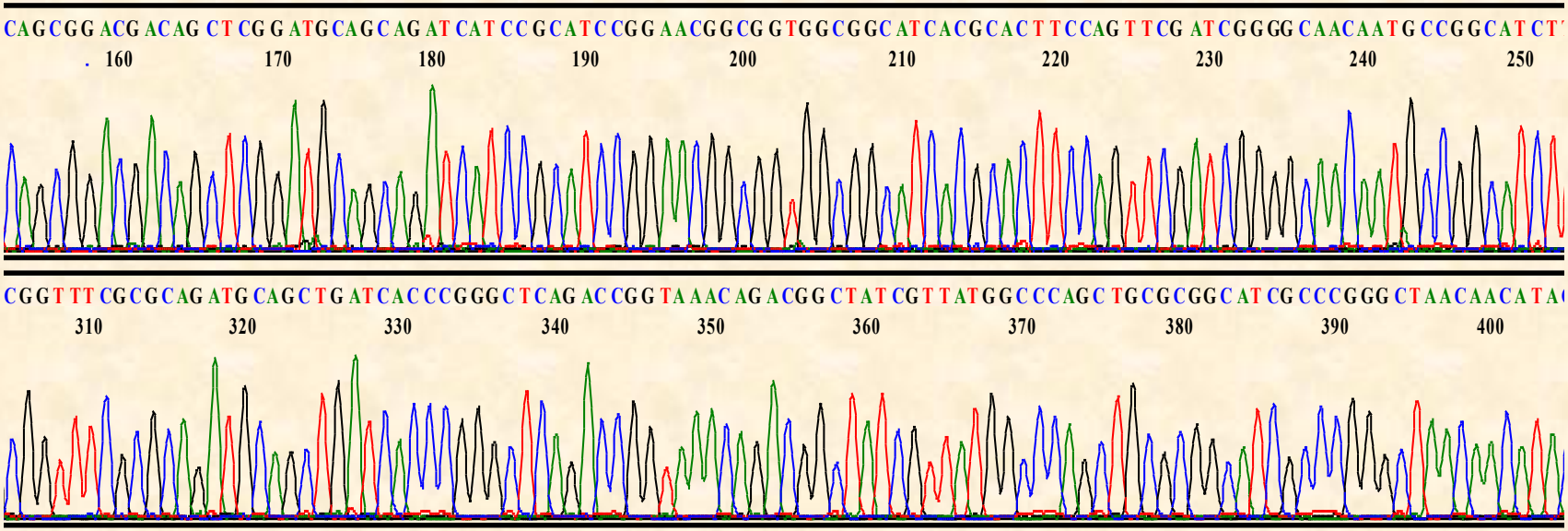
Predikce genů



GATAGCGTAATGATCGGCTGGCTGCCGCATTTCATGCTGGTTTCCCAACGAAAATAACCGCTCACGGTGCCATCACGATCGCACACCGCAAAATCGGCGG

TACAGGTGGTCGCGCCCGCCGCCAGCACATCGCTGCGCCAATAATGATCTTTCAGCGGACGACAGCTCGGATGCAGCAGATCATCCGCATCCGGAACGGC

GGTGGCGGCATCACGCACTTCCAGTTCGATCGGGGCAACAATGCCGGCATCTTTCAGGGCAAAGCGAATAAACAGCACGCTCACTTCCGCGCGCAGCGCC

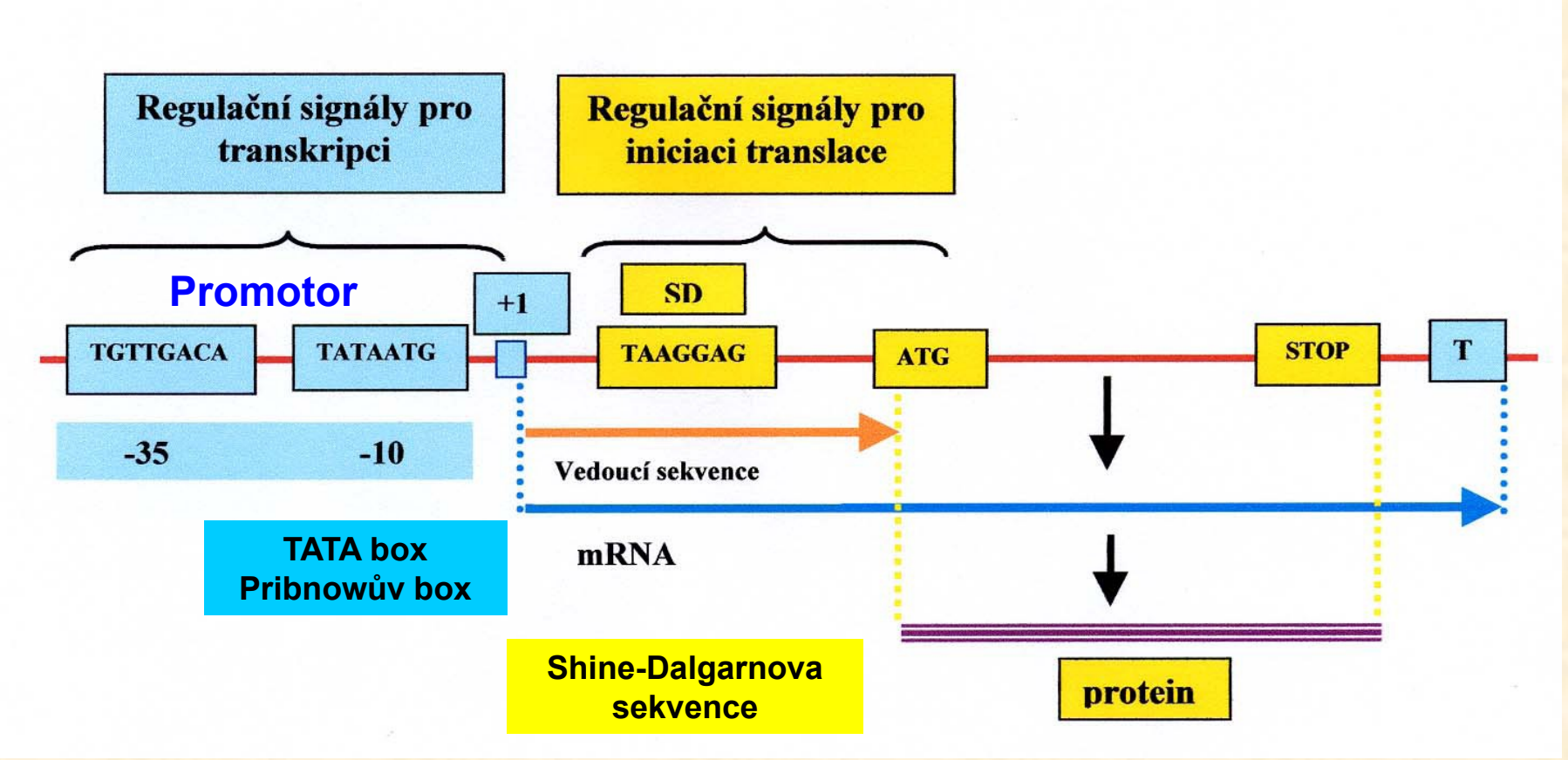
AGCGCGGTTTCGCGCAGATGCAGCTGATCACCCGGGCTCAGACCGGTAAACAGACGGCTATCGTTATGGCCCAGCTGCGCGGCATCGCCCGGGCTAACAA

CATACAGGTGGCGACCATCAATCACGGTCGGGGCGGCCGGATCACGGCTGGCTTCCGGATAGGCGCTCAGCAGGGTAACGGCATCCACAATCACCAGCAT

Opravdu ORF kóduje protein?

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

Translační a transkripční signální sekvence



Prokaryota

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

Algoritmy a nástroje pro identifikaci genů

- **Každý program má výhody a nevýhody – rozumné použít více predikčních nástrojů.**

  GeneMark

  GlimmerM

  GRAIL

  GenScan

  Fgenes

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.

Prokaryotické geny

- **Velmi jednoduchý přístup k predikci genů**

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

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

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

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!

Eukaryotické geny
Jednobuněčná eukaryota

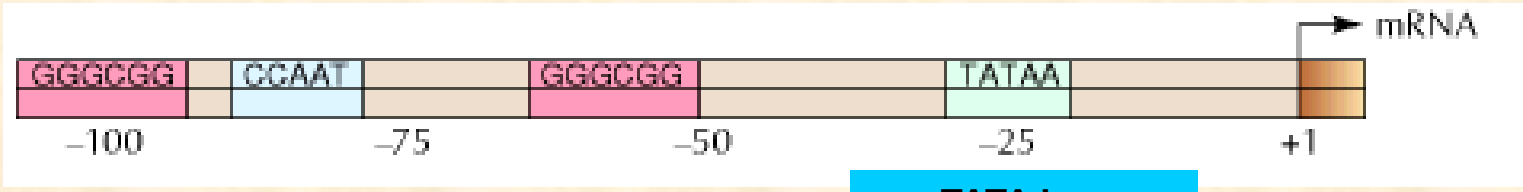
- **Genomy jednobuněčných eukaryot se výrazně liší (frekvence intronů, jak velká část genomu je tvořená geny kódujícími proteiny).**

- ***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.**

Translační a transkripční signální sekvence

Regulační signály pro transkripci



GC box

GC box

TATA box
Hognessův box

Promotor RNA-polymerasy II

Regulační signály pro iniciaci translace

(gcc)gccRccAUGG

Kozak sequence
Sekvence Kozakové

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.

Eukaryota

Mnoho intronů, dlouhé intergenové úseky

***Ab initio* STATISTICKÉ METODY**

IDENTIFIKOVANÉ EXONY

SEKVENČNÍ HOMOLOGIE