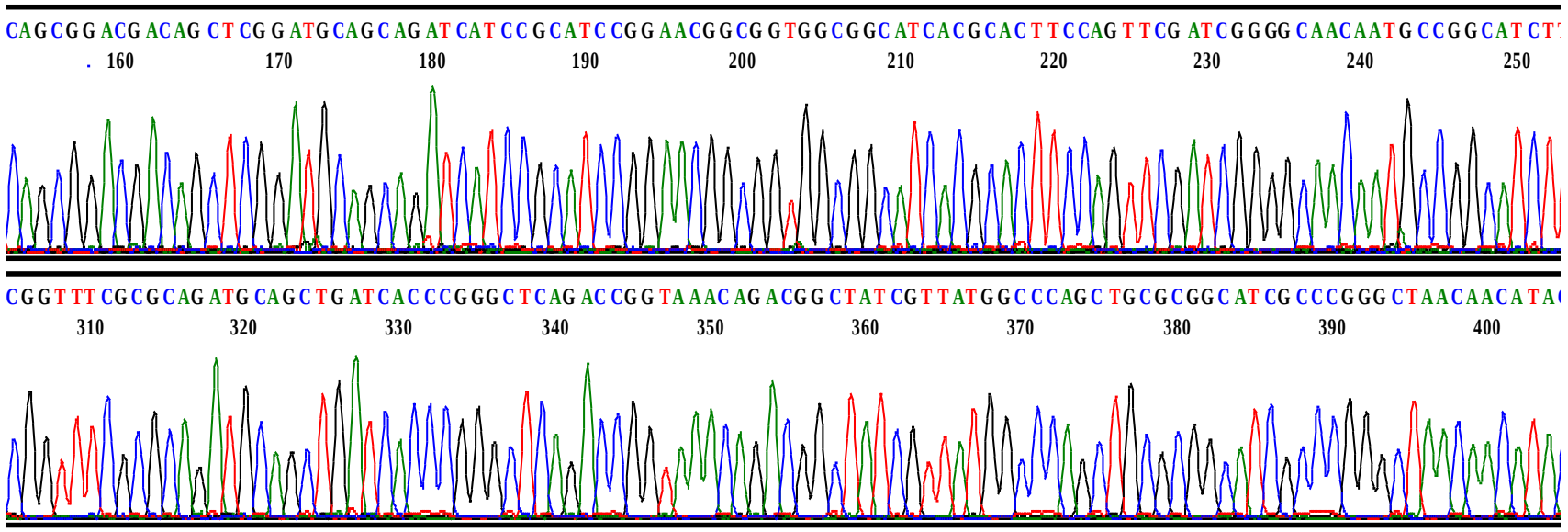
Predikce genů

C2131 Úvod do bioinformatiky, jaro 2024

Lenka Malinovská

Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

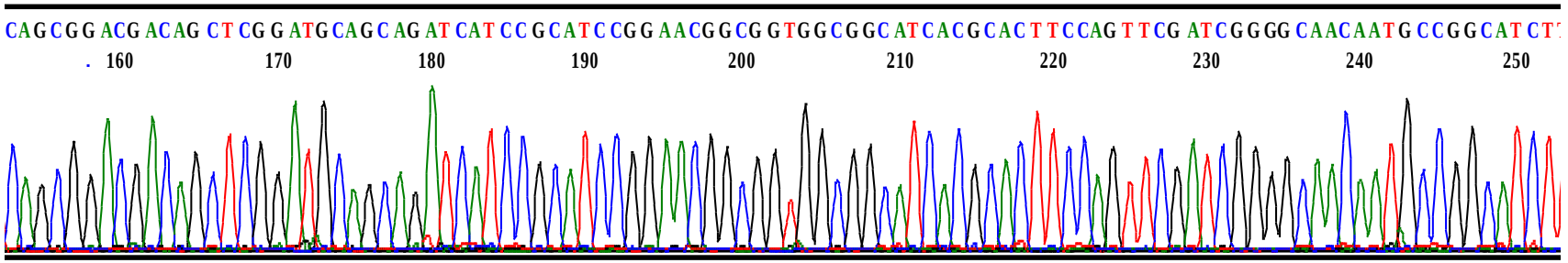
TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

Molekulárně biologická data



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

Identifikace a anotace genů a proteinů

Predikce genů

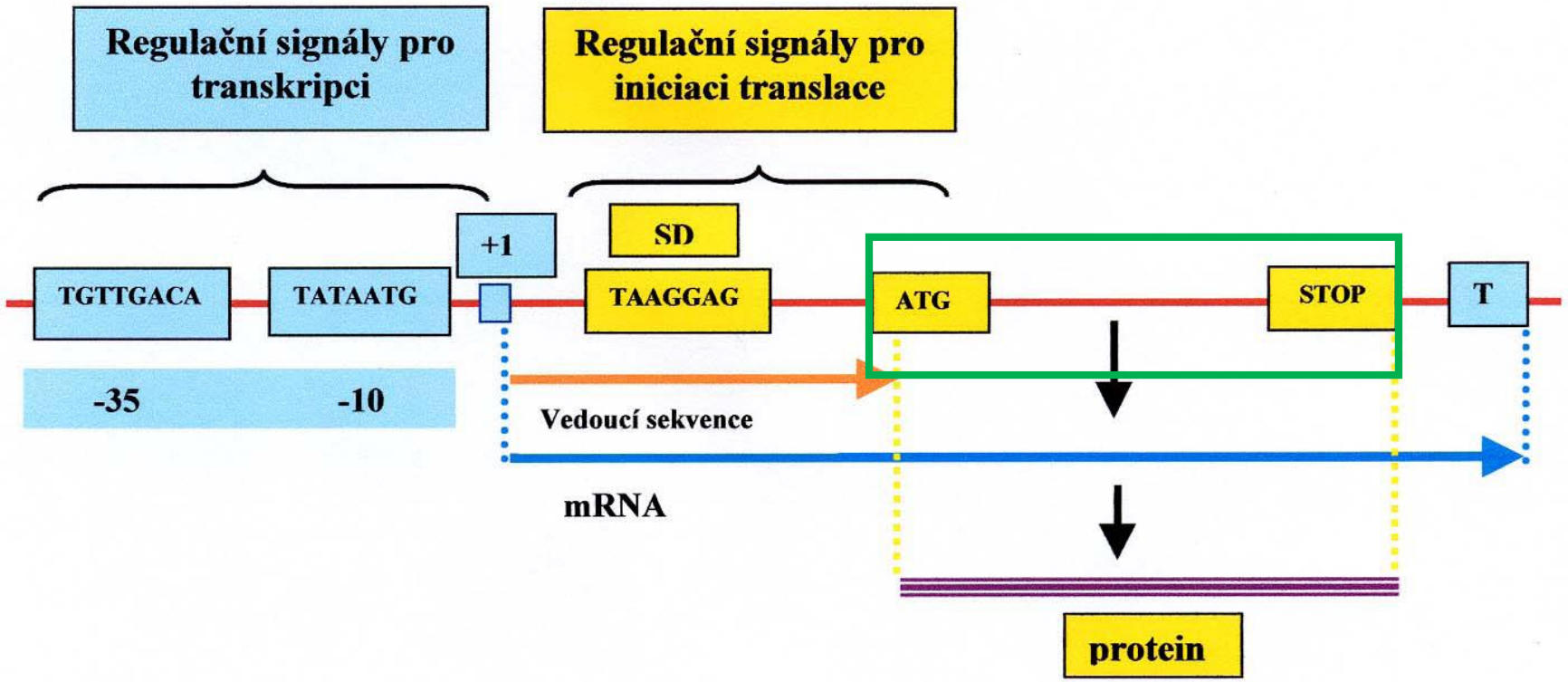
Table 1
Software commonly used for bacterial genome annotation and comparison

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 je **prvním krokem** v anotaci genů a genomů.

Predikce genů

- Predikce genů je **prvním krokem** v anotaci genů a genomů.
- Zahrnuje identifikaci **ORF** - otevřených čtecích rámců.

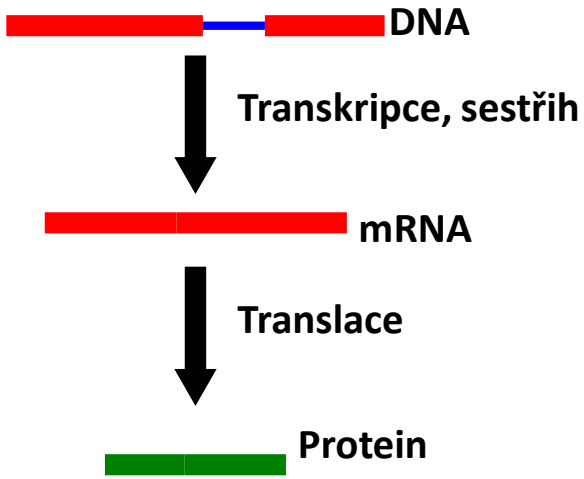


Predikce genů

- Predikce genů je **prvním krokem** v anotaci genů a genomů.
- Zahrnuje identifikaci **ORF** - otevřených čtecích rámců (Jako predikce „genů“ se mnohdy označuje právě pouze predikce ORF).
- V případě eukaryot (složené geny) predikce zahrnuje také identifikaci **exonů/intronů**, tj. míst sestřihu.

Introny *– DNA-sekvence složeného genu, jejichž přepisy se při posttranskripční úpravě sestřihem z primárního transkriptu vyštěpují a nepřecházejí tedy do výsledné mRNA.*
Exony *– DNA-sekvence složeného genu, které se při sestřihu nevyštěpují, ale spojují a přecházejí do výsledné mRNA.*

Rosypal, 2003



Predikce genů

Sekvence v databázi:

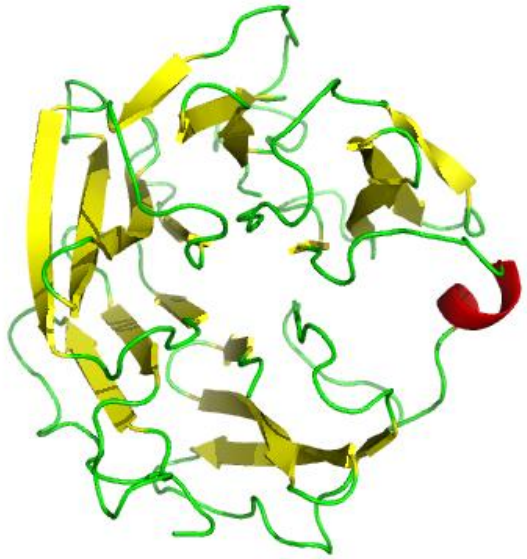
```
ATGAAATTGCTTCACTTCGTCCTGTTTTTCCAGGCCTCACTCCTTCCAGTAGGCTCCCTCGCGCAAGAGG
GTGGAAACGTCACGGATTCTGAAGTCCAGACTATCCCTGGTTGGTTCTCACTTCTCGTCCTTCCCACCGC
CGTTTAGATCATCTTGGCGGGTTGAAGATCGTGGGATGATACAAGAGAGATTGTCTGACTATTATTTCTT
TAGGTACCGGGATAGCAGCCGTCAACTCTGTCAATTTGTTGCGGATCTATAGCCAAGACATACTCGGTGG
CATCCGCGAGGCCAGATTTGAAGGCTATTGGAGCGGAGGACTTTTGAACGACACGATTGCAAAGGCCAAG
ACCAATTCATCAATTGCTGCCGCCTCTGATGATCTAGAACTAGTAAGCAGTACTATATCACTCATGCATC
CGGAGTATGCACAAGTATTTACTTCAATCTTAAGATCCGCGTCTACTATCTCTCACCGAATAACACTCTA
GGCGAAGCAGCATCTGATTCCCAGGGGGGGTGGTACACTGGCTCCCTCAACCACTATCAGTTTCGGGTTG
CATCTCATTCGAGGCTGGCTGCAGTATTTGTTCCCGGAATTCGAAGGCCAAGCTTACGTGTATATGCCCA
GCTCCCGGATAACAGTGTACAGGAGTTTGGATATGATGGTAAGCCGCCATGATCTCCAGTCCCGTCCTGC
TCCCCATATTTCAGTTAGAACCATTAAAGTTGCTAATCATCCGAAAGTTGGTCGTGGATGGGAACGCCTC
GATAATTTTGGCCCTGCTCTACCTGGTACAGCTATAGCAGCGTTAACATATACCACTGGTCTACGAAGAT
CAGACATTCGGTAAAACTTATCTCTCCCCCTCCTATCCACCTGAAACTTCTTGAGAACGACTCGGCTTTC
CCGCTTCTCTGCTGTATTCACCGAAACTAATGACTCTTTCAATCTTGTTTCCCCCCACAGTGTCTATTTC
CAAGCCACAAACCGCCACGTCGTTGAAAGGATTTATGATAGCCGATCGTGGAGTGATGGAGGCATCGTCG
TGCGAACCGCCAAGCCACGAACTCCTCTAGCTGCCACCAGCTTCTTGATGGTACCAGGAAATCCTCAGAG
TGTCCGAGTGTACTATGGCACCGAAGACAACCGTATCCTTGAAAAGGGGACCGAAGGCGGCCTTTACTGG
TACGATGGCGCATTCGAGCATTCAGTCATCCCTGGTTCTCAGGTAGCTACCGTAGATTGGGGAAATGGAG
GAGACTTCAATATCAGAGTCTATATTCAAGACGGAGCATTTAAGAACGGGATAAGTGAATGGGCTTGGTT
CCGCCGCTTATGGCGCCGAGGAGTCATTGCTATCCCTCCTGCATAA
```

Annotation Pipeline :: Eukaryotic Annotation Propagation Pipeline

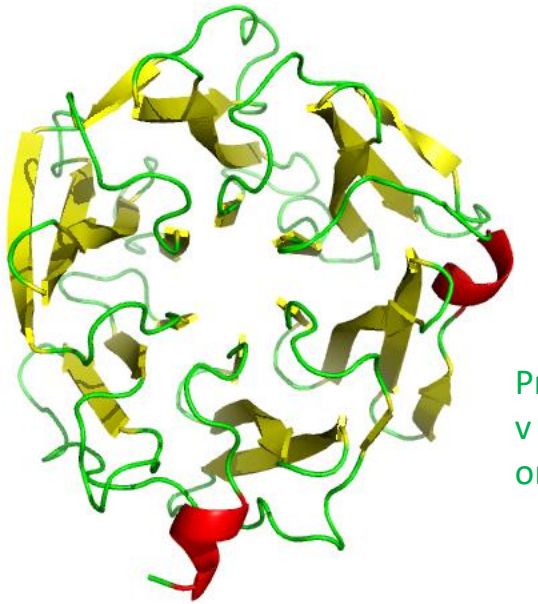
Neúspěch při expresi syntetického genu
v hostitelském organismu, protein nebyl produkován

Srovnání s homologním charakterizovaným
genem/proteinem na sekvenční a strukturní úrovni,
úprava exonů/intronů

```
ATGAAACTGCTGCATTTTGTGCTGTTTTTTCAGGCGAGCCTGCTGCCGGTGGGCAGCCTG
GCGCAGGAAGGCGGCAACGTGACCGATAGCGAAGTGCAGACCATTCCGGGCACCGGCATT
GCGGCGGTGAACAGCGTGAACCTGCTGCGCATTTATAGCCAGGATATTCTGGGCGGCATT
CGCGAAGCGCGCTTTGAAGGCTATTGGAGCGGCGGCCTGCTGAACGATACCATTGCGAAA
GCGAAAACCAACAGCAGCATTGCGGCGGCGAGCGATGATCTGGAACTGATTCGCGTGTAT
TATCTGAGCCCGAACAACACCCTGGGCGAAGCGGCGAGCGATAGCCAGGGCGGCTGGTAT
ACCGGCAGCCTGAACCATTATCAGTTTCGCGTGGCGAGCCATAGCCGCCTGGCGGCGGTG
TTTGTGCCGGGCATTCGCCGCCCGAGCCTGCGCGTGTATGCGCAGCTGCCGGATAACAGC
GTGCAGGAATTTGGCTATGATGTGGGCCGCGGCTGGGAACGCCTGGATAACTTTGGCCCG
GCGCTGCCGGGCACCGCGATTGCGGCGCTGACCTATACCACCGGCCTGCGCCGCAGCGAT
ATTCGCGTGTATTTTCAGGCGACCAACCGCCATGTGGTGGAACGCATTTATGATAGCCGC
AGCTGGAGCGATGGCGGCATTGTGGTGCGCACCGCGAAACCGCGCACCCCGCTGGCGGCG
ACCAGCTTTCTGATGGTGCCGGGCAACCCGCAGAGCGTGCGCGTGTATTATGGCACCGAA
GATAACCGCATTCTGGAAAAAGGCACCGAAGGCGGCCTGTATTGGTATGATGGCGCGTTT
GAACATAGCGTGATTCCGGGCAGCCAGGTGGCGACCGTGGATTGGGGCAACGGCGGCGAT
TTTAACATTCGCGTGTATATTCAGGATGGCGCGTTTAAAAACGGCATTAGCGAATGGGCG
TGGTTTCGCCGCCTGTGGCGCCGCGGCGTGATTGCGATTCCGCCGGCGTAA
```



Predikce struktury
proteinu



Protein úspěšně produkován
v prokaryotickém hostitelském
organismu

Predikce genů

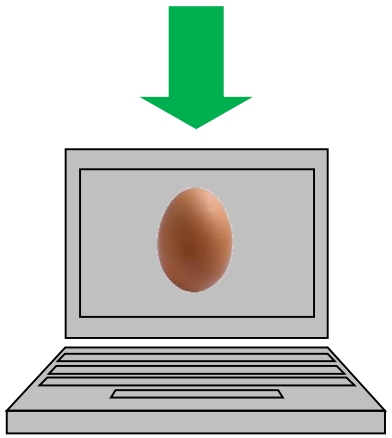
- Predikce genů je **prvním krokem** v anotaci genů a genomů.
- Zahrnuje identifikaci **ORF** - otevřených čtecích rámců (Jako predikce „genů“ se mnohdy označuje právě pouze predikce ORF).
- V případě eukaryot (složené geny) predikce zahrnuje také identifikaci **exonů/intronů**, tj. míst sestřihu. **Problematická**, vzniká velké množství chyb.
- Predikce genů se velmi často soustředí na geny kódující proteiny.
- Predikce genů u prokaryot funguje **výrazně lépe** než u eukaryot (souvislost s organizací genomu prokaryot).

Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC



NCBI National Center for Biotechnology Information

EMBL-EBI

DDBJ DNA Data Bank of Japan

GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

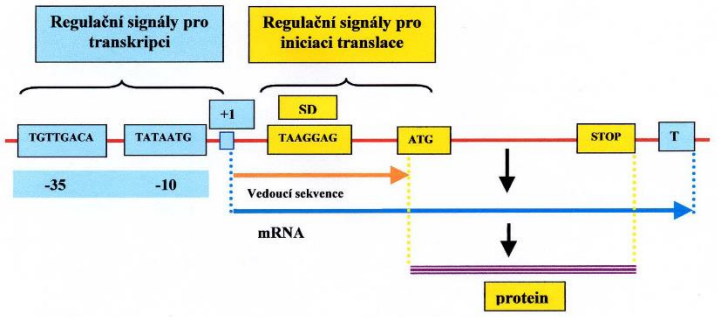
LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKGGKIRIE
IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPRE--ATLNSGNGKIRFE
LPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVI
LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGKGKVRVV
lPPn-aFg---lanaad-Qtiklfid-D-p-PAAtfkgag-----l-t-tlnSgnGkiRve

ASANGRQSATDARLAPLSAGD------TVWLGWLGAEDGADADYNDGIVILQWPIT
VSVNGKPSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG
VMANGRPSRLGSRQVDIFKKS--------YFGIIGSEDGADDDYNDGIVFLNWPLG
VTANGKPSKIGSRQVDIFKKT--------YFGLVGSEDGGDGDYNDGIAILNWPLG
vsaNGrpSat--R---ifkks------tvyfGivgsEDGaDaDYNDGIviLqWPig

Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.

Regulační a signální sekvence: startovní/stop kodon, sestřihové signály, RBS (vazebné místo pro ribozom), polyadenylační signál.
Kodon=triptet (délka genu je v násobcích tří).
Nukleotidové složení kódujících a nekódujících oblastí se liší.



Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.
- **Metody založené na homologii** – sekvenční podobnost se známými geny/proteiny. ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**) = nejspolehlivější predikce. Problém – unikátní geny bez známých homologů (většinou nejzajímavější).
- Kombinace obou postupů

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.

Table 1 Some prokaryotic genomes

Organism	Domain	Size (base pairs)	Genes	Comments
Nanoarchaeum equitans	*Archaea*	490 885	552	Smallest known cellular genome
Mycoplasma genitalium	*Bacteria*	580 070	470	Smallest genome among *Bacteria*; human pathogen
Chlamydia trachomatis	*Bacteria*	1 042 519	894	Intracellular parasite of humans
Aquifex aeolicus	*Bacteria*	1 551 335	1544	Hyperthermophile, autotroph
Methanothermobacter thermoautotrophicus	*Archaea*	1 751 377	1855	Methanogen, thermophile
Halobacterium salinarium	*Archaea*	2 571 010	2630	Extreme halophile
Sulfolobus solfataricus	*Archaea*	2 992 245	2977	Hyperthermophile, acidophile
Bacillus subtilis	*Bacteria*	4 214 810	4100	Produces endospores
Pseudomonas aeruginosa	*Bacteria*	6 264 403	5570	Metabolically versatile; can be a pathogen
Bradyrhizobium japonicum	*Bacteria*	9 105 828	8317	Nitrogen-fixing bacterium; forms root nodules on soybean plants
Escherichia Coli	*Bacteria*	4 639 221	4288	Model organism for molecular biology

How small can a genome get and still run a living organism? Researchers now say that a symbiotic bacterium called *Carsonella ruddii*, which lives off sap-feeding insects, has taken the record for smallest genome with just 159,662 'letters' (or base pairs) of DNA and 182 protein-coding genes. At one-third the size of previously found 'minimal' organisms, it is smaller than researchers thought they would find.

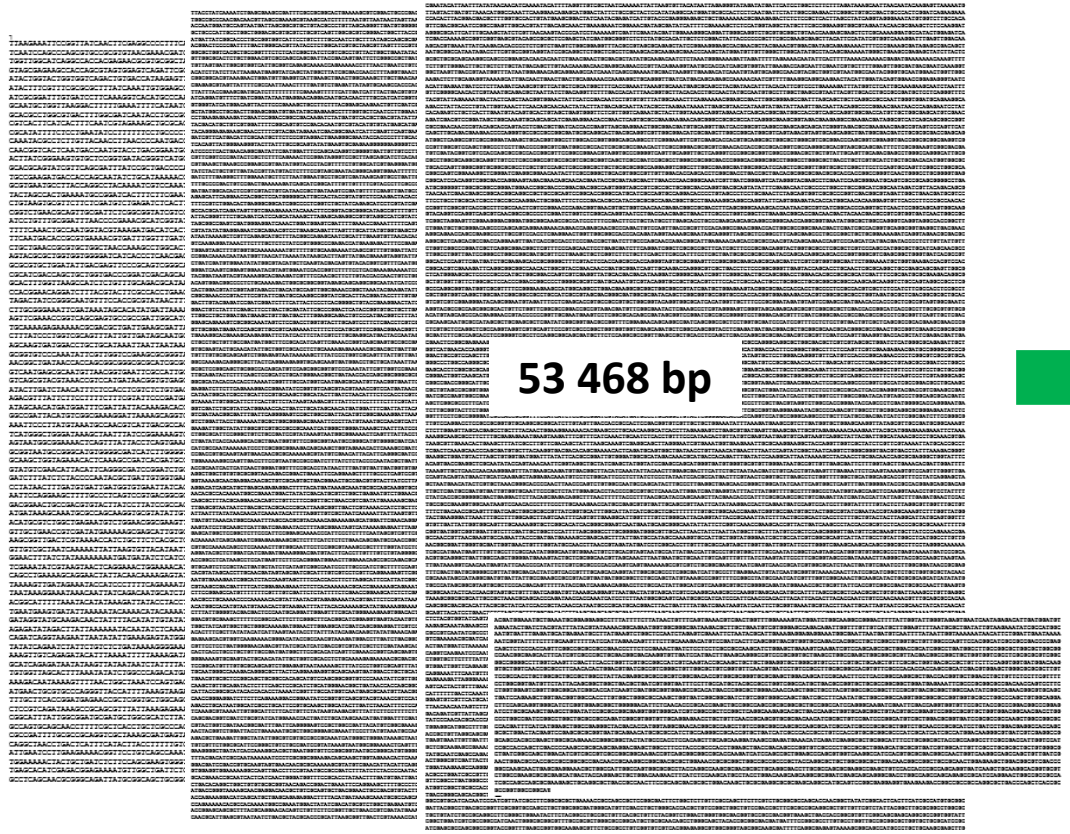
https://1url.cz/MK8Kx

Bacteriology

Michael T Madigan, *Southern Illinois University, Carbondale, Illinois, USA*
Deborah O Jung, *Southern Illinois University, Carbondale, Illinois, USA*

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.



GeneMarkS

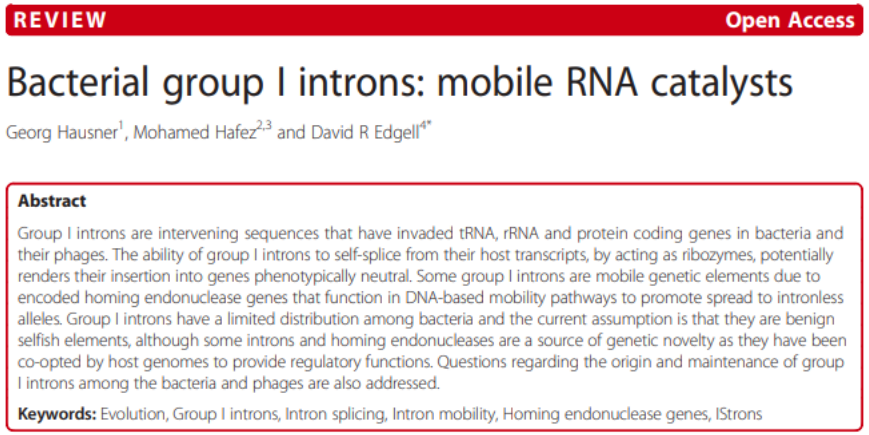
Predikováno 56 genů (ORF)

Predicted genes

Gene #	Strand	LeftEnd	RightEnd	Gene Length	Class
1	-	1	1515	1515	1
2	+	1694	2359	666	1
3	+	2739	4025	1287	1
4	-	4087	4293	207	2
5	+	4618	5175	558	1
6	-	5165	5905	741	1
7	+	6086	7663	1578	1
8	+	7980	8390	411	1
9	+	8659	11067	2409	2
10	+	11100	11438	339	1
11	-	11567	11947	381	1
12	+	12036	12896	861	1
13	+	13004	13969	966	1
14	+	14077	14373	297	1
15	+	14688	15659	972	1
16	-	15777	16586	810	1
17	-	16639	17652	1014	1
18	+	18318	19328	1011	1
19	-	19468	19977	510	1
20	+	20206	20763	558	1
21	-	20753	21493	741	1
22	+	21674	23251	1578	1
23	+	23568	23978	411	1
24	-	24333	24542	210	2
25	-	24663	25337	675	2
26	-	25334	25777	444	2
27	-	25857	25985	129	1
28	+	26345	26908	564	1
29	-	26913	27311	399	2
30	+	27310	27867	558	1
31	-	27857	28597	741	1
32	+	28778	30355	1578	1
33	+	30672	31082	411	1
34	+	31418	32542	1125	2
35	+	32598	33749	1152	2
36	-	33731	34219	489	1
37	-	34224	35651	1428	1
38	+	35750	36097	348	1
39	-	36151	36669	519	1
40	-	36712	37302	591	1
41	-	37299	38456	1158	1
42	-	38482	39597	1116	1
43	-	39680	40657	978	1
44	+	40665	40958	294	2
45	+	41018	41827	810	2
46	+	41983	42534	552	2
47	+	42531	43736	1206	1
48	+	43807	44532	726	1
49	+	44846	45520	675	1
50	+	45706	46338	633	2
51	+	46823	47128	306	2
52	+	47366	47719	354	1
53	+	47716	48165	450	1
54	+	48432	49709	1278	1
55	+	50339	52468	2130	1
56	+	52475	>53467	993	1

Predikce genů u prokaryot

- **Prokaryotické genomy**: malé (0,5 až 10 Mbp) a kompaktní, vysoká hustota genů, 90 % genomu je kódující, jeden gen připadá přibližně na 1000 nukleotidů.
- **Prokaryotické geny:** ORF je nepřerušovaný úsek DNA mezi startovním kodonem (ATG, GTG, TTG, CTG) a stop kodonem (TAA, TGA, TAG). Prokaryotické geny neobsahují introny (Dobře, můžou obsahovat introny).



REVIEW Open Access

Bacterial group I introns: mobile RNA catalysts

Georg Hausner[1], Mohamed Hafez[2,3] and David R Edgell[4*]

Abstract

Group I introns are intervening sequences that have invaded tRNA, rRNA and protein coding genes in bacteria and their phages. The ability of group I introns to self-splice from their host transcripts, by acting as ribozymes, potentially renders their insertion into genes phenotypically neutral. Some group I introns are mobile genetic elements due to encoded homing endonuclease genes that function in DNA-based mobility pathways to promote spread to intronless alleles. Group I introns have a limited distribution among bacteria and the current assumption is that they are benign selfish elements, although some introns and homing endonucleases are a source of genetic novelty as they have been co-opted by host genomes to provide regulatory functions. Questions regarding the origin and maintenance of group I introns among the bacteria and phages are also addressed.

Keywords: Evolution, Group I introns, Intron splicing, Intron mobility, Homing endonuclease genes, IStrons

Group II introns in the bacterial world

Francisco Martínez-Abarca and Nicolás Toro*
Grupo de Ecología Genética, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Profesor Albareda 1, 18008 Granada, Spain.

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

- **Prokaryotické genomy**: malý obsah nekódujících úseků umožňuje „manuální“ identifikaci ORF.

1) **Překlad prokaryotické DNA do proteinové sekvence.**

2) **Identifikace potenciálních ORF.**

3) **Ověření spolehlivosti predikce – je identifikovaný ORF skutečně součástí genu?**

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

1) Překlad prokaryotické DNA do proteinové sekvence.

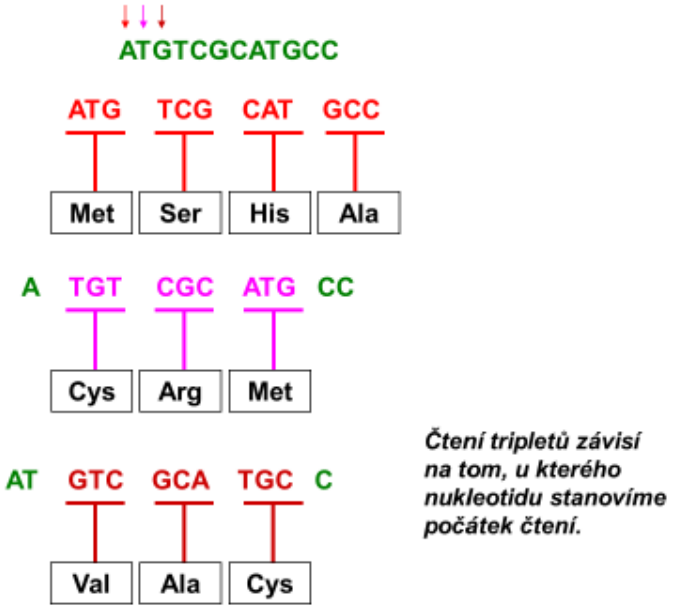
The table shows the 64 codons and the amino acid for each. The direction of the mRNA is 5' to 3'.

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

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

1) Překlad prokaryotické DNA do proteinové sekvence.



Čtení tripletů závisí na tom, u kterého nukleotidu stanovíme počátek čtení.

5´ ATGCGCAGGAATGCATAG 3´
3´ TACGCGTCCTTACGTATC 5´

Překlad DNA sekvence – od 5´konce

5´ ATGCGCAGGAATGCATAG 3´

5´ CTATGCATTCCTGCGCAT 3´

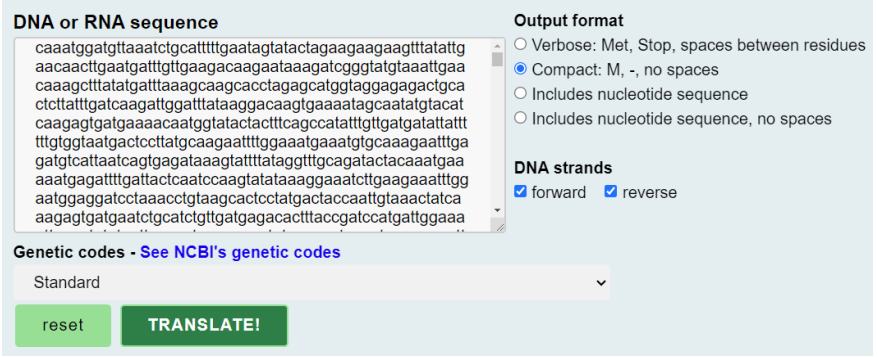
Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

1) Překlad prokaryotické DNA do proteinové sekvence.

- **Translate (ExPASy)**

  https://web.expasy.org/translate/



- **ORF Finder (NCBI)**

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

Translate

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

DNA or RNA sequence

caaatggatgttaaatctgcatttttgaatagtatactagaagaagaagtttatattg
aacaacttgaatgatttgttgaagacaagaataaagatcgggtatgtaaattgaa
caaagctttatatgatttaaagcaagcacctagagcatggtaggagagactgca
ctcttatttgatcaagattggatttataaggacaagtgaaaatagcaatatgtacat
caagagtgatgaaaacaatggtatactactttcagccatatttgttgatgatattatt
tttgtggtaatgactccttatgcaagaattttggaaatgaaatgtgcaaagaatttga
gatgtcattaatcagtgagataaagtattttataggtttgcagatactacaaatgaa
aaatgagattttgattactcaatccaagtatataaaggaaatcttgaagaaatttgg
aatggaggatcctaaacctgtaagcactcctatgactaccaattgtaaactatca
aagagtgatgaatctgcatctgttgatgagacactttaccgatccatgattggaaa

Output format

- ○ Verbose: Met, Stop, spaces between residues
- ◉ Compact: M, -, no spaces
- ○ Includes nucleotide sequence
- ○ Includes nucleotide sequence, no spaces

DNA strands

☑ forward ☑ reverse

Genetic codes - See NCBI's genetic codes

Standard

reset **TRANSLATE!**

Standard
Vertebrate mitochondrial
Yeast mitochondrial
Mold, protozoan and coelenterate mitochondrial, mycoplasma/spiroplasma
Invertebrate mitochondrial
Ciliate, dasycladacean and hexamita nuclear
Echinoderm and flatworm mitochondrial
Euplotid nuclear
Alternative yeast nuclear
Ascidian mitochondrial
Alternative flatworm mitochondrial
Blepharisma nuclear
Chlorophycean mitochondrial
Trematode mitochondrial
Scenedesmus obliquus mitochondrial
Pterobranchia mitochondrial

Standard

reset **TRANSLATE!**

Translate

5'3' Frame 1

AEMPSVYAR-PERC-RQF-P-KT-KIFRYAIR-GSAAFIGGQQPYVDRTERFLQGIYRG-LPVPAFCCPEEKYWKRSEIMVFCDGKQAKRRRKAGYH
EHQGTKRSYSENTGNNAAGILCQYCKPICGSWNKKSKYRF-PET-LFIR-LCRRRKQ-ICSNCSQIDCKKTWSNSI-PIIPIWRLWSWKNTPGTGCW
S-SKKSVP--SCTLFII-KVYPAIYLCCQSTQTD-ICEFLSDGRCTDY--YPVLIRKISYAGQLLPYF-SPASERKTDYPYFR-GACRHYGYPRQNC
FPFQMGTFCRNQIAGPFYKKTDH-R-TKQRRNCSSGRYAGLPCCRSQNQCKRTDWSN-LSDRLLYSI-ERP-S-IAERND-QNCCQPEKSHQYSLHP
GSGL-LFRN-KRAAAFKNKKKRDRITKTACDVFLKRIHQFHLY-NR-RNGRKRPFYSNVCL-YHQRRIEN--RNQEIRKRSY-KNQTVNNH-ALK-L
K-RIVVFFIFYLVLSLKVFFINTHMKILMVCLGNICRSPLAEGIMKTKVPDNFVVDSAGTISLHEGEHPDKRMNSLRSGEYFGETNQMVNLEGLTIT
DTEYTHPYVDWHYHENAYFTFLLQGTMTEGNRKETYGCSAGTLLYHHWEDPHYN

5'3' Frame 2

QKCLQFMRDNLNAAEDNSDLKKLEKSFDMLFDKVQPLSLVANNLTLIVPSDFYKEYIEDNYLSLLSAALKKNIGKGVKLWYSVMENRPKGEEKPVTM
NIKGQSVPTPKTQETMPQGFSANIVNPFVVPGIRKVNIDSNLKPDYSFDSYVEGESNKFAATVARSIAKRPGATAFNPLFLYGGYGVGKTHLGQAVG
LEVKNQFPDKVVLYLSSEKFIQQFISAAKAHKQTEFANFYQMVDVLIIDDIQFLSGKSATQDSFFHIFDHLHQNGKQIILTSDKAPADIMDIQDRIV
SRFKWGLSAEIKSPDLSTRRQIIEDKLSRDGIVLPGDMLDFLAAEAKTNVRELIGVINSVIAYSTVYKRDLSLELLKETINRIAANQKKVINIPYIQ
EVVCDYFGIKKEQLLSKTRKREIALPRQLAMYFSKEFTNSTFTKIGEEMGGKDHSTVMYACDTIKDVSKIDKEIKKYVKDLTERIKQ-IITEL-NN-
NEESLYSSFFI-FCR-KSFL-ILI-KY-WFVWEIYAEVRWQKEL-KQKYRTTLW-TQQEPFHCTKENIRIKE-TAFVAANILGKPIKWSIWKD-PSP
IRSILILT-TGIIMKMLISLFFYRAP-QKETEKKPTAVLQGHYCIIIGKIRTII

5'3' Frame 3

RNAFSLCAIT-TLLKTILTLKNLKNLSICYSIRFSRFHWWPTTLR-SYRAIFTRNI-RIITCPCFLLP-RKILEKE-NYGIL-WKTGQKAKKSRLP-
TSRDKAFLLRKHRKQCRRDSLPIL-THLWFLE-EK-I-ILT-NLTIHSIVM-KEKAINLQQL-PDRLQKDLEQQHLTHYSYMEVMELEKHTWDRLLV
LK-KISSLIKLYFIYHLKSLSSNLSLLPKHTNRLNLRISIRW-MY-LLMISSSYQENQLRRTASSIFLITCIRTENRLSLLQIRRLQTLWISKTELF
PVSNGDFLQKSNRRTFLQEDRSLKIN-AETELFFREICWTSLLPKPKPM-EN-LE-LTQ-SLTLQYIRETLVLNC-KKRLTELLPTRKKSSIFLTSR
KWFVIISELKKSSCFQKQEKERSHYQDSLRCISQKNSPIPPLLK-VKKWEEKTILQ-CMLVIPSKTYRKLIKKSRNT-KILLKESNSK-SLSFKIIK
MKNRCILHFLFSFVVKSLFYKYSYENINGLSGKYMQKSAGRRNYENKSTGQLCGRLSRNHFIARRRTSG-KNEQPS-RRIFWGNQSNGQFGRIDHHR
YGVYSSLRRLALS-KCLFHFSSTGHHDRRKQKRNLRLFCRDIIVSSLGRSAL-S

3'5' Frame 1

RL-CGSSQ--YNNVPAEQP-VSFLFPSVMVPCRRKVK-AFS--CQST-G-VYSVSVMVNPSKLTI-LVSPKYSPLRRLFILLSGCSPSCNEMVPAES
TTKLSGTFVFIIPSASGLLHIFPRQTINIFI-VFIKKTFNDKTK-KMKNTTILHFNYFKAQ-LFTV-FFQ-DLLRIS-FLYQFSIRL-WYHKHTLL-
NGLFLPFLHLF--RWNW-ILLRNTSQAVLVMRSLFFLFLKAAALF-FRNNHKPLPGCKEY--LFSGWQQFC-SFLSAIQD-GLSYIL-SKRSLS-LL
QSVLLHWFWLRQQGSPAYLPEEQFRLCLVYLQ-SVFL-KGPAI-FLQKVPI-NGKQFCLGYP-CLQAPYLK-G-SVFRSDAGDQKYGRSCPA-LIFL
IRTGYHQ-SVHLPSDRNSQIQSVCVLWQQR-IAG-TFQMINKVQLYQGTDFLLQDQQPVPGVFFQLHNLHIGIMG-MLLLQVFLQSIWLQLLQIYCF
LLLHNYRMNSQVSG-NLYLLFLFQEPQMGLQYWQRIPAALFPVFSE-ERFVP-CSW-PAFLRLLACFPSQNTIISLLFQYFSSGQQKAGTGNYPLYI
PCKNRSVRST-GCWPPMKAAEPYRIAYRKIFQVF-GQNCLQQRSGYRA-TEGIS

3'5' Frame 2

DYSADLPNDDTIMSLQNSRRFLFCFLLSWCPVEEK-NKHFHDNASLRKDEYTPYR-WSILPN-PFDWFPQNIRRYEGCSFFYPDVLLRAMKWFLLSL
PQSCPVLLFS-FLLPADFCIYFPDKPLIFSYEYL-KRLLTTKLNKK-RIQRFFILIIILKLSDYLLFDSFSKIFYVFLDFFINFRYVFDGITSIHYCR
MVFSSHFFTYFSKGGIGEFF-EIHRKLSW-CDLSFSCF-KQLLFFNSEIITNHFLDVRNIDDFFLVGSNSVNRFFQQFKTKVSLIYCRVSDH-VNYS
NQFSYIGFGFGSKEVQHISRKNNSVSA-FIFNDLSSCRKVRRFDFCRKSPFETGNNSVLDIHNVCRRLI-SKDNLFSVLMQVIKNMEEAVLRS-FS-
-ELDIINNQYIYHLIEIRKFSLFVCFGSRDKLLDKLFR--IKYNFIRELIFYFKTNSLSQVCFSNSITSI-E-WVKCCCSRSFCNRSGYSCCKFIAF
SFYITIE-IVRFQVRIYIYFSYSRNHKWVYNIGRESLRHCFLCFRSRNALSLDVHGNRLFFAFWPVFHHRIP-FHSFSNIFLQGSRKQGQVIILYIF
LVKIARYDQRKVVGHQ-KRLNLIE-HIERFFKFFKVRIVFSSVQVIAHKLKAFL

3'5' Frame 3

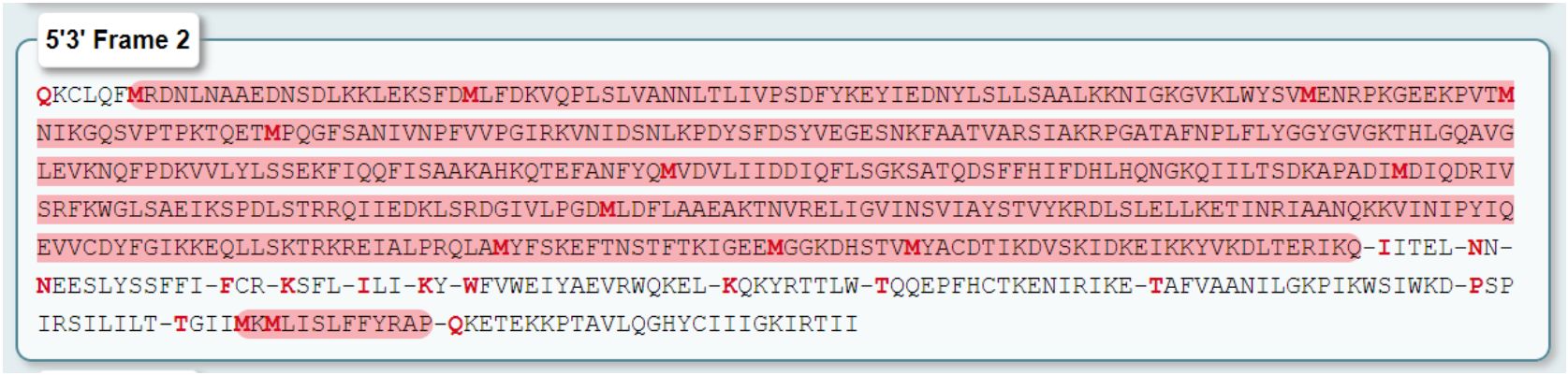
IIVRIFPMMIQ-CPCRTAVGFFSVSFCHGAL-KKSEISIFMIMPVYVRMSILRIGDGQSFQIDHLIGFPKIFAATKAVHSFIRMFSFVQ-NGSC-VY
HKVVRYFCFHNSFCQRTSAYISQTNH-YFHMSIYKKDF-RQN-IKNEEYNDSSF-LF-SSVIIYCLILSVRSFTYFLISLSIFDTSLMVSQAYITVE
WSFPPISSPILVKVELVNSFEKYIASCLGNAISLFLVFESSCSFLIPK-SQTTSWM-GILMTFFWLAAILLIVSFSNSRLRSLLYTVE-AITELITP
ISSLTLVLASAARKSSISPGRTIPSLLSLSSMICLLVERSGDLISAESPHLKRETILSWISIMSAGALSEVRIICFPF-CR-SKIWKKLSCVADFPD
KNWISSIISTSTI--KFANSVCLCALAAEINCWINFSDDK-STTLSGN-FFTSRPTACPRCVFPTP-PPYRNNGLNAVAPGLFAIDLATVAANLLLS
PST-LSNE-SGFRLESIFTFLIPGTTNGFTILAENPCGIVSCVFGVGTLCPLMFMVTGFSSPFGLFSITEYHNFTPFPIFFFRAAESRDR-LSSIYS
L-KSLGTINVRLLATNESG-TLSNSISKDFSSFLRSELSSAAFRLSRIN-RHFC

Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

2) **Identifikace potenciálních ORF.**

- Jak dlouhý má být „rozumný“ ORF? Stop kodon se v nekódující sekvenci náhodně vyskytuje přibližně každých 20 kodonů. V úvahu se tedy berou ORF delší než **třicet kodonů** (reálně i delší).
- Empirické pravidlo: Správný ORF = **nejdelší** ORF odpovídající danému úseku DNA.

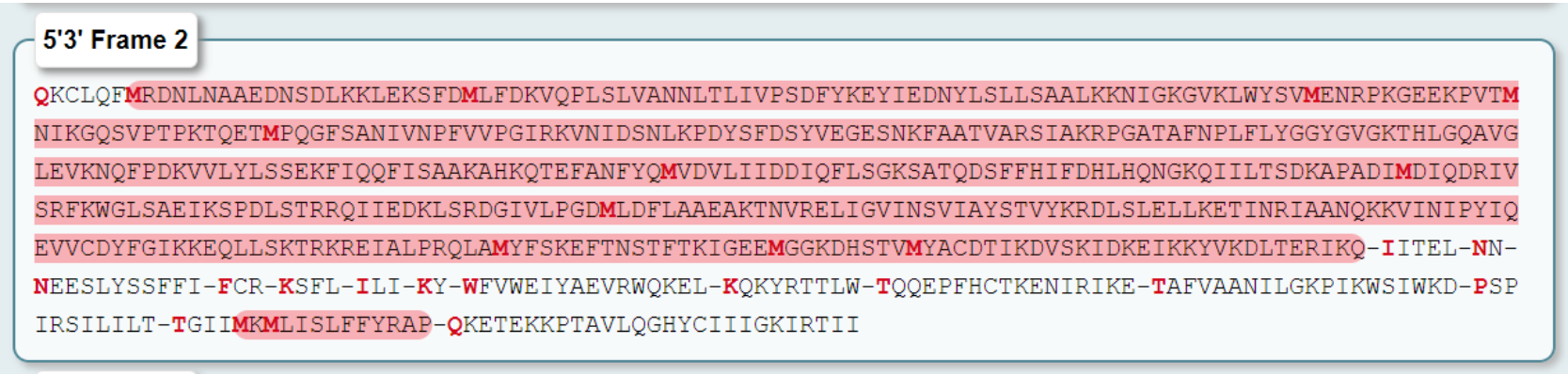


Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

3) Ověření spolehlivosti predikce – je identifikovaný ORF skutečně součástí genu?

- Kóduje ORF protein **podobný** již popsanému proteinu?
- Vyskytují se před/za ORF typické signální sekvence?
- Statistické parametry sekvence: preference kodonů, obsah GC



ORF Finder

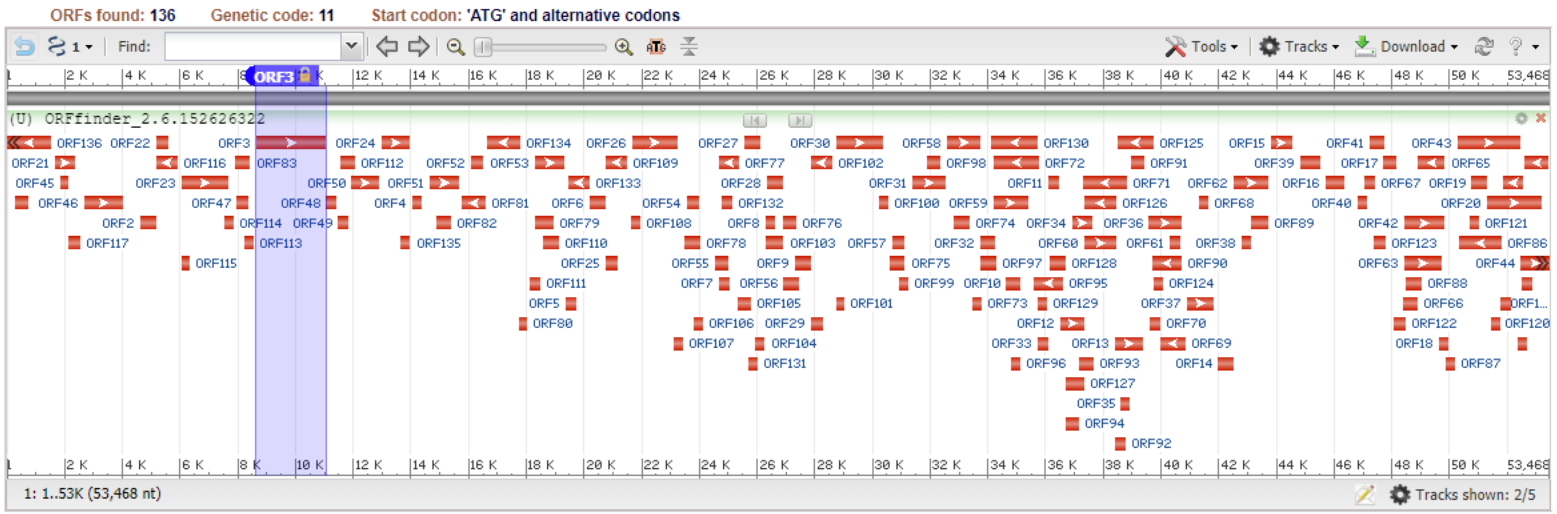
Open Reading Frame Finder

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

Choose Search Parameters

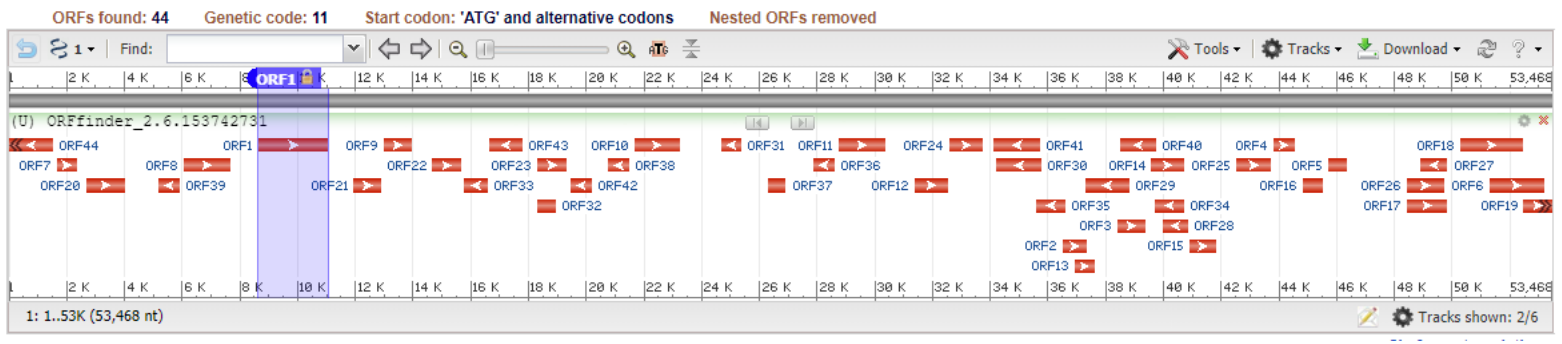
- Minimal ORF length (nt): 300
- Genetic code: 11. Bacterial, Archaeal and Plant Plastid
- ORF start codon to use:
  - "ATG" only
  - "ATG" and alternative initiation codons
  - Any sense codon
- Ignore nested ORFs:

Sequence



Choose Search Parameters

- Minimal ORF length (nt): 600
- Genetic code: 11. Bacterial, Archaeal and Plant Plastid
- ORF start codon to use:
  - "ATG" only
  - "ATG" and alternative initiation codons
  - Any sense codon
- Ignore nested ORFs:



ORF Finder

ORF1 (804 aa) Display ORF as... Mark

```
>lcl|ORF1
MDMIHPGSSLDKAINNTRVKNVSTDVKHGQIQERKRNFIYKKNDDISSRF
KLYSSLVKQKNATEDVVLIGKMILDEVRSYRTIHNDRNIVSNSGNWKTSF
LCNLARLLYSIFNGSNYFCSREGENNSSSSTLLTIHQPEKQELLQQKSIK
HLPTSNNIDGYIKIRKTRGAEDQTTTITQSLIINELLKVDRNTIPFQKIS
ELNDIIHSYENMQIKNSRKGIEILVKQGELLSSLINVNKGNKQLSDNASK
IINLLGIEYQSHKVDIEPFIHAVWVAGAPPDNTFSYITAFLNTYKDYTYL
LWDPNAFGAAKFSGILKNIAMNYAIMRLRRTNHLAEEMNEVILKIQNIQN
ETIEFKETRERLKELENRYKSLTSETKEKFNVFFLESMIGMQDNYFTYCI
SNGISNTDDISRLDFLTNVLKLSPEVQNDFKSTVEKNKRDIDLLKNTISQ
KWDRFQLRDINTLESFKKPQDYFFYQQEMLLRWNYAAASDQVRINILKEY
GGIYTDTDILPAYSDKVSQIINEKSDDKRFFEDLKLRRIISESILSLIKG
EKYSIKHDGLDETTLNQLNNILSEIEKLTIDDYFKPVETKVVRDTFKIFK
RYQKWTENTWNIRGNNNFMLTHKGSDFILSGQKKQYLLQRIRDNISYNNL
FYTTEDLKSLNNVAIGGIPAKKYLEHGLFSEYRQDGTIPYVVSTLNISGP
DMIMRQMKKYYKSLGRIGEVHIKDNKLSDVNFLGVYASSNKDNKSFNWLN
PVSVGINDITPDDESSWAVRNNDINKILFEKINCHVPEKMDELRAQGYHF
KVRT
```

ORF1

SmartBLAST

BLAST

Marked set (0)

SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

Introduction

SmartBLAST processes your protein query to present a concise summary of the five best protein matches from well-studied reference species in the landmark database (described below). If possible, the matches will be from different organisms. If SmartBLAST cannot find five matches in the landmark database, it will uses matches from the protein non-redundant (nr) database. SmartBLAST produces these results using a combination of an optimized BLASTP search, a new implementation of BLAST meant to find closely related matches, and a multiple alignment. Additionally, SmartBLAST presents Conserved Domain Database matches to your query. Additional matches to the nr database are presented lower in the report.

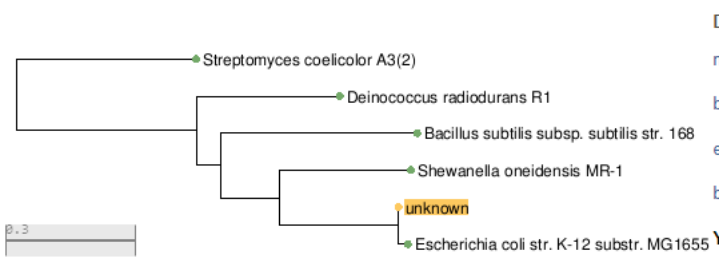
SmartBLAST is under active development and may change with little or no notice.

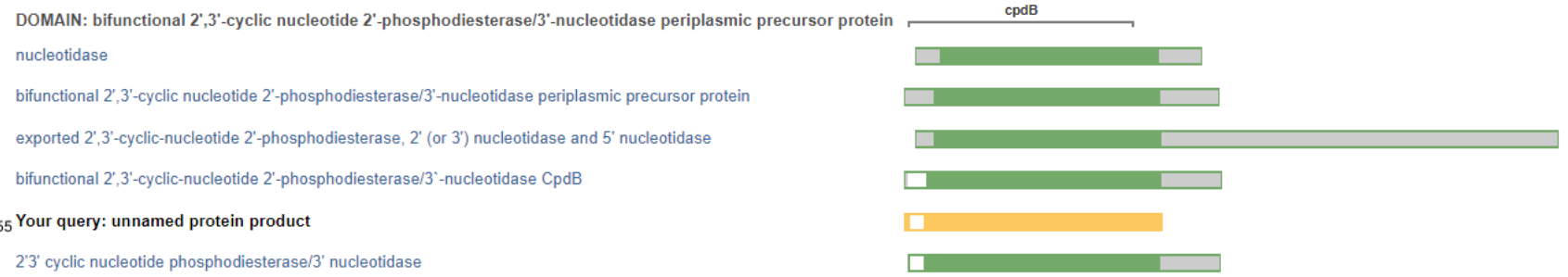
Landmark Database

The landmark database includes proteomes from 27 genomes spanning a wide taxonomic range. This search set is produced using the best available genomic assemblies for each organism with the following procedure. First, the most recent representative assembly from each organism is identified. Second, all proteins annotated on each assembly are downloaded and compiled into the landmark BLAST database. The result is a taxonomically diverse non-redundant set of proteins supported by genomic assemblies.

Query: unnamed protein product

Query length: 531 aa





Predikce genů u prokaryot – základní postupy

(bez využití specializovaných programů)

3) Ověření spolehlivosti predikce – je identifikovaný ORF skutečně součástí genu?

- Kóduje ORF protein **podobný** již popsanému proteinu?
- Vyskytují se před/za ORF typické signální sekvence?
- Statistické parametry sekvence: preference kodonů, obsah GC

Obsah GC – zastoupení G a C v sekvenci NA (genom, gen, část genu, fragment, syntetický oligonukleotid). Vyšší obsah GC párů je asociován s vyšší stabilitou DNA.

Obsah GC

Obsah GC – zastoupení G a C v sekvenci NA (genom, gen, část genu, fragment, syntetický oligonukleotid). Vyšší obsah GC párů je asociován s vyšší stabilitou DNA.

- Velmi rozdílný pro různé prokaryotické genomy (25%-75%).
- Adaptace na vysokou teplotu?
- Adaptace na životní podmínky?

Base composition bias might result from competition for metabolic resources

Eduardo P.C. Rocha and Antoine Danchin

High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes

Laurence D. Hurst[1*] and Alexa R. Merchant[2]

species	GC_{coding} (%)	GC_3 (%)	thermophilic
Aeropyrum pernix[a]	57.50	66.40	yes
Archaeoglobus fulgidus[a]	49.37	58.42	yes
Methanobacterium thermoautotrophicum[a]	50.46	56.59	yes
Methanococcus jannaschii[a]	31.84	24.74	yes
Pyrococcus abyssi[a]	45.16	50.31	yes
Pyrococcus horikoshii[a]	42.32	42.97	yes
Aquifex aeolicus[b]	43.58	47.93	yes
Bacillus subtilis[b]	44.32	44.61	no
Borrelia burgdorferi[b]	29.31	20.82	no
Campylobacter jejuni[b]	32.82	18.96	no
Chlamydia muridarum[b]	39.13	29.92	no
Chlamydia pneumoniae[b]	41.30	34.88	no
Chlamydia trachomatis[b]	41.61	34.30	no
Deinococcus radiodurans[b]	65.72	84.02	no
Escherichia coli[b]	51.37	54.90	no
Haemophilus influenzae[b]	38.76	29.09	no
Helicobacter pylori[b]	39.56	41.95	no
Mycobacterium tuberculosis[b]	65.81	79.67	no
Mycoplasma genitalium[b]	31.64	23.01	no
Mycoplasma pneumoniae[b]	41.05	42.08	no
Neisseria meningitidis[b]	50.14	55.49	no
Rickettsia prowazekii[b]	30.59	18.47	no
Synechocystis sp.[b]	48.66	49.99	no
Thermotoga maritima[b]	46.45	52.62	yes
Treponema pallidum[b]	52.52	54.10	no
Ureaplasma urealyticum[b]	35.20	16.97	no
Vibrio cholerae[b]	47.17	49.08	no

Obsah GC

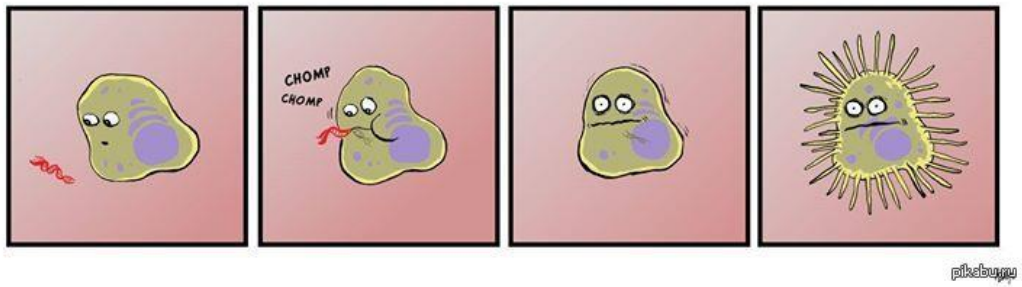
Obsah GC – zastoupení G a C v sekvenci NA (genom, gen, část genu, fragment, syntetický oligonukleotid). Vyšší obsah GC párů je asociován s vyšší stabilitou DNA.

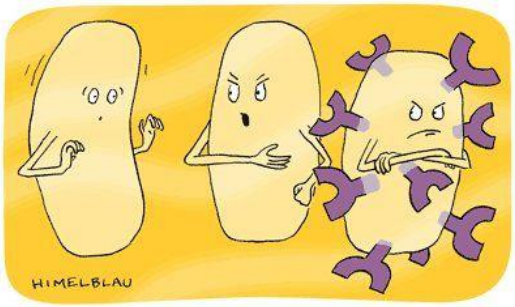
- Velmi rozdílný pro různé prokaryotické genomy (25%-75%).
- Variabilita je nejvyšší v třetí kodonové pozici (GC_3).
- Adaptace na vysokou teplotu?
- Adaptace na životní podmínky?
- Využití: identifikace genů získaných horizontálním přenosem.

Horizontal gene transfer: building the web of life

Shannon M. Soucy[1], Jinling Huang[2] and Johann Peter Gogarten[1,3]

Abstract | Horizontal gene transfer (HGT) is the sharing of genetic material between organisms that are not in a parent–offspring relationship. HGT is a widely recognized mechanism for adaptation in bacteria and archaea. Microbial antibiotic resistance and pathogenicity are often associated with HGT, but the scope of HGT extends far beyond disease-causing organisms. In this Review, we describe how HGT has shaped the web of life using examples of HGT among prokaryotes, between prokaryotes and eukaryotes, and even between multicellular eukaryotes. We discuss replacement and additive HGT, the proposed mechanisms of HGT, selective forces that influence HGT, and the evolutionary impact of HGT on ancestral populations and existing populations such as the human microbiome.





"Don't pick it up," I say, and he says, "It's just a *plasmid*, what harm could it do?" Well just look at him now....God knows *what* protein he's expressing!

Metody predikce genů

- Dva hlavní přístupy: metody ***ab initio***/metody založené na **homologii** (sekvenční).
- ***Ab initio*** – predikce genů založená pouze na **sekvenci**, jejích vlastnostech a statistických parametrech.
- **Metody založené na homologii** – sekvenční podobnost se známými geny/proteiny. ORF kóduje protein, který je podobný již dříve popsanému proteinu (prohledávání **DATABÁZÍ** pomocí **ALIGNMENTU**) = nejspolehlivější predikce. Problém – unikátní geny bez známých homologů (většinou nejzajímavější).
- Kombinace obou postupů
- **Specializované predikční programy**

Predikce genů u prokaryot

Markovovy modely

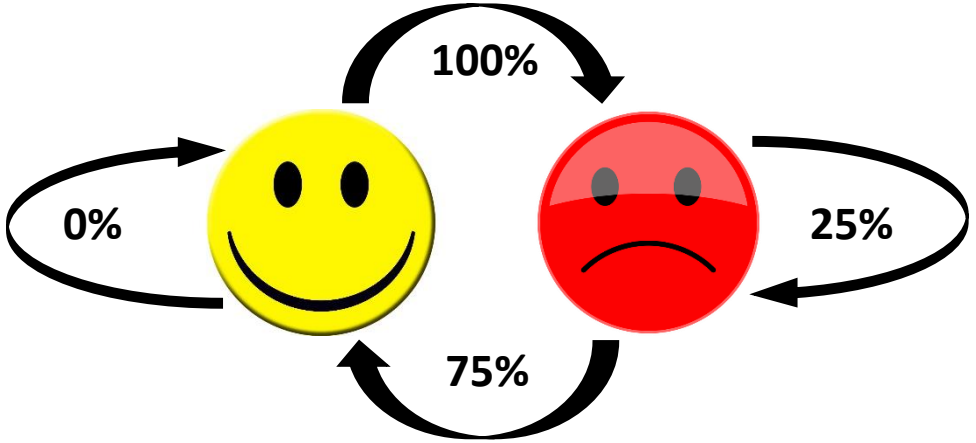
- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak:
  Současný stav závisí pouze na minulém stavu.
  Budoucí stav závisí pouze na současném stavu.

Spokojenost studenta:
Jak se bude cítit zítra?

Spokojenost studenta:
Jak se bude cítit zítra?

Student se vyskytuje ve dvou stavech.
Několik dní ho pozorujeme:

Vytvoříme Markovův model popisující studenta



Predikce genů u prokaryot

Markovovy modely

- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak:
  Současný stav závisí pouze na minulém stavu.
  Budoucí stav závisí pouze na současném stavu.

Spokojenost studenta:
Jak se bude cítit zítra?

Spokojenost studenta:
Jak se bude cítit zítra?

Student se vyskytuje ve dvou stavech.
Několik dní ho pozorujeme:

Vytvoříme Markovův model popisující studenta

Predikce genů u prokaryot

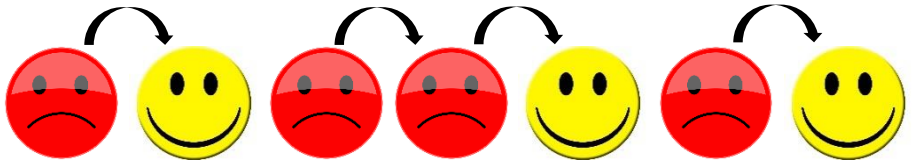
Markovovy modely

- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak:
  Současný stav závisí pouze na minulém stavu.
  Budoucí stav závisí pouze na současném stavu.

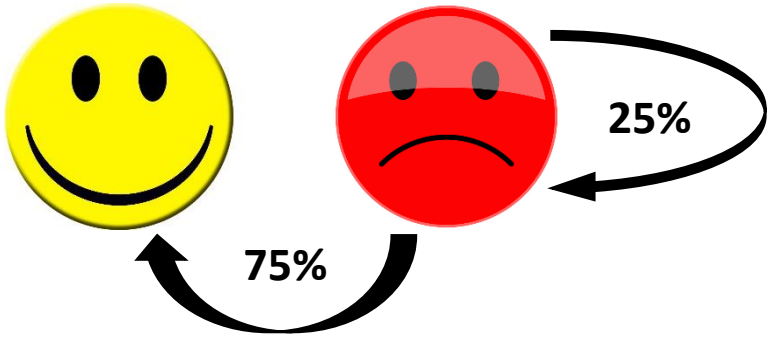
Spokojenost studenta:
Jak se bude cítit zítra?

Spokojenost studenta:
Jak se bude cítit zítra?

Student se vyskytuje ve dvou stavech.
Několik dní ho pozorujeme:



Vytvoříme Markovův model popisující studenta



Predikce genů u prokaryot

Markovovy modely

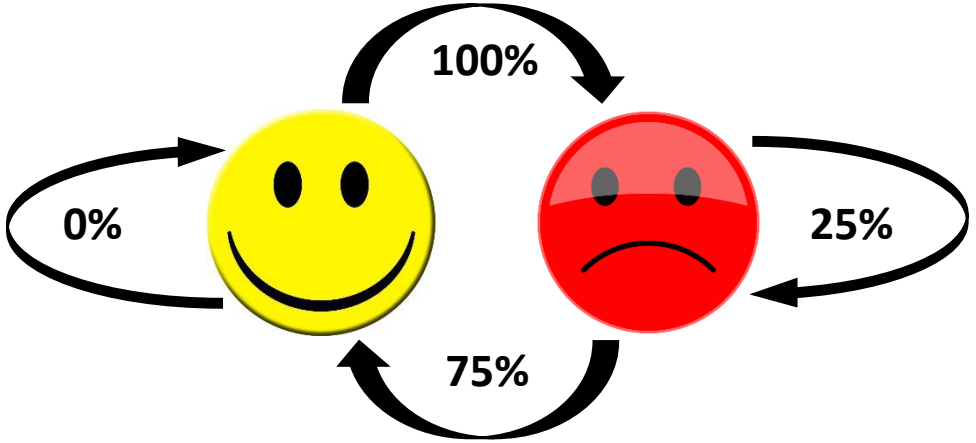
- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak:
  Současný stav závisí pouze na minulém stavu.
  Budoucí stav závisí pouze na současném stavu.

Spokojenost studenta:
Jak se bude cítit zítra?

Spokojenost studenta:
Jak se bude cítit zítra?

Student se vyskytuje ve dvou stavech.
Několik dní ho pozorujeme:

Vytvoříme Markovův model popisující studenta



Predikce genů u prokaryot

Markovovy modely

- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak: Současný stav závisí na minulém stavu.

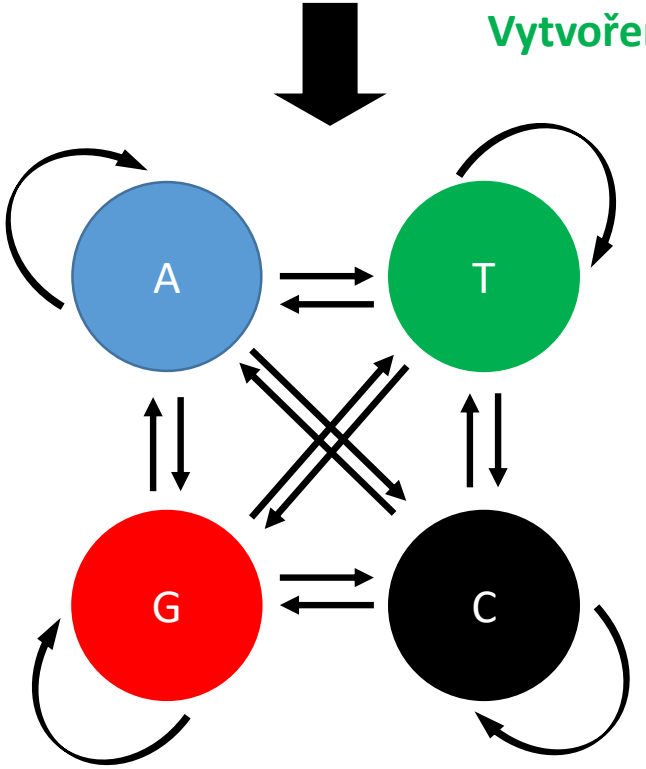
- Kódující sekvence má jiné rozložení nukleotidů než nekódující. Kódující sekvence není náhodná. Výskyt dané báze v sekvenci závisí na předcházející bázi.

- Lépe: Výskyt báze v sekvenci závisí na *k* předcházejících bázích (Markovovy modely vyšších řádů). Kódující sekvence je složená z kodonů (triplety). Lépe ji tedy popisuje Markovův model druhého řádu (výskyt báze v sekvenci závisí na dvou předcházejících bázích), neboli distribuce **tripletů** (kodonů) v kódující sekvenci **není náhodná**. Statistika dále praví, že dvojice kodonů mají tendenci korelovat, ještě lépe tedy fungují Markovovy modely pátého řádu, neboli distribuce **hexamerů** v kódující sekvenci je **mnohem méně náhodná**. Kódující oblast odhalíme přesněji.

Sekvence známých genů

```
atg ctg gtg att gtg gat gcc gtt acc ctg
ctg agc gcc tat ccg gaa gcc agc cgt gat
```

Vytvoření modelu



Predikce genů u prokaryot

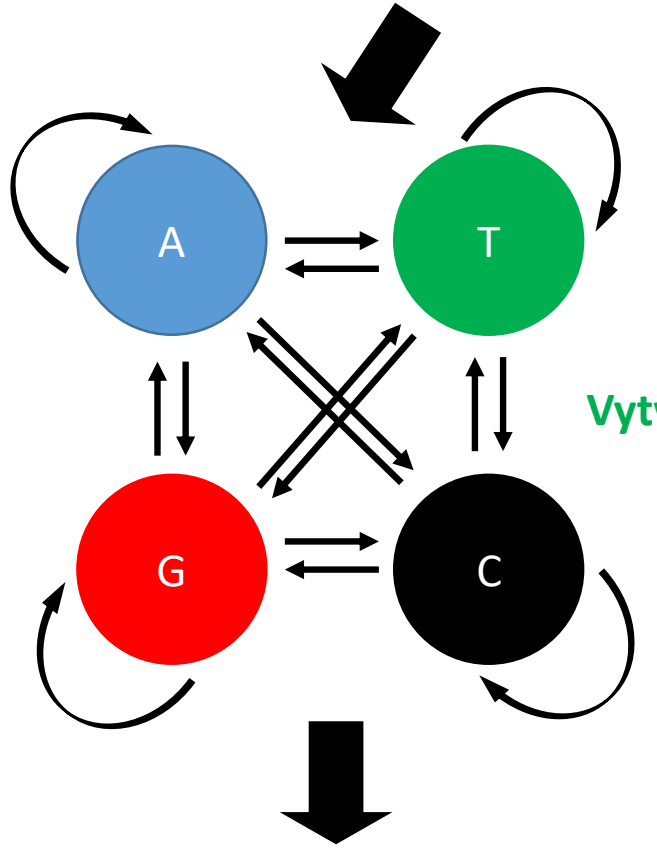
Markovovy modely

- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak: Současný stav závisí na minulém stavu.

- Kódující sekvence má jiné rozložení nukleotidů než nekódující. Kódující sekvence není náhodná. Výskyt dané báze v sekvenci závisí na předcházející bázi.

- Lépe: Výskyt báze v sekvenci závisí na *k* předcházejících bázích (Markovovy modely vyšších řádů). Kódující sekvence je složená z kodonů (triplety). Lépe ji tedy popisuje Markovův model druhého řádu (výskyt báze v sekvenci závisí na dvou předcházejících bázích), neboli distribuce **tripletů** (kodonů) v kódující sekvenci **není náhodná**. Statistika dále praví, že dvojice kodonů mají tendenci korelovat, ještě lépe tedy fungují Markovovy modely pátého řádu, neboli distribuce **hexamerů** v kódující sekvenci je **mnohem méně náhodná**. Kódující oblast odhalíme přesněji.

Sekvence známých genů



Vytvoření modelu

gttccgg**atgcggatgatctgctgcatccgagctgtcgtccgg**
aaagatcattattggcgcagcgatgtgctggcggcgggcgcga
ccacctgtaccgccgattttgcggtgtgcgatcgtgatggcac
cgtgagcggttattttcgttgggaaaccagcattgaaattgcg
ggcagccagccggataccaaacagccgggctttaaaccgagca
gcgatcgcaatggcaactttagcctgccgccgaataccgcctt
taagcgatagctctatgcgaacgcgttgcggatcgtcagatct
gaaactgtttatt

Které části zkoumané sekvence odpovídají modelu – jsou kódující?

Predikce genů u prokaryot

Markovovy modely

GLIMMER
Microbial Gene-Finding System

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

CCB » Software » Glimmer

ABOUT GLIMMER

Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA. The IMM approach is described in our original Nucleic Acids Research paper on Glimmer 1.0 and in our subsequent paper on Glimmer 2.0. The IMM is a combination of Markov models from 1st through 8th-order, where the order used is determined by the amount of data available to train the model. In addition, the positions used as context for the model need not immediately precede the predicted postion but are determined by a decision procedure based on the predictive power of each position in the training data set (which we term an Interpolated Context Model or ICM). The models for coding sequence are 3-periodic nonhomogenous Markov models. Improvements made in version 3 of Glimmer are described in the third Glimmer paper.

Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and has been used to annotate the genomes of thousands of bacterial, archaeal, and viral genomes around the world.

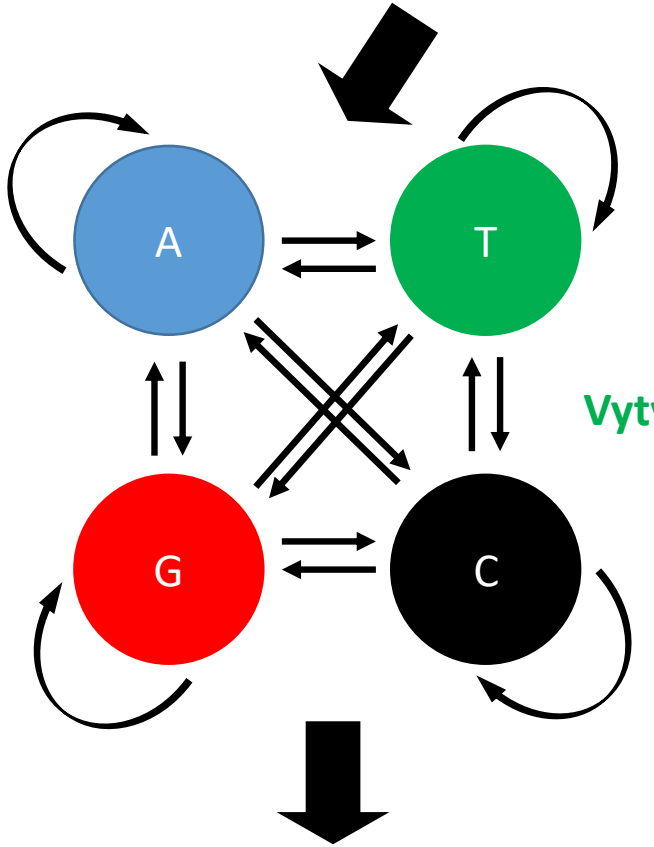
http://ccb.jhu.edu/software/glimmer/index.shtml

Microbial gene identification using interpolated Markov models

Steven L. Salzberg[1,2,*], Arthur L. Delcher[3], Simon Kasif[4] and Owen White[1]

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, [2]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA, [3]Department of Computer Science, Loyola College in Maryland, Baltimore, MD 21210, USA and [4]Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Sekvence známých genů



Vytvoření modelu

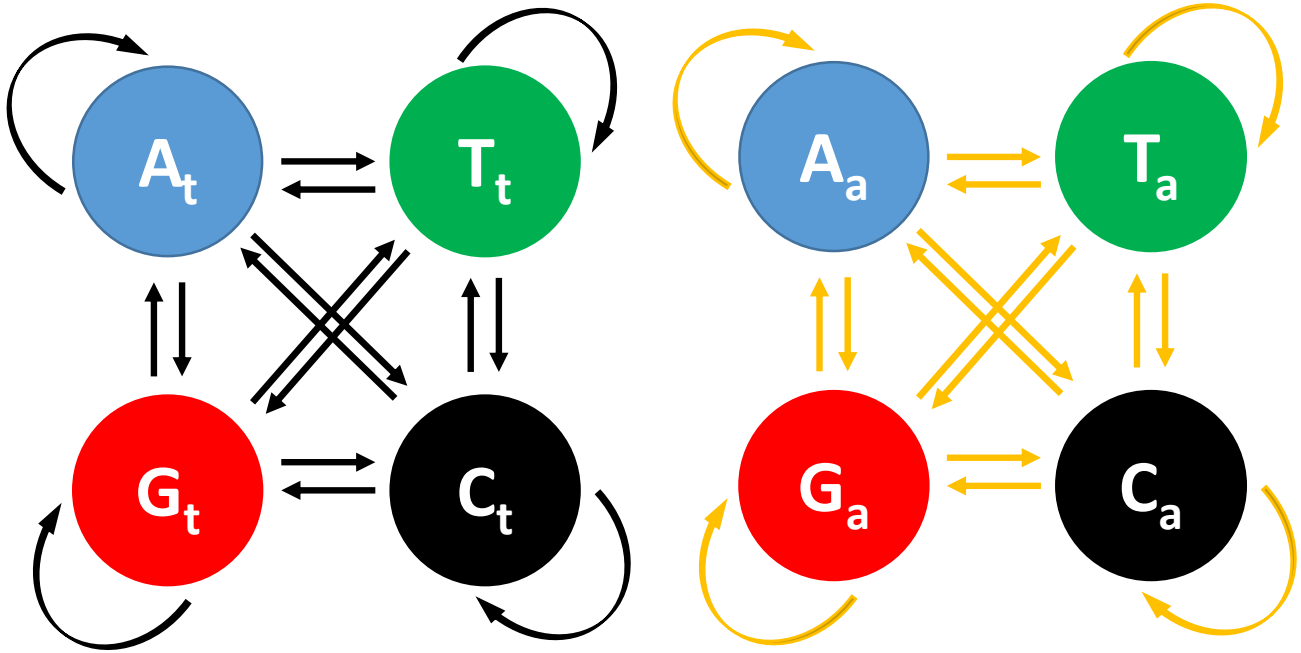
gttccgg**atgcggatgatctgctgcatccgagctgtcgtccgg**
aaagatcattattggcgcagcgatgtgctggcggcgggcgcga
ccacctgtaccgccgattttgcggtgtgcgatcgtgatggcac
cgtgagcggttattttcgttgggaaaccagcattgaaattgcg
ggcagccagccggataccaaacagccgggctttaaaccgagca
gcgatcgcaatggcaactttagcctgccgccgaataccgcctt
taagcgatagctctatgcgaacgcgttgcggatcgtcagatct
gaaactgtttatt

Které části zkoumané sekvence odpovídají modelu – jsou kódující?

Predikce genů u prokaryot

Markovovy modely

- Markovův proces: Proces bez paměti, tj. stav systému v budoucnu závisí pouze na současném stavu. Současný stav udává pravděpodobnost, s jakou systém přejde do jiných stavů, přičemž nezáleží na tom, jak se systém do současného stavu dostal. Nebo jinak: Současný stav závisí na minulém stavu.
- Prokaryotické geny: typické, atypické
- **Typické**: 100 až 500 aminokyselin, zastoupení nukleotidů typické pro daný organismus. **Atypické**: kratší nebo delší, odlišné zastoupení nukleotidů (možný horizontální přenos genů).

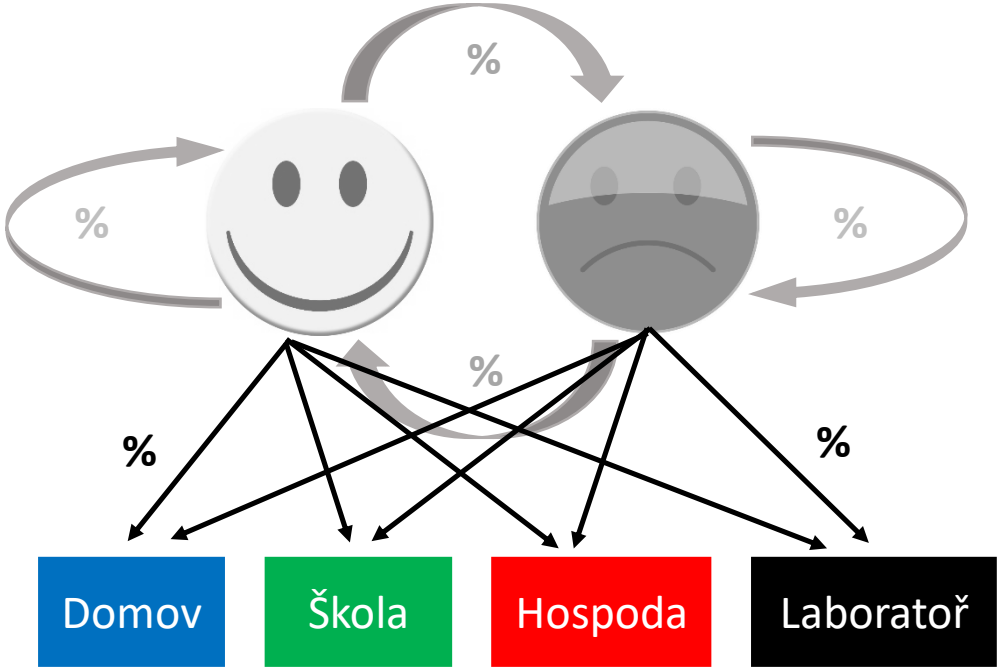


Pro přesný popis všech genů v genomu jsou nutné dva Markovovy modely. Možným řešením je také využití **skrytých Markovových modelů.**

Predikce genů u prokaryot

Skryté Markovovy modely

- Skrytý Markovův model: Jednotlivé stavy mohou generovat různé znaky s definovanou pravděpodobností. Stavy jsou **skryté**, vidíme pouze **znaky**, které generují.



DDHLLŠLDDHHHHDHDHDHŠLLLDL

Jaký je nejpravděpodobnější průchod skrytými stavy?

DDHLLŠLDDHHHHDHDHDHŠLLLDL

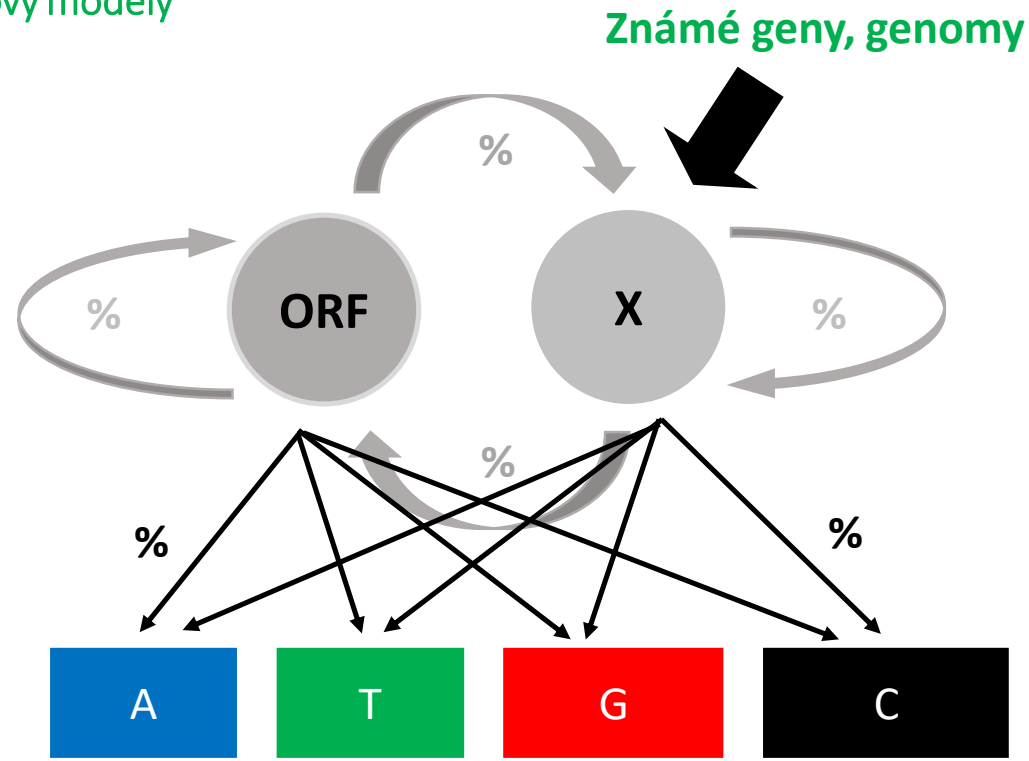
D: 0,1
Š: 0,3
H: 0,1
L: 0,5

D: 0,2
Š: 0,05
H: 0,7
L: 0,05

Predikce genů u prokaryot

Skryté Markovovy modely

- Skrytý Markovův model: Jednotlivé stavy mohou generovat různé znaky s definovanou pravděpodobností. Stavy jsou **skryté**, vidíme pouze **znaky**, které generují.




```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

Jaký je nejpravděpodobnější průchod skrytými stavy?

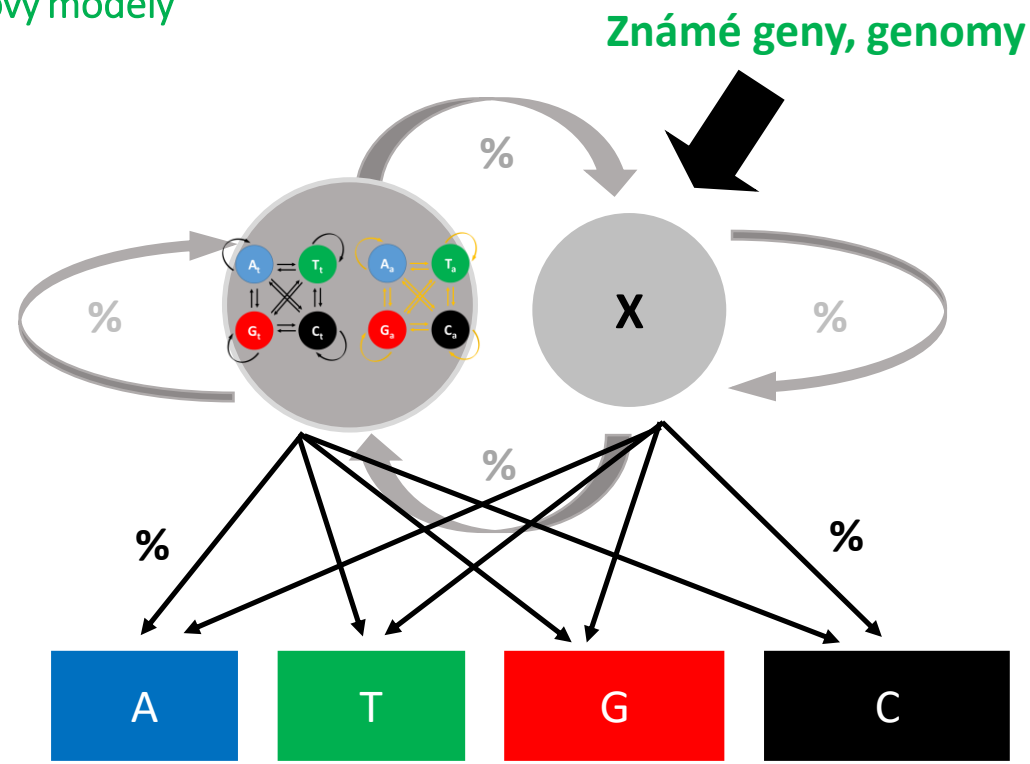
```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

```
MRMICCIRAVVRKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSI
EIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKR
```

Predikce genů u prokaryot

Skryté Markovovy modely

- Skrytý Markovův model: Jednotlivé stavy mohou generovat různé znaky s definovanou pravděpodobností. Stavy jsou **skryté**, vidíme pouze **znaky**, které generují.




```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

Jaký je nejpravděpodobnější průchod skrytými stavy?

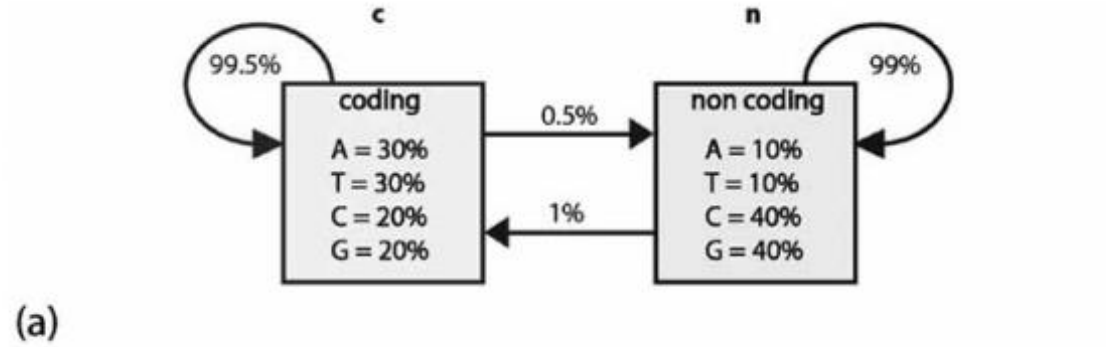
```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

```
MRMICCIRAVVRKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSI
EIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKR
Protein třídy 1, typický
```

Predikce genů u prokaryot

Skryté Markovovy modely

- Skrytý Markovův model: Jednotlivé stavy mohou generovat různé znaky s definovanou pravděpodobností. Stavy jsou **skryté**, vidíme pouze **znaky**, které generují.

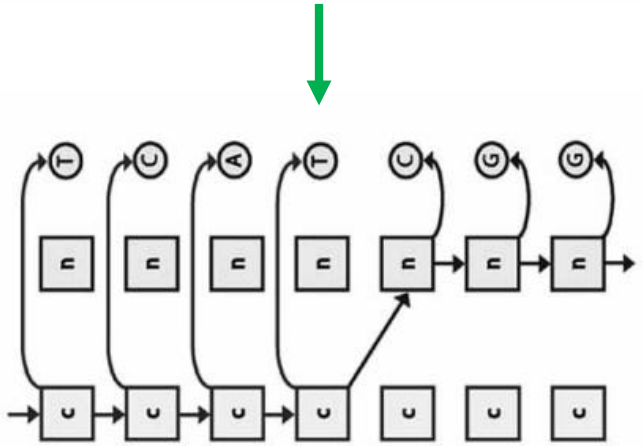



```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

Jaký je nejpravděpodobnější průchod skrytými stavy?

```
gttccggatgcggatgatctgctgcatccgagctgtcgtccggaaagatcattattggcgcagc
gatgtgctggcggcgggcgcgaccacctgtaccgccgattttgcggtgtgcgatcgtgatggca
ccgtgagcggttattttcgttgggaaaccagcattgaaattgcgggcagccagccggataccaa
acagccgggctttaaaccgagcagcgatcgcaatggcaactttagcctgccgccgaataccgcc
tttaagcgatagctctatgcgaacgcgttgcggatcgtcagatctgaaactgtttatt
```

```
MRMICCIRAVVRKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSI
EIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKR
Protein třídy 1, typický
```



Current Bioinformatics, 2007, 2, 49-61

Hidden Markov Models in Bioinformatics

Predikce genů u prokaryot

Skryté Markovovy modely

GeneMark

A family of gene prediction programs developed at
Georgia Institute of Technology , Atlanta, Georgia, USA.

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

GeneMark

GeneMark developed in 1993 was the first gene finding method recognized as an efficient and accurate tool for genome projects. GeneMark was used for annotation of the first completely sequenced bacteria, *Haemophilus influenzae*, and the first completely sequenced archaea, *Methanococcus jannaschii*. The GeneMark algorithm uses species specific inhomogeneous Markov chain models of protein-coding DNA sequence as well as homogeneous Markov chain models of non- coding DNA. Parameters of the models are estimated from training sets of sequences of known type. The major step of the algorithm computes a posteriory probability of a sequence fragment to carry on a genetic code in one of six possible frames (including three frames in complementary DNA strand) or to be "non-coding".

GeneMark.hmm (prokaryotic)

GeneMark.hmm algorithm was designed to improve the gene prediction quality, particularly to improve GeneMark in finding exact gene starts. The idea was to integrate the GeneMark models into naturally designed hidden Markov model framework with gene boundaries modeled as transitions between hidden states. Additionally, the ribosome binding site model is used to make the gene start predictions more accurate. In evaluations by different groups it was shown that GeneMark.hmm is significantly more accurate than GeneMark in exact gene prediction. From 1998 until now GeneMark.hmm and its self-training version, GeneMarkS, are the standard tools for gene identification in new prokaryotic genomic sequences, including metagenomes.

© 1998 Oxford University Press Nucleic Acids Research, 1998, Vol. 26, No. 4 1107–1115

GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashin and Mark Borodovsky[1,*]

School of Biology and [1]Schools of Biology and Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received August 14, 1997; Revised and Accepted December 30, 1997

Table 5. Results of GeneMark.hmm predictions for 10 complete bacterial genomes

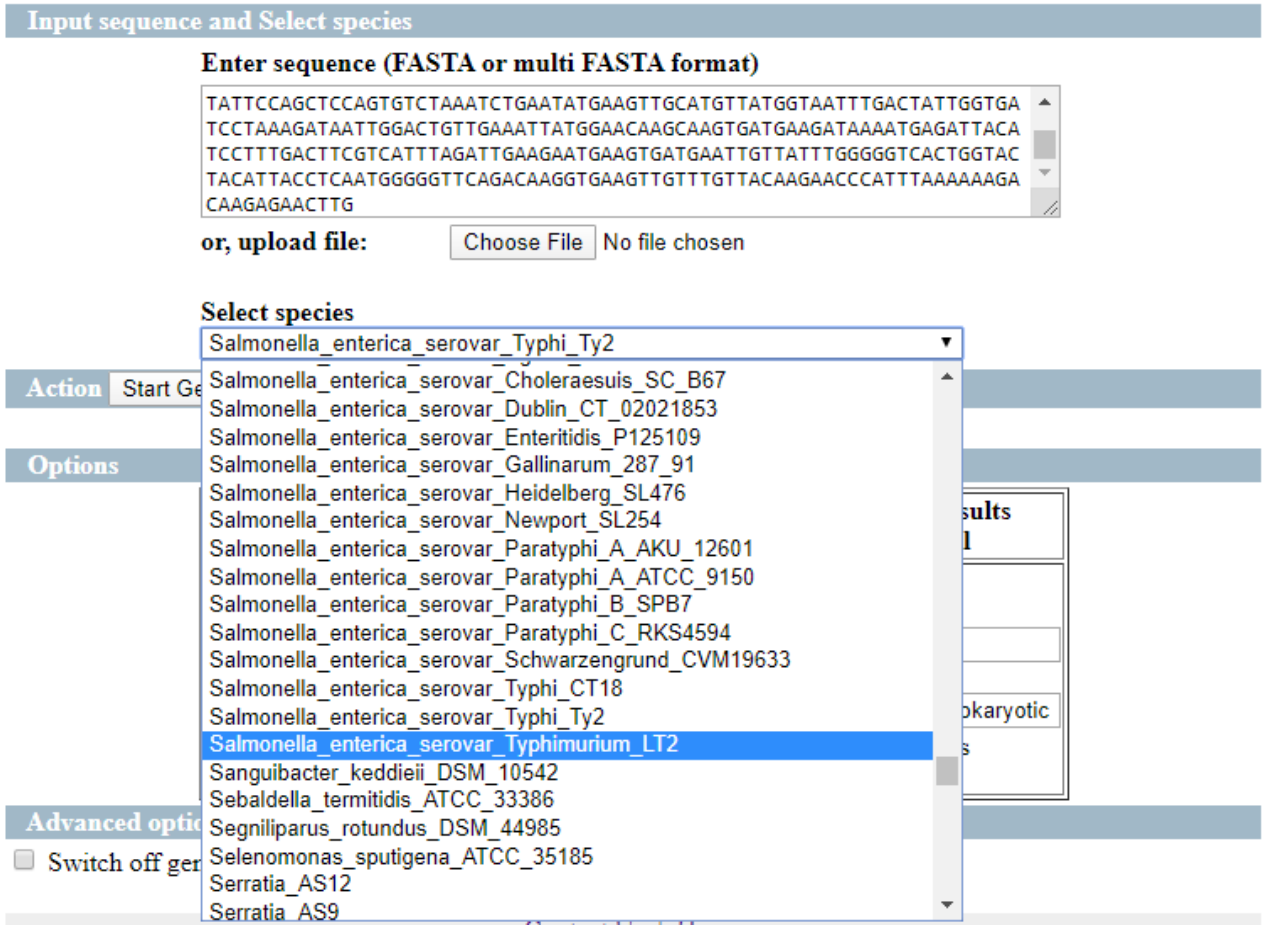
Genome	Genes annotated	Genes predicted	Exact prediction (%)	Missing genes (%)	Wrong genes (%)
A.fulgidus	2407	2530	73.1	10.8 (2.0)	15.1
B.subtilis	4101	4384	77.5	3.6 (2.8)	9.8
E.coli	4288	4440	75.4	5.0 (2.7)	8.2
H.influenzae	1718	1840	86.7	3.8 (3.2)	10.2
H.pylori	1566	1612	79.7	6.0 (4.4)	8.7
M.genitalium	467	509	78.4	9.9 (1.7)	17.3
M.jannaschii	1680	1841	72.7	4.6 (0.8)	12.9
M.pneumoniae	678	734	70.1	7.8 (4.1)	13.6
M.thermoauthotrophicum	1869	1944	70.9	5.0 (3.5)	8.6
Synechocystis	3169	3360	89.6	4.0 (1.5)	9.4
Averaged	21 943	23 194	78.1	5.4 (2.7)	10.4

The second and third columns show the number of genes annotated in GenBank and the corresponding number of genes predicted, respectively. 'Exact prediction' is a fraction of annotated genes for which both the 5′-end and the 3′-end were predicted exactly. 'Missing genes' is a fraction of annotated genes for which neither the 5′-end nor the 3′-end was predicted exactly; in this column the numbers in brackets show the missing genes after using the combined program (GeneMark.hmm + GeneMark). 'Wrong genes' is a fraction of predicted genes for which no annotated analog was found. All measures are expressed as percentages. The data shown are the results obtained after post-processing procedure (RBS recognition).

Predikce genů u prokaryot

Skryté Markovovy modely

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Feb 12 08:45:26 2020
Sequence file name: seq.fna
Model file name:
/home/genemark/parameters/prokaryotic/Salmonella_enterica_serovar_T
yphi_Ty2/GeneMark_hmm_combined.mod
RBS: true
Model information: Salmonella_enterica_serovar_Typhi_Ty2

FASTA definition line: empty-fasta-def-line
Predicted genes
   Gene     Strand     LeftEnd     RightEnd       Gene      Class
    #                                           Length
    1         -          598         1155          558        1
    2         +         3001         4008         1008        2
    3         +         4316         5092          777        2
    4         +         5478         5606          129        2
    5         +         6507         6689          183        2


    >gene_1|GeneMark.hmm|185_aa|-|598|1155  >empty-fasta-def-line
    MMAQVGDKLEYIDLLLIHSPLPGKTKRLESWKVLQDAVEKGWIKNIGVSNYGKHHIEELL
    TNATIPPAVNQIEISPWCMRQDLATWCLSKGINVEAYAPLTHGNKLQVNNTEFQEIMQKY
    NKSAAQILIKWSLQKGYIPLPKTKTPSRLKENLSVDDFELTNEEIKAIDQPDAYEPTDWE
    CTDAP

    >gene_2|GeneMark.hmm|335_aa|+|3001|4008 >empty-fasta-def-line
    MAIKIGINGFGRIGRLVLRVALGRKDIEVVAVNDPFIAPDYAAYMFKYDSTHGRYKGEVT
    ASGDDLVIDGHKIKVFQERDPANIPWGKSGVDYVIESTGVFTKLEGAQKHIDAGAKKVII
    TAPSADAPMFVVGVNEDKYTPDLKIISNASCTTNCLAPLAKVVNDTFGIEEGLMTTVHSI
    TATQKTVDGPSHKDWRGGRTASGNIIPSSTGAAKAVGKVIPELNGKLTGMSLRVPTTDVS
    VVDLTVRLKKAASYEEIAQAIKKASEGPLKGVLGYTEDAVVSTDFLGSSYSSIFDEKAGI
    LLSPTFVKLISWYDNEYGYSTRVVDLLEHVAKASA

    >gene_3|GeneMark.hmm|258_aa|+|4316|5092 >empty-fasta-def-line
    LIINQTFKIYKSNSSLSSSIKLTKKQLEKFRITLCKYLYQRYQQTLQIEDDLLVQYELYR
    DIVPELTIYNNSIEEHSQYKIQYLIIKCANNRLISRQVFIFVSNDDDFTLVLNRITNHPS
    INEYILELLEDINDYPLIIKPLDLPDYLIPNLVDQLGNKLDSLGDLQLVYSSSTSNRLRN
    FIIDIPKNDLPKLSSDGKLYQDIVKFMYNNTKIKFEKLRIEKFINNLINIASDGKFKLII
    NDNQLIWFLIESIKSSLQ

    >gene_4|GeneMark.hmm|42_aa|+|5478|5606  >empty-fasta-def-line
    MIFPQLPQLKKNILTNKERLQKTMTTLLKEPNMMKLKKPLNL

    >gene_5|GeneMark.hmm|60_aa|+|6507|6689  >empty-fasta-def-line
    VAFYTPTFKHSQKVQNNNKLLLMVTKIQTTIGFSKELEDNLIMILLVTPLTLNIFLTVCM
```



Predikce genů u prokaryot

Markovovy modely

Co když není model pro můj organismus v seznamu GeneMark?

- Lze použít model pro **blízce příbuzný** organismus.
- Lze využít **heuristický model** (pro krátké sekvence).
- Lze využít **„self-training“** algoritmus (pro dostatečně dlouhé sekvence).

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Mar 25 10:09:08 2020
Sequence file name: seq.fna
Model file name: /home/genemark/parameters/prokaryotic/Escherichia_coli__BL21_Gold_DE3_pLysS_AG_/GeneMark_hmm_combined.mod
RBS: true
Model information: Escherichia_coli__BL21_Gold_DE3_pLysS_AG_

FASTA definition line: empty-fasta-def-line
Predicted genes
   Gene    Strand    LeftEnd    RightEnd      Gene     Class
    #                                       Length
    1        +          <3         314         312         1
    2        +         318        1604        1287         1
    3        -        1698        2471         774         1
    4        -        2550        3980        1431         1
    5        +        4249        5202         954         1
    6        +        5313        5903         591         1
    7        -        5960       >6244         285         1
```

```
GeneMark.hmm PROKARYOTIC (Version 3.26)
Date: Wed Mar 25 10:13:05 2020
Sequence file name: seq.fna
Model file name: /home/genemark/parameters/prokaryotic/Pseudomonas_aeruginosa_PAO1/GeneMark_hmm_combined.mod
RBS: true
Model information: Pseudomonas_aeruginosa_PAO1

FASTA definition line: empty-fasta-def-line
Predicted genes
   Gene    Strand    LeftEnd    RightEnd      Gene     Class
    #                                       Length
    1        +          <3         314         312         2
    2        +         318        1604        1287         2
    3        -        1698        2471         774         2
    4        -        2550        3980        1431         2
    5        +        4249        5202         954         2
    6        +        5313        5903         591         2
    7        -        5960       >6244         285         2
```

Heuristic Models

Computer methods of accurate gene finding in DNA sequences require models of protein coding and non-coding regions derived either from experimentally validated training sets or from large amounts of anonymous DNA sequence. A heuristic method for derivation of parameters of inhomogeneous Markov models of protein coding regions, was proposed in 1999. The heuristic method utilizes the observation that parameters of the Markov models used in GeneMark can be approximated by the functions of the sequence G+C content. Therefore, a short DNA sequence sufficient for estimation of the genome G+C content (a fragment longer than 400 nt) is also sufficient for derivation of parameters of the Markov models used in GeneMark and GeneMark.hmm. Models built by the heuristic approach could be used to find genes in small fragments of anonymous prokaryotic genomes, such as metagenomic sequences, as well as in genomes of organelles, viruses, phages and plasmids. This method can also be used for highly inhomogeneous genomes where adjustment of the Markov models to local DNA composition is needed. The heuristic method provides an evidence that the mutational pressure that shapes G+C content is the driving force of the evolution of codon usage pattern.

Heuristické řešení *– přibližné řešení založené na zkušenosti, poučeném odhadu nebo empirických poznatcích. Dá nám rozumné výsledky rozumně rychle.*

Metagenomy

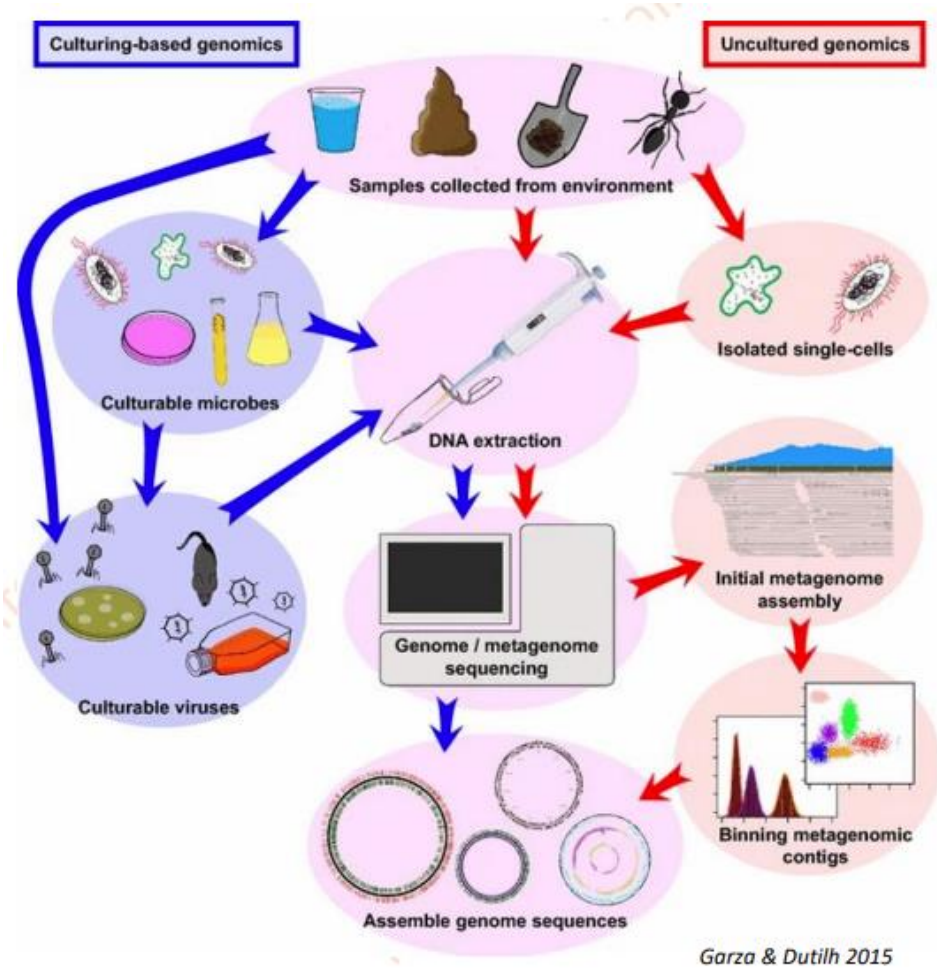
- Mnoho organismů nelze získat v **izolovaném stavu**.
- Půdní společenstva, mořská společenstva, střevní mikrobiom.
- Metagenomika se zabývá sekvenacemi **komplexních vzorků.**

*„A metagenomic sample is a heterogeneous mixture of **rather short sequences** originated from a shotgun sequencing of a microbial community. A vast majority (99%) of microbial species in a given community are likely to be non-cultivable.“*

MetaGeneMark

Predikční program specializovaný na metagenomy

http://exon.gatech.edu/GeneMark/meta_gmhmmp.cgi



Predikce genů u prokaryot

Markovovy modely

Co když není model pro můj organismus v seznamu GeneMark?

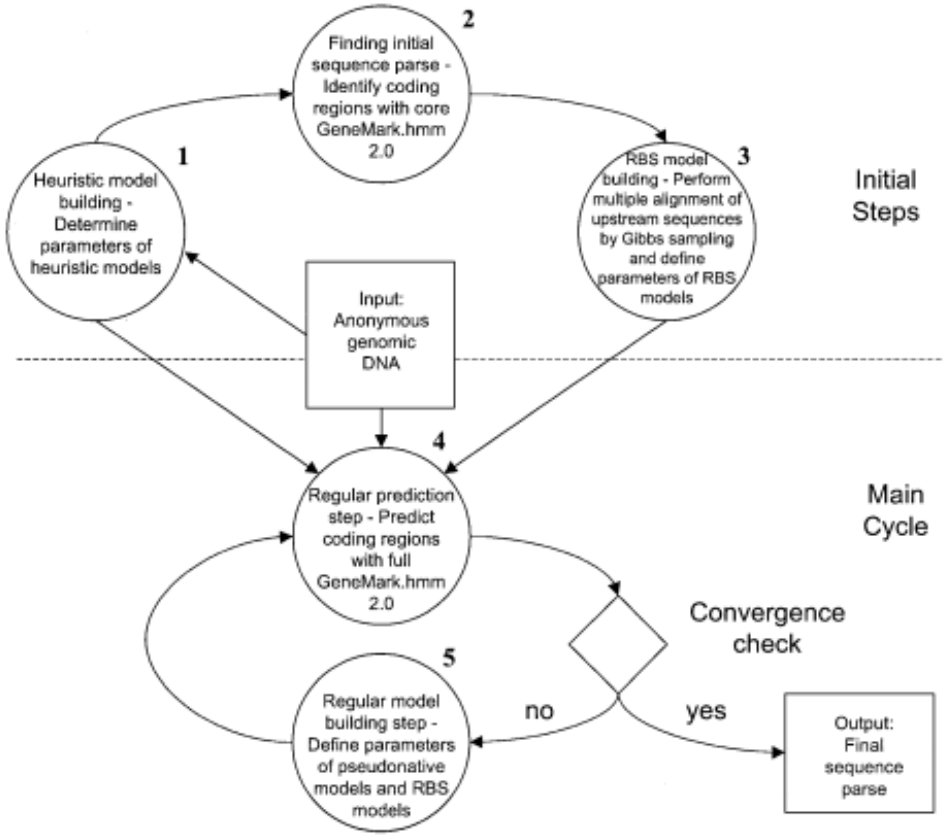
- Lze použít model pro **blízce příbuzný** organismus.
- Lze využít **heuristický model** (pro krátké sekvence).
- Lze využít **„self-training“** algoritmus (pro dostatečně dlouhé sekvence).

Heuristic Models

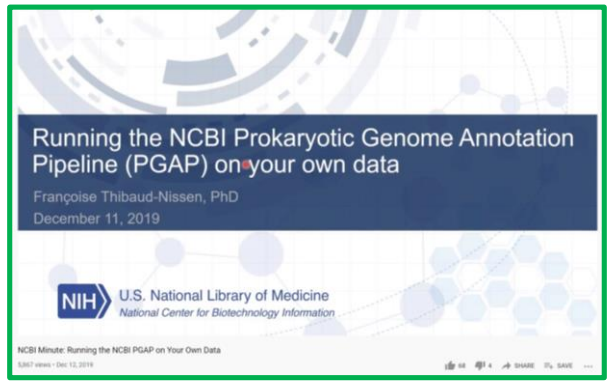
Computer methods of accurate gene finding in DNA sequences require models of protein coding and non-coding regions derived either from experimentally validated training sets or from large amounts of anonymous DNA sequence. A heuristic method for derivation of parameters of inhomogeneous Markov models of protein coding regions, was proposed in 1999. The heuristic method utilizes the observation that parameters of the Markov models used in GeneMark can be approximated by the functions of the sequence G+C content. Therefore, a short DNA sequence sufficient for estimation of the genome G+C content (a fragment longer than 400 nt) is also sufficient for derivation of parameters of the Markov models used in GeneMark and GeneMark.hmm. Models built by the heuristic approach could be used to find genes in small fragments of anonymous prokaryotic genomes, such as metagenomic sequences, as well as in genomes of organelles, viruses, phages and plasmids. This method can also be used for highly inhomogeneous genomes where adjustment of the Markov models to local DNA composition is needed. The heuristic method provides an evidence that the mutational pressure that shapes G+C content is the driving force of the evolution of codon usage pattern.

GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions

John Besemer[1], Alexandre Lomsadze[1,3] and Mark Borodovsky[1,2,*]



* Vždycky přemýšlím, jestli je GeneMark vlastně pojmenován podle Markovových modelů nebo podle šéfa skupiny…



Anotace genomů u prokaryot

PGAP is now available as a stand-alone software package. You can annotate your genomes on your own machine, local cluster or the Cloud! Get started by watching a short video!

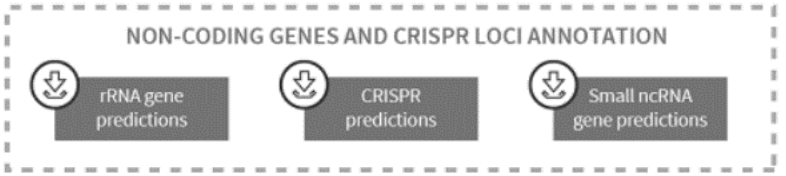
NCBI Prokaryotic Genome Annotation Pipeline

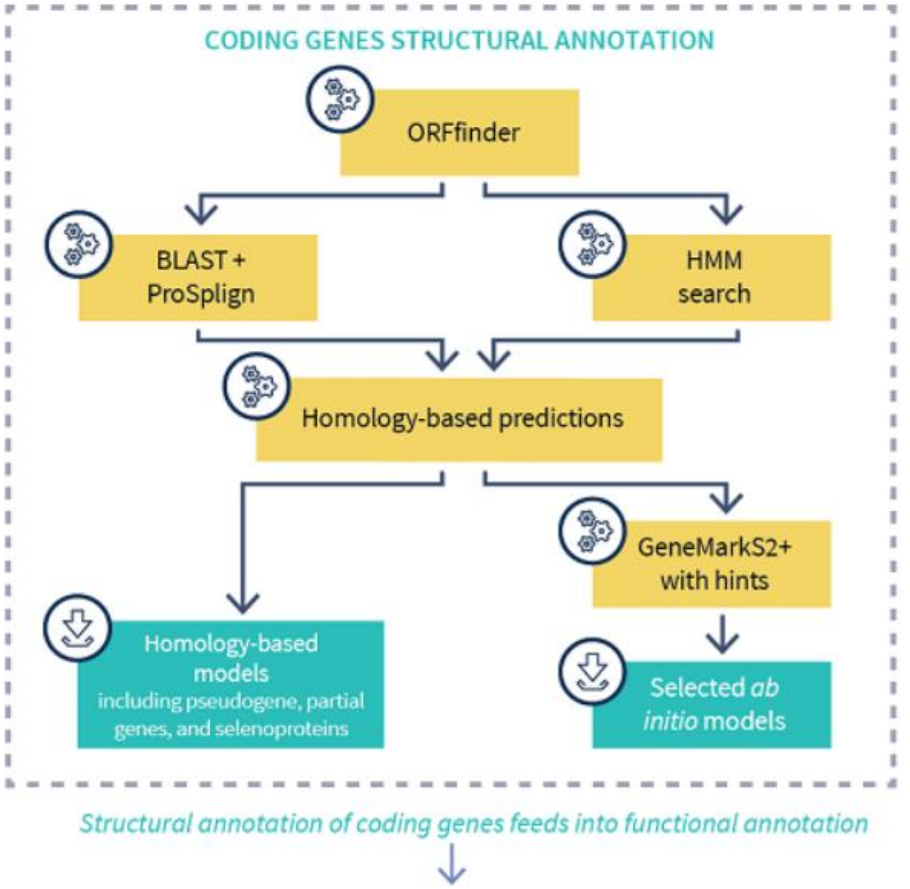
The **NCBI Prokaryotic Genome Annotation Pipeline** (PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

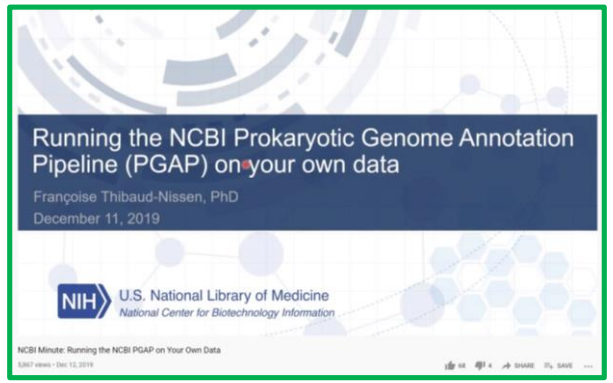
Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality (Li W, O'Neill KR et al 2021, Haft DH et al 2018, Tatusova T et al 2016). Structural and functional annotation uses Protein Family Models, a hierarchical collection of evidence composed of Hidden Markov Model-based and BLAST-based protein families (HMMs and BlastRules) and Conserved Domain Database architectures(CDDs). HMMs, BlastRules and CDDs are used to assign names, gene symbols, publications and EC numbers to the prokaryotic RefSeq proteins that meet the criteria for inclusion in a family. HMMs and BlastRules contribute to structural annotation.

https://www.ncbi.nlm.nih.gov/genome/annotation_prok/







Anotace genomů u prokaryot

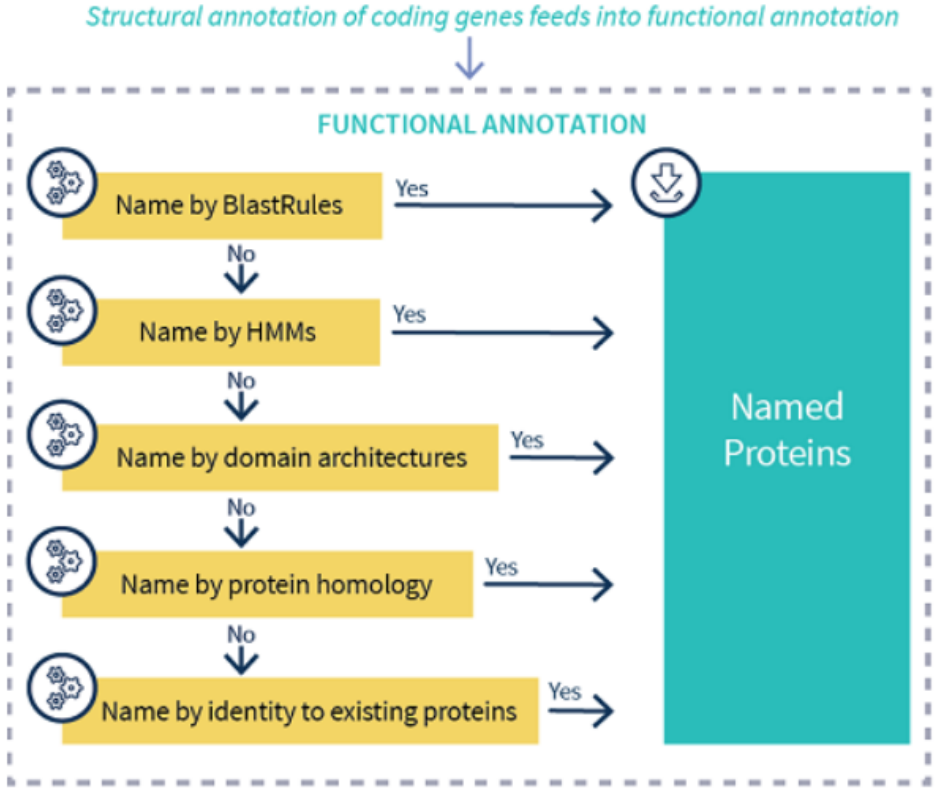
PGAP is now available as a stand-alone software package. You can annotate your genomes on your own machine, local cluster or the Cloud! Get started by watching a short video!

NCBI Prokaryotic Genome Annotation Pipeline

The **NCBI Prokaryotic Genome Annotation Pipeline** (PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality (Li W, O'Neill KR et al 2021, Haft DH et al 2018, Tatusova T et al 2016). Structural and functional annotation uses Protein Family Models, a hierarchical collection of evidence composed of Hidden Markov Model-based and BLAST-based protein families (HMMs and BlastRules) and Conserved Domain Database architectures(CDDs). HMMs, BlastRules and CDDs are used to assign names, gene symbols, publications and EC numbers to the prokaryotic RefSeq proteins that meet the criteria for inclusion in a family. HMMs and BlastRules contribute to structural annotation.

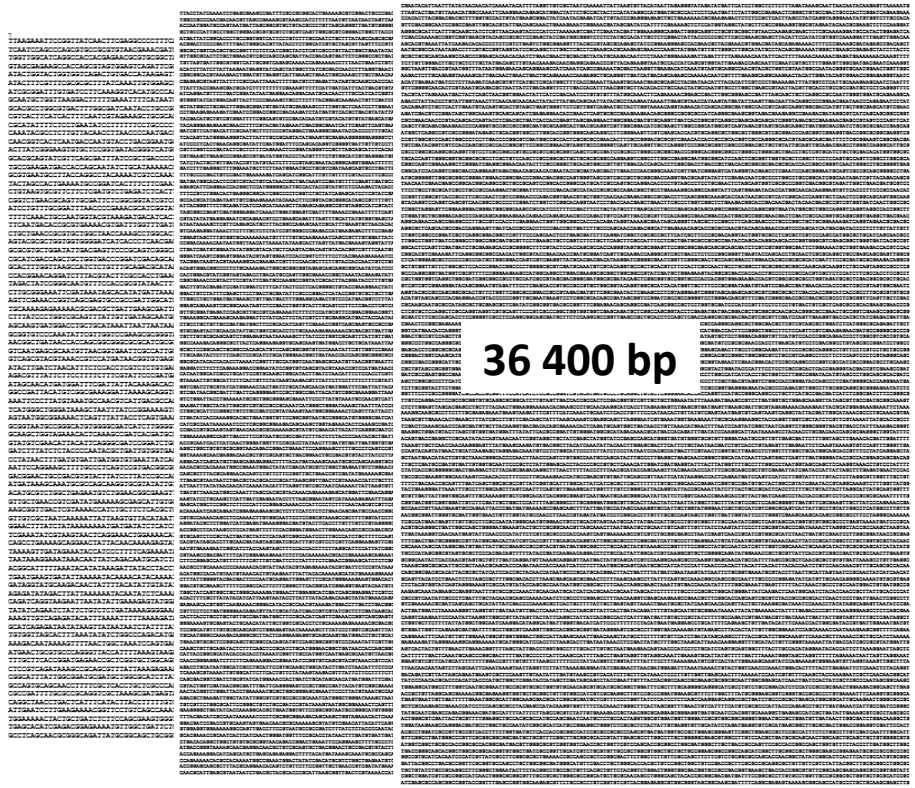


https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

- Mnoho informací je pouze **PREDIKOVÁNO** (s využitím bioinformatiky). **Je nutné vyvarovat se slepého spoléhání na informace uvedené v databázi!**
- **Chyby se mohou šířit – nové sekvence s neznámou funkcí jsou často anotovány na základě sekvenční podobnosti s již existujícími záznamy v databázi!** Chybná anotace může ovlivnit celou skupinu podobných sekvencí!

Predikce genů u eukaryot

- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, > 90 % genomu může být nekódující, jeden gen u člověka připadá *přibližně* na 100 kbp.



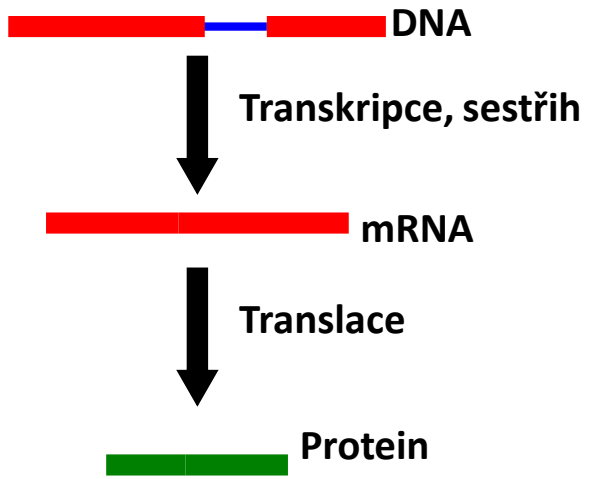
GENSCAN

Predikovány 2 kódující sekvence, z toho jedna neúplná

```
>/tmp/02_12_20-11:25:12.fasta|GENSCAN_predicted_peptide_2|262_aa
MGENWGASDTGDEAKPDPAMACSPEVPGRLLVGQDTAPRGGAEVTGSRGDGHHRFPALAP
DRHREPRPEQGGTQPAEGRGLDSHETEETEKGEMEEMETGKTEGREEMEKGELGENGRAS
DAGMRQSQTQPRSAVPREDTAPGGAGGLYDSEPGKEQRPEVVPSTVPTGRPAQAEGSDPT
RHRSPVCRSPETHILWLTAVRPLGRRRPHVAQTAPLGLKPADKATHPARRCCVATAEGPR
TTFPMTHGQTLAQQGSLRPGAV
>/tmp/02_12_20-11:25:12.fasta|GENSCAN_predicted_peptide_3|1286_aa
XRPLVPSPAEERVLNLPAVVVASSFLLSHLSVGVGVPCATVDARDFVCLASPPQQHHHVG
LGAGGVSCSGSYSEEGLKPGSGTHIHQLGPPVSSFVFPATLLKILINSRIWSAGWKISVW
QSGAWFIDGAFPLRPHGVEGACGCPLYWKGPLFYGAGGERTGSVSVHKFVAMWRKILQNC
HDDAAKFVHLLMSPGCNYLVQEDFVPFLQRPHRTPHGRAGVYTTRLVSSHPQDVVNTHPG
LSFLKEASEFHSRYITTVIQRIFYAVNRSWSGRITCAELRRSSFLQPGGLGASPQRPRAQ
AAWVLWQNVALLEEEADINQLTEFFSYEHFYVIYCKFWELDTDHDLLIDADDLARHNDHG
QDAVCGRAALFLTHCGLKRGAAPWLVRPRDTGDRRDGRPGCGTFSWPKLTATVLSLLPPV
AVPLRPEYSALYKSVLEPRSALRSVDFMLDLALLDGKACPFYQDDRQDLLRSSHTVRTAQ
APDGSHCPCEDPGPSPHRCLTGRKVQKEGKISYADFVWFLISEEDKKTPTSIEYWFRCMD
LDGDGALSMFELEYFYEEQCRRLDSMAIEALPFQDCLCQMLDLVKPRTEGKITLQDLKRC
KLANVFFDTFFNIEKYLDHEQKEQISLLRDGDSGGPELSDWEKYAAEEYDILVAEETAGE
PWEDGPGGRLPSRSGSDVGHGHRSPRPSDFGAAQNTFWLPLGSRPMGSMDVHFNPCSPIS
RPVRFLPSSTGPLRPTLGRLSCVSVLLVVIVCTVALTWQPTSGSPQLPSLCLLLSDTLRQ
LWPPDLAWMVSTPTAARLCVCGTASSTLGSTAFTVSTPAPAVGAQAPPPGAFVEPASSAR
SALLAQRLELKVGSGDCLVLEVGGRAACVSRKTAQGLLRAAGHVFAIPSLQVRGRAQPCG
AEAECAALPAGPEALLRGALTAGRRGPVRVRMRGRGPGAAVTPPARTPPRGRSPPWVPGP
ASVERVRVRTTVTGGSRAPQSWTGASTTGCVNVLCKRTKPFLAITLVRVTGGDHGNNPVL
THDKTLPQETEVAASTRQGRAKPGPPREQHTHPVHPPAAVAVKEADAQGGPVTPEERLRG
LQVVGRRGSTSRGNAAVPKVSSVSPGAPLNSRMPPPGSAKGQDPQQQDSHNPRCPGQCRG
RACTPTPLPEKWGGPVGGPRGRRRCGETQSPAVMPSTCSGGPFPRGLTCAGGQGPHKTST
RGTLTVASVPPEARMFQGSAPHPLNDNRGSPALRGRHPLEIPEQGPDPAHPSPPSRGAFV
SEPRVFTGAAPGPPRSSSMSSVPGGP
```

Predikce genů u eukaryot

- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, > 90 % genomu může být nekódující.
- **Eukaryotické geny:** skládají se z exonů a intronů. Podléhají sestřihu, může probíhat alternativní sestřih.
- Exony mohou být velmi **krátké**, introny velmi **dlouhé**.



Nízká hustota genů, exony/introny, alternativní sestřih:

Hledání jehly v kupce sena, přičemž jehla je rozlámaná na kousky.
Kousky jehly je nutné najít a SPRÁVNĚ poslepovat dohromady.

Predikce genů u eukaryot

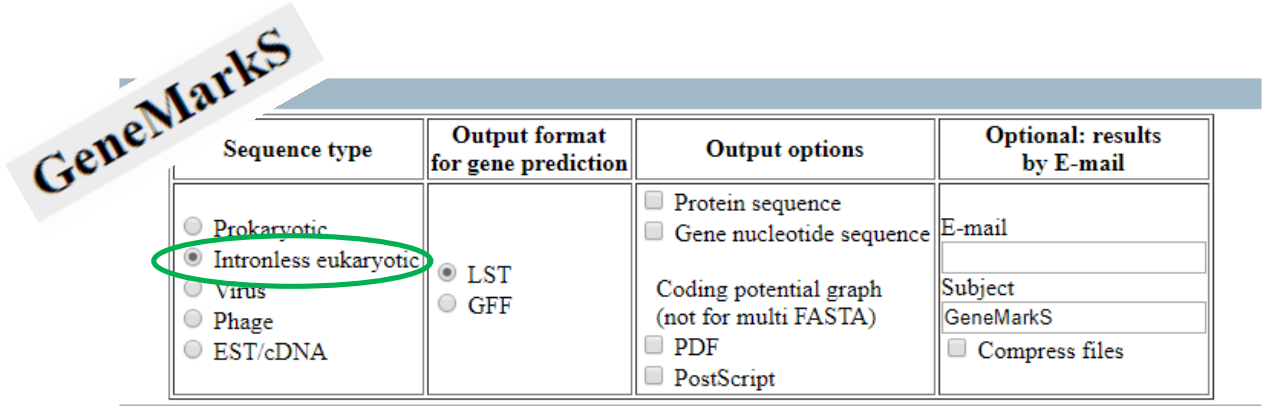
- **Eukaryotické genomy**: velké až obrovské (10 Mbp až 670 Gbp). Mohou mít velmi nízkou hustotu genů, > 90 % genomu může být nekódující.
- **Eukaryotické geny:** skládají se z exonů a intronů. Podléhají sestřihu, může probíhat alternativní sestřih.
- Exony mohou být velmi **krátké**, introny velmi **dlouhé**.

Co pomáhá při predikci:

Signální sekvence, sestřihová místa (GT/AG), zastoupení nukleotidů v kódujících/nekódujících oblastech, ATG.

Predikce genů u eukaryot

- 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.
- Pro některá jednobuněčná eukaryota je možné použít stejné postupy jako pro prokaryota.



Predikce genů u eukaryot

- 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.

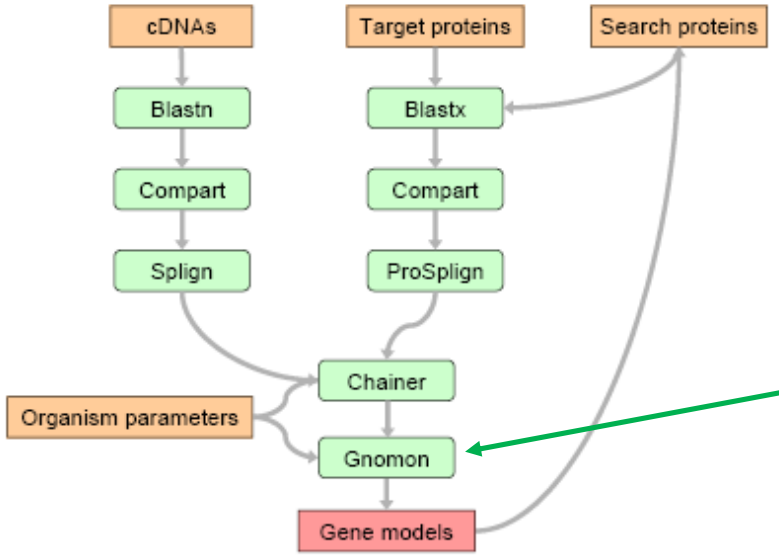
Slime mold = hlenka
Fuligo septica
Dog vomit slime mold

Hlenky jsou záhadné houby v podobě blitky, škraloupu, slizu či průjmu, které se dokážou pohybovat, mají paměť a navzájem komunikují. Tato hlenka leze po stohu poblíž Slezských Rudoltic. | *Foto: DENÍK/František Kuba*

https://www.denik.cz/z_domova/hlenky-na-severu-moravy20090715.html

Metody predikce genů u eukaryot

- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- ***Ab initio*** – Např. HMM (skryté Markovovy modely)



***Gnomon**, the NCBI eukaryotic gene prediction tool*

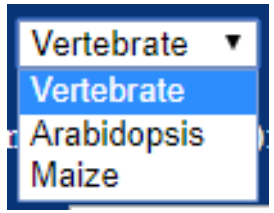
„The core algorithm of the ab initio prediction capability of Gnomon is based on Genscan."

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

Komplexní model struktury genu (HMM + transkripční, translační, sestřihové signály).

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



Eukaryotic Genome Annotation at NCBI: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/

Metody predikce genů u eukaryot

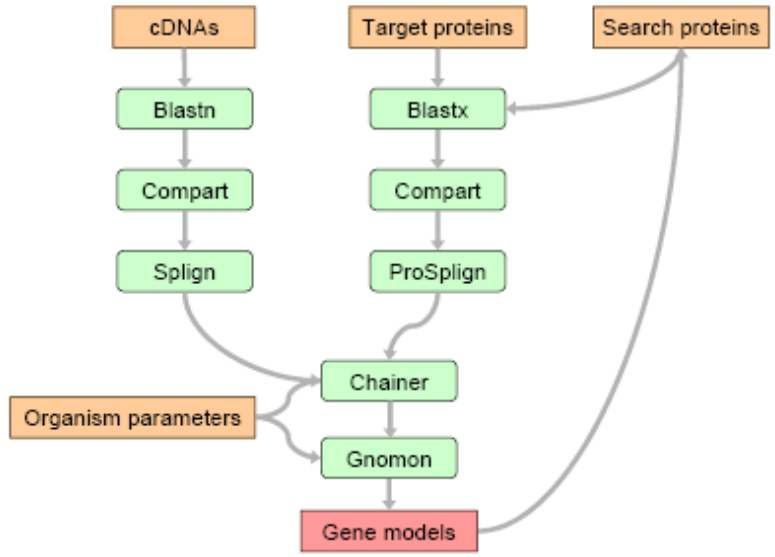
- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- Metody založené na **homologii** – exonové sekvence příbuzných druhů jsou konzervované. Potenciální exony jsou porovnány se sekvencemi v databázi. Nelze použít pro nové geny bez homologů v databázi.
- Metody založené na **konsenzu** (shoda mínění, vzájemný souhlas) – porovnání výstupů z více různých predikčních programů. Výběr shodných výsledků – omezení falešně pozitivních výsledků. Problém: nižší citlivost, vynechání některých genů.

Metody predikce genů u eukaryot

- Metody ***ab initio***/metody založené na **homologii**/metody založené na **konsenzu**.
- V praxi často využívány **kombinace** přístupů, *ab initio* + homologie. Využití **experimentálních** dat – proteiny, RNA sekvence, geny (ze zkoumaného organismu nebo homologní), „spliced alignments“.

***Gnomon**, the NCBI eukaryotic gene prediction tool*

Before we start a genome annotation we collect several data sets. First we collect all available cDNA for the studied organism and sometimes cDNA for closely related organisms. Then we generate a Target protein set and a Search protein set. The former is a collection of the proteins that we believe should be found on the genome. Usually this includes all known proteins for the studied organism and several sets of known proteins for other, well studied genomes. The latter set is a much wider collection of eukaryotic proteins. We try to align on the genome all proteins from the Target Protein Set. The proteins from the Search Protein Set are aligned only if they are similar enough to predicted models, in which case these additional alignments are used in refining the models. In addition to the sequences used for the homology search we create an organism specific parameter set which is used for evaluation of the *ab initio* scores.



https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/

Metody predikce genů u eukaryot

Predicting Genes in Single Genomes with AUGUSTUS

Katharina J. Hoff[1,2] and Mario Stanke[1,2]

[1]University of Greifswald, Institute of Mathematics and Computer Science, Greifswald, Germany
[2]Corresponding authors: *katharina.hoff@uni-greifswald.de*; *mario.stanke@uni-greifswald.de*

AUGUSTUS is a tool for finding protein-coding genes and their exon-intron structure in genomic sequences. It does not necessarily require additional experimental input, as it can be applied in so-called *ab initio* mode. However, extrinsic evidence from various sources such as transcriptome sequencing or the annotations of closely related genomes can be integrated in order to improve the accuracy and completeness of the annotation. AUGUSTUS can be applied to single genomes, or simultaneously to several aligned genomes. Here, we describe steps required for training AUGUSTUS for the annotation of individual genomes and the steps to do the actual structural annotation. Further, we describe the generation and integration of evidence from various sources of extrinsic evidence.

http://bioinf.uni-greifswald.de/webaugustus/

http://exon.gatech.edu/GeneMark/gmep_plus_instructions.html

ABSTRACT

We have made several steps toward creating a fast and accurate algorithm for gene prediction in eukaryotic genomes. First, we introduced an automated method for efficient *ab initio* gene finding, GeneMark-ES, with parameters trained in iterative *unsupervised* mode. Next, in GeneMark-ET we proposed a method of integration of unsupervised training with information on intron positions revealed by mapping short RNA reads. Now we describe GeneMark-EP, a tool that utilizes another source of external information, a protein database, readily available prior to the start of a sequencing project. A new specialized pipeline, ProtHint, initiates massive protein mapping to genome and extracts hints to splice sites and translation start and stop sites of potential genes. GeneMark-EP uses the hints to improve estimation of model parameters as well as to adjust coordinates of predicted genes if they disagree with the most reliable hints (the -EP+ mode). Tests of GeneMark-EP and -EP+ demonstrated improvements in gene prediction accuracy in comparison with GeneMark-ES, while the GeneMark-EP+ showed higher accuracy than GeneMark-ET. We have observed that the most pronounced improvements in gene prediction accuracy happened in large eukaryotic genomes.

GeneMark-ET integrates into GeneMark-ES information on mapped RNA-Seq reads. **GeneMark-EP+** integrates into GeneMark-ES information on cross-species protein sequences. **GeneMark-ETP** integrates into GeneMark-ES both types of external information, RNA reads and cross-species proteins.

GeneMark-ETP
Tomas Bruna, Alexandre Lomsadze and Mark Borodovsky
"GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistence with Extrinsic Data"
bioRxiv, 2023, January 5

Ab initio predikce genů u eukaryot

RESEARCH ARTICLE Open Access

A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms

2020

Nicolas Scalzitti, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch and Julie D. Thompson*

Abstract

Background: The draft genome assemblies produced by new sequencing technologies present important challenges for automatic gene prediction pipelines, leading to less accurate gene models. New benchmark methods are needed to evaluate the accuracy of gene prediction methods in the face of incomplete genome assemblies, low genome coverage and quality, complex gene structures, or a lack of suitable sequences for evidence-based annotations.

Results: We describe the construction of a new benchmark, called G3PO (benchmark for Gene and Protein Prediction PrOgrams), designed to represent many of the typical challenges faced by current genome annotation projects. The benchmark is based on a carefully validated and curated set of real eukaryotic genes from 147 phylogenetically disperse organisms, and a number of test sets are defined to evaluate the effects of different features, including genome sequence quality, gene structure complexity, protein length, etc. We used the benchmark to perform an independent comparative analysis of the most widely used ab initio gene prediction programs and identified the main strengths and weaknesses of the programs. More importantly, we highlight a number of features that could be exploited in order to improve the accuracy of current prediction tools.

Conclusions: The experiments showed that ab initio gene structure prediction is a very challenging task, which should be further investigated. We believe that the baseline results associated with the complex gene test sets in G3PO provide useful guidelines for future studies.

Keywords: Genome annotation, Gene prediction, Protein prediction, Benchmark study

We used the G3PO benchmark to compare the accuracy and efficiency of five widely used ab initio gene prediction programs, namely Genscan, GlimmerHMM, GeneID, Snap and Augustus. Our initial comparison highlighted the difficult nature of the test cases in the G3PO benchmark, since 68% of the exons and 69% of the Confirmed protein sequences were not predicted with 100% accuracy by all five gene prediction programs. Different benchmark tests were then designed in order to identify the main strengths and weaknesses of the different programs, but also to investigate the impact of the genomic environment, the complexity of the gene structure, or the nature of the final protein product on the prediction accuracy.

- Fungují „relativně dobře“ pro „průměrné proteiny“.
- Problém s identifikací krátkých proteinů, dlouhých proteinů, genů s mnoha exony (>20), genů z méně studovaných druhů.
- Nezbytné pro identifikaci dosud neznámých proteinů…☹

„Take-home message“

- **Dva hlavní přístupy: metody *ab initio*/metody založené na homologii (sekvenční podobnosti).**
- **Predikce často zjednodušena na predikci ORFs kódujících proteiny.**
- **K predikci jsou využívány: obsah GC, zastoupení kodonů, signální sekvence (startovní/stop kodony, RBS, sestřihové signály).**
- **Markovovy (skryté) modely.**
- **Predikce genů u prokaryot (malé, kompaktní genomy, bez intronů) funguje výrazně lépe než u eukaryot (velké komplexní genomy, introny, alternativní sestřih).**

Použitá a doporučená literatura

REVIEW Open Access

Bacterial group I introns: mobile RNA catalysts

Georg Hausner[1], Mohamed Hafez[2,3] and David R Edgell[4*]

Group II introns in the bacterial world

Francisco Martínez-Abarca and Nicolás Toro*

Grupo de Ecología Genética, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Profesor Albareda 1, 18008 Granada, Spain.

Base composition bias might result from competition for metabolic resources

Eduardo P.C. Rocha and Antoine Danchin

High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes

Laurence D. Hurst[1*] and Alexa R. Merchant[2]

Horizontal gene transfer: building the web of life

Shannon M. Soucy[1], Jinling Huang[2] and Johann Peter Gogarten[1,3]

ESSENTIAL BIOINFORMATICS,
Jin Xiong, 2006

Microbial gene identification using interpolated Markov models

Steven L. Salzberg[1,2,*], Arthur L. Delcher[3], Simon Kasif[4] and Owen White[1]

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, [2]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA, [3]Department of Computer Science, Loyola College in Maryland, Baltimore, MD 21210, USA and [4]Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

RESEARCH ARTICLE Open Access

Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters

Paul Gagniuc[1*] and Constantin Ionescu-Tirgoviste[2]

© 1998 Oxford University Press Nucleic Acids Research, 1998, Vol. 26, No. 4 1107–1115

GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashin and Mark Borodovsky[1,*]

School of Biology and [1]Schools of Biology and Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received August 14, 1997; Revised and Accepted December 30, 1997

Predicting Genes in Single Genomes with AUGUSTUS

Katharina J. Hoff[1,2] and Mario Stanke[1,2]

Bacteriology

Michael T Madigan, *Southern Illinois University, Carbondale, Illinois, USA*

Deborah O Jung, *Southern Illinois University, Carbondale, Illinois, USA*

ENCYCLOPEDIA OF LIFE SCIENCES © 2007, John Wiley & Sons, Ltd. www.els.net

GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions

John Besemer[1], Alexandre Lomsadze[1,3] and Mark Borodovsky[1,2,*]

[1]School of Biology and [2]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA and [3]Gene Probe, Inc., 883 Heritage Place, Atlanta, GA 30033-4103, USA