Announcements

- Homework 1 is published on the website, submit until Tuesday November 9 22:00

Journal club: groups

- Groups published on the course website
- MS teams has a channel for each group, use it to communicate within group (chat, online meetings, document sharing)
- Group 4 has three members who do not speak Slovak, two of them are not located in Bratislava

Journal club: meeting

- Everybody first reads the assigned paper individually,
  then organize a group meeting, where you discuss the paper
  (particularly any portions which you did not understand),
  plan writing of the journal club report
- The first group meeting should occur no later than Nov. 23.
  It can take place in MS Teams or in person.
- Announce the first meeting at 1 day in advance
  (time and location or link) in the group channel chat
- After the meeting, post a short summary to the group channel:
  who participated, what did you agree upon, any problems
- You can arrange a consultation with us if needed.
- You do not need to report any additional meetings.

Journal club report

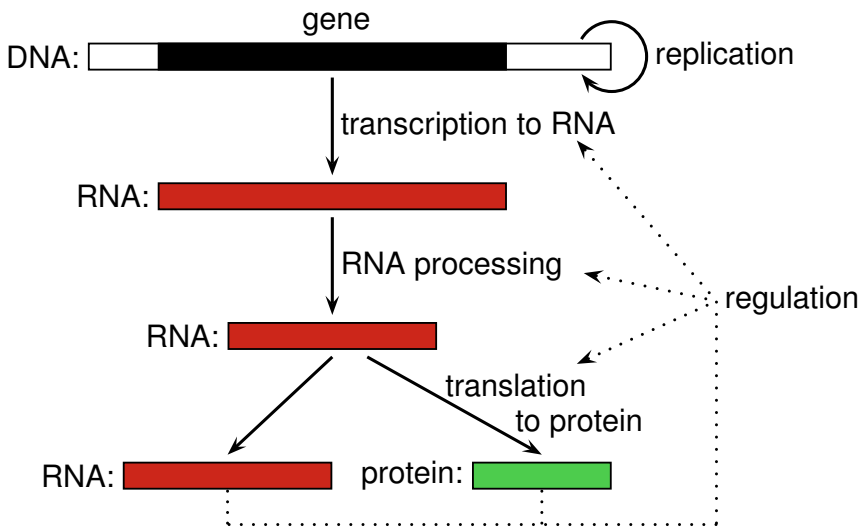
- The main methods and results of the article in your own words
- Understandable for students of this course (both computer scientists and biologists)
- You do not have to cover the entire content of the article in the report and, conversely, you can use other resources
- Try to express your own view of the topic, do not strictly follow the text of the article
- The recommended length is about 1-2 pages per person, one coherent text
- The report should list the members of the group who have actively participated. They will get the same points (the rest zero)

Gene finding

Tomáš Vinař

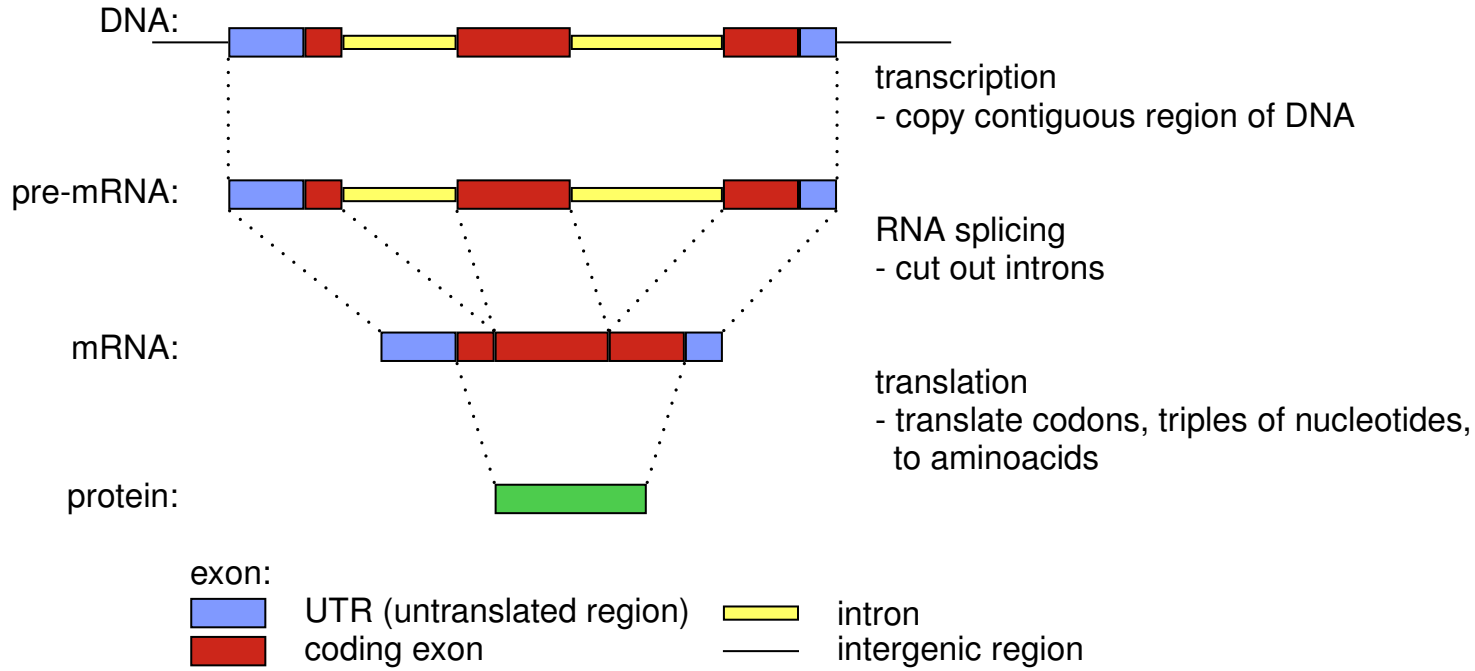
October 21, 2021

What to do with sequenced and assembled genomes?

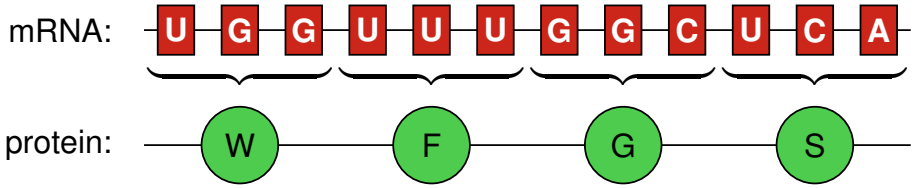


- protein-coding genes (today's lecture)
- RNA genes
- signals for regulation of transcription, splicing, etc.
- pseudogenes (non-functional copies of genes)
- sequence repeats

Protein synthesis and translation in Eukaryotes



Translation: three nucleotides (codon) $\rightarrow$ aminoacid in the protein



Human genome

- protein-coding genes
  - cca 20,000, cover approx. 40% of genome length
  - cca 10 exons in each gene
  - exons cover approx. 2% of genome length
  - coding exons approx. 1.2%

- sequence repeats
  - cover approx. 49% of genome length

Bioinformatics problem: Gene finding

Goal: find all protein-coding genes in the genome
(assemble a catalogue of all proteins)

Simplifying assumpations:

- no alternative splicing, no overlapping genes
- we are not searching for untranslated regions (UTRs) at the beginning and the end of the gene

Bioinformatics problem: Gene finding

Input: DNA sequence

```
cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
```

Bioinformatics problem: Gene finding

Goal: mark each base as intron/exon/intergenic

cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca

tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca

tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca

ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta

aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc

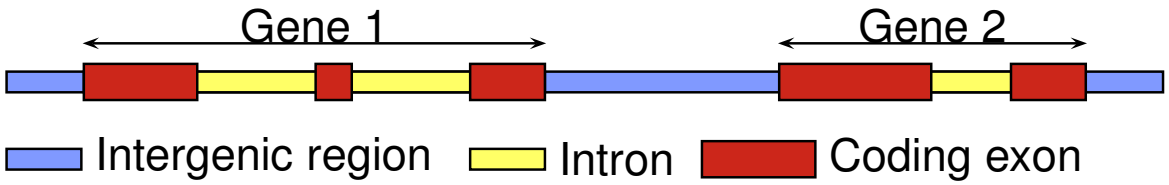
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc

gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc

gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc

tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg

acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac



Bioinformatics problem: gene finding

Input: DNA sequence

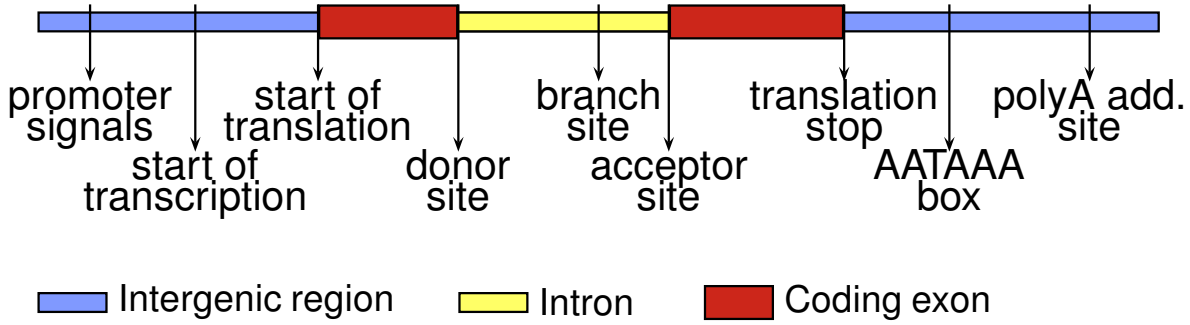
Goal: mark each base as intron/exon/intergenic

- Still not a well-defined problem!
  How to recognize a gene?

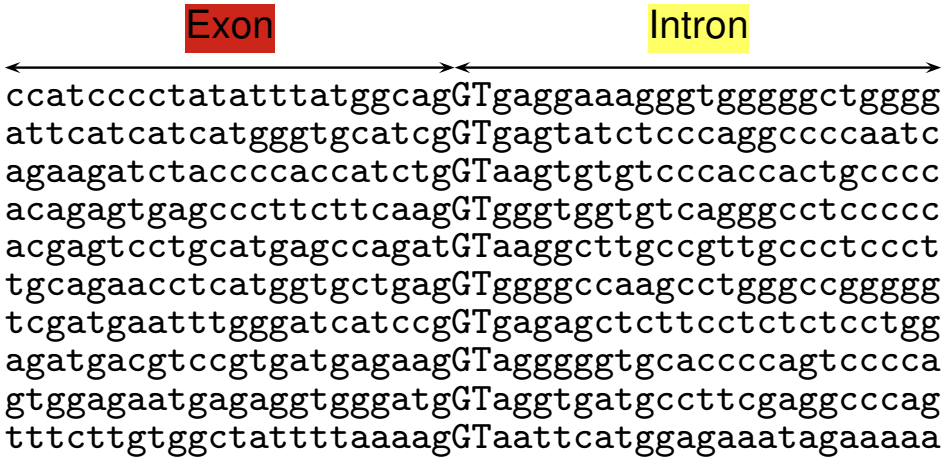
How to recognize a gene?

Signals at the exon boundaries:

short strings that serve as binding sites for the transcription machinery



Example of a signal: donor splice site



How to recognize a gene?

Sequence composition:

- different k-mer frequency in coding and non-coding regions,
- coding regions are 3-periodic,
- stop codons (TAA, TGA, TAG) appear only at the end of the last exon.

Example: in human genome, exons are GC rich

		a	c	g	t
coding exon	0	0.26	0.26	0.32	0.16
	1	0.30	0.24	0.20	0.26
	2	0.17	0.32	0.31	0.20
intron		0.26	0.22	0.22	0.30
intergenic		0.27	0.23	0.23	0.27

Bioinformatics problem: Gene finding

Input: DNA sequence

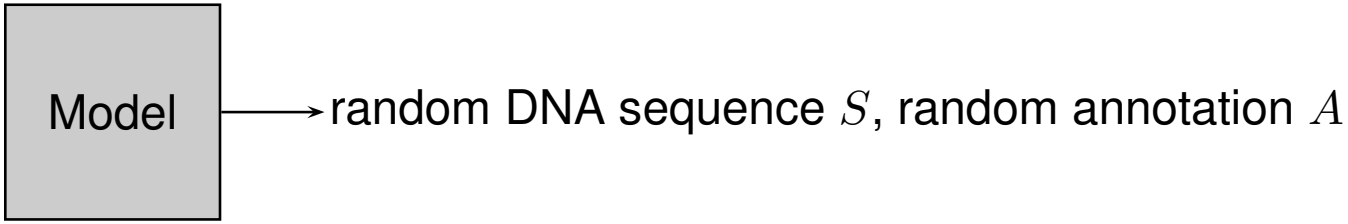
Goal: mark each base as intron/exon/intergenic

- Not a well-defined problem!
  How to recognize a gene?
- No information **by itself** can uniquely determine which parts of the sequence correspond to genes.
- Want a **scoring scheme** that will tell us how well a particular annotation corresponds to our knowledge about gene structure.
- Then we are looking for an annotation (or a segmentation of the sequence into non-overlapping regions representing individual genes) with the **maximum score.**
- We use **probabilistic models** to define such scoring scheme.

Probabilistic model of protein-coding genes

No single source of information uniquely determines genes

Combine all sources using probablistic models

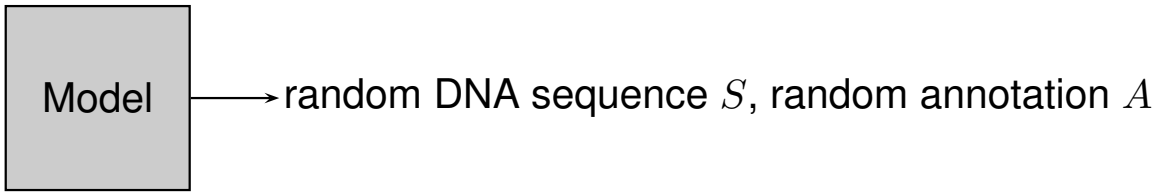


$\Pr(S, A)$ – probability model generates pair (S, A).

Model with high probability generates pairs with properties similar to the real genes

Using a probabilistic model: for a new sequence S find the most probable annotation $A = \arg\max_A \Pr(A|S)$

Probabilistic model of protein-coding genes



Using a probabilistic model: for a new sequence S find the most probable annotation $A = \arg\max_A \Pr(A|S)$

Toy example: sequences of length 2

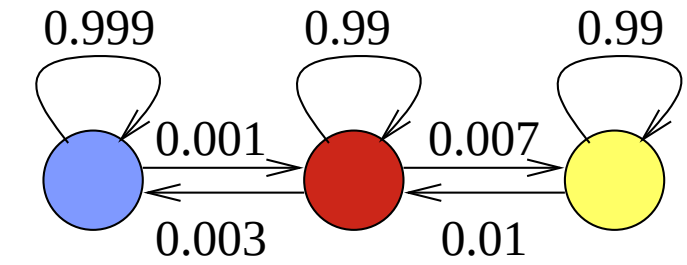
Table of probabilities for 16 sequences, 9 annotations (sums to 1)

The most probable annotation for $S =$aa is aa.

aa	0.008	ac	0.009	ag	0.0085	...
aa	0	ac	0	...		
aa	0.011	...				
aa	0					
aa	0.009					
aa	0					
aa	0.007					
aa	0					
aa	0.010					

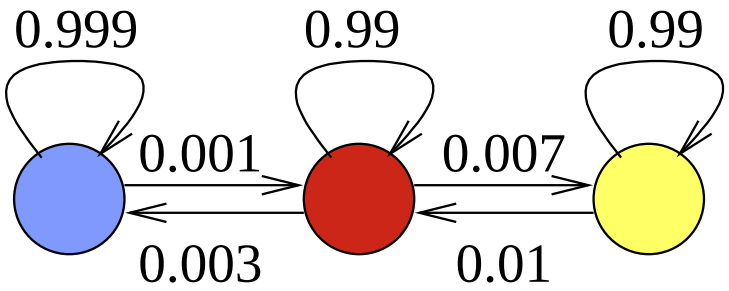
Hidden Markov model (HMM)

Way of defining models for longer sequences.



- Finite-state automaton, states e.g. exon, intron, intergenic
- Generates sequences and annotations base-by-base
- In each step, in the current state, randomly generate a single base according to the state's emmission table
- Then randomly transition to the next state according to the probabilities on edges (transition probabilities)

Hidden Markov model (HMM)



a:0.27	a:0.24	a:0.26
c:0.23	c:0.27	c:0.22
g:0.23	g:0.28	g:0.22
t:0.27	t:0.21	t:0.30

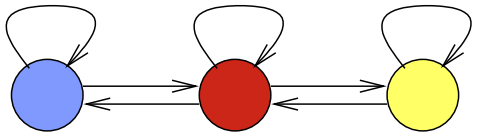
Assume the model starts in the blue state

Example:

$$\Pr(\text{aca}) = 0.27 \cdot 0.001 \cdot 0.27 \cdot 0.99 \cdot 0.24 = 0.000017$$

$$\Pr(\text{aca}) = 0.27 \cdot 0.999 \cdot 0.23 \cdot 0.999 \cdot 0.27 = 0.017$$

Notation



Sequence $S_1, \ldots, S_n$

Annotation $A_1, \ldots, A_n$

Model parameters:

Transition probability $a(u, v) = \Pr(A_{i+1} = v | A_i = u)$,

Emission probability $e(u, x) = \Pr(S_i = x | A_i = u)$,

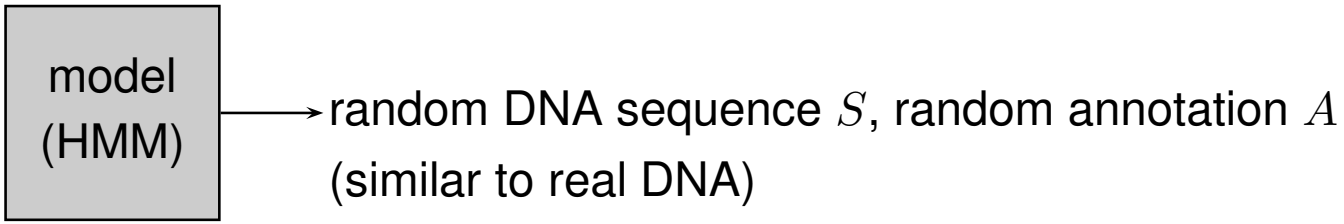
Starting probability $\pi(u) = \Pr(A_1 = u)$.

a			
	0.99	0.007	0.003
	0.01	0.99	0
	0.001	0	0.999

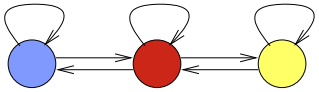
e	a	c	g	t
	0.24	0.27	0.28	0.21
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

Resulting probability: $\Pr(A_1, \ldots, A_n, S_1, \ldots, S_n) =$
$\pi(A_1)e(A_1, S_1) \prod_{i=2}^{n} a(A_{i-1}, A_i)e(A_i, S_i)$

Finding genes with HMMs

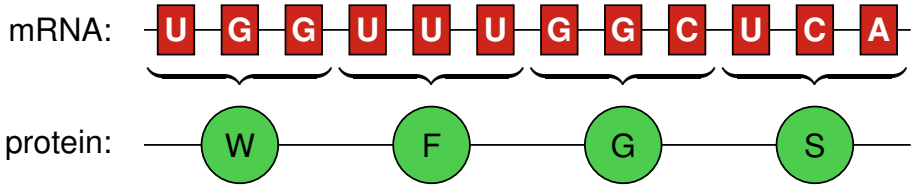


$\Pr(S, A)$ – probability that the model generates pair (S, A).

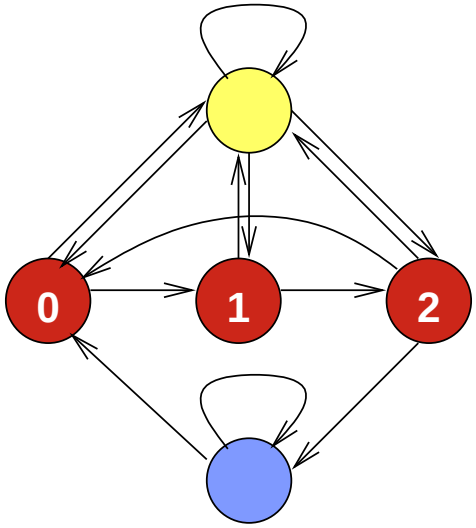
- **Determine states and transitions of the model:** by hand based on your knowledge about the gene structure 
- **Parameter training:** emission and transition probabilities are determined based on the real sequences with known genes (**training set**)
- **Use:** for a new sequence S, find the most probable annotation $A = \arg\max_A \Pr(A|S)$
  Viterbi algorithm in $O(nm^2)$ (dynamic programming)

Gene finding HMM: 3-periodic exons

three nucleotides (codon) $\rightarrow$ aminoacid in the protein



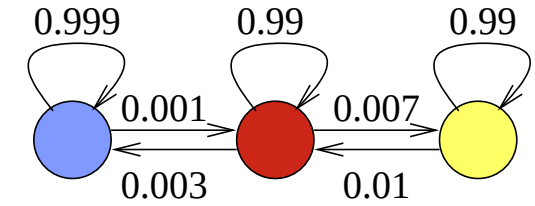
Instead of a single state for exon, use three states in a cycle

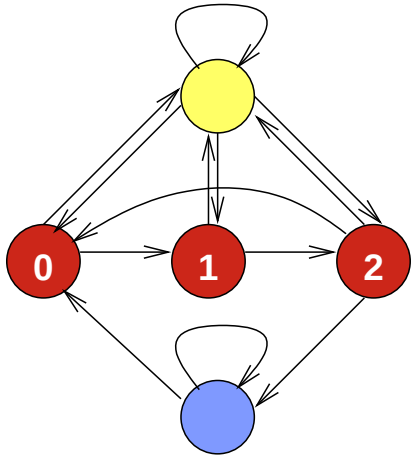


a	0	1	2		
0	0		0		0
1	0	0			0
2		0	0		
					0
		0	0	0	

$\Pr(A_i|A_{i-1})$

Emission probabilities of new states will differ



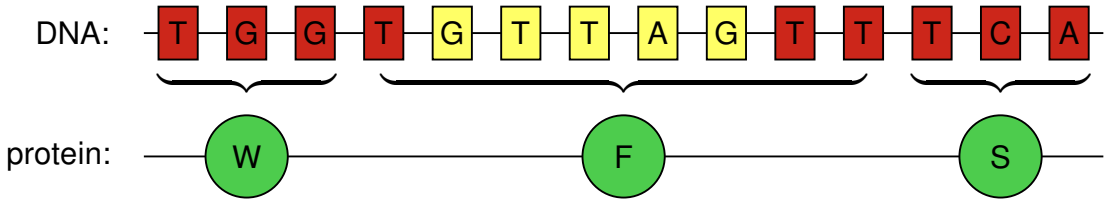


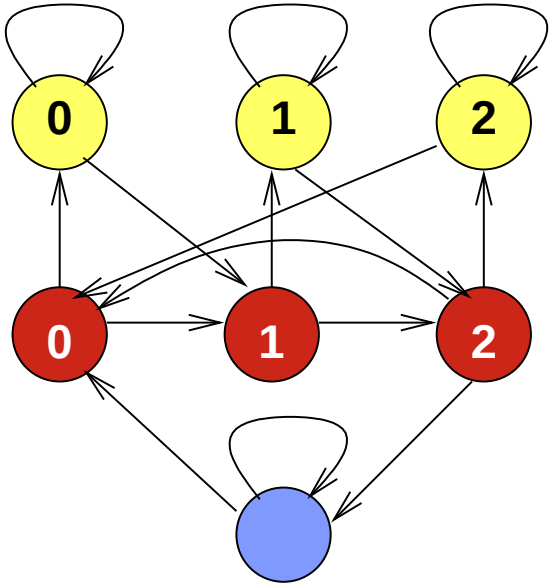
e	a	c	g	t
	0.24	0.27	0.28	0.21
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

e	a	c	g	t
0	0.26	0.26	0.32	0.16
1	0.30	0.24	0.20	0.26
2	0.17	0.32	0.31	0.20
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

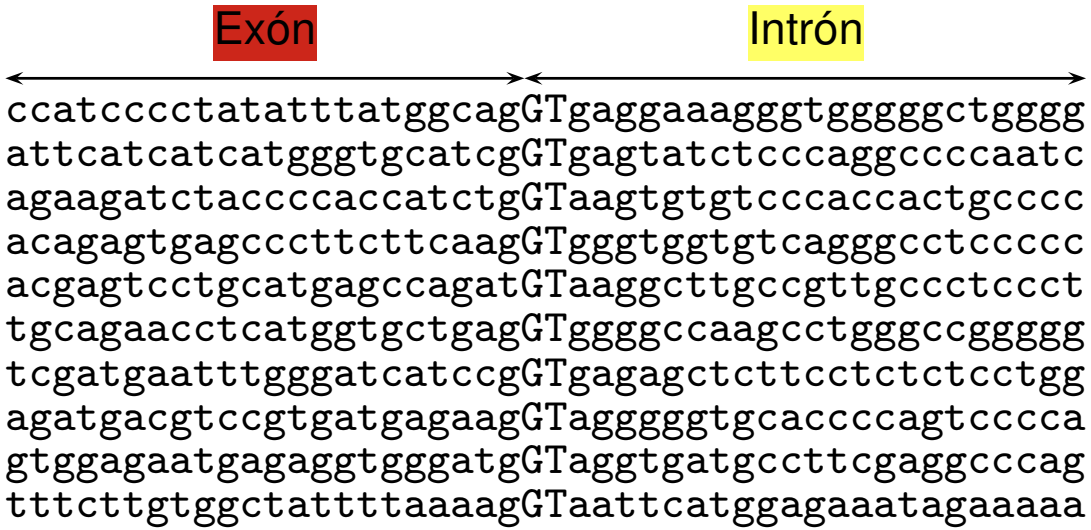
Gene finding HMM: consistent codons

Intron can interrupt a codon in the middle, but we must continue where we left off.

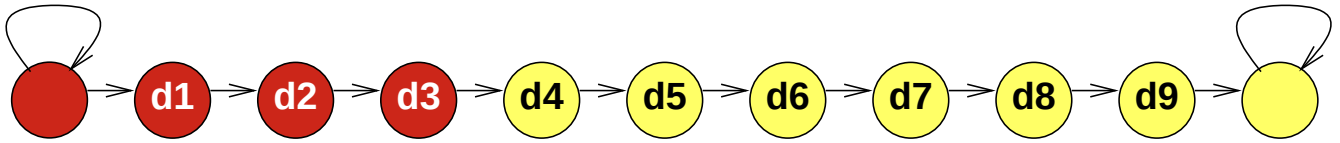




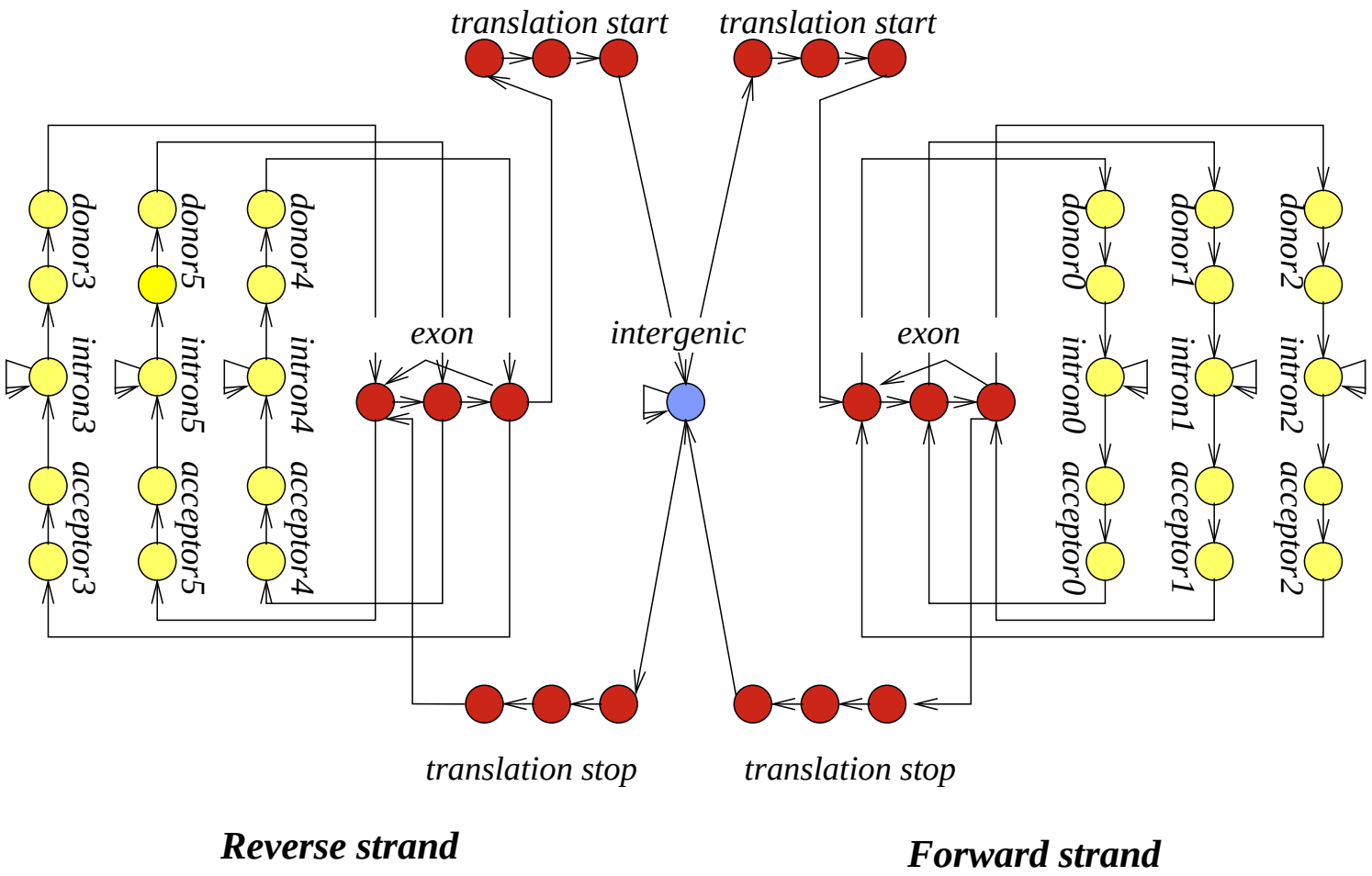
Gene finding HMM: signals



Add a sequence of states between exon and intron:



Gene finding HMM: a complete model



Higher order states

Order 0: emission table e contains $\Pr(S_i|A_i)$

Order 1: e contains $\Pr(S_i|A_i, S_{i-1})$

A_i	S_{i-1}	a	c	g	t
	a	0.24	0.23	0.34	0.19
	c	0.30	0.31	0.13	0.26
	g	0.27	0.28	0.28	0.17
	t	0.13	0.28	0.38	0.21
	a	0.30	0.18	0.27	0.25
	c	0.32	0.28	0.06	0.35
	g	0.27	0.22	0.27	0.24
	t	0.20	0.21	0.26	0.33

...

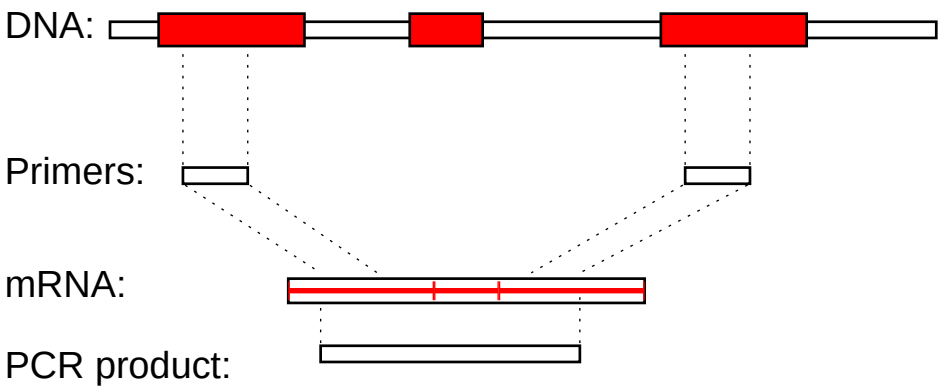
For exons, introns, etc. use orders 4-5.

Experimental verification of predicted genes

Transcription and splicing

- RNA-Seq: sequencing of all mRNAs extracted from the cell
- RT PCR: targeted verification of a specific gene using specifically designed primers

Problems: difficult to find genes that are expressed under special conditions, i.e. embryonic development
genomic DNA contamination, non-unique mapping to the genome



Experimental verification of predicted genes

Translation, existence of the protein

- Mass spectrometry
- Detection based on antibodies
- Other methods specific to individual proteins

Examples of gene finging programs

Based only on DNA sequence:

HMMGene [Krogh, 1997], Genscan [Burge and Karlin, 1997], GeneZilla [Majoros et al., 2004], ExonHunter [Brejová et al., 2005], Augustus [Stanke and Waack, 2003]

Prokaryotes:

GeneMark [Lukashin and Borodovsky, 1998], Glimmer [Delcher et al., 1999].

Examples of gene finding programs

Comparison of multiple sequences:

Twinscan [Korf et al., 2001], Exoniphy [Siepel and Haussler, 2004], N-SCAN [Gross and Brent, 2006]
(Twinscan extended to multiple genomes).

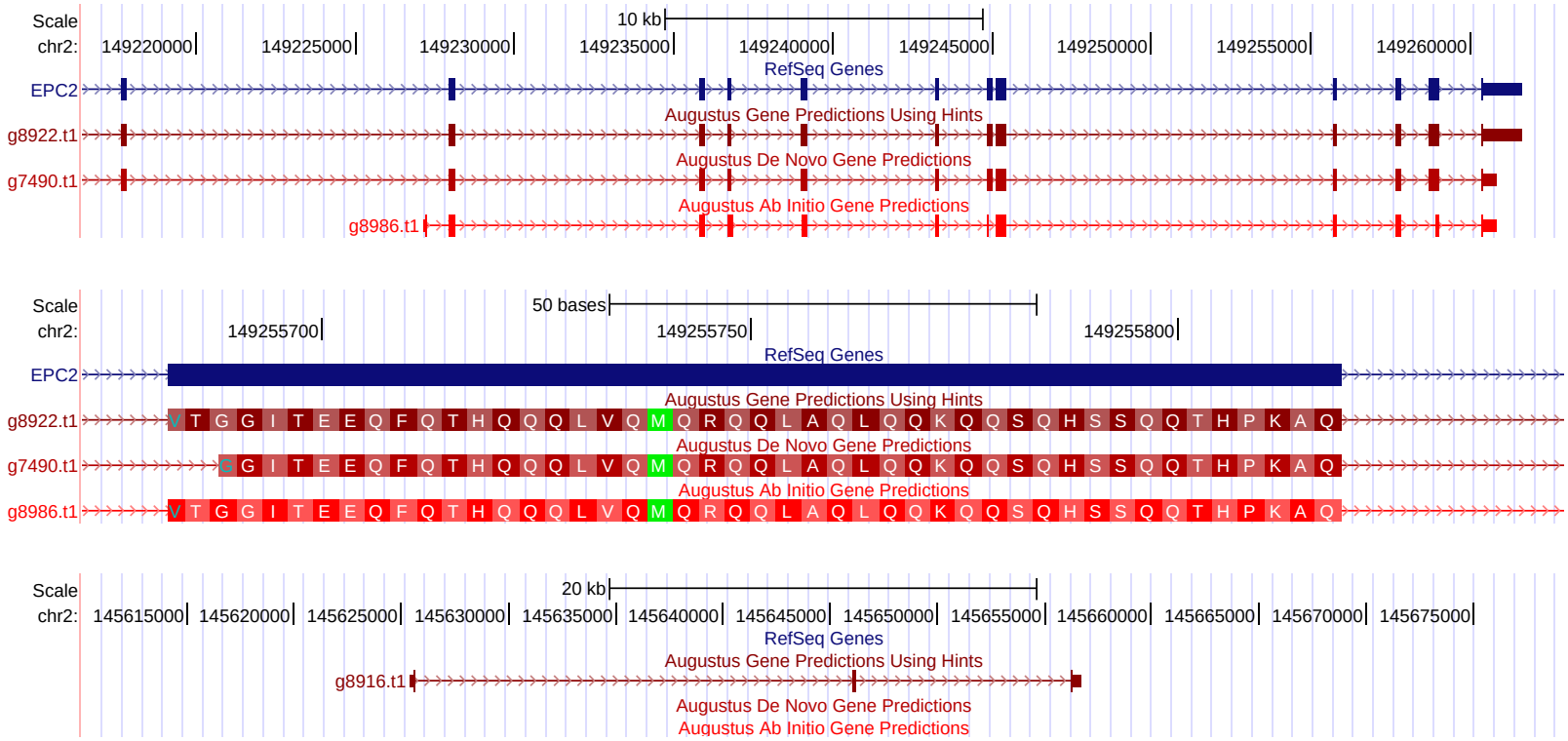
Integration of additional information: (RNA-seq, proteins from related genomes, etc.)
ExonHunter [Brejová et al., 2005], Augustus [Stanke et al., 2006], Jigsaw [Allen and Salzberg, 2005], Fgenesh++ [Solovyev et al., 2006].

Limitations of gene finders

- Alternative splicing: one gene can produce different mRNAs; gene finders typically only find one

- Overlapping genes (including genes in introns)
- Atypical genes (unusual signal, short or long exons or introns)
- Untranslated regions (UTR) are difficult

Gene finders often make errors



Best results on human genome: [Guigo et al 2006]

20% genes, 60% exons correct based on DNA

35% genes, 65% exóns correct based on comparisons

70% génes, 85% exóns correct with additional info

How many genes in human genome?

Before 2001: **50 000–140 000** genes

2001: draft human genome: **30 000–40 000** genes

2004: completed human genome: **20 000–25 000** genes

2007: Ensembl, RefSeq, VEGA catalog: **24 500** genes

[Clamp a kol. 2007] claims only **20 500** correct

Are there genes about which we don't know yet?

2010: RefSeq **22 333** genes

uncertainty of ± 1000 [Pertea, Salzberg 2010]

Individuals can differ in tens of genes

2012: Project ENCODE estimates **20 687** protein coding genes,

on average 6 transcripts per gene,

plus 8 800 short and 9 600 long RNA genes

Summary

- Newly sequenced genomes need to be annotated:
  determine functions of individual segments of the genome
- Example of annotation: finding genes that code for proteins
- Hidden Markov models are suitable for gene finding
- Models make a lot of errors, but they at least give us the basic understanding of location an number of genes, we can study their function