

# Predikce genů

**Pro zajímavost...**

**Důležité...**

# Molekulárně biologická data

- **Výkonné technologie:**

Automatické sekvencování

MALDI-TOF

NMR spektroskopie

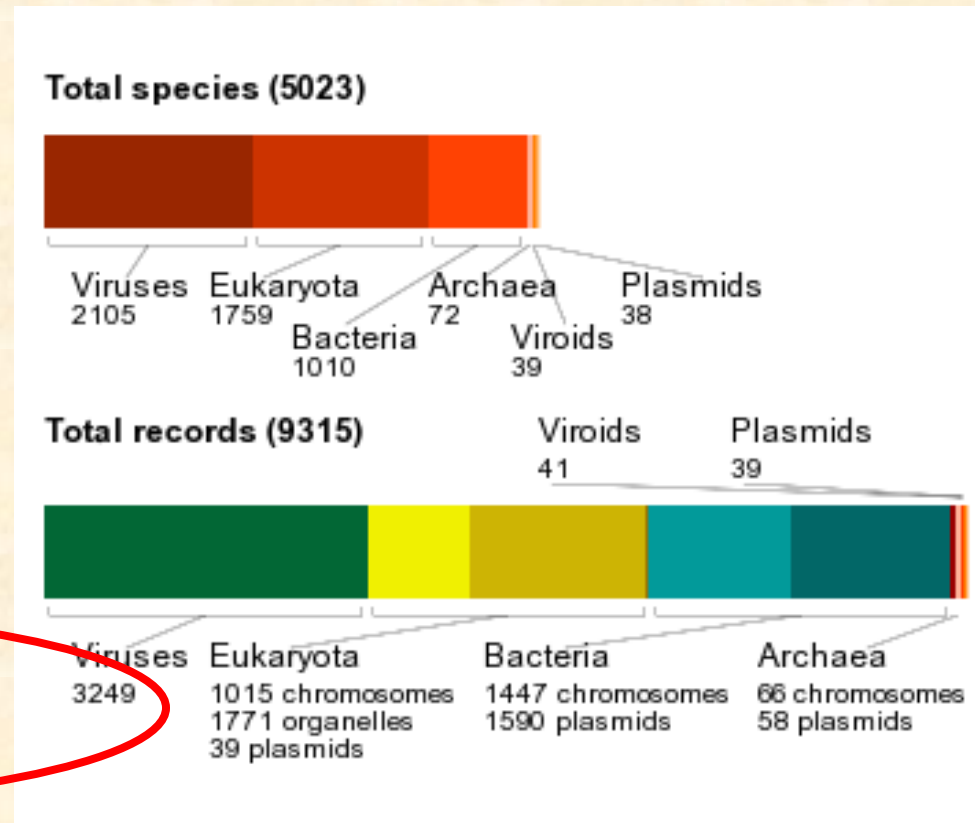
Proteinová krystalografie

**Výrazný nárůst množství biologických dat.**

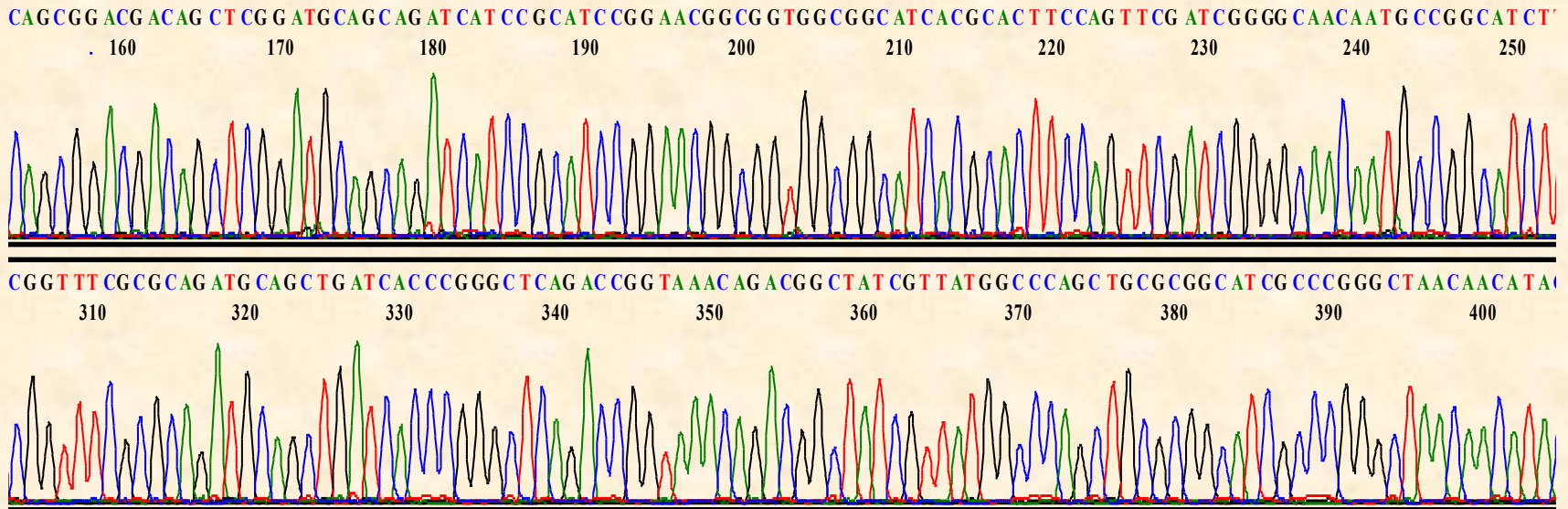
# Rozdělení molekulárně biologických databází

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

**Genomové zdroje**



# Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGATTTTCATGCTGGTTTCCCAACGAAAAAACCCTCACGGTGCCATCACGATCGCACACCGCAAATCGGCGG  
TACAGGTGGTCGCGCCCGCCGACACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC  
GGTGGCGGCATCACGCACCTTCCAGTTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACCTTCCGCGGCAGCGCC  
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCAGCTGCGGGCATCGCCCGGGCTAACAA  
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

GATAGCGTAATGATCGGCTGGCTGCCGATTTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCCGAAAAATCGGCGG  
TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC  
GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCGCGCGCAGCGCC  
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
TGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTTCGTCCGCTGAAAGATCATTATTGG
CGCAGCGATGTGCTGGCGGGCGGGCGGACCACCTGTACCGCCGATTTTTCGGTGTGCGATCGTGATGGCACCG
TGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATAACCAAACAGCCGGGCTT
TAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCTTCTATGCG
AACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTA
ACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAA
CGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATAACCGTGTGGCTGGGCTGGCTG
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://www.expasy.org/tools/dna.html>

- **ORF Finder (NCBI)**

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

# ExpASy

<http://www.expasy.org/tools/dna.html>

[Site Map](#)

Search ExpASy

[Contact us](#)

Search  for

The ExpASy Server requires Javascript to be fully functional. You may not see all the information available for this page ([More information](#)).



## ExpASy Proteomics Server

The ExpASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [References](#) / [Linking to ExpASy](#)).

[\[Databases\]](#) [\[Tools & Software\]](#) [\[Education & Services\]](#) [\[Links\]](#)  
[\[Announcements\]](#) [\[Mirror Sites\]](#) [\[Job openings\]](#)

### Databases

- [UniProt Knowledgebase \(Swiss-Prot and TrEMBL\)](#) - Protein knowledgebase
- [ViralZone](#) - Portal to viral UniProtKB/Swiss-Prot entries new
- [PROSITE](#) - Protein families and domains
- [SWISS-2DPAGE](#) - Two-dimensional polyacrylamide gel electrophoresis
- [World-2DPAGE Repository](#) - A public standards-compliant repository for gel-based proteomics data published in the literature
- [MIAPEGelDB](#) - A public repository for MIAPE Gel electrophoresis documents
- [ENZYME](#) - Enzyme nomenclature

### Tools and software packages

- [Proteomics and sequence analysis tools](#)
  - Identification and characterization (Aldente, FindMod, Popitam, Phenyx, pI/Mw, ProtParam...)
  - **DNA -> Protein**
  - Similarity searches (BLAST...)
  - Pattern and profile searches (ScanProsite...)
  - Post-translational modification and topology prediction
  - Primary structure analysis
  - Secondary and tertiary structure tools (Swiss-PdbViewer...)
  - Alignment and Phylogenetic analysis
- [Melanie / ImageMaster](#) - Software for 2-D PAGE analysis

# ExpASy

<http://www.expasy.org/tools/dna.html>

[Site Map](#)


[Search ExpASy](#)

[Contact us](#)

Search  for

The ExpASy Server requires Javascript to be fully functional. You may not see all the information available for this page ([More](#))

## DNA -> Protein

- [Translate](#)  - Translates a nucleotide sequence to a protein sequence
- [Transeq](#) - Nucleotide to protein translation from the EMBOSS package
- [Graphical Codon Usage Analyser](#) - Displays the codon bias in a graphical manner
- [BCM search launcher](#) - Six frame translation of nucleotide sequence(s)
  
- [Reverse Translate](#) - Translates a protein sequence back to a nucleotide sequence
- [\(Reverse\)-Transcription and Translation Tool](#)
  
- [Genewise](#) - Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors
- [LabOnWeb](#) - Elongation, expression profiles and sequence analysis of ESTs using Compugen LEADS clusters

- [UniProt Knowledgebase \(Swiss-Prot and TrEMBL\)](#) - Protein knowledgebase
- [ViralZone](#) - Portal to viral UniProtKB/Swiss-Prot entries new
- [PROSITE](#) - Protein families and domains
- [SWISS-2DPAGE](#) - Two-dimensional polyacrylamide gel electrophoresis
- [World-2DPAGE Repository](#) - A public standards-compliant repository for gel-based proteomics data published in the literature
- [MIAPEGelDB](#) - A public repository for MIAPE Gel electrophoresis documents
- [ENZYME](#) - Enzyme nomenclature

- [Proteomics and sequence analysis tools](#)
  - [Identification and characterization \(Aldente, FindMod, Popitam, Phenyx, p/Mw, ProtParam...\)](#)
  - [DNA -> Protein](#)
  - [Similarity searches \(BLAST...\)](#)
  - [Pattern and profile searches \(ScanProsite...\)](#)
  - [Post-translational modification and topology prediction](#)
  - [Primary structure analysis](#)
  - [Secondary and tertiary structure tools \(Swiss-PdbViewer...\)](#)
  - [Alignment and Phylogenetic analysis](#)
- [Melanie / ImageMaster](#) - Software for 2-D PAGE analysis

# ExPASy

<http://www.expasy.org/tools/dna.html>

**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:  ▾

or

# Translate Tool - Results of translation

## 5'3' Frame 1

V CW **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 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 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

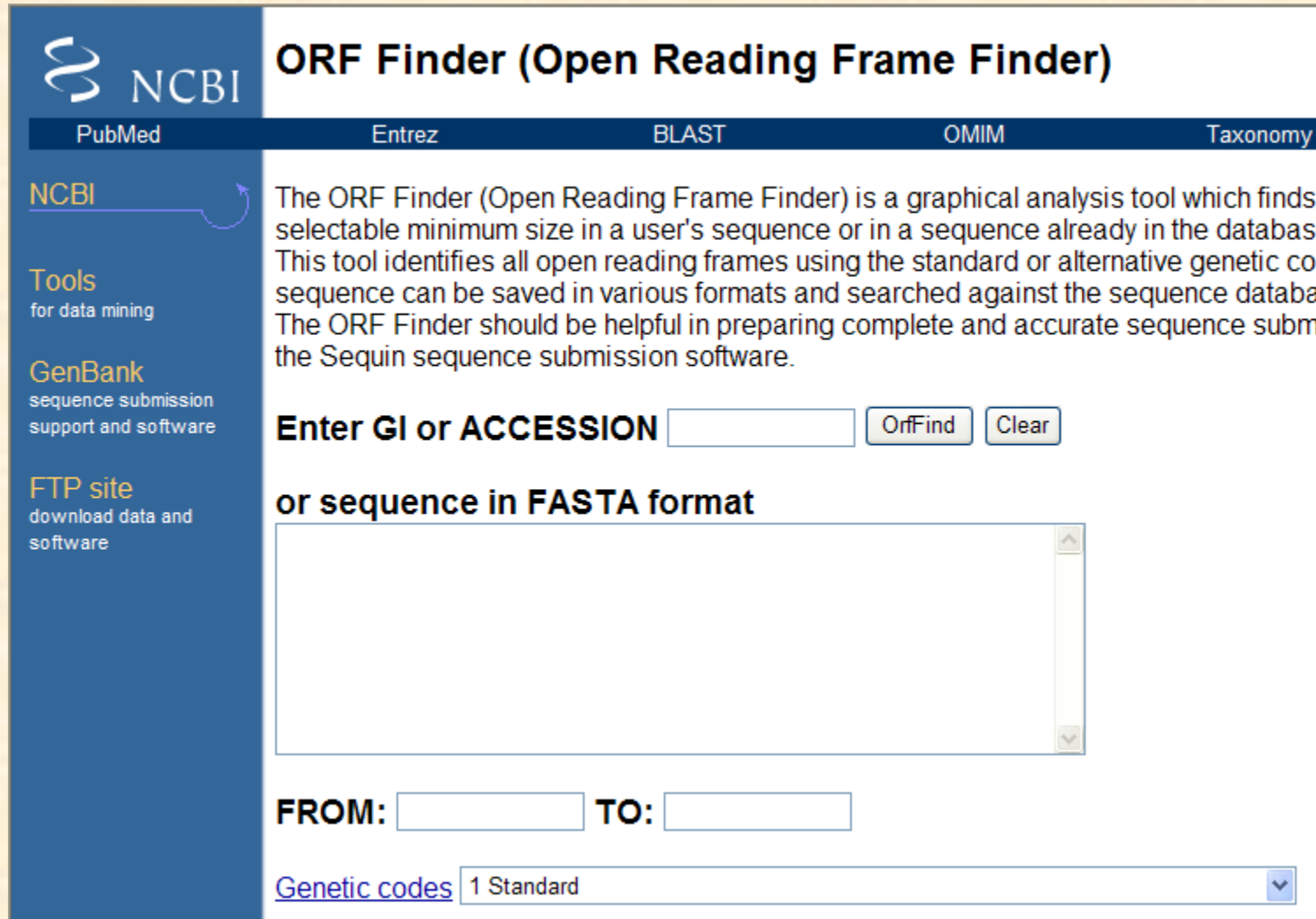
P L G N R P L Q N N N A I I I I R I G T I F R A Q P A Q P H G I A R A Q R R Q T G I G R A L T A V R A R F N T N F T T F A I Q G K Q T H T I F A V T H K G G G R F R R I I N K Q F Q I L T I R R V R I E D R F K G G I R R Q A K V A I A I A A R F K A R L F G I R L A A R N F N A G F P T K I T A H G A I T I 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 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)

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



The screenshot shows the NCBI ORF Finder web interface. At the top left is the NCBI logo. A navigation bar contains links for PubMed, Entrez, BLAST, OMIM, and Taxonomy. A left sidebar lists 'NCBI', 'Tools for data mining', 'GenBank sequence submission support and software', and 'FTP site download data and software'. The main content area is titled 'ORF Finder (Open Reading Frame Finder)' and contains a descriptive paragraph. Below the text are input fields for 'GI or ACCESSION' and 'or sequence in FASTA format', along with 'OrfFind' and 'Clear' buttons. At the bottom, there are 'FROM:' and 'TO:' input fields, and a 'Genetic codes' dropdown menu set to '1 Standard'.

**ORF Finder (Open Reading Frame Finder)**

PubMed Entrez BLAST OMIM Taxonomy

**NCBI**

**Tools**  
for data mining

**GenBank**  
sequence submission  
support and software

**FTP site**  
download data and  
software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic code. The results can be saved in various formats and searched against the sequence database. The ORF Finder should be helpful in preparing complete and accurate sequence submissions for the Sequin sequence submission software.

**Enter GI or ACCESSION**

**or sequence in FASTA format**

**FROM:**  **TO:**

Genetic codes 1 Standard



# ORF Finder (NCBI)

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

**ORF Finder (Open Reading Frame Finder)**

PubMed Entrez BLAST OMIM Taxonomy

**NCBI**

**Tools**  
for data mining

**GenBank**  
sequence submission support and software

**FTP site**  
download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic code. The sequence can be saved in various formats and searched against the sequence database. The ORF Finder should be helpful in preparing complete and accurate sequence submissions for the Sequin sequence submission software.

**Enter GI or or sequence**

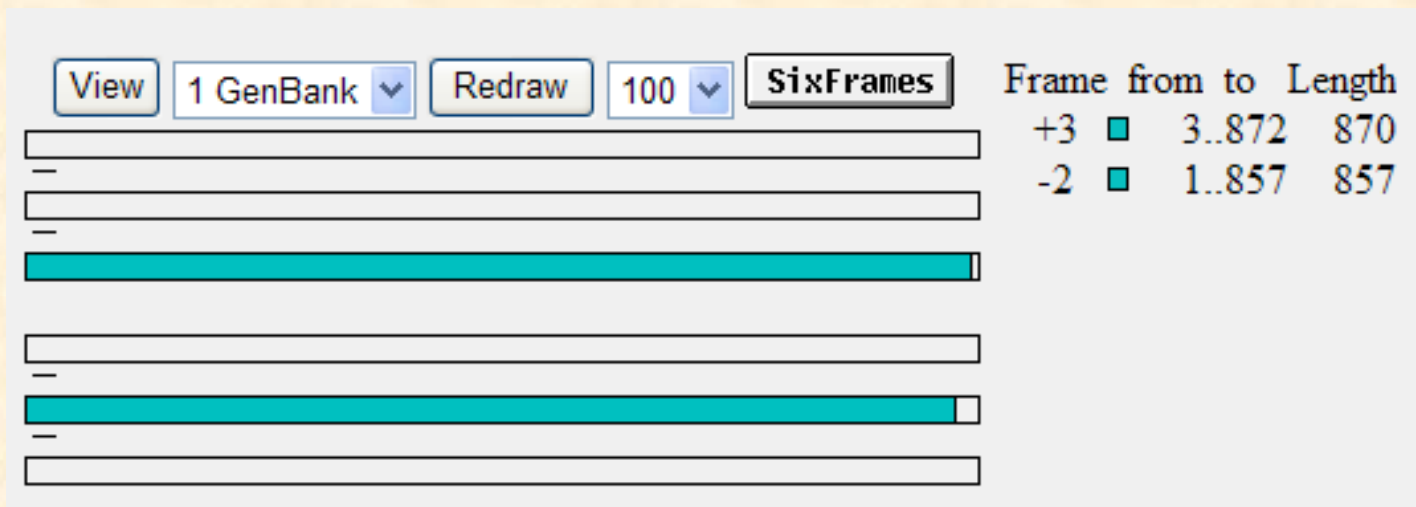
**FROM:**

Genetic codes

- [The Standard Code](#)
- [The Vertebrate Mitochondrial Code](#)
- [The Yeast Mitochondrial Code](#)
- [The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code](#)
- [The Invertebrate Mitochondrial Code](#)
- [The Ciliate, Dasycladacean and Hexamita Nuclear Code](#)
- [The Echinoderm and Flatworm Mitochondrial Code](#)
- [The Euplotid Nuclear Code](#)
- [The Bacterial and Plant Plastid Code](#)
- [The Alternative Yeast Nuclear Code](#)
- [The Ascidian Mitochondrial Code](#)
- [The Alternative Flatworm Mitochondrial Code](#)
- [Blepharisma Nuclear Code](#)
- [Chlorophycean Mitochondrial Code](#)
- [Trematode Mitochondrial Code](#)
- [Scenedesmus Obliquus Mitochondrial Code](#)
- [Thraustochytrium Mitochondrial Code](#)

# ORF Finder (NCBI)

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



## 5' Frame 3

**Met**LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDA AQLGHNDSRLFTGLSPGDQLHLRETALALRAEVS VLFIRFALKDAGIVAPI  
ELEVRDAATAVPDADDLLHPSCRPLKDHYWRS DVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGS QPDTKQPGFKPSSDRNGN  
FSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNS EDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLG A  
EDGADADYNDGIVILQWPIT **Stop**W

## 3' Frame 2

PLGNRPLQNNNAIIIRIGTIFRAQPAQPHGIARAQRRTGIGRALTAVRARENTNFTTFAIQGKQTH TIFAVTHKGGGRFRRIINKQFQILT I  
RRVRIEDRFKGGIRRQAKVAIAIAARFKARLFGIRLAARNFNAGFPTKITAHGAITIAHRKIGGTGGRARRQHIAAPI **Met**IFQRTTAR **Met**QQII  
RIRNGGGGITHFQFDRGNNAGIFQGKANKQHAHFRAQRQRGFAQ **Met**QLITRAQTGKQTAIV **Met**AQLRGIARANNIQVATINHGRGGRITA  
GFRIGAQQGNGIHNHQH

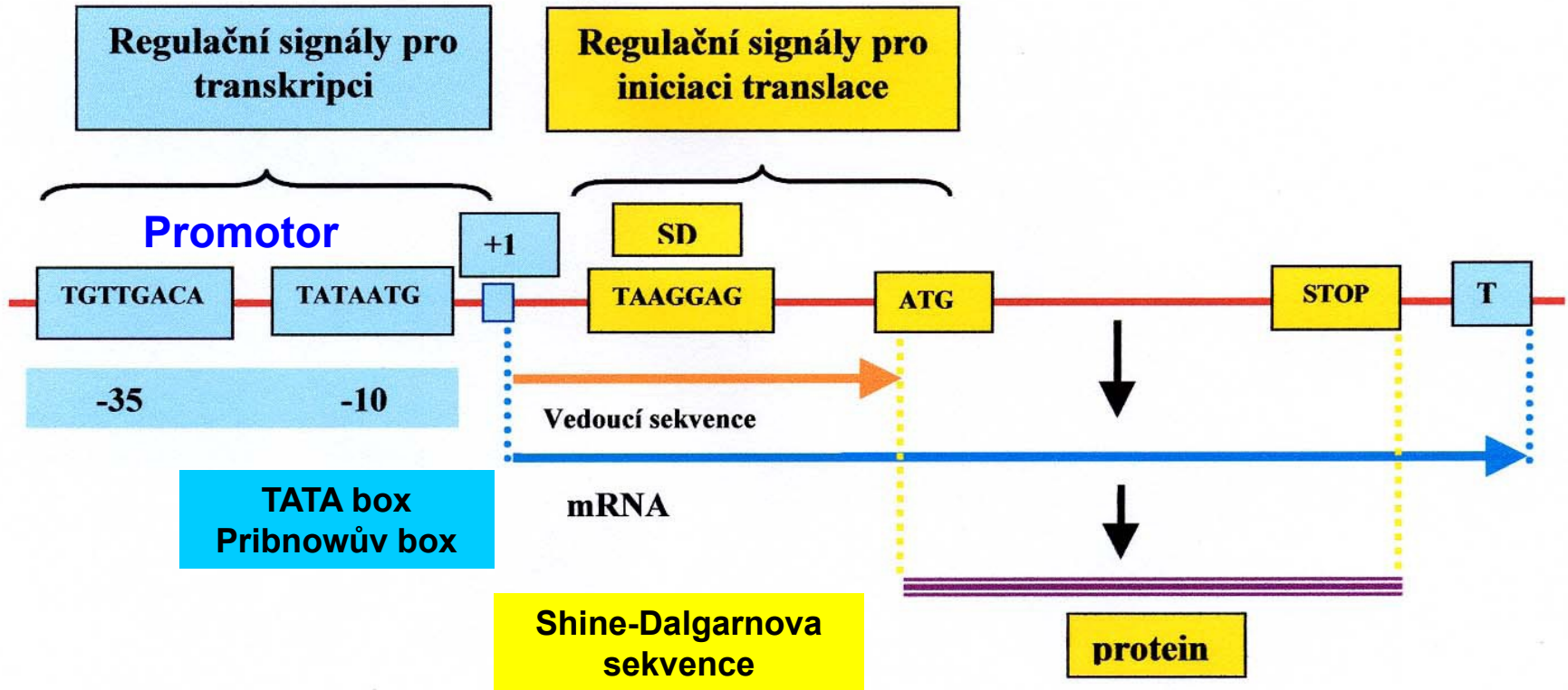
# Prokaryotické geny

- **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?

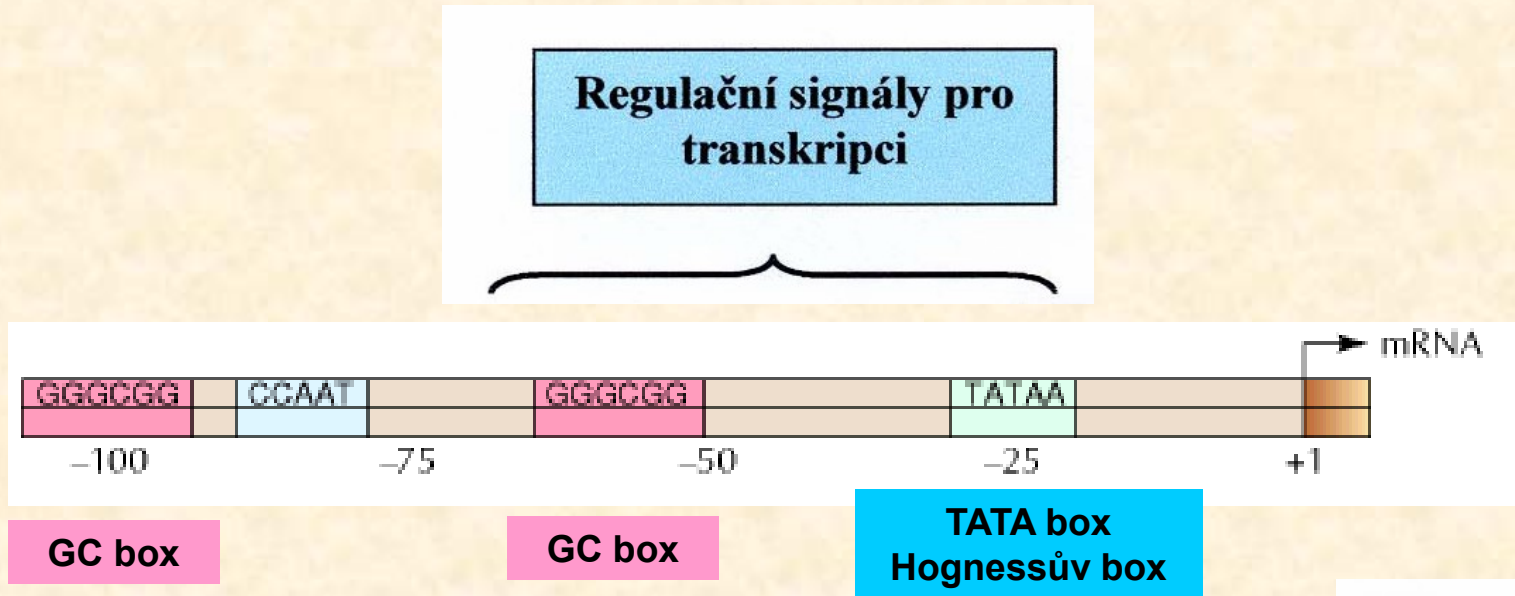
- **ORF kóduje protein, který je podobný již dříve popsanému proteinu** (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**).
- **ORF má typický obsah GC nebo frekvenci kodonů.** Srovnání s charakteristickými vlastnostmi známých genů ze stejného organismu.
- **Před ORF se nachází typické RBS (ribosome-binding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

# Translační a transkripční signální sekvence



## Prokaryota

# Translační a transkripční signální sekvence



Promotor RNA-polymerasy II



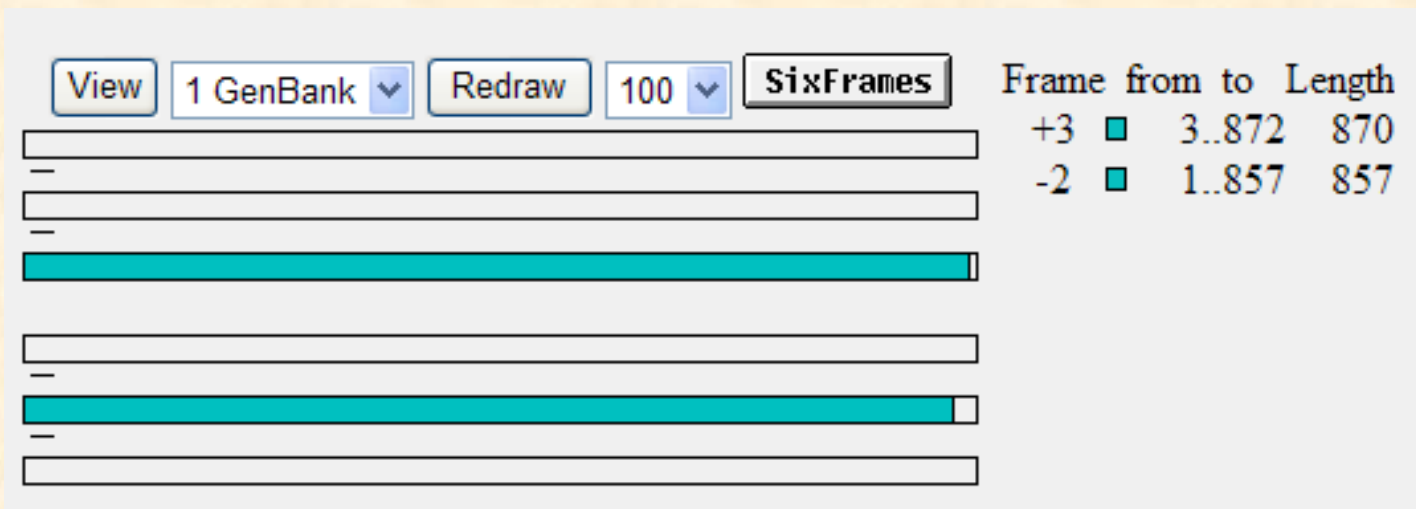
# Eukaryota

# 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.**

# ORF Finder (NCBI)

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



## 5' Frame 3

**Met**LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKDAGIVAPI  
ELEVRDAATAVPDADDLLHPSCRPLKDHYWRSVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGN  
FSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSGDVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLG  
AEDGADADYNDGIVILQWPIT **Stop** W

## 3' Frame 2

PLGNRPLQNNNAIIIRIGTIFRAQPAQPHGIARAQRRQTGIGRALTAVRARFNTNFTTFAIQGKQTHTIFAVTHKGGGRFRRIINKQFQILT  
RRVRIEDRFKGGIRRQAKVAIAIAARFKARLFGIRLAARNFNAGFPTKITAHGAIHIAHRKIGGTGGRARRQHIAAPI **Met**IFQRTTAR **Met**QQII  
RIRNGGGGITHFQFDRGNNAGIFQGKANKQHAHFRAQRQRGFAQ **Met**QLITRAQTGKQTAIV **Met**AQLRGIARANNIQVATINHGRGGRITA  
GFRIGAQQGNGIHNHQH



# ORF Finder (NCBI)

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

Program **blastp** Database **nr** **BLAST**  with parameters **Cognitor**

View 1 GenBank Redraw 100 SixFrames

	Frame from to	Length
+3	3..872	870
-2	1..857	857

Length: 289 aa

```
3 atgctggtgattgtggatgccgttacacctgctgagcgcctatccg
M L V I V D A V T L L S A Y P
48 gaagccagccgtgatccggccgccccgaccgtgattgatggtcgc
E A S R D P A A P T V I D G R
93 cacctgtatgtttagccccggcgatgccgcgcagctgggccat
H L Y V V S P G D A A Q L G H
138 aacgatagccgtctgtttaccggtctgagcccggtgatccgtg
N D S R L F T G L S P G D Q L
183 catctgcgcgaaaccgcgctggcgctgcccgcggaagtgagcgtg
H L R E T A L A L R A E V S V
228 ctgtttattcgctttgccctgaaagatgccggcattgttgccccg
L F I R F A L K D A G I V A P
273 atcgaactggaagtgcgtgatgccgccaccgccgttccggatgcg
I E L E V R D A A T A V P D A
318 gatgatctgctgcacccgagctgtcgtccgctgaaaagatcattat
D D L L H P S C R P L K D H Y C
363 tggcgagcagatgtctggcgcgccgagaccacctgtgaccgcc
W R S D V L A A G A T T C T A
408 gattttgcggtgtgcgatcgtgatggcaccgtgagcggttat
D F A V C D R D G T V S G Y F
453 cgttgggaaaccagcattgaaattgcccggcagccagccggatcc
R W E T S I E I A G S Q P D T
498 aaacagccgggctttaaccgagcagcagatcgcaatggcaactt
K Q P G F K P S S D R N G N F
543 agcctgccgccaataaccgctttaaagcgatcttctatgcgaac
S L P P N T A F K A I F Y A N
588 gcggcgagatcgtcaggatctgaaactgtttattgatgagccg
A A D R Q D L K L F I D D A P
633 gaaccggccgccaccttggggtaaccagcgaagatgggtgtgctg
E P A A T F V G N S E D G V R
678 ctgtttaccctgaatagcaaaaggtggtaaaattcgattgaaagcg
L F T L N S K G G K I R I E A
723 agcgcgaaccggcgtcagagcgcgaccgatgccgctctggcgccg
S A N G R Q S A T D A R L A P
768 ctgagcggggcgataaccgtgtggctgggctgggctgggcgcgaa
L S A G D T V W L G W L G A E
813 gatggtgccgatgcggattataatgatggcattgttattctgcag
D G A D A D Y N D G I V I L Q
858 tggccgattacctaa 872
W P I T *
```

# 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**.



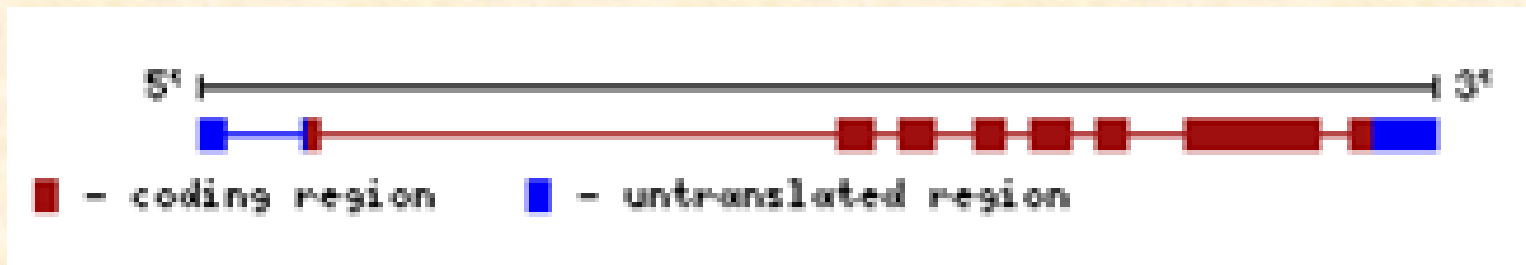
Glyceraldehyd-3-fosfát-dehydrogenasa  
*Candida albicans*

# 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

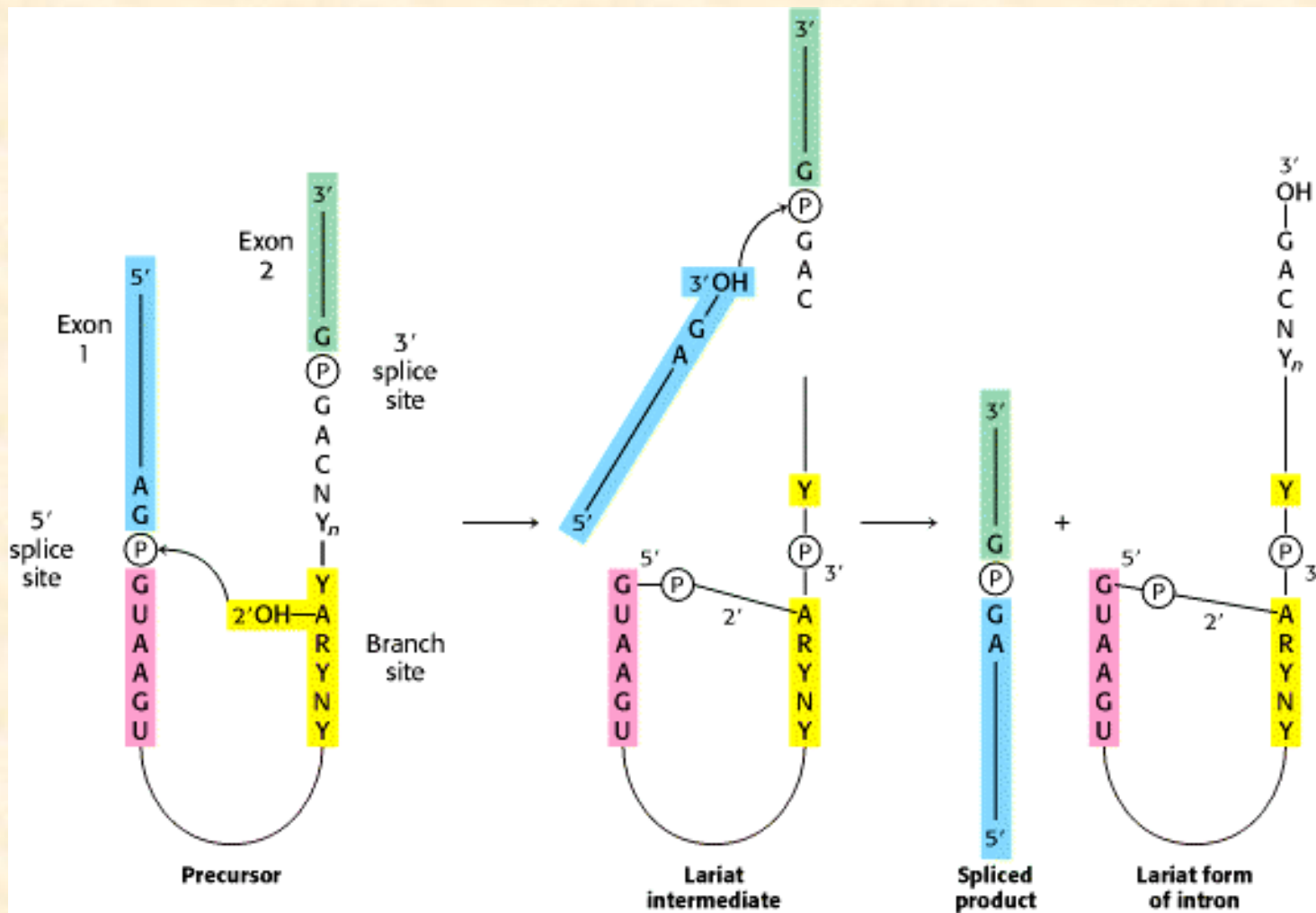
## Mnohobuněčná eukaryota

- **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.]

# 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 probablistický 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/glimmer/glmr_form.html">http://www.tigr.org/tdb/glimmer/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.