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

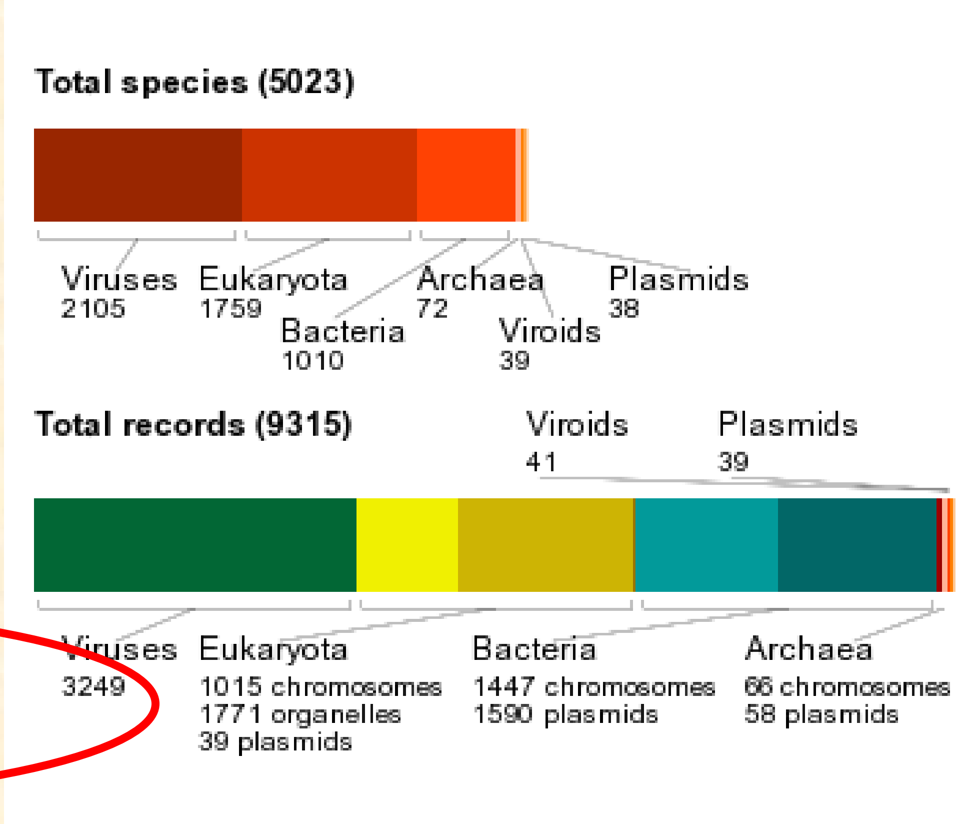
- **Databáze:**

  Primární

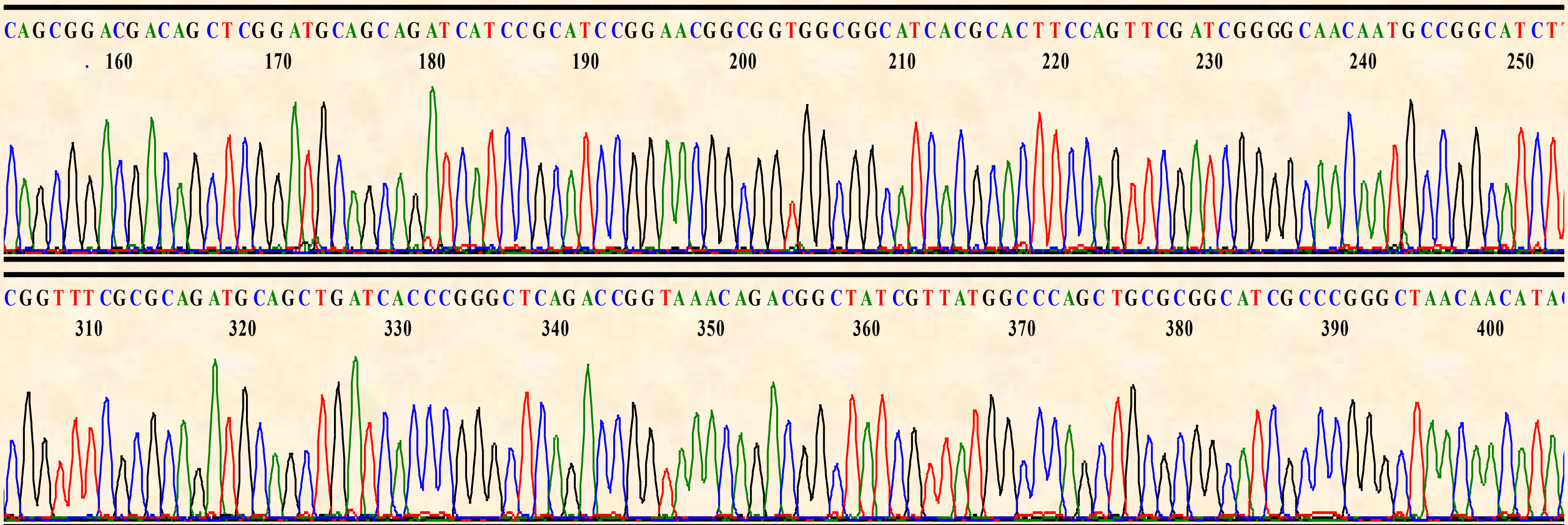
  Sekundární

  Strukturní

  Genomové zdroje



Molekulárně biologická data



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

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
CGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGGAAG
TGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCGTGA
TGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGG
CGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCG
TGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAACAGCCGGGCTT
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 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
		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 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
		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
		AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
		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
		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

- **ExPASy**

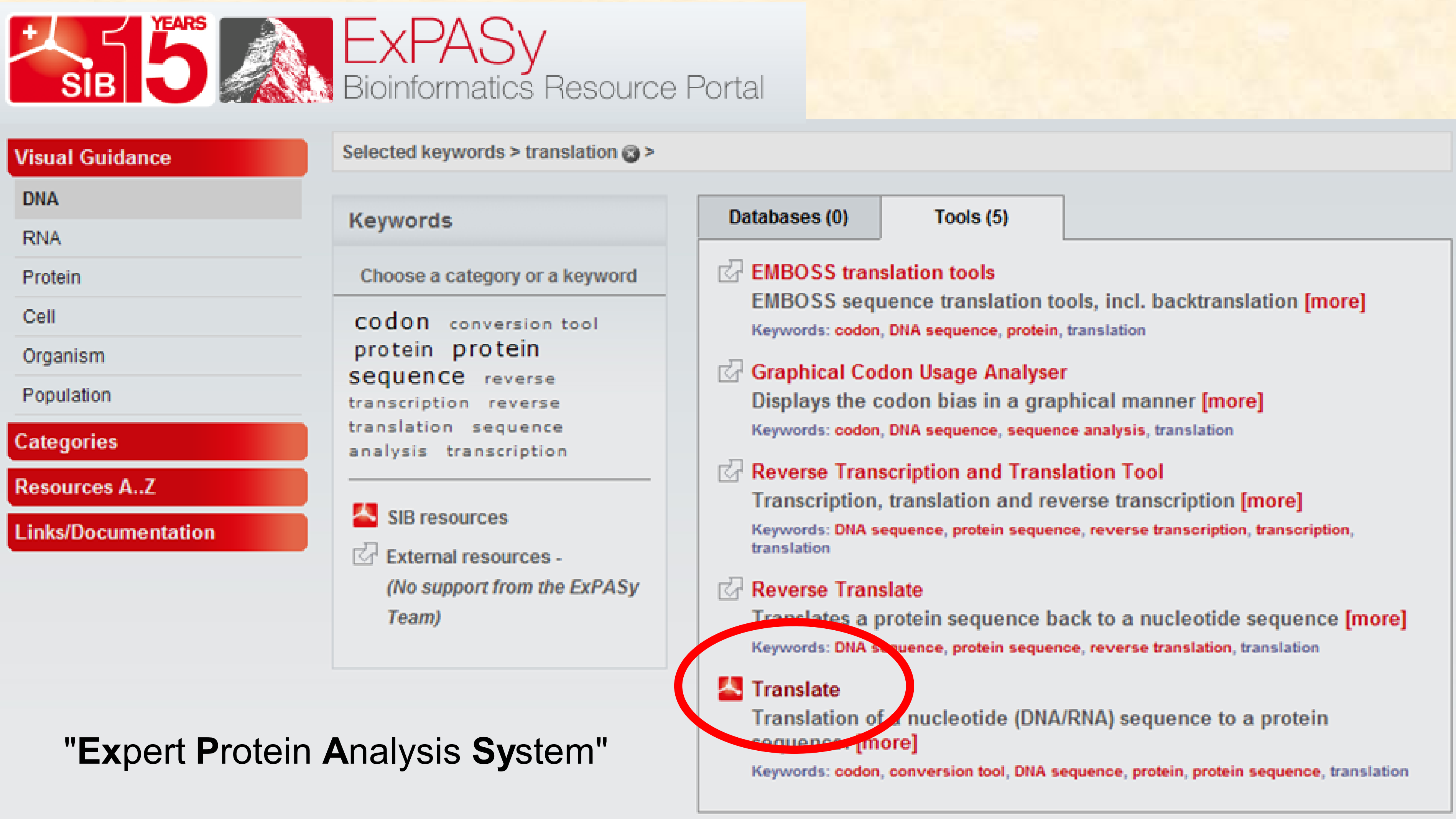
  http://web.expasy.org/translate/

- **ORF Finder (NCBI)**

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

ExPASy

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



"**Ex**pert **P**rotein **A**nalysis **Sy**stem"

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
TAGCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCG
CGGAAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAA
GTGCGTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGA
TCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATC
GTGATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACC
AAACAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAA
AGCGATCTTCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGG
CCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATT
CGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGA
TACCGTGTGGCTGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTC
TGCAGTGGCCGATTACCTAATGGG
```

Output format: Verbose ("Met", "Stop", spaces between residues)

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
V C W **Stop** L W **Met** P L P C **Stop** A P I R K P A V I R P P R P **Stop** L **Met** V A T C **Met** L L A R A **Met** P R S W A I T I A V C L P V **Stop** A R V I S C I C A K P R W R C A R K
Stop A C C L F A L P **Stop** K **Met** P A L L P R S N W K C V **Met** P P P P F R **Met** R **Met** I C C I R A V V R **Stop** K I I I G A A **Met** C W R R A R P P V P P I L R C A I V **Met** A P
Stop A V I F V G K P A L K L R A A S R I P N S R A L N R A A I A **Met** A T L A C R R I P P L K R S S **Met** R T R R I V R I **Stop** N C L L **Met** **Met** R R N R P P P L W V T A K **Met**
V C V C L P **Stop** I A K V V K F V L K R A R T A V R A R P **Met** P V W R R **Stop** A R A I P C G W A G W A R K **Met** V P **Met** R I I **Met** **Met** A L L F C S G R L P N G

5'3' Frame 2
Y A G D C G C R Y P A E R L S G S Q P **Stop** S G R P D R D **Stop** W S P P V C C **Stop** P G R C R A A G P **Stop** R **Stop** P S V Y R S E P G **Stop** S A A S A R N R A G A A R
G S E R A V Y S L C P E R C R H C C P D R T G S A **Stop** C R H R R S G C G **Stop** S A A S E L S S A E R S L L A Q R C A G G G R D H L Y R R F C G V R S **Stop** W H R E
R L F S L G N Q H **Stop** N C G Q P A G Y Q T A G L **Stop** T E Q R S Q W Q L **Stop** P A A E Y R L **Stop** S D L L C E R G G S S G S E T V Y **Stop** **Stop** C A G T G R H L C G
Stop Q R R W C A S V Y P E **Stop** Q R W **Stop** N S Y **Stop** S E R E R P S E R D R C P S G A A E R G R Y R V A G L A G R G R W C R C G L **Stop** **Stop** W H C Y S A V A D
Y L **Met**

5'3' Frame 3
Met L V I V D A V T L L S A Y P E A S R D P A A P T V I D G R H L Y V V S P G D A A Q L G H N D S R L F T G L S P G D Q L H L R E T A L A L R A E V S V L F I R F A L K D
A G I V A P I E L E V R D A A T A V P D A D D L L H P S C R P L K D H Y W R S D V L A A G A T T C T A D F A V C D R D G T V S G Y F R W E T S I E I A G S Q P D T K Q P
G F K P S S D R N G N F S L P P N T A F K A I F Y A N A A D R Q D L K L F I D D A P E P A A T F V G N S E D G V R L F T L N S K G G K I R I E A S A N G R Q S A T D A R L
A P L S A G D T V W L G W L G A E D G A D A D Y N D G I V I L Q W P I T **Stop** W

3'5' Frame 1
P I R **Stop** S A T A E **Stop** Q C H H Y N P H R H H L P R P A S P A T R Y R P R S A A P D G H R S R S D G R S R S L Q Y E F Y H L C Y S G **Stop** T D A H H L R C Y P Q R W
R P V P A H H Q **Stop** T V S D P D D P P R S H R R S L **Stop** R R Y S A A G **Stop** S C H C D R C S V **Stop** S P A V W Y P A G C P Q F Q C W F P N E N N R S R C H H D R
T P Q N R R Y R W S R P P P A H R C A N N D L S A D D S S D A A D H P H P E R R W R H H A L P V R S G Q Q C R H L S G Q S E **Stop** T A R S L P R A A P A R F R A D
A A D H P G S D R **Stop** T D G Y R Y G P A A R H R P G **Stop** Q H T G G D H Q S R S G R P D H G W L P D R R S A G **Stop** R H P Q S P A Y

3'5' Frame 2
P L G N R P L Q N N N A I I I I R I G T I F R A Q P A Q P H G I A R A Q R R Q T G I G R A L T A V R A R F N T N F T T F A I Q G K Q T H T I F A V T H K G G G R F R R I I N K Q F
Q I L T I R R V R I E D R F K G G I R R Q A K V A I A I A A R F K A R L F G I R L A A R N F N A G F P T K I T A H G A I T I A H R K I G G T G G R A R R Q H I A A P I **Met** I F Q R T
T A R **Met** Q Q I I R I R N G G G G I T H F Q F D R G N N A G I F Q G K A N K Q H A H F R A Q R Q R G F A Q **Met** Q L I T R A Q T G K Q T A I V **Met** A Q L R G I A R A N N I Q V
A T I N H G R G G R I T A G F R I G A Q Q G N G I H N H Q H

3'5' Frame 3
H **Stop** V I G H C R I T **Met** P S L **Stop** S A S A P S S A P S Q P S H T V S P A L S G A R R A S V A L **Stop** R P F A L A S I R I L P P L L F R V N R R T P S S L L P T K V A A
G S G A S S I N S F R S **Stop** R S A A F A **Stop** K I A L K A V F G G R L K L P L R S L L G L K P G C L V S G W L P A I S **Met** L V S Q R K **Stop** P L T V P S R S H T A K S A
V Q V V A P A A S T S L R Q **Stop** **Stop** S F S G R Q L G C S R S S A S G T A V A A S R T S S S I G A T **Met** P A S F R A K R I N S T L T S A R S A S A V S R R C S **Stop** S
P G L R P V N R R L S L W P S C A A S P G L T T Y R W R P S I T V G A A G S R L A S G **Stop** A L S R V T A S T I T S I

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) :

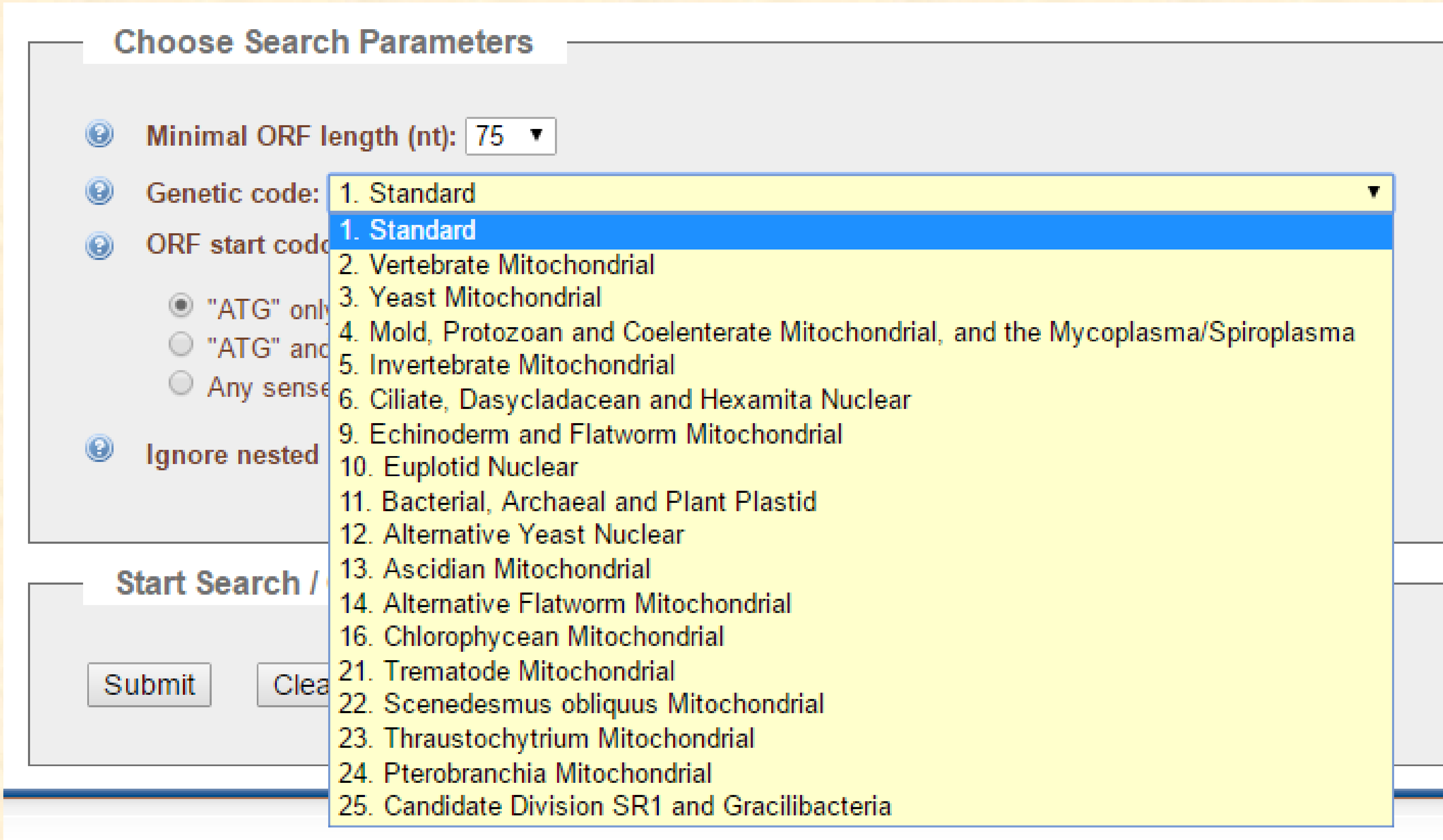
- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

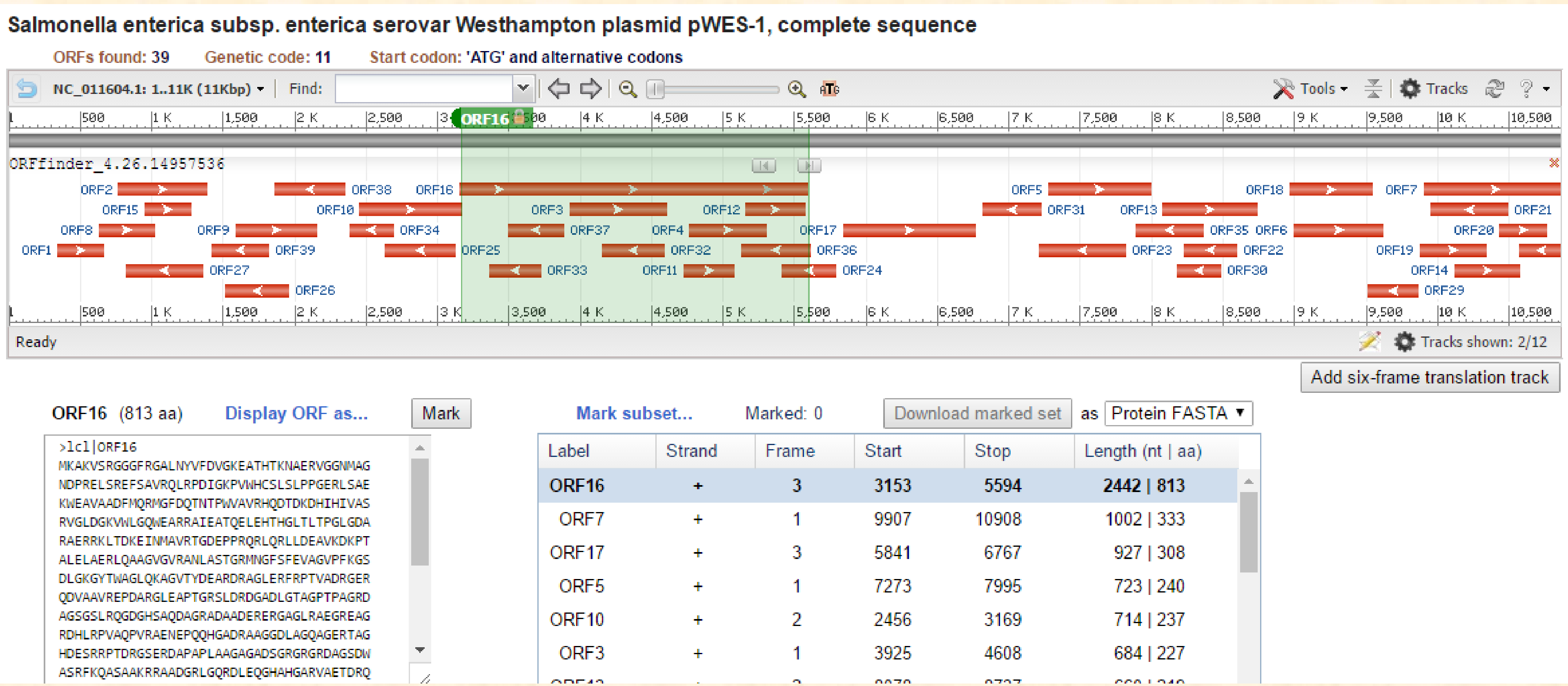
Enter accession number, gi, or nucleotide sequence in FASTA format:

From: **To:**

ORF Finder (NCBI)
https://www.ncbi.nlm.nih.gov/orffinder/



ORF Finder (NCBI)
https://www.ncbi.nlm.nih.gov/orffinder/



Prokaryotické geny

- **Velmi jednoduchý přístup k predikci genů**

  Zjednodušení vede k chybám, ale jejich množství je POMĚRNĚ MALÉ.

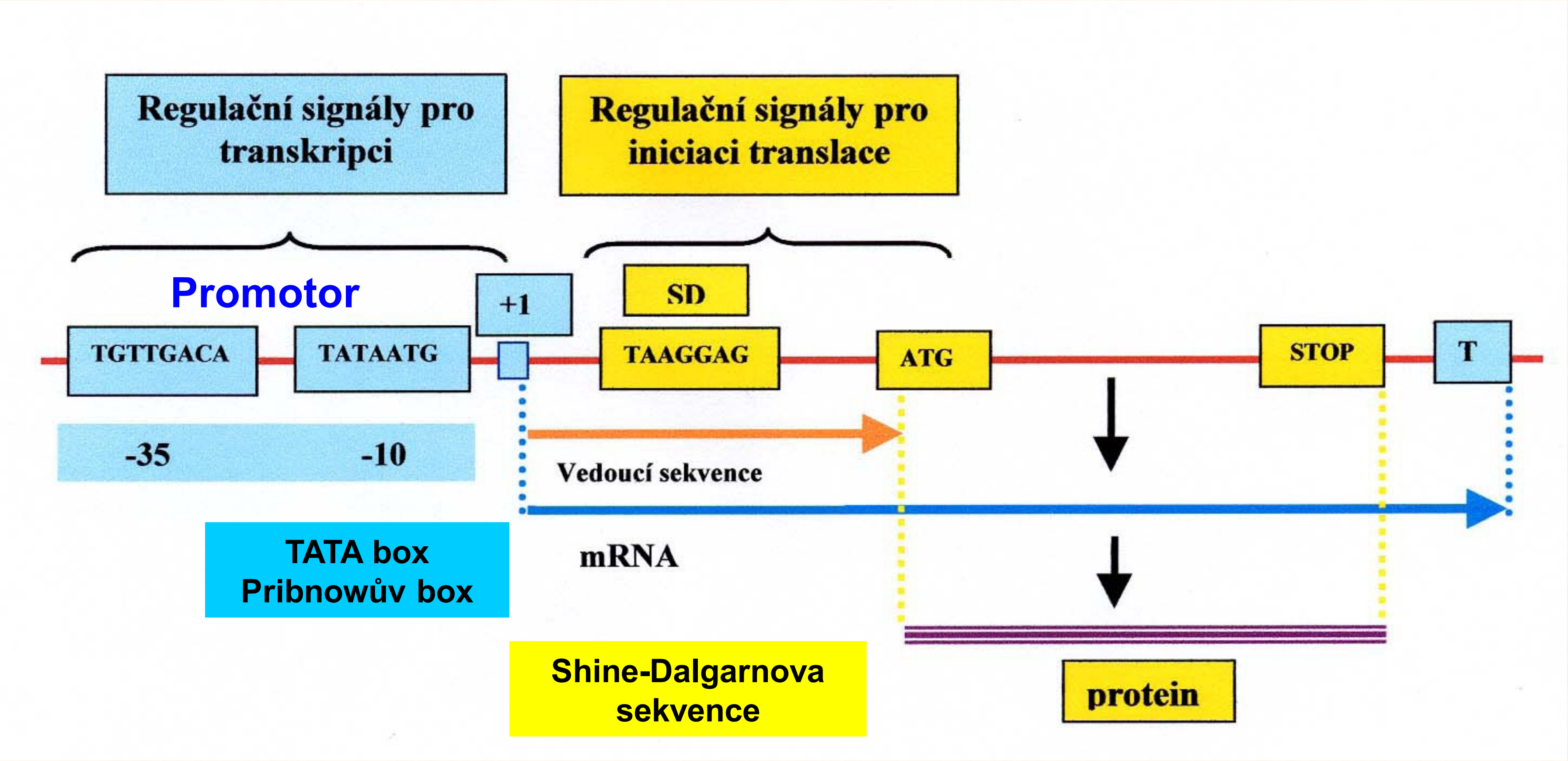
- **Chyby mohou vznikat při SEKVENCOVÁNÍ DNA.**

  Přidání/odstranění startovního a/nebo stop kodonu může vést ke ZKRÁCENÍ, PRODLOUŽENÍ nebo úplnému VYNECHÁNÍ genu.

Opravdu ORF kóduje protein?

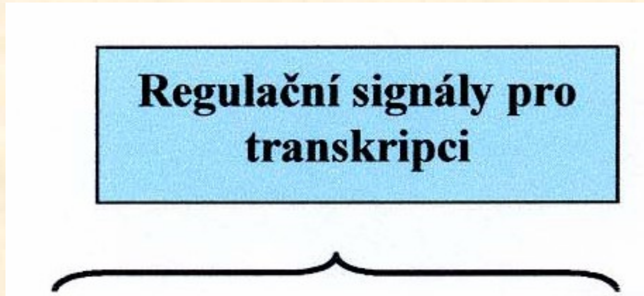
- ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání DATABÁZÍ pomocí ALIGNMENTU).
- ORF má typický obsah GC nebo frekvenci kodonů. Srovnání s charakteristickými vlastnostmi známých genů ze stejného organismu.
- Před ORF se nachází typické RBS (ribosome-binding site) nebo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

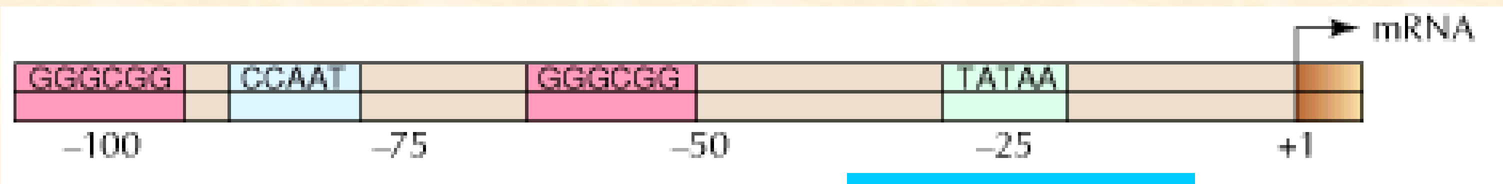
Translační a transkripční signální sekvence



Prokaryota

Translační a transkripční signální sekvence





GC box

GC box

TATA box
Hognessův box

Promotor RNA-polymerasy II

(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 HLYVVSPGDA AQLGHNDSRL FTGLSPGDQL
     HLRETALALR AEVSVLFIRF ALKDAGIVAP IELEVRDAAT AVPDADDLLH PSCRPLKDHY
     WRSDVLAAGA TTCTADFAVC DRDGTVSGYF RWETSIEIAG SQPDTKQPGF KPSSDRNGNF
     SLPPNTAFKA IFYANAADRQ DLKLFIDDAP EPAATFVGNS EDGVRLFTLN SKGGKIRIEA
     SANGRQSATD ARLAPLSAGD TVWLGWLGAE DGADADYNDG IVILQWPIT
//
```

Sequence in FASTA format

BLAST BLAST submission on ExPASy/SIB

Sequence analysis tools: ProtParam, ProtScale, Compute pI/Mw,

ScanProsite

Direct Submission to SWISS-MODEL

ORF Finder (NCBI)
https://www.ncbi.nlm.nih.gov/orffinder/

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

```
>lcl|ORF16
MKAKVSRGGGFRGALNYVFDVGKEATHTKNAERVGGNMAG
NDPRELSREFSAVRQLRPDIGKPVWHCSLSLPPGERLSAE
KWEAVAADFMQRMGFDQTNTPWVAVRHQDTDKDHIHIVAS
RVGLDGKVWLGQWEARRAIEATQELEHTHGLTLTPGLGDA
RAERRKLTDKEINMAVRTGDEPPRQRLQRLLDEAVKDKPT
ALELAERLQAAGVGVRANLASTGRMNGFSFEVAGVPFKGS
DLGKGYTWAGLQKAGVTYDEARDRAGLERFRPTVADRGER
QDVAAVREPDARGLEAPTGRSLDRDGADLGTAGPTPAGRD
AGSGSLRQGDGHSAQDAGRADAADERERGAGLRAEGREAG
RDHLRPVAQPVRAENEPQQHGADRAAGGDLAGQAGERTAG
HDESRRPTDRGSERDAPAPLAAGAGADSGRGRGRDAGSDW
ASRFKQASAAKRRAADGRLGQRDLEQGHAHGARVAETDRQ
```

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

Eukaryotické geny
Jednobuněčná eukaryota

- **Genomy jednobuněčných eukaryot se výrazně liší** (frekvence intronů, jak velká část genomu je tvořená 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

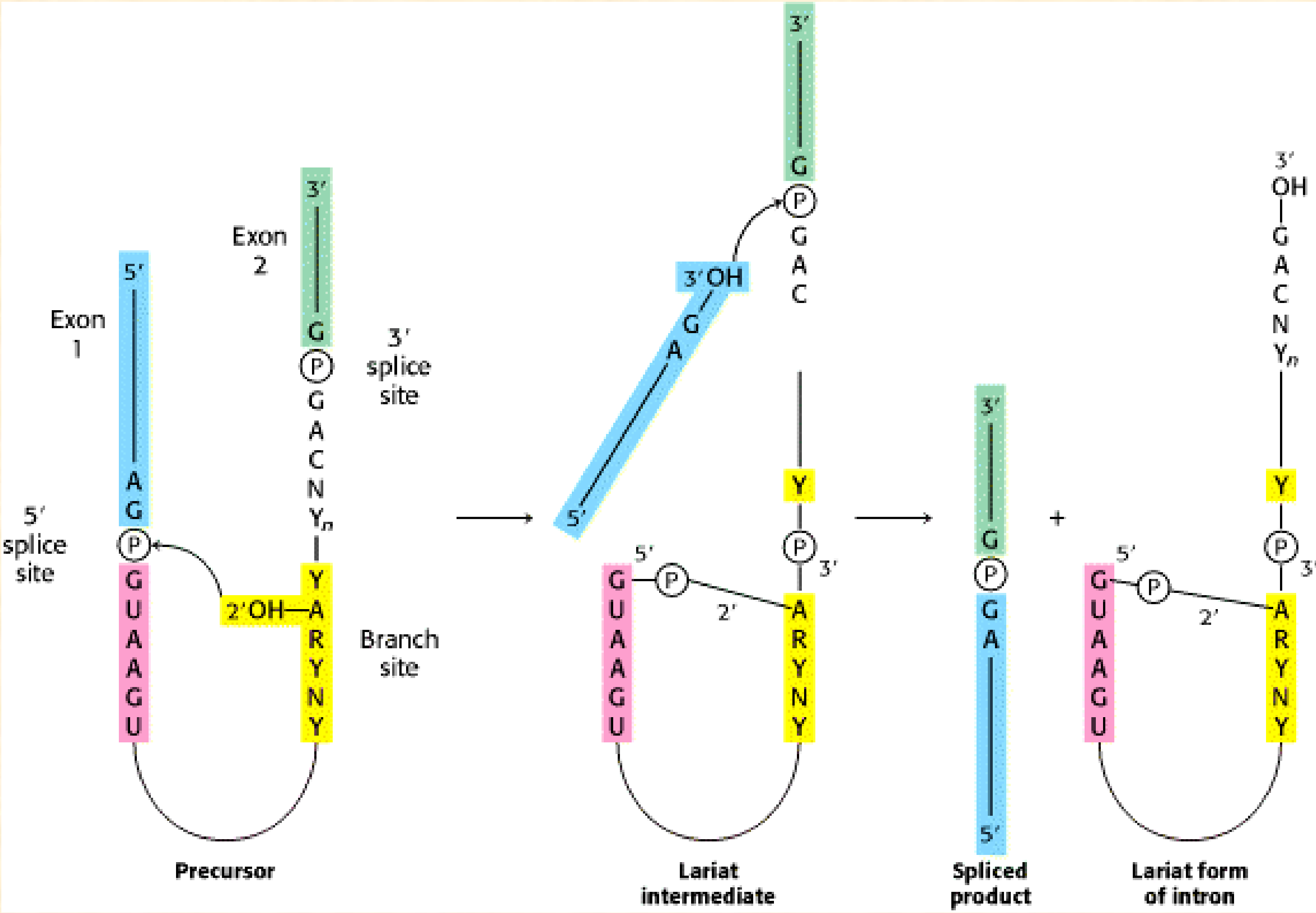
Eukaryotické geny
Mnohobuněčná eukaryota

- **Rozpoznání exonů/intronů**

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

- **Chyby při rozpoznávání exonů/intronů**

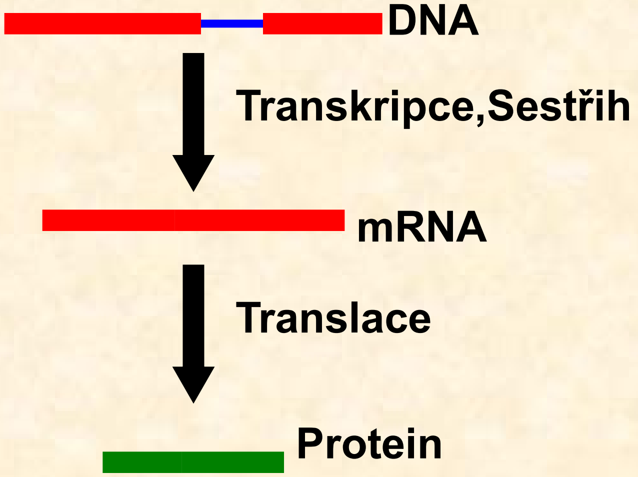
Velké množství chyb. Dlouhé introny – určeny jako intergenové úseky. Krátké intergenové useky – určeny jako introny.



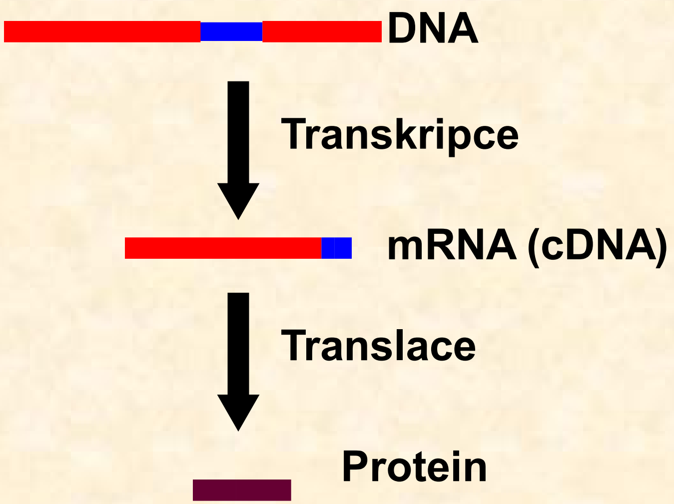
Splicing Mechanism Used for mRNA Precursors. The upstream (5′) exon is shown in blue, the downstream (3′) exon in green, and the branch site in yellow. Y stands for a purine nucleotide, R for a pyrimidine nucleotide, and N for any nucleotide. The 5′ splice site is attacked by the 2′-OH group of the branch-site 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)



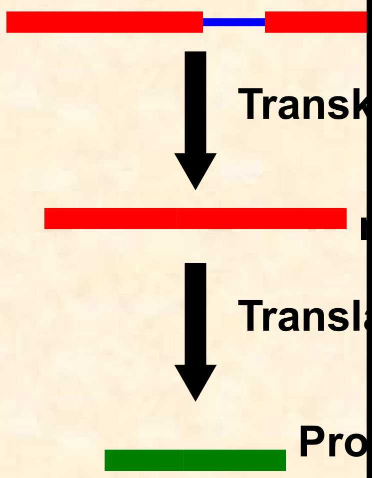
MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS
NCSSSMLSRC

MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVWs wdqvtylktt cyvngyftds ncsssmlsrc

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)

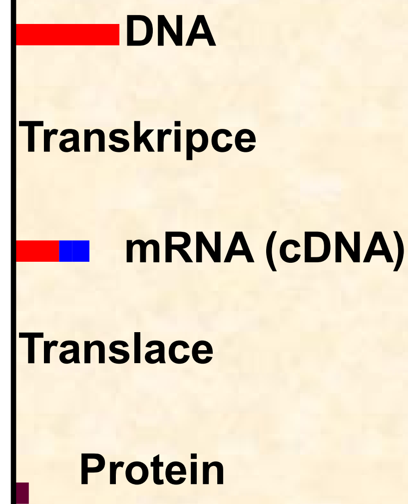


Chybná predikce intronu?

Alternativní sestřih?

Různé kmeny/životní podmínky/buněčný cyklus?

Chyba při přípravě cDNA knihovny?



MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVCS WDQVTYLKTT CYVNGYFTDS
NCSSSMLSRC

MADPEVEADG ELDLEKRASA QTCKIVNVDT
YVNCRYDAKL DAGAIFGFPK GEKLTFACWK
HGDCYNGVWs wdqvtylktt cyvngyftds ncsssmlsrc

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

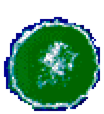
Gene Prediction in Bacteria, Archaea and Metagenomes

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

GeneScan

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

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

```
               \\|//
               (o o)
-. .-.   .-oOOo~(_)~oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.
||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\
|/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-'
```

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Algoritmy a nástroje pro identifikaci genů

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

*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.**