Molekulárně biologická data

- **Výkonné technologie:**

  Automatické sekvenování

  MALDI-TOF

  NMR spektroskopie

  Proteinová krystalografie

Výrazný nárůst množství biologických dat.

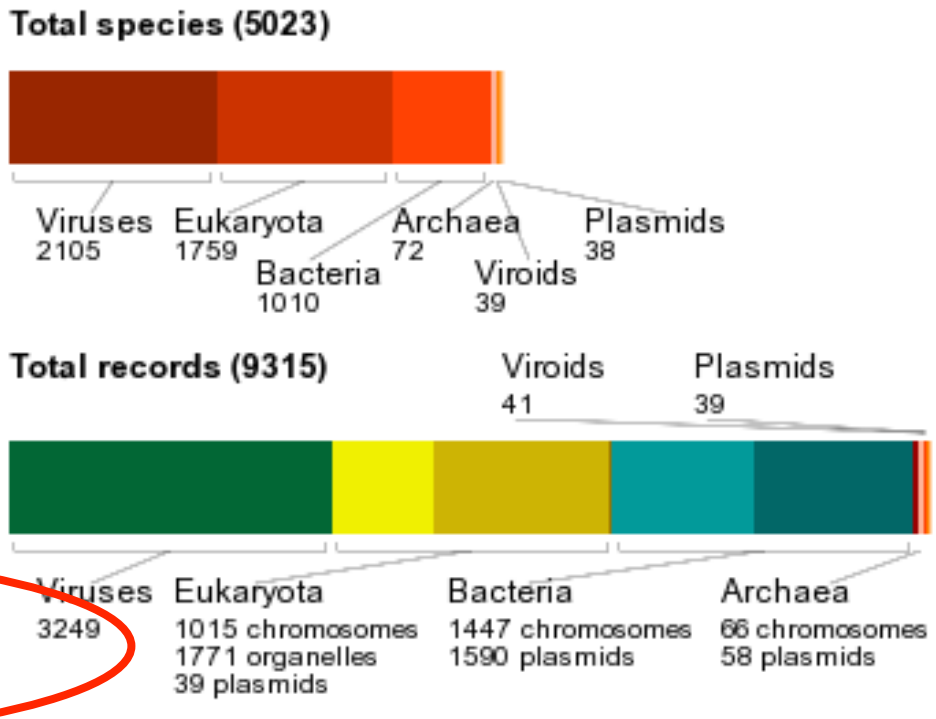
Rozdělení molekulárně biologických databází

- **Databáze:**

  Primární

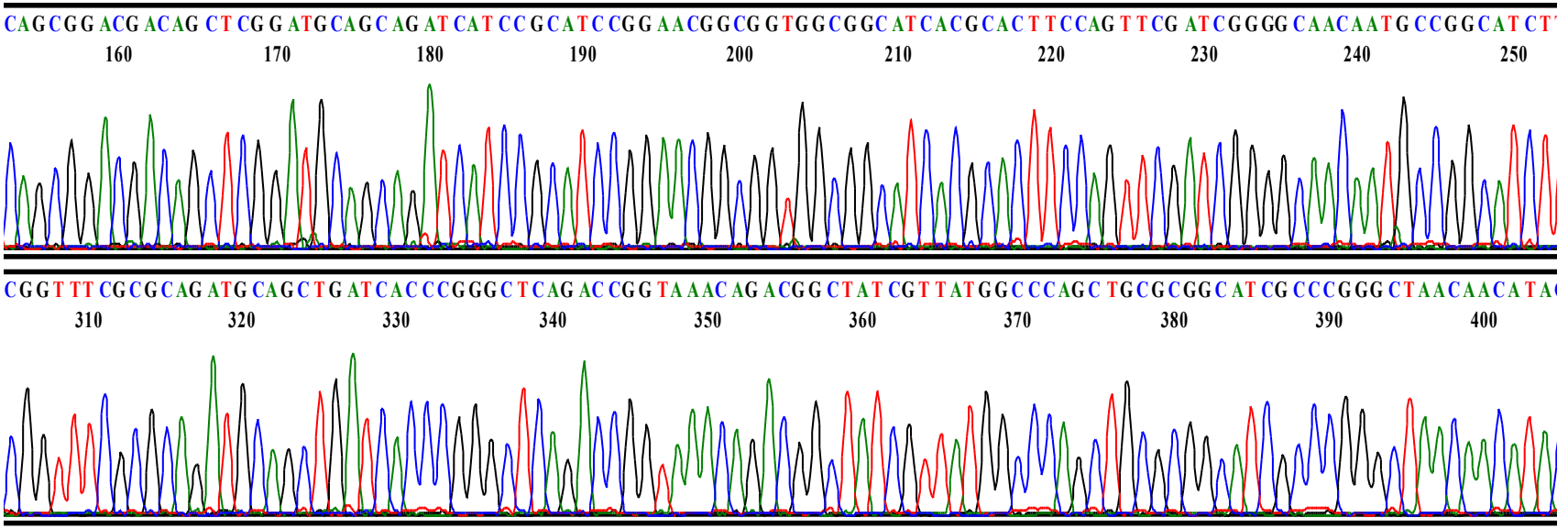
  Sekundární

  Strukturní



Genomové zdroje

Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

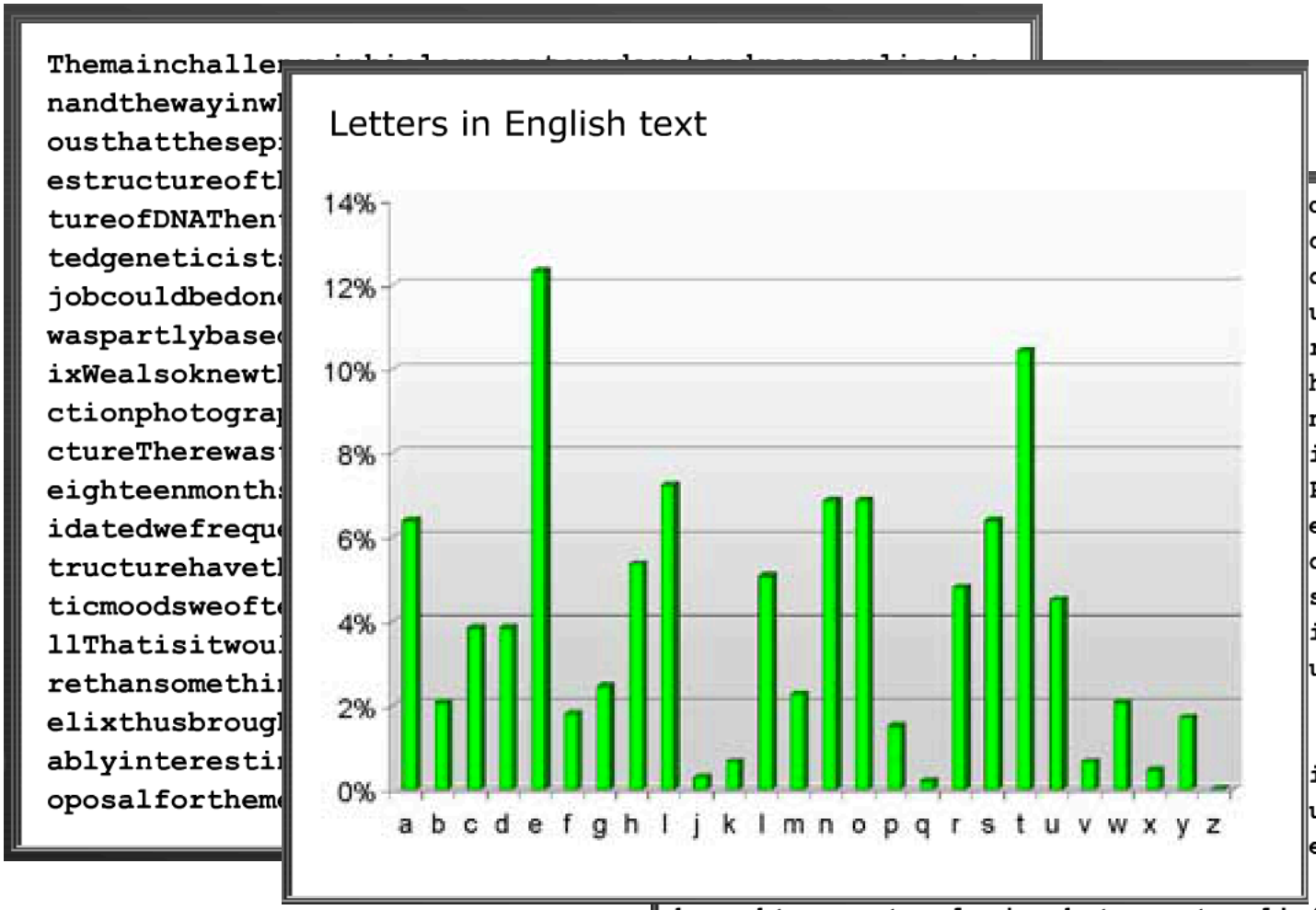
„Syrové“ sekvence DNA

Identifikace a anotace genů a proteinů

- Posloupnost písmen může (a nemusí) mít význam

sekvence nukleotidů, počítačové 0 a 1, v běžném jazyku

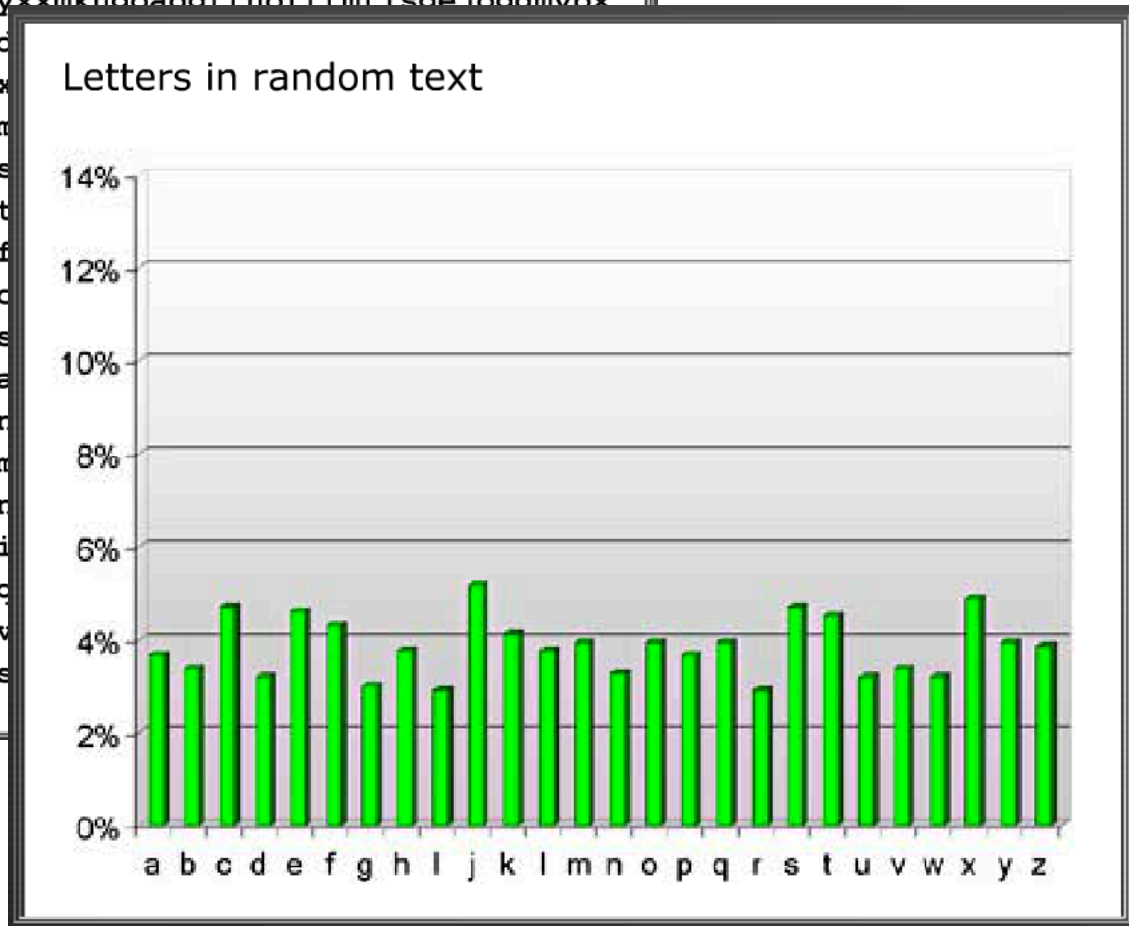
Smysluplná sekvence?



Frekvenční analýza

Smysluplná sekvence?

jcvbyfmmktllkrfsuogqfoqzpjklhvzgnkifytjtbjavafjlvqnlyf
ozkcbjbwkdyueayklxkietjzclpgrknxhjdnqitaxyvuorfxgihkyr
rcxummzwuoxzujxjzyrzbsebpzfxjwjrxapzpyaqcneijgdwtpsweo
tjqqepnltykhvmfelmhshvyxxmkngoadgttnofttmtlsgejoggmybx
djpdipxftmdhyothcvoixod
yhkyfgkyqvghibnyjamluox
cczxvnkzcxyuxrfwdosxqsm
vktgjxhhvrvwxtfiudbvqjs
syiqexibxtsvyxepvdocaht
egzdkhegrcwmwtselofmfyf
asesfptktyacpxlmmqjjqtc
iecnowaemfmrpqcbretesns
ildrxuepplewxrqujadbwla
bxxdihdyspvfccjdneaeacn
yupyekrqpcjalsehvnzsnqm
ggeyhpwobwtaatwgxcamjun
lqpogupltfpbwjjahdkbwhi
xehqemciyakfkpwcycjddsc
nqmqqloukfrfpwbxyluffpv
ogncujkyjujorbpssmweqfs



Frekvenční analýza

Smysluplná sekvence?

01010101010101010101
01010101010101010101
0101
0101
0101
01010101010101010101

Sekvence nemůže být současně náhodná i smysluplná!

Náhodná nebo smysluplná?

Frekvenční analýza

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

01010101010101010101

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, …

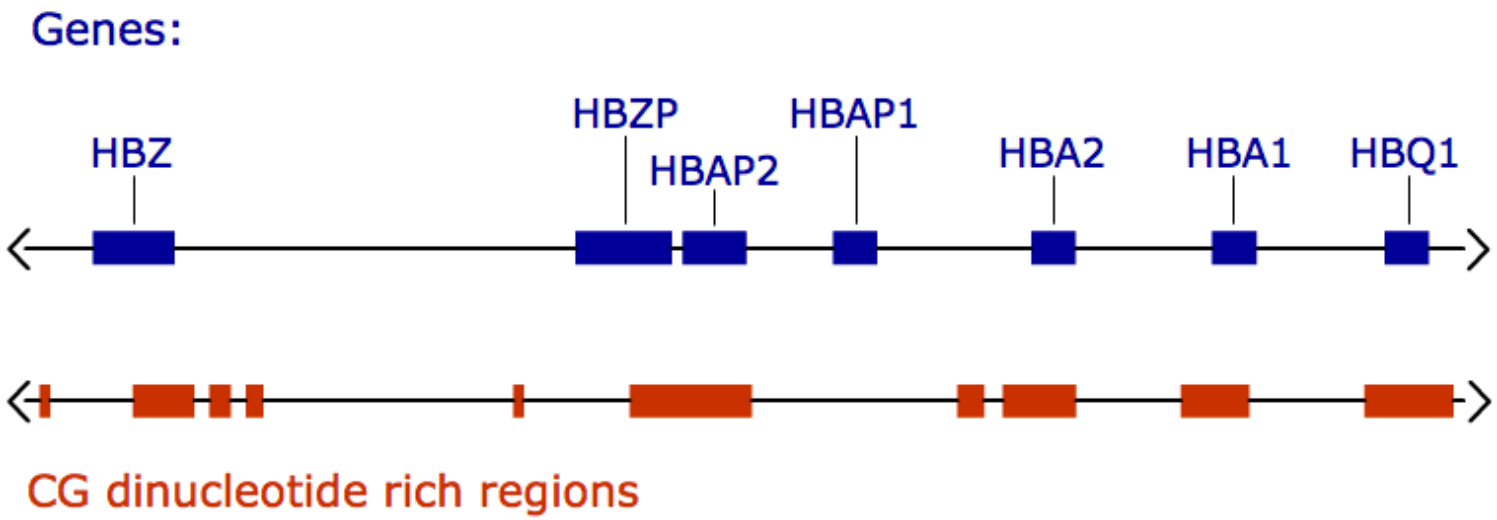


Table 1
Software commonly used for bacterial genome annotation and comparison

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.**

GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCC
CGACCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATA
GCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGG
AAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGC
GTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATT
ATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATG
GCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAACAGC
CGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCT
TCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCT
TTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAG
CGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGC
TGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGA
TTACCTAATGGG

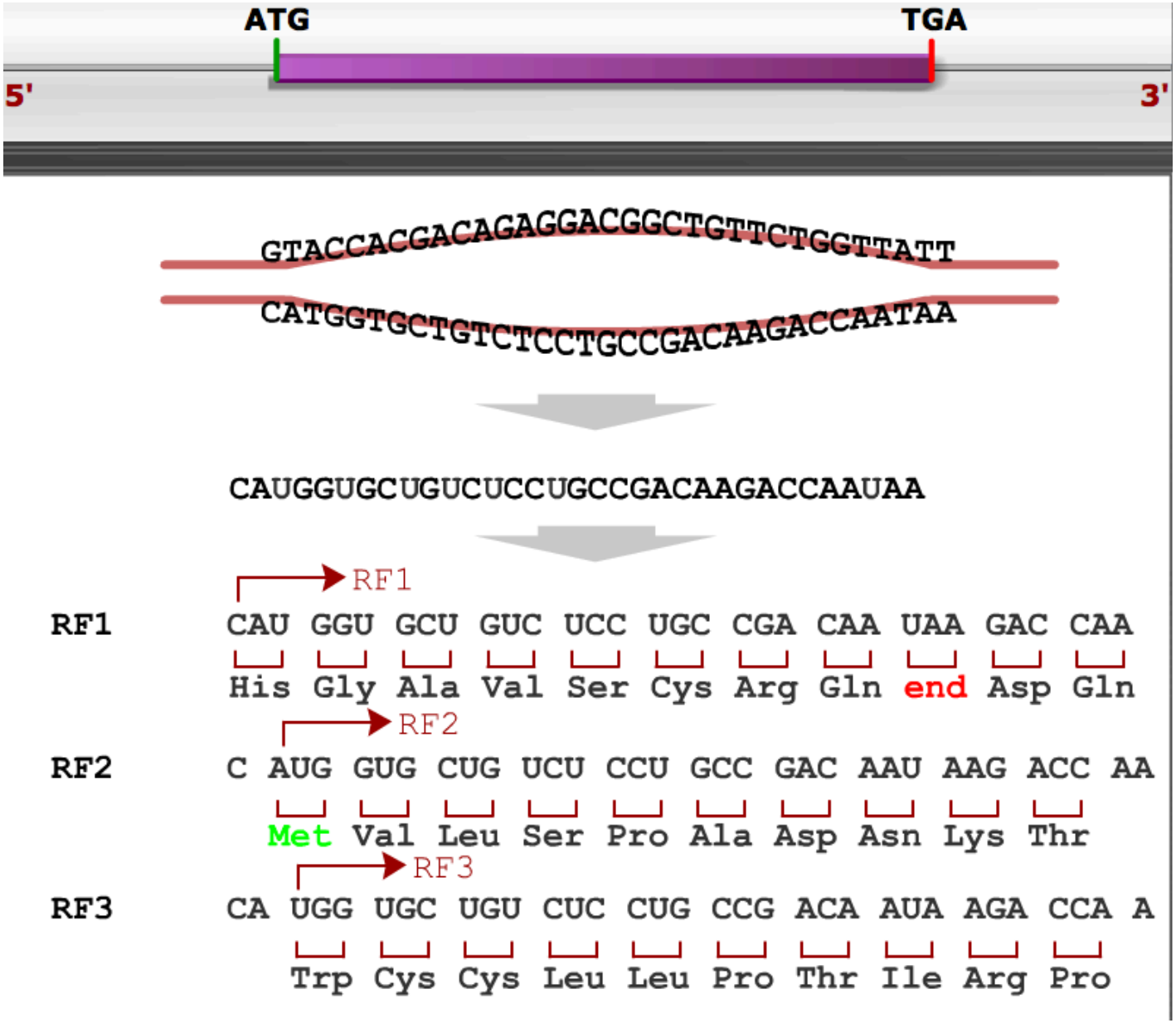
nonpolar | polar | basic | acidic | (stop codon)

Překlad DNA sekvence

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

1st base	2nd base U	2nd base C	2nd base A	2nd base G
U	UUU (Phe/F) Phenylalanine	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



Překlad DNA sekvence

- **ExPASy**

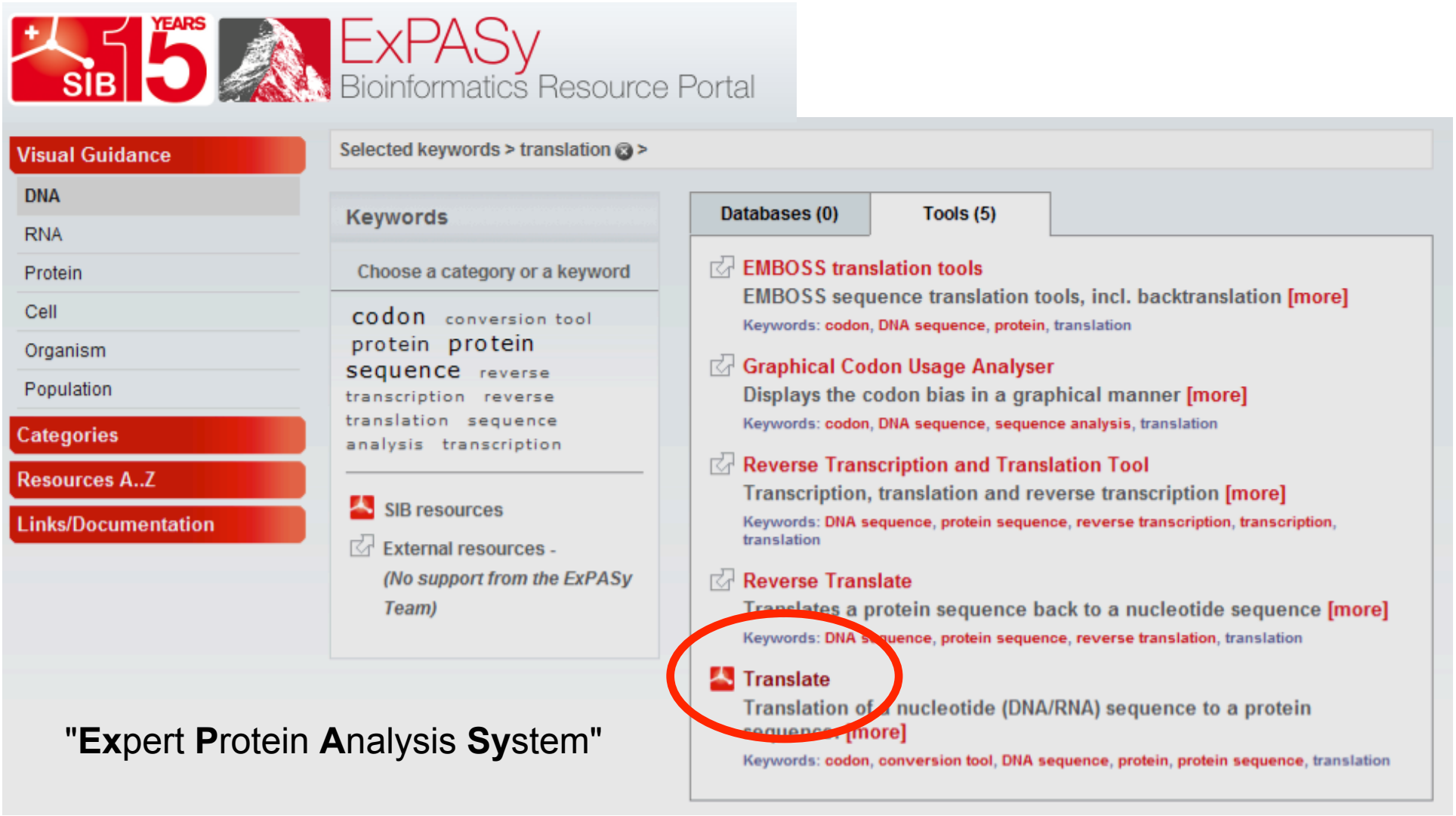
  http://web.expasy.org/translate/

- **ORF Finder (NCBI)**

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

ExPASy

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



"**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).

```
GTATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCC
CGACCGTGATTGATGGTCGCCACCTGTATGTTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATA
GCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGCTGGCGCTGCGCGCGG
AAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGC
GTGATGCCGCCACCGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATT
ATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTGTACCGCCGATTTTGCGGTGTGCGATCGTGATG
GCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAACAGC
CGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCT
TCTATGCGAACGCGGCGGATCGTCAGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCT
TTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGTGGTAAAATTCGTATTGAAG
CGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGC
TGGGCTGGCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGA
TTACCTAATGGG
```

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**

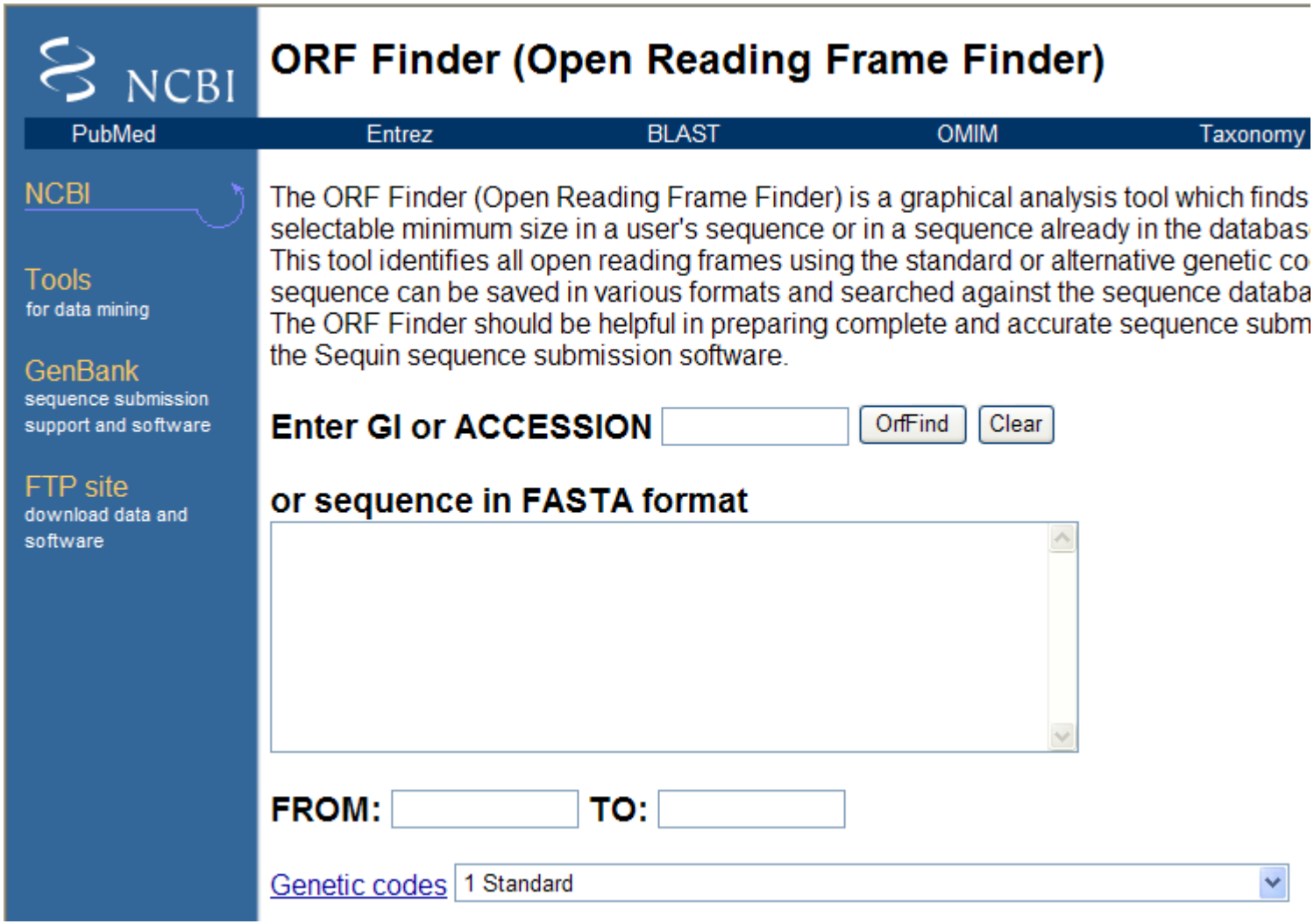
5'3' Frame 3
Met L V I V D A V T L L S A Y P E A S R D P A A P T V I D G R H L Y V V S P G D A A Q L G H N D S R L F T G L S P G D Q L H L R E T A L A L R A E V S V L F I R F A L K D A G I V A P I E L E V R D A A T A V P D A D D L L H P S C R P L K D H Y W R S D V L A A G A T T C T A D F A V C D R D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

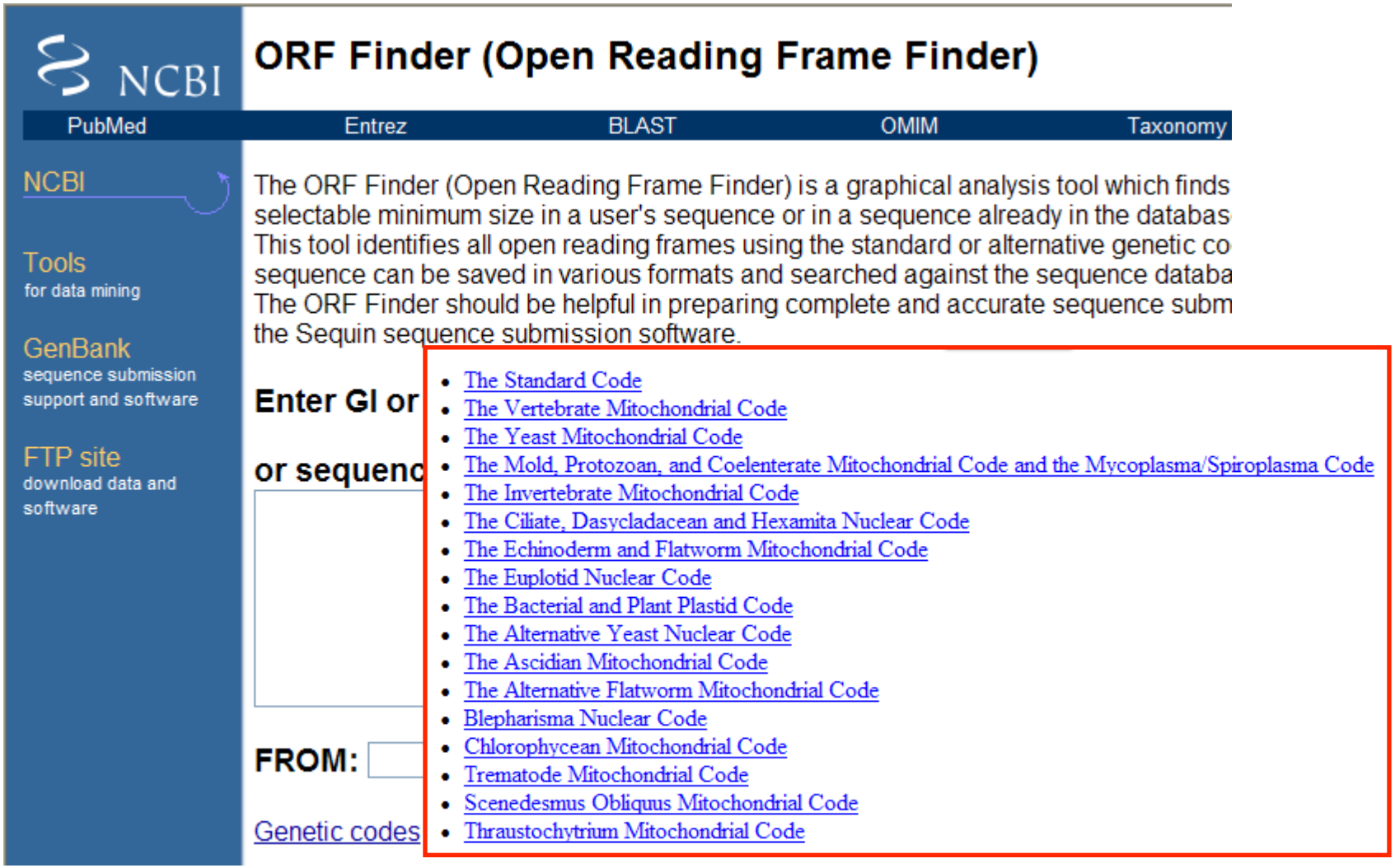
3'5' Frame 2
P L G N R P L Q N N N A I I I I R I G T I F R A Q P A Q P H G I A R A Q R R Q T G I G R A L T A V R A R F N T N F T T F A I Q G K Q T H T I F A V T H K G G G R F R R I I N K Q F Q I L T I R R V R I E D R F K G G I R R Q A K V A I A I A A R F K A R L F G I R L A A R N F N A G F P T K I T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

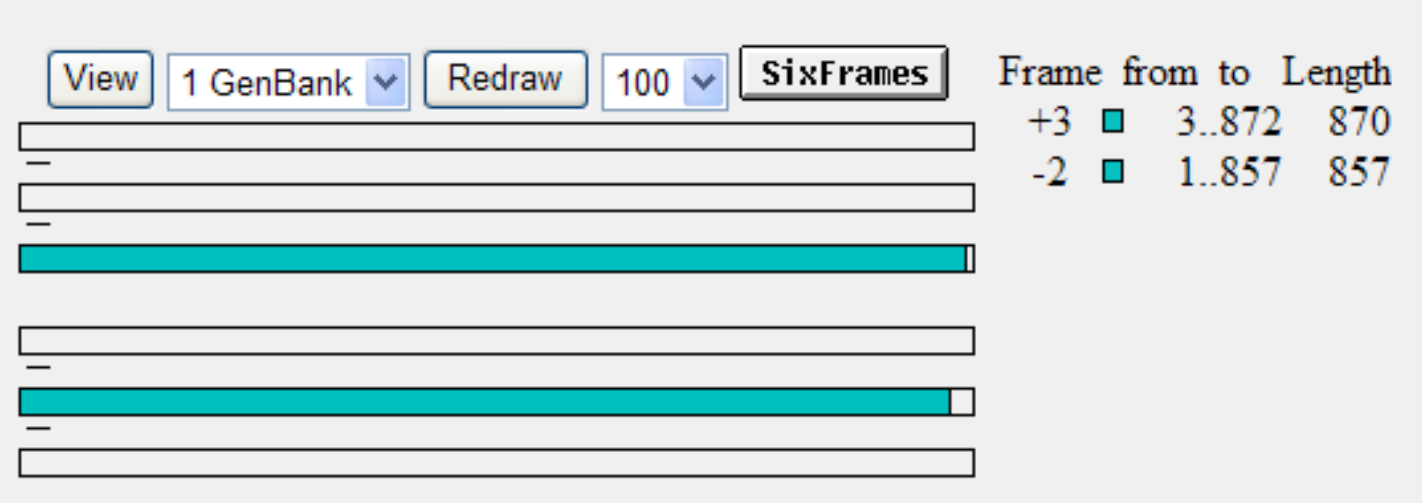
ORF Finder (NCBI)
http://www.ncbi.nlm.nih.gov/gorf/gorf.html



ORF Finder (NCBI)
http://www.ncbi.nlm.nih.gov/gorf/gorf.html



ORF Finder (NCBI)
http://www.ncbi.nlm.nih.gov/gorf/gorf.html



5'3' Frame 3

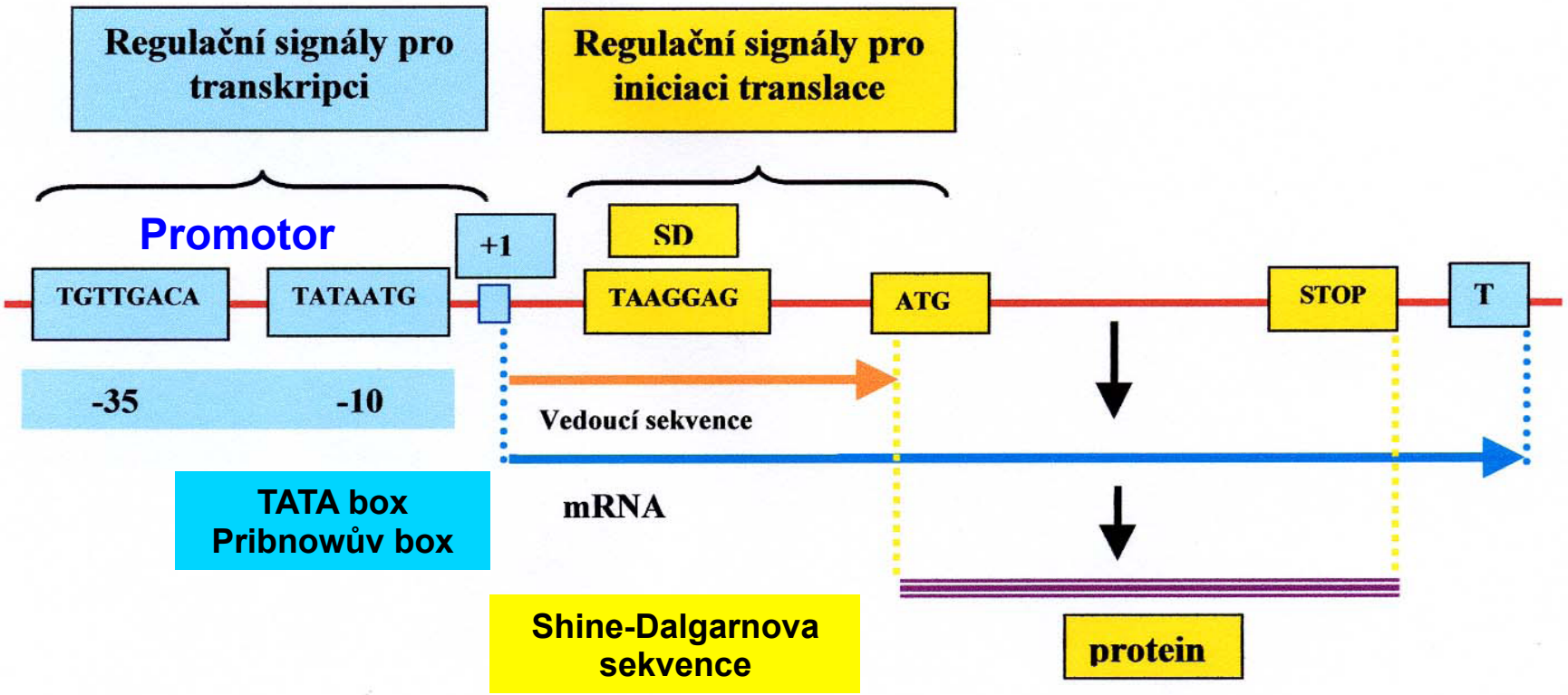
Met L V I V D A V T L L S A Y P E A S R D P A A P T V I D G R H L Y V V S P G D A A Q L G H N D S R L F T G L S P G D Q L H L R E T A L A L R A E V S V L F I R F A L K D A G I V A P I
E L E V R D A A T A V P D A D D L L H P S C R P L K D H Y W R S D V L A A G A T T C T A D F A V C D R D G T V S G Y F 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 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

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.

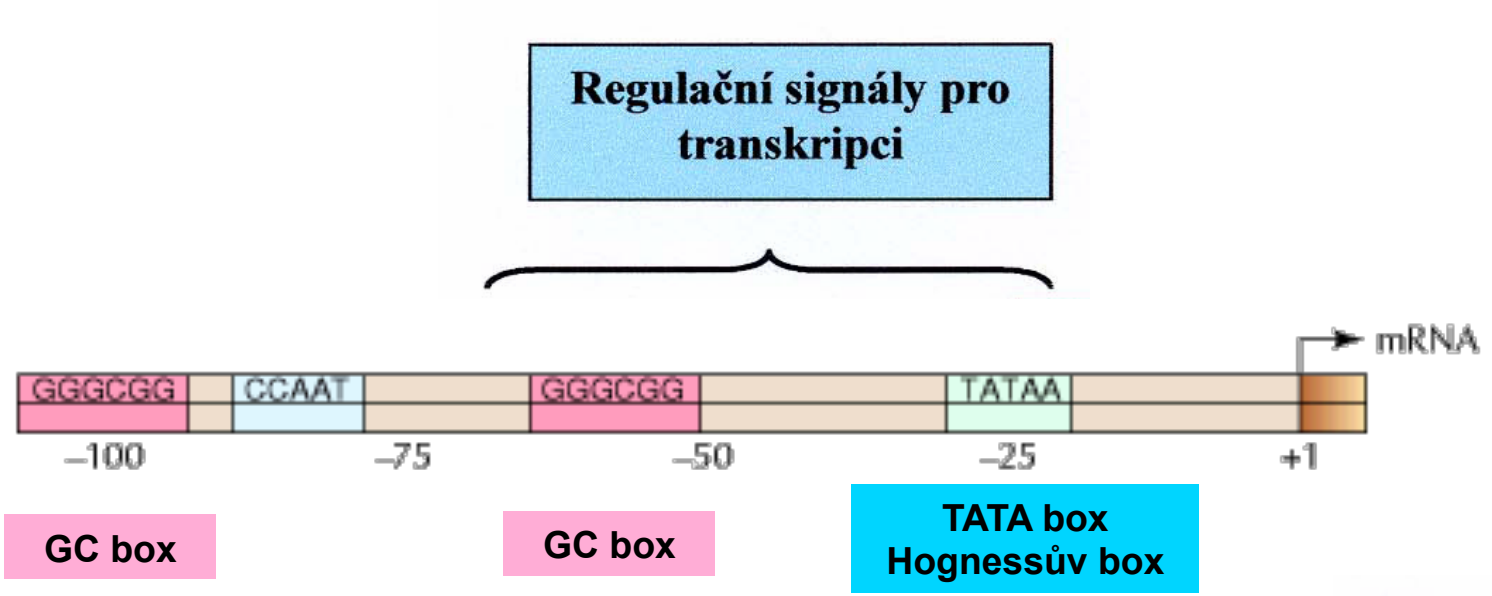
Translační a transkripční signální sekvence



Prokaryota

oblast bohatá na puriny
~ cca 8 bází upstream

Translační a transkripční signální sekvence



Promotor RNA-polymerasy II

Regulační signály pro iniciaci translace

(gcc)gccRccAUGG

Kozak sequence
Sekvence Kozakové

Eukaryota

Promotor sequences

~ggcctataaaattctctttccattgtgtttcag|tgca~

~tatataaataagctgcatactcggtctctcag|actg~

~gcgtataaaagcatgccagccctcactgcctttatttc|gaat~

~ggtataaatcacttgctcgtctgccatgcag|ctcg~

~ttataaattcaaatttctccgtctctcaccctgcagatgc~

~cctataaaagcgagtgagccgtgtctattctag|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 nukleoticu pak ke ZMĚNĚ ČTECÍHO RÁMCE

Experimental vs. database sequence

```
PLL       -------MPNPDNTEAYVAGEVEIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW 53
PLU0732   MKKEPIKMPNPDNTEAYVAGEVAIENSAIALSGIVSVANNADNRLEVFGVSTDSAVWHNW 60

PLL       QTAPLPNSSWAGWNKFNGVVTSKPAVHRNSDGRLEVFVRGTDNALWHNWQTAADTNTWSS 113
PLU0732   QTAPLPNSSWAGWNKFNGVVTSKPAVHRNSDGRLEVFVRSTDNALWHNWQTAADTNTWSS 120

PLL       WQPLYGGITSNPEVCLNSDGRLEVFVRGSDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP 173
PLU0732   WQPLYGGITSNPEVCLNSDGRLEVFARGTDNALWHIWQTAAHTNSWSNWKSLGGTLTSNP 180

PLL       AAHLNADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVINNCDGRLEV 233
PLU0732   AAHINADGRIEVFARGADNALWHIWQTAAHTDQWSNWQSLKSVITSDPVVIGNCDGRLEV 240

PLL       FARGADSTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGQLEVFARGADNTLWRT 293
PLU0732   FARGADNTLRHISQIGSDSVSWSNWQCLDGVITSAPAAVKNISGRLEVFARGADNTLWRT 300

PLL       WQTSHNGPWSNWSSFTGIIASAPTVAKNSDGRIEVFVLGLDKALWHLWQTTSSTTSSWTT 353
PLU0732   WQTSQNGPWSNWSSFTGIIASAPTVAKNSDGRIEVFVLGLDKALWHLWQTTSSTTSSWTT 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

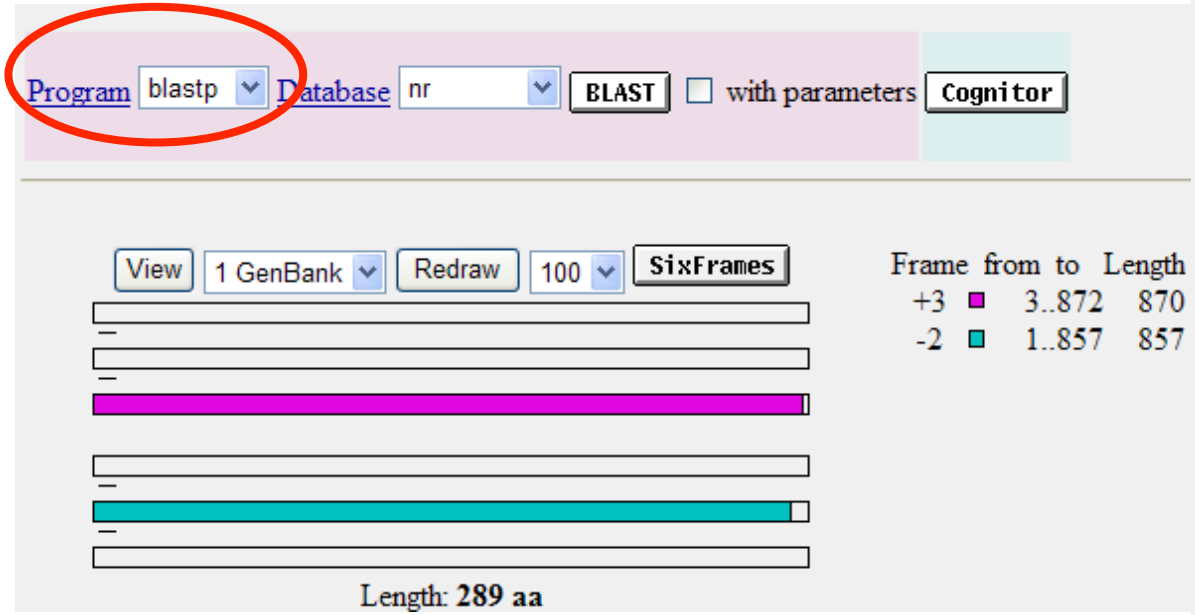
BLAST BLAST submission on ExPASy/SIB

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

ScanProsite

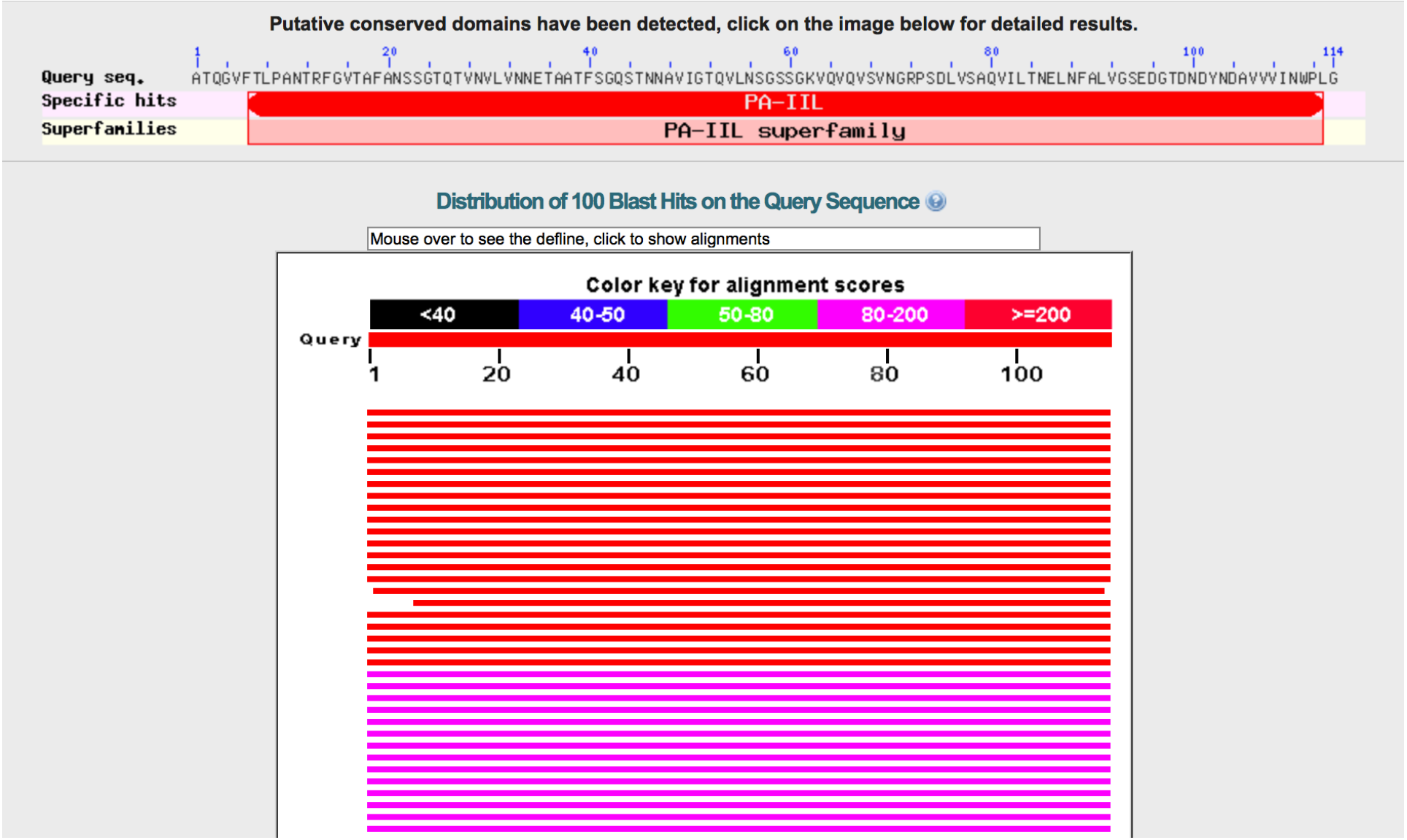
Direct Submission to SWISS-MODEL

ORF Finder (NCBI)
http://www.ncbi.nlm.nih.gov/gorf/gorf.html




```
  3 atgctggtgattgtggatgccgttaccctgctgagcgcctatccg
    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 cacctgtatgttgttagcccgggcgatgccgcgcagctgggccat
    H  L  Y  V  V  S  P  G  D  A  A  Q  L  G  H
138 aacgatagccgtctgtttaccggtctgagcccgggtgatcagctg
    N  D  S  R  L  F  T  G  L  S  P  G  D  Q  L
183 catctgcgcgaaaccgcgctggcgctgcgcgcggaagtgagcgtg
    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 gatgatctgctgcatccgagctgtcgtccgctgaaagatcattat
    D  D  L  L  H  P  S  C  R  P  L  K  D  H  Y
363 tggcgcagcgatgtgctggcggcgggcgcgaccacctgtaccgcc
    W  R  S  D  V  L  A  A  G  A  T  T  C  T  A
408 gattttgcggtgtgcgatcgtgatggcaccgtgagcggttatttt
    D  F  A  V  C  D  R  D  G  T  V  S  G  Y  F
453 cgttgggaaaccagcattgaaattgcgggcagccagccggatacc
    R  W  E  T  S  I  E  I  A  G  S  Q  P  D  T
498 aaacagccgggctttaaaccgagcagcgatcgcaatggcaacttt
    K  Q  P  G  F  K  P  S  S  D  R  N  G  N  F
543 agcctgccgccgaataccgcctttaaagcgatcttctatgcgaac
    S  L  P  P  N  T  A  F  K  A  I  F  Y  A  N
588 gcggcggatcgtcaggatctgaaactgtttattgatgatgcgccg
    A  A  D  R  Q  D  L  K  L  F  I  D  D  A  P
633 gaaccggccgccacctttgtgggtaacagcgaagatggtgtgcgt
    E  P  A  A  T  F  V  G  N  S  E  D  G  V  R
678 ctgtttaccctgaatagcaaaggtggtaaaattcgtattgaagcg
    L  F  T  L  N  S  K  G  G  K  I  R  I  E  A
723 agcgcgaacggccgtcagagcgcgaccgatgcccgtctggcgccg
    S  A  N  G  R  Q  S  A  T  D  A  R  L  A  P
768 ctgagcgcgggcgataccgtgtggctgggctggctgggcgcggaa
    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



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    LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV  67
            LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV
Sbjct  1    LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIGTQVLNSGSSGKVQVQV  60

Query  68   SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  114
            SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG
Sbjct  61   SVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  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="LPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVIG
TQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVI
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 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.**

Eukaryotické geny
Mnohobuněčná eukaryota

- **Mnohobuněčná eukaryota**

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

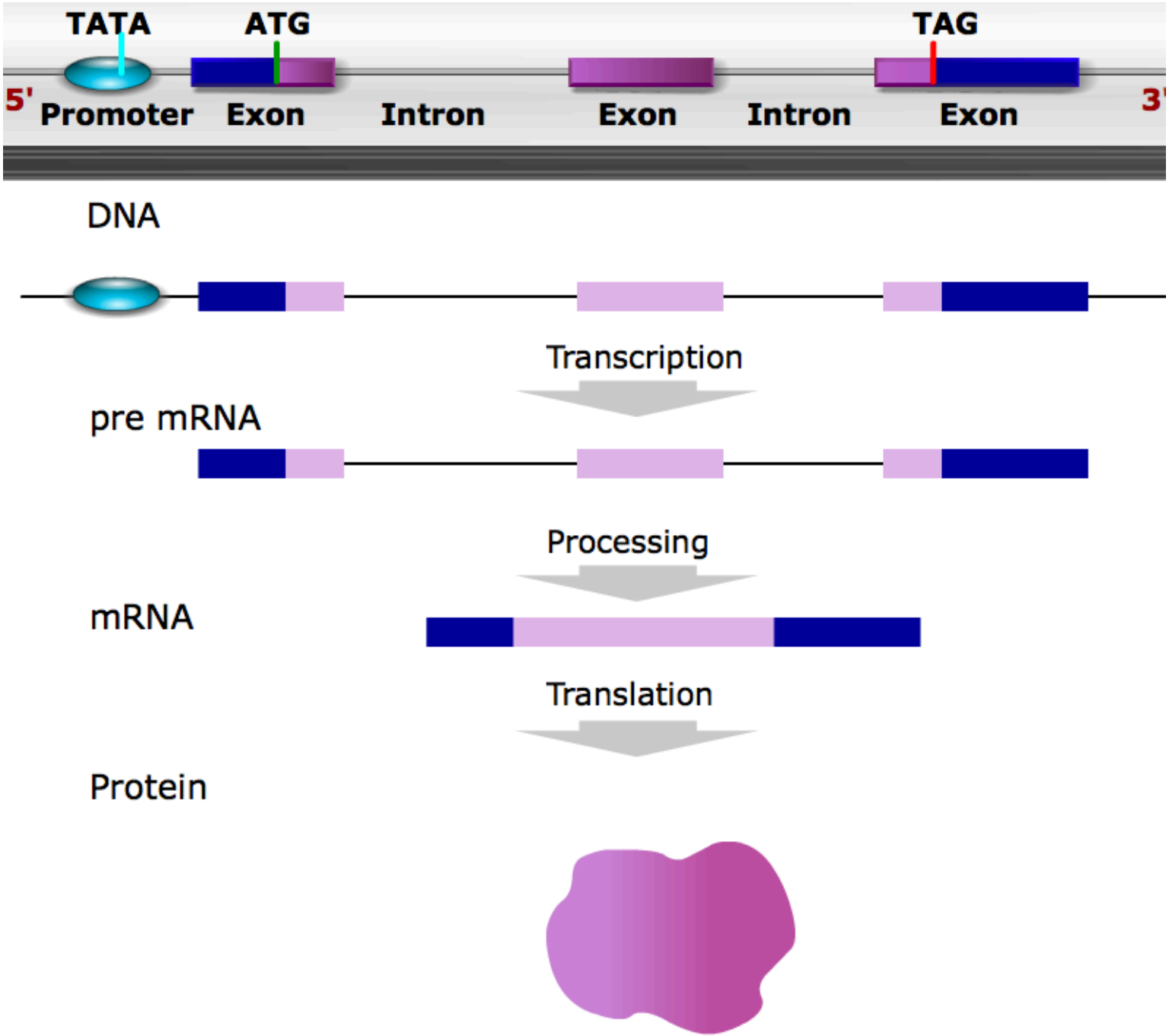
Glyceraldehyd-3-fosfát-dehydrogenasa
Candida albicans

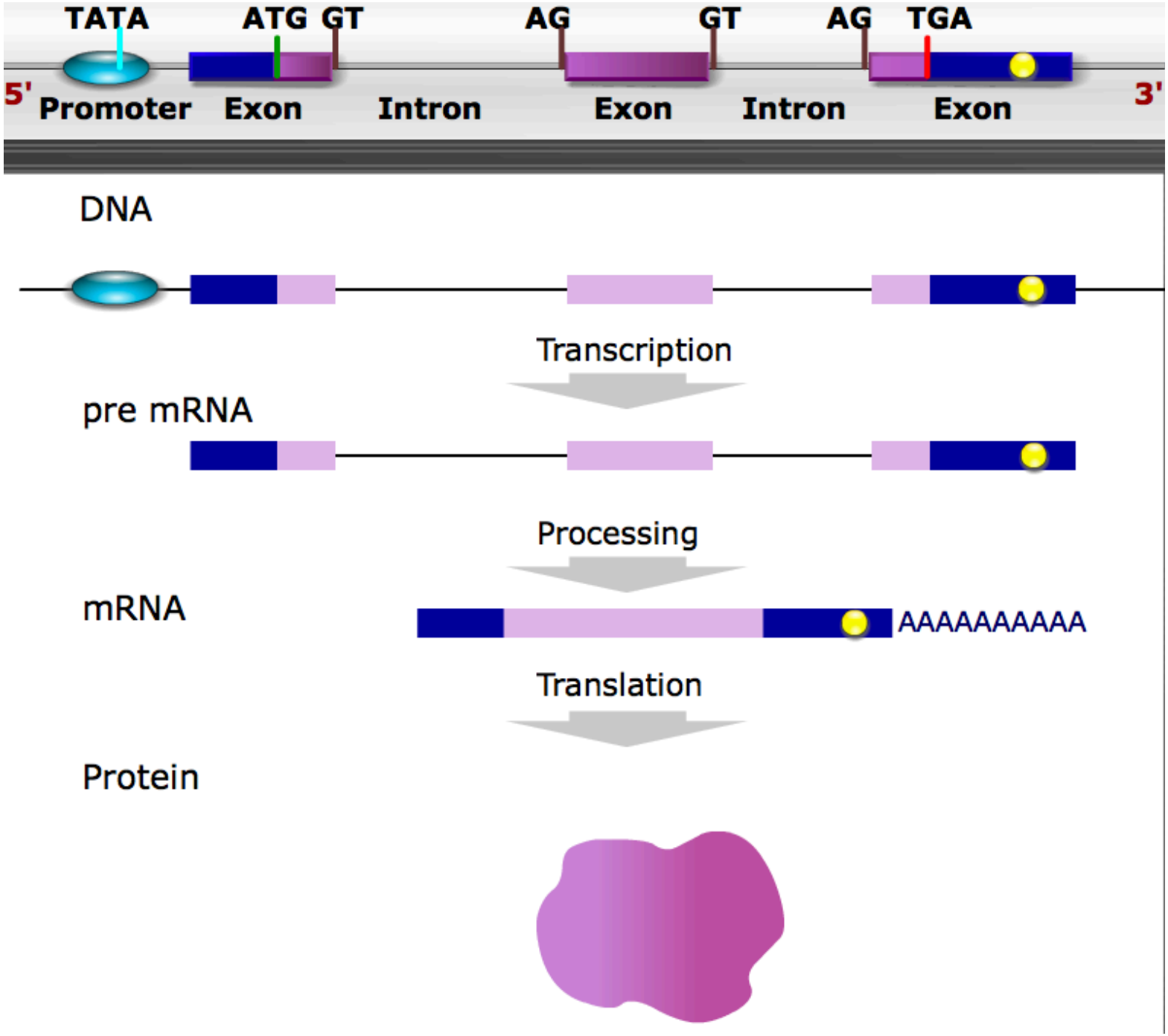
Eukaryotické geny
Mnohobuněčná eukaryota

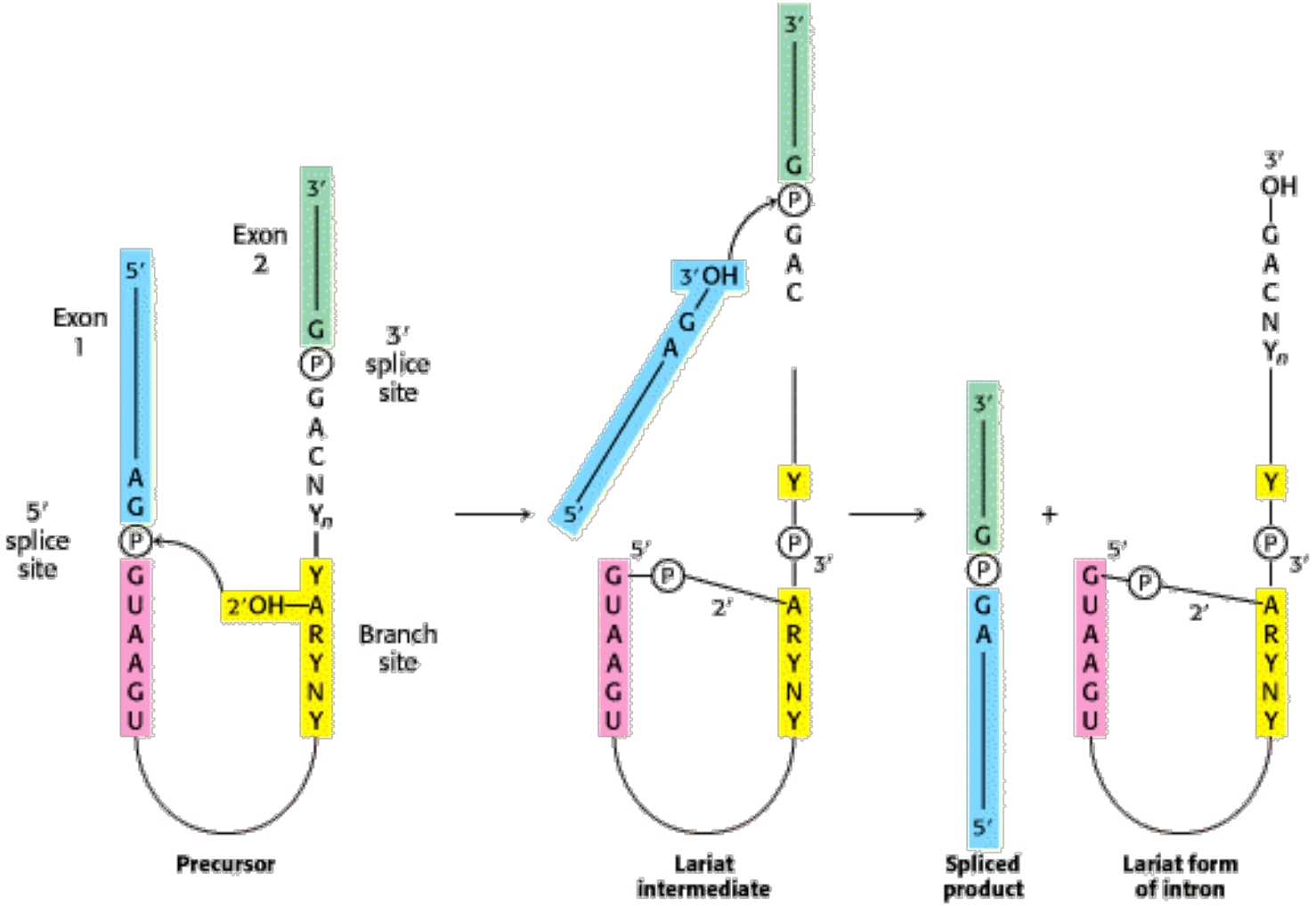
- **Mnohobuněčná eukaryota**

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

Glyceraldehyd-3-fosfát-dehydrogenasa
Homo sapiens

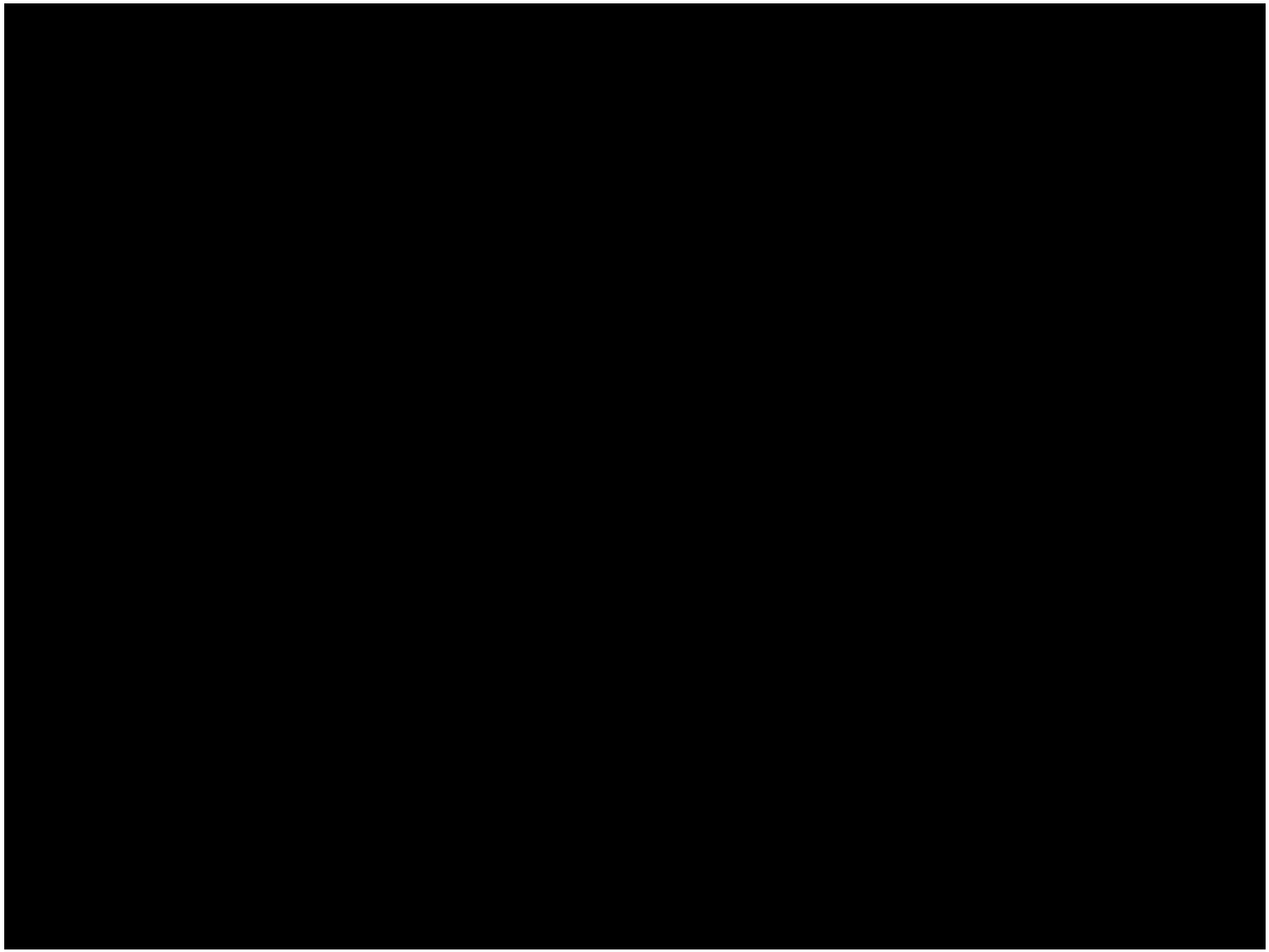

TATA
ATG
TAG
5'
Promoter
Exon
Intron
Exon
Intron
Exon
3'
DNA
Transcription
pre mRNA
Processing
mRNA
Translation
Protein


TATA
ATG
GT
AG
GT
AG
TGA
5'
Promoter
Exon
Intron
Exon
Intron
Exon
3'
DNA
Transcription
pre mRNA
Processing
mRNA
AAAAAAAAAA
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 useky – 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)

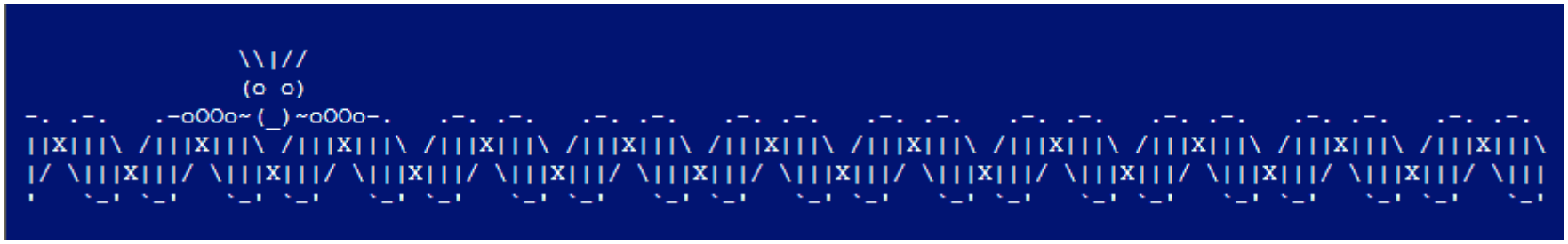
  Primární analýza velkých úseků eukaryotické genomové DNA

GeneScan

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

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA



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.

Úkol – deadline 21.dubna !

- DEFINITION fucose-specific lectin [Arthroderma otae CBS 113480].

- ACCESSION XP_002846975
- VERSION XP_002846975.1
- DBSOURCE REFSEQ: accession XM_002846929.1