

Molekulárně biologická data

- **Výkonné technologie:**

Automatické sekvenování

MALDI-TOF

NMR spektroskopie

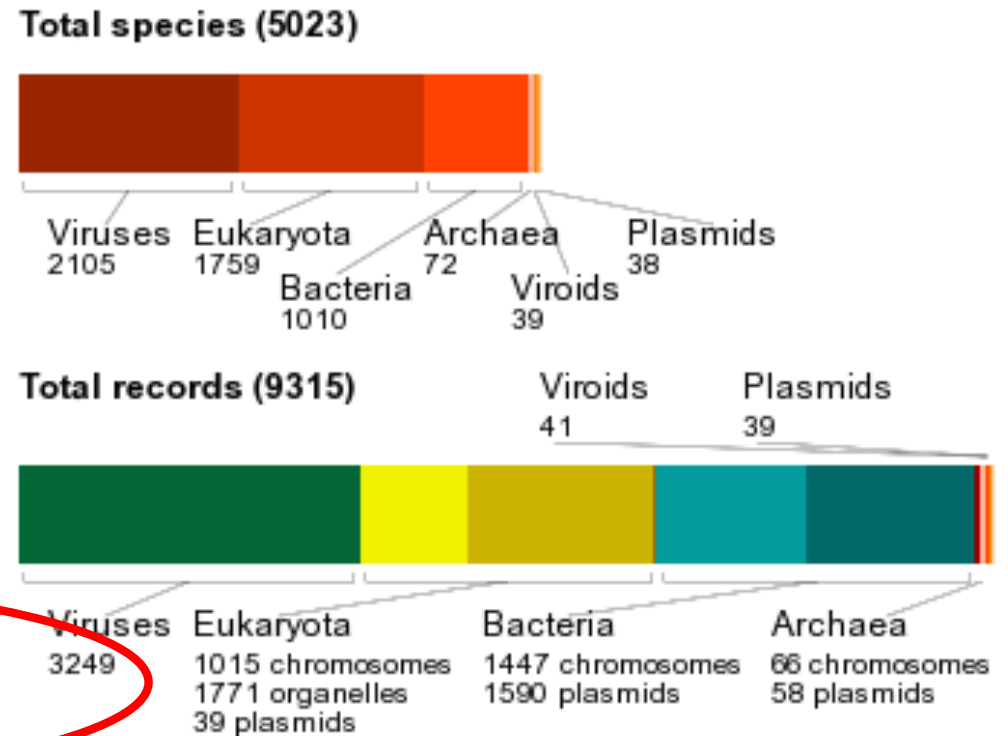
Proteinová krystalografie

Výrazný nárůst množství biologických dat.

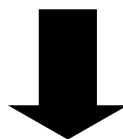
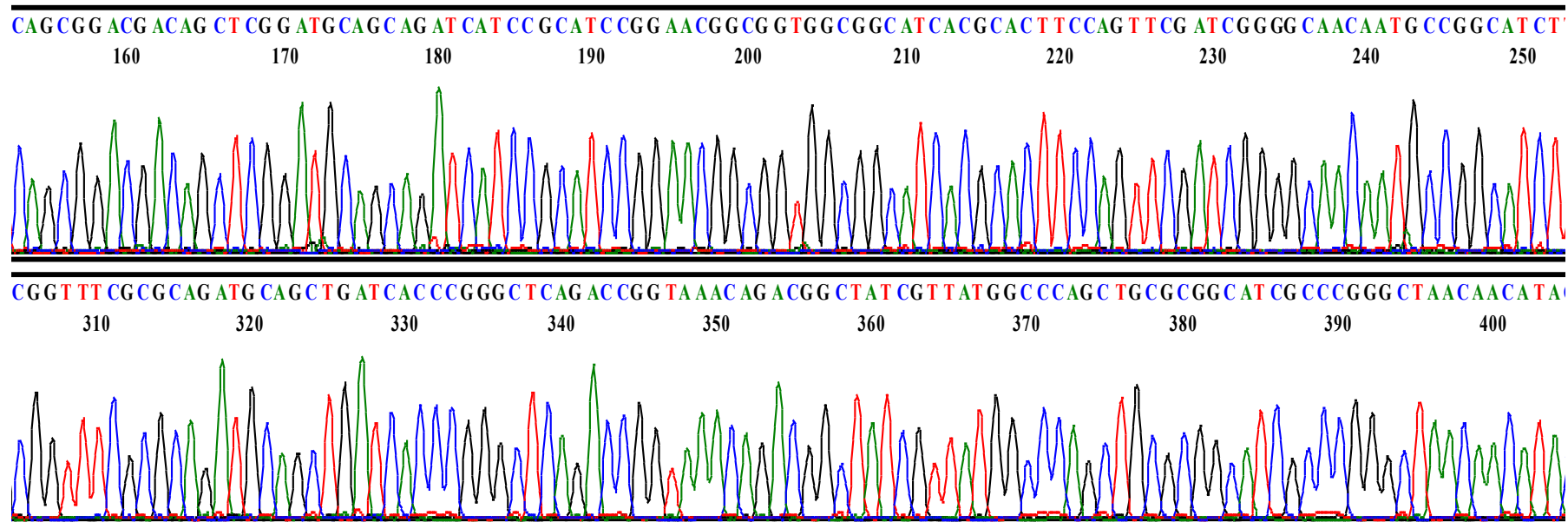
Rozdělení molekulárně biologických databází

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

Genomové zdroje



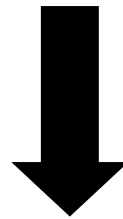
Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGATTTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG
TACAGGTGGTCGCGCCC GCCAGCACATCGCTGCGCCAATAATGATCTTT CAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC
GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGGCAGCGCC
AGCGCGGTTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

GATAGCGTAATGATCGGCTGGCTGCCGCAATTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG
TACAGGTGGTCGCGCCCGCCGAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC
GGTGGCGGCATCACGCACCTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACCTCCGCGCGCAGCGCC
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCAGCTGCGCGGCATCGCCCGGGCTAACAA
CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

„Syrové“ sekvence DNA



Identifikace a anotace genů a proteinů

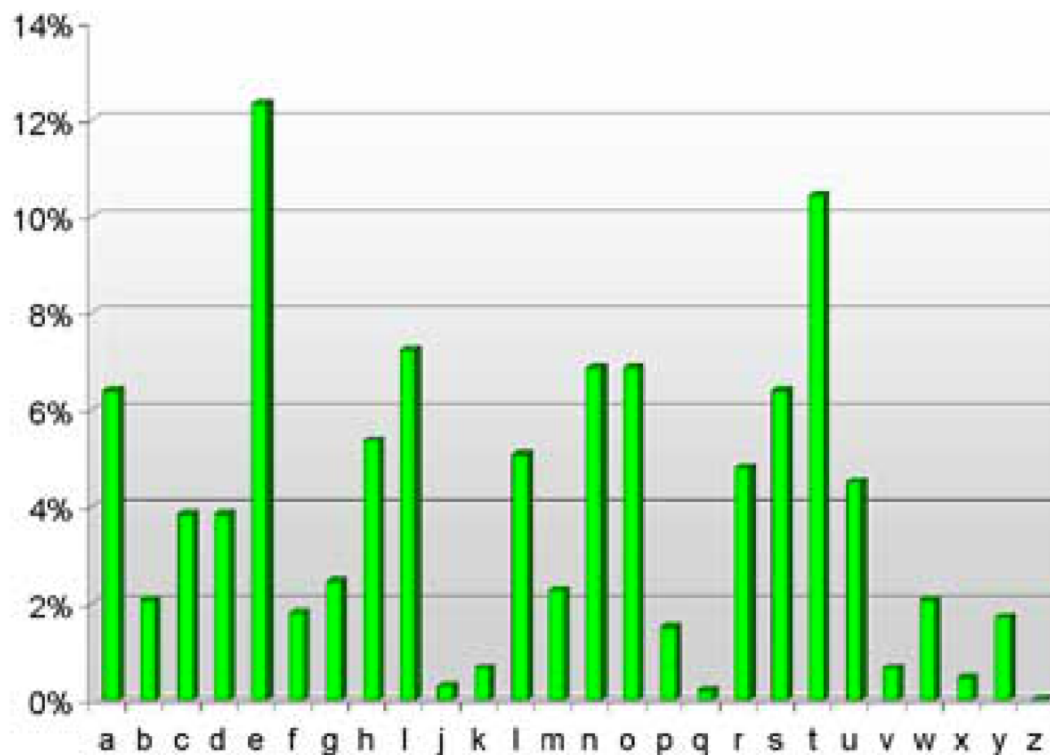
- Posloupnost písmen může (a nemusí) mít význam

sekvence nukleotidů, počítačové 0 a 1, v běžném jazyku

Smysluplná sekvence?

The main challenge was to understand gene control and the way in which proteins are made. One of the problems was that these proteins were not known to have a simple linear structure of DNA. Then James Watson and Francis Crick, with Rosalind Franklin and Maurice Wilkins, showed that the structure of DNA was a double helix. This was a great discovery. Then geneticists like Gregor Mendel and others showed that the job could be done. The structure of DNA was partly based on the work of others. We also knew that DNA was a long chain of nucleotides. The structure of DNA was a double helix. There was a lot of interest in this. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery.

Letters in English text



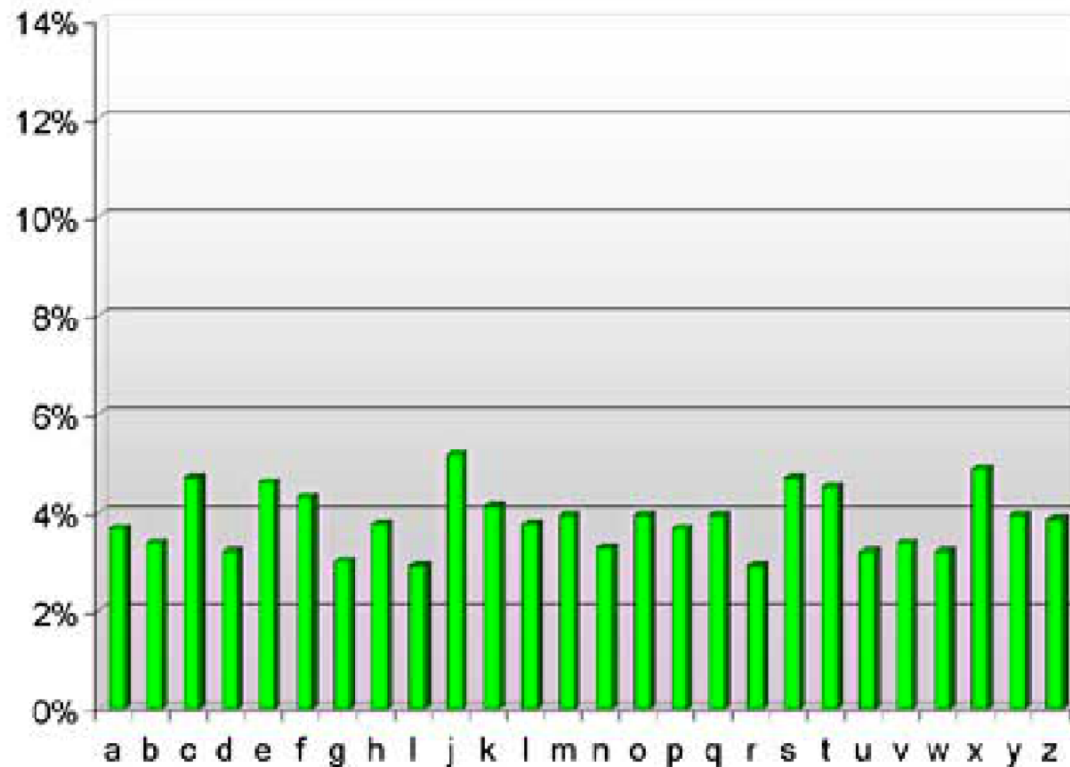
Understand gene control. protein problems could be solved by the use of the gene structure of DNA. Then we were interested in his lab, we thought it would take a few months. Pauling's feat in solving the structure of DNA was a great discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery. It was a big discovery.

It brought us not only joy but great relief. It was unbelievably interesting and immediately allowed us to make a serious proposal for the mechanism of gene duplication.

Smysluplná sekvence?

jcvbyfmmktllkrfsuogqfoqzpjklhvzgnkifytjtbjavafjlvqnlyf
ozkcbjwbkdyueayklxkietjzclpgrknxhjdngitaxyvuorfxgihkyr
rcxummzwooxzujxjzrzbsebpzfxjwjrxapzpyaqcnei jgdwtpsweo
tjqqepnltykhvmfelmhshvyxxmkngoadattnofttmtlsaeioqamvbx
djpdipxftmdhyothcvoixod
yhkyfgkyqvghibnyjamluox
cczxvnkzcxyuxrfdosxqsm
vktgjxhhvrvwxtfiudbvqjs
syiqexibxtsvyxepvdocaht
egzdkhegrcwmwtse lofmyf
asesfptktyacpxlmmqjjqtc
iecnowaemfmrpqcbretesns
ildrxuepplewrxqujadbwla
bxxdihdyspvfccjdneaeacr
yupyekrqpcjalsehvnzsnqm
ggeyhpwobwtaatwgxcamjun
lqpogupltfpbwjjahdkbwhi
xehqemciyakfkpwcycjddsc
nqmqq loukfrfpwbxyluffpv
ogncujkyjujorbps smweqfs

Letters in random text



Frekvenční analýza

Smysluplná sekvence?

01010101010101010101

01010101010101010101

01010101010101010101

01010101010101010101

01010101010101010101

01010101010101010101

Sekvence nemůže být současně
náhodná i smysluplná!

Náhodná nebo smysluplná?

Frekvenční analýza

číslo	počet	poměr
0	10 (60)	50%
1	10 (60)	50%

010101010101010101

číslo	počet	poměr
0	10	50%
1	10	50%

Očekávaná frekvenční analýza párů pro náhodnou sekvenci

číslo	počet	poměr
00		25%
11		25%
01		25%
10		25%

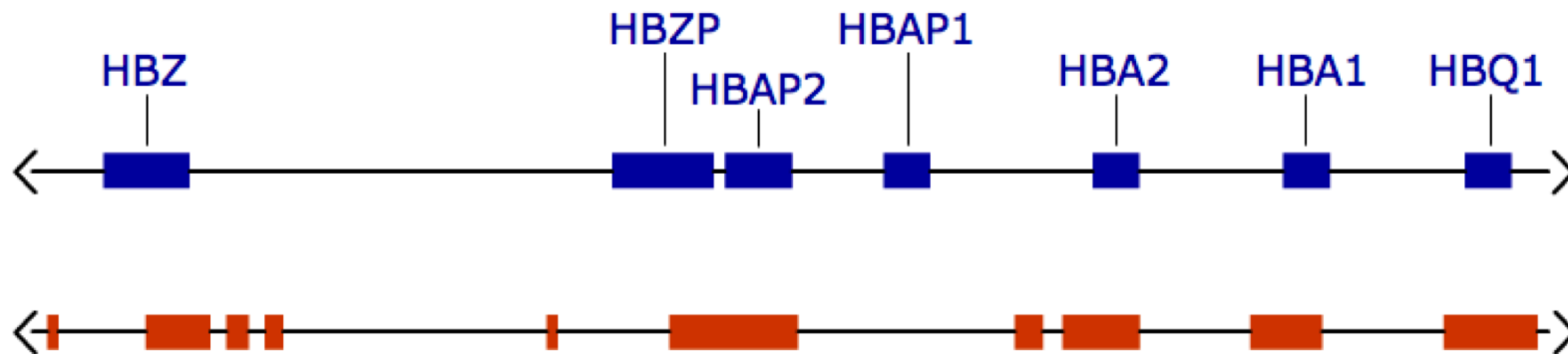
Frekvenční analýza párů pro výše uvedenou sekvenci

číslo	počet	poměr
00	0	0%
11	0	0%
01	10	53%
10	9	47%

K čemu je to dobré?

- Obsah GC je např. vyšší v genových částech než intergenových
- GC ostrůvky se objevují v oblastech regulujících transkripci, ...

Genes:



CG dinucleotide rich regions

Table 1
Software commonly used for bacterial genome annotation and comparison

<i>DNA level annotation</i>		
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
RNAmmmer	http://www.cbs.dtu.dk/services/RNAmmmer/	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
<i>Protein level annotation</i>		
BLAST	http://www.ncbi.nlm.nih.gov/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
<i>Comparative genomic tools</i>		
Mauve	http://ge1.lahabs.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
TGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCCTGA
TGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGG
CGCAGCGATGTGCTGGCGGCCGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCG
TGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATAACCAAACAGCCGGGCTT
TAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCTTCTATGCG
AACGCGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTA
ACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAGCGAGCGCGAA
CGGCCGTGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATAACCGTGTGGCTGGGCTGGCTG
GGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAATGGG

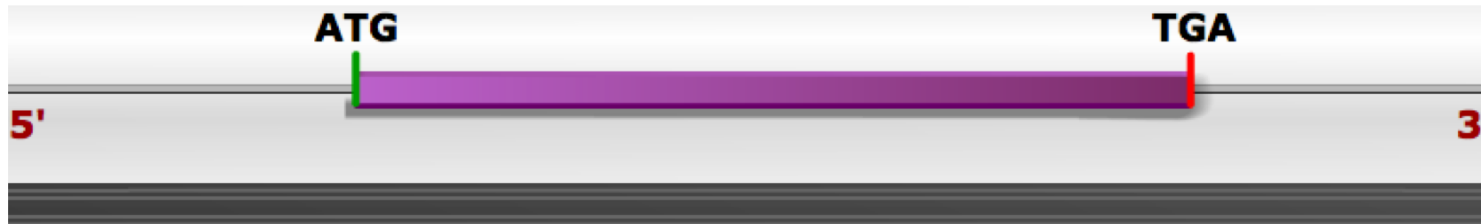
nonpolar polar basic acidic (stop codon)

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 ^[A]	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



~~GTACCACGACAGAGGACGGCTGTTCTGGTTATT~~
~~CATGGTGCTGTCTCCTGCCGACAAGACCAATAA~~



CAUGGUGCUGUCUCCUGCCGACAAGACCAUAA



RF1 → RF1
 RF1 CAU GGU GCU GUC UCC UGC CGA CAA UAA GAC CAA
 └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘
 His Gly Ala Val Ser Cys Arg Gln **end** Asp Gln
 RF2 → RF2
 RF2 C AUG GUG CUG UCU CCU GCC GAC AAU AAG ACC AA
 └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘
 Met Val Leu Ser Pro Ala Asp Asn Lys Thr
 RF3 → RF3
 RF3 CA UGG UGC UGU CUC CUG CCG ACA AUA AGA CCA A
 └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘ └─┘
 Trp Cys Cys Leu Leu Pro Thr Ile Arg Pro

Překlad DNA sekvence

- **ExPASy**

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

- **ORF Finder (NCBI)**

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

ExpASy

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

The screenshot displays the ExpASy Bioinformatics Resource Portal interface. At the top, there is a header with the SIB 15 YEARS logo and the text 'ExpASy Bioinformatics Resource Portal'. Below the header, there is a navigation menu on the left with categories like 'Visual Guidance', 'Categories', 'Resources A..Z', and 'Links/Documentation'. The main content area shows search results for the keyword 'translation'. The results are organized into two columns: 'Databases (0)' and 'Tools (5)'. The 'Tools (5)' column lists several tools, with the 'Translate' tool highlighted by a red circle. The 'Translate' tool description is: 'Translation of a nucleotide (DNA/RNA) sequence to a protein sequence. [more]'. Other tools listed include 'EMBOSS translation tools', 'Graphical Codon Usage Analyser', 'Reverse Transcription and Translation Tool', and 'Reverse Translate'.

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
TAGCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCG
CGGAAGTGAGCGTGCTGTTTATTTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAA
GTGCGTGATGCCGCCACCGCCGTTCCGGATGCCGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGA
TCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGGACCCACTGTACCGCCGATTTTTCGGGTGTGCGATC
GTGATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACC
AAACAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAA
AGCGATCTTCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGG
CCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATT
CGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGA
TACCGTGTGGCTGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCCGATTATAATGATGGCATTGTTATTC
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** PALLPRSNWKC **V** **Met** PPPFR **Met** R **Met** ICCIRAVVR **Stop** KIIIGAA **Met** CWRRARPPVPPILRCAIV **Met** AP **Stop** AVIFVVGKPAKLRAASRIPNSRALNRAAIA **Met** ATLACRRIPPLKRSS **Met** RTRRIVRI **Stop** NCLL **Met** Met RRNRPPPLWWTAK **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** SAASELSSAERSLLAQRCAGGGRDHLYRRFCGVRS **Stop** WHRE **Stop** RLFSLGNQH **Stop** NCGQPAGYQTAGL **Stop** TEQRSQWQL **Stop** PAAEYRL **Stop** SDLLCERGGSSGSETVY **Stop** **Stop** CAGTGRHLCG **Stop** QRRWCASVYPE **Stop** QRW **Stop** NSY **Stop** SERERP SERDRCP SGAAERGRYRVAGLAGRGRWCRCGL **Stop** **Stop** WHCYS AVADYL **Met**

5'3' Frame 3

Met LVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKD **Stop** AGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQP **Stop** GFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARL **Stop** APLSAGDTVWLGWLGAEADGADADYNDGIVILQWPIT **Stop** W

3'5' Frame 1

PIR **Stop** SATAE **Stop** QCHHYNPHRHHLPRPAS PATRYRPRSAAPDGHRSDGRSRSLQYEFYHLCYSG **Stop** TDAHHLRCYPQRW **Stop** RPVPAHHQ **Stop** TVSDPDDPPRSHRRSL **Stop** RRYSAAG **Stop** SCHCDRCSV **Stop** SPAVWYPAGCPQFQCWFPNENNRSRCHHDR **Stop** TPQNR RYRWSRPPPAHRCANNDL SADDSSDAADHPHPERRWRHHALPVRS GQQCRHLSGQSE **Stop** TARSLPRAAPARFRAD **Stop** AADHPGSDR **Stop** TDGYRYGPAARHRPG **Stop** QHTGGDHQSRSGRPDHGWLPDRRSAG **Stop** RHPQSPAY

3'5' Frame 2

PLGNRPLQNNNAIIIRIGTIFRAQPAQPHGIARAQRRQTGIGRALTAVRARENTNFTTFAIQGKQHTHTIFAVTHKGGGRFRRIINKQF **Stop** QILTIRRVRIEDRFKGGIRRQAKVAIAIAARFKARLFGIRLAARNFNAGFPTKITAHGAITIAHRKIGGTGGRARRQHIAAPI **Met** IFQRT **Stop** TAR **Met** QQIIRIRNGGGGITHFQFDRGNAGIFQGKANKQHAHFRAQRQRGFAQ **Met** QLITRAQTGKQTAIV **Met** AQLRGIARANNIQV **Stop** ATINHGRGGRITAGFRIGAQQGNGIHNHQH

3'5' Frame 3

H **Stop** VIGHCRIT **Met** PSL **Stop** SASAPSSAPSQPSHTVSPALSGARRASVAL **Stop** RPFALASIRILPPLLFRVNR RTPSSLLPTKVAA **Stop** GSGASSINSFRS **Stop** RSAAFA **Stop** KIALKAVFGGRLKLPLRSLLGLKPGCLVSGWLP AIS **Met** LVSQRK **Stop** PLTVPSRSHTAKSA **Stop** VQVVAPAASTSLRQ **Stop** **Stop** SFSGRQLGCSRSSASGTAVAASRTSSSIGAT **Met** PASFRAKRINSTLTSARSASAVSRCS **Stop** S **Stop** PGLRPVNRRLSLWPSCAASPGLTTYRWRPSITVGAAGSRLASG **Stop** ALSRVTASTITSI

ORF Finder (NCBI)

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

The screenshot shows the NCBI ORF Finder web interface. On the left is a blue sidebar with the NCBI logo and navigation links for PubMed, Entrez, BLAST, OMIM, and Taxonomy. Below these are links for NCBI, Tools for data mining, GenBank (sequence submission support and software), and an FTP site (download data and software). The main content area is titled "ORF Finder (Open Reading Frame Finder)" and contains a descriptive paragraph: "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." Below the text are input fields for "Enter GI or ACCESSION" with "OrfFind" and "Clear" buttons, a large text area for "or sequence in FASTA format", and "FROM:" and "TO:" input fields. At the bottom, there is a dropdown menu for "Genetic codes" currently 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 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 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 accession number

or sequence

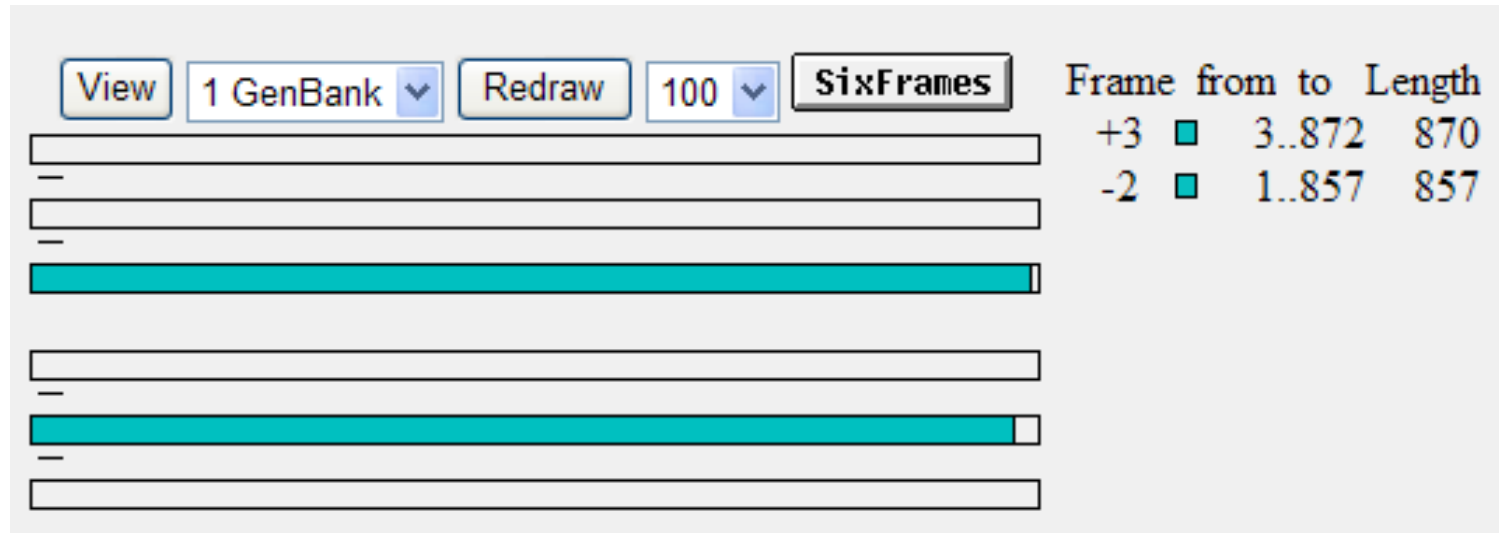
FROM:

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>

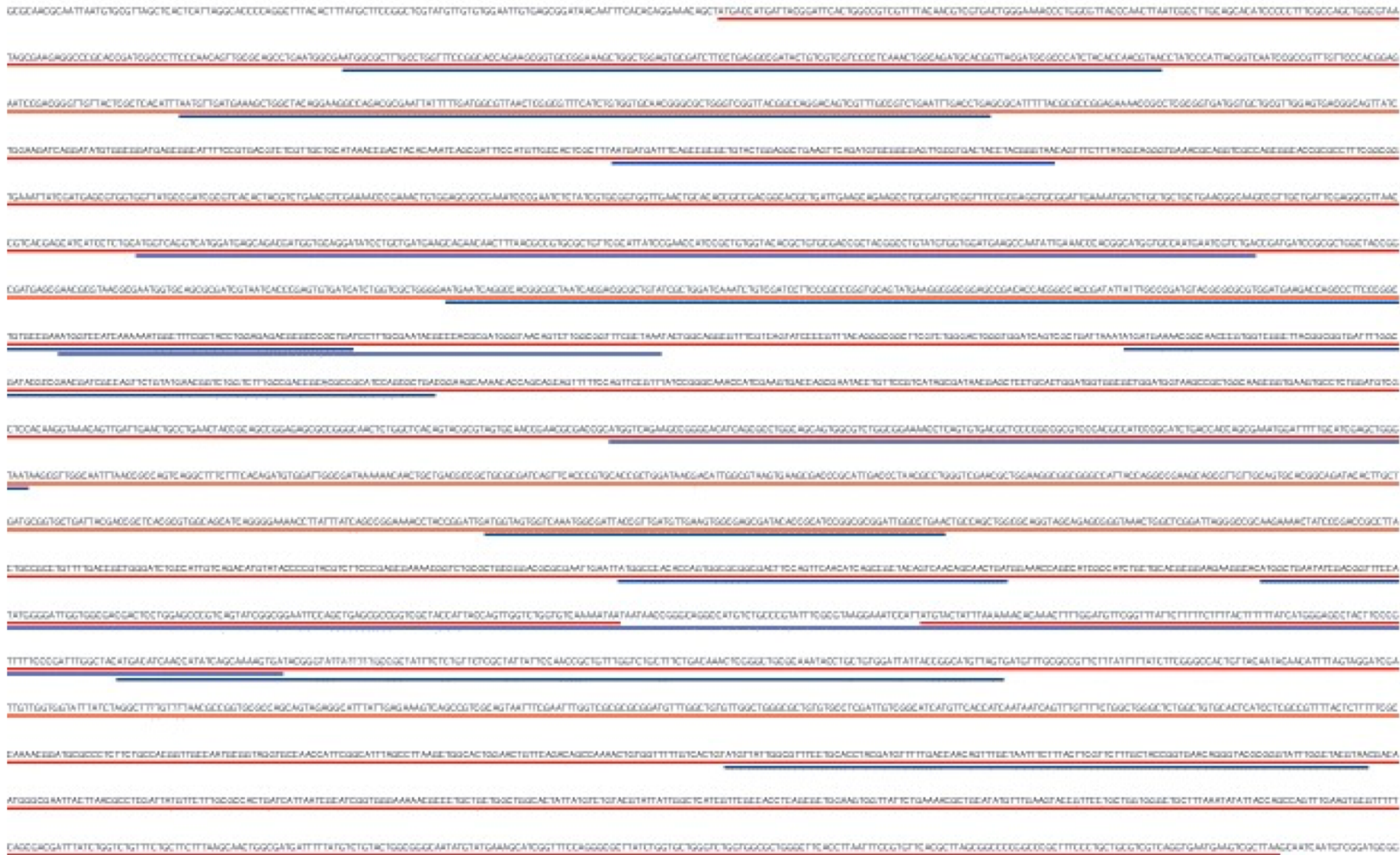


5' Frame 3

MetLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDA AQLGHNDSRLFTGLSPGDQLHLRETALALRAEVS VLFIRFALKDAGIVAPI
ELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGN
FSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLG
AEDGADADYNDGIVILQWPIT **Stop** W

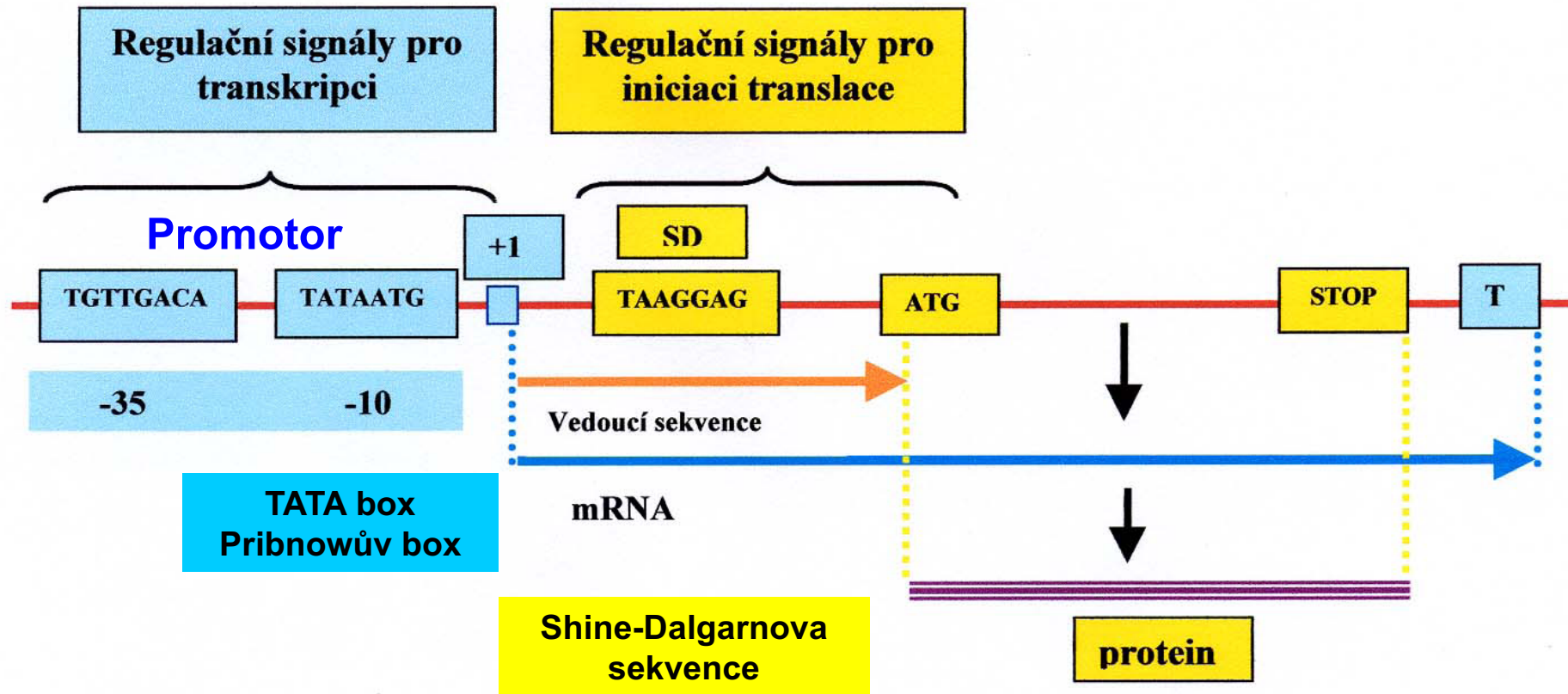
3' Frame 2

PLGNRPLQNNNAIIIRIGTIFRAQPAQPHGIARAQRRQTGIGRALTAVRARENTNFTTFAIQGKQHTHTIFAVTHKGGGRFRRIINKQFQILTI
RRVRIEDRFKGGIRRQAKVAIAIAARFKARLFGIRLAARNFNAGFPTKITAHGAITIAHRKIGGTGGRARRQHIAAPI **Met**IFQRTTAR **Met**QQII
RIRNGGGGITHFQFDRGNAGIFQGKANKQHAHFRAQRQRGFAQ **Met**QLITRAQTGKQTAIV **Met**AQLRGIARANNIQVATINHGRGGRITA
GFRIGAQQGNGIHNHGH



The diagram shows 4522 bp of the lactose operon of *Escherichia coli* with all ORFs longer than 50 codons marked. The sequence contains two real genes - *lacZ* and *lacY* - indicated by the red lines. These real genes cannot be mistaken because they are much longer than the spurious ORFs, shown in blue.

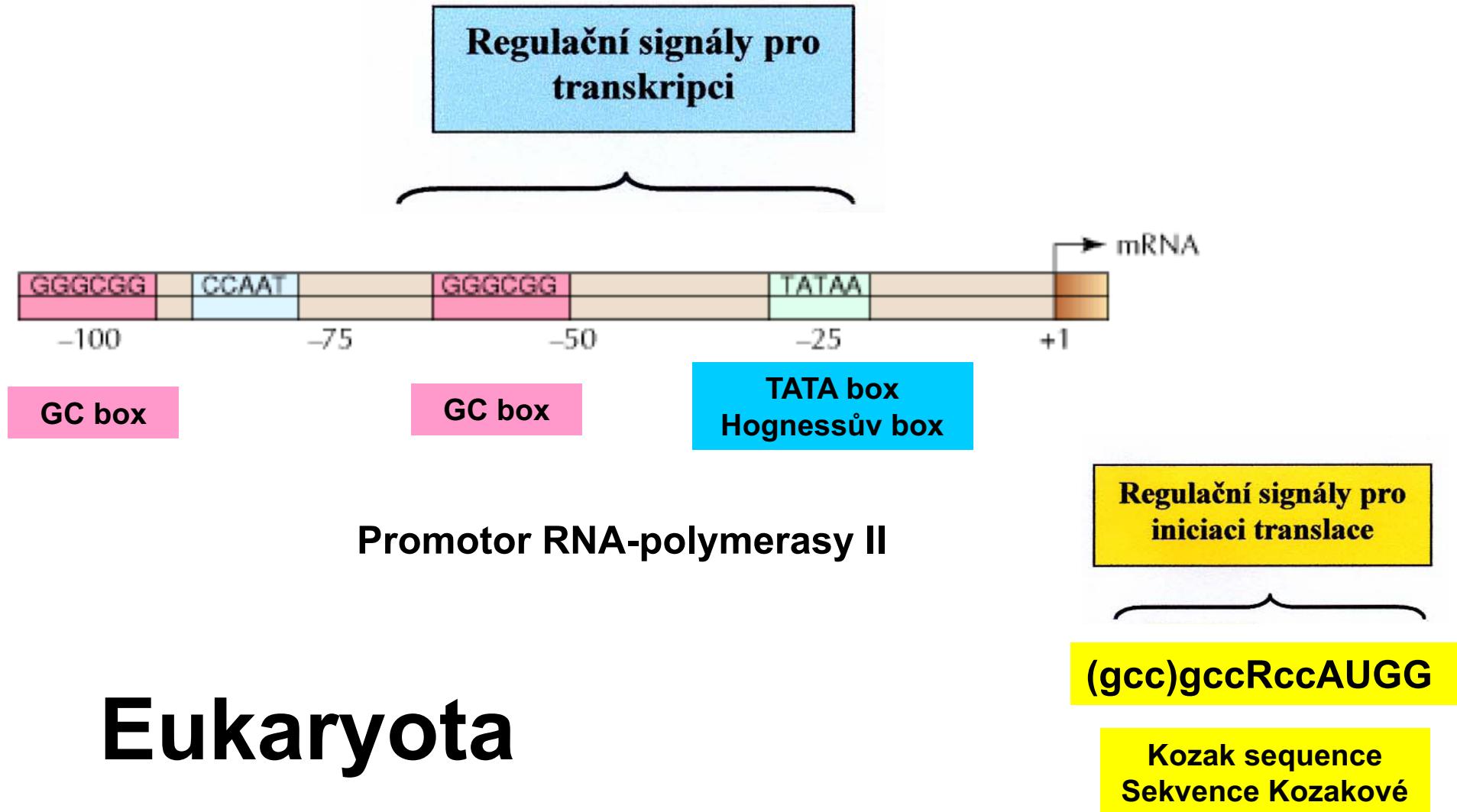
Translační a transkripční signální sekvence



Prokaryota

oblast bohatá na puriny
~ cca 8 bází upstream

Translační a transkripční signální sekvence



Promotor sequences

~ggcctataaaaattctctttccattgtgtttcag | tgca~

~tatataaataagctgcatactcggctctctcag | actg~

~gcgtataaaaagcatgccagccctcactgcctttatttc | gaat~

~ggtataaaatcacttgctcgtctgccatgcag | ctcg~

~ttataaaattcaaatttctccgtctctcaccctgcagatgc~

~cctataaaaagcgagtgagccgtgtctattctag | gcgg~

Prokaryotické geny

- **Velmi jednoduchý přístup k predikci genů**

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

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

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

Vynechání-inzerce nukleotidu pak ke **ZMĚNĚ ČTECÍHO RÁMCE**

Experimental vs. database sequence

PLL	-----MPNPDNTEAYVAGEVEIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW	53
PLU0732	MKKEPIKMPNPDNTEAYVAGEVAIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW	60
PLL	QTAPLPNSSWAGWNKFNGVVT SKPAVHRNSDGRLEV FVRGTDNALWHNWQTAADTNTWSS	113
PLU0732	QTAPLPNSSWAGWNKFNGVVT SKPAVHRNSDGRLEV FVRSTDNALWHNWQTAADTNTWSS	120
PLL	WQPLYGGITSNPEVCLNSDGRLEV FVRGSDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP	173
PLU0732	WQPLYGGITSNPEVCLNSDGRLEV FARGTDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP	180
PLL	AAHLNADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVINNC DGRLEV	233
PLU0732	AAHINADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVIGNC DGRLEV	240
PLL	FARGADSTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGQLEVFARGADNTLWRT	293
PLU0732	FARGADNTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGRLEV FARGADNTLWRT	300
PLL	WQTS HNGPWSNWSSFTGIIASAPTVAKNSDGRIEV FVLGLDKALWHLWQTTSSTTSSWTT	353
PLU0732	WQTS QNGPWSNWSSFTGIIASAPTVAKNSDGRIEV FVLGLDKALWHLWQTTSSTTSSWTT	360
PLL	WALIGGITLIDASVI-	368
PLU0732	WALIGGITLIDASVIK	376

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.**

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



ScanProsite



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

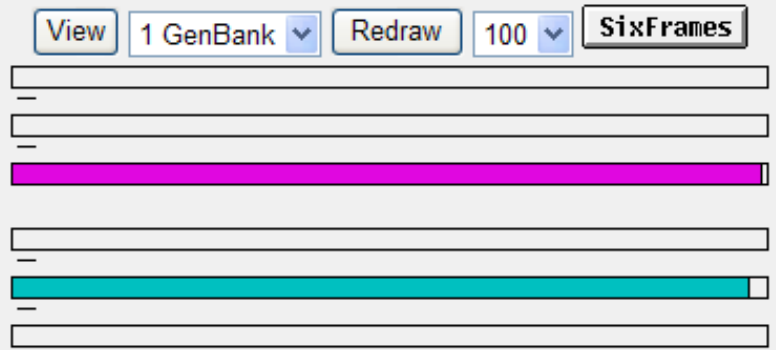


Direct Submission to SWISS-MODEL

ORF Finder (NCBI)

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

Program **blastp** Database **nr** with parameters



Length: 289 aa

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

```
3 atgctggtgattgtggatgccgttacctgctgagcgcctatccg
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 cacctgtatggtttagcccgggcgatgccgcgcagctgggccc
H L Y V V S P G D A A Q L G H
138 aacgatagccgtctgtttaccggctctgagccccgggtgatcagctg
N D S R L F T G L S P G D Q L
183 catctgcgcgaaaccgctggcgtgcgcgcggaagtgagcgtg
H L R E T A L A L R A E V S V
228 ctgtttattcgtttgccctgaaagatgccggcattgttgccccg
L F I R F A L K D A G I V A P
273 atcgaactggaagtgcgtgatgccgccaccgcttccggatgcg
I E L E V R D A T A V P D A
318 gatgatctgctgcacccgagctgtcctcctgaaagatcattat
D D L L H P S C R P L K D H Y
363 tggcgcagcagatgtgctggcggggggcgaccacctgtaccgcc
W R S D V L A A G A T T C T A
408 gattttgcgggtgtgcgatcgtgatggcaccgtgagcgggtat
D F A V C D R D G T V S G Y F
453 cgttgggaaaccagcattgaaattgcgggcagccagccgatacc
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 agcctgcccgcgaataccgcctttaagcagatcttctatgccaac
S L P P N T A F K A I F Y A N
588 gcggcggatcgtcaggatctgaaactgtttattgatgatgcgcgc
A A D R Q D L K L F I D D A P
633 gaaccggccgccaccttggggtaacagcgaagatggtgtgcgt
E P A A T F V G N S E D G V R
678 ctgtttaccctgaatagcaaagggtgtaaaattcgtattgaagcg
L F T L N S K G G K I R I E A
723 agcgcgaacggccgtcagagcgcgaccgatgcccgctctggcgccg
S A N G R Q S A T D A R L A P
768 ctgagcgcgggcgataaccgtgtggctgggctgggctggcgcgaa
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 *
```

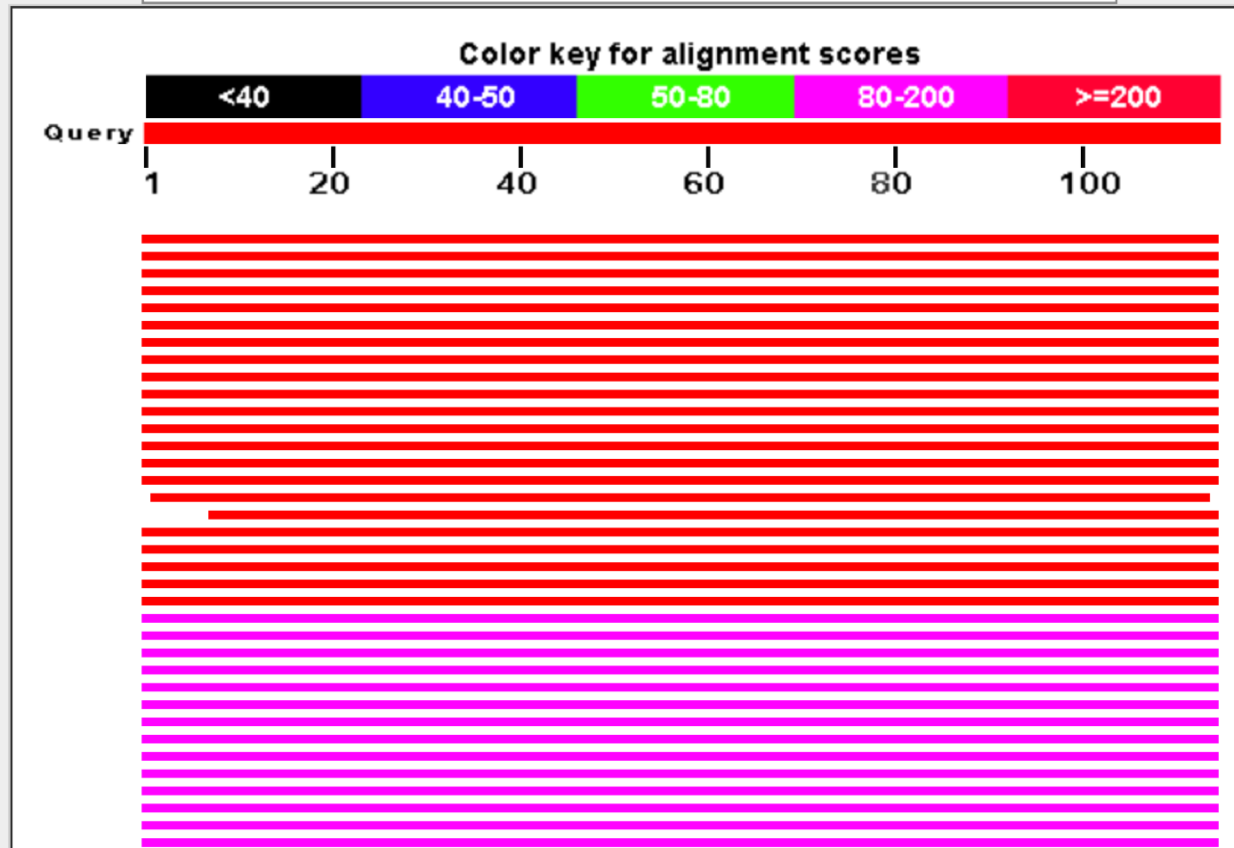
Searching for PA-IIL protein

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq.	ATQGVFTLPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNNAVIGTQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSSEGDNDYNDVAVVWINWPLG
Specific hits	PA-IIL
Superfamilies	PA-IIL superfamily

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
207 bits(527)	3e-66	Compositional matrix adjust.			107/107(100%)
	107/107(100%)		0/107(0%)		

Query	8	LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGOSTNNAVIGTQVLNSGSSGKVQVQV	67
		LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGOSTNNAVIGTQVLNSGSSGKVQVQV	
Sbjct	1	LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGOSTNNAVIGTQVLNSGSSGKVQVQV	60

Query	68	SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDVAVVINWPLG	114
		SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDVAVVINWPLG	
Sbjct	61	SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDVAVVINWPLG	107

LOCUS NZ_JUUU01000485 5873 bp DNA linear CON 21-AUG-2015
DEFINITION Pseudomonas aeruginosa strain 744_PAER 959_5873_75941, whole genome
shotgun sequence.
ACCESSION NZ_JUUU01000485 NZ_JUUU00000000

.....

gene complement(5548..>5873)
/locus_tag="ADF63_RS25535"
CDS complement(5548..>5873)
/locus_tag="ADF63_RS25535"
/inference="EXISTENCE: similar to AA
sequence:RefSeq:WP_009876850.1"
/note="Derived by automated computational
analysis using gene prediction method: Protein Homology."
/codon_start=3
/transl_table=11
/product="fucose-binding lectin"
/protein_id="WP_049233417.1"
/db_xref="GI:896235191"

/translation="LPANTRFGVTAFAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIG
TQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDVAVVVI
NWPLG"

Chyby

- Nejčastější
 - – chyby v sekvenaci
 - – špatná predikce -alternace startovního kodonu
 - – shot gun sekvenace

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.**

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

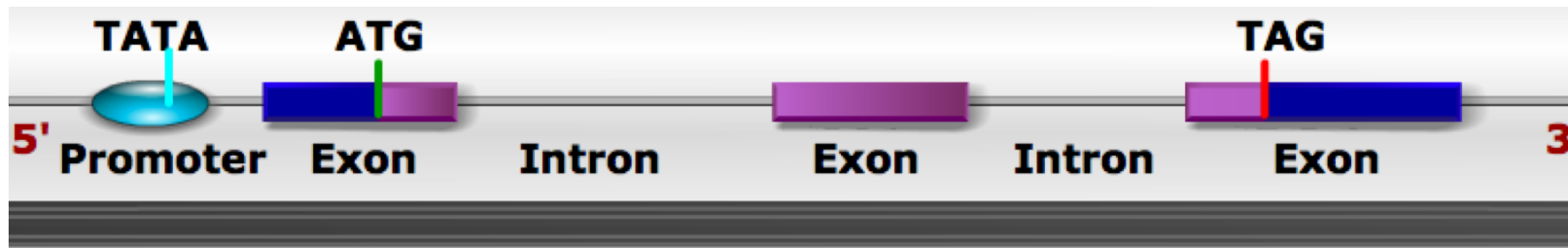
Mnohobuněčná eukaryota

- Mnohobuněčná eukaryota**

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



Glyceraldehyd-3-fosfát-dehydrogenasa
Homo sapiens



DNA



Transcription

pre mRNA



Processing

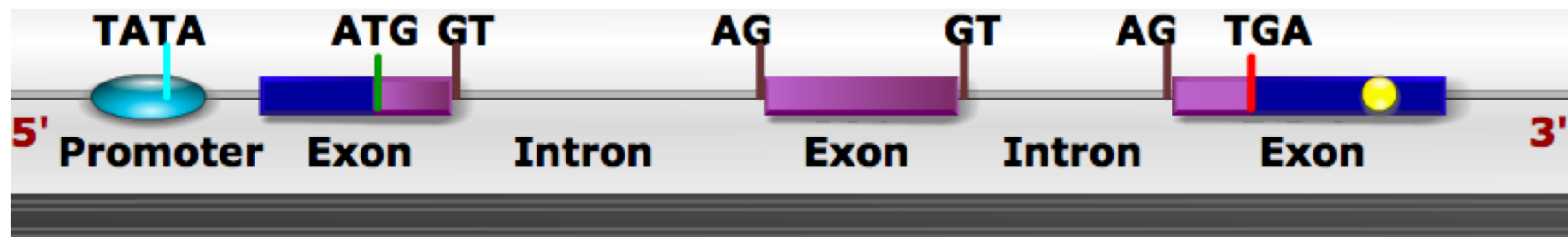
mRNA



Translation

Protein





DNA



Transcription

pre mRNA



Processing

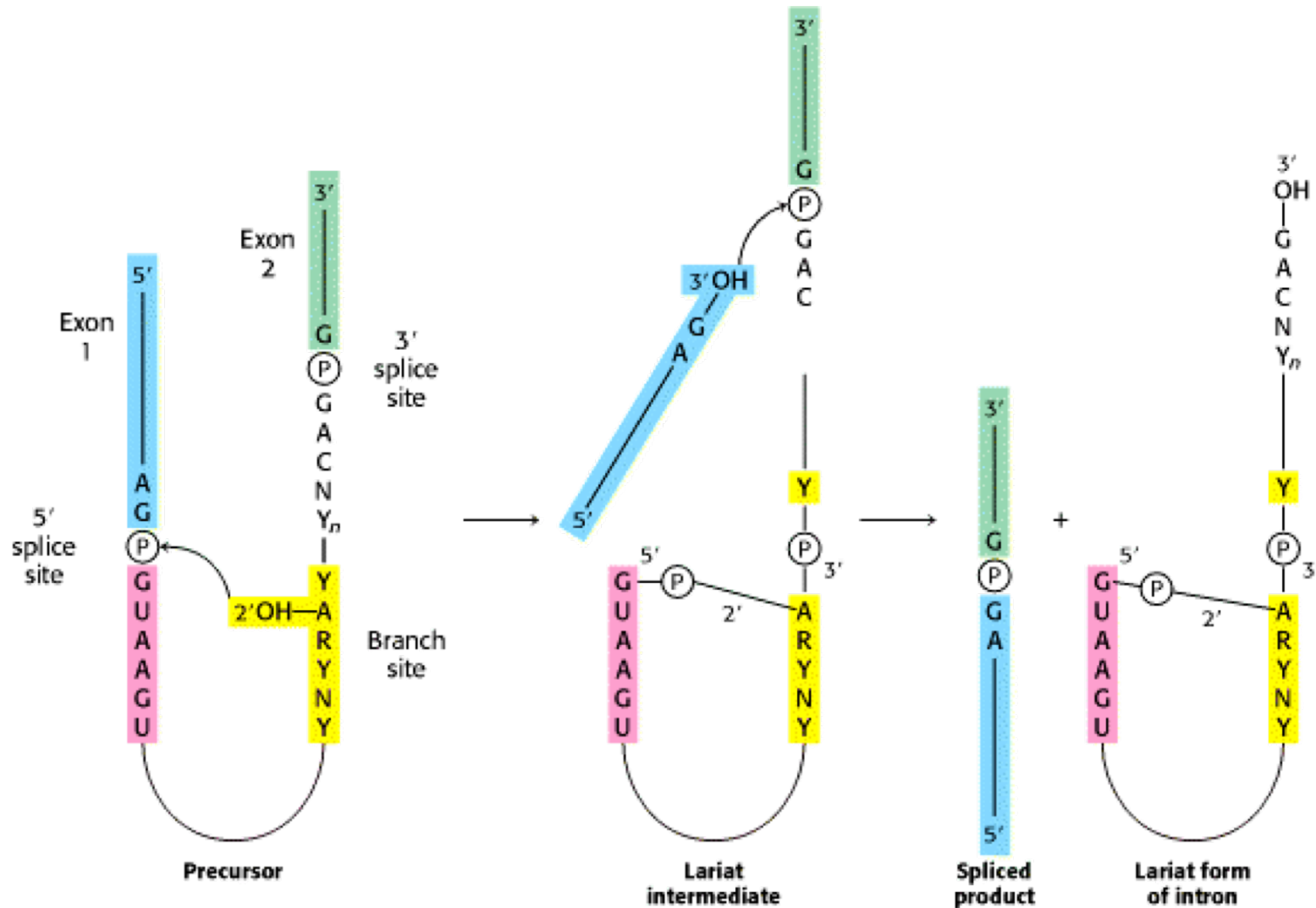
mRNA



Translation

Protein

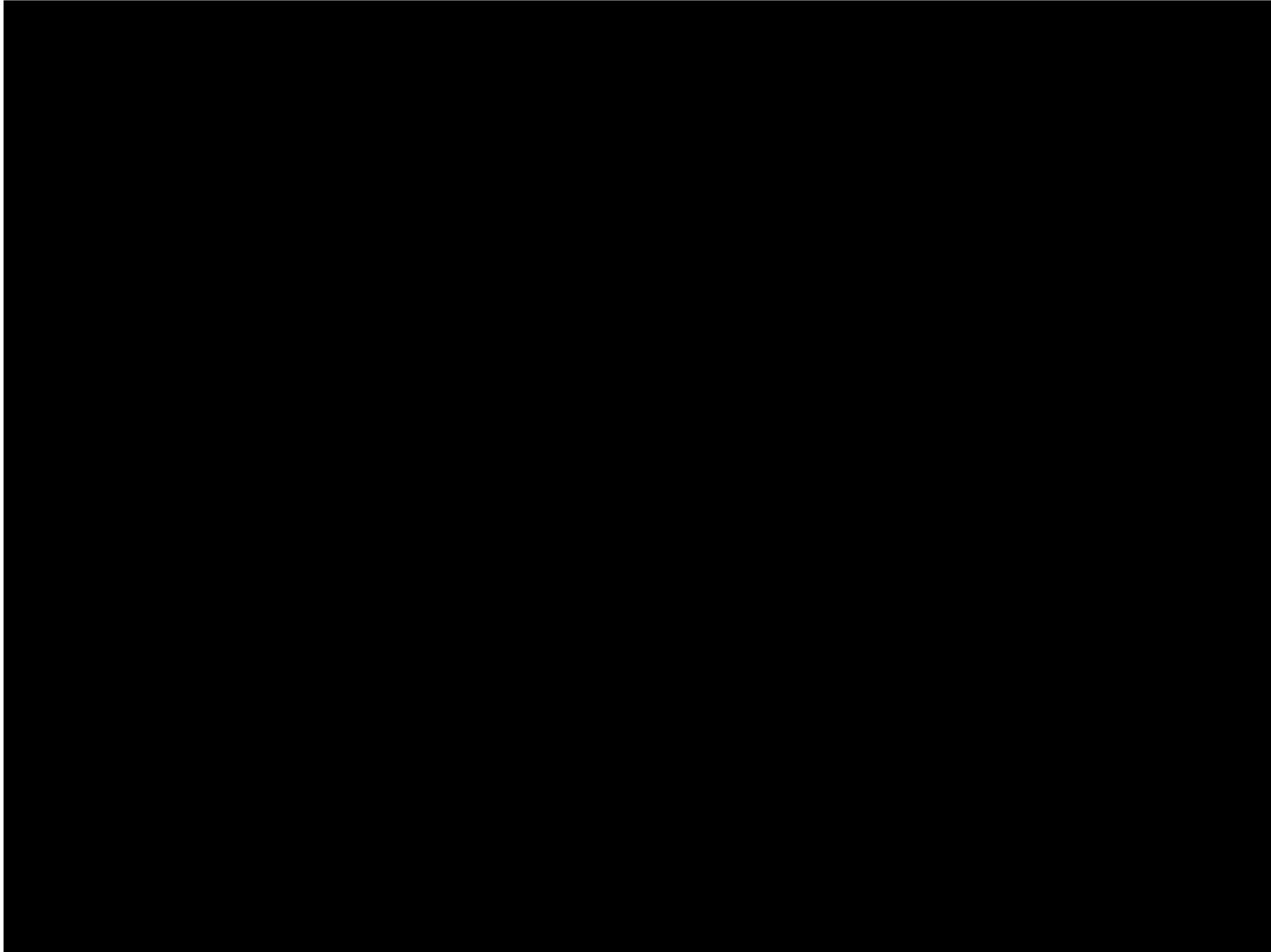




Splicing Mechanism Used for mRNA Precursors. The upstream (5') exon is shown in blue, the downstream (3') exon in green, and the branch site in yellow. R stands for a purine nucleotide, Y for a pyrimidine nucleotide, and N for any nucleotide. The 5' splice site is attacked by the 2'-OH group of the branch-site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

Eukaryotické geny

Mnohobuněčná eukaryota



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 úseky – určeny jako introny.

Algoritmy a nástroje pro identifikaci genů

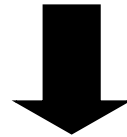
- **Predikce genů na základě sekvenční homologie** – vyhledávání v databázích pomocí algoritmů.
- **Predikce genů *ab initio*** – predikce na základě statistických parametrů DNA sekvence.
- **Většina běžně používaných metod kombinuje oba dva přístupy.**

Prokaryota

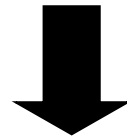
ATG.....TAA

Bez intronů

SEKVENČNÍ HOMOLOGIE



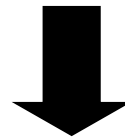
**IDENTIFIKOVANÉ GENY VYUŽITY
PRO „TRÉNOVÁNÍ“ STATISTICKÉ
METODY**



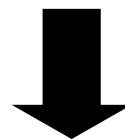
**ANALÝZA ZBÝVAJÍCÍCH
ČÁSTÍ GENOMU**

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	http://www1.imim.es/geneid.html	
FGENESH	Human, mouse, Drosophila, rice	HMM	http://www.softberry.com/berry.phtml?topic=fgenes&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, <i>C. elegans</i>	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.

Úkol – deadline 27.dubna

- DEFINITION fucose-specific lectin [Arthroderma otae CBS 113480].
- ACCESSION XP_002846975
- VERSION XP_002846975.1
- DBSOURCE REFSEQ: accession XM_002846929.1