

# Predikce genů

**Pro zajímavost...**

**Důležité...**

# Molekulárně biologická data

- **Výkonné technologie:**

Automatické sekvencování

MALDI-TOF

NMR spektroskopie

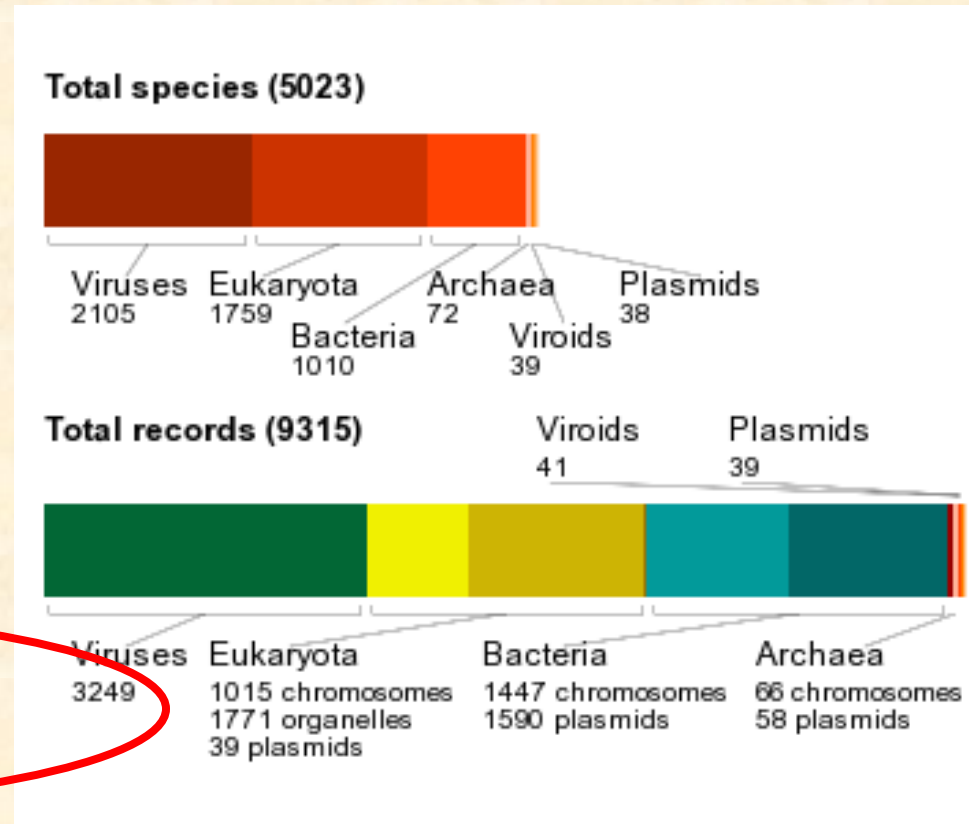
Proteinová krystalografie

**Výrazný nárůst množství biologických dat.**

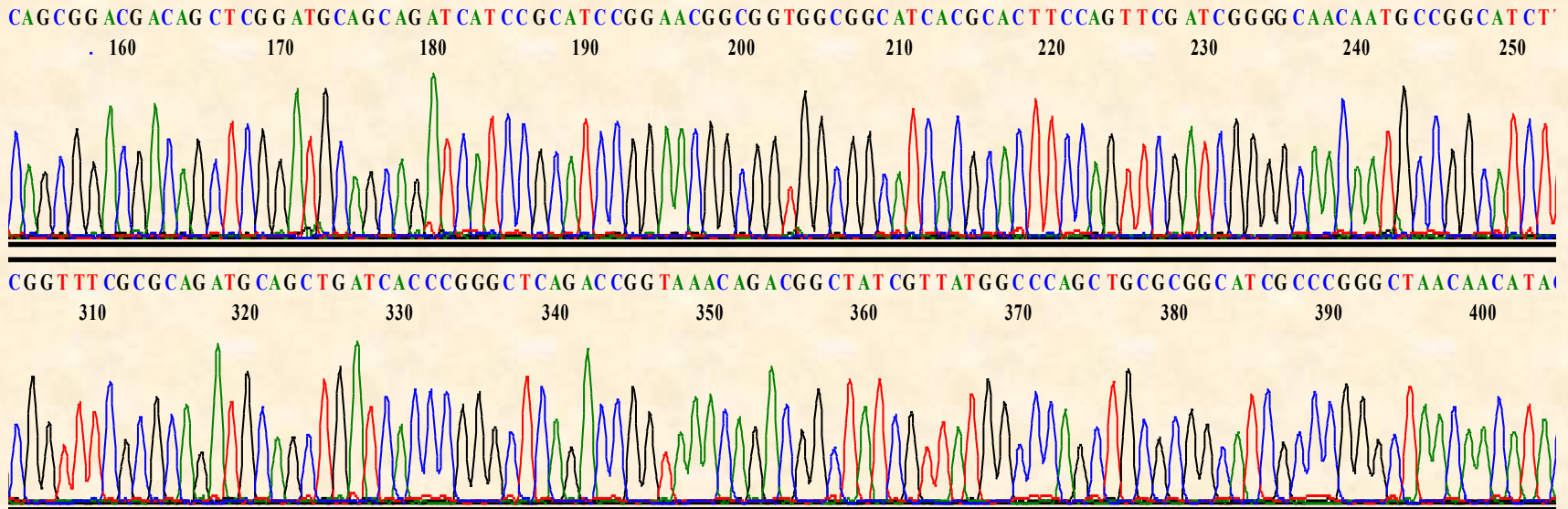
# Rozdělení molekulárně biologických databází

- **Databáze:**
  - Primární
  - Sekundární
  - Strukturní

**Genomové zdroje**



# Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGATTTTCATGCTGGTTTCCCAACGAAAAA TAACCGCTCACGGTGCCATCACGATCGCACACCGCAAATCGGCGG  
TACAGGTGGTCGCGCCCGCCGACACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC  
GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGGCAGCGCC  
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCAGCTGCGCGGCATCGCCCGGGCTAAACAA  
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

GATAGCGTAATGATCGGCTGGCTGCCGATTTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCCGAAAAATCGGCGG  
TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC  
GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCGCGCGCAGCGCC  
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACC GGTAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA  
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAAATCACCAGCAT

**„Syrové“ sekvence DNA**



**Identifikace a anotace genů a proteinů**

Table 1  
Software commonly used for bacterial genome annotation and comparison

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



# Predikce genů kódujících proteiny

- **Prokaryotické geny**
- Nepřerušované úseky DNA mezi **startovním kodonem** (ATG, GTG, TTG, CTG) a **stop kodonem** (TAA, TGA, TAG).
  
- **Eukaryotické geny**
- Přerušovány **introny**. Průměrná délka exonu je 50 kodonů, některé jsou mnohem kratší.
- Některé introny extrémně dlouhé, geny zabírají mbp v genomové DNA.

**Predikce eukaryotických genů je  
mnohem složitější než predikce  
genů prokaryotických a  
představuje **STÁLE**  
**NEVYŘEŠENÝ** problém!**



# Prokaryotické geny

- **Prokaryotický gen = nejdelší ORF odpovídající danému úseku DNA.**

GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCC  
GACCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATAGC  
CGTCTGTTTACCGGTCTGAGCCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGGAAG  
TGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCGTGA  
TGCCGCCACCGCCGTTCCGGATGCGGGATGATCTGCTGCATCCGAGCTGTTCGTCCGCTGAAAGATCATTATTGG  
CGCAGCGATGTGCTGGCGGGCGGGCGGACCACCTGTACCGCCGATTTTTCGGGTGTGCGATCGTGATGGCACCG  
TGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATAACCAAACAGCCGGGCTT  
TAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCTTCTATGCG  
AACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTA  
ACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAA  
CGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTGGCTG  
GGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAATGGG

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'.

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

# Překlad DNA sekvence

- **ExPASy**

<http://web.expasy.org/translate/>

- **ORF Finder (NCBI)**

<https://www.ncbi.nlm.nih.gov/orffinder/>

# ExpASy

<http://www.expasy.org/vg/index/dna>



- Visual Guidance
- DNA
- RNA
- Protein
- Cell
- Organism
- Population
- Categories
- Resources A..Z
- Links/Documentation


Selected keywords > translation < >


**Keywords**

Choose a category or a keyword






codon conversion tool  
protein protein  
sequence reverse  
transcription reverse  
translation sequence  
analysis transcription

---

 SIB resources

 External resources -  
(No support from the ExpASy Team)

**Databases (0)**      **Tools (5)**

-  **EMBOSS translation tools**  
EMBOSS sequence translation tools, incl. backtranslation [more]  
Keywords: codon, DNA sequence, protein, translation
-  **Graphical Codon Usage Analyser**  
Displays the codon bias in a graphical manner [more]  
Keywords: codon, DNA sequence, sequence analysis, translation
-  **Reverse Transcription and Translation Tool**  
Transcription, translation and reverse transcription [more]  
Keywords: DNA sequence, protein sequence, reverse transcription, transcription, translation
-  **Reverse Translate**  
Translates a protein sequence back to a nucleotide sequence [more]  
Keywords: DNA sequence, protein sequence, reverse translation, translation
-  **Translate**  
Translation of a nucleotide (DNA/RNA) sequence to a protein sequence [more]  
Keywords: codon, conversion tool, DNA sequence, protein, protein sequence, translation

"Expert Protein Analysis System"

# ExPASy

<http://web.expasy.org/translate/>

**Translate** is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).

```
GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCC
CCGACCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGA
TAGCCGTCTGTTTACCGGTCTGAGCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCG
CGGAAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCGATCGAACTGGAA
GTGCGTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGA
TCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATC
GTGATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACC
AAACAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAA
AGCGATCTTCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGG
CCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATT
CGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGA
TACCGTGTGGCTGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTC
TGCAGTGGCCGATTACCTAATGGG
```

Output format:  ▼

Reset

or

TRANSLATE SEQUENCE

# Translate Tool - Results of translation

Open reading frames are highlighted in red. Please select one of the following frames - in the next page, you will be able to select your initiator and retrieve your amino acid sequence:

## 5'3' Frame 1

VCW **Stop** LW **Met** PLPC **Stop** APIRKPAVIRPPRP **Stop** L **Met** VATC **Met** LLARA **Met** PRSWAITIAVCLPV **Stop** ARVISCICAKPRWRCARK **Stop** ACCLFALP **Stop** K **Met** PALLPRSNWKCV **Met** PPPPFR **Met** R **Met** ICCIRAVVR **Stop** KIIIGAA **Met** CWRRARPPVPPILRCAIV **Met** AP **Stop** AVIFVGKPAKLRASRIPNSRALNRAAIA **Met** ATLACRRIPPLKRSS **Met** RTRRIVR **Stop** NCLL **Met** Met RRRNPPLWVTAK **Met** VCVCLP **Stop** IAKVVKFVLKRARTAVRARP **Met** PVWRR **Stop** ARAIPCGWAGWARK **Met** VP **Met** RII **Met** Met ALLFCSGRLPNG

## 5'3' Frame 2

YAGDCGCRYPAERLSGSQP **Stop** SGRPDRD **Stop** WSPPVCC **Stop** PGR CRAAGP **Stop** R **Stop** PSVYRSEPG **Stop** SAASARNRAGAAR **Stop** SERAVYSLCPCRCRHCCPDRTGSA **Stop** CRHRRSGCG **Stop** SAASELSSAERSLLAQRCAAGGRDHLRYRFCGVRS **Stop** WHRE **Stop** RFLSLGNQH **Stop** NCGQPAGYQTAGL **Stop** TEQRSQWQL **Stop** PAAEYRL **Stop** SDLLCERGGSSGSETVY **Stop** **Stop** CAGTGRHLCG **Stop** QRRWCASVYPE **Stop** QRW **Stop** NSY **Stop** SERERPSEDRCPSGAAERG RYRVAGLAGRGRWCRCGL **Stop** **Stop** WHCYSAVADYL **Met**

## 5'3' Frame 3

**Met** LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKD **Stop** AGIVAPIELEVRDAATAVDPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFVCDRDGTVSGYFRWETSIEIAGSQPDTKQP **Stop** GFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARL **Stop** APLSAGDTVWLGWLG AEDGADADYNDGIVILQWPIT **Stop** W

## 3'5' Frame 1

PIR **Stop** SATAE **Stop** QCHHYNPHRHHLPRPASPATRYRPRSAAPDGHRSRSDGRSRSLQYEFYHLCYSG **Stop** TDAHHLRCYPQRW **Stop** RPVPAHHQ **Stop** TVSDPDDPPRSHRSL **Stop** RRYSAAG **Stop** SCHCDRCV **Stop** SPAVWYPAGCPQFCWF PENNRSRCHHDR **Stop** TPQNR RYRWSRPPAHR CANNDSLADDSSDAADHPHPERRWRHHALPVRSGQQCRHLSGQSE **Stop** TARSLPRAAPARFRAD **Stop** AADHPGSDR **Stop** TDGYRYGPAARHRPG **Stop** QHTGGDHQSRSGRPDHGWLPDRRSAG **Stop** RHPQSPAY

## 3'5' Frame 2

PLGNRPLQNNNAIIIRIGTIFRAQPAQPHGIARAQRRQTGIGRALTAVRARFNTNFTTFAIQGGQHTHTIFAVTHKGGGRFRRIINKQF **Stop** QILTIRRVRIEDRFKGGIRRQAKVAIAIARFKARLFGIRLAARNFNAGFPTKITAHGAITIAHRKIGGTGGRARRQHIAAPI **Met** IFQR **Stop** TAR **Met** QQIIRIRNGGGGITHFQFDRGNNAGIFQGKANKQHAHFRAQRQRGFAQ **Met** QLITRAQTGKQTAIV **Met** AQLRGIARANNIQV **Stop** ATINHGRGGRITAGFRIGAQQGNGIHNHQH

## 3'5' Frame 3

H **Stop** VIGHCRIT **Met** PSL **Stop** SASAPSSAPSQPSHTVSPALSGARRASVAL **Stop** RPFALASIRILPPLLFRVNRRTPSLLPTKVAA **Stop** GSGASSINSFRS **Stop** RSAAFA **Stop** KIALKAVFGGRLKLPRLSLLGLKPGCLVSGWLP AIS **Met** LVSQRK **Stop** PLTVPSRSHTAKSA **Stop** VQVVAPAASTSLRQ **Stop** **Stop** SFSGRQLGCSRSSASGTAVAASRTSSSIGAT **Met** PASFRAKRINSTLTSARSASAVSRRC **Stop** S **Stop** PGLRPVNRRLSLWPSCAASPGLTTYRWRPSITVGAAGSRLASG **Stop** ALSRVTASTISI



# ORF Finder (NCBI)

<https://www.ncbi.nlm.nih.gov/orffinder/>

## Open Reading Frame Finder


ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.


This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).

**Examples** (click to set values, then click Submit button) :

- [NC\\_011604](#) Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- [NM\\_000059](#); genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

## Enter Query Sequence

 Enter accession number, gi, or nucleotide sequence in FASTA format:

 From:  To:



# ORF Finder (NCBI)

<https://www.ncbi.nlm.nih.gov/orffinder/>

## Choose Search Parameters

Minimal ORF length (nt):

Genetic code:

1. Standard

ORF start code:

1. Standard

"ATG" only

"ATG" and

Any sense

Ignore nested

2. Vertebrate Mitochondrial

3. Yeast Mitochondrial

4. Mold, Protozoan and Coelenterate Mitochondrial, and the Mycoplasma/Spiroplasma

5. Invertebrate Mitochondrial

6. Ciliate, Dasycladacean and Hexamita Nuclear

9. Echinoderm and Flatworm Mitochondrial

10. Euplotid Nuclear

11. Bacterial, Archaeal and Plant Plastid

12. Alternative Yeast Nuclear

13. Ascidian Mitochondrial

14. Alternative Flatworm Mitochondrial

16. Chlorophycean Mitochondrial

21. Trematode Mitochondrial

22. Scenedesmus obliquus Mitochondrial

23. Thraustochytrium Mitochondrial

24. Pterobranchia Mitochondrial

25. Candidate Division SR1 and Gracilibacteria

Start Search /

Submit

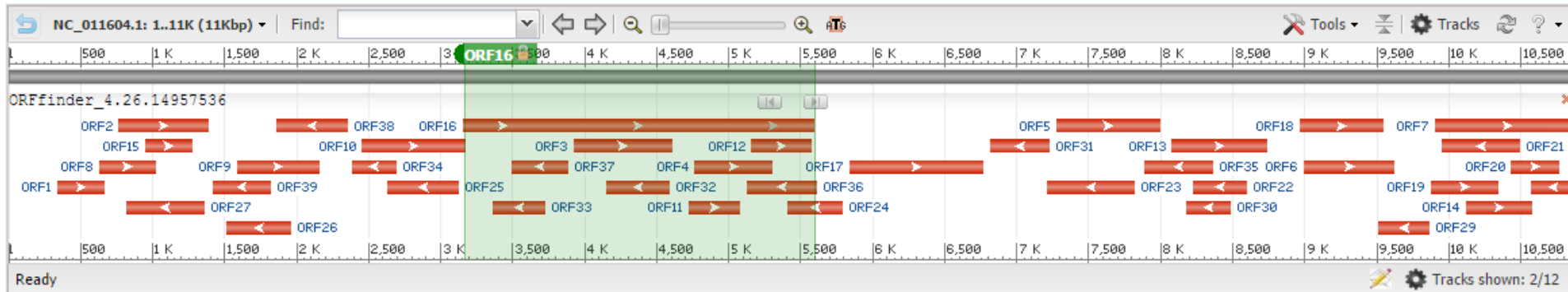
Clear

# ORF Finder (NCBI)

<https://www.ncbi.nlm.nih.gov/orffinder/>

Salmonella enterica subsp. enterica serovar Westhampton plasmid pWES-1, complete sequence

ORFs found: 39 Genetic code: 11 Start codon: 'ATG' and alternative codons



Add six-frame translation track

ORF16 (813 aa) Display ORF as... Mark

Mark subset... Marked: 0 Download marked set as Protein FASTA

```
>lcl|ORF16
MKAKVSRGGGFRGALNYVFDVGKEATHTKNAERVGGNMG
NDPRELSREFSAVRLRPDVGKPVNHCSLSLPPGERLSAE
KWEAAADFVQRMGFDQNTNPWAVRHQDTDKDHIHIVAS
RVGLDGKVMWLGQNEARRAIEATQELEHTHGLTLPGLGDA
RAERRKLTDKKEINMAVRTGDEPPRQRLQRLLEAVKDKPT
ALELAERLQAAGVGVVLANLASTGRMNGFSFEVAGVPPKGS
DLGKGYTWAGLQKAGVTYDEARDRAGLERFRPTVADRGER
QVVAAVREPDARGLEAPTGRSLDRDGDADLGTAGPTAGRD
AGSGSLRQGDGHSADQAGRADAADERERAGLRAEGREAG
RDHLRPVAQPVRAENEPQHGADRAAGGDLAQAGERTAG
HDESRRPTDRGSEADAPAPLAAGAGADSGRGRDAGSDW
ASRFKQASAAKRRRAADGRLQQRDLEQGHAGARVAETDRQ
```

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF16	+	3	3153	5594	2442   813
ORF7	+	1	9907	10908	1002   333
ORF17	+	3	5841	6767	927   308
ORF5	+	1	7273	7995	723   240
ORF10	+	2	2456	3169	714   237
ORF3	+	1	3925	4608	684   227

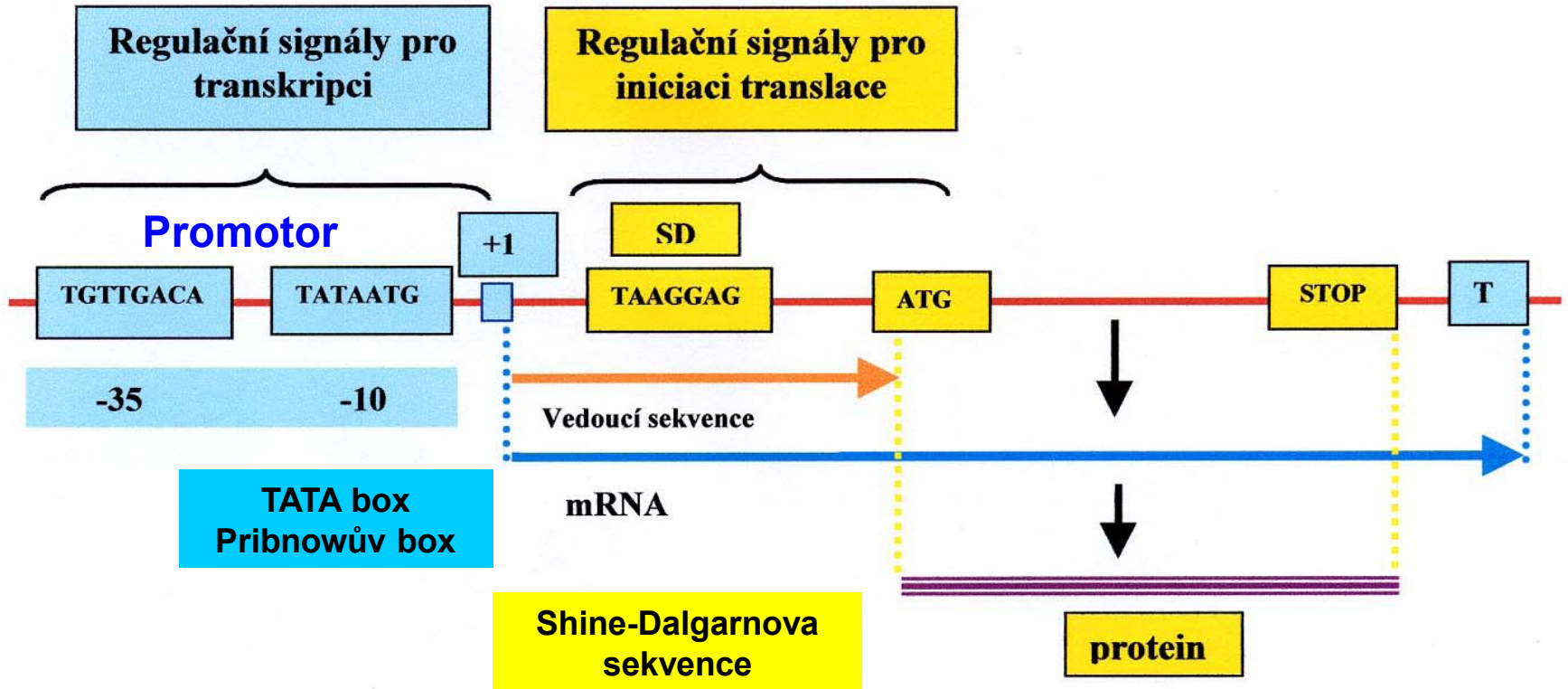
# Prokaryotické geny

- **Velmi jednoduchý přístup k predikci genů**  
Zjednodušení vede k chybám, ale jejich množství je **POMĚRNĚ MALÉ**.
- **Chyby mohou vznikat při SEKVENCOVÁNÍ DNA.**  
Přidání/odstranění startovního a/nebo stop kodonu může vést ke **ZKRÁCENÍ**, **PRODLOUŽENÍ** nebo úplnému **VYNECHÁNÍ** genu.

# Opravdu ORF kóduje protein?

- **ORF kóduje protein, který je podobný již dříve popsanému proteinu** (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**).
- **ORF má typický obsah GC nebo frekvenci kodonů.** Srovnání s charakteristickými vlastnostmi známých genů ze stejného organismu.
- **Před ORF se nachází typické RBS (ribosome-binding site) nebo promotor.**

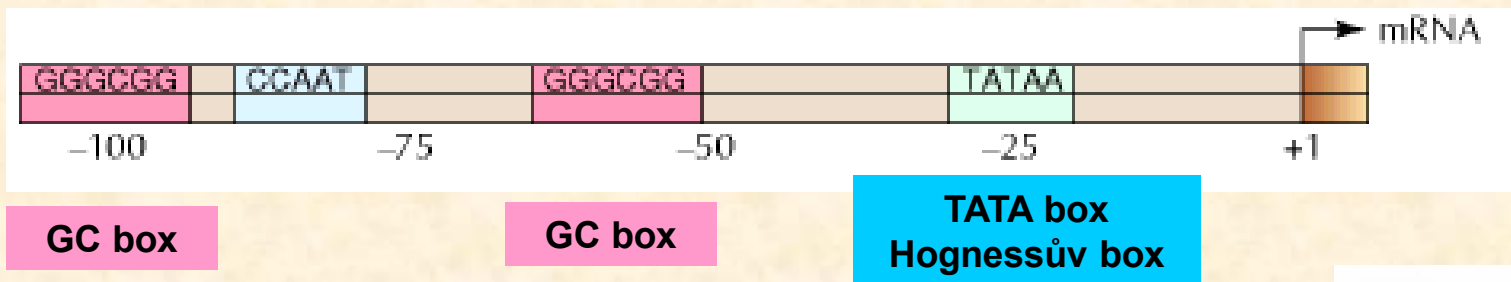
# Translační a transkripční signální sekvence



## Prokaryota

# Translační a transkripční signální sekvence

Regulační signály pro transkripci



Promotor RNA-polymerasy II

Regulační signály pro iniciaci translace

(gcc)gccRccAUGG

Kozak sequence  
Sekvence Kozakové

## Eukaryota



# 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.**



# Translate Tool - Results of translation

```
ID VIRT18492 Unreviewed; 289 AA.
AC VIRT18492;
DE Translation of nucleotide sequence generated on ExPASy
DE on 08-May-2014 by 147.251.28.220.
CC -!- This virtual protein sequence will automatically be deleted
CC from the server after a few days.
DR SWISS-2DPAGE; VIRT18492; VIRTUAL.
SQ SEQUENCE 289 AA; 266AF312C81FBE3D CRC64.
MLVIVDAVTL LSAYPEASRD PAAPTVIDGR HLYVVS PGDA AQLGHNDSRL FTGLSPGDQL
HLRETALALR AEVSVLFIRF ALKDAGIVAP IELEVRDAAT AVPDADDDLH PSCRPLKDHY
WRSDVLAAGA TTCTADFAVC DRDGTVSGYF RWETSIEIAG SQPDTKQPGF KPSSDRNGNF
SLPNTAFKA IFYANAADRQ DLKLFIDDAP EPAATFVGNS EDGVRLFTLN SKGGKIRIEA
SANGRQSATD ARLAPLSAGD TVWLGWLGAE DGADADYNDG IVILQWPIT
```

//

Sequence in [FASTA format](#)

[BLAST](#) BLAST submission on ExPASy/SIB



ScanProsite



Sequence analysis tools: [ProtParam](#), [ProtScale](#), [Compute pI/Mw](#),



[Direct Submission to SWISS-MODEL](#)

# ORF Finder (NCBI)

<https://www.ncbi.nlm.nih.gov/orffinder/>

ORF16 (813 aa) [Display ORF as...](#) [Mark](#)

```
>|c1|ORF16
MKAKVSRGGGFRGALNYVFDVGKEATHTKNAERVGGNMAG
NDPRELSREFSAVRQLRPDIGKPVWHCSLSLPPGERLSAE
KWEAVAADFMRMGFDQTNTPWAVRHQDQDKDHIHIVAS
RVGLDGKVNLGQWEARRAIEATQLEHTHGLTTPGLGDA
RAERRKLT DKEINMAVRTGDEPPRQLRLLDEAVKDKPT
ALELAERLQAAGVGVANLASTGRMNGFSFEVAGVPPFKGS
DLGKGYTWAGLQKAGVTYDEARDRAGLERFRPTVADRGER
QDVAAVREPDARGLEAPTGRSLDRDGADLGTAGPTPAGRD
AGSGSLRQGDGHSAQDAGRADAADERERAGLRAEGREAG
RDHLRPVAQPVRAENEPQQHGADRAAGDLAGQAGER TAG
HDESRRPTDRGSESDAPAPLAAGAGADSGRGRDRDAGSDW
ASRFKOASAAKRBADGRLGQRDLEQGHAGARVAETDRQ
```

[SmartBLAST ORF16](#)

[BLAST ORF16](#)

[BLAST marked set](#)

BLAST Database:

UniProtKB/Swiss-Prot (swissprot) ▼

UniProtKB/Swiss-Prot (swissprot)

Reference proteins (refseq\_protein)

Non-redundant protein sequences (nr)

[Go back to the submitting page...](#)

[Mark subset...](#) Marked: 0 [Download marked set](#) as [Protein FASTA](#) ▼

Label	Strand	Frame	Start	Stop	Length (nt   aa)
ORF2	+	1	766	1392	627   208
ORF23	-	1	7818	7210	609   202
ORF18	+	3	8961	9551	591   196
ORF9	+	2	1586	2155	570   189
ORF27	-	1	1362	814	549   182
ORF21	-	1	10497	9955	543   180
ORF4	+	1	4765	5301	537   178
ORF25	-	1	3129	2629	501   166
ORF36	-	3	5617	5123	495   164
ORF38	-	3	2353	1859	495   164
ORF35	-	3	8362	7866	477   159

# Eukaryotické geny

## Jednobuněčná eukaryota

- **Genomy jednobuněčných eukaryot se výrazně liší** (frekvence intronů, jak velká část genomu je tvořena geny kódujícími proteiny).
- *Saccharomyces cerevisiae* – 67% genomu je protein-kódující, jen 4% obsahují introny.
- Hlenky – průměrný gen obsahuje 3,7 intronu.
- **Pro některá jednobuněčná eukaryota (kvasinky) je možné použít stejné postupy jako pro prokaryota.**





**Slime mold = hlenka**

***Fuligo septica***

**Dog vomit slime mold**

# Eukaryotické geny

## Mnohobuněčná eukaryota

- **Mnohobuněčná eukaryota**

Komplexní organizace genomu, geny separovány dlouhými **INTERGENOVÝMI** úseky, geny obsahují množství **INTRONŮ**, i velmi **DLOUHÝCH**.



Glyceraldehyd-3-fosfát-dehydrogenasa  
*Candida albicans*

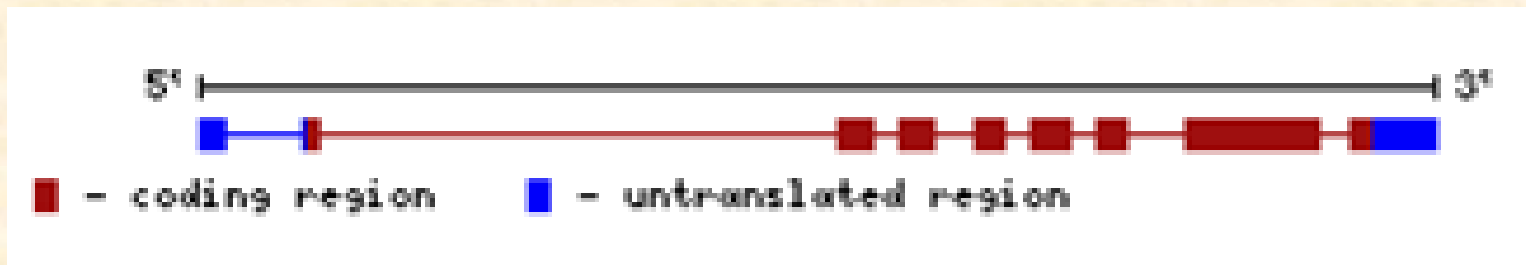


# Eukaryotické geny

## Mnohobuněčná eukaryota

- **Mnohobuněčná eukaryota**

Komplexní organizace genomu, geny separovány dlouhými **INTERGENOVÝMI** úseky, geny obsahují množství **INTRONŮ**, i velmi **DLOUHÝCH**.



Glyceraldehyd-3-fosfát-dehydrogenasa  
*Homo sapiens*

# Eukaryotické geny

## Mnohobuněčná eukaryota

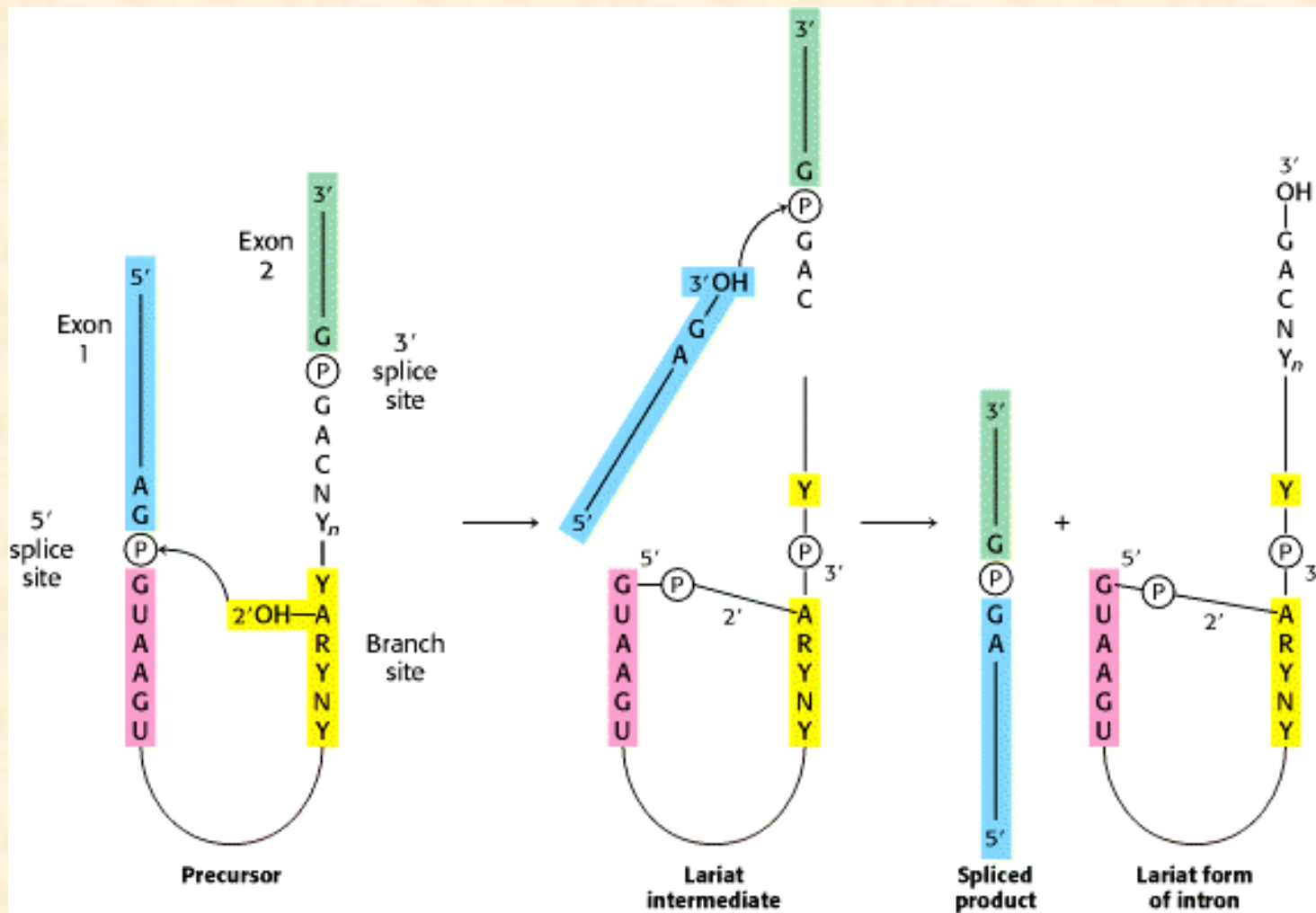
- **Rozpoznání exonů/intronů**

Identifikace míst sestřihu: **GT** na 5 konci, **AG** na 3 konci.

- **Chyby při rozpoznávání exonů/intronů**

Velké množství chyb. Dlouhé introny – určeny jako intergenové úseky. Krátké intergenové useky – určeny jako introny.



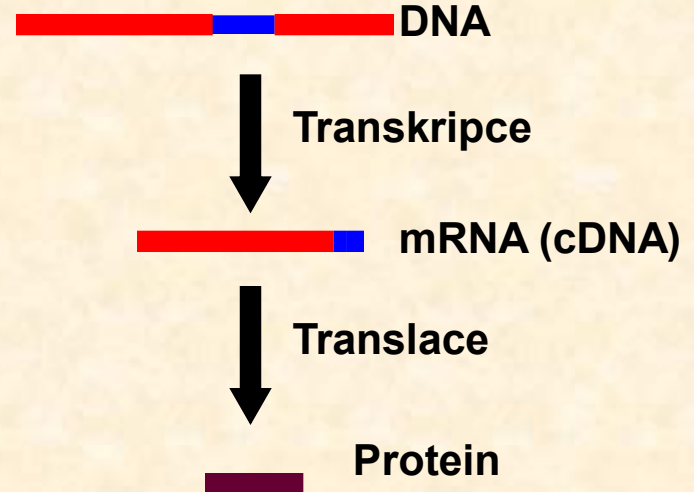
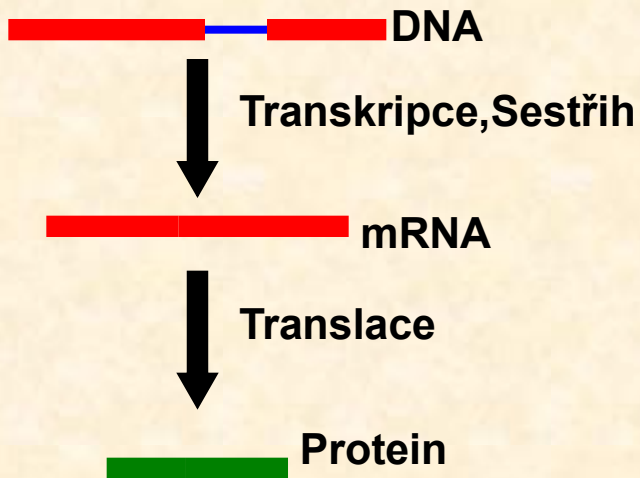


**Splicing Mechanism Used for mRNA Precursors.** The upstream (5') exon is shown in blue, the downstream (3') exon in green, and the branch site in yellow. Y stands for a purine nucleotide, R for a pyrimidine nucleotide, and N for any nucleotide. The 5' splice site is attacked by the 2'-OH group of the branch-site adenosine residue. The 3' splice site is attacked by the newly formed 3'-OH group of the upstream exon. The exons are joined, and the intron is released in the form of a lariat. [After P. A. Sharp. *Cell* 2(1985):3980.]

# Predikce genů – příklad z praxe

Hypotetický gen/protein,  
predikovaný při anotaci genomu  
*Aspergillus fumigatus* Af293

Identifikace genu/proteinu  
na úrovni mRNA (příprava cDNA  
pro klonování)



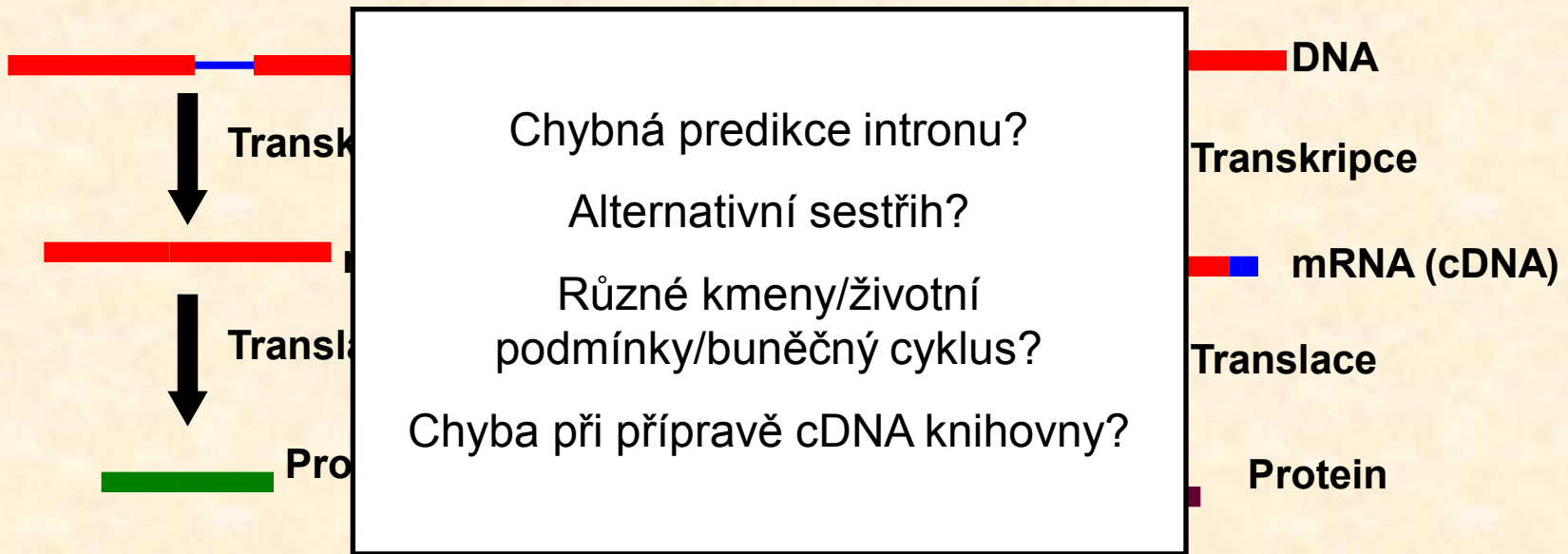
MADPEVEADG ELDLEKRASA QTCKIVNVDT  
YVNCRYDAKL DAGAIFGFPK GEKLTFCWK  
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS  
NCSSMLSRC

MADPEVEADG ELDLEKRASA QTCKIVNVDT  
YVNCRYDAKL DAGAIFGFPK GEKLTFCWK  
HGDCYNGVW

# Predikce genů – příklad z praxe

Hypotetický gen/protein,  
predikovaný při anotaci genomu  
*Aspergillus fumigatus* Af293

Identifikace genu/proteinu  
na úrovni mRNA (příprava cDNA  
pro klonování)



MADPEVEADG ELDLEKRASA QTCKIVNVDT  
YVNCRYDAKL DAGAIFGFPK GEKLTFCWK  
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS  
NCSSSMLSRC

MADPEVEADG ELDLEKRASA QTCKIVNVDT  
YVNCRYDAKL DAGAIFGFPK GEKLTFCWK  
HGDCYNGVW

# Algoritmy a nástroje pro identifikaci genů

- **Predikce genů na základě sekvenční homologie** – vyhledávání v databázích pomocí algoritmů.
- **Predikce genů *ab initio*** – predikce na základě statistických parametrů DNA sekvence.
- **Většina běžně používaných metod kombinuje oba dva přístupy.**

# Prokaryota

ATG.....TAA

Bez intronů

**SEKVENČNÍ HOMOLOGIE**



**IDENTIFIKOVANÉ GENY VYUŽITY  
PRO „TRÉNOVÁNÍ“ STATISTICKÉ  
METODY**



**ANALÝZA ZBÝVAJÍCÍCH  
ČÁSTÍ GENOMU**

# Eukaryota

Mnoho intronů, dlouhé intergenové úseky  
*Ab initio* STATISTICKÉ METODY



IDENTIFIKOVANÉ EXONY



SEKVENČNÍ HOMOLOGIE

# Algoritmy a nástroje pro identifikaci genů

- Každý program má výhody a nevýhody –  
rozumné použít více predikčních nástrojů.

**GeneMark**

**GlimmerM**

**GRAIL**

**GenScan**

**Fgenes**



# Algoritmy a nástroje pro identifikaci genů

- **GeneMark**

<http://exon.gatech.edu/GeneMark>

Využívá **Markovovy** modely

Vyžaduje parametry specifické pro daný organismus = nutné „natrénování“ pomocí známých genů

Varianty pro prokaryotické, eukaryotické, virové sekvence

# GeneMark

<http://exon.gatech.edu/GeneMark>

## Gene Prediction in Bacteria, Archaea and Metagenomes



For bacterial and archaeal gene prediction we recommend to use a parallel combination of [GeneMark-P\\*](#) and [GeneMark.hmm-P](#) with pre-computed models.

A novel genome can be analyzed either by the program with [Heuristic models](#) (if the sequence is shorter than 100 kb) or by the self-training program [GeneMarks\\*](#) (aka GeneMark.hmm-PS).

Metagenomic sequences can be analyzed by our [new program](#) with updated heuristic models.

## Gene Prediction in Eukaryotes



For eukaryotic gene prediction you can use the parallel combination of [GeneMark-E\\*](#) and [GeneMark.hmm-E](#).

For a novel genome (the one whose name is not in the list of available models) you can install and run locally GeneMark.hmm-ES, the self-training program (just 10MB sequence is needed for training).

## Gene Prediction in Viruses, Phages and Plasmids



For novel virus, phage and plasmid gene prediction you can use either the [Heuristic approach](#) (if the sequence is shorter than 50 kb) or the self-training program [GeneMarks](#) (aka GeneMark.hmm-PS). Both options will run the parallel combination of GeneMark and GeneMark.hmm.

# Algoritmy a nástroje pro identifikaci genů

- **GeneScan**

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

**Komplexní model** struktury genu (transkripční, translační, sestřihové signály + statistické vlastnosti kódujících a nekódujících úseků)

Primární analýza velkých úseků eukaryotické genomové DNA



# Algoritmy a nástroje pro identifikaci genů

Program	Organism	Algorithm*	Website	Homology
GeneID	Vertebrates, plants	DP	<a href="http://www1.imim.es/geneid.html">http://www1.imim.es/geneid.html</a>	
FGENESH	Human, mouse, Drosophila, rice	HMM	<a href="http://www.softberry.com/berry.phtml?topic=fgenes&amp;group=programs&amp;subgroup=gfind">http://www.softberry.com/berry.phtml?topic=fgenes&amp;group=programs&amp;subgroup=gfind</a>	
GeneParser	Vertebrates	NN	<a href="http://beagle.colorado.edu/~eesnyder/GeneParser.html">http://beagle.colorado.edu/~eesnyder/GeneParser.html</a>	EST
Genie	Drosophila, human, other	GHMM	<a href="http://www.fruitfly.org/seq_tools/genie.html">http://www.fruitfly.org/seq_tools/genie.html</a>	protein
GenLang	Vertebrates, Drosophila, dicots	Grammar rule	<a href="http://www.cbil.upenn.edu/genlang/genlang_home.html">http://www.cbil.upenn.edu/genlang/genlang_home.html</a>	
GENSCAN	Vertebrates, Arabidopsis, maize	GHMM	<a href="http://genes.mit.edu/GENSCAN.html">http://genes.mit.edu/GENSCAN.html</a>	
GlimmerM	Small eukaryotes, Arabidopsis, rice	IMM	<a href="http://www.tigr.org/tdb/glimmerm/glmr_form.html">http://www.tigr.org/tdb/glimmerm/glmr_form.html</a>	
GRAIL	Human, mouse, Arabidopsis, Drosophila	NN, DP	<a href="http://compbio.ornl.gov/Grail-bin/EmptyGrailForm">http://compbio.ornl.gov/Grail-bin/EmptyGrailForm</a>	EST, cDNA
HMMgene	Vertebrates, <i>C. elegans</i>	CHMM	<a href="http://www.cbs.dtu.dk/services/HMMgene/">http://www.cbs.dtu.dk/services/HMMgene/</a>	
AUGUSTUS	Human, Arabidopsis	IMM, WWAM	<a href="http://augustus.gobics.de/">http://augustus.gobics.de/</a>	
MZEF	Human, mouse, Arabidopsis, Fission yeast	Quadratic discriminant analysis	<a href="http://rulai.cshl.org/tools/genefinder/">http://rulai.cshl.org/tools/genefinder/</a>	

\*DP, dynamic programming; NN, neural network; MM, Markov model; HMM, Hidden Markov model; CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM.

# Shrnutí

- Predikce prokaryotických genů **mnohem** jednodušší než u eukaryotických.
- Predikce genů ***ab initio***/na základě sekvenční homologie.
- Nutné **kombinovat** oba přístupy.
- Rozumné využívat **více** predikčních programů.