

2. Translating the corresponding mRNA into the corresponding protein sequence

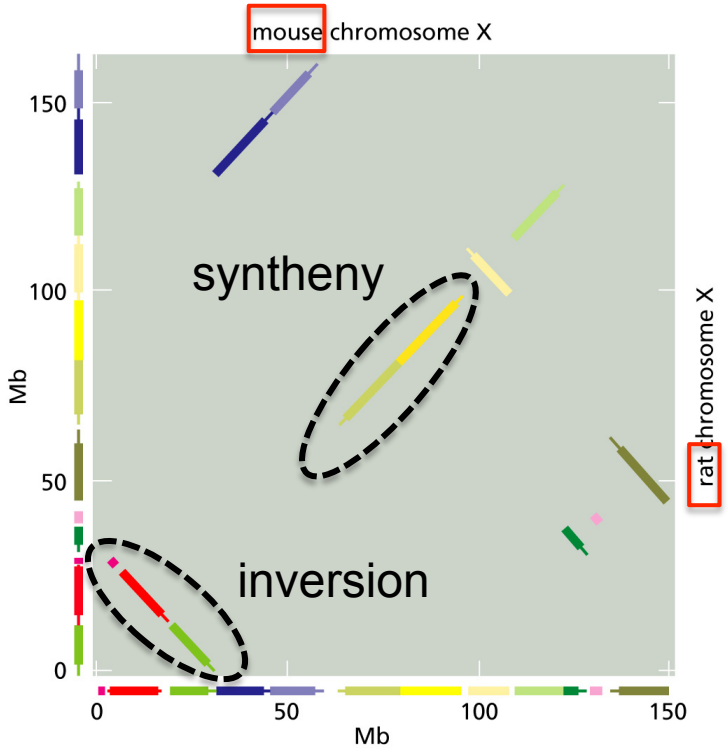
3. comparing with many protein sequences with the aim of finding a similar protein

Eukaryotic genes: similarity-based (extrinsic) methods

When the genome sequence of a closely related organism is available, a **genome-to-genome** sequence aligment can be very powerful in determining the status of uncertain genes prediction

The alignment of two genomes can be not trivial, as large scale rearrangements may have occurred also between closely related species

However, regions of **syntheny** where the genomes are sufficiently similar to make their common evolutionary ancestry apparent can be of help



Methods for gene search in eukaryotes

The main aim of the gene search is being able to distinguish between coding and non coding DNA and locate genes

1. Identifying ORFs (Open Reading Frames) – not feasible
2. *Signals* search (e.g. promoters and splicing sites)
3. Difference in the *nucleotide composition*
4. Homology search (sequence similarity)

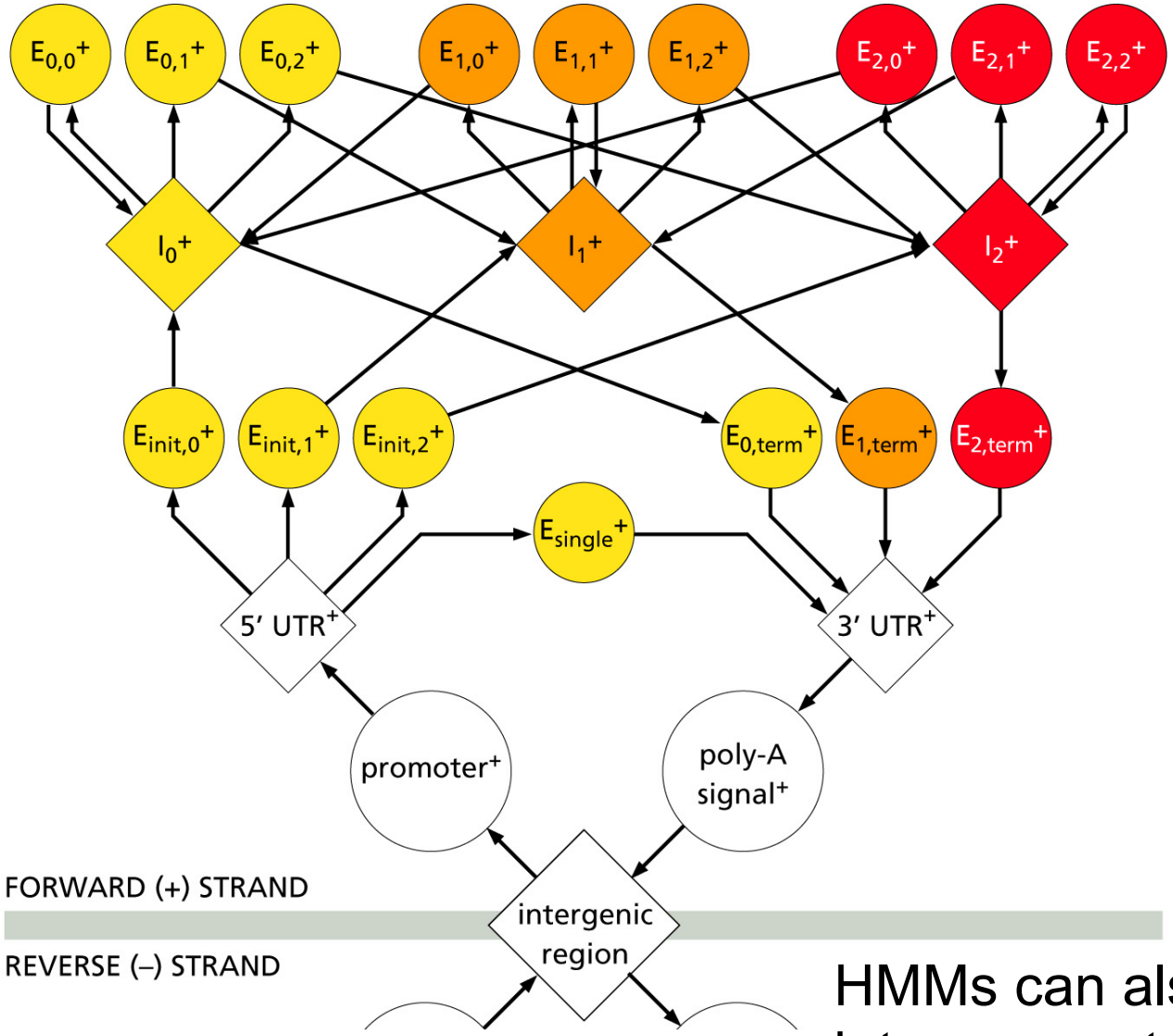
All the methods above are put together to predict a complete eukaryotic gene structure!

Methods for gene prediction in eukariotes

Many programs predict individual gene components and then combine them automatically into a gene prediction

Approaches used for doing that are mainly Hidden Markov Models (HMMs) and dynamic programming

Methods for gene prediction in eukaryotes



Simplified scheme of an HMM for eukaryotic gene prediction (GenScan).
$E_{0,1}$ is an exon in frame 0 (previous exon ended with a complete codon) which ends with the 1st nt of a codon and so on…

HMMs can also be modified to take into account results of homology searches (e.g. with BLAST)

Furthermore in eukaryotes…

In addition to the nuclear genome, specialized organelles in eukaryotic cells, such as **mitochondria** (generating energy by aerobic respiration in most animals, plants and fungi) and **chloroplasts** (hosting the process of photosynthesis in plants and algae) feature **their own genomes**

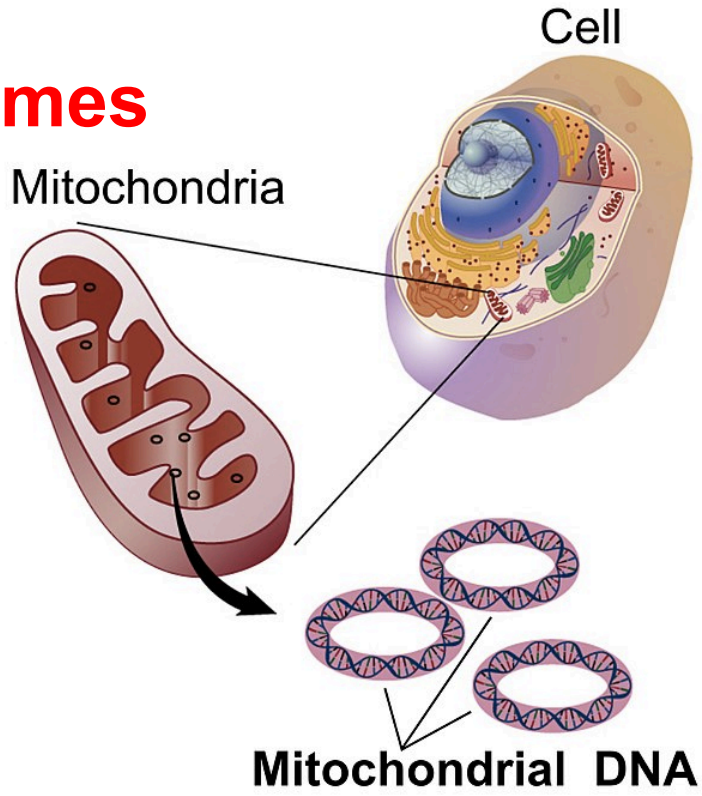
These genome encode only some of the proteins specific to **mitochondria** and **chloroplasts** (most of their proteins are encoded by genes in the cell nucleus)

Mitochondria and **chloroplasts** are believed to be the relics of prokaryotic organisms embedded by the ancestors of the eukaryotic cells (by Horizontal Gene Transfer)

Mitochondrial genomes

In animals, mitochondrial DNA (mtDNA) is a small circular chromosome (11 to 28 Kbp)

Mitochondria, and therefore the mtDNA, are passed exclusively from mother to offspring through the egg cell



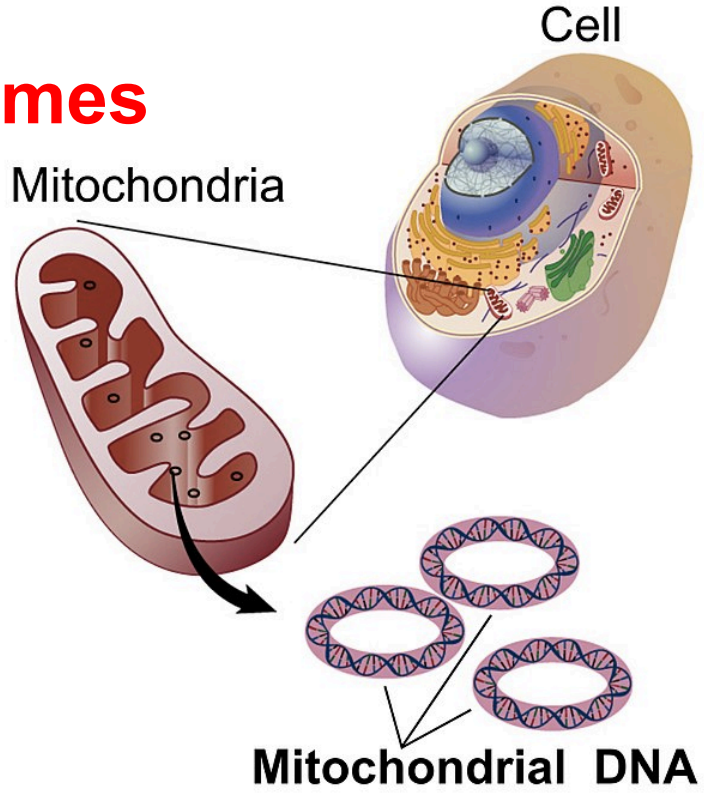
Human mtDNA was the first significant part of the human genome to be sequenced, it includes 16,569 base pairs and encodes altogether for two rRNAs, 22 tRNAs, and 13 proteins, all involved in the oxidative phosphorylation process

Several mutations in mtDNA (protein or RNA-coding) are associated to serious genetic diseases, including types of diabetes, cancer, neurodegenerative diseases etc.

Mitochondrial genomes

In animals, mitochondrial DNA (mtDNA) is a small circular chromosome (11 to 28 Kbp)

Mitochondria, and therefore the mtDNA, are passed exclusively from mother to offspring through the egg cell



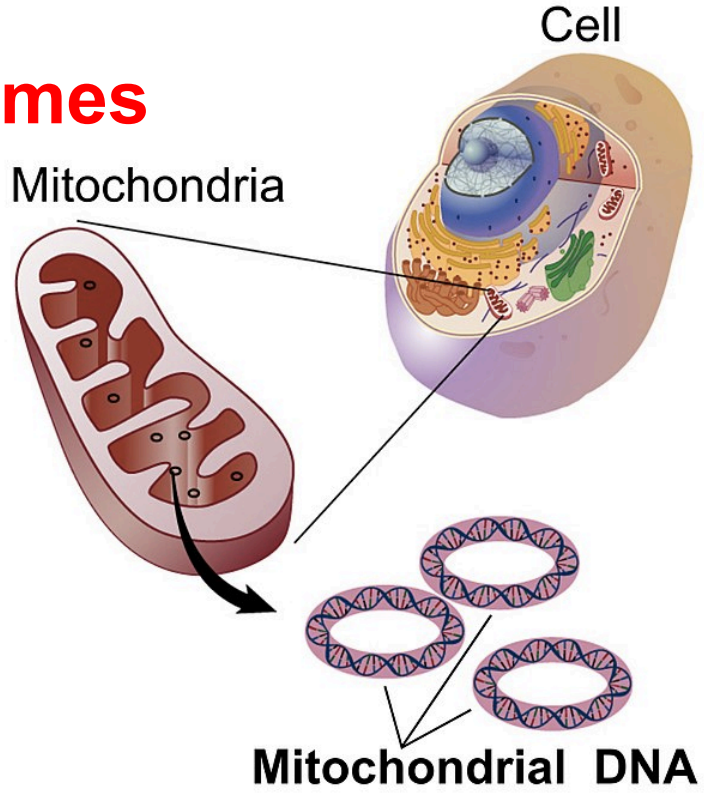
mtDNA evolves faster than nuclear DNA (by a factor of ≈10, in terms of substitutions per bp per million years), therefore it represents a pillar of phylogenetics and evolutionary biology

It permits an examination of the relatedness of populations, and so has become important in the study of the distribution of species and ecosystems in geographic space and through geological time (biogeography)

Mitochondrial genomes

In most mitochondria, protein-coding genes do not contain introns

Many mitochondria – not plant mitochondria however – translate the **TGA codon** as "tryptophan (W)", not "stop"



FIRST BASE	SECOND BASE U	SECOND BASE C	SECOND BASE A	SECOND BASE G	THIRD BASE
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Trp in many mitochondria

How we evaluate the correctness of a prediction?

		Predicetd as a splice site?	
		YES	*NO*
Is it a splice site?	*YES* **POSITIVE (P)**	TP	FN
	NO **NEGATIVE (N)**	FP	TN

Let's define true and false positives/negatives:

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

How we evaluate the correctness of a prediction?

		Predicetd as a splice site?	
		YES	*NO*
Is it a splice site?	*YES* POSITIVE (P)	TP	FN
	NO NEGATIVE (N)	FP	TN

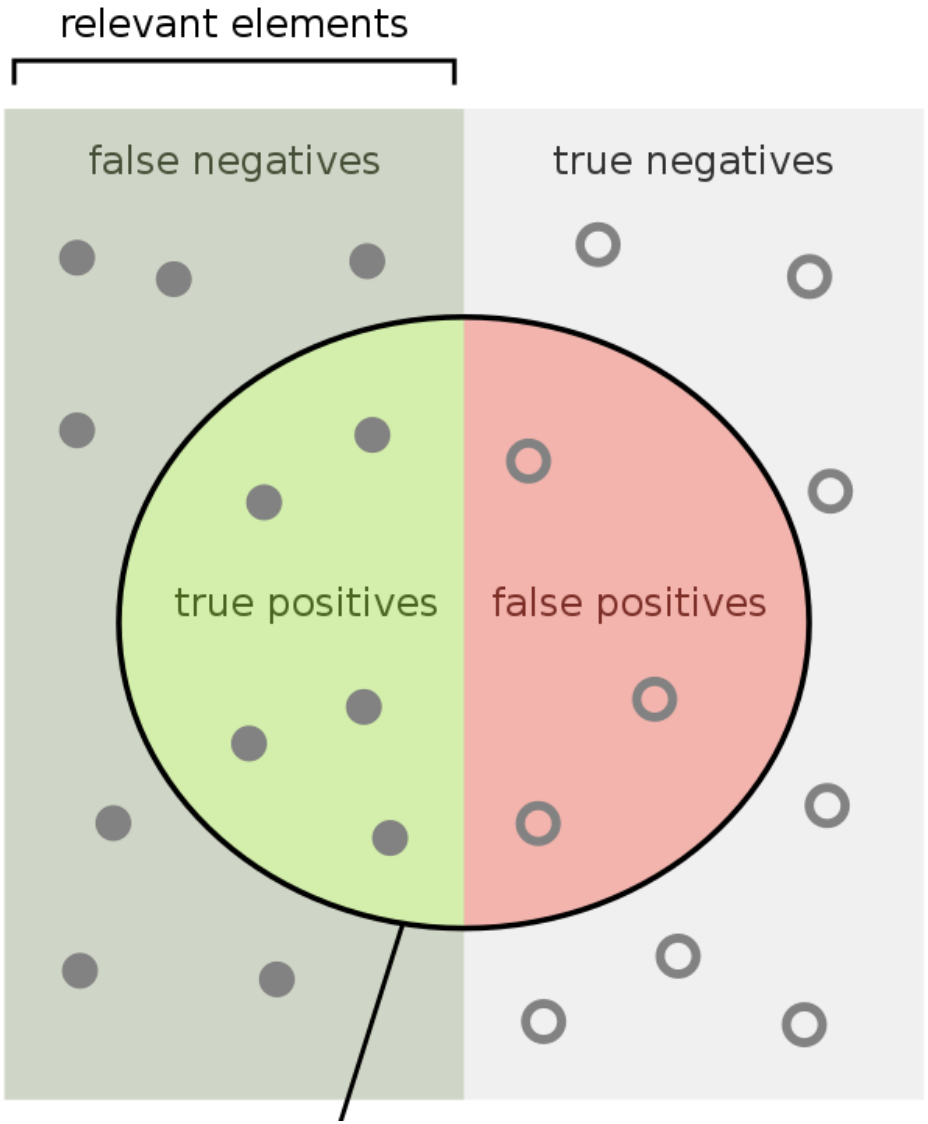
sensitivity, recall, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

How we evaluate the correctness of a prediction?



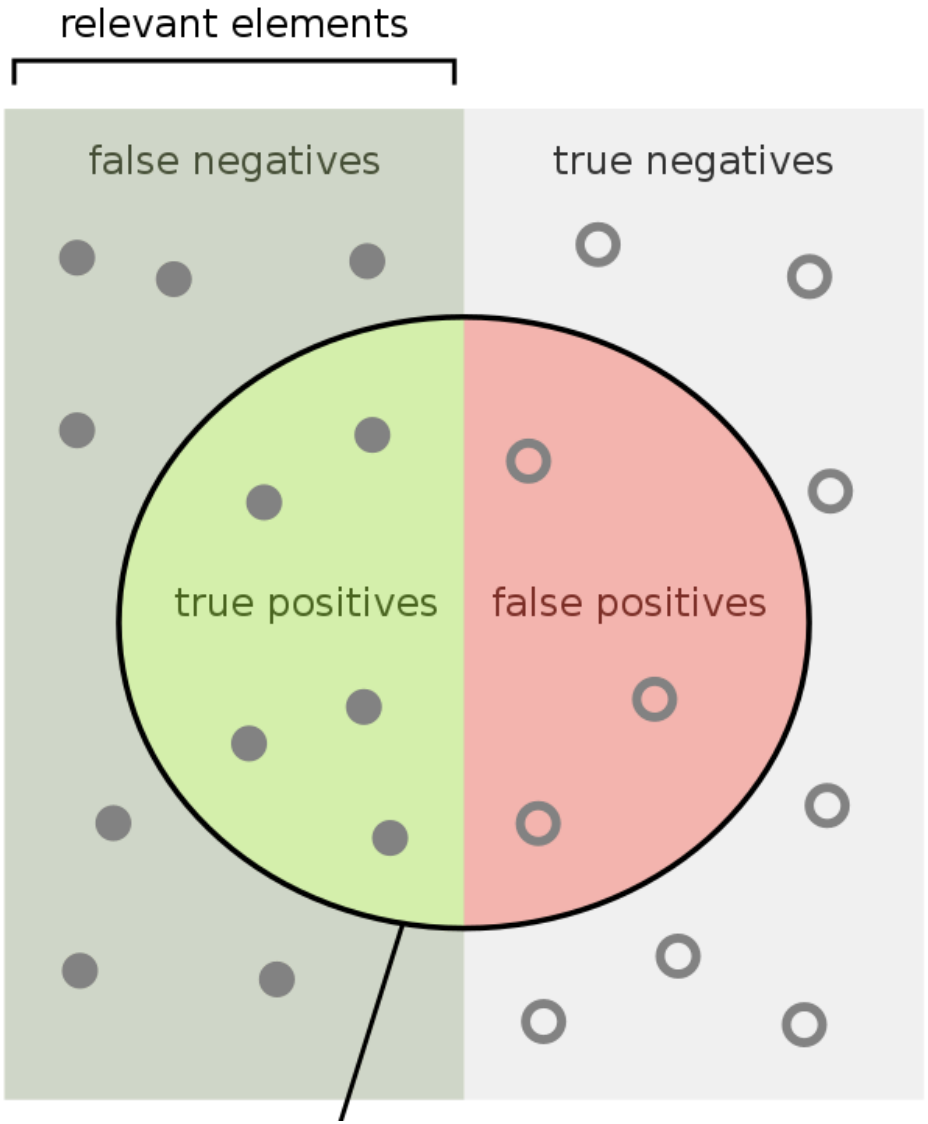
sensitivity, recall, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$\text{Sensitivity} = \frac{\text{green}}{\text{green} + \text{dark green}}$$

Sensitivity is a measure of how good a model is at PICKING UP relevant elements (e.g. splice sites)

How we evaluate the correctness of a prediction?



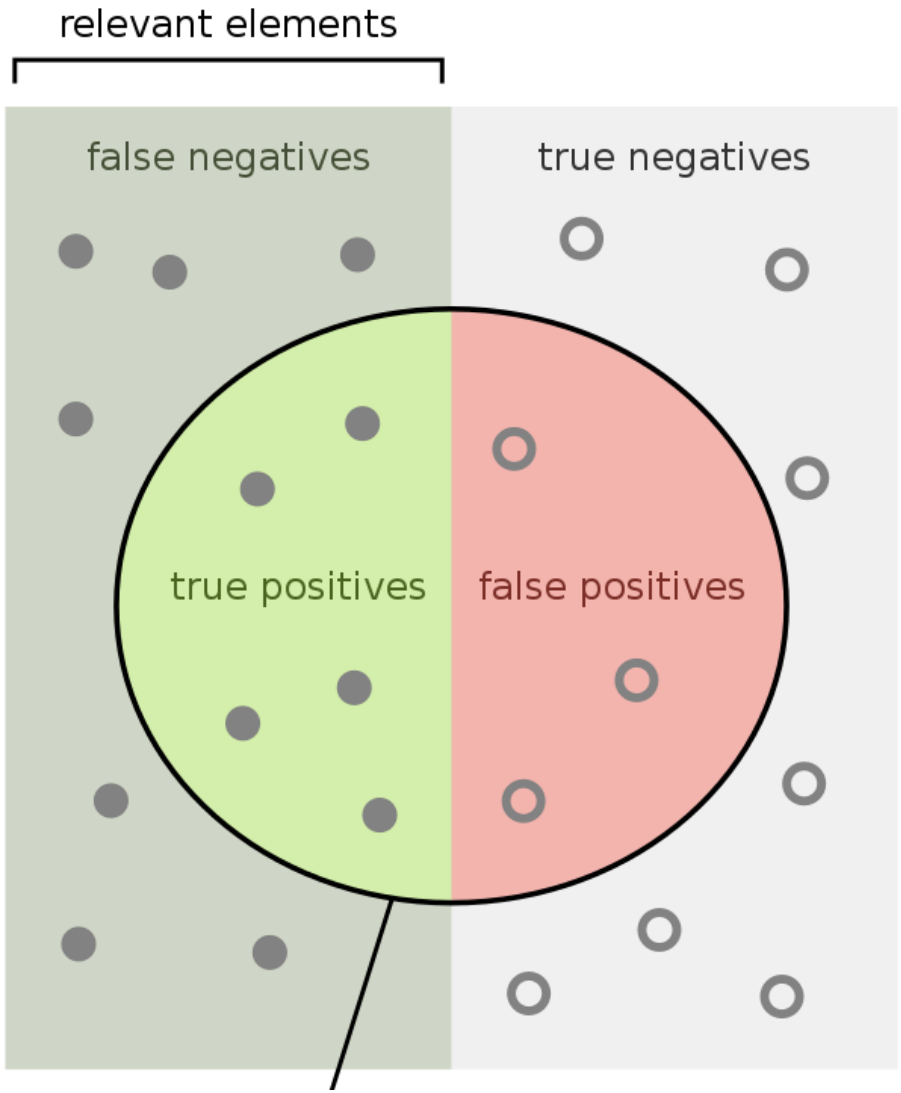
sensitivity, recall, or true positive rate (TPR)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

Example of high sensitivity: few splice sites (green) are not identified

Sensitivity is a measure of how good a model is at PICKING UP relevant elements (e.g. splice sites)

How we evaluate the correctness of a prediction?



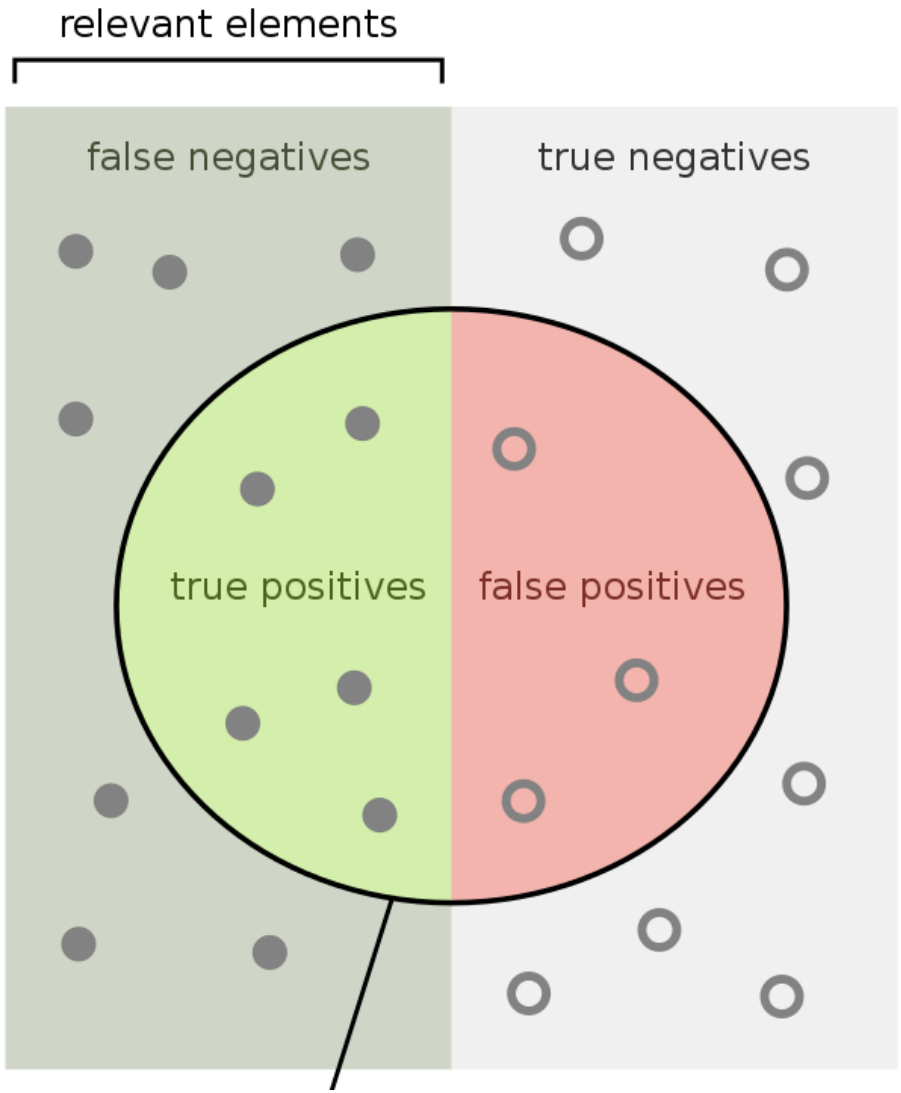
specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

$$\text{Specificity} = \frac{\quad}{\quad}$$

Specificity is a measure of how good a model is at DISCERNING between relevant & non relevant elem. (e.g splice/non splice sites)

How we evaluate the correctness of a prediction?



specificity, selectivity or true negative rate (TNR)

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

Example of high specificity: few non splice sites (red) are predicted as splice sites

Specificity is a measure of how good a model is at DISCERNING between relevant & non relevant elem. (e.g splice/non splice sites)

Matthews correlation coefficient (MCC)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{1/2}}$$

The Matthew's correlation coefficient (MCC) is a single parameter accounting for the overall prediction success (including sensitivity and specificity)

When the FP and FN are zero (i.e. zero incorrect predictions), then MCC equals 1 (max possible value)

When the TP and TN are zero (i.e. all incorrect predictions), then MCC equals 0 (min possible value)

Matthews correlation coefficient (MCC)

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{1/2}}$$

		S_n (%)	S_p (%)	MCC
Testing set I				
5′SS				
	Our method	67.15	71.71	0.389
	ASSP	56.63	62.66	0.193
3′SS				
	Our method	64.22	71.88	0.362
	ASSP	58.16	65.10	0.233

Example of performance for the prediction of alternative splice sites in human (NAR 2006)

Classification model	MCC	Sensitivity	Specificity
TSSPlant	0.310	0.976	0.231
TransPrise	0.791	0.872	0.919

Example of methods performance for the prediction of promoter signals in chromosome 2 of Oryza sativa
(PeerJ 2019)

Lessons 8-9. Content

1. An overview of genomes and genes organization. The genome and gene structure in prokaryotes is way less complex than in eukaryotes

2. Gene finding in prokaryote genomes. Methods can be either intrinsic or extrinsic (homology-based). The intrinsic methods, based on ORF and signals search and on differences in the nt composition of coding & non-coding regions, have a relatively low error rate (are quite efficient) for prokaryotes

3. Gene finding in eukaryote genomes. Way more complex. ORF search is not efficient, while the other above methods can be used and combined in a method to predict a complete eukaryotic gene structure, mainly by HMMs. Extrinsic methods can be of great help.