Gene finding

Lecture 12.

Motivation

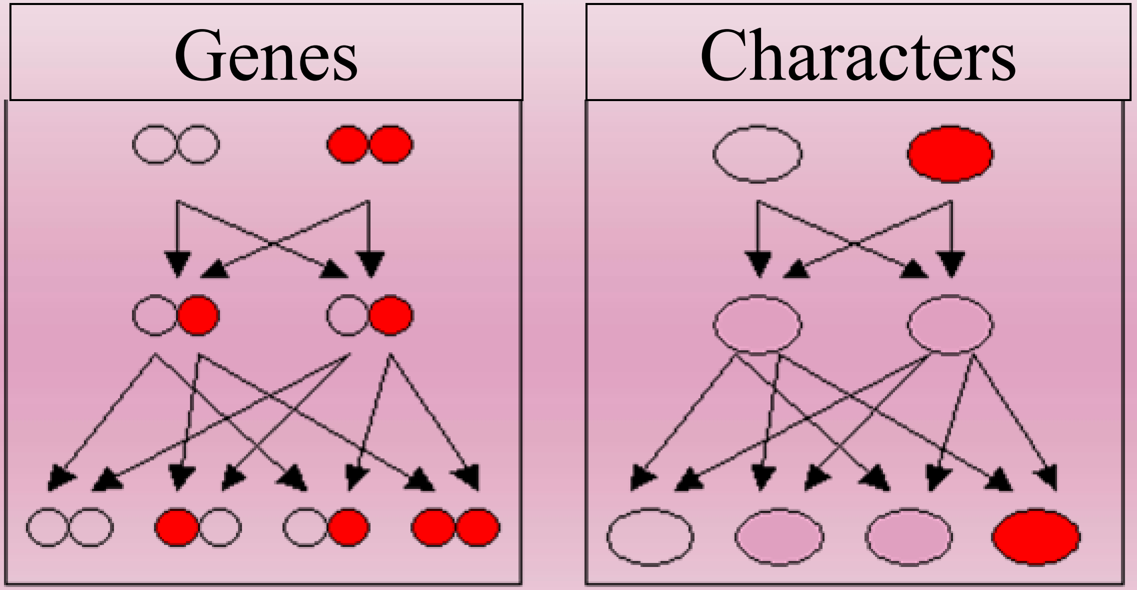
- Not all sequences of DNA are *coding,* namely are used as a template for protein. In the human genome only 2-3% sequences are coding.
- You cannot read DNA sequence and tell which part codes for a protein, therefore we need to do it automatically

Background

- Gene expression – the process by which a protein is generated according to the information in a DNA sequence.
- It involves 2 steps:
  - Transcription: produces an mRNA using DNA sequence as a template
  - Translation: synthesizes the protein according to information coded in the mRNA

Recap:From gene to protein

- *Phenotype* – an outward expression of **discrete** genetic characteristics. Proteins are responsible for phenotype



How information from the sequence of nucleotides is converted into a sequence of aminoacids ?

Protein Synthesis: Transcription

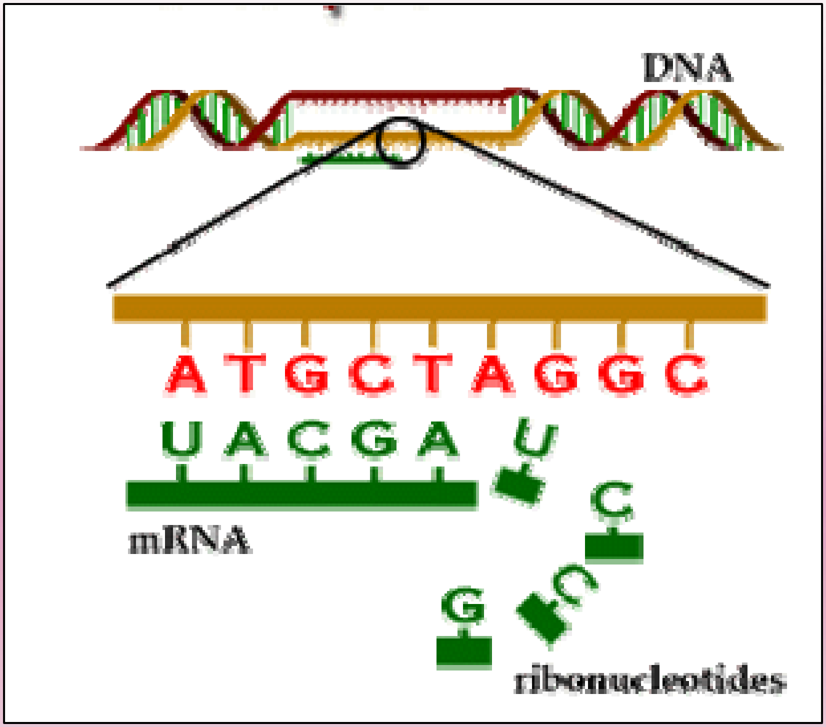
- RNA – **R**ibo**N**ucleic **A**cid -

  a short and unstable polymer of the same nucleotides as DNA:

  Adenine, **C**ytosine, **G**uanine, **U**racil (instead of **T**hymine)

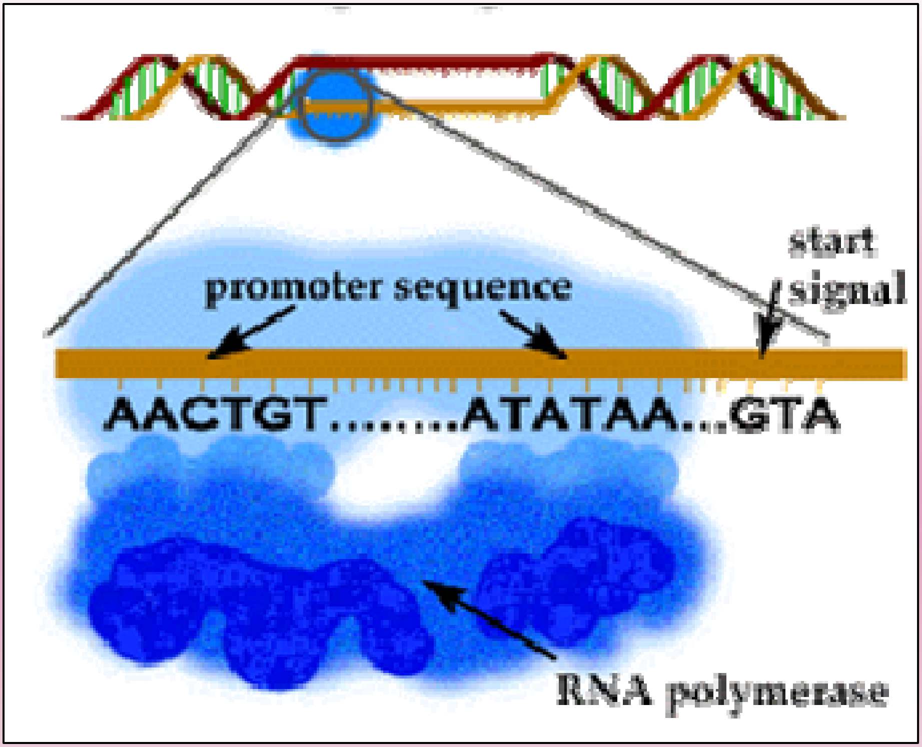
Messenger-RNA, **m-RNA**

- Copy of the template strand of DNA is made in the cell nucleus
- The copy moves into cytoplasm



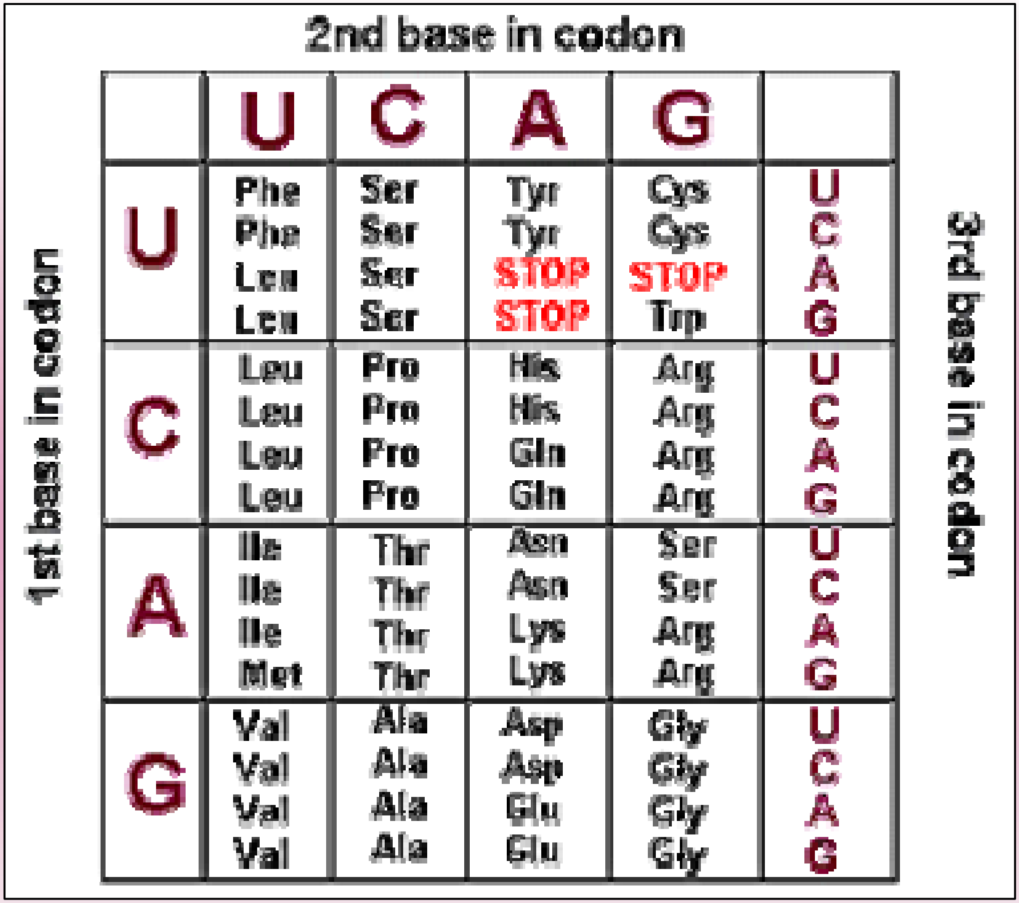
Protein Synthesis: Transcription

- Initiation with binding of the RNA polymerase to the ***promoter*** site (comparatively conserved sequences).
- The synthesis starts at ***start codon*** GTA (which then become bases CAU on the RNA molecule).



Genetic Code

- There are 4^3=64 possible triplets – *codons*, but only 20 aminoacids and 3 stop codons.
- The code is degenerative: different triplets code for the same aminoacid
- Important in keeping the proteins functional

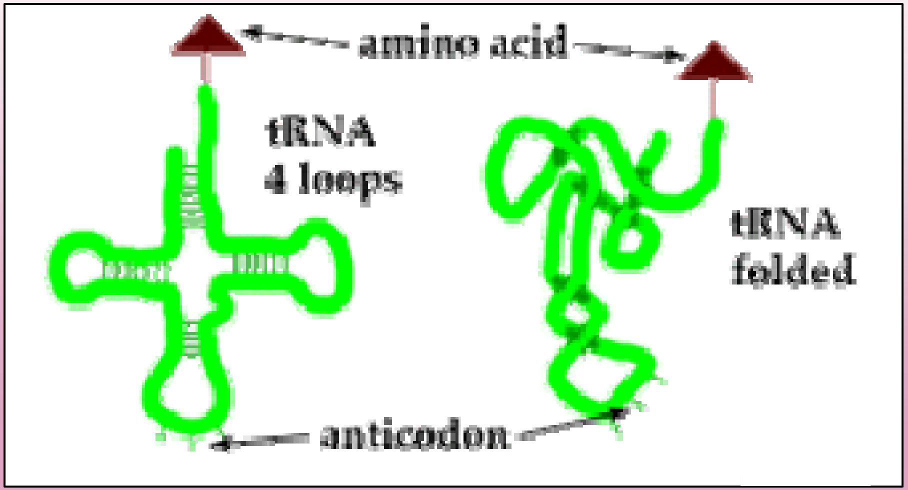


2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
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

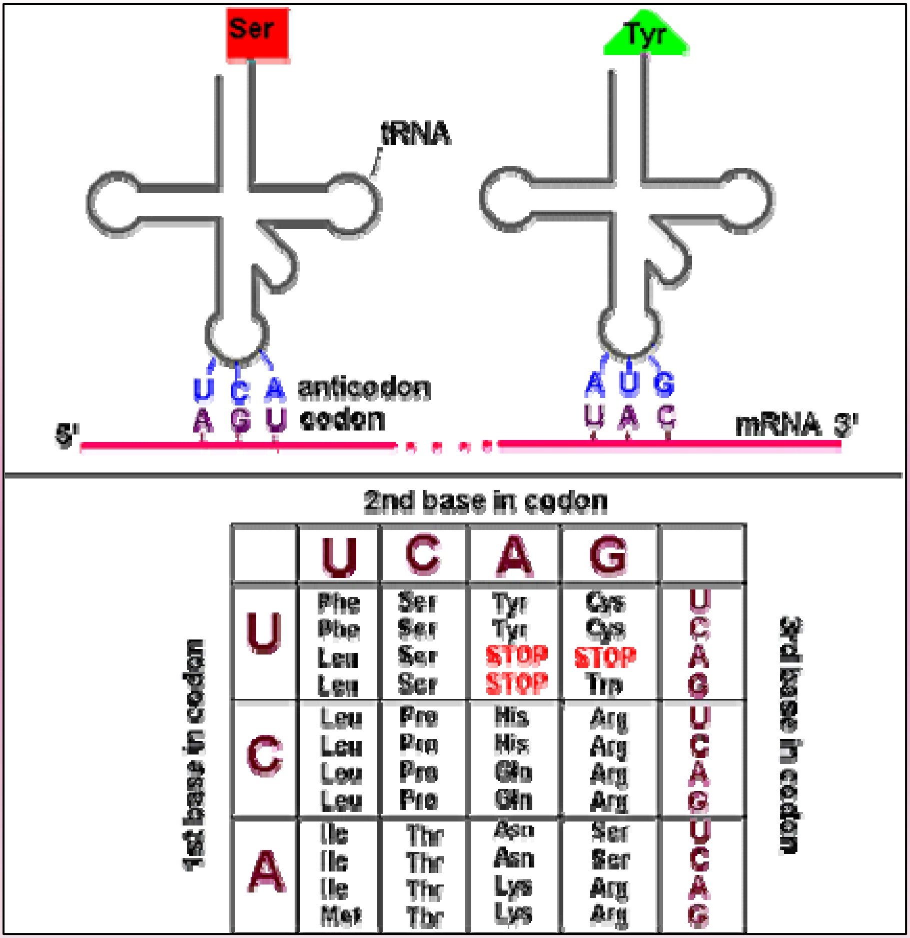
Protein Synthesis: Transport RNA

- **t-RNAs** are short
- Fold into a cloverleaf secondary structure
- Hydrogen bonds hold into an L-shaped tertiary structure



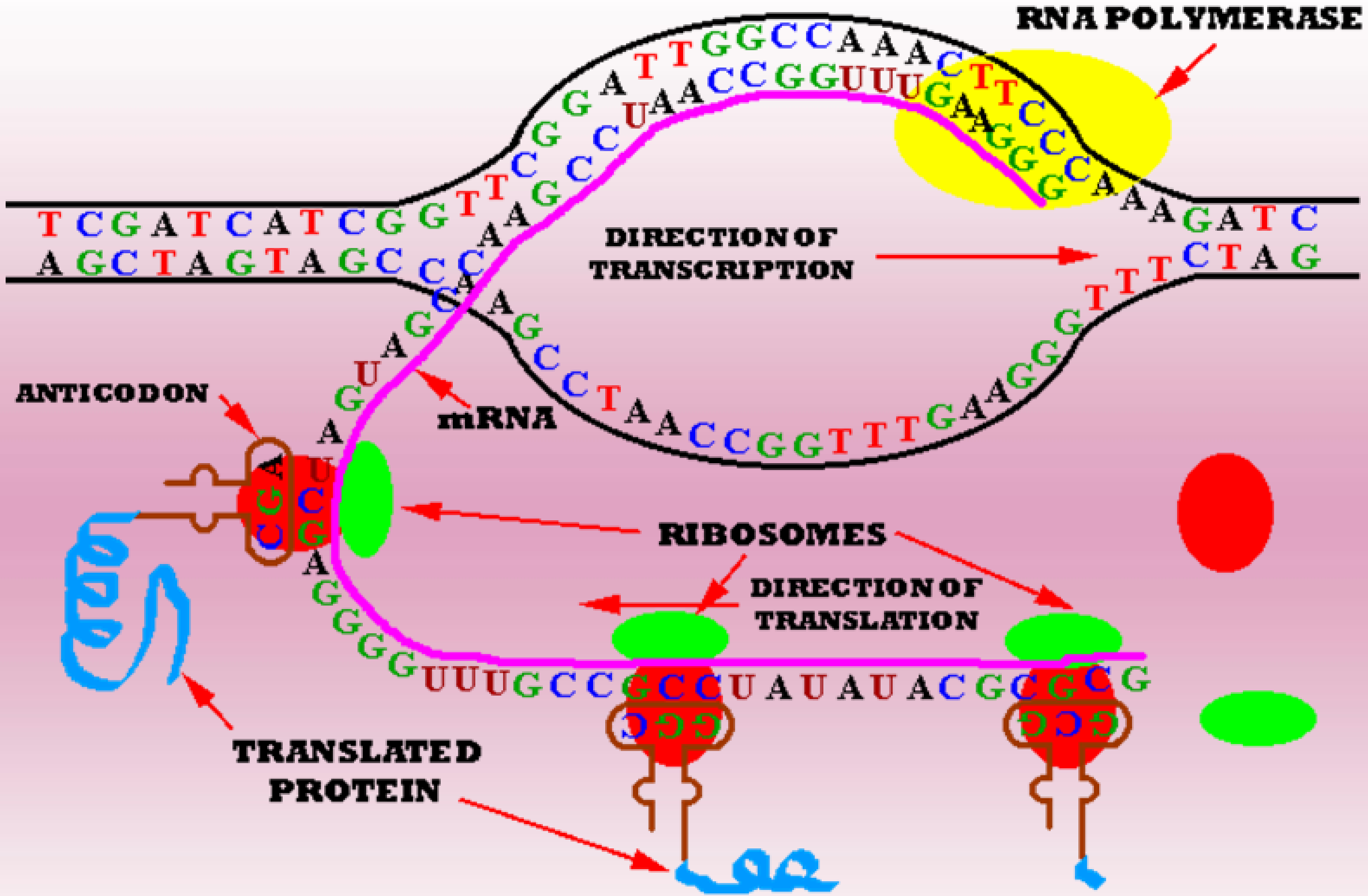
Protein Synthesis: Transport RNA

The anticodon is complementary to the triplet encoding the attached aminoacid, according to the genetic code



1st base in codon	2nd base in codon				3rd base in codon
	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

Protein Synthesis



The central dogma of molecular biology

DNA contains the complete genetic information that defines the structure and function of an organism. Proteins are formed using the genetic code of the DNA. Three different processes are responsible for the inheritance of genetic information and for its conversion from one form to another :

1. Replication

2. Transcription

3. Translation

The central dogma of molecular biology

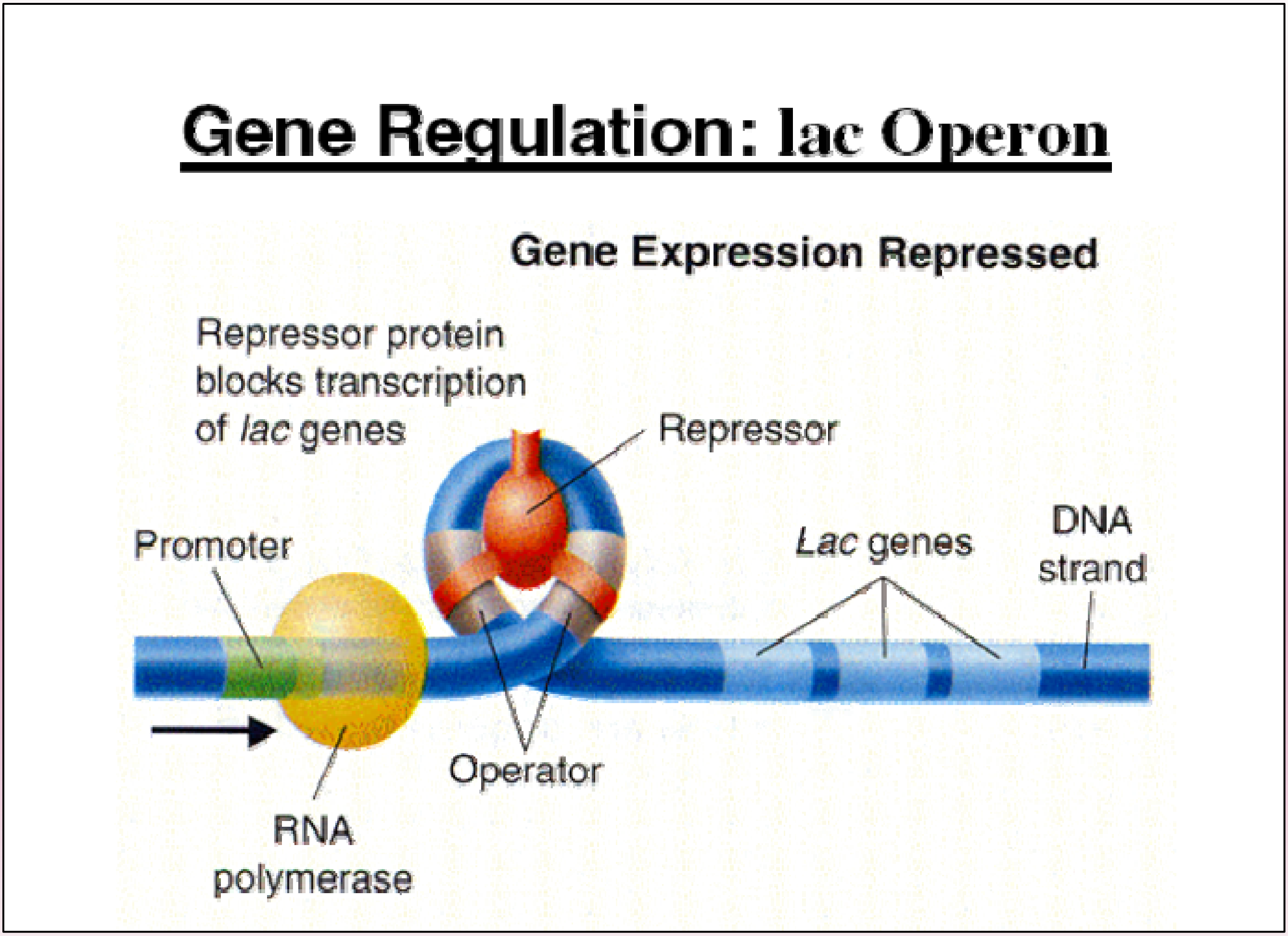
General	Special	Unknown
DNA → DNA	RNA → DNA	protein → DNA
DNA → RNA	RNA → RNA	protein → RNA
RNA → protein	DNA → protein	protein → protein

The direction of the information flow:

DNA->RNA->Protein,

never Protein->DNA

Regulation of gene expression

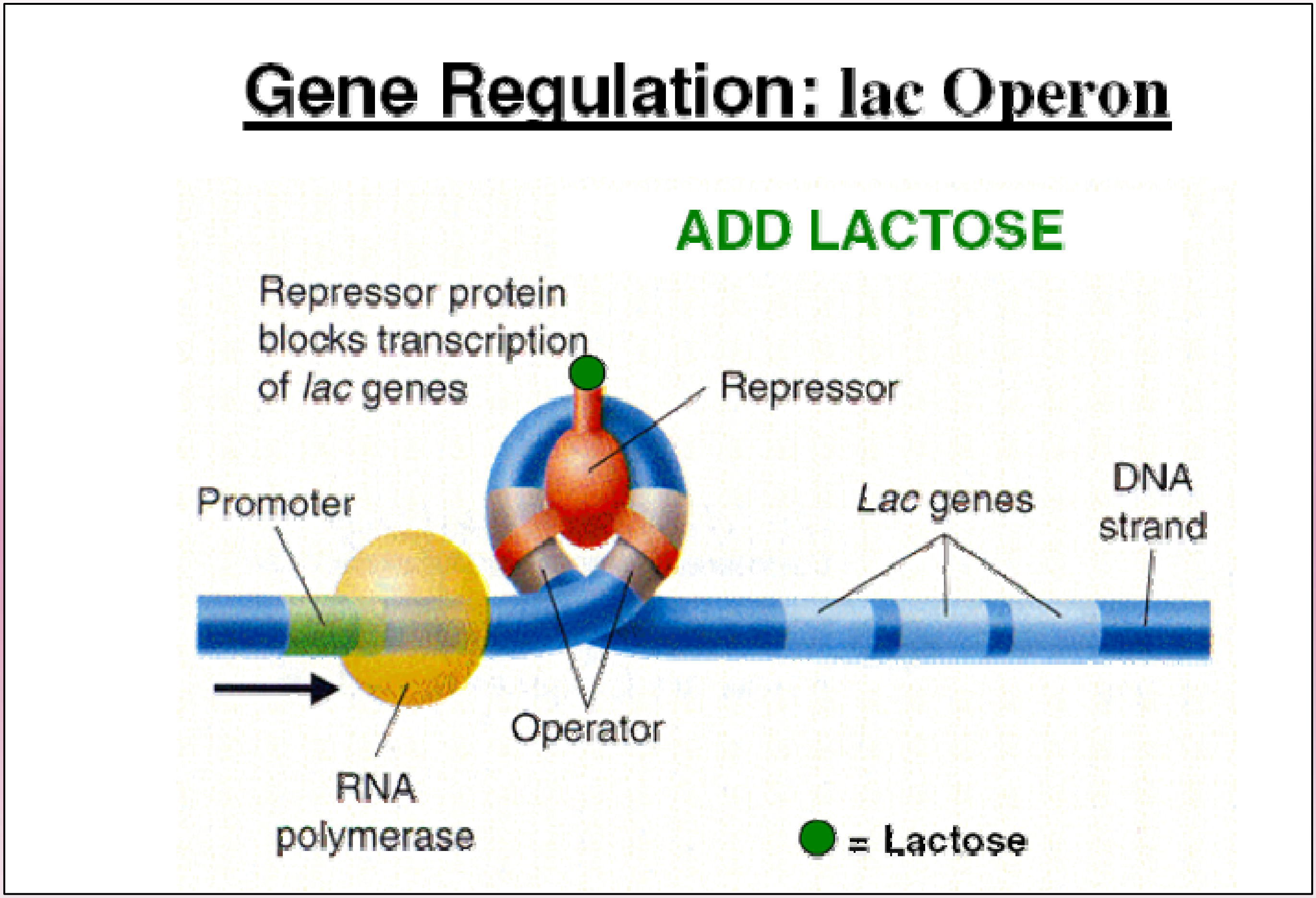


Gene Regulation: lac Operon

Gene Expression Repressed

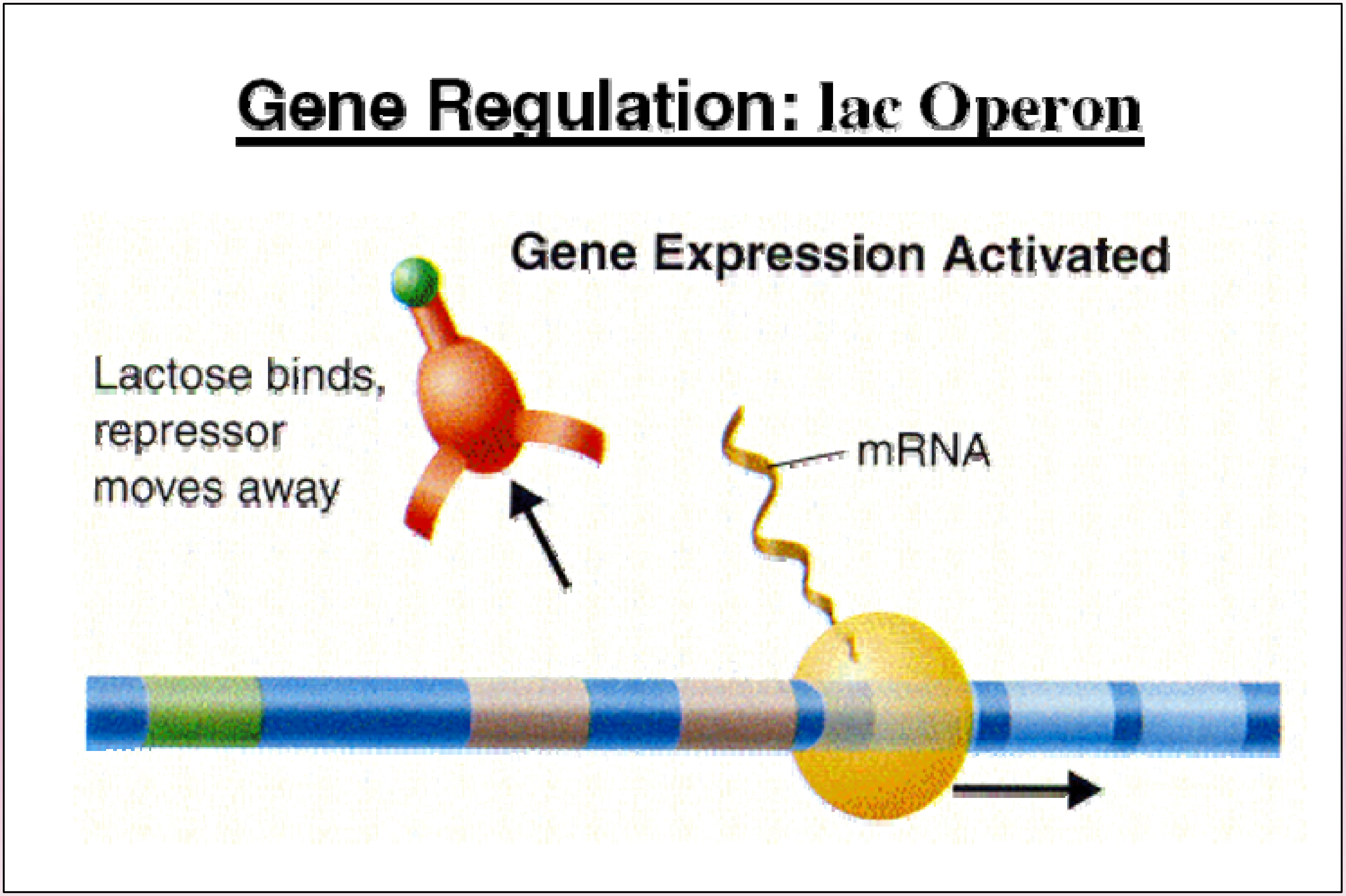
Repressor protein blocks transcription of *lac* genes

Regulation of gene expression



Gene Regulation: lac Operon

Regulation of gene expression

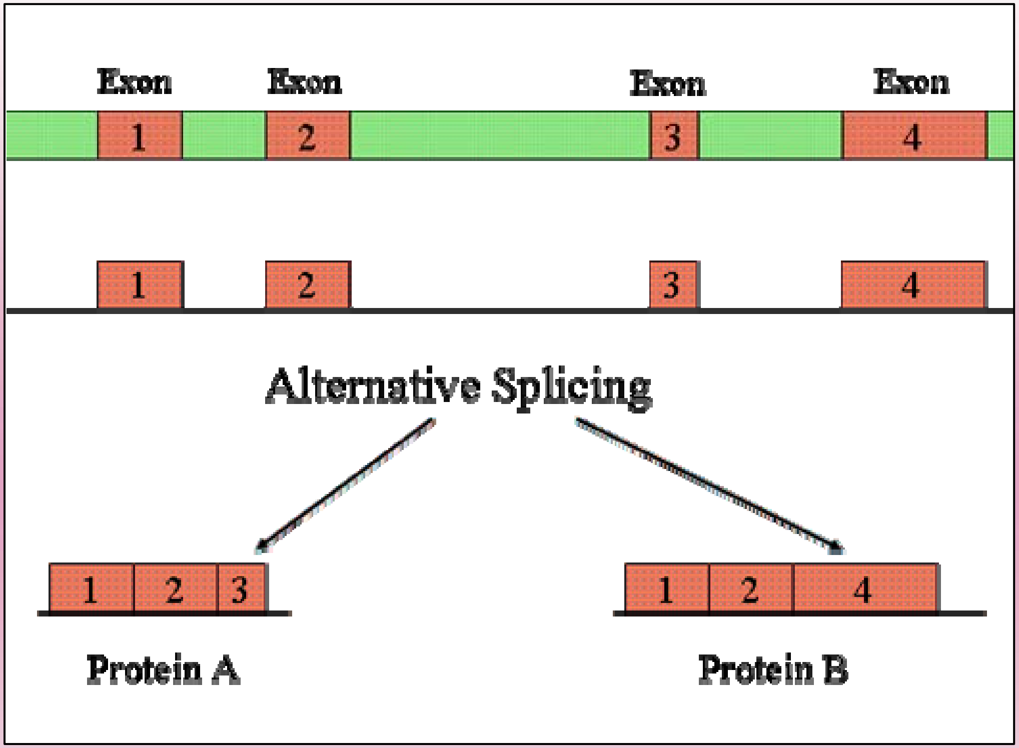


Gene Regulation: lac Operon

Gene Expression Activated
Lactose binds, repressor moves away
mRNA

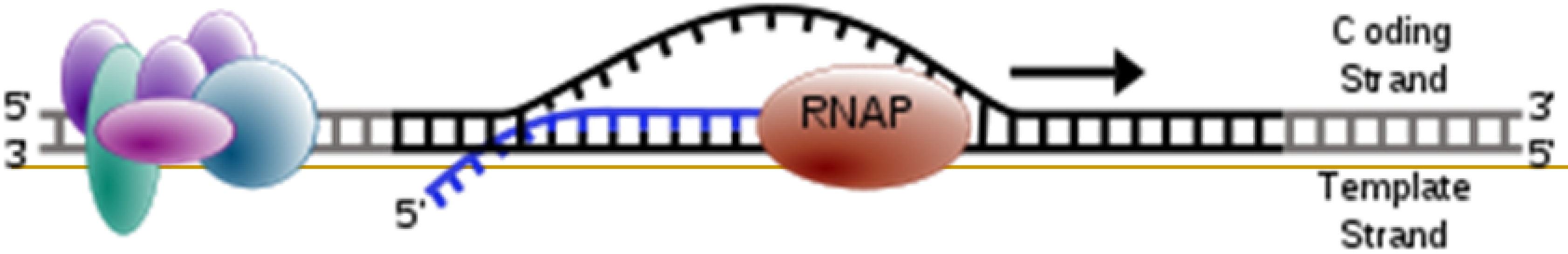
Gene to Protein: Complications

- Collinearity between the linear order of nucleotides and the linear order of aminoacids – did not persist after the 70-s:
  - Overlapping genes – different proteins from the same overlapping sequence of DNA
  - Interleaving *exons* (coding) and *introns* (non-coding) regions
  - Alternative splicing of exons



Transcription

- Transcription starts when an enzyme – RNA polymerase – recognizes and binds to a DNA molecule at a specific site – called *promoter*.
- The direction of the transcription is from the 5’ to the 3’ end of a single strand of DNA. This direction is called *downstream*.
- The promoter is located upstream from the desired start of the transcription.



Transcription signals

- Transcription start site – CAP signal – always A or G
- TATA-box (sequence TATAAA) – 50 nucleotides upstream from the RNA-polymerase binding site. Serves for binding of the TATA-protein, which helps to create RNA-polymerase – DNA complex.
- Termination signal: polyadenylation signal: AATAA

Only 70% of human promoters contain these core signal sequences.

It is hard to determine the start and the end of a gene on DNA

Translation

- The mRNA is translated into a sequence of aminoacids, which folds into a protein.
- In eukaryotes, the primary transcript undergoes splicing: introns are cut off and are discarded, and the exons are spliced together into a final coding mRNA sequence, which serves as a template for the protein sequence. Introns can be 70-30,000 bp long.
- Example: final m-RNA: this is a gene
- Primary transcript:

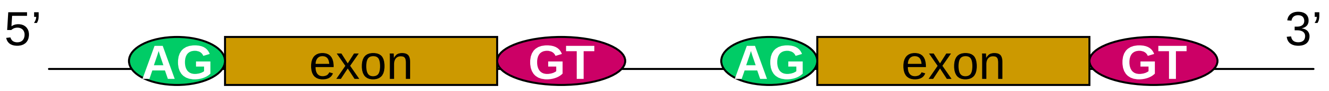
and this was is a long beautiful gene again

introns: in black, exons: in red

Translation signals

- Kozak signal: gccrccATGc – initiation of translation in vertebrates - where ATG is a start codon and r is a purine (A or G)
- Termination codon: TGA, TAA, TAG

Splicing signals

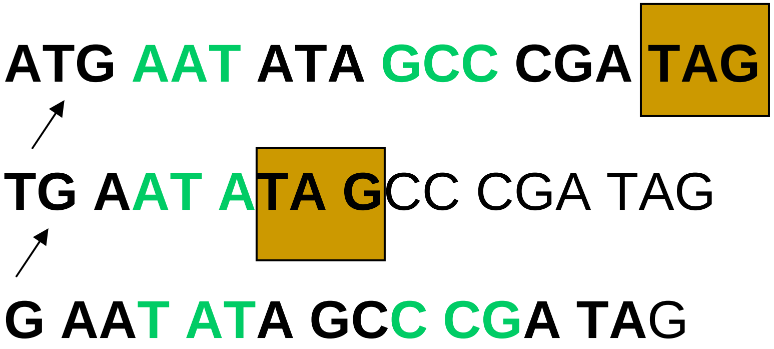


- Splicing is performed by complexes called spliceosomes: they consist of protein and snRNA. The snRNA recognizes the splice sites by complementary binding, the recognition must be precise.
- 5'-end of an intron – donor site – GT
- 3'-end of an intron – acceptor site - AG

Complications

- Promoters are not uniform. The possible reason – to control the level of gene expression for various genes. It was shown that there is a 80% correlation between the conservation of a promoter region and the binding energy of RNA-polymerase.
- Thus, finding signals is inherently stochastic – probabilistic - problem
- The average human gene is 30,000 bp long, but the dystrophin gene is 2 million bp long
- The mammalian genome contains on average 225 bp per each 1 kbp of sequences which can encode protein, but are never transcribed – pseudogenes
- One sequence can encode several different proteins – alternative splicing, open reading frames

Open reading frame



ATG AAT ATA GCC CGA TAG

TG AAT ATA GCC CGA TAG

G AAT ATA GCC CGA TAG

Stop codons:
TAA, TAG, TGA

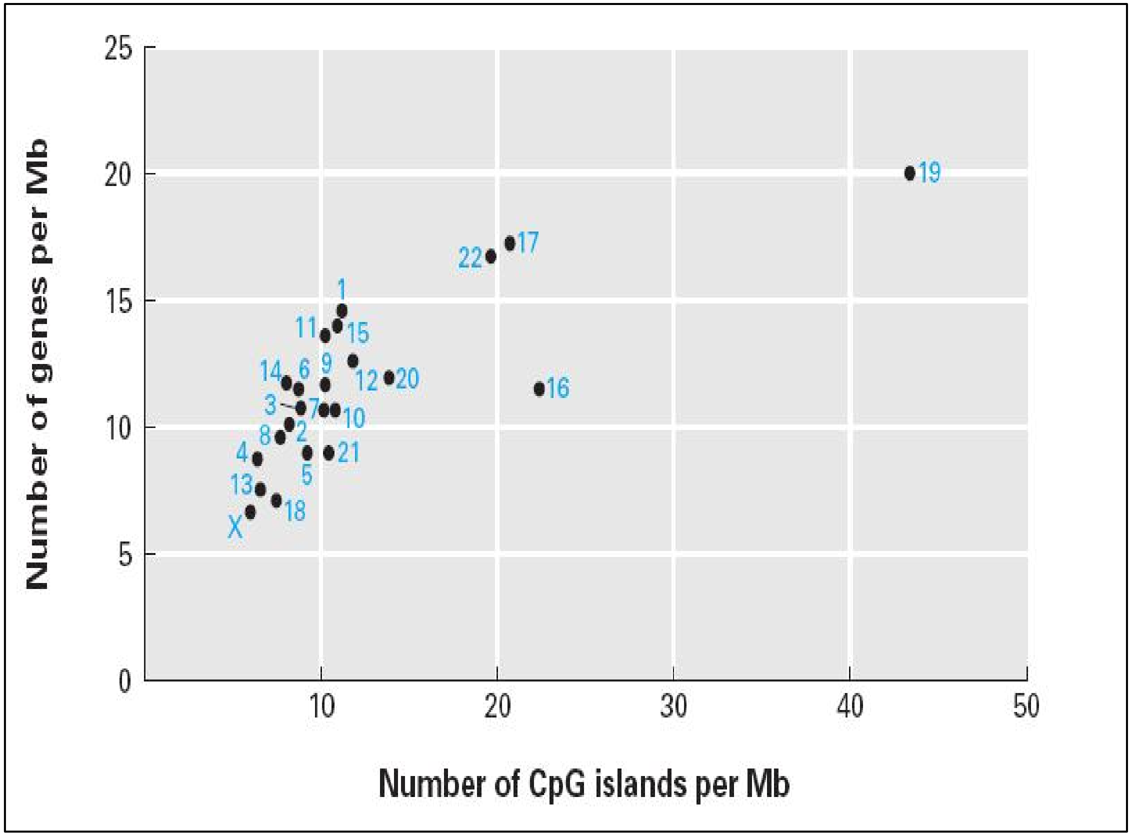
Start codon ATG

- 3 possible reading frames for each coding mRNA
- 3 out of 64 combinations are stop codons – each 64/3=21 codon – stop, the genes are in general larger than this (~300 aminoacids)
- The search for long reading frames fails to detect short exons

Main approaches r

- Ab initio
- Similarity-based
- Mixed

Ab initio (from a sequence alone) gene prediction



- Find long open reading frames
- Remove from candidates the sequences not flanked by the splicing signals
- Check distribution of di-nucleotides (CpG islands) or tri-nucleotides (codon usage around splice sites)

Similarity-based gene prediction

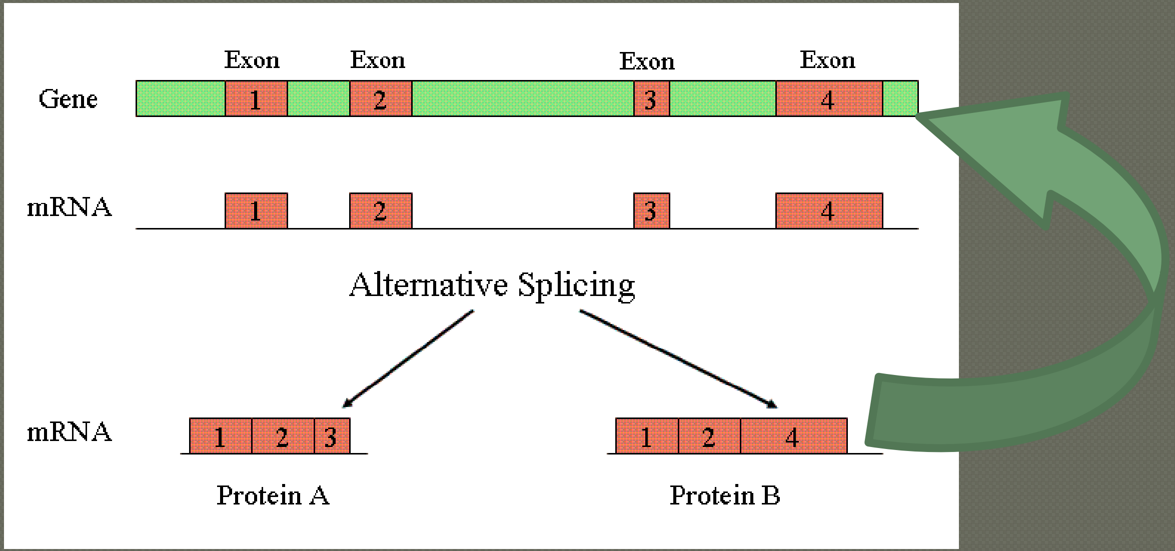
- Suppose we know the coding sequence for the mouse protein
- Find similar sequence in human genome – this is a similarity-based gene prediction.
- The problem is that the exons are of different sizes and of different order in different species, even though they may encode for a very similar protein

Reverse gene prediction

- Suppose we know the sequence of mRNA which encodes a protein. We convert it to cDNA and we determine its sequence.
- The problem now is to locate the sequence in the genome: cDNA consists of an unknown number of exons of an unknown length.

Choosing the scoring scheme

- C-DNA example
  - If we extract m-RNA from the working cell at any stage of its working cycle, and then we transcribe it backwards into DNA by using viral reverse transcriptase, we obtain c-DNA
  - Now we want to align the obtained c-DNA with the region of genome in order to locate the gene



Choosing the scoring scheme for c-DNA alignment

- If the spaces are penalized on the unit bases, that would align c-DNA substrings close together rather than allowing large gaps corresponding to introns

- The number of mismatches in the aligned regions should not be large, since these regions are just a transcript of the original genomic DNA

Scores for c-DNA alignment

- The following scoring scheme solves the problem:
  - Constant gap weight
  - Heavy penalty for mismatches

- The optimal alignment can be induced to cut up c-DNA to match the exons on the DNA sequence they have originated from

Another solution: Exon chaining problem

- Given a set of putative exons, find a maximal set of non-overlapping putative exons

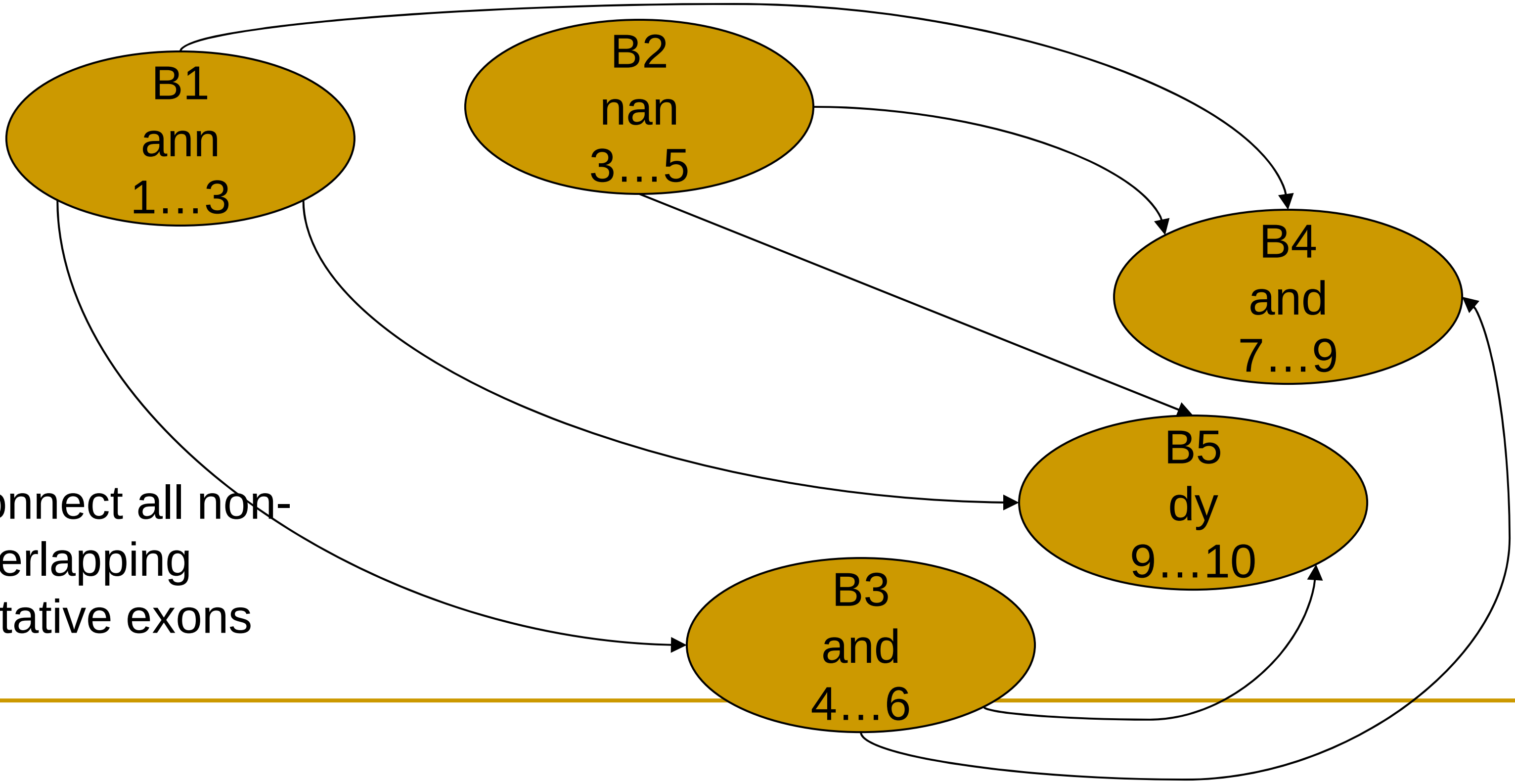
Locating exons

G	1	2	3	4	5	6	7	8	9	10
	a	n	n	a	n	d	a	n	d	y

cDNA	1	2	3	4	5
	n	a	n	n	y

Step 1. Determine putative exons – by probabilistic methods plus signals

G	1	2	3	4	5	6	7	8	9	10
	a	n	n	a	n	d	a	n	d	y



Connect all non-overlapping putative exons

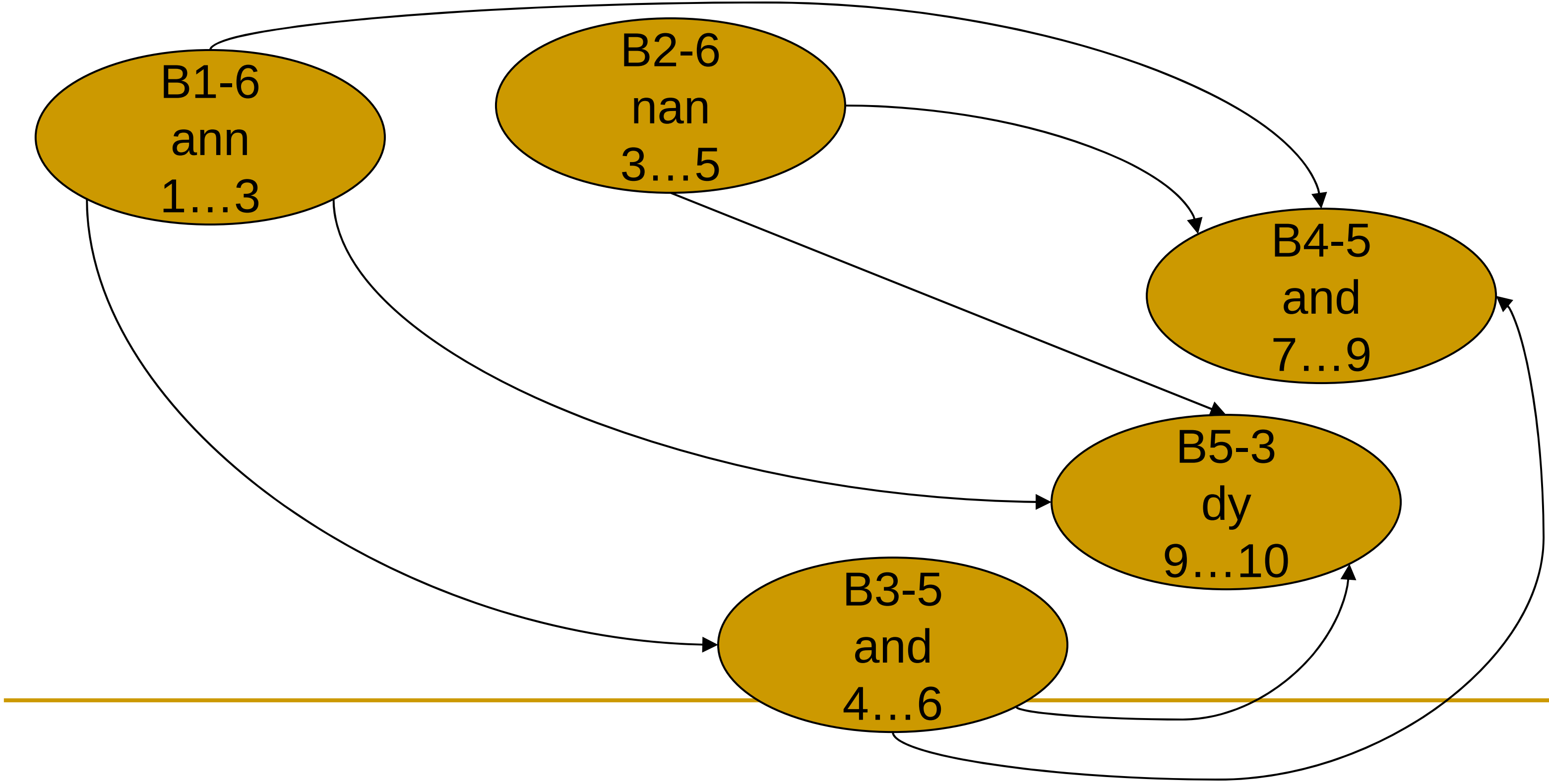
Step 2. Compute local alignment score for each node

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

1 – for mismatch

-1 – for indel



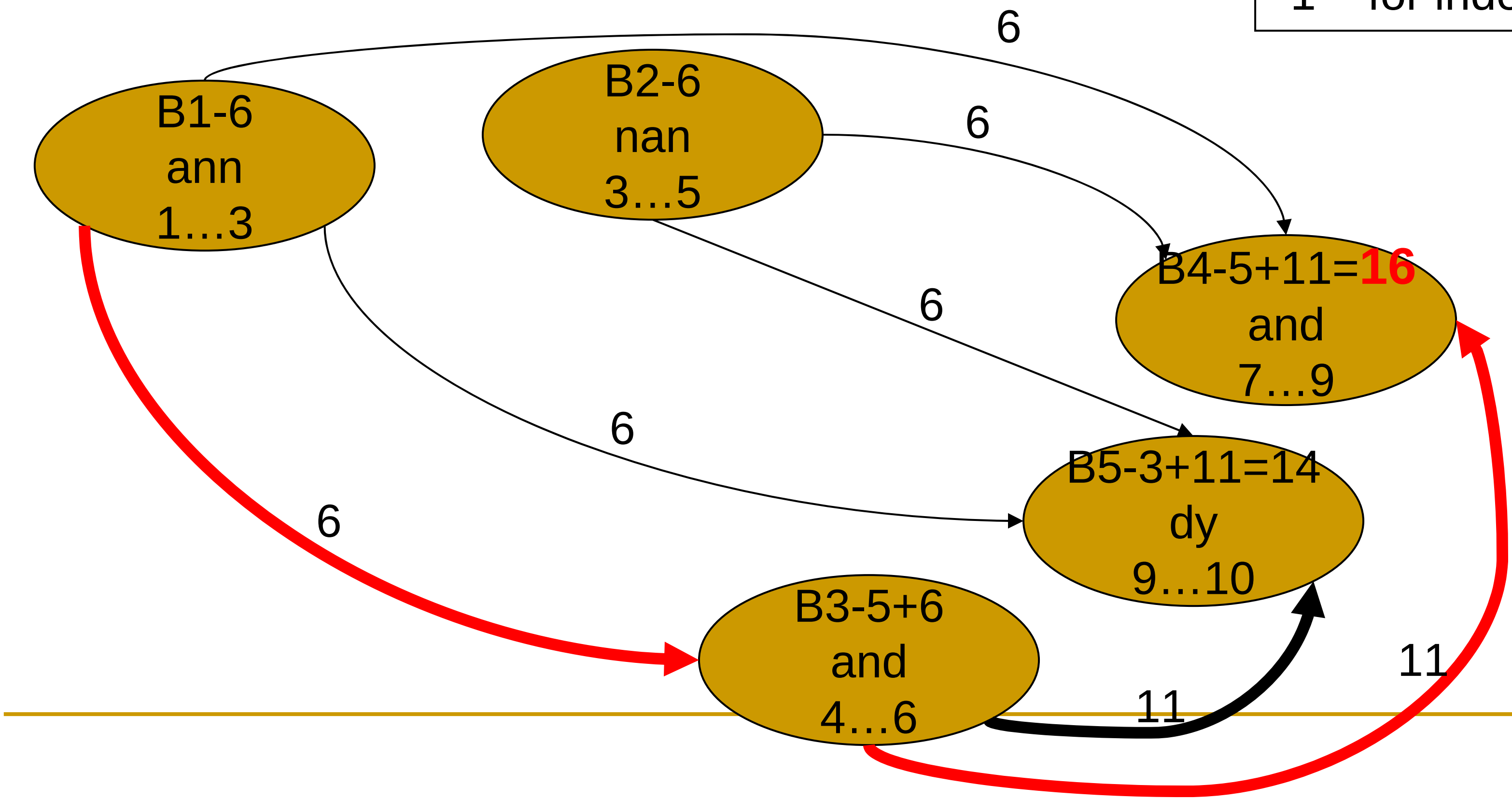
Step 3. Find the path with max weight by dynamic programming

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

1 – for mismatch

-1 – for indel



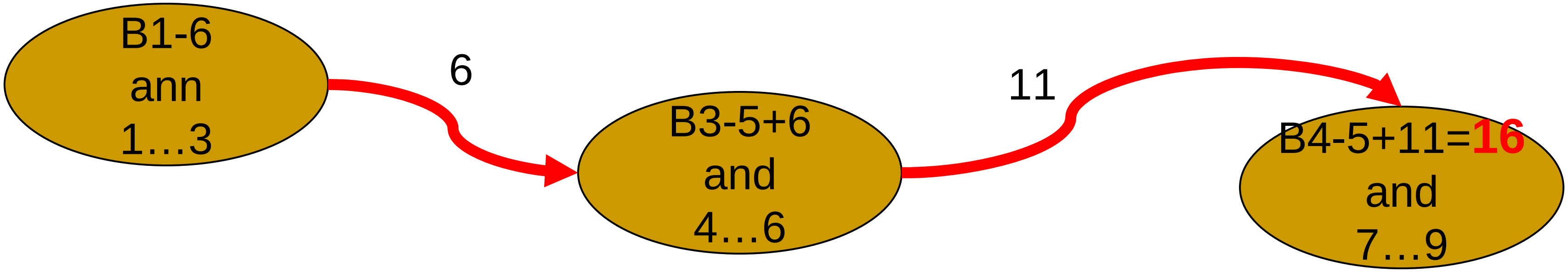
The solution: annandand

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

1 – for mismatch

-1 – for indel



G	1	2	3	4	5	6	7	8	9	10
	a	n	n	a	n	d	a	n	d	y

Do you see the problem with this solution?

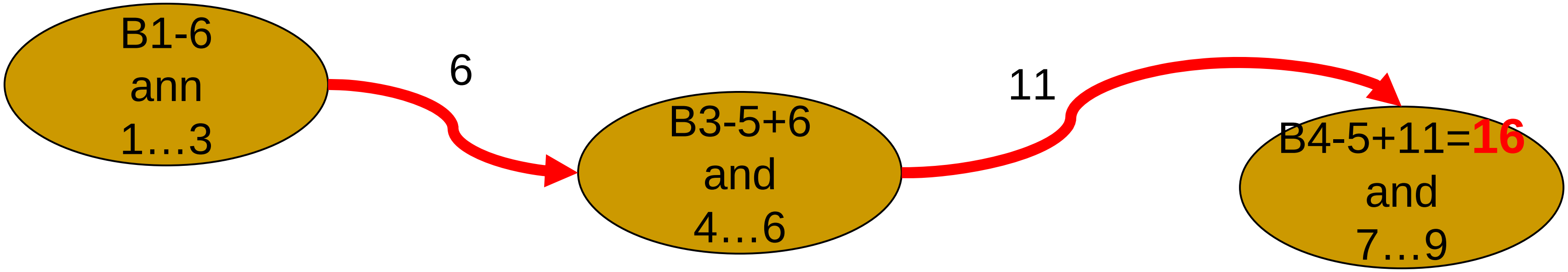
The solution: annandand

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

1 – for mismatch

-1 – for indel



G	1	2	3	4	5	6	7	8	9	10
	a	n	n	a	n	d	a	n	d	y

The solution is incorrect, since we gave weight based on local similarity of putative exons, not taking into account the relative order of substrings in cDNA

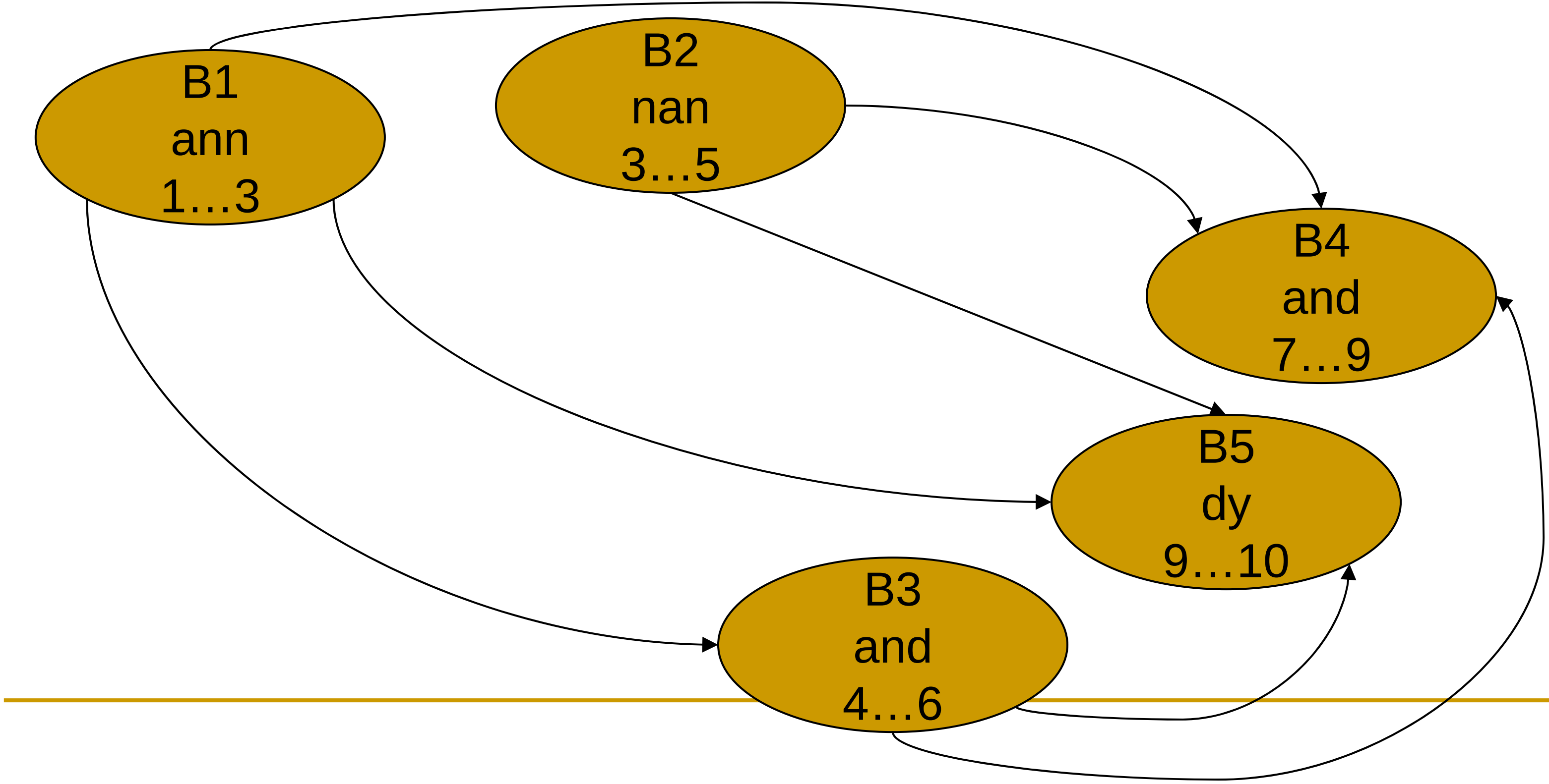
What we really want: the path through the graph with an *overall maximum total* similarity to cDNA

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

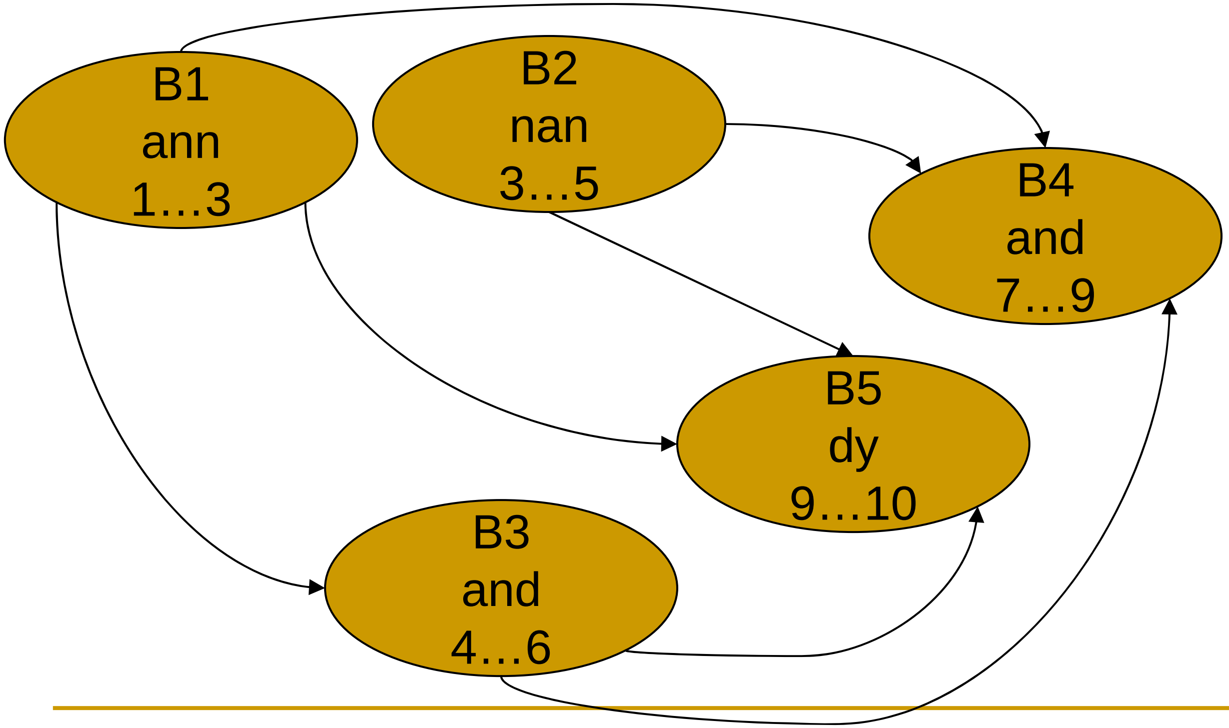
1 – for mismatch

-1 – for indel



Spliced alignment algorithm

cDNA	1	2	3	4	5
	n	a	n	n	y



The main idea:

1. Compute the total alignment score for each path in the graph by normal dynamic programming

2. When adding the next block, for each value in the DP table choose max between all predecessors (if they exist)

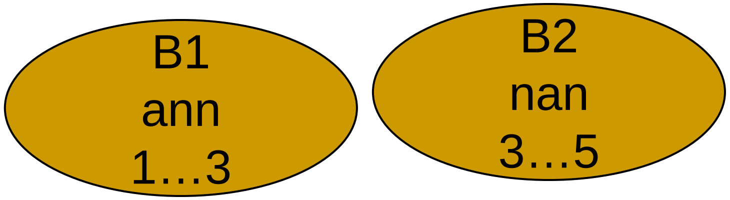
Spliced alignment algorithm – demo 1. Compute similarity score between cDNA and the nodes without parents

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

1 – for mismatch

-1 – for indel


B1
ann
1…3

B2
nan
3…5

B1		n	a	n	n	y
	0	-1	-2	-3	-4	-5
a	-1	1	1	0	-1	-2
n	-2	1	2	3	2	1
n	-3	0	2	4	5	4

B2		n	a	n	n	y
	0	-1	-2	-3	-4	-5
n	-1	2	1	0	-1	-2
a	-2	1	4	3	2	1
n	-3	0	2	6	5	4

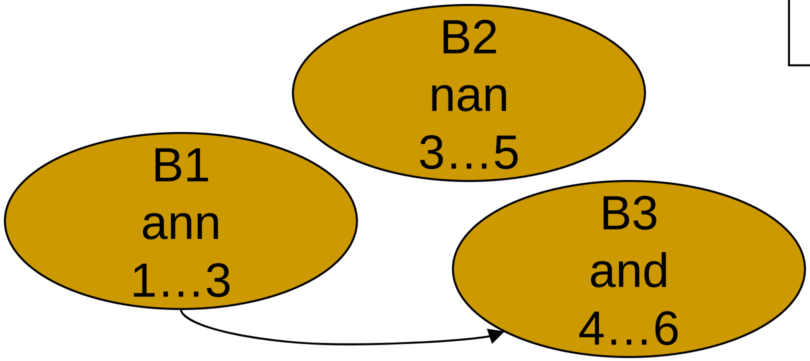
Spliced alignment algorithm – demo 2. Block 3 has only 1 incoming edge – so continue computing the DP table for concatenation of B1 and B3

cDNA	1	2	3	4	5
	n	a	n	n	y

2 – for match

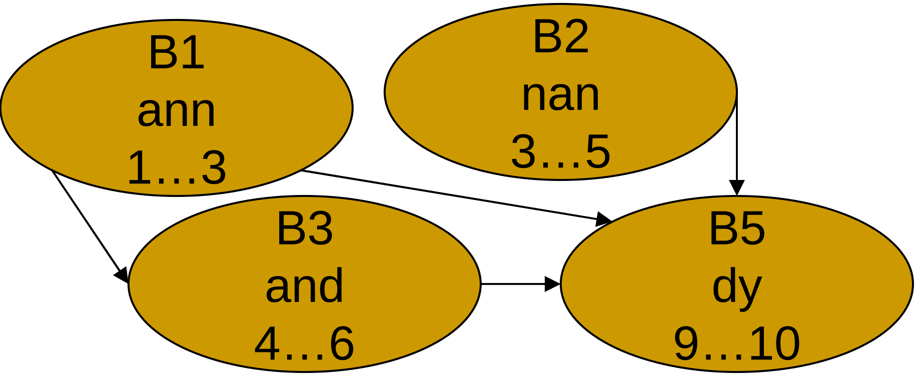
1 – for mismatch

-1 – for indel



			n	a	n	n	y
		0	-1	-2	-3	-4	-5
B1	a	-1	1	1	0	-1	-2
	n	-2	1	2	3	2	1
	n	-3	0	2	4	5	4
B3	a	-4	-1	2	3	4	4
	n	-5	-2	1	4	5	4
	d	-6	-3	0	3	4	6

Spliced alignment algorithm – demo 3. Block B5 has 3 incoming edges. Continue DP table with B5. In each cell choose max between 3 values of parent blocks



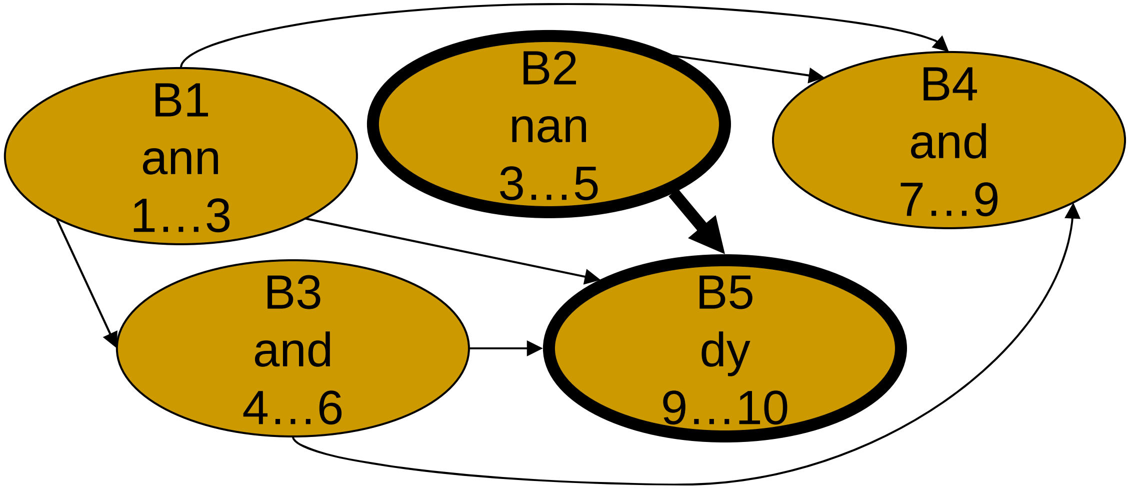
	B1, B3		n	a	n	n	y
		0	-1	-2	-3	-4	-5
	a	-1	1	1	0	-1	-2
	n	-2	1	2	3	2	1
	n	-3	0	2	4	5	4
	a	-4	-1	2	3	4	4
	n	-5	-2	1	4	5	4
	d	-6	-3	0	3	4	6
B5	d						
	y						

	B1		n	a	n	n	y
		0	-1	-2	-3	-4	-5
	a	-1	1	1	0	-1	-2
	n	-2	1	2	3	2	1
	n	-3	0	2	4	5	4
B5	d						
	y						

	B2		n	a	n	n	y
		0	-1	-2	-3	-4	-5
	n	-1	2	1	0	-1	-2
	a	-2	1	4	3	2	1
	n	-3	0	2	6	5	4
B5	d	-4	-1	1	5	7	6
	y	-5	-2	0	4	6	9

Spliced alignment algorithm – solution.

Path B2-B5 has a highest total score: 9.



Path B2, B5 – total score 9

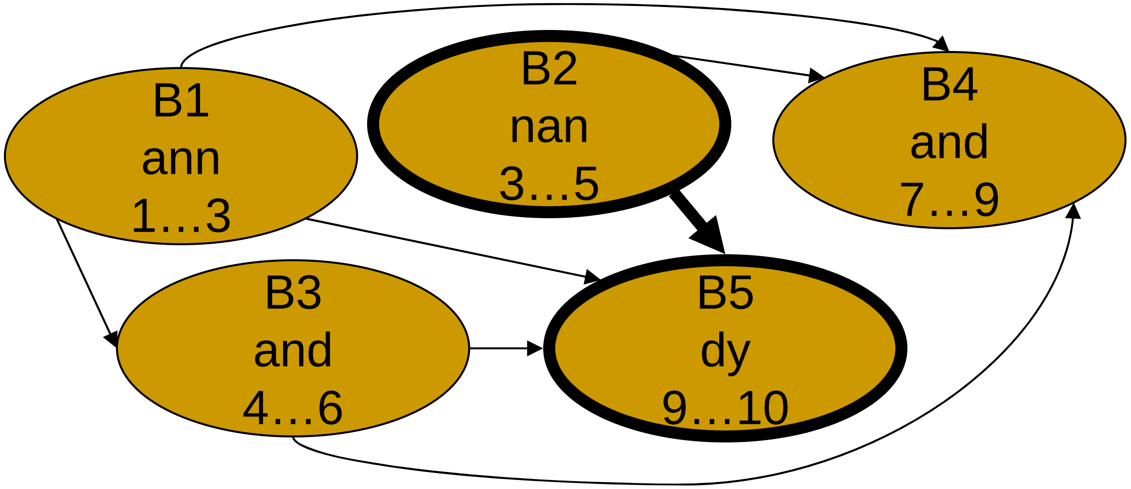
n	a	n	d	y
n	a	n	n	y
2	2	2	1	2

G	1	2	3	4	5	6	7	8	9	10
	a	n	**n**	**a**	**n**	d	a	n	**d**	**y**

The gene consists of 2 exons

Spliced alignment algorithm – solution.

Path B2-B5 has a highest total score between all possible paths. Check



Path B1, B5 – total score 6

-	a	n	n	d	y
n	a	n	n	-	y
-1	2	2	2	-1	2

Path B2, B4 – total score 8

n	a	n	a	n	d
n	a	n	-	n	y
2	2	2	-1	2	1

Path B1, B3, B5 – total score 5

a	n	n	a	n	d	a	n	d
-	-	n	a	n	-	-	n	y
-1	-1	2	2	2	-1	-1	2	1

Path B1, B3, B4 – total score 3

a	n	n	a	n	d	d	y
-	-	n	a	n	n	-	y
-1	-1	2	2	2	-1	-1	2

Steps of finding protein-coding DNA ab initio: summary

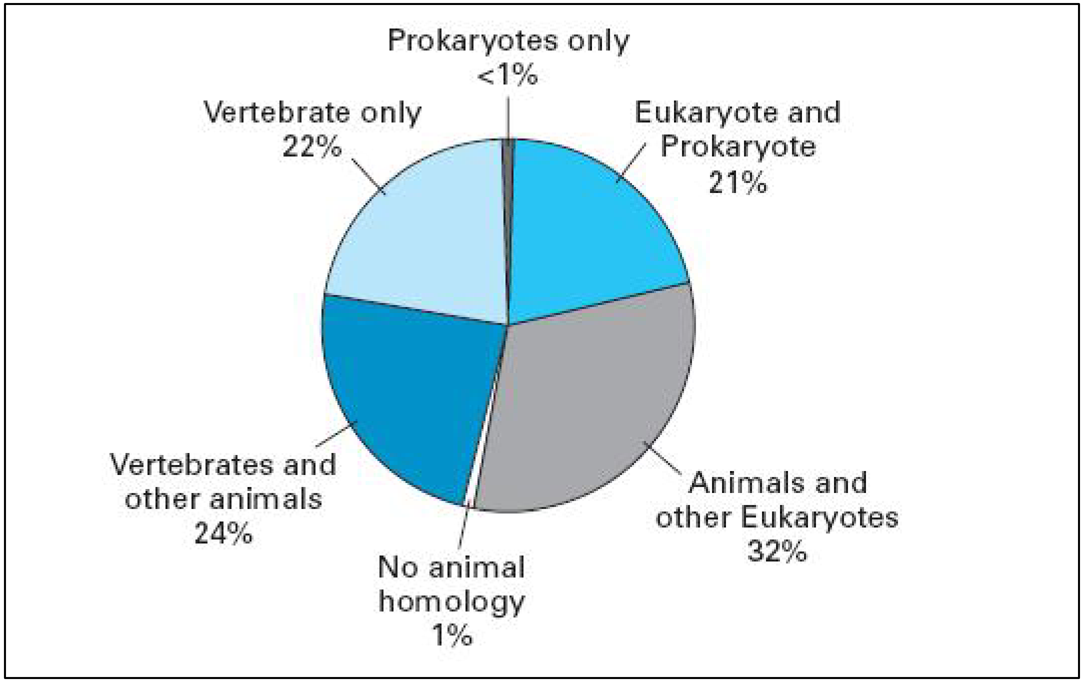
- Partition genomic DNA G into open reading frames, according to the occurrence of stop codons
- Leave only sufficiently large ORFs
- For each such ORF, analyze the content of dinucleotides using HMM for CpG islands
- Leave only sequences which contain CpG islands
- Check for start and stop signals of transcription: promoter region, polyA tail
- The remaining sequences are candidates for being protein-coding genes

am : locating specific gene (Eukaryotes)

- By physical mapping, determine in what chromosome region the gene of interest occurs
- Sequence mRNA (cDNA) of a target gene
- Perform ab initio search for coding sequences in the specified region (see previous slide), produce the set of putative genes
- In each putative gene, find putative exons (by searching for flanking dinucleotides)
- Find the chain of non-overlaping putative exons with the best total score between this chain and the target cDNA
- Or, do the alignment of the putative gene with the target cDNA using specific score matrix

The results of gene prediction in Human genome

The predicted human proteins – what makes us unique?


Prokaryotes only
<1%
Vertebrate only
Eukaryote and
Prokaryote
Vertebrates and
other animals
No animal
homology
Animals and
other Eukaryotes