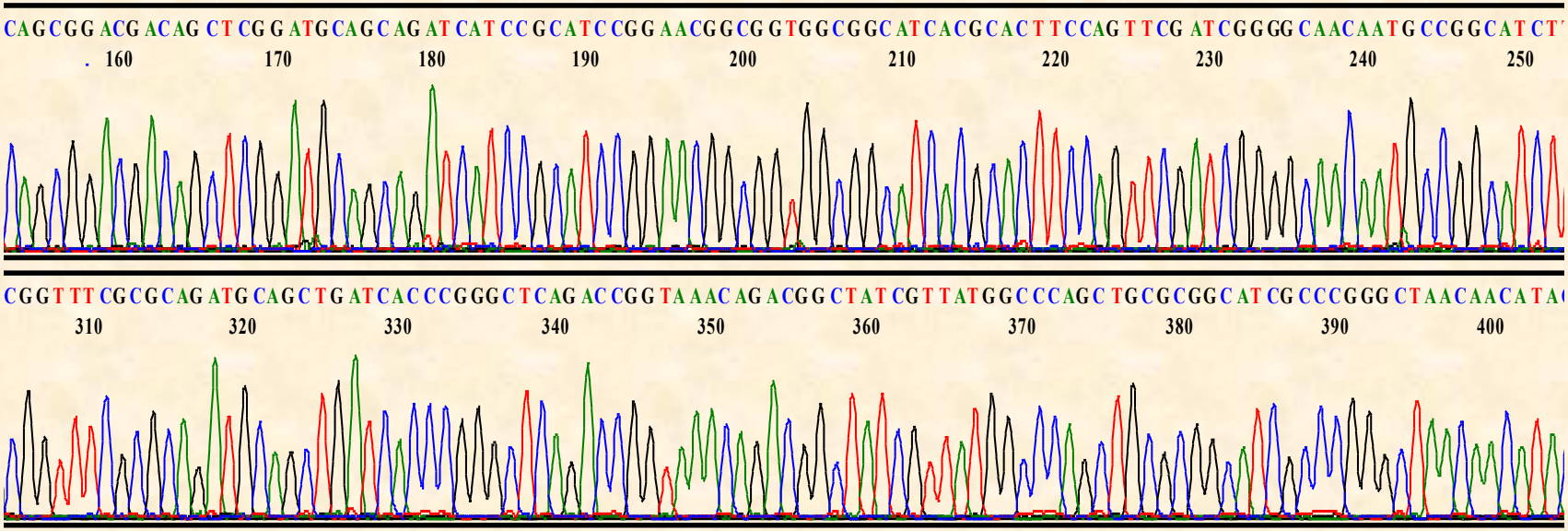
Predikce genů



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

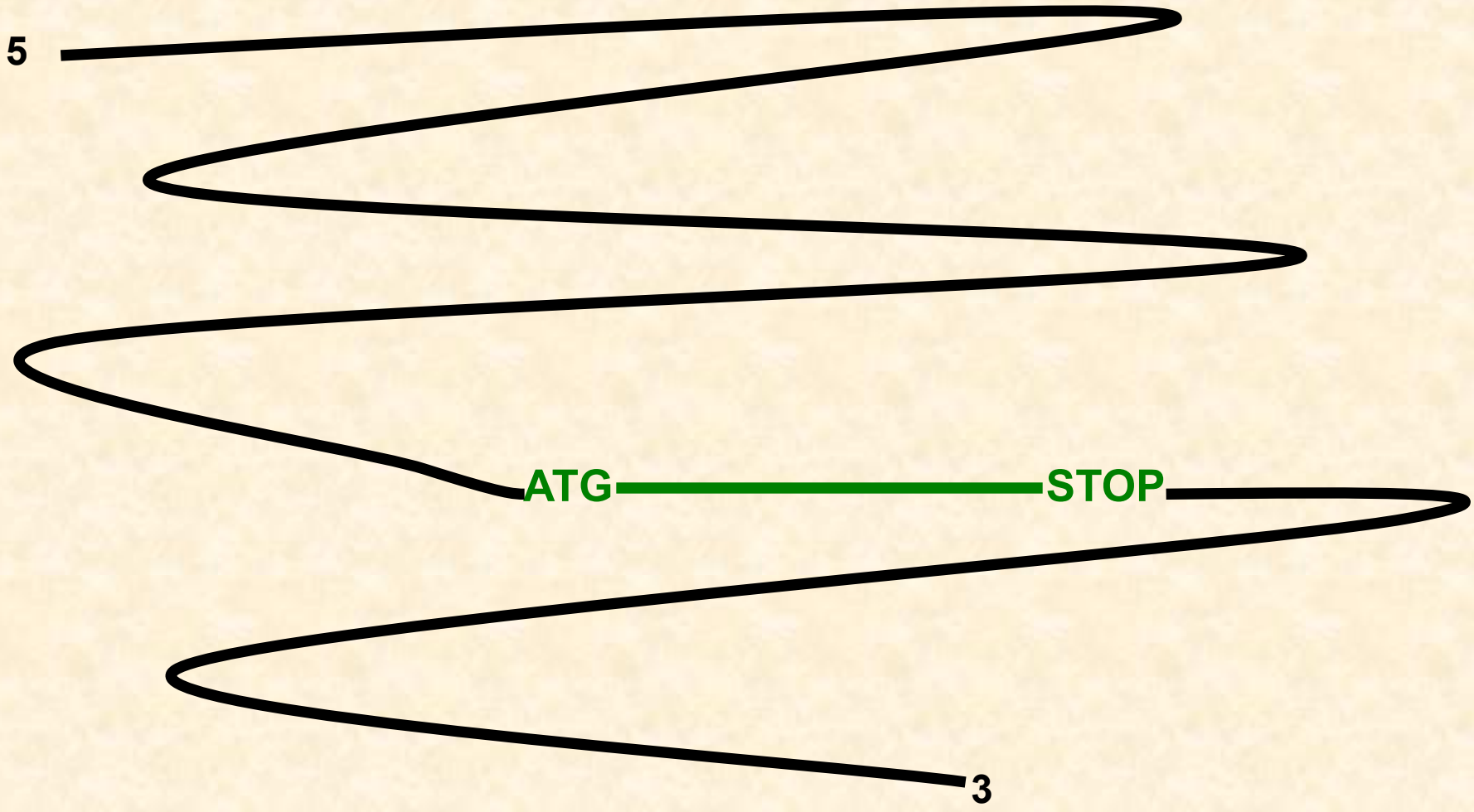
„Syrové“ sekvence DNA

Identifikace a anotace genů a proteinů

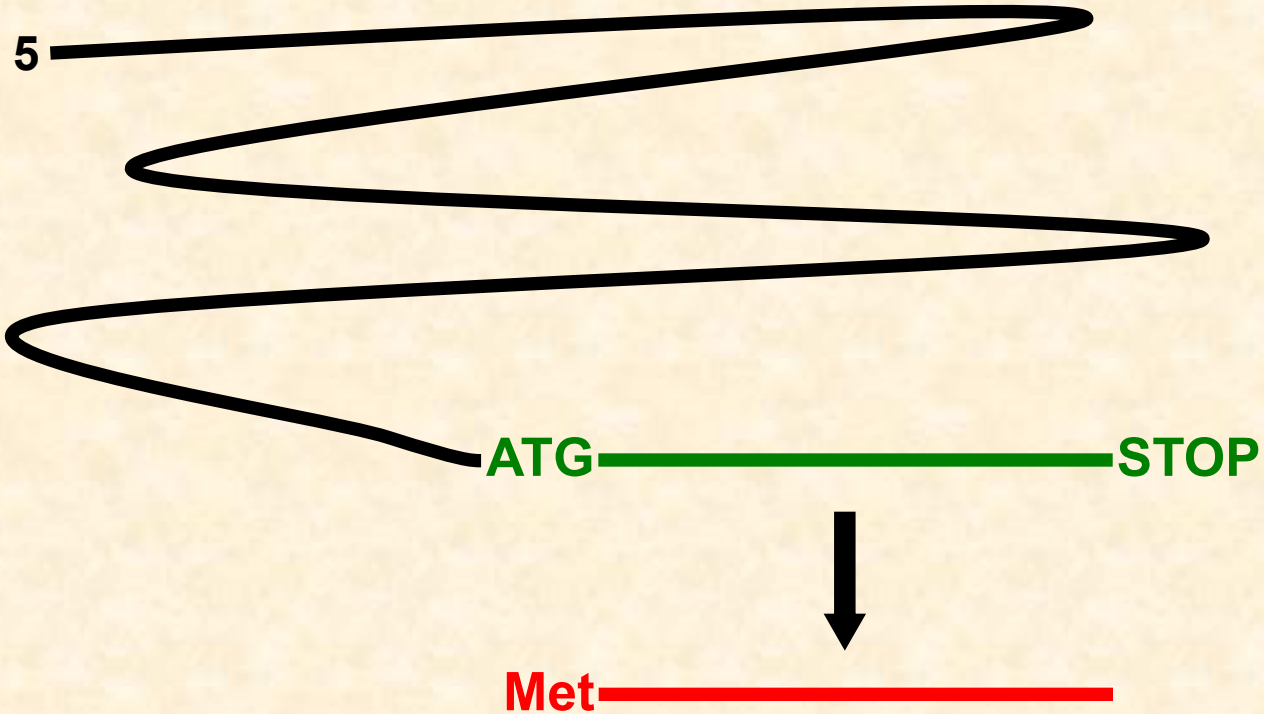
Table 1
Software commonly used for bacterial genome annotation and comparison

DNA level annotation		
GeneMark	http://exon.gatech.edu/genemark/	Protein gene prediction
Glimmer	http://www.genomics.jhu.edu/Glimmer/	Protein gene prediction
SHOW	http://genome.jouy.inra.fr/ssb/SHOW/	Protein gene prediction
tRNAscan-SE	http://lowelab.ucsc.edu/tRNAscan-SE/	tRNA gene prediction
RNAmmer	http://www.cbs.dtu.dk/services/RNAmmer/	rRNA gene prediction
RepSeek	http://www.abi.snv.jussieu.fr/%98public/RepSeek/	Search for approximate repeats in complete DNA sequences
IslandPath	http://www.pathogenomics.sfu.ca/islandpath/	Identification of genomic islands
Protein level annotation		
BLAST	http://www.ebi.ac.uk/blast/	Compare a novel sequence with those contained in nucleotide and protein databases
InterProScan	http://www.ebi.ac.uk/InterProScan/	Search for domains/motifs in the InterPro database
COGNITOR	http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html	Compare a query sequence to the COG (Cluster of Orthologous Groups of proteins) database
PRIAM	http://bioinfo.genopole-toulouse.prd.fr/priam/	Detection of enzymatic function in a fully sequenced genome, based on all sequences available in the ENZYME database
GOAnno	http://bips.u-strasbg.fr/GOAnno/	BLAST search on the Gene Ontology database
PSORTb	http://www.psort.org/psortb/	Prediction of bacterial protein subcellular localization
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Prediction of transmembrane helices in protein sequences
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Prediction of signal peptide cleavage sites in protein sequences
Comparative genomic tools		
Mauve	http://gel.ahabs.wisc.edu/mauve/	Multiple genome alignments in the presence of large-scale evolutionary events
MOSAIC	http://mig.jouy.inra.fr/mig/mig_eng/presentation/project/mosaic	Define the set of backbones and loops in closely related bacterial genomes
ACT	http://www.sanger.ac.uk/Software/ACT/	Comparative genome analysis and visualization tools for multiple genome alignments
CGAT	http://mbgd.genome.ad.jp/CGAT/	
MaGe	http://www.genoscope.cns.fr/agc/mage/	Computation of gene order conservation (syntenies) between available bacterial genomes
Pathologic	http://biocyc.org/	Metabolic network reconstruction and comparative pathway analysis
PUMA2	http://compbio.mcs.anl.gov/puma2/	Metabolic pathway reconstruction
The SEED	http://theseed.uchicago.edu/FIG/	Comparative analysis and annotation tools using the subsystem approach
STRING	http://string.embl.de/	Search Tool for the Retrieval of Interacting Proteins
PyPhy	http://www.cbs.dtu.dk/staff/thomas/pyphy/	Reconstruction of phylogenetic relationships of complete microbial genomes
HoSeqI	http://pbil.univ-lyon1.fr/software/HoSeqI/	Automatically assign sequences to homologous gene families from the HOGENOM database

Gen



Gen

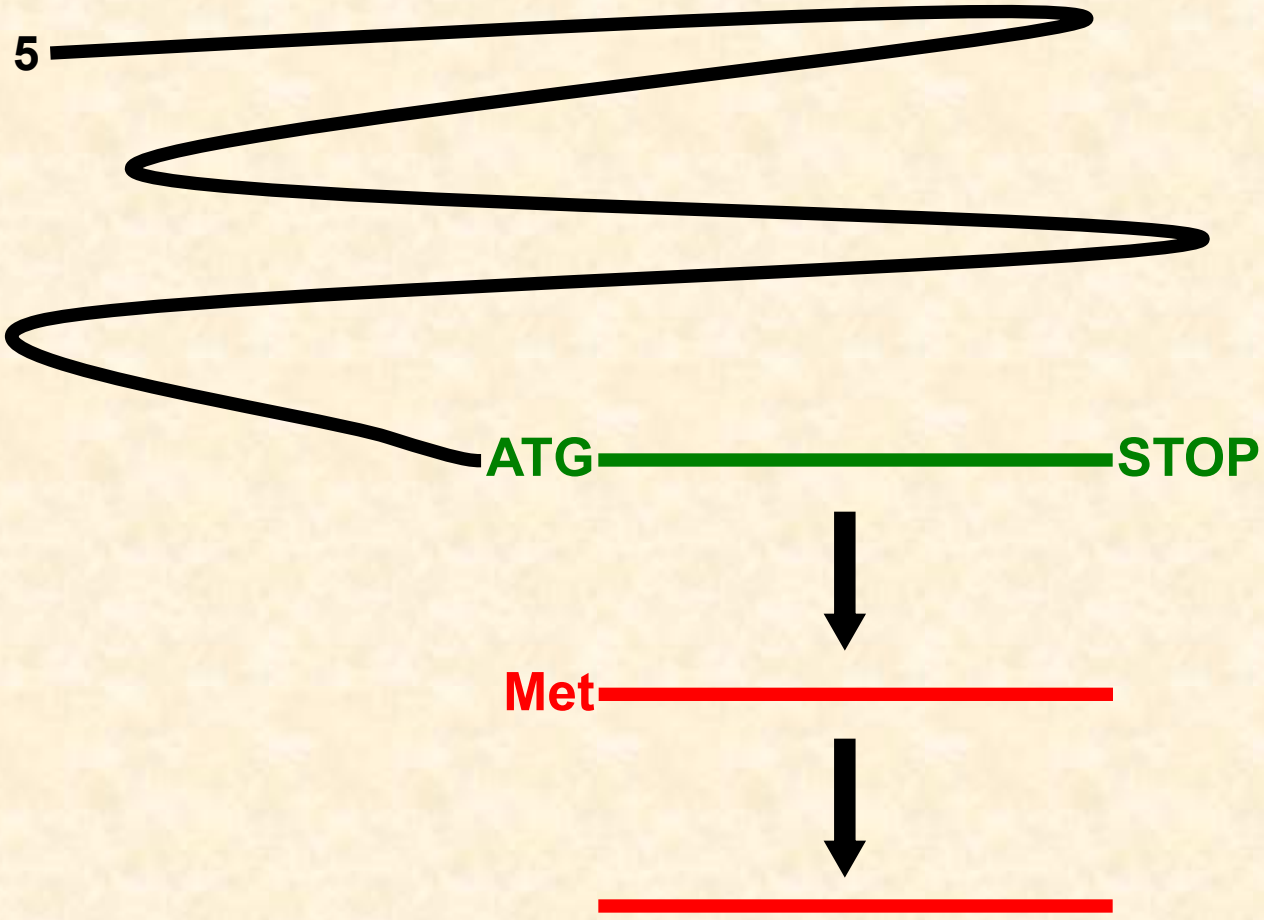


DNA

RNA

Protein

Gen



DNA

RNA

Protein

Predikce genů kódujících proteiny

- **Prokaryotické geny**

  Nepřerušované úseky DNA mezi startovním kodonem (ATG, GTG, TTG, CTG**) a stop kodonem (TAA, TGA, TAG).**

- **Úsek kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání DATABÁZÍ pomocí ALIGNMENTU).**

Překlad DNA sekvence

- **ExPASy**
  http://www.expasy.org/tools/dna.html

- **ORF Finder (NCBI)**

  http://www.ncbi.nlm.nih.gov/gorf/gorf.html

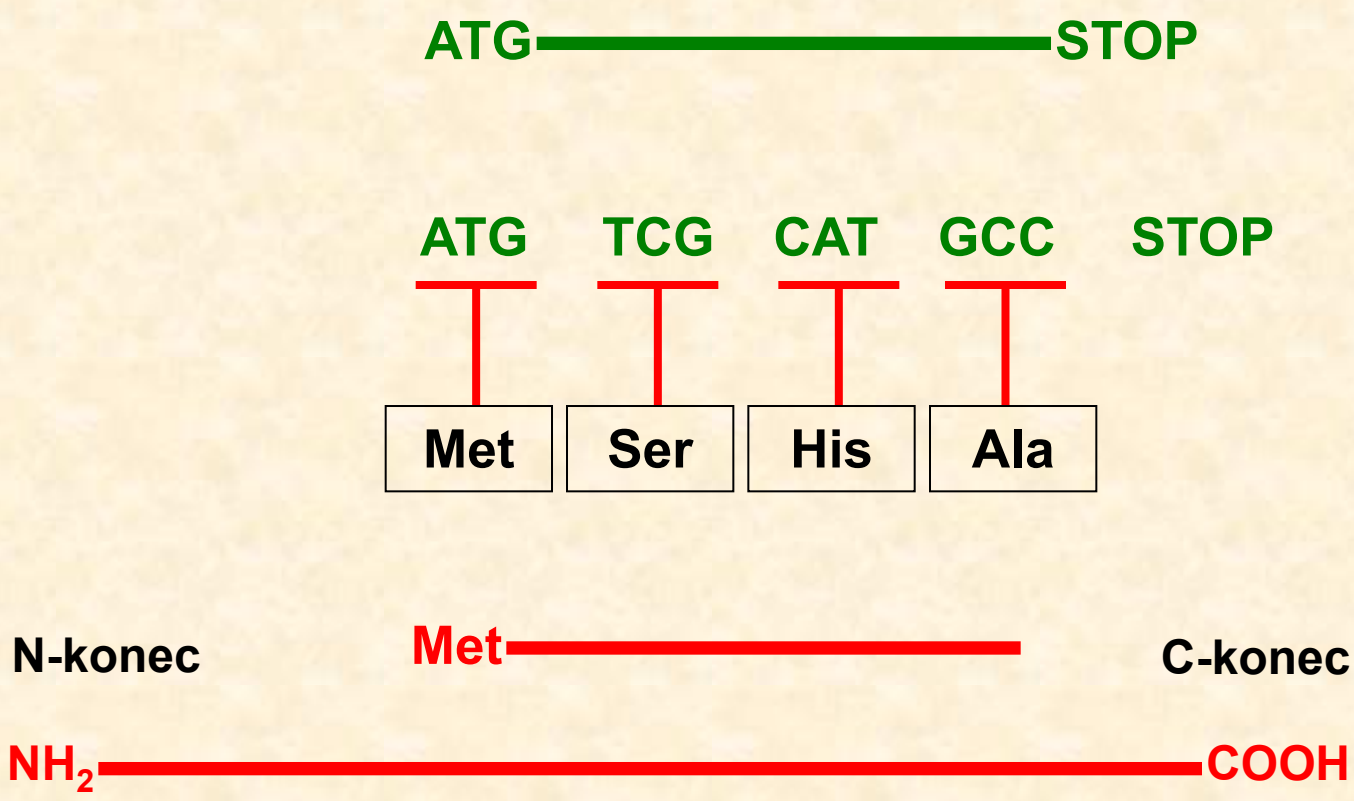
nonpolar | polar | basic | acidic | (stop codon)

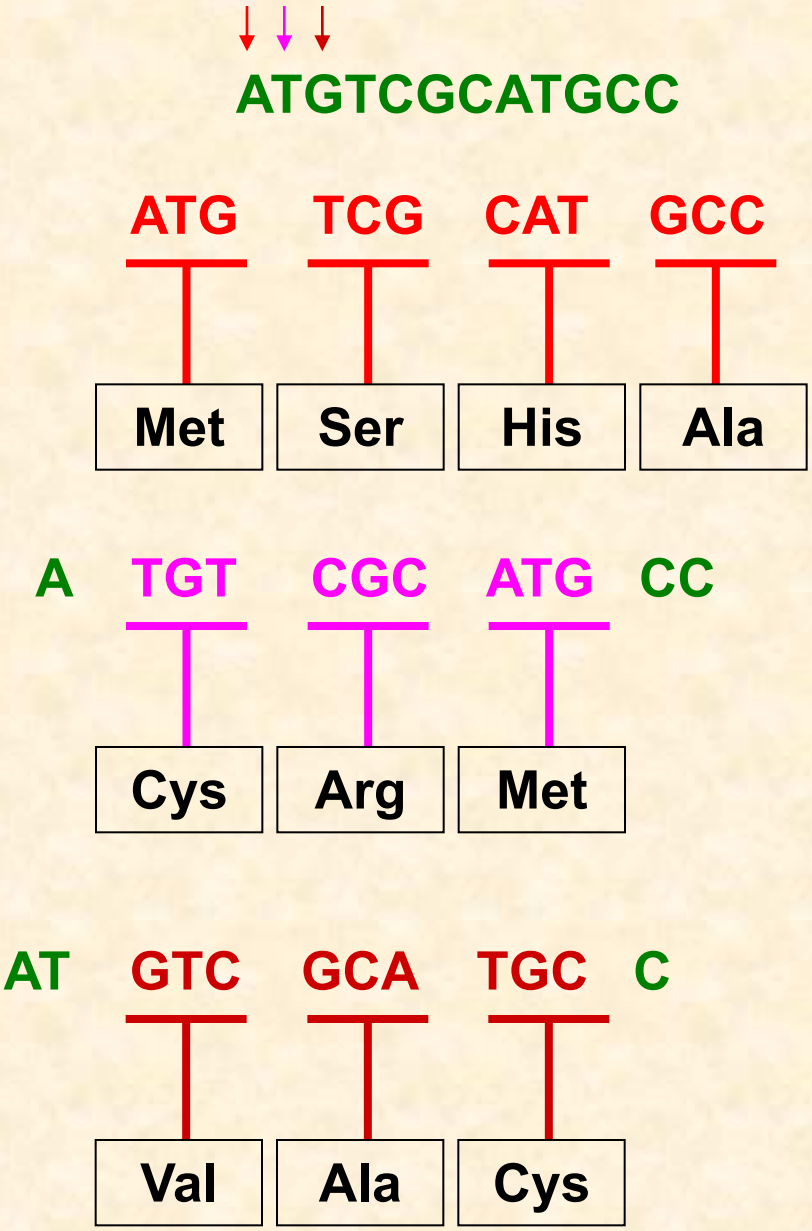
Překlad DNA sekvence

The table shows the 64 codons and the amino acid for each. The direction of the mRNA is 5' to 3'.

1st base	2nd base: U	2nd base: C	2nd base: A	2nd base: G
U	UUU (Phe/F) Phenylalanine UUC (Phe/F) Phenylalanine	UCU (Ser/S) Serine UCC (Ser/S) Serine	UAU (Tyr/Y) Tyrosine UAC (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine UGC (Cys/C) Cysteine
U	UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (*Stop*)	UGA Opal (*Stop*)
U	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (*Stop*)	UGG (Trp/W) Tryptophan
C	CUU (Leu/L) Leucine CUC (Leu/L) Leucine	CCU (Pro/P) Proline CCC (Pro/P) Proline	CAU (His/H) Histidine CAC (His/H) Histidine	CGU (Arg/R) Arginine CGC (Arg/R) Arginine
C	CUA (Leu/L) Leucine CUG (Leu/L) Leucine	CCA (Pro/P) Proline CCG (Pro/P) Proline	CAA (Gln/Q) Glutamine CAG (Gln/Q) Glutamine	CGA (Arg/R) Arginine CGG (Arg/R) Arginine
A	AUU (Ile/I) Isoleucine AUC (Ile/I) Isoleucine	ACU (Thr/T) Threonine ACC (Thr/T) Threonine	AAU (Asn/N) Asparagine AAC (Asn/N) Asparagine	AGU (Ser/S) Serine AGC (Ser/S) Serine
A	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
A	AUG (Met/M) Methionine, *Start* [A]	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU (Val/V) Valine GUC (Val/V) Valine	GCU (Ala/A) Alanine GCC (Ala/A) Alanine	GAU (Asp/D) Aspartic acid GAC (Asp/D) Aspartic acid	GGU (Gly/G) Glycine GGC (Gly/G) Glycine
G	GUA (Val/V) Valine GUG (Val/V) Valine	GCA (Ala/A) Alanine GCG (Ala/A) Alanine	GAA (Glu/E) Glutamic acid GAG (Glu/E) Glutamic acid	GGA (Gly/G) Glycine GGG (Gly/G) Glycine

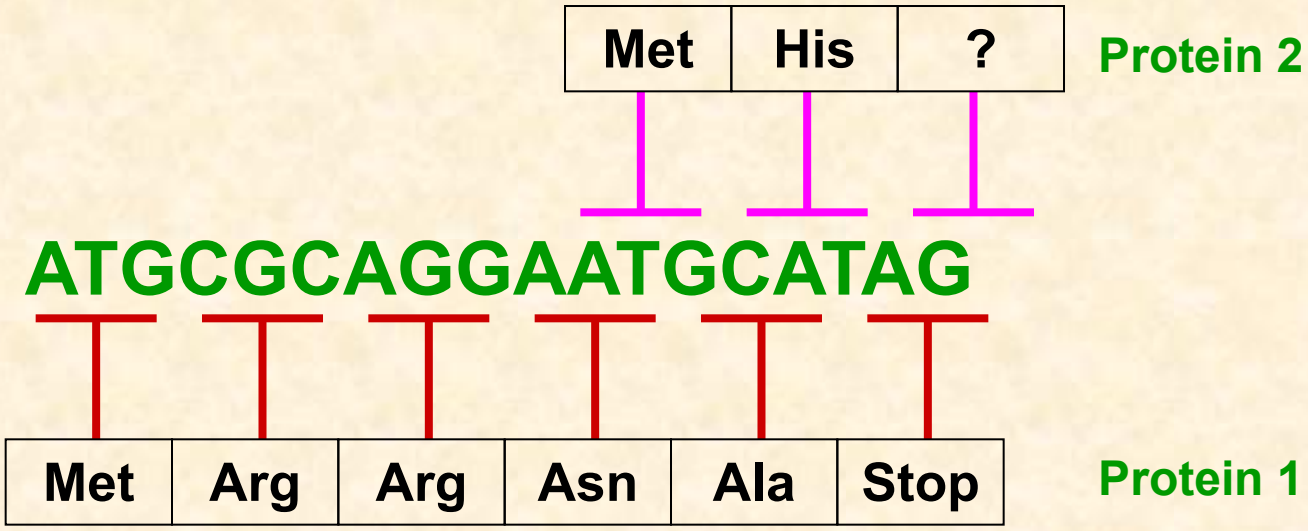
Překlad DNA sekvence





Čtení tripletů závisí na tom, u kterého nukleotidu stanovíme počátek čtení.

ATGCGCAGGAATGCATAG sekvence DNA



Překlad DNA sekvence – reverzní čtecí rámce

5 ATGCGCAGGAATGCATAG 3

3 TACGCGTCCTTACGTATC 5

A-T

G-C

Překlad DNA sekvence – od 5 konce

5 ATGCGCAGGAATGCATAG 3

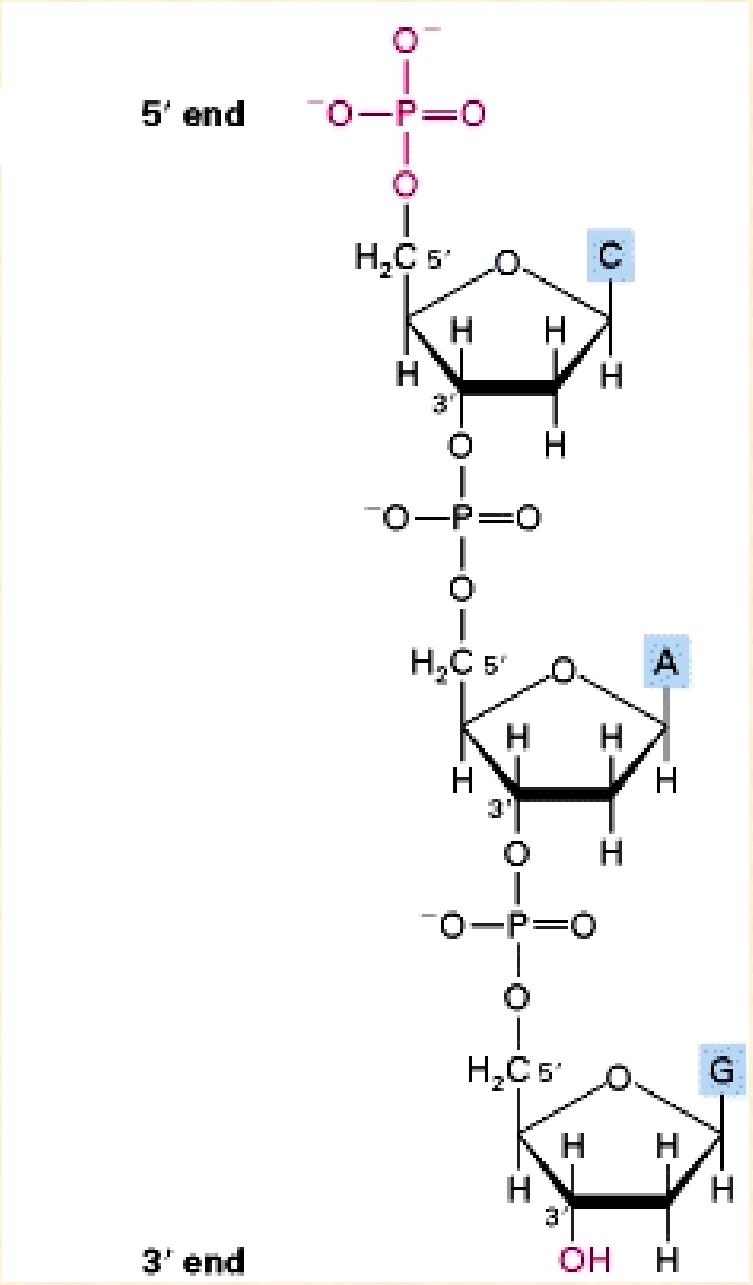
5 CTATGCATTCCTGCGCAT 3

Opravdu ORF kóduje protein?

- ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání DATABÁZÍ pomocí ALIGNMENTU) = nejspolehlivější ověření.

- Nástroje pro překlad DNA jsou propojeny s prohledáváním databází.

Polynukleotidový řetězec

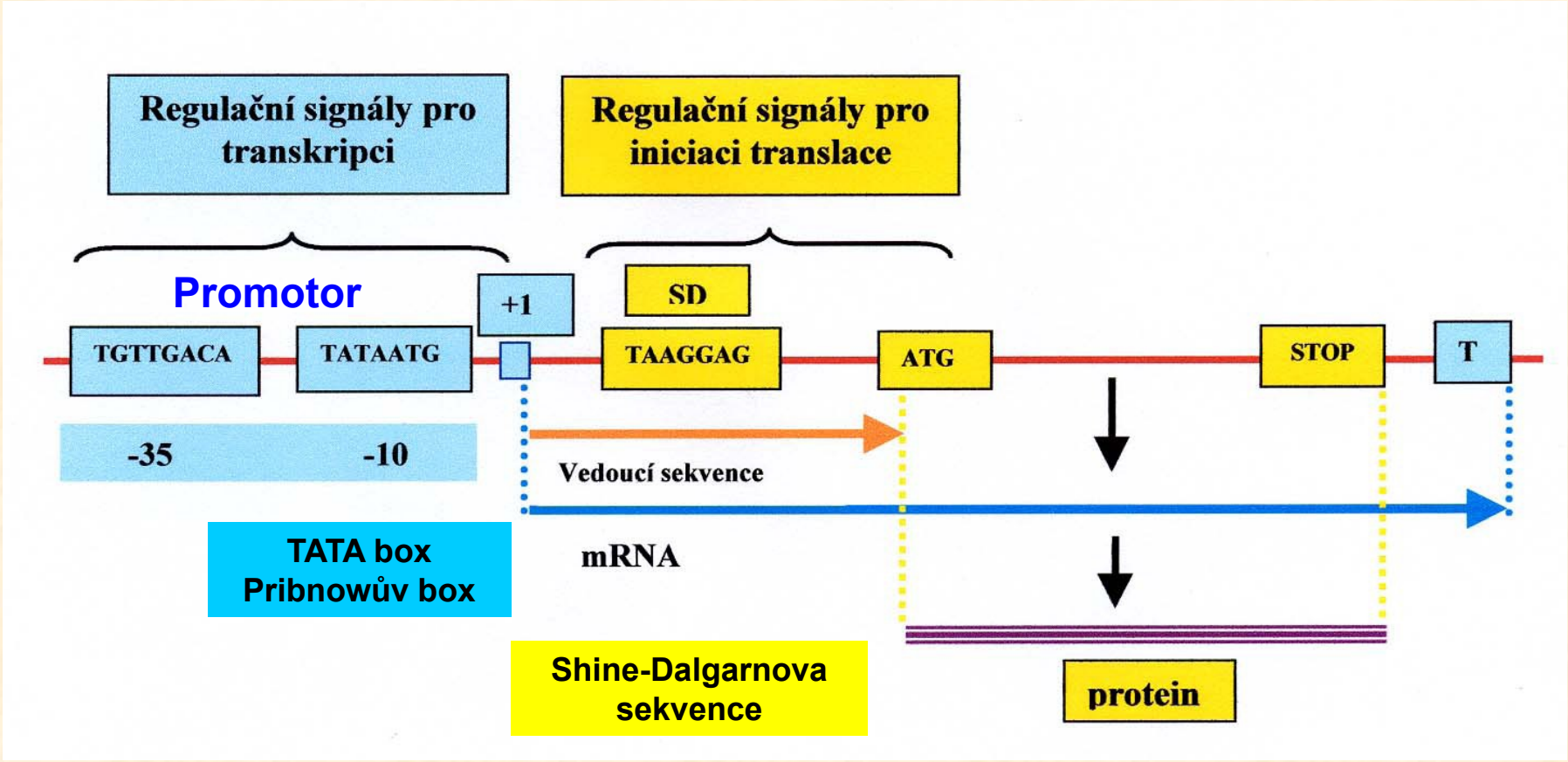


5 -P-C-A-G-3 -OH

5 ATGCGTTGACGTTGCACGTGCAA
GTCGCAGTCGATGCGATGCTGAC
GTACGTGCGTACGATGCGTCGTA
CGTGCTGACGTCGTACGT 3

5 → 3

Translační a transkripční signální sekvence



Prokaryota