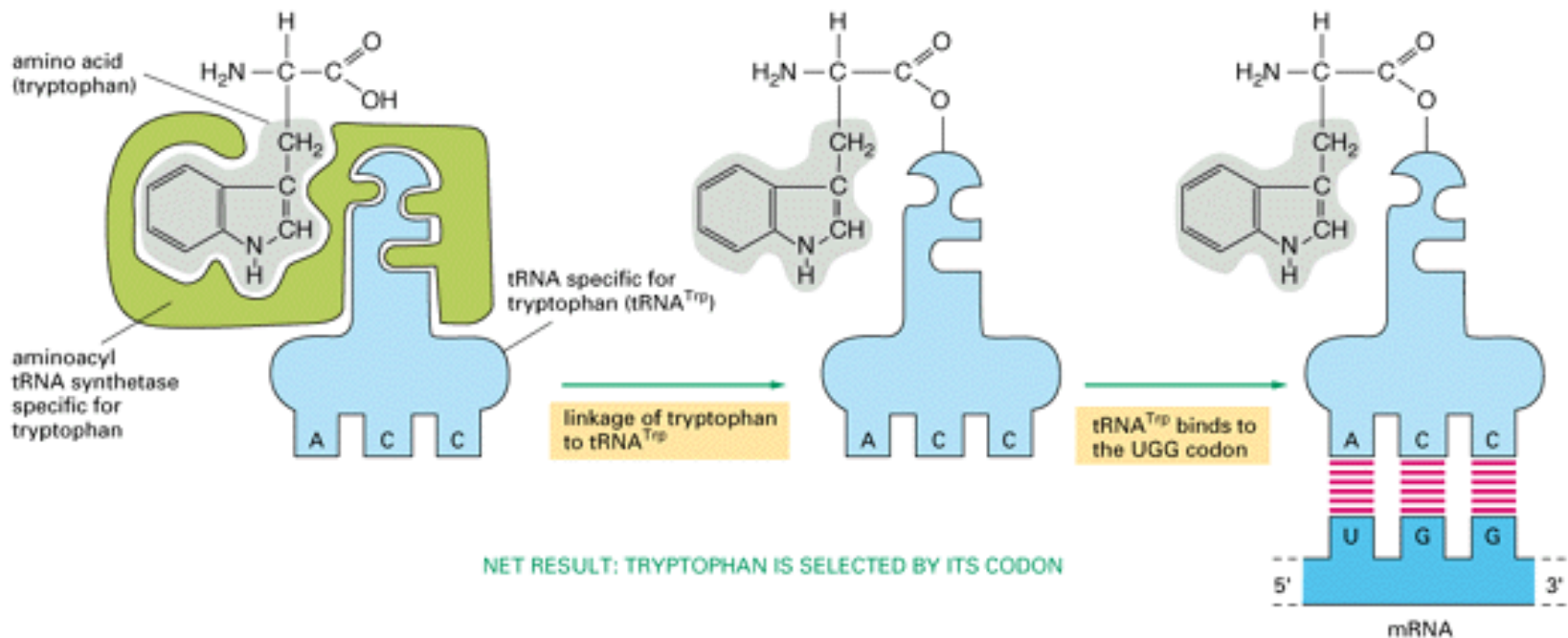


The transfer RNAs (tRNAs)

- Each tRNA has an anticodon and a corresponding amino acid attached at the 3' end
- There are ~31 tRNAs (varying according to species) for 61 codons



The genetic code

1st position (5' end) ↓	2nd position				3rd position (3' end) ↓
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

- The UAA, UAG and UGA codons are **stop** codons because there are no corresponding tRNAs (except exceptions...)
- The UGA codon can code for **selenocysteine** depending on the biological context (21st aa)
- The UAG codon can code for **pyrrolysine** depending on the biological context (22nd aa)
- The initiation methionine (Met or M) is coded by AUG (except exceptions...)
- **Conceptual translation:** translation according to the genetic code without experimental verification

The 21st amino acid: selenocysteine (Sec)

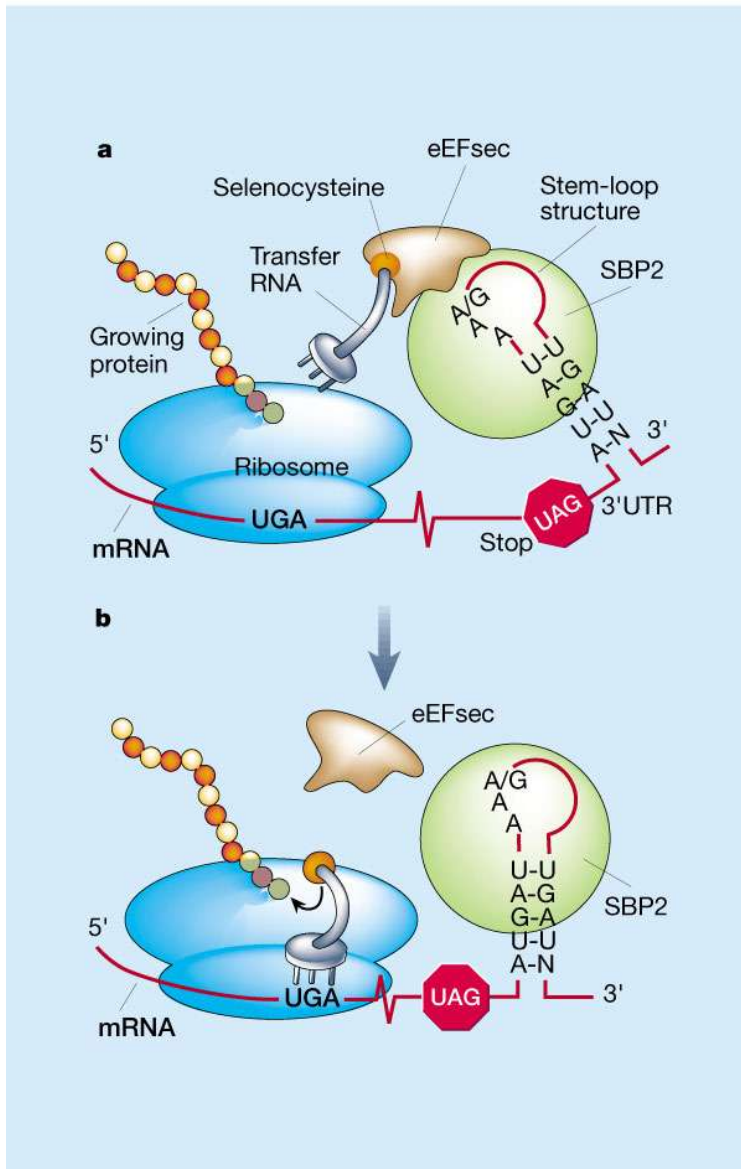
Eukaryotic example

- The mRNA has a stem-loop structure in the 3' end
- The stem-loop is recognised by 2 specific proteins (SBP2 and then eEFsec)
- eEFsec recruits the tRNA which carries the selenocysteine
- The selenocysteine is incorporated at the position equivalent to the STOP codon
- The protein synthesis can continue...

Atkins and Gesteland (2000), *Nature*, 407, 463-465

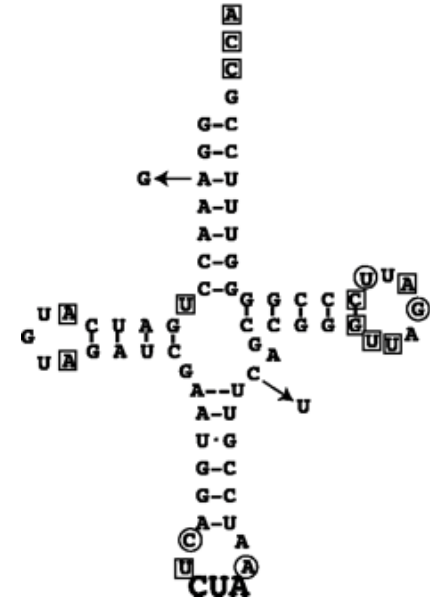
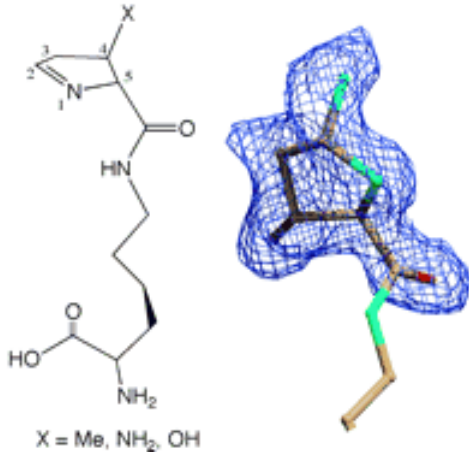
General

- Selenocysteine is found in e.g. certain glutathione peroxidases, 5' deiodianase, thioredoxin reductase, formate dehydrogenase, hydrogenases...



The 22nd amino acid: pyrrolysine

- tRNA complementary to a UAG stop codon carries a Lysine (Lys or K) converted into pyrrolysine



- Found in Archaea and Eubacteria
- The insertion mechanism is not known
- Pyrrolysine is found in methanogen methyltransferase

Srinivasan *et al.* (2002), *Science*, 296, 1459-1462

Hao *et al.* (2002), *Science*, 296, 1462-1466

The genetic code is almost universal

Codons	Code « universel »	Codes mitochondriaux			
		Mammifères	Drosophile	Levures	Végétaux
UGA	STOP	<i>Trp</i>	<i>Trp</i>	<i>Trp</i>	STOP
AUA	Ile	<i>Met</i>	<i>Met</i>	<i>Met</i>	Ile
CUA	Leu	Leu	Leu	<i>Thr</i>	Leu
AGA } AGG }	Arg	<i>STOP</i>	<i>Ser</i>	Arg	Arg

Other “nuclear” exceptions : ciliates, euplotides, bacteria, blepharisma (macronuclear)

Translation of the amino terminal sequence of human albumin
Using different genetic codes

Standard:	MKWVTFISLLFLFSSAYSRG
Yeast mito:	MKWVTFISTTFTFSSAYSRG
Mammal mito:	MKWVTFISLLFLFSSAYS*G
Insect mito:	MKWVTFISLLFLFSSAYSSG
Plant mito:	MKWVTFISLLFLFSSAYSRG

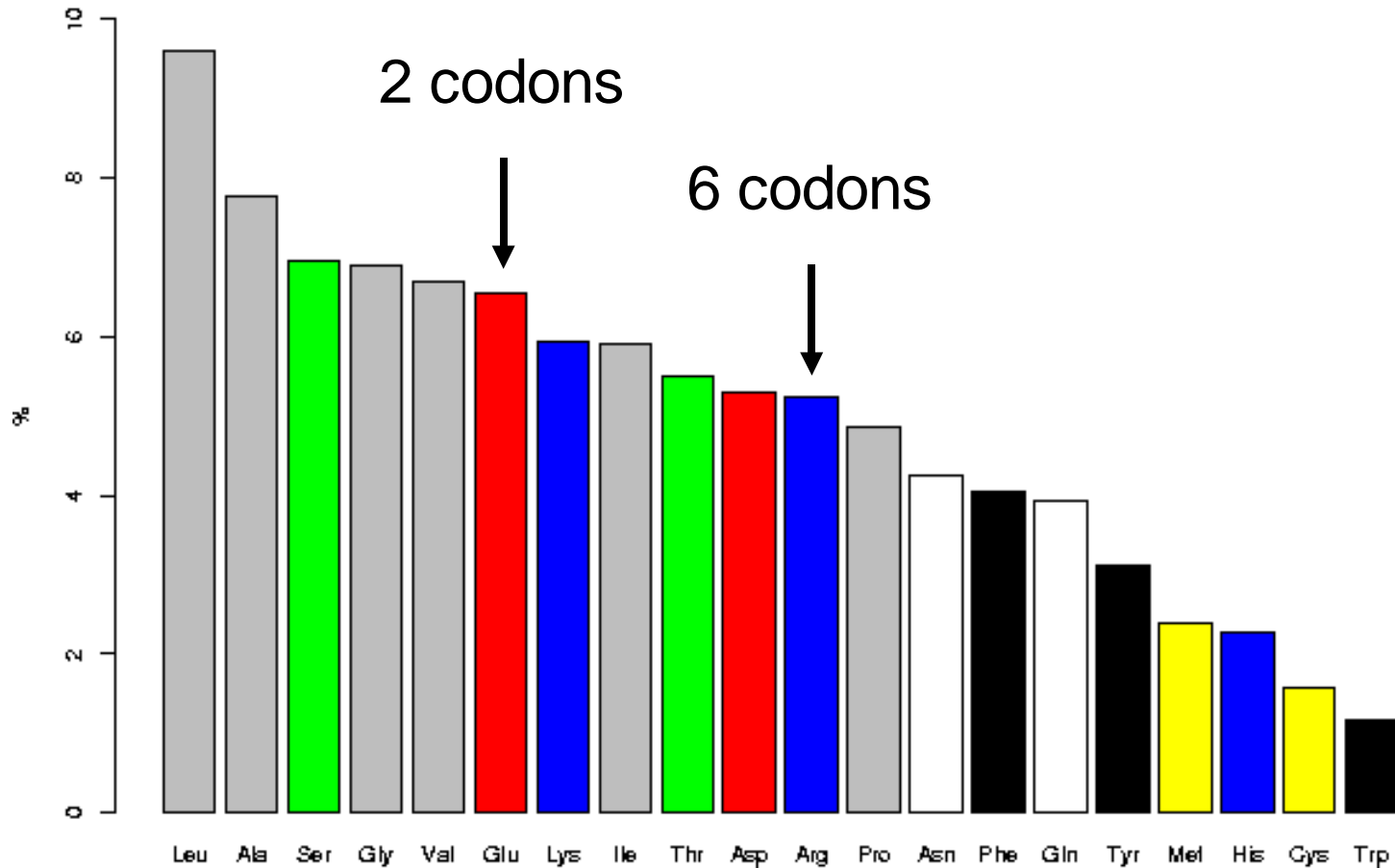
Degeneracy of the genetic code

GCA	AGA									UUA					AGC					
GCC	AGG									UUG					AGU					
GCG	CGA						GGA		AUA	CUA				CCA	UCA	ACA				GUA
GCU	CGC						GGC		AUC	CUC				CCC	UCC	ACC				GUC
	CGG	GAC	AAC	UGC	GAA	CAA	GGG	CAC	AUU	CUG	AAA		UUC	CCG	UCG	ACG		UAC		GUG
	CGU	GAU	AAU	UGU	GAG	CAG	GGU	CAU	AUU	CUU	AAG	AUG	UUU	CCU	UCU	ACU	UGG	UAU		GUU
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

- Many alternative codons for the same amino acid differs only in the 3rd position
- The most frequent amino acids do not necessarily have the highest number of codons

Amino acid frequencies in Swiss-Prot

Amino acid composition



Legend: gray = aliphatic, red = acidic, green = small hydroxy, blue = basic, black = aromatic, white = amide, yellow = sulfur

<http://www.expasy.org/sprot/relnotes/relstat.html>

Codon usage

- Specific codon usage in specific organisms
 - Relative tRNA abundance
- Specific codon usage in specific genes
- Important parameter in gene finding programmes

Example: Codon frequencies (%) for serine (Ser or S) codons in different organisms

Codon	<i>E. coli</i>	Fruitfly	Man	Maise	Yeast
AGT	3	1	10	3	5
AGC	20	23	34	30	4
TCG	4	17	9	22	1
TCA	2	2	5	4	6
TCT	34	9	13	4	52
TCC	37	48	28	37	33

The most frequently used codons have the highest chance of being present in the CDS -> this is used in gene prediction

Codon usage

Codon usage database: <http://www.kazusa.or.jp/codon/>

Homo sapiens [gbpri]: 55194 CDS's (24298072 codons)

fields: [triplet] [frequency: **per thousand**] (Amino acid)

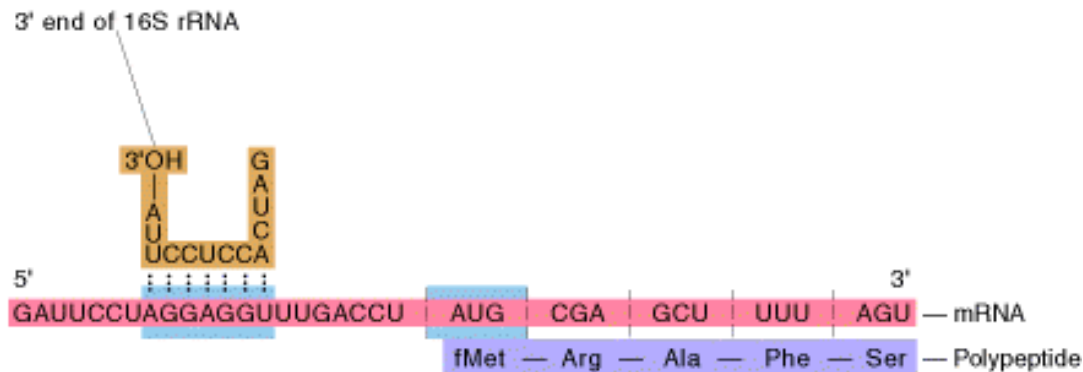
UUU 17.1 (Phe)	UCU 14.7 (Ser)	UAU 12.1 (Tyr)	UGU 10.0 (Cys)
UUC 20.6 (Phe)	UCC 17.6 (Ser)	UAC 15.5 (Tyr)	UGC 12.2 (Cys)
UUA 7.5 (Leu)	UCA 12.0 (Ser)	UAA 0.7 (STOP)	UGA 1.5 (STOP)
UUG 12.6 (Leu)	UCG 4.4 (Ser)	UAG 0.6 (STOP)	UGG 12.7 (Trp)
CUU 13.0 (Leu)	CCU 17.3 (Pro)	CAU 10.5 (His)	CGU 4.6 (Arg)
CUC 19.8 (Leu)	CCC 20.1 (Pro)	CAC 15.0 (His)	CGC 10.7 (Arg)
CUA 7.8 (Leu)	CCA 16.7 (Pro)	CAA 12.0 (Gln)	CGA 6.3 (Arg)
CUG 39.8 (Leu)	CCG 6.9 (Pro)	CAG 34.1 (Gln)	CGG 11.6 (Arg)
AUU 16.1 (Ile)	ACU 13.0 (Thr)	AAU 16.7 (Asn)	AGU 11.9 (Ser)
AUC 21.6 (Ile)	ACC 19.4 (Thr)	AAC 19.5 (Asn)	AGC 19.3 (Ser)
AUA 7.7 (Ile)	ACA 15.1 (Thr)	AAA 24.1 (Lys)	AGA 11.5 (Arg)
AUG 22.2 (Met)	ACG 6.1 (Thr)	AAG 32.2 (Lys)	AGG 11.4 (Arg)
GUU 11.0 (Val)	GCU 18.6 (Ala)	GAU 21.9 (Asp)	GGU 10.8 (Gly)
GUC 14.6 (Val)	GCC 28.4 (Ala)	GAC 25.6 (Asp)	GGC 22.5 (Gly)
GUA 7.2 (Val)	GCA 16.1 (Ala)	GAA 29.0 (Glu)	GGA 16.4 (Gly)
GUG 28.4 (Val)	GCG 7.5 (Ala)	GAG 39.9 (Glu)	GGG 16.3 (Gly)

Coding GC 52.45% 1st letter GC 56.04% 2nd letter GC 42.37% 3rd letter GC 58.93%

Translation initiation

- The methionine (Met or M) or formylmethionine (f-Met; in bacteria) is always the first amino acid in a polypeptide chain
 - Even if the first codon is not AUG but CUG, GUG or UUG
 - Even when the Met is often (in ~50% of all proteins) post-translationally removed
- There are at least two Met tRNAs, initiator tRNA-Met and internal Met tRNA-Met

Shine Dalgarno

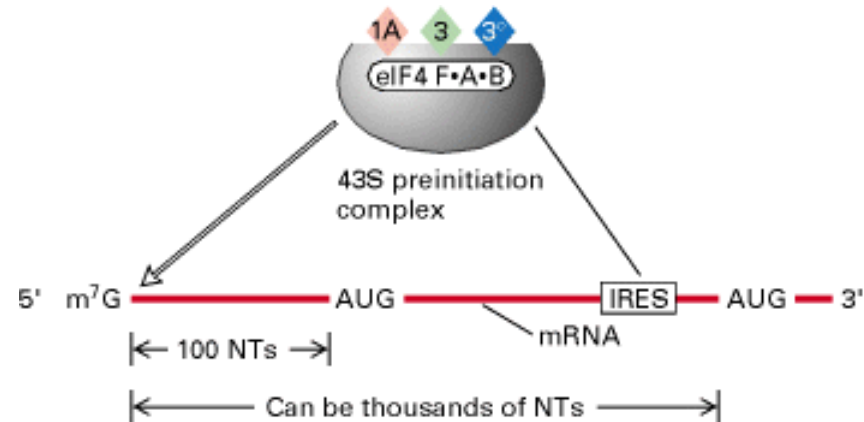


AGCACGAGGGGAAAUCUGAUGGAACGCUAC *E. coli trpA*
 UUUGGAUGGAGUGAAACGAUGGCGAUUGCA *E. coli araB*
 GGUAACCAGGUAAACAACCAUGCGAGUGUUG *E. coli thrA*
 CAUUCAGGGUGGUGAAUGUGAAACCAGUA *E. coli lacI*
 AAUCUUGGAGGCUUUUUUUAUGGUUCGUUCU ϕ X174 phage A protein
 UAACUAAGGAUGAAAUGCAUGUCUAAGACA Q β phage replicase
 UCCUAGGAGGUUUGACCUAUGCGAGCUUUU R17 phage A protein
 AUGUACUAAGGAGGUUGUAUGGAACAACGC λ phage *cro*

┌──────────┐ ┌──────────┐
 Pairs with Pairs with
 16S rRNA initiator tRNA

Alternative initiation

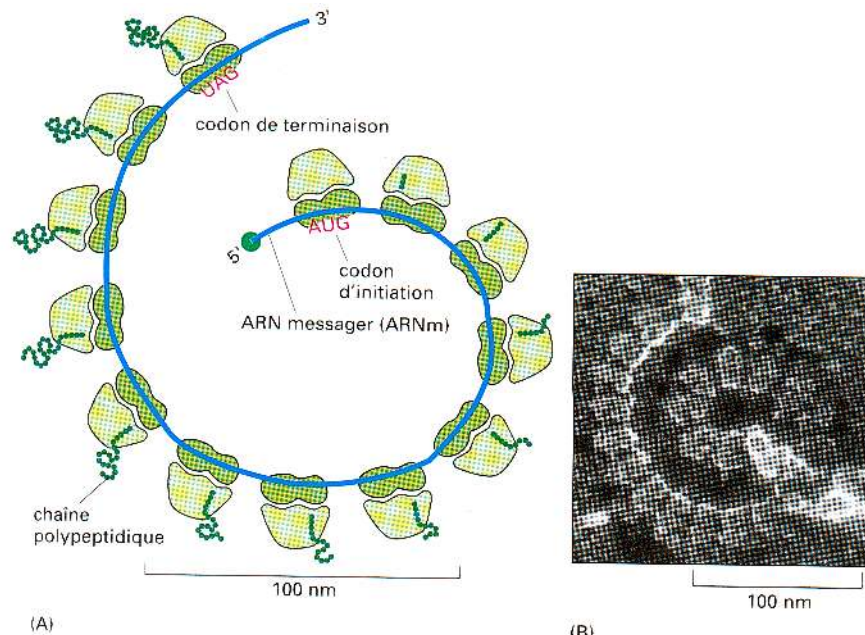
- The first Met codon in a favourable environment is not necessarily the initiator Met used by the cell
- This is more prevalent in prokaryotes than in eukaryotes



- In eukaryotes initiation occurs near 5' capped end (normal situation) or at internal ribosome entry sites (much less frequently)
- The ribosome slides along the mRNA until an acceptable AUG codon is reached, usually within about 100 nucleotides

Polyribosomes (polysomes) (prokaryotes and eukaryotes)

During protein synthesis: from 20 seconds and several minutes: multiple initiations



~80 nucleotides between 2 ribosomes

Eukaryotes: 10 ribosomes per mRNA

Prokaryotes: until 300 ribosomes per mRNA

Termination

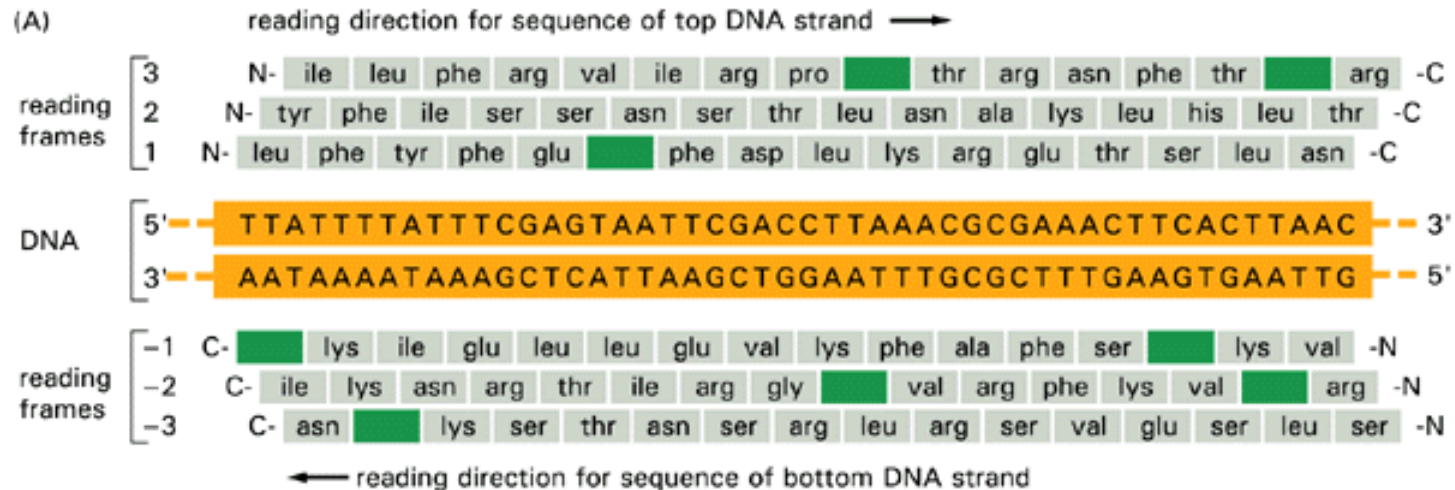
- UAG, UAA and UGA are STOP codons
- As there are no tRNAs binding to ribosomes, they simply fall off here
- In random sequence with 50% GC, the STOP codons are present on average every 21 nucleotides, in 70% GC, it is every 52 nucleotides
- STOP codons are statistically very underrepresented

Exceptions (stop codons that do not stop)

- tRNA suppressors
- The genetic code in certain organelles
- Selenocysteine and pyrrolysine

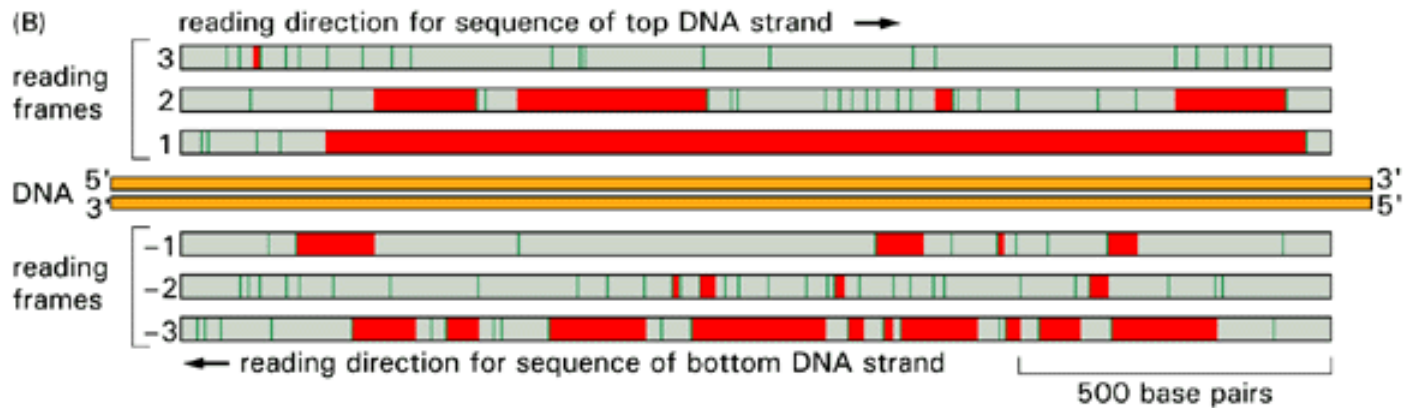
Reading frames 1

- A piece of doublestranded DNA can be translated in 6 reading frames, 3 in phase (direct or forward) and 3 in reversed phase (complementary or inverted)
 - This is a **conceptual translation**

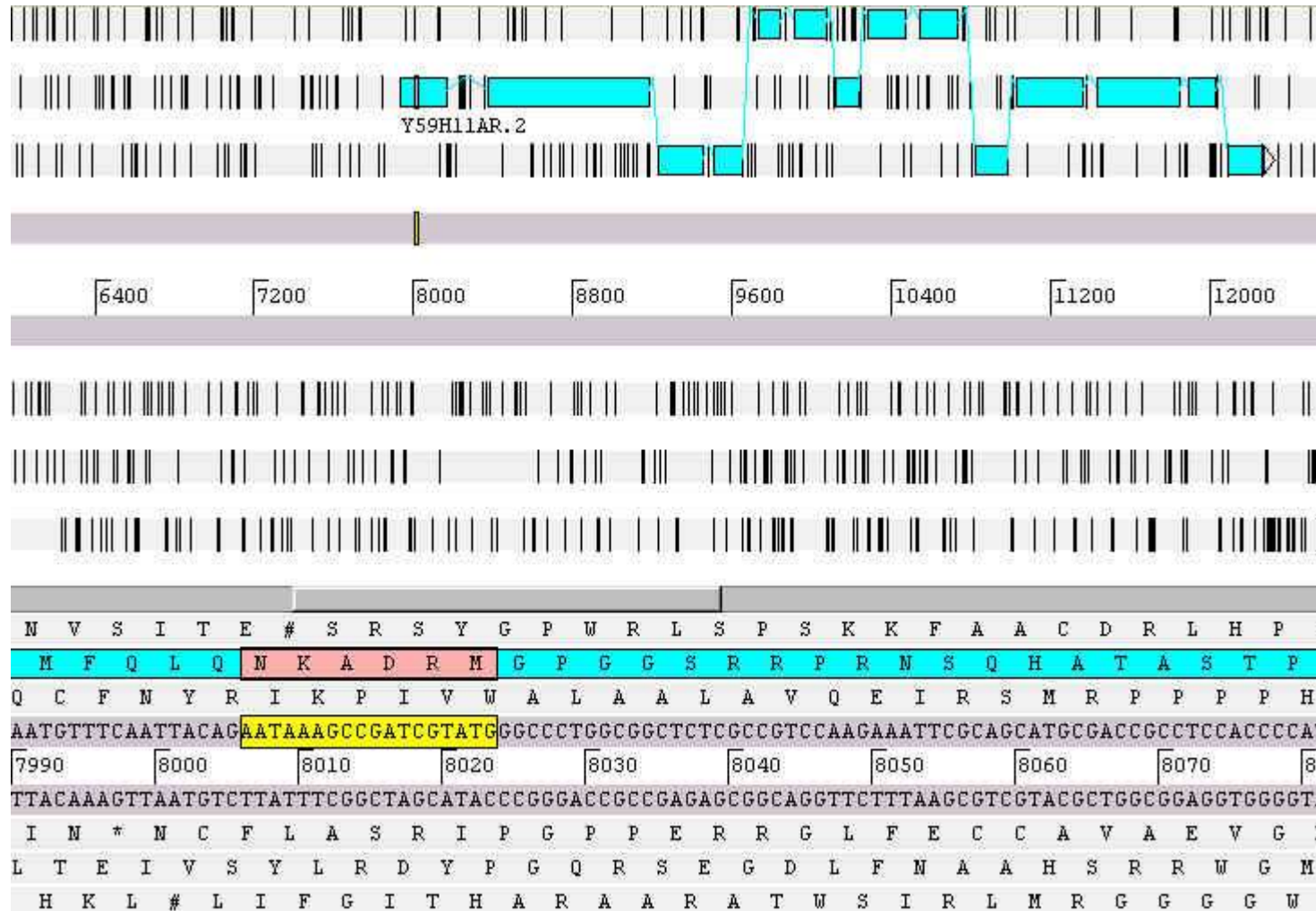


Reading frames 2

- Normally only one reading frame at a time codes for a functional protein
- In the cell, the reading frame is determined at the moment of translational initiation and it is maintained by the translation mechanism until the termination signal (STOP codon).



Reading frames of a eukaryotic gene



Reading frame, some definitions

- ORF: Open Reading Frame
 - A series of codons, including the initiation codon and the STOP codon, which codes for a potential or known protein (theoretical minimum: 21 aa)
- CDS: CoDing Sequence, Coding DNA Segment
 - DNA or RNA region of which the nucleotide sequence determines the amino acid sequence of a protein
- **All CDS's are ORFs, but all ORFs are not CDS's!**
- Frameshift: a change of reading frame caused by an insertion or a deletion of 1 or 2 nucleotides (indel)



How do you identify the correct reading frame and the CDS ?

GENE PREDICTION

Summary

- A gene can code for both RNA and proteins
- Prokaryotes have 90% coding and 10% non-coding, no introns and genes can be polycistronic
- Mammals have <10% coding and >90% non-coding, many introns, splice variants and are monocistronic
 - Introns are the biggest challenge for gene finders
 - Finding the correct ATG follows closely after

Summary 2

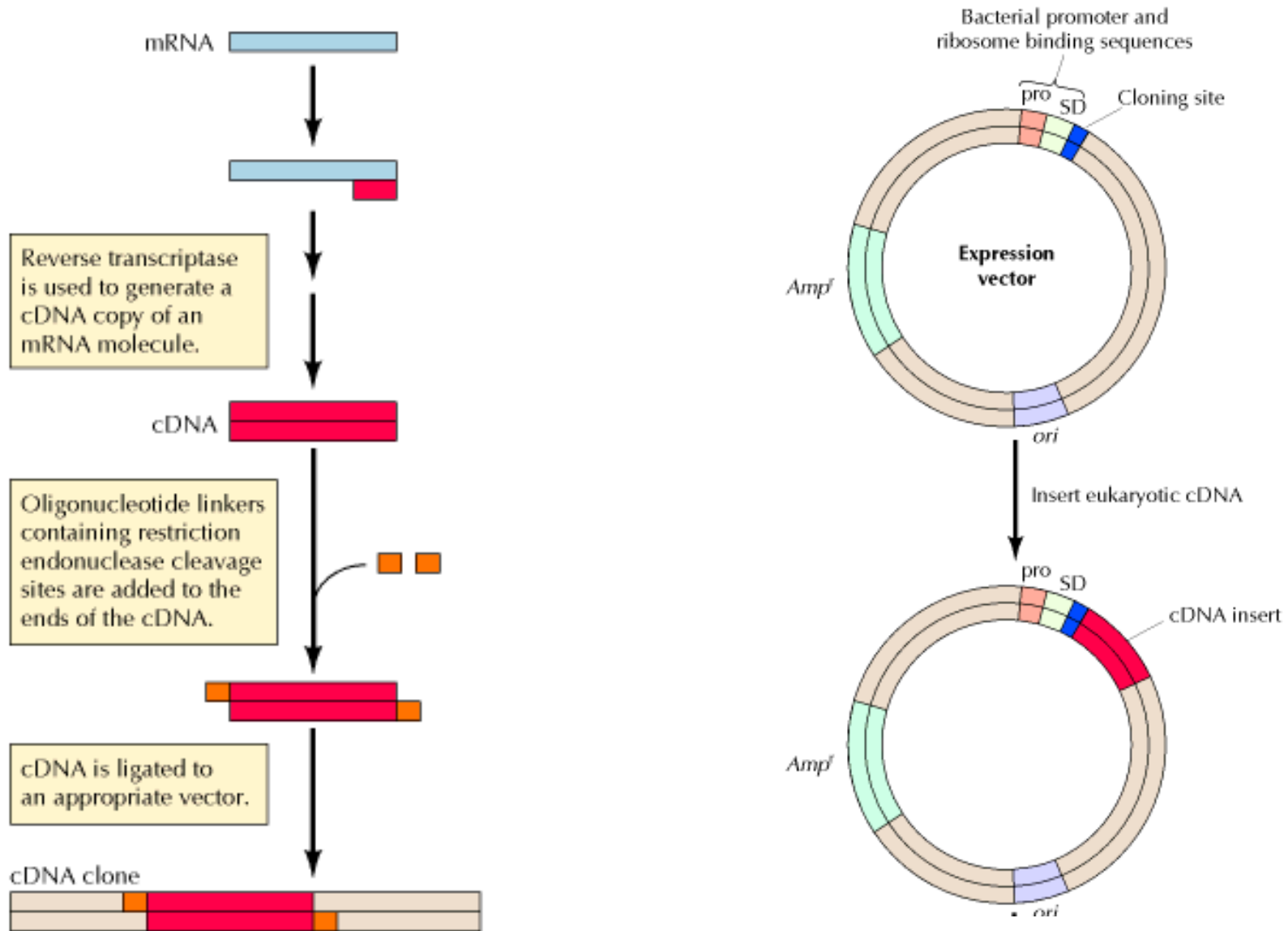
A protein coding gene is composed of

- The ORF **Crucial for gene finding**
- Promotor region,
 - Enhancers, repressors **Very little conservation, not used**
 - Shine Dalgarno (prokaryotes) **Used**
 - CpG islands (eukaryotes)
 - TATA and CAAT box **Used**
 - Terminator, Poly A signal
- ATG **Very important**
- Splice site signals **Conserved, but very short sequences**
 - Donor and acceptor sites, branch points
- TAG, TGA, TAA **Of course**

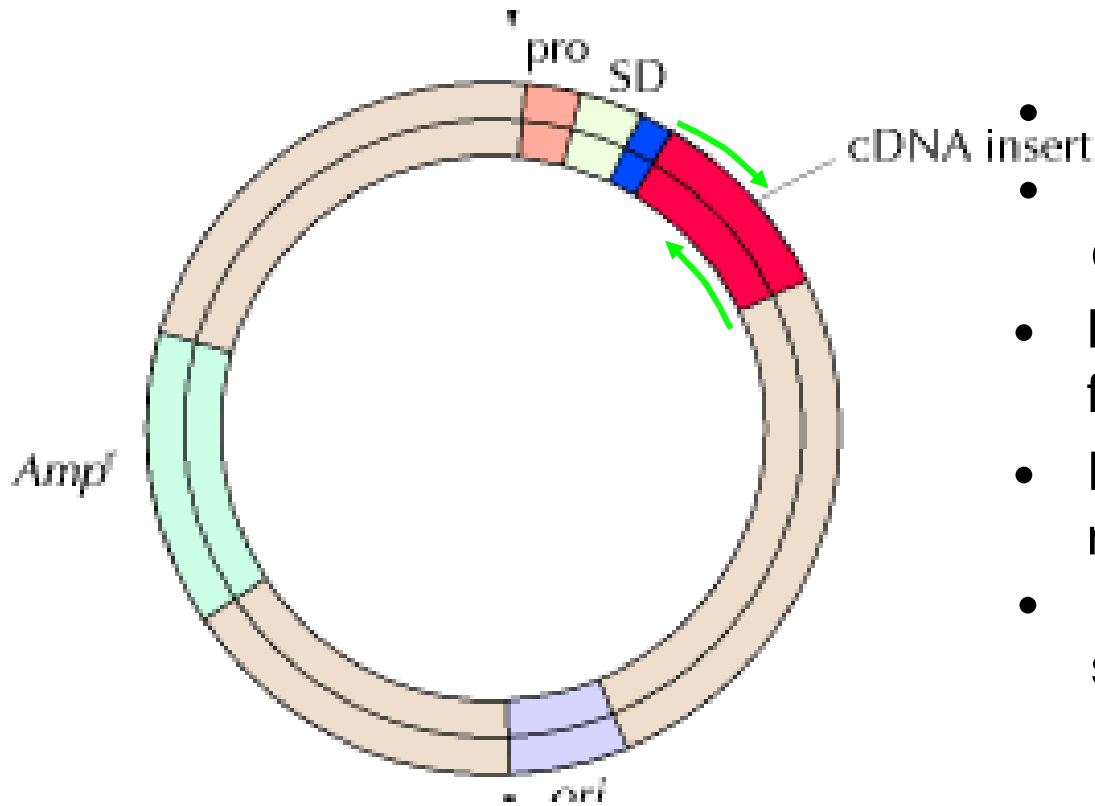
Other gene properties

- Codon frequencies **Used especially in early methods**
- Hexanucleotide frequencies **Idem**
- Conservation between species **Used in some methods, more and more now**

Making cDNA libraries



Making ESTs



- Produce cDNA library
- Pick as many (ideally all) clones as possible
- Make a single sequence run from the 3' end (polyA primer)
- Read as long as possible with no manual verification
- Ideally also make a single sequence run from the 5' end

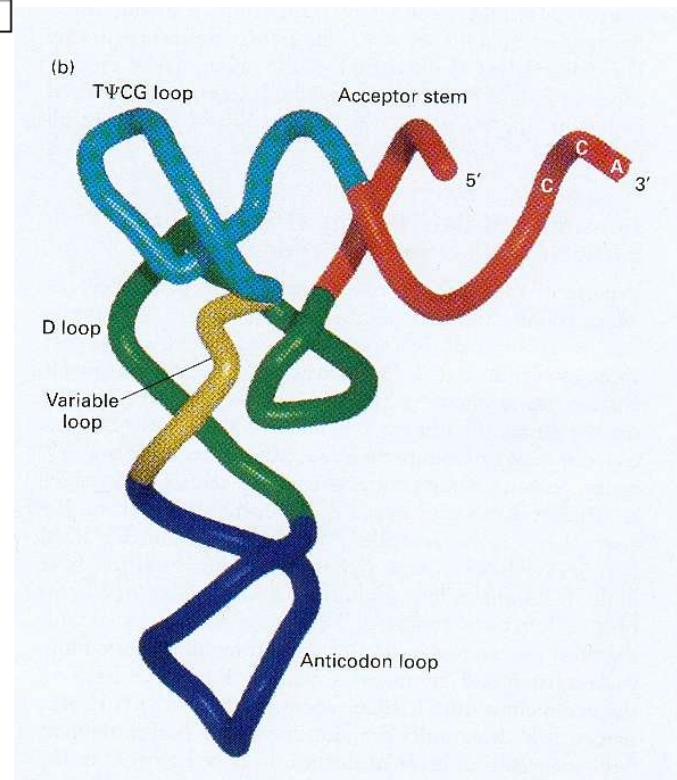
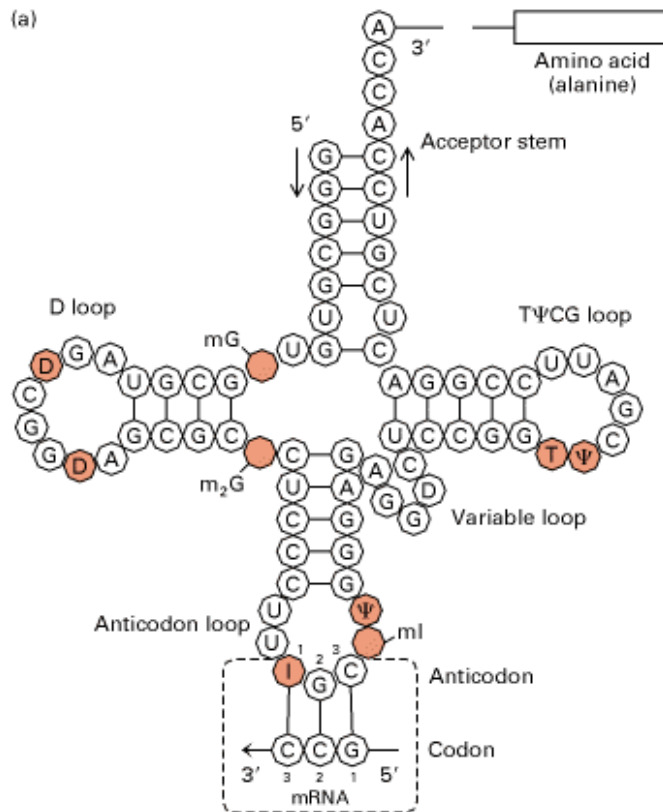
More on ESTs

- ESTs are short (2-500 bp) sequences
- ESTs are used as a tag of expressed genes
- Ideally ESTs are used to find both 3' AND 5' extremities of genes
- Clones used for ESTs are kept and can be ordered to allow later sequencing and of the entire insert
- EST clones are ideally full length, but this is FAR FROM always the case
- EST give information on the structure of the gene and alternative splicing
- BUT...
 - only one sequence run per sample => many errors including indels
 - deficit of weakly expressed genes
 - only genes with polyA are picked up
 - many partial clones
 - contaminations - artefacts

Gene prediction

Transfer RNAs (tRNA)

- There are ~ 31 different tRNAs; each composed of 75 - 95 nucleotides;
- Many million copies in the cytoplasm; they are different in the mitochondria and in the chloroplasts (different genetic codes);
- Indispensable in protein synthesis;



Prediction of tRNA genes in genomic sequences

- tRNA have conserved sequence elements;
- Programs use a combination of patterns searches; probabilistic methods and (for eukaryotes) search for Pol III promoters;
- Current programs (tRNAscan) find 99% of true tRNAs with a false positive rate of less than one per 15 billion nucleotides of random sequence.

Detection of genes in prokaryotic genomes

- In the era of massive prokaryotic genome sequencing, the computer has become a major tool in the analysis of the information content of DNA
- Today we have sequences of more than 125 entire Eubacterial (109) and Archaeal (16) genomes
- For a large majority of these genomes, the genes have only been detected by informatic tools

The first developments in the 1980s

- Fickett, Gribskov, Staden (1,2,3), demonstrate the intrinsic properties of sequences, independent of transcription or translation signals (1982-84)
- Demonstration by Borodovsky et al. of the utility of Markov chains for sequence analysis and gene prediction (1986). (4)
- Description of GeneMark in 1993 by Borodovsky.(5)
- The first tests of gene finding algorithms on the genome of *E. coli* (unfinished at the time). (6, 7)

Approaches

- **Intrinsic properties**

- Codon usage for the organism concerned
- Amino acid preferences
- Base position frequencies: the frequencies for each of the 4 nucleotides in each of the 3 codon positions
- G+C content in the organism concerned
- transcription signals (RBS)

- **Extrinsic properties**

- Homology: similarity to known genes present in a database

Problems

- The exact prediction of start sites: often the start site corresponding to the longest possible ORF is chosen
- Alternative start codons: UUG, GUG, CUG
- Overlap between the genes: 1 base TGATG (TGA=stop; ATG=met) or 4 bases ATGA
- In the beginning the prediction programmes did not take overlap into account. But the new versions do

GeneMark

<http://opal.biology.gatech.edu/GeneMark/>

- Developed from 1986 by Borodovsky
- Published in 1993. (5)
- Used in 1993 to analyse 176 kb of *E. coli* genome (7)
- Central for the method: the coding and non-coding regions of a DNA sequence are treated as subsequences which follows different rules of "nucleotide ordering"

How does it work? 1/2

- Combines specific Markov models for coding and non-coding regions, the use of which is decided by a Bayesian decision function
- The algorithm is capable of distinguishing between 3 types of DNA sequence:
 - 3) Sequence coding for a protein in the 3 forward reading frames (states 1-3)
 - 4) Sequence coding for a protein in the 3 reverse reading frames (states 4-6)
 - 5) Non-coding sequence (state 7)

How does it work? 2/2

- Given a sequence S , the probability is calculated for S either coding for a protein in each of the 6 reading frames or being not coding (P_i , $i=1\dots7$)
- If one of these 7 probabilities is > 0.5 , S is considered belonging to the state in question.

Improvements to GeneMark

1/2

- An improved version, GeneMark.hmm (8), allows a more precise definition of the borders of the gene
- A probabilistic model to find ribosomal binding sites (RBS) have been developed from 325 genes with known RBS, leading to the definition of the consensus sequence AGGAG, which is complementary to a pentamer situated in the far 3' end of 16S rRNA (*E. coli*)

Improvements to GeneMark

2/2

- GeneMark.hmm 2.0 uses as supplementary parameter the length of the "spacer", the "spacer" being the sequence between the last nucleotide of the RBS and the first nucleotide of the gene
- GeneMark.hmm 2.0 can predict genes with any length overlap
- GeneMarkS (9), an even more improved version...

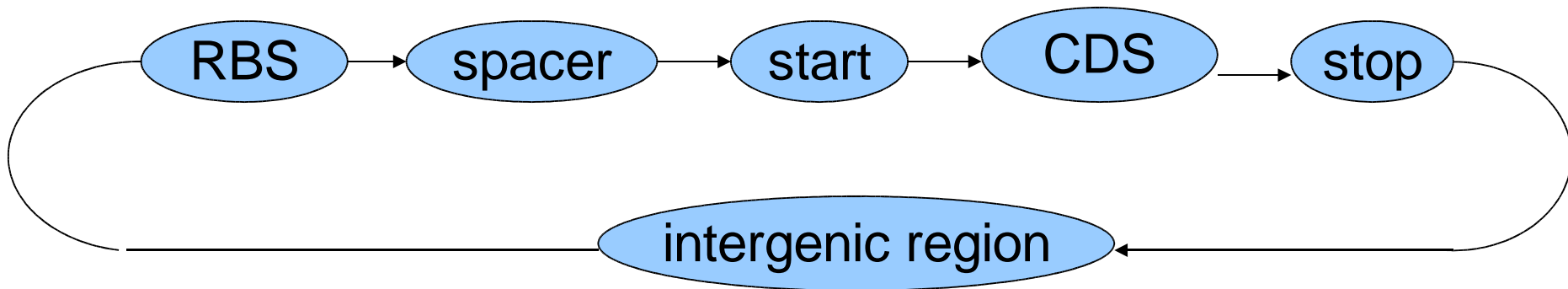


diagram of the GeneMark.hmm procedure

Glimmer 1/2

Gene Locator and Interpolated Markov Modeler

<http://www.tigr.org/softlab/glimmer/glimmer.html>

- Developed in 1998 by TIGR ⁽¹⁰⁾, specifically for finding genes in Eubacterial and Archaeal genomes
- Makes use of IMM (Interpolated Markov Models), a generalisation of Markov chains, for the identification of the coding regions and to distinguish them from the non-coding regions
- Used to analyse the complete genomes from *H. pylori*, *B. burgdorferi*, *T. pallidum*, *C. trachomatis*, etc. (all the genomes sequenced by TIGR)
- Glimmer 2.0 is better at coping with overlapping genes than version 1.0 ⁽¹¹⁾.

Glimmer 2/2

- By default, Glimmer defines the start site of the gene so the genes becomes as long as possible
- An option allows the calculation of the hybridisation maximum between 16S rRNA and a given region situated upstreams of a potential start codon.

The case of G+C rich genomes

1/2

For ex. *Rhizobium meliloti*, *Mycobacterium tuberculosis*

Why is this a problem?

The stop codons (UAA, UAG, UGA) are GC poor

- the average distance between stop codons increases which results in an increasing number of long random open reading frames
- but this does not necessarily mean that these ORFs are coding

The third codon position is normally GC rich

- if the 3rd base in a codon in reading frame 1 is a G or C, and this nucleotide corresponds to a base in the 1st position on the opposite strand, this 1st base will also be a G or C. One will thus never find stop codons in this reading frame as these latter all start with U!

The case of G+C rich genomes

1/2

- So not only will the programmes detect the long ORFs in one reading frame but they will also find (long) ORFs in other reading frames! Superposition of ORFs.

FrameD

<http://www.toulouse.inra.fr/FrameD.html>

- FrameD (12) reutilises the same input information used by Glimmer, giving more weight to the Interpolated Markov Models (IMMs)
- FrameD allows the detection of frame shifts
- Its graphical output can include homology information to proteins found in protein databases
- A test on *B. subtilis*, a GC poor organism produces the same result as Glimmer
- It is currently impossible to detect genes overlapping on the 2 opposite strands

State of the art

- Some sequencing groups do high quality annotation work: TIGR, Sanger Centre
- Certain groups are not devoted to high throughput genomic sequencing and, therefore, have less experience; their work may reflect this lack of experience
- For others we ask ourselves: but what on earth did they do to get that result??

References 1

1. Fickett J.W. (1982) Nucl. Acids Res. 10, 5303-5318.
2. Gribskov M., Devereux J. and Burgess R.R. (1984) Nucl. Acids Res. 12, 539-549.
3. Staden R. (1984) Nucl. Acids Res. 12, 551-567.
4. Borodovsky M. et al. (1986), Molecular Biology 20, 826-833.
5. Borodovsky M. and McIninch J., (1993) Comput. Chem. 17 (2), 123-133.
6. Borodovsky M., Rudd K.E., Koonin E.V. (1994) Nucl. Acids Res. 22, 4756-4767
7. Blattner F.R. et al. (1993) Nucl. Acids Res. 21,5408-17.

References 2

8. Lukashin A.V. and Borodovsky M. (1998) Nucl. Acids Res. 26, 1107-1115.
9. Besemer J., Lomsadze A. and Borodovsky M. (2001) Nucl. Acids Res. 29, 2607-2618.
10. Salzberg S.L., Delcher A. L., Kasif S. and White O. (1998) Nucl. Acids Res. 26, 544-548.
11. Delcher A.L., Harmon D., Kasif S., White O. and Salzberg S.L. (1999) Nucl. Acids Res. 27, 4636-4641.
12. Schiex T., Thébault P. and Kahn D., in JOBIM'2000 conferences Proceedings pp 321-328.

Prediction of genes in eukaryotic genomes

Difficulties of the task

Increasing complexity of the genomes parallel with that of the organisms

Gene scarcity

Intron-exon

Alternative splicing

The possibility of intron encoded genes

Pseudo-genes

Finding genes in the human genome

Example: chromosome 6

- 166,880,988 bp (6% of genome)
- 2190 gene structures found
 - 772 known (mean 9.95 exons per gene)
 - 500 novel CDS and transcripts (mean 5.81 and 3.81 exons per gene)
 - 285 putative genes (mean 2.86 exons per gene)
 - 633 pseudogenes
- 2.5 splice variants per known or novel gene (splice variants for ~40% of genes)
- Genes cover 42% of sequence
 - mean size 31195 bp
- Exons cover 2.2% of sequence
 - mean size 281 bp
- There are 9.2 genes per Mb of sequence
- Repeats cover 43.95%

Finding genes in the human genome

More statistics from chromosome 6

- MHC cluster 43 genes pr Mb
- Longest coding exon: 9114 bp
- Most exons: 101
- Longest intron (first in gene): 479 kb
- Largest gene: 1.4 Mb, 12 exons
- Most splice variants: 16 (genes with more than 1000 variants are known, neurexins)

Gene prediction approaches

3 different approaches are used:

ab initio

prediction based on sequence similarity

comparative genomics

ab initio prediction 1/2

- Methods based on rules, signal detection or statistical models
- *ab initio* methods can find entirely new genes, which do not resemble any known sequence or domain
- methods are often species specific

ab initio prediction 2/2

1) Rules or signals

- Promoters (short motifs TATA, CAAT, GC)
- Consensus for cleavage for the addition of polyA
- Start and stop codons
- Splice donor and acceptor sites, branching point

2) Statisitc models

- Codon bias for synonymous codons
 - According to organism
 - According to isochore (%GC)

Prediction based on sequence similarity

- Utilisation of ESTs: alignment of ESTs with the genomic sequence obeying the rules of splicing
- Utilisation of protein sequences: comparison of the genomic sequence with the protein sequence databases
- Utilisation of profiles and HMMs: comparison of the genomic sequence with fingerprints of protein domains following a defined model of splicing

Comparative genomics

- Phylogenetic approach: comparison of two genomes diverged during evolution
- Ideal divergence time: 300 million years (human-bird)
- Sequence conservation => selective pressure => function
- Allows the detection of all functional elements (transcribed or non-transcribed gene, regulatory element...)
- But...
 - Conserved function, but which?
 - The approach is dependent on the sequencing status for the other genomes

Prediction programmes 1/2

- FGENES (Gene-Finder): discriminatory analysis and optimisation by dynamic programming
- GeneID: hierarchical selection by filtering (GeneID+ implements sequence homologies)
- GRAIL (Gap3): neural network and dynamic programming
- GeneBuilder (GeneView): utilises sequence homologies
- GeneParser: optimisation by neural network and dynamic programming
- GenLang: based on a formal grammar
- Xpound: based on a probabilistic mathematical model

Prediction programmes 2/2

- Genscan: generalised HMM
- GeneMark.hmm: HMM and dynamic programming
- Genie: generalised HMM
- HMMgene: HMM with "conditional maximum of likelihood"
- Morgan: decision trees, dynamic programming and Markov chains
- MZEF: discriminatory analysis
- GeneWise: HMM and sequence homology information
- Procrustes: method based sequence homology information
- Twinscan: method based on comparative genomics

Evaluation of gene finding methods

- A test set of 195 genomic mammal sequences containing known gene structures were built
- They were all newer than the programmes, i.e. the sequences had not been used to train these
- All sequences including uncanonical features, atypical start codon or splice site dinucleotides were excluded
- When testing the 7 programmes only features predicted on the forward strands were taken into account

Sensitivity and specificity

- We define **sensitivity** as the proportion of coding nucleotides that are correctly predicted as coding:
$$S_n = TP / (TP + FN)$$
- and **specificity** as the proportion of nucleotides predicted as coding that are actually coding:
$$S_p = TP / (TP + FP)$$
- TP=True Positive, TN=True Negative,
FP=False Positive, FN=False Negative

Prediction of nucleotides and exons

Table 1. Nucleotide and Exon Level Accuracy

Programs	No. of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	Sp	AC	CC	ESn	ESp	(ESn+ESp)/2	ME	WE	PCa	PCp	OL
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark .hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
GeneScan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.32	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69	0.46	0.41	0.43 ± 0.28	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

For each sequence in the HMR195 dataset, the exons predicted on the forward (+) strand were compared to the annotated exons. The standard measures of predictive accuracy on nucleotide and exon level were calculated for each sequence and averaged over all sequences for which they were defined. This was done separately for each of the programs tested.

(No. of sequences) number of sequences effectively analyzed by each program; in parentheses is the number of sequences where the absence of gene was predicted; (Sn) nucleotide level sensitivity; (Sp) nucleotide level specificity; (AC) approximate correlation; (CC) correlation coefficient; (ESn) exon level sensitivity; (ESp) exon level specificity; (ME) missed exons; (WE) wrong exons; (PCa) proportion of real exons that were partially predicted (only one exon boundary correct); (PCp) proportion of predicted exons that were only partially correct; (OL) proportion of predicted exons that overlap an actual exon. AC and (ESn+ESp)/2 are given with standard deviation.

Prediction of signals

Table 2. Accuracy versus Signal Type

Programs	Signal type			
	start codon (195)	acceptor site (753)	donor site (753)	stop codon (195)
FGENES	0.67 (0.63)	0.80 (0.77)	0.85 (0.82)	0.75 (0.72)
GeneMark.hmm	0.46 (0.60)	0.81 (0.75)	0.82 (0.78)	0.57 (0.64)
Genie	0.56 (0.57)	0.77 (0.82)	0.78 (0.83)	0.72 (0.73)
Genscan	0.61 (0.78)	0.87 (0.80)	0.90 (0.84)	0.76 (0.86)
HMMgene	0.75 (0.78)	0.81 (0.85)	0.83 (0.87)	0.78 (0.81)
Morgan	0.43 (0.43)	0.66 (0.57)	0.65 (0.56)	0.39 (0.39)
MZEF	—	0.59 (0.65)	0.66 (0.73)	—

For each program, the proportion of actual signals identified correctly (the upper number) and the proportion of predicted signals that are correct (the lower number) are averaged over all signals belonging to a particular type. The number in parenthesis in the header of each column represents the number of signals of each type in the HMR195 dataset.

Prediction of nucleotides and exons in DNA with different C+G content

Table 3. Accuracy versus G + C Content

C + G content	<40%(14)		40-50%(69)		50-60%(93)		>60%(19)	
	AC	(Esn+Esp)/2	AC	(Esn+Esp)/2	AC	(Esn+Esp)/2	AC	(Esn+Esp)/2
FGENES	0.84	0.70	0.81	0.64	0.85	0.71	0.87	0.66
GeneMark.hmm	0.79	0.48	0.80	0.46	0.87	0.62	0.85	0.48
Genie	0.85	0.69	0.85	0.60	0.92	0.75	0.87	0.79
Genscan	0.94	0.80	0.91	0.66	0.91	0.74	0.88	0.70
HMMgene	0.91	0.76	0.90	0.73	0.92	0.79	0.91	0.77
Morgan	0.65	0.29	0.72	0.49	0.69	0.43	0.69	0.37
MZEF	0.66	0.71	0.65	0.50	0.70	0.62	0.58	0.53

The HMR195 dataset was partitioned according to the G + C% content of the sequences. The number in parenthesis in the header of each column represents the number of sequences belonging to each partition. For each program, AC and (ESn+Esp)/2 are averaged over all sequences belonging to the particular partition for which they are defined.

Prediction of exons with different length

Table 4. Accuracy versus Exon Length

Programs	Length range of exons in bp						
	0-24 (22)	25-49 (49)	50-74 (91)	75-99 (130)	100-199 (440)	200-299 (91)	300+ (125)
FGENES	0.45 (0.33)	0.55 (0.42)	0.71 (0.64)	0.80 (0.75)	0.80 (0.81)	0.71 (0.61)	0.59 (0.66)
GeneMark.hmm	0.05 (0.12)	0.39 (0.51)	0.60 (0.58)	0.77 (0.72)	0.75 (0.73)	0.67 (0.62)	0.46 (0.45)
Genie	0.27 (0.18)	0.53 (0.47)	0.60 (0.66)	0.80 (0.81)	0.70 (0.83)	0.71 (0.68)	0.69 (0.69)
Genscan	0.18 (0.29)	0.45 (0.81)	0.68 (0.79)	0.89 (0.85)	0.84 (0.76)	0.87 (0.71)	0.66 (0.65)
HMMgene	0.23 (0.42)	0.59 (0.76)	0.64 (0.75)	0.79 (0.77)	0.80 (0.85)	0.78 (0.72)	0.77 (0.74)
Morgan	0.30 (0.14)	0.37 (0.14)	0.38 (0.31)	0.61 (0.57)	0.51 (0.57)	0.51 (0.41)	0.42 (0.35)
MZEF	0.00 (0.00)	0.16 (0.44)	0.32 (0.45)	0.40 (0.58)	0.49 (0.73)	0.45 (0.53)	0.12 (0.26)

The HMR195 dataset was partitioned according to the length of the annotated exons. The number in parenthesis in the header of each column represents the number of actual exons belonging to each partition. For each program, CRa (the proportion of real exons that are correctly predicted [the upper number]) and CRp (the proportion of predicted exons that are correct [the number in parentheses]) are averaged over all sequences belonging to that particular partition.

Prediction of different exon types

Table 5. Accuracy versus Exon Type

Programs	Exon type			
	Initial (152)	Internal (601)	terminal (152)	single (43)
FGENES	0.64 (0.55)	0.79 (0.78)	0.66 (0.58)	0.58 (0.83)
GeneMark.hmm	0.40 (0.48)	0.78 (0.72)	0.52 (0.51)	0.30 (0.65)
Genie	0.49 (0.45)	0.76 (0.82)	0.61 (0.57)	0.70 (0.68)
Genscan	0.57 (0.71)	0.87 (0.76)	0.67 (0.73)	0.63 (0.83)
HMMgene	0.68 (0.72)	0.78 (0.83)	0.70 (0.73)	0.77 (0.79)
Morgan	0.35 (0.35)	0.55 (0.46)	0.36 (0.36)	—
MZEF	—	—	—	—

The HMR195 dataset was partitioned according to the type of the annotated exons. The number in parenthesis in the header of each column represents the number of actual exons belonging to each partition. For each program, CRa (the upper number) and CRp (the number in parentheses), are averaged over all sequences belonging to that particular partition.

Reliability of probability scores

Table 6. Accuracy versus Probability

Programs	Probability range of predicted exons				
	0.00–0.50	0.50–0.75	0.75–0.90	0.90–0.95	0.95+
Genscan	0.32 (112)	0.45 (159)	0.75 (132)	0.84 (93)	0.94 (481)
HMMgene	0.32 (91)	0.65 (173)	0.79 (136)	0.83 (96)	0.95 (406)
MZEF	—	0.43 (111)	0.54 (104)	0.64 (72)	0.74 (258)

The HMR195 dataset was partitioned according to probability of the predicted exons. For each program, CRp (proportion of predicted exons that are correct) is averaged over all sequences belonging to that particular partition. The number in parenthesis is the number of exons belonging to each partition.

Gene finding tools used in the exercises

- NetGene2
 - Neural network

<http://www.cbs.dtu.dk/services/NetGene2/>
- HMMGene
 - HMM that incorporate coding statistics

<http://www.cbs.dtu.dk/services/HMMgene/>
- Genebuilder

<http://www.itba.mi.cnr.it/webgene/>
- Genscan
 - General probabilistic model

<http://genes.mit.edu/GENSCAN.html>

References

Rogic S, Mackworth AK, Ouellette FB. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* 2001 **11**(5):817-32.

Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* 2000 **10**(10):1631-42.

Burset M, Guigo R. Evaluation of gene structure prediction programs. *Genomics.* 1996 **34**(3):353-67.

<http://linkage.rockefeller.edu/wli/gene/programs.html>