

# **Počítačové vyhledávání genů a funkčních oblastí na DNA**

# Hodnota genomových sekvencí záleží na kvalitě anotace

- Anotace – Charakterizace vlastností genomů
  - s použitím výpočetních a experimentálních metod
- Hledání genů:
  - Predikce – Kde jsou geny lokalizovány?
  - Podobnost – Jak geny vypadají?
  - Domény – Jakou funkci mají kódované proteiny?
  - Funkce – V jakých metabolických drahách?
  - Evidence – Experimentální důkaz genu
    - Transkriptom
    - Proteom

# Hledání genů

- Geny tvoří **obsahovou složku** genomu
  - Jedinečné sekvence odpovědné za funkční produkt
    - Variabilní délka
    - Strukturní geny
      - jednoduché
      - složené z exonů a intronů
  - Geny pro funkční RNA
    - rRNA (ribosomal RNA)
    - tRNA, tmRNA (transfer RNA)
    - snRNA (small nuclear)
    - snoRNA (small nucleolar)
    - RNAi (interfering RNA)
    - CRISPR lokusy
  - Regulační sekvence (ori, promotory, terminátory)

# Co nás zajímá při hledání genu

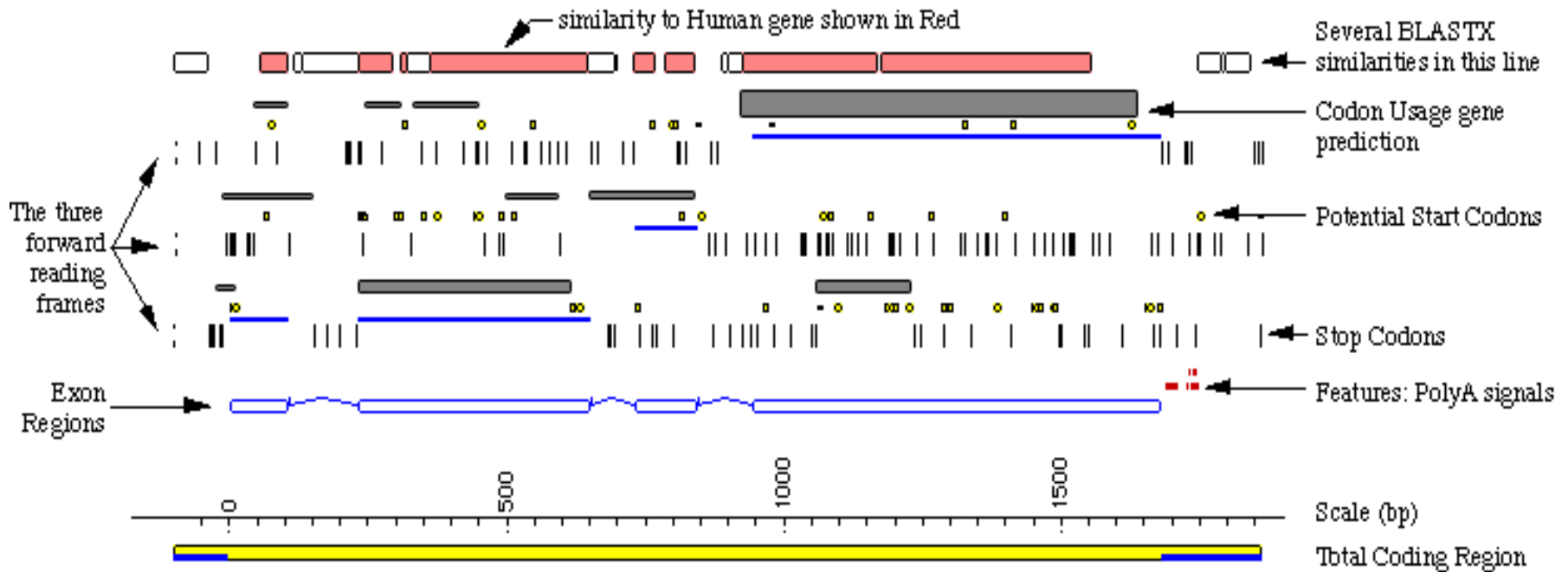
U necharakterizované sekvence DNA zjišťujeme:

- Která oblast kóduje protein
- Který DNA řetězec je kódující
- Který čtecí rámeček je využíván
- Jaké jsou koordináty genu
- Kde jsou hranice exonů a intronů
- Kde se nacházejí regulační sekvence
- Jaká je modulární struktura genomů

# Přístupy pro hledání genů

1. Metody založené na hledání podobností s již popsányými geny
2. Metody srovnávací genomiky
  - Srovnání více dokončených genomů
  - Hledání konzervativních oblastí
3. Využití algoritmů a statistických metod pro analýzu sekvence
4. Integrované přístupy, automatické anotace

# Integrovaný přístup při expertní anotaci genomů

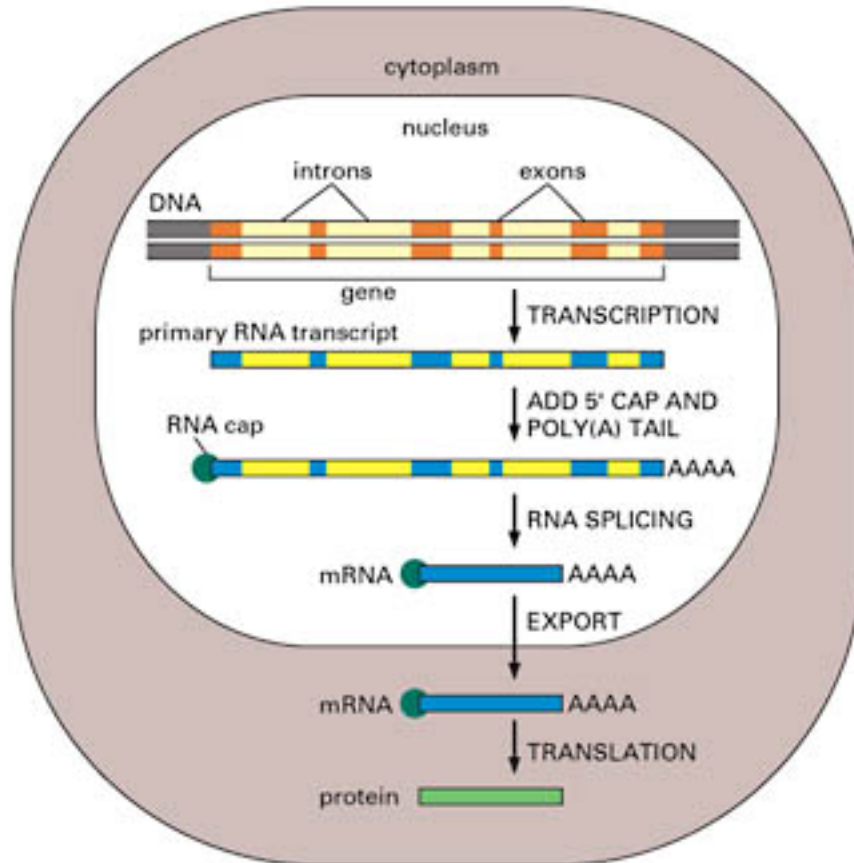


# Příklady velikostí genomů

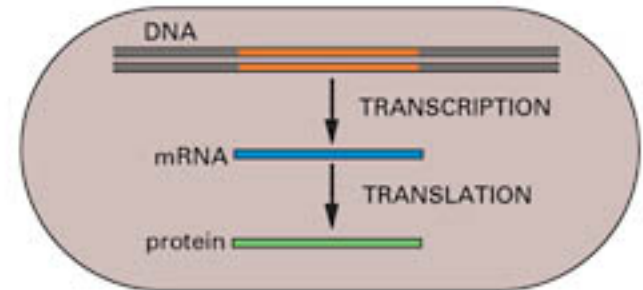
Druh	Velikost	Genů	Genů na Mb
<i>H. sapiens</i>	3 200 Mb	22 000	7
<i>D. melanogaster</i>	137 Mb	13 338	97
<i>C. elegans</i>	85,5 Mb	18 266	214
<i>A. thaliana</i>	115 Mb	25 800	224
<i>S. cerevisiae</i>	15 Mb	6 144	410
<i>E. coli</i>	4,6 Mb	4 300	934

# Prokaryotický versus eukaryotický gen

(A) EUCARYOTES



(B) PROCARYOTES





# Prokaryotický versus eukaryotický gen vyžadují odlišné přístupy

- Prokaryota
  - malé genomy  $0.5 - 10 \cdot 10^6$  bp
  - Vysoká hustota kódujících sekvencí (>90%)
  - Žádné introny (vyjímky Archea, fágy)
  - Hledání otevřených čtecích rámců
  - Doplněno např. hledáním signálů pro vazebná místa ribozómu
  - Operony: jeden transkript, mnoho genů
  - Úspěšnost cca 99 %
  - Problémy: překrývající se ORFs, krátké geny, místa TSS a promotory
- Eukaryota
  - Velké genomy  $10^7 - 10^{10}$  bp
  - Nízká hustota kódujících sekvencí (<50%)
  - UTRs
  - Struktura intron/exon
  - Statistické modely frekvencí nukleotidů
  - Sledování závislostí přítomných ve struktuře kodonů
  - Obsah GC
  - Přesnost dosahuje cca 50 %
  - Problémy: mnoho!
    - postranskripční modifikace
    - alternativní sestřih

# 1. Metody založené na hledání podobností s již popsányými geny

- Založené na konzervativním charakteru sekvencí s určitou funkcí
- Využívají nástroje pro lokální nebo globální přiložení sekvencí (BLAST, FASTA, LAGAN, AVID, atd.)
- Nemohou identifikovat geny, které nejsou v databázi (~50% genů)
- Omezení u sekvencí s nízkou podobností

# Metody založené na hledání podobností

- Databáze
  - Proteiny
  - cDNA (evidence RNA)
  - EST, UniGene
- Nástroje pro párové přiložení sekvencí umožňující analýzu genů
  - Hledání genů na základě podobnosti sekvencí proteinů
    - blastx
    - tblastn
    - fastX
    - genomové prohlížeče
- Neumožňuje nalézt geny, které nejsou v databázi

# Odhalení genů eukaryot s použitím ESTs

- Expressed Sequence Tags (ESTs) reprezentují **sekvence exprimovaných genů** (cDNA).
- Jestliže se oblast shoduje s EST s vysokou stringencí, pravděpodobně se jedná o gen
  - EST podává přesnou predikci hranic exonů.

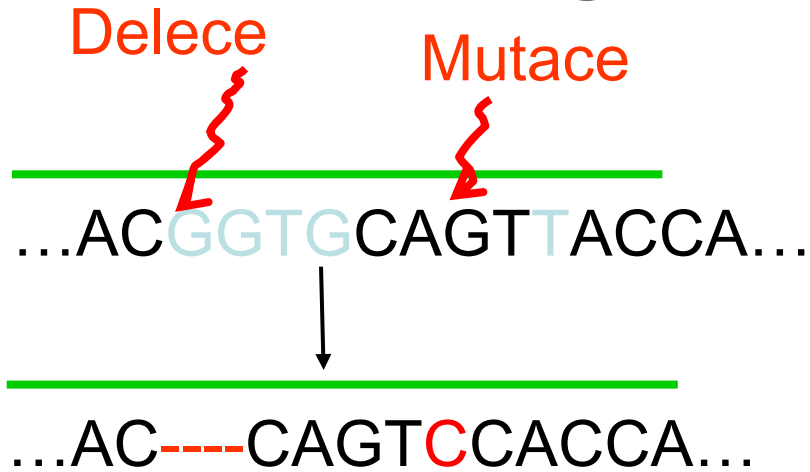
## 2. Srovnávací genomika – hledání na základě homologie

- Hledání založené na předpokladu, že kódující sekvence jsou více konzervativní než nekódující
- Dva přístupy:
  - intra-genomický (genové rodiny)
  - inter-genomický (mezi druhy)
- Mnohonásobné přiložení homologických oblastí
  - exony
  - regulační oblasti
- Obtížné stanovení limitů podobnosti a optimální evoluční vzdálenosti

# Co je srovnáváno?

- **Lokalizace genů v genomu**
- **Struktura genů**
  - Počet exonů
  - Délky exonů
  - Délky intronů
  - Podobnost sekvencí
- **Vlastnosti genů**
  - Místa sestřihu
  - Využití kodonů
  - Konzervované sekvence

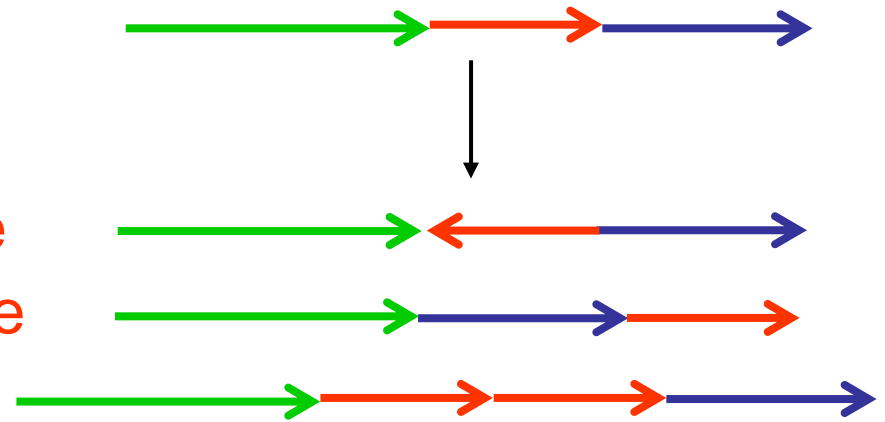
# Evoluce na úrovni DNA - Problém globálního přiložení



Výsledná sekvence

PŘESKUPENÍ

Inverze  
Translokace  
Duplikace

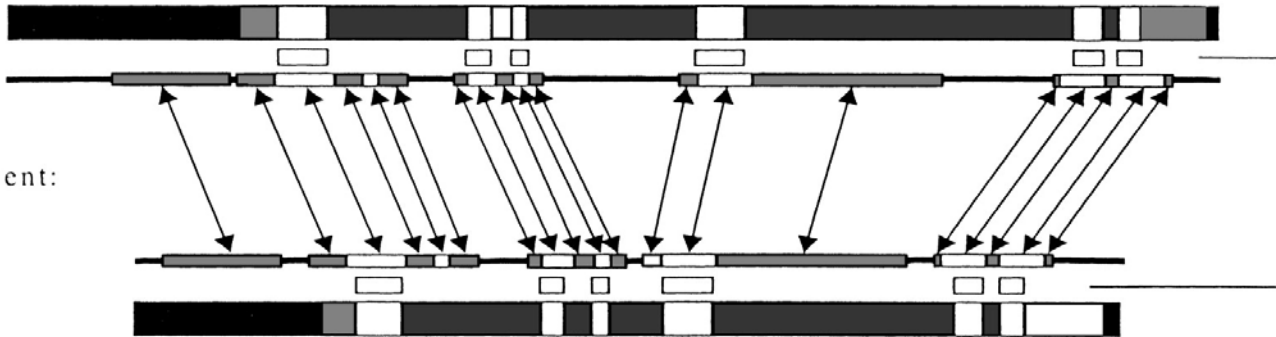


Nalezení nejefektivnější transformace jedné sekvence do druhé vyžaduje využití přístupů pro identifikaci přestaveb

# Příklad srovnání lokusů a chromozómů

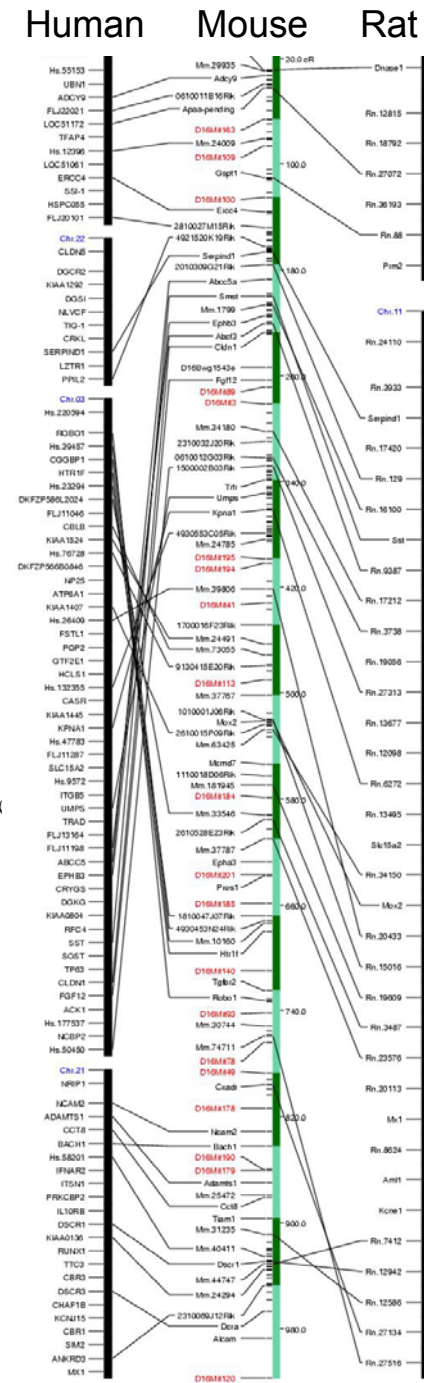
Charakterizace rozdílů umožňuje odhalit mechanismy změn

Human Locus: HUMPCNA



Mouse Locus: MMPCNAG

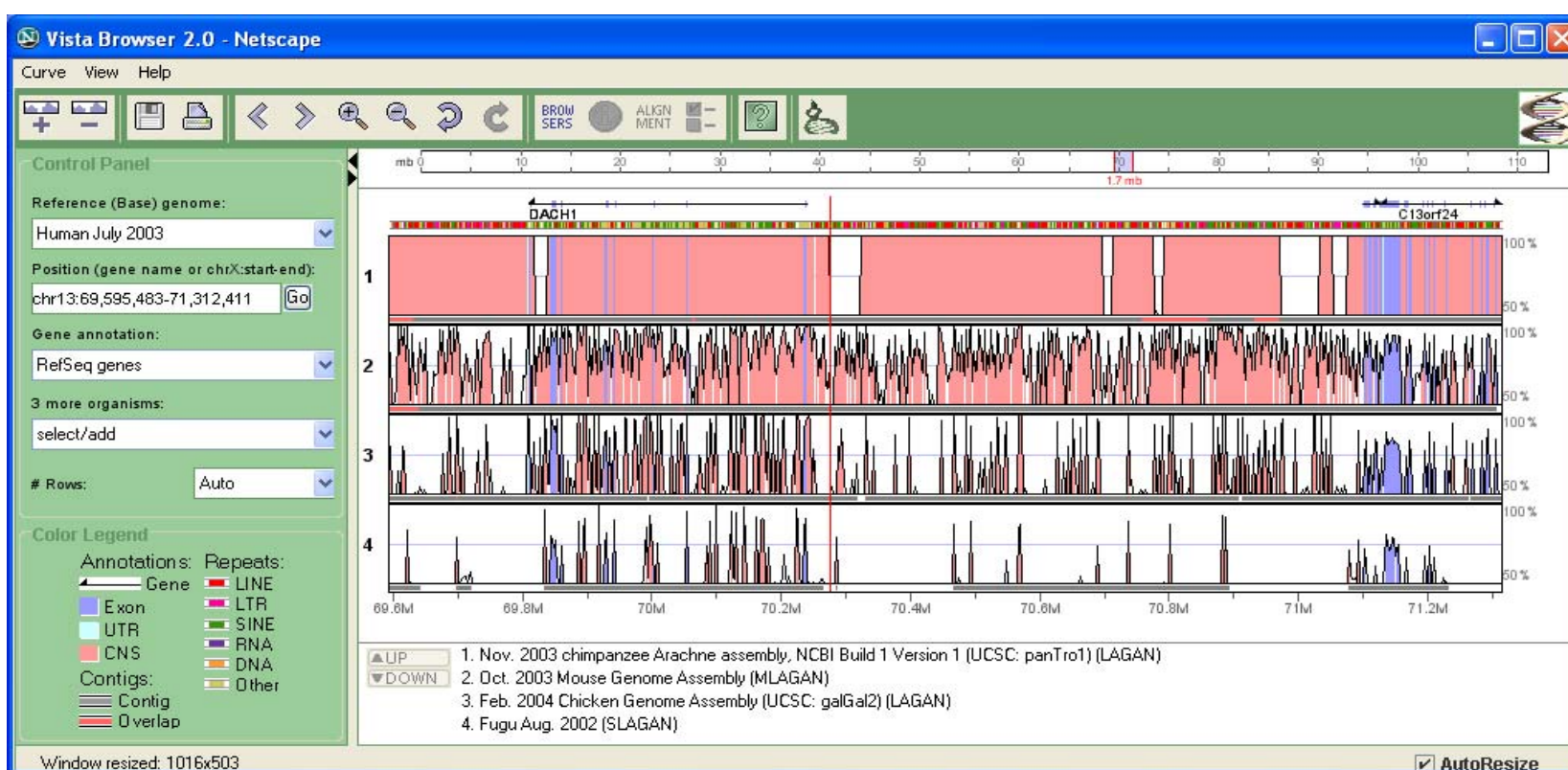
Alignment:





# Proč používat přístupy srovnávací genomiky ?

- Konzervovanost sekvencí v průběhu značných evolučních vzdáleností značí specifickou funkci (geny, funkční-regulační oblasti)
- Ztráta konzervovanosti během krátkých evolučních vzdáleností značí adaptivní evoluci



- šimpanz
- myš
- kuře
- Fugu

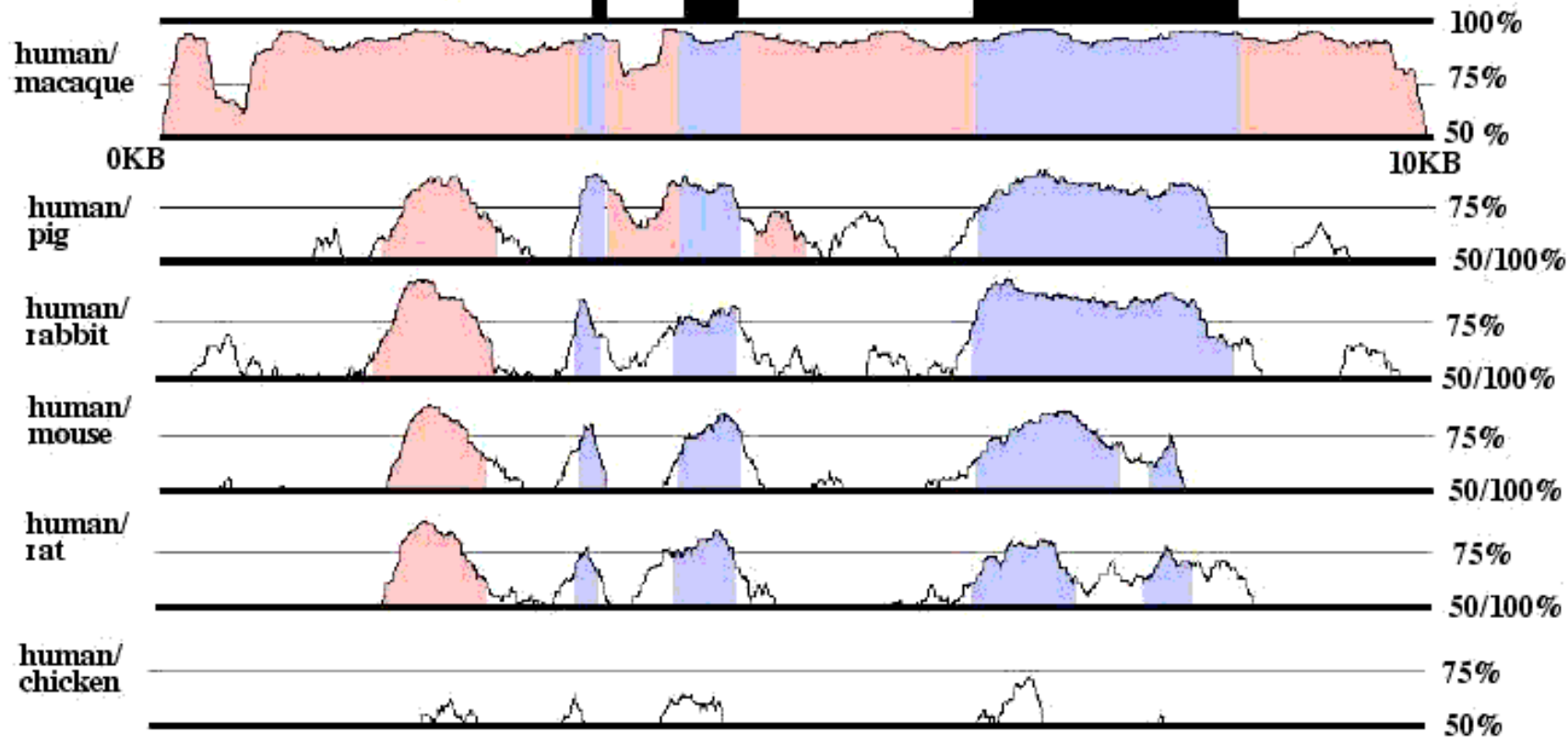
# Konzervativní charakter regulačních oblastí a exonů

## Multi-Species Comparative Analysis

Liver  
Enhancer



Apolipoprotein AI gene



# Základní zdroje a přístupy

- Databáze
  - NCBI: Genomy, Geny, Proteiny, SNPs, ESTs, Taxonomie, atd.
  - databáze genomových center
- Analytický software
  - Databázové dotazy (nalezení podobných sekvencí), algoritmy pro přiložení, shluková analýza, vyhledávání repetice, predikce genů
- Algoritmy pro dlouhá globální přiložení
  - lokální přiložení s rozšířeným vkládáním mezer – citlivé, ale málo specifické pro dlouhé sekvence
    - BLASTZ
    - BLAT
  - globální přiložení
    - AVID
    - LAGAN
    - S-LAGAN
    - MAVID, MLAGAN

# AVID

- Umožňuje srovnání pouze homologních sekvencí bez duplikací, inverzí nebo translokací
- Pokud je aplikován na celé genomy, vyžaduje předem přípravu a identifikaci vzájemně si odpovídajících regionů

# LAGAN

## (Limited Area Global Alignment)

- Umožňuje srovnat mnohem delší sekvence než AVID v důsledku jiného algoritmu pro identifikaci vzájemně odpovídajících si úseků
- Používá se společně s následným lokálním přiložením dlouhých sekvencí (BLAT)
  - rat – mouse
  - rat - human

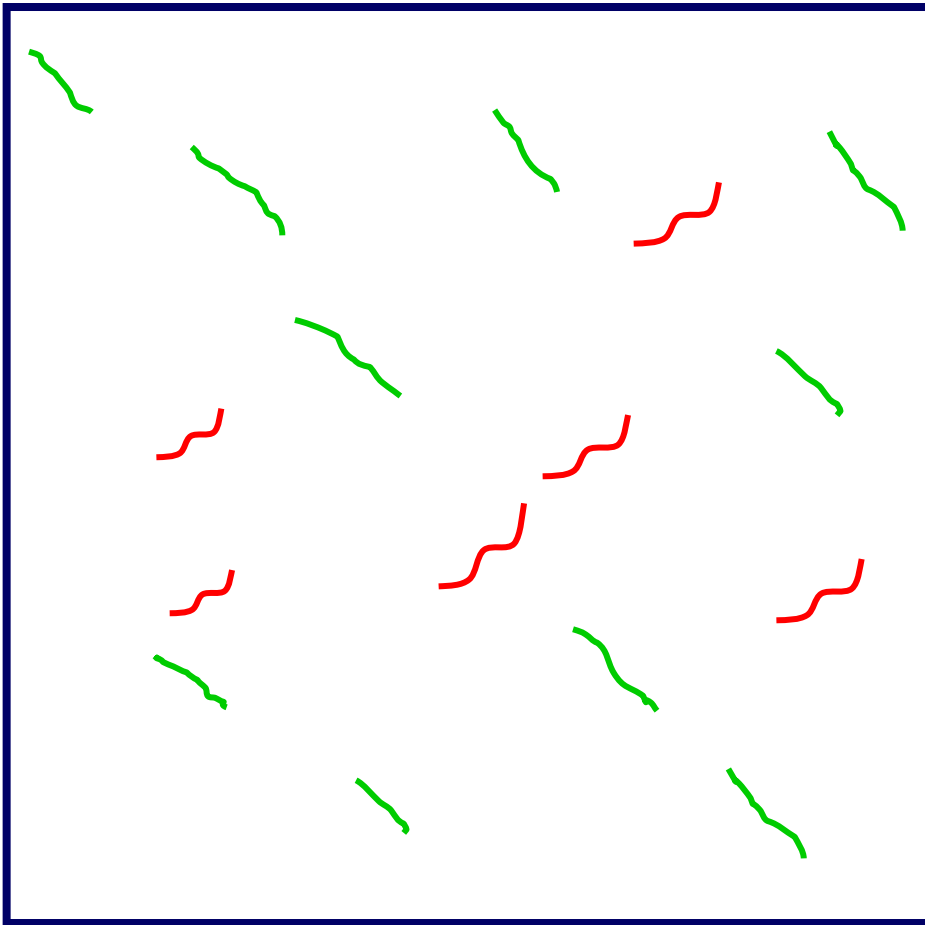
# Multi-LAGAN (MLAGAN)

- V porovnání s LAGAN provádí navíc mnohonásobná globální přiložení
- Nejprve provede přiřazení více příbuzných genomů a následně přiřazuje genomy více fylogeneticky vzdálené
- Umožňuje konstrukci fylogenetických stromů na základě globálního přiložení genomů

# Shuffle-LAGAN (S-LAGAN)

- Slouží pro globální přiložení kompletních sekvencí genomů
- Detekuje genomová přeskupení a inverze
- Poskytuje přiřazení všech kombinací vložených sekvencí

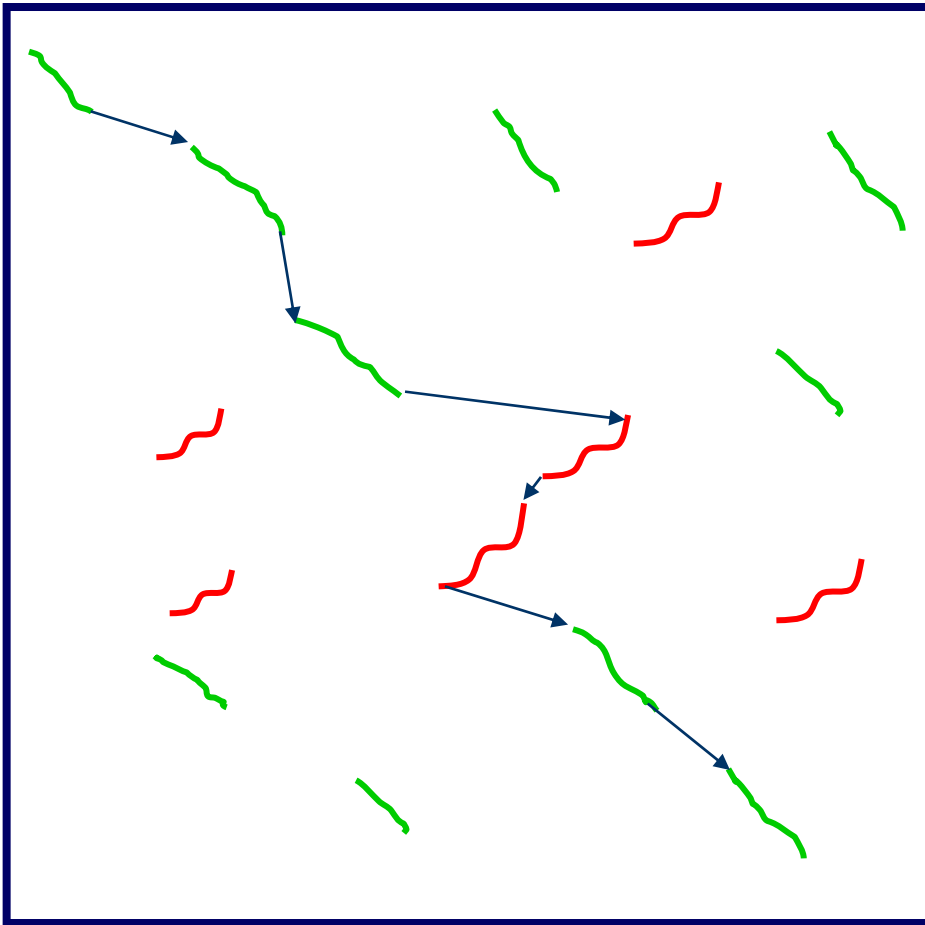
# S-LAGAN: příklad algoritmu



1. Nalezení lokálních přiřazení
2. Sestavení hrubé mapy homologií
3. Globální přiřazení dle odpovídajících si částí

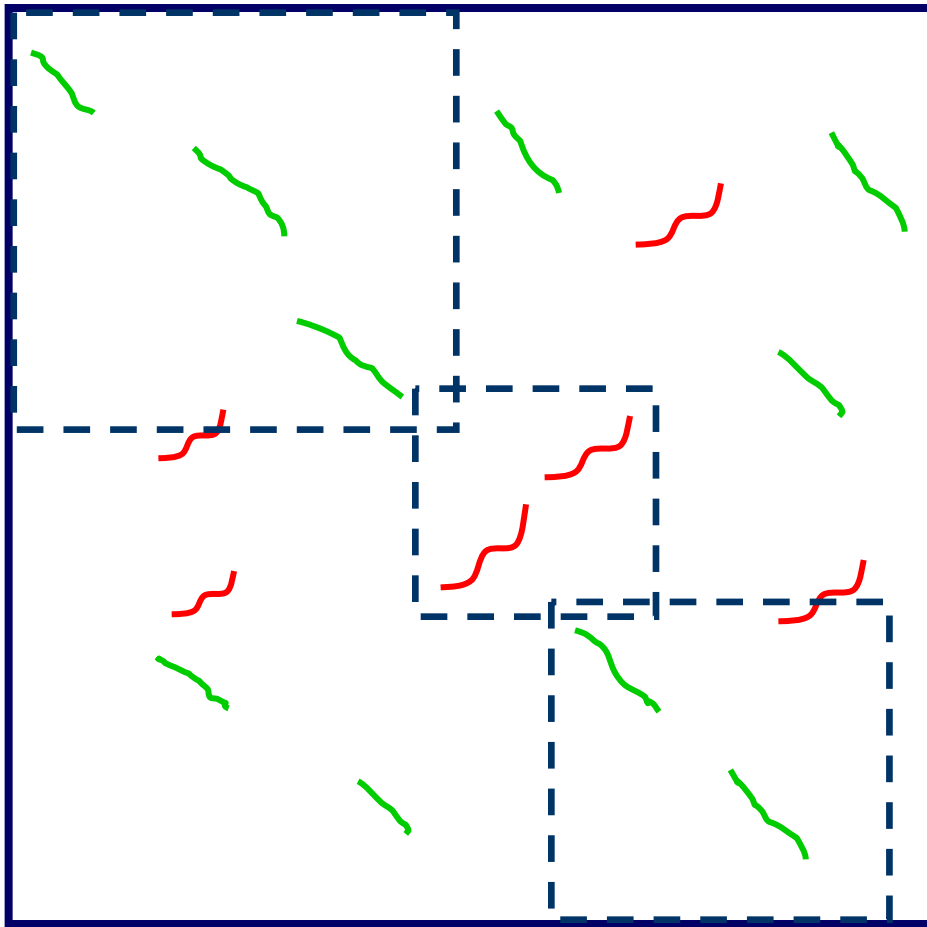


# S-LAGAN: příklad algoritmu



1. Nalezení lokálních přiřazení
2. Sestavení hrubé mapy homologií
3. Globální přiřazení dle odpovídajících si částí

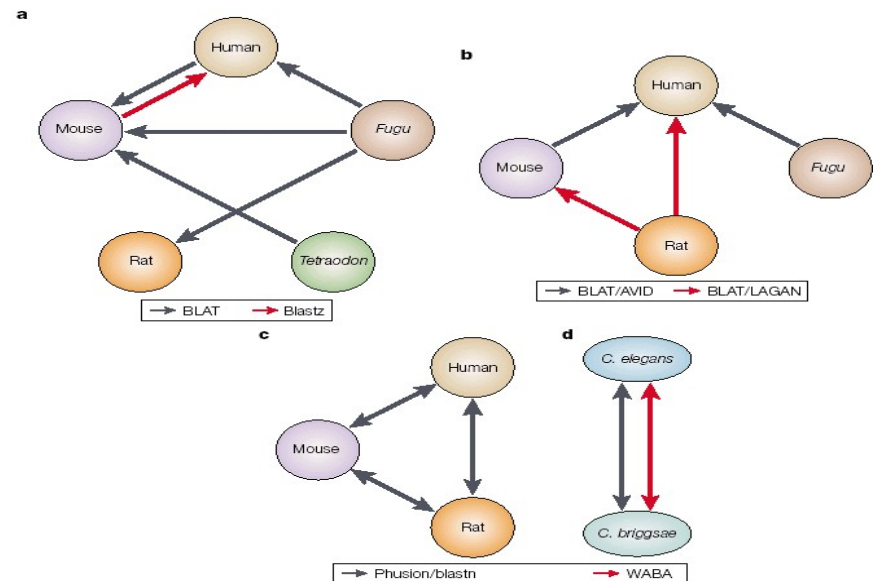
# S-LAGAN: příklad algoritmu



1. Nalezení lokálních přiřazení
2. Sestavení hrubé mapy homologií
3. Globální přiřazení dle odpovídajících si částí

# Precomputed alignments

- U významných skupin organismů jsou k dispozici rozsáhlá mezidruhová srovnání
  - UC Santa Cruz/PennState (translated BLAT or BLASTZ)
  - Berkeley Genome Pipeline (BLAT/AVID)
  - Ensembl (Phusion/Blastn)
  - Vista Genome Server (LAGAN/SLAGAN/AVID)
  - NMPDR (National Microbial Pathogen Data Resource)





# Vista Tools

<http://genome.lbl.gov/vista/index.shtml>



Tools for Comparative Genomics



About Us



Cite Us



Contact Us

[VISTA Home](#)

[Custom Alignment](#)

[Browser](#)

[Enhancer DB](#)

[Downloads](#)

[Publications](#)

[Help](#)

This web site will be down for maintenance on Tuesday Nov. 11, 2014. Sorry for the inconvenience.

VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species.

## Submit Your Sequences

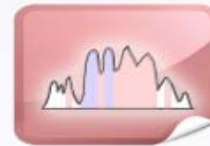
mVISTA



- » [mVISTA](#)  
Align and compare your sequences from multiple species
- » [rVISTA](#)  
Locate regulatory sequences in your data using comparative sequence analysis and transcription factor binding site search.
- » [qVISTA](#)  
Compare your sequences against whole-genome assemblies.
- » [wgVISTA](#)  
Align pair of sequences up to 10Mb long (finished or draft) including microbial whole-genome assemblies.

## Precomputed Alignments

VISTA Browser



- » [VISTA-Point](#)  
Access complete data and visual presentation of pairwise and multiple alignments of whole genome assemblies.
- » [VISTA Browser](#)  
Examine pre-computed pairwise and multiple alignments of whole genome assemblies.
- » [Whole Genome rVISTA](#)  
Identify transcription factor binding sites that are conserved between species and over-represented in upstream regions of groups of genes.
- » [Microbial Genomes](#)  
Access pre-computed full scaffold alignments for microbial genomes through the VISTA component of [IMG](#).

## New tool from VISTA family!



**VISTA Region Viewer (RViewer)** is an interactive on-line tool for comparing and prioritizing genomic intervals.

## Updates

### April 2014

Updated the [Sorghum](#), [Monkey flower](#), [Moss](#), [Maize](#), [Medicago](#), [Switchgrass](#), and [Soybean](#) assemblies, and added 5 new plants: [C. grandiflora](#), [Drummond's rockcress](#), [Turnip mustard](#), [A. halleri](#), and [Hall's parietgrass](#).

180 New whole-genome plant alignments are added to [VISTA Browser](#).

### August 2013

Updated the [C. elegans](#) and [C. briggsae](#) assemblies, and added 5 new worms: [C. brenneri](#), [C. remanei](#), [C. japonica](#), [C. sp. 11](#), and [C. angaria](#).

» [Vista News Archive](#)

## » [Enhancer DB](#)



Experimentally validated human noncoding fragments with gene enhancer activity as assessed in transgenic mice.  
<http://enhancer.lbl.gov/>

## » [JGI Genome Portal](#)



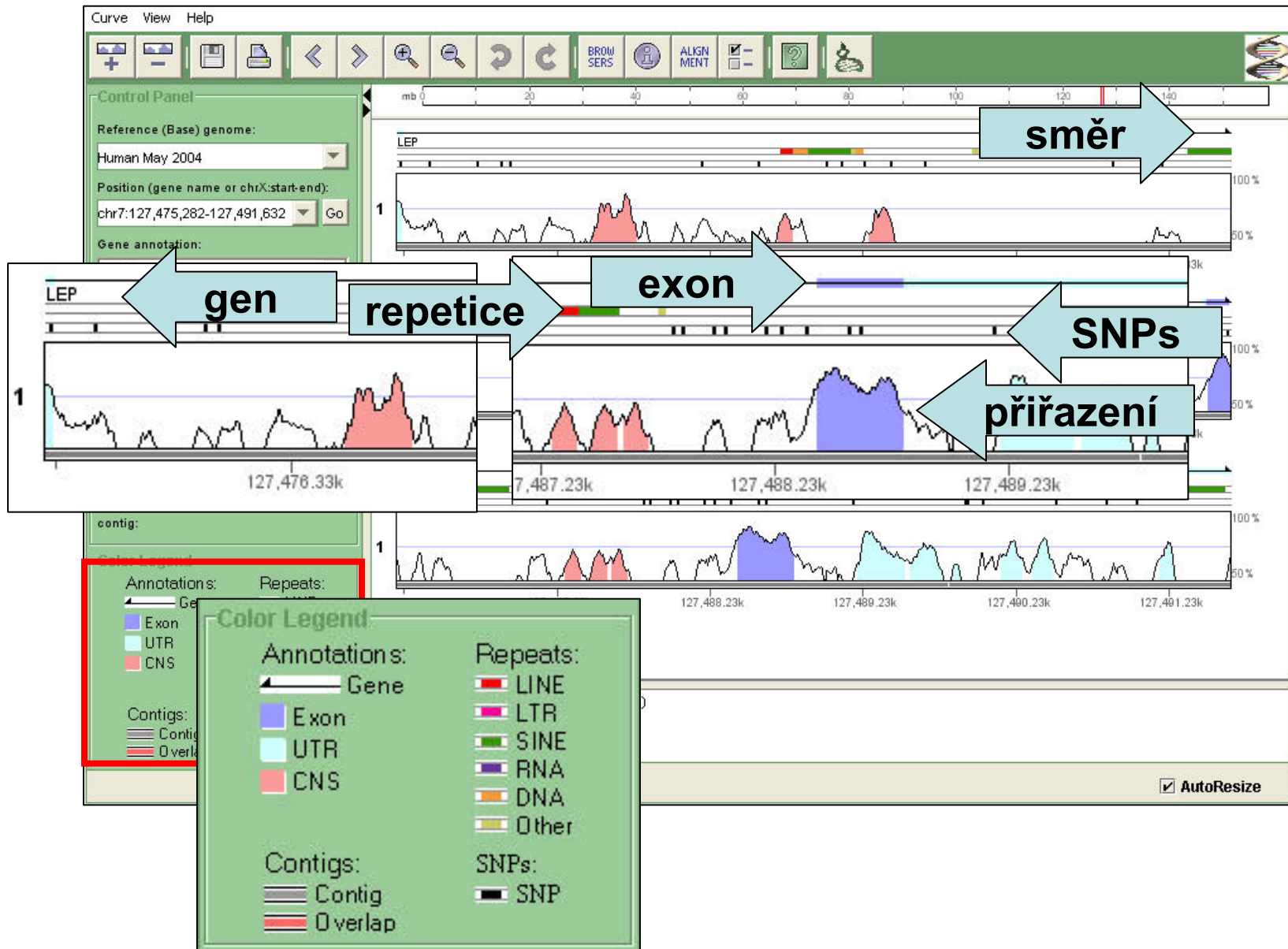
Find VISTA alignments for a number of genomes sequenced in the Department of Energy Joint Genome Institute <http://genome.jgi-psf.org/>

## » [Other Projects](#)



[Phylo-VISTA](#)  
[TreeQ-Vista](#)  
[PGA](#)

# VISTA Browser: Alignment Details



# 3. Predikce kódující oblasti na základě hledání signálů

## *(ab initio)*

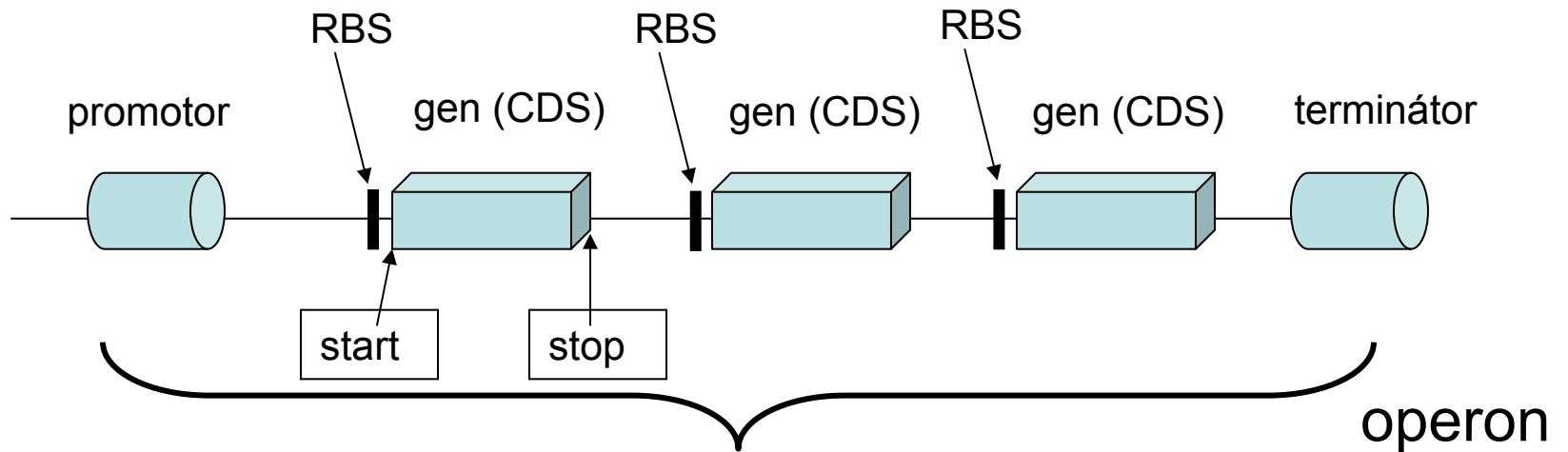
- Využívá pouze sekvenční data a výpočetní přístupy integrující analýzu sekvence a detekci signálů
- Prokaryota
  - Hledání otevřených čtecích rámců doplněné hledáním konzervativních signálů v transkripčních jednotkách
  - **ORF Finder (Open Reading Frame Finder)**  
<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>
- Eukaryota
  - Predikce promotorů
  - Predikce polyA-signálů
  - Predikce míst sestřihu a start/stop kodonů

# ***Klíčové signály pro odhalení genů***

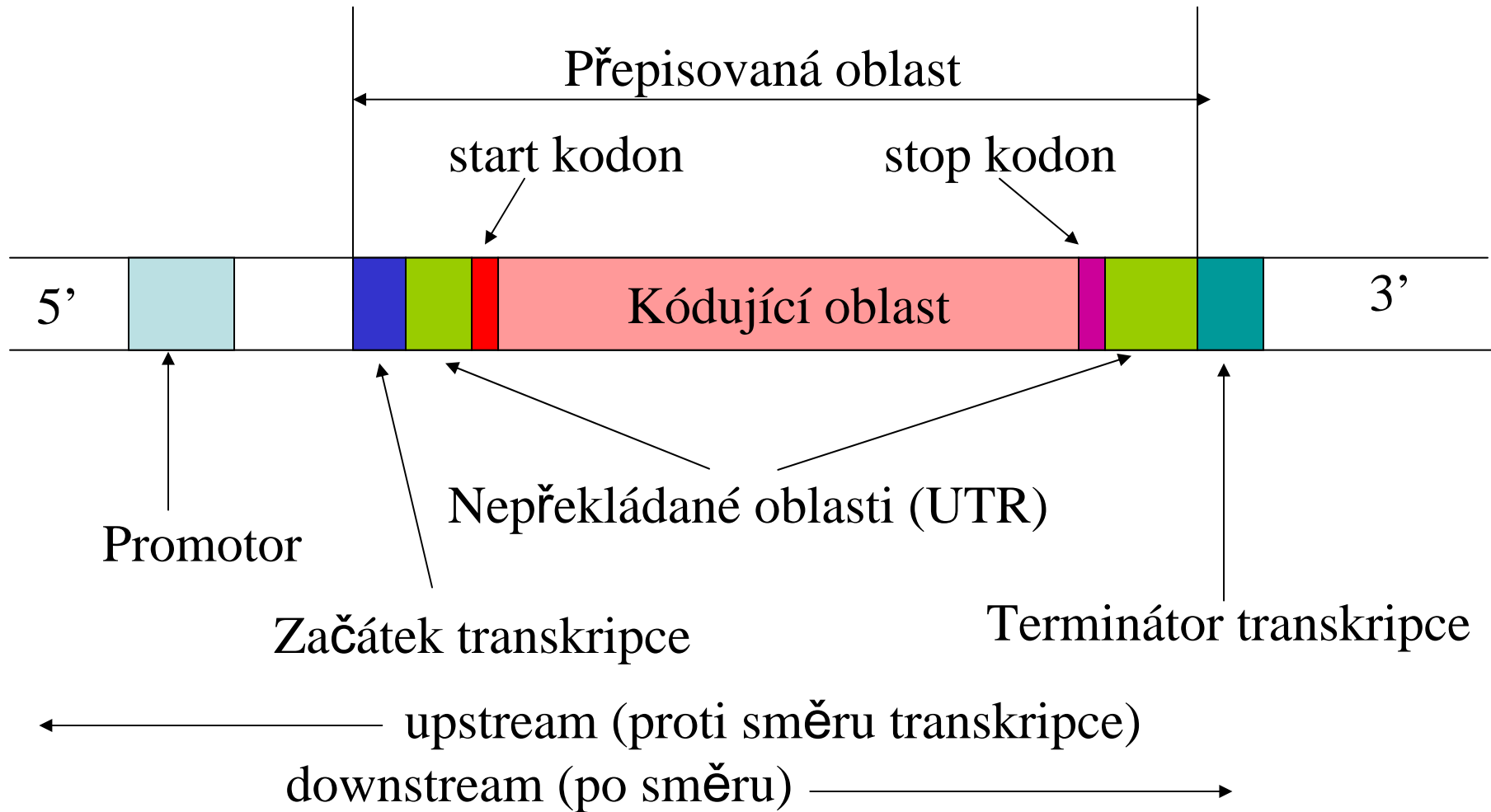
- iniciační a terminační kodony
- promotory
- vazebná místa pro ribozómy (RBS)
- místa sestřihu
- terminátory transkripce
- polyadenylační místa
- vazebná místa pro transkripční faktory



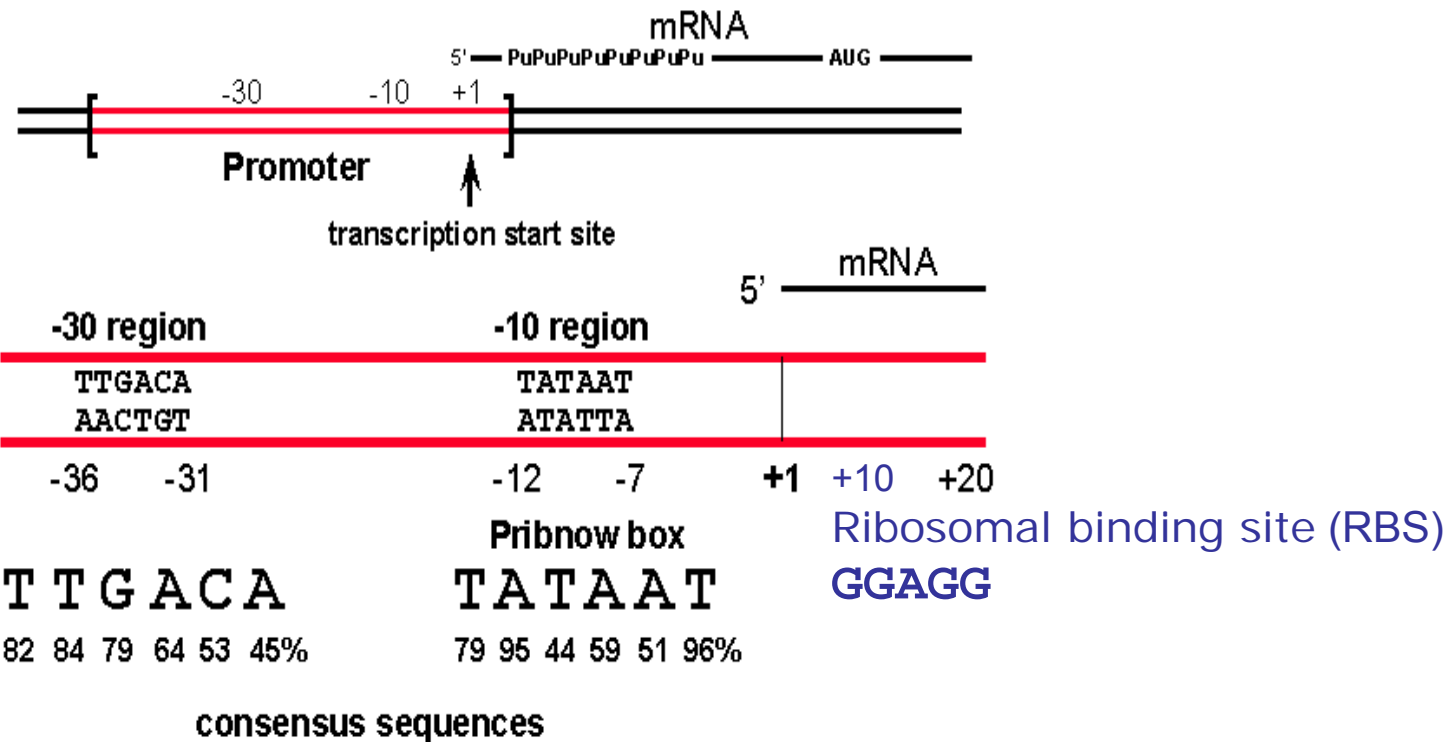
# Struktura prokaryotické transkripční jednotky



# Struktura prokaryotického genu



# Konzervativní struktury v promotoru prokaryot



# Signály v jednoduchém strukturním genu

*fem* gene

```
1 ATATGGTCAGTGCATATAAAAATTTGTTATCATTAGAGTAATTAAGGTCATTTAATAACTTTTGGGAATCA 70
71 ATTGGAGGTTCTCATATGTTATCTTTTAGTCAAATAGAAGTCATAGCTTAGAACAACTTTTAAAAGAAG 140
141 GATATTCACAAATGGCTGATTTAAATCTCTCCCTAGCGAACGAAGCTTTTCCGATAGAGTGTGAAGCATG 210
211 CGATTGCAACGAAACATATTTATCTTCTAATTCAACGAATGAATCATTAGACGAGGAGATGTTTATTTAG 280
281 CAGATTTATCACCAGTACAGGGATCTGAACAAGGGGGAGTCAGACCTGTAGTCATAATTCAAATGATAC 350
351 TGGTAATAAATATAGTCCTACAGTTATTGTTGCGGCAATAACTGGTAGGATTAATAAAGCGAAAATACCG 420
421 ACACATGTAGAGATTGAAAAGAAAAAGTATAAGTTGGATAAAGACTCAGTTATATTATTAGAACAAATTC 490
491 GTACACTTGATAAAAAACGATTGAAAGAAAAACTGACGTA CT TATCCGATGATAAAATGAAAGAAGTAGA 560
561 TAATGCACTAATGATTAGTTTAGGGCTGAATGCAGTAGCTCACCAGAAAAATTAGGCGGTCTATTATATGT 630
631 ATTTTTTCAGAGATAAATAAAATATTGATATAAAAGACAATAACTTTATAATAATTATAACTATTTCTAAA 700
701 TTCTGTACGAAGAATTTTCTTATAAACAAAGATTTTAGCAAATACCAGTTATGATATTCATATTTTTTAT 770
771 TATAAAAGGATGTCTTAAGTTTTTTTAGGCTTTAGGTATTCCATCCTAAAGTTTTTTTTTAGCTTAAAAGTA 840
841 TCATCTACAGCAAATTTGCAAACGACAAAATTTGATAAGTGCAATTAATAAATGTTAGTAAGTGAATCAT 910
911 AATTATCCTTGCTTAAGCATTGCTTTGTAAGGGAAGTGAGGAGGCAACTAATCG 965
```

rsbU gene

putative promotor

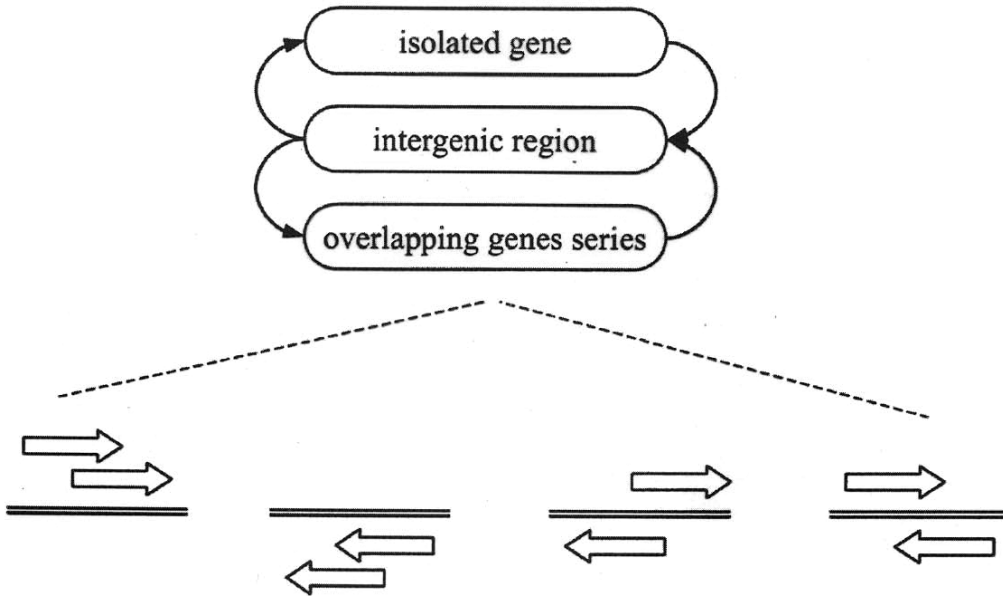
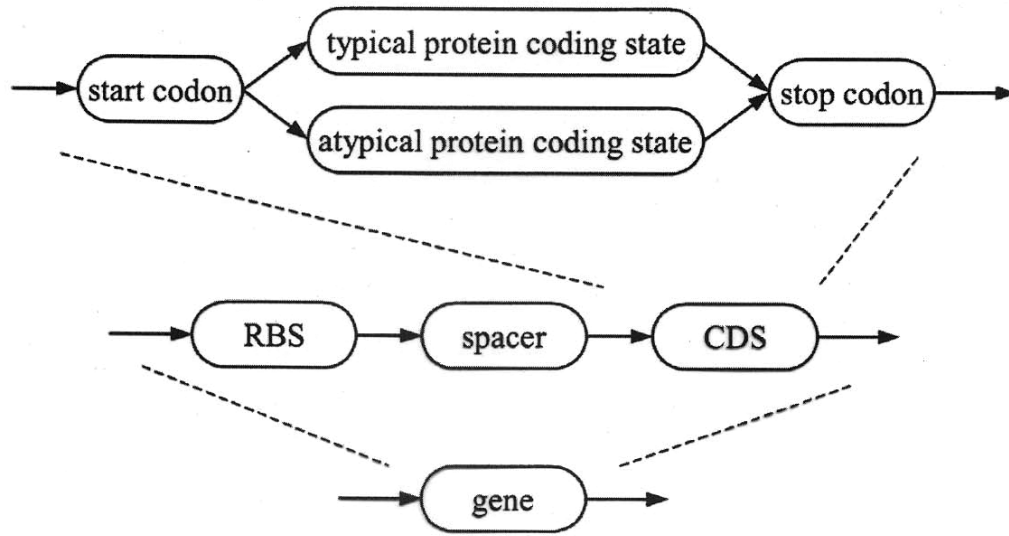
putative RBS

start

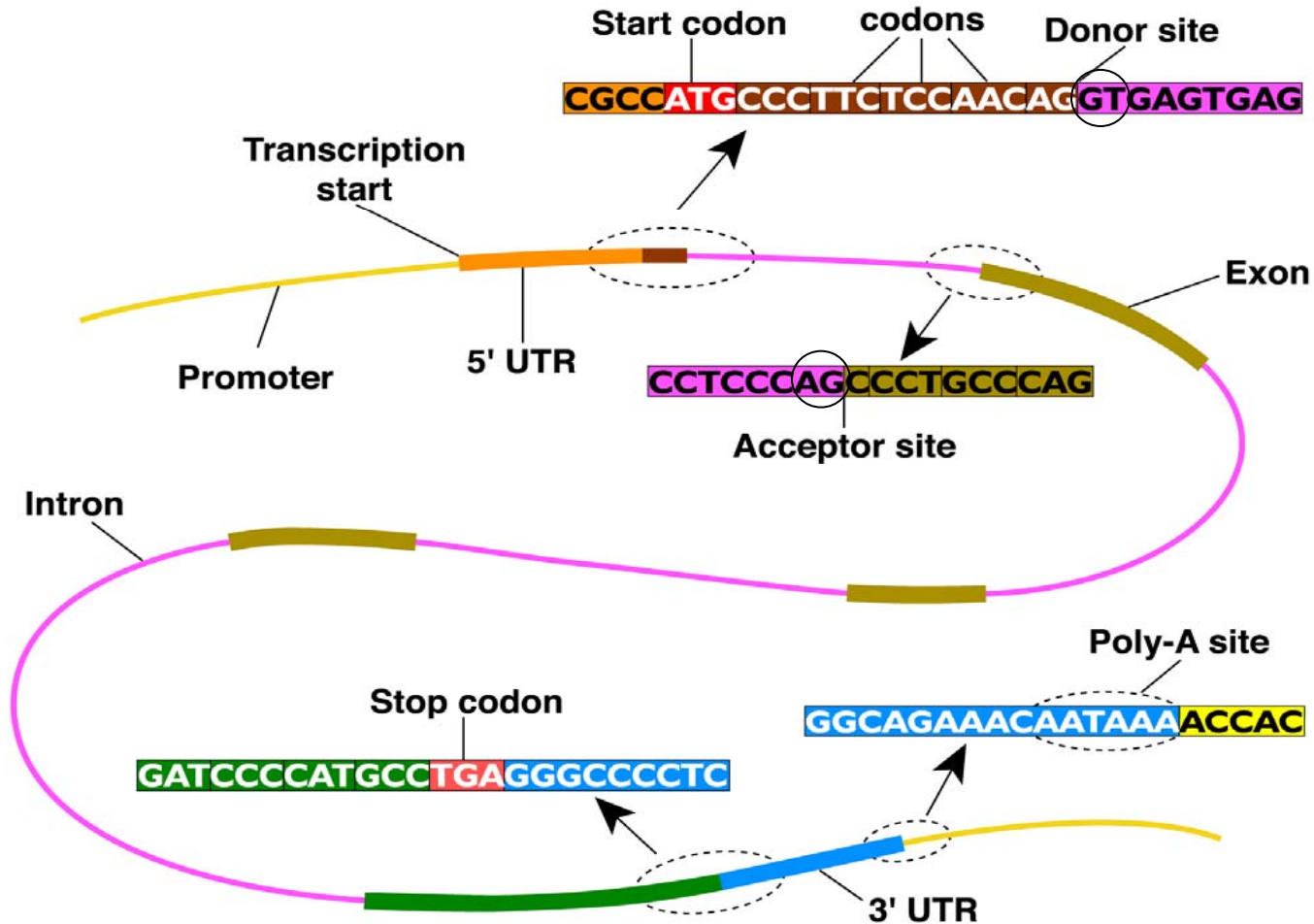
stop

terminator

# Model pro hledání jednoduchých genů



# Signály – senzory ve struktuře eukaryotického genu



# Metody pro vyhledávání signálů

- hledání konvenční sekvence spolu s možnostmi přípustných odchylek
- použití vážených matic
  - každá pozice vzoru signálu připouští shodu s jakýmkoli zbytkem
  - různé zbytky mají v každé pozici přiřazenou jinou významnost

# Příklad konsenzní sekvence signálu

- Získána výběrem nejčastěji se vyskytující báze v každé pozici mnohonásobného přiložení příslušné subsekvence našeho zájmu

TACGAT

TATAAT

TATAAT

GATACT

TATGAT

TATGTT

konsensus sequence

TATAAT

konsensus (IUPAC)

TATRNT

- Vede ke ztrátě informací a získání mnoha falešně pozitivních i negativních výsledků



# Příklad poziční vážené matice

- Vyjadřuje frekvenci každé báze v každé pozici příslušné sekvence

TACGAT		1	2	3	4	5	6
TATAAT	A	0	6	0	3	4	0
TATAAT	C	0	0	1	0	1	0
GATACT	G	1	0	0	3	0	0
TATGAT	T	5	0	5	0	1	6
TATGTT							

- Skóre každého předpokládaného místa je vyjádřeno součtem hodnot z matice (převáděno na pravděpodobnosti)
- Nevýhody:
  - Je vyžadována hraniční hodnota
  - Předpokládá nezávislost sousedících bází

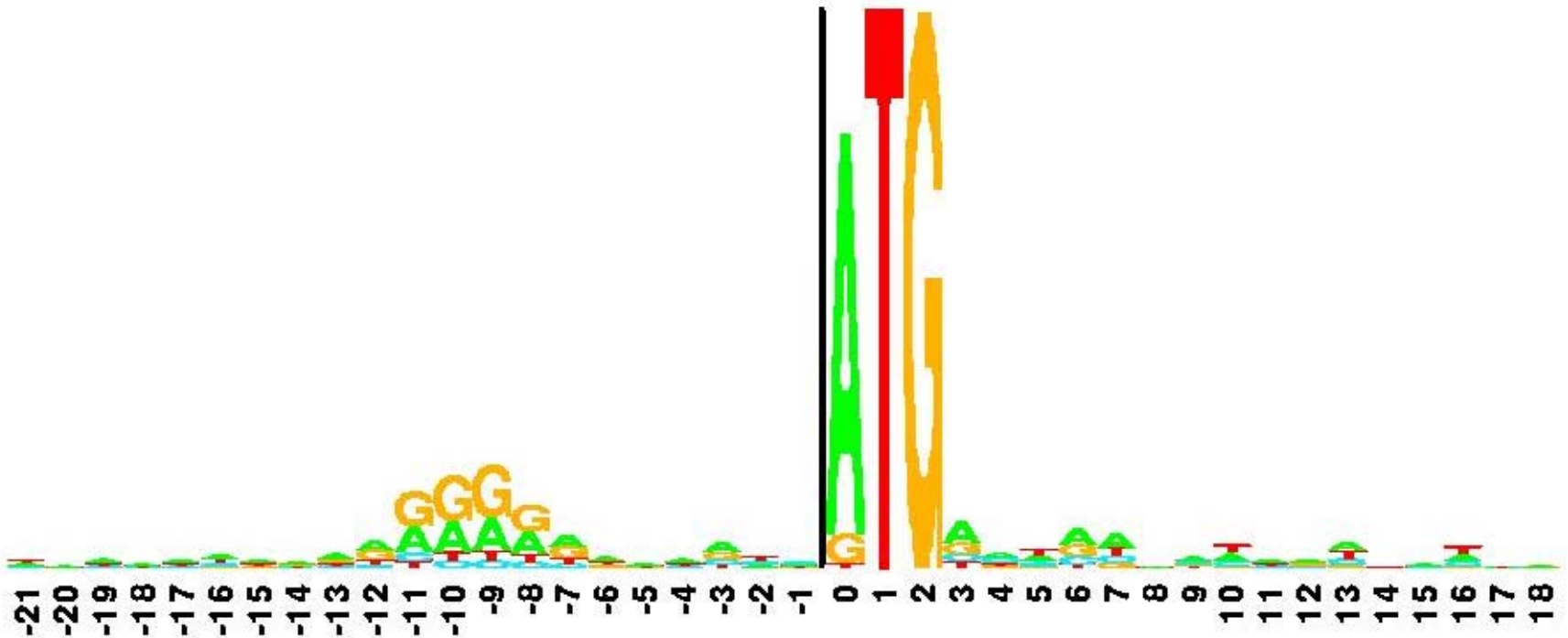
A



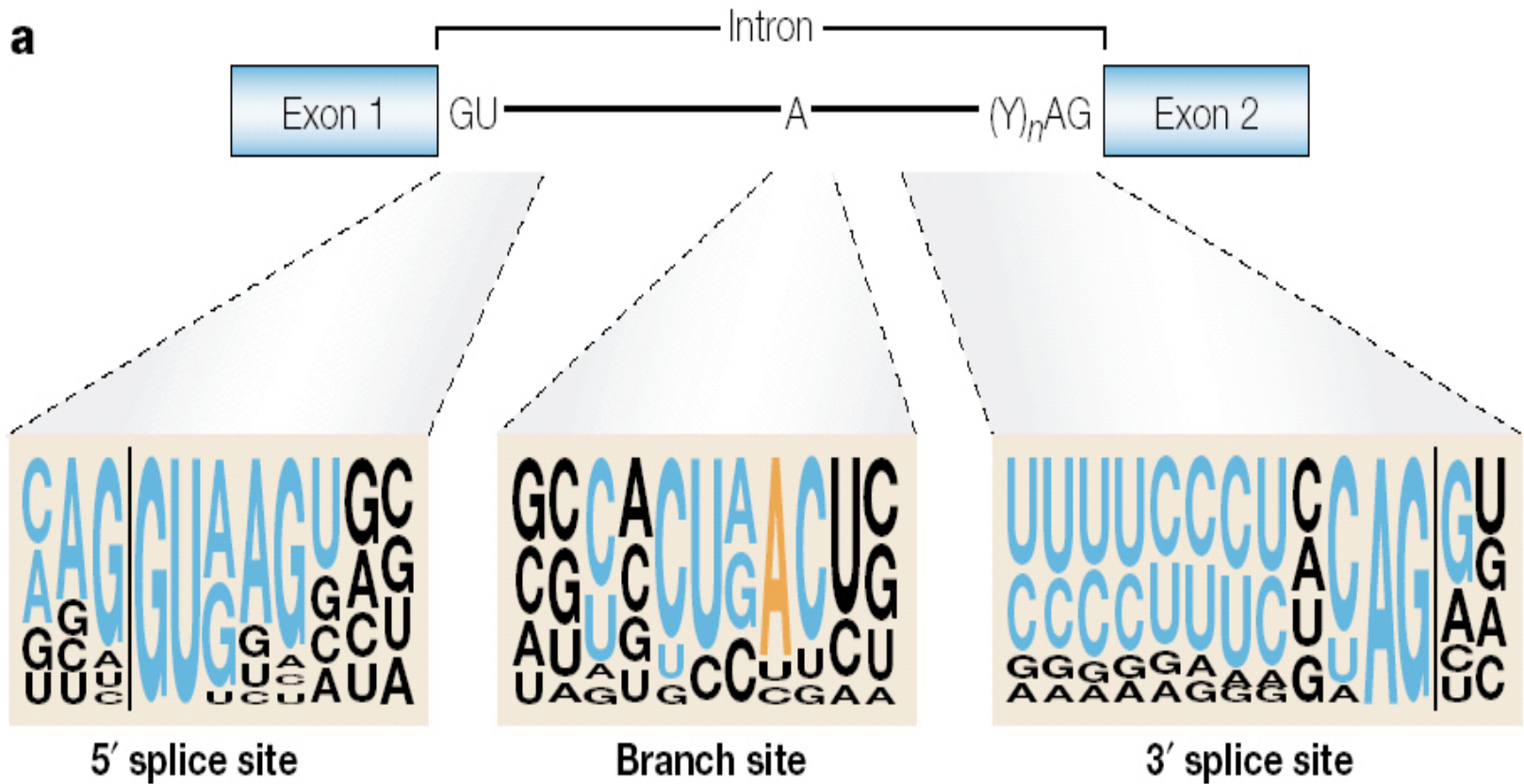
Příklad signálu

RBS (vazebné  
místo pro ribozóm)

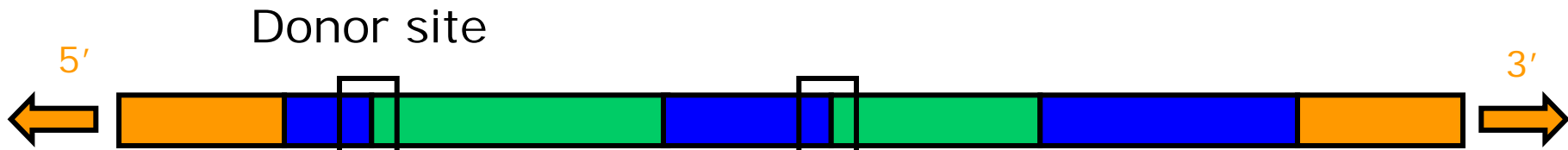
# Vazebné místo pro ribozóm (RBS) a iniciační kodon ATG u *E. coli*



# Predikce míst sestřihu

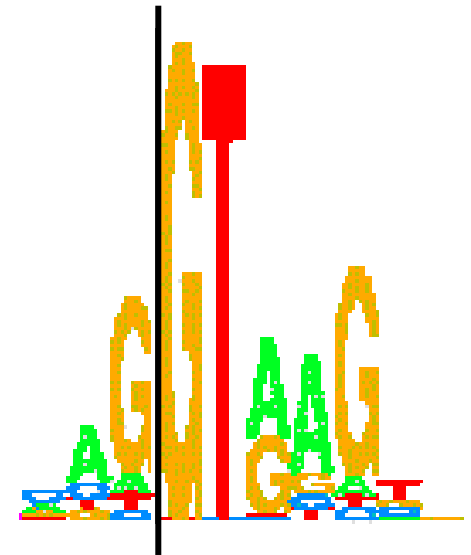


# Pozičně vážená matice pro odvození donorového místa sestřihu

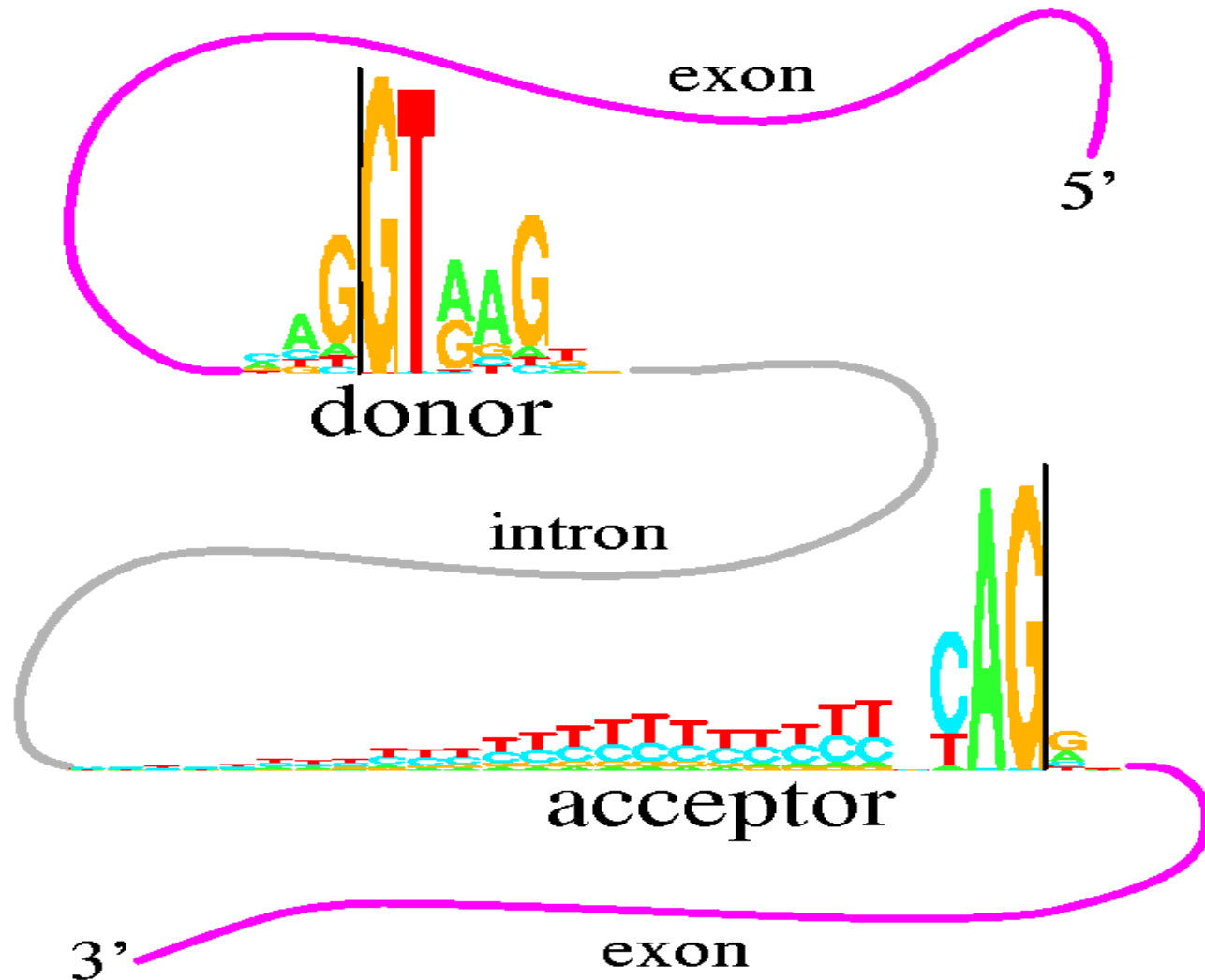


Position

%	-8	...	-2	-1	0	1	2	...	17
A	26	...	60	9	0	1	54	...	21
C	26	...	15	5	0	1	2	...	27
G	25	...	12	78	99	0	41	...	27
T	23	...	13	8	1	98	3	...	25



# Příklad signálů: místa sestřihu (myš)



# Statistická analýza sekvence predikovaného genu

- Důležité je posouzení charakteru sekvence
  - délka genu
  - frekvence využití kodonů
  - obsah GC (indikace horizontálního přenosu)
  - GC skew a AT skew
    - $GC\ skew = (G - C)/(G + C)$
    - $AT\ skew = (A - T)/(A + T)$
  - statistické modely modely frekvencí nukleotidů (využití hexamerů)
  - periodicitu nukleotidů

# Testování exonů – využití kodonů

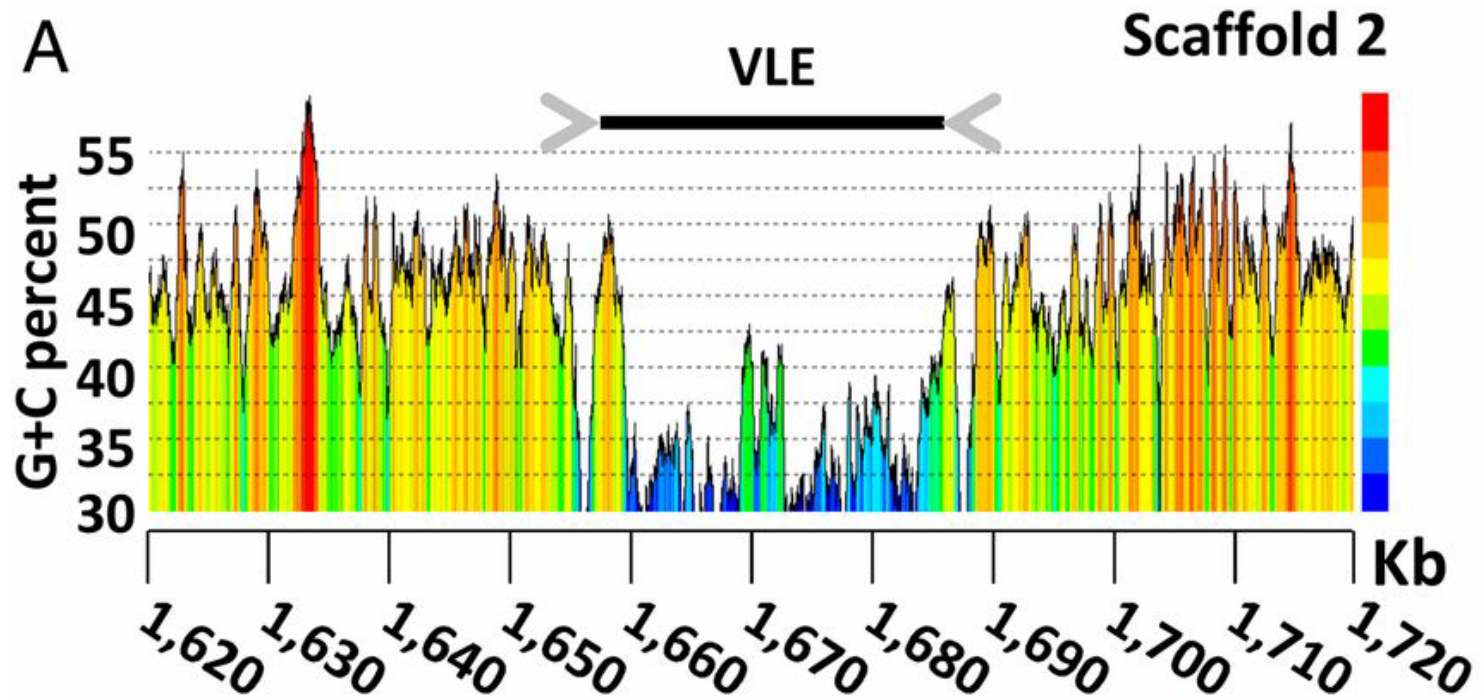
<u>AA</u>	<u>codon</u>	<u>/1000</u>	<u>frac</u>
Ser	TCG	<b>4.31</b>	0.05
Ser	TCA	11.44	0.14
Ser	TCT	15.70	0.19
Ser	TCC	17.92	0.22
Ser	AGT	12.25	0.15
Ser	AGC	19.54	0.24
Pro	CCG	<b>6.33</b>	0.11
Pro	CCA	17.10	0.28
Pro	CCT	18.31	0.30
Pro	CCC	18.42	0.31

<u>AA</u>	<u>codon</u>	<u>/1000</u>	<u>frac</u>
Leu	CTG	<b>39.95</b>	0.40
Leu	CTA	<b>7.89</b>	0.08
Leu	CTT	12.97	0.13
Leu	CTC	20.04	0.20
Ala	GCG	<b>6.72</b>	0.10
Ala	GCA	15.80	0.23
Ala	GCT	20.12	0.29
Ala	GCC	<b>26.51</b>	0.38
Gln	CAG	<b>34.18</b>	0.75
Gln	CAA	11.51	0.25

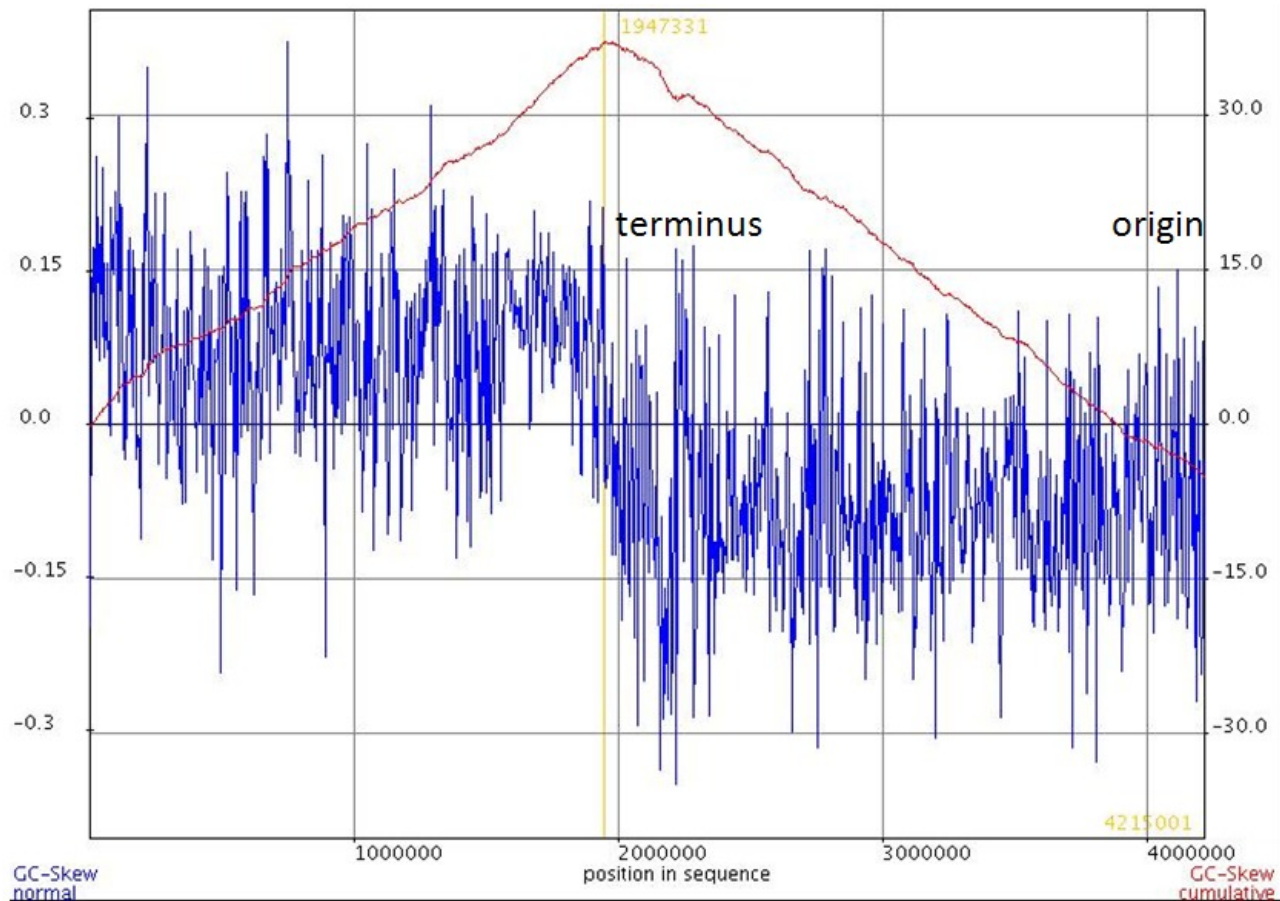
Codon usage database: <http://www.kazusa.or.jp/codon/>



# Obsah G+C – příklad využití pro identifikaci mobilního elementu

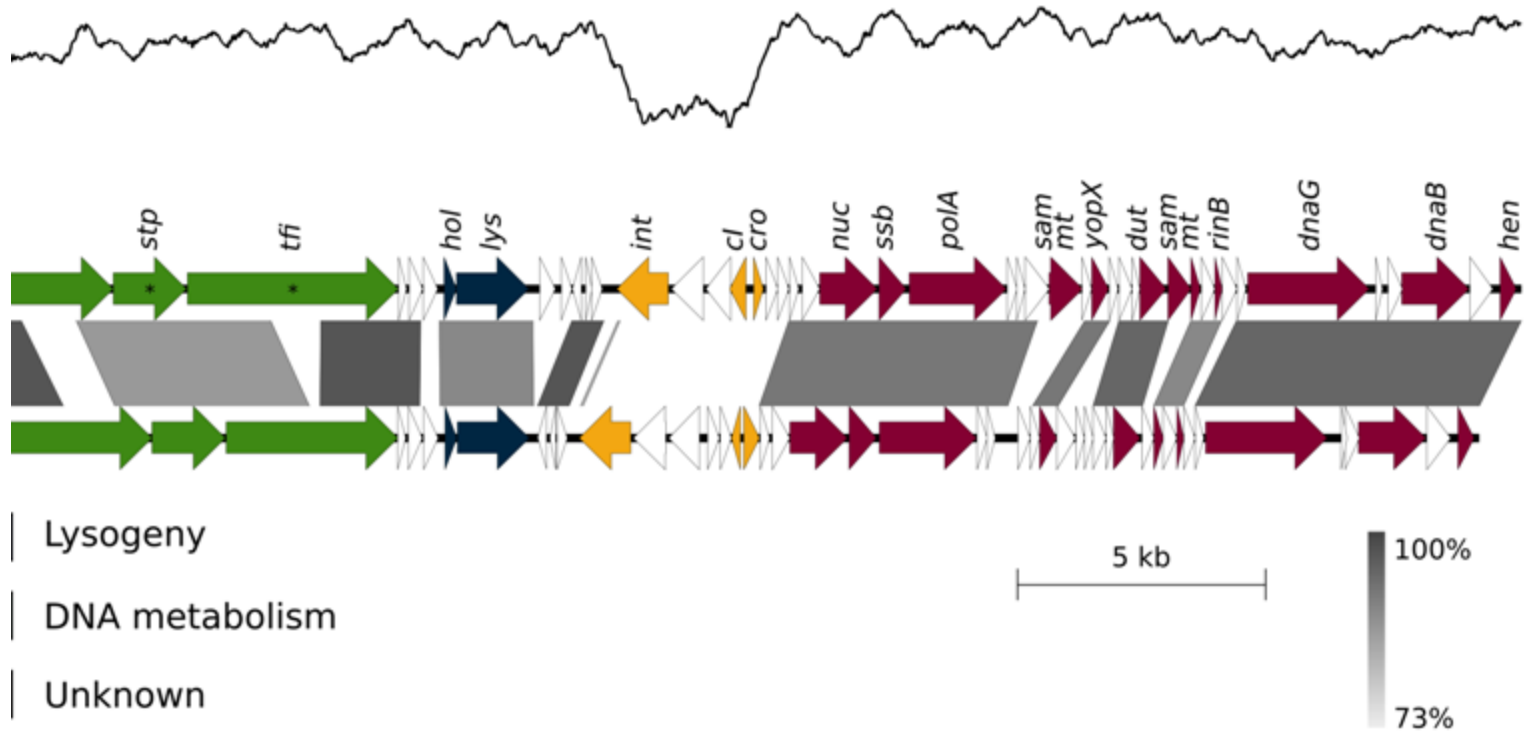


# GC skew – příklad využití pro identifikaci počátku replikace



# AT skew – příklad využití pro identifikaci kódujícího vlákna DNA

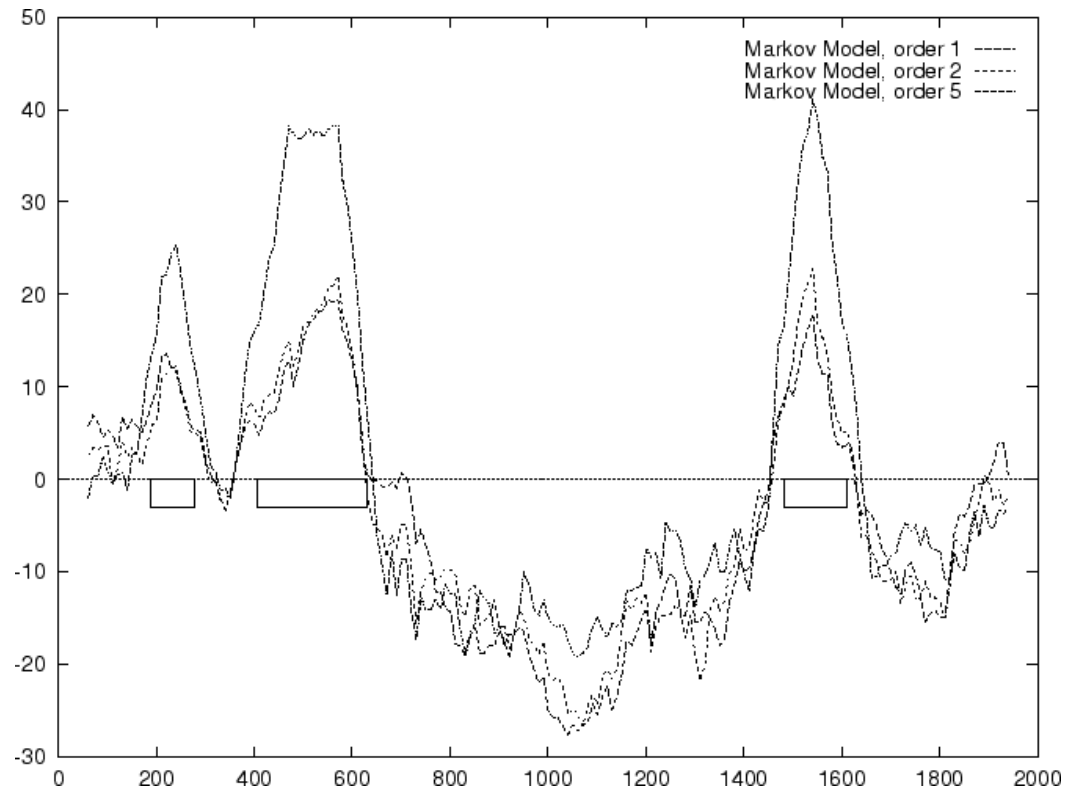
## AT skew



# Frekvence hexamerů - rozlišení mezi kódujícími a nekódujícími oblastmi

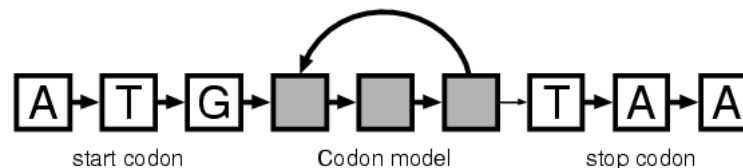
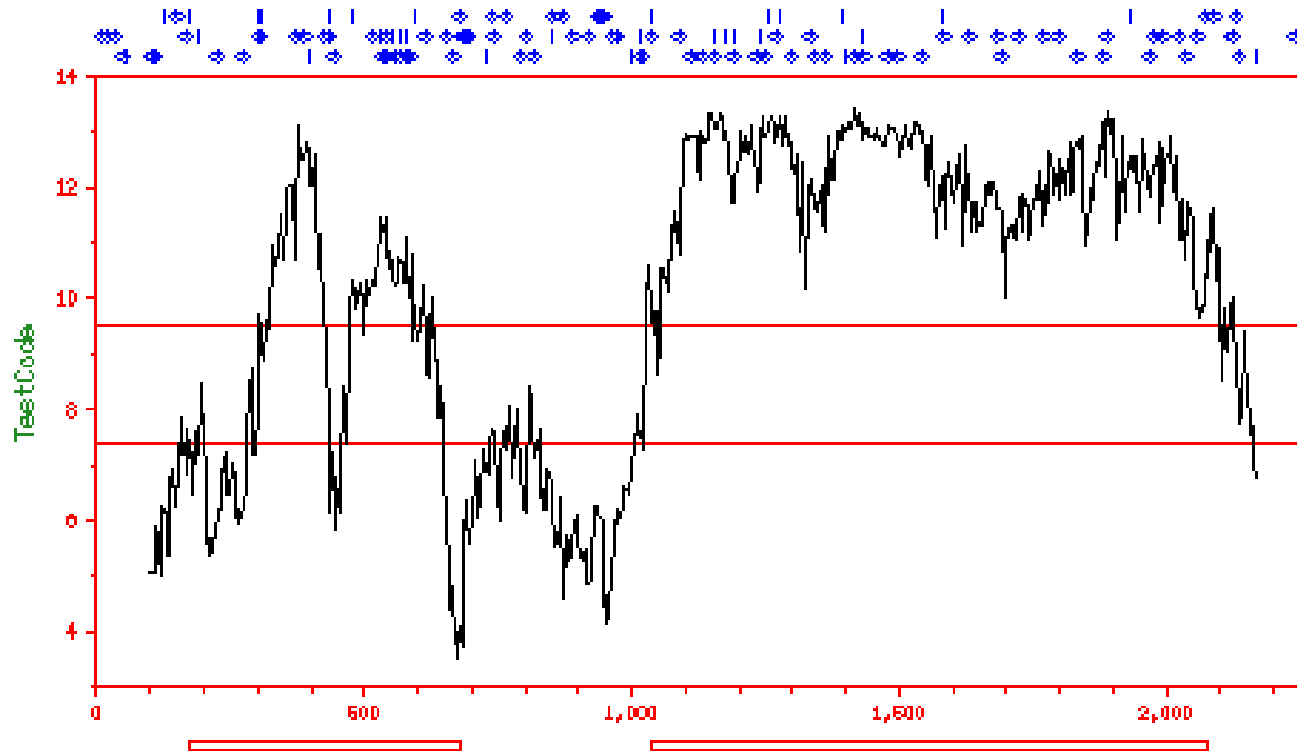
- Rozdíly v distribuci jiných oligonukleotidů než kodonů (tri-nukleotidů)

- odráží závislosti mezi sousedními aminokyselinami v proteinech



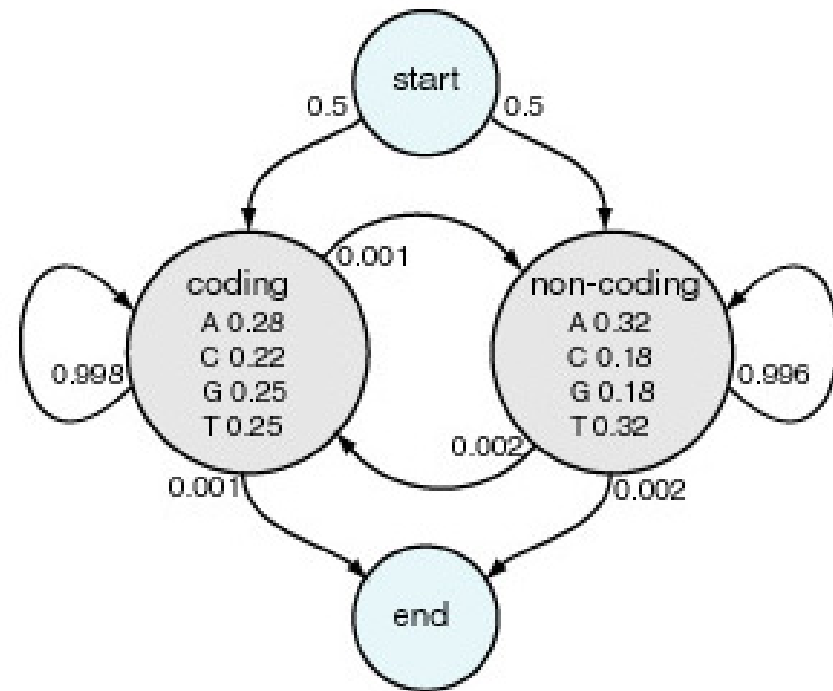
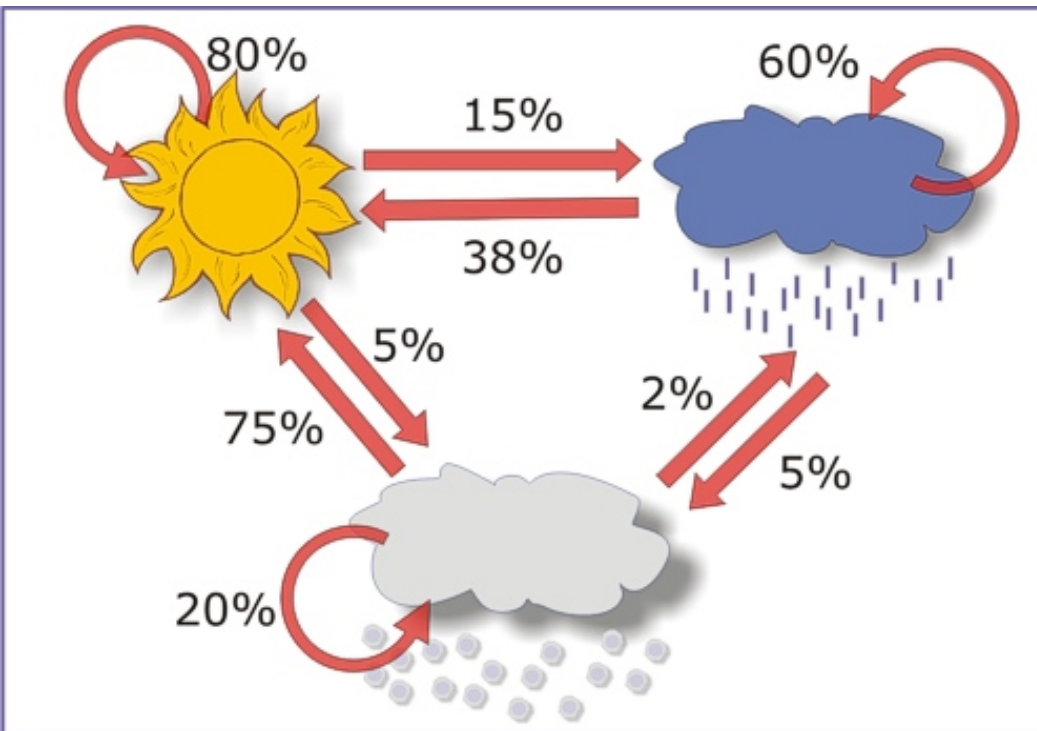
# Testování ORF – frekvence nukleotidů

TESTCODE of: gb\_ba:Ecoomp1 ck: 778, 1 to: 2270  
Window: 200 bp October 6, 1998 10:54

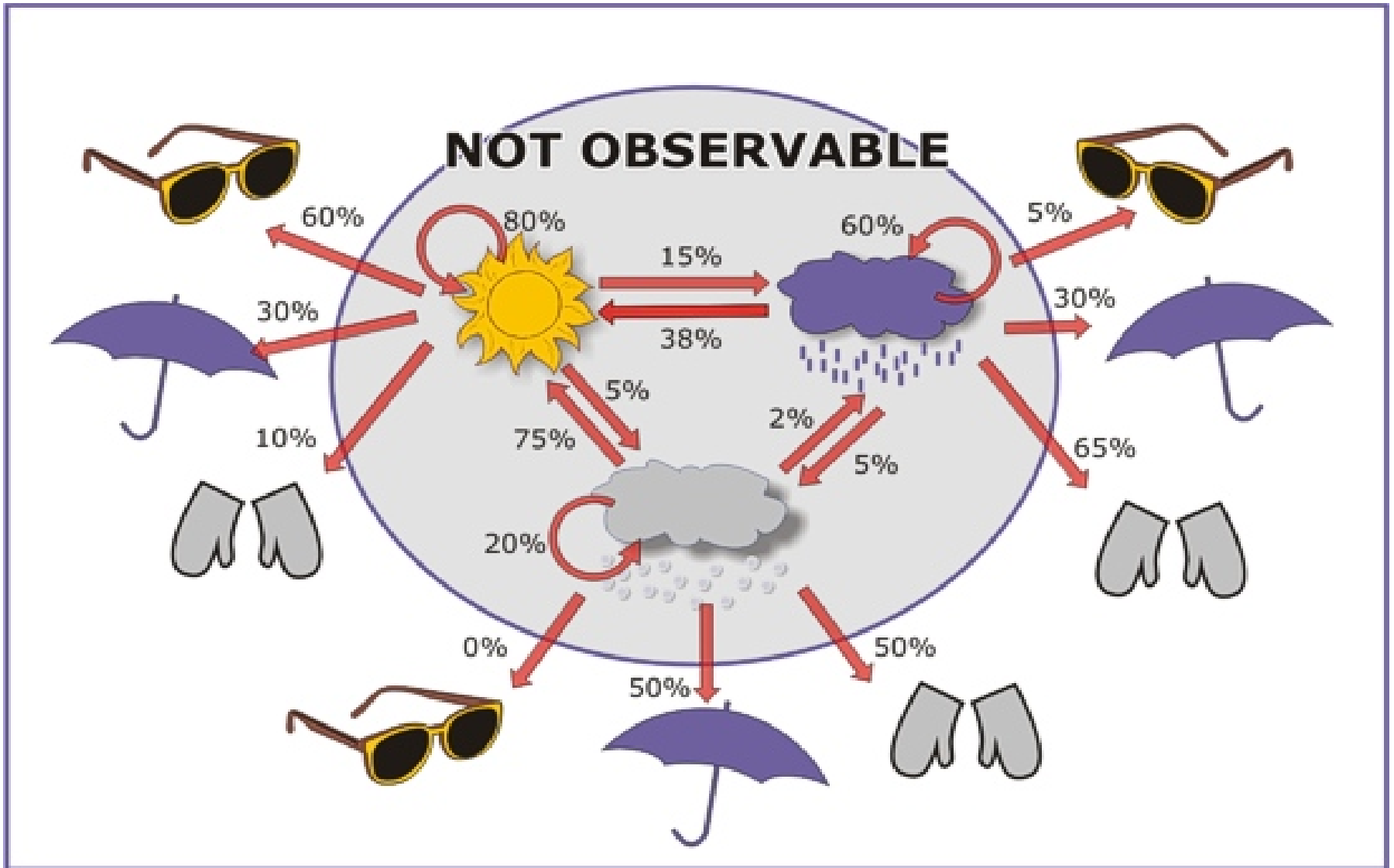


# Markovovy modely

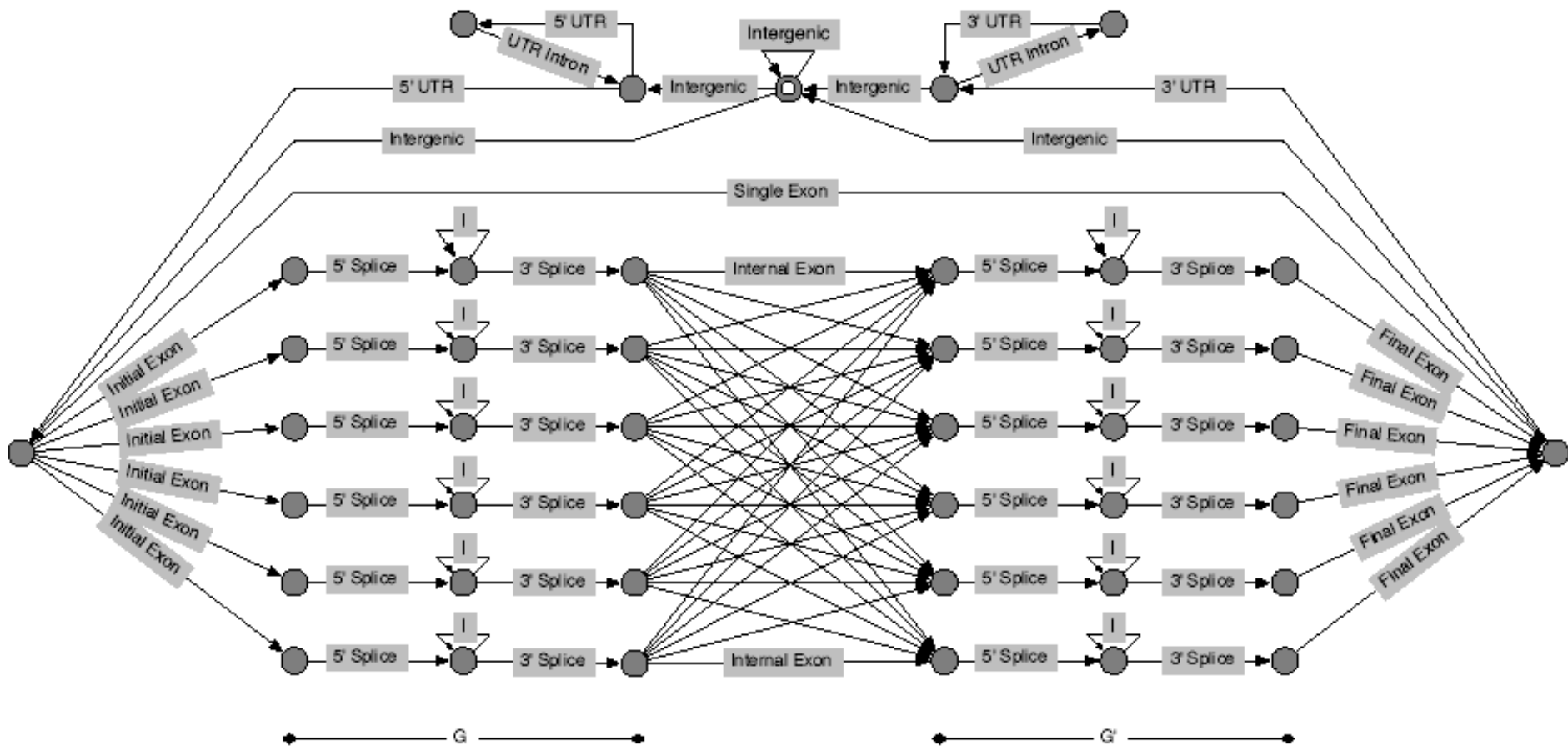
- Vyjadřují pravděpodobnost sekvenčních událostí
- Nejčastěji používané statistické modely pro hledání genů



# Hidden Markov Models (HMM)



# Příklad komplexního algoritmu se skrytými Markovovými modely (HMM)





# Populární programy pro predikci genů

- Programy využívající explicitní pravidla
  - GeneFinder
- Programy založené na „Hidden Markov Models“
  - GeneMark
  - Glimmer
  - GenScan
  - TwinScan
- Programy využívající neuronové sítě
  - Grail,
  - GrailEXP

# GeneMark

<http://opal.biology.gatech.edu/GeneMark/>

## GeneMark

A family of gene prediction programs developed by Mark Borodovsky's [Bioinformatics Group](#) at the [Georgia Institute of Technology](#), Atlanta, Georgia, USA.

### What's New:

Gene identification in novel eukaryotic genomes by self-training algorithm:  
[GeneMark.hmm-ES](#)

Supported  
by NIH



## Gene Prediction in Bacteria, Archaea and Metagenomes



For bacterial and archaeal gene prediction you can use the parallel combination of [GeneMark-P](#) and [GeneMark.hmm-P](#). For a novel genome you can use either the [Heuristic models](#) option (if the sequence is shorter than 200 kb) or the self-training program [GeneMarkS](#) (aka [GeneMark.hmm-PS](#)).

## 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 run [GeneMark.hmm-ES](#), the self-training program (just 10MB sequence is needed for training).

## Gene Prediction in Viruses



For gene prediction in novel viruses and phages you can use [GeneMark.hmm](#). Viral genome annotations are accessible via [VIOLIN](#) database.

## Gene Prediction in EST and cDNA



To analyze ESTs and cDNAs you can use [GeneMark-E](#).

Powered by IBM



## Borodovsky Group

### Gene Prediction Programs

- [GeneMark](#)
- [GeneMark.hmm](#)
- [GeneMarkS](#)
- [Heuristic models](#)
- [Frame-by-Frame](#)

### Information

- [Background](#)
- [References](#)
- [In GenBank](#)
- [FAQ](#)
- [Contact](#)

### Databases of predicted genes

- Prokaryotes [Closed, Updating](#)
- [Viruses/Phages \(VIOLIN\)](#)

### Models for Gene Prediction

- [Download](#)

# Glimmer

[http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer\\_3.cgi](http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi)



## Microbial Genomes



[HOME](#) | [SEARCH](#) | [SITE MAP](#) | [Genome Project](#) | [Genome](#) | [Prokaryotic Projects](#) | [Collaborators](#) | [gMap](#) | [ProtMap](#) | [TaxPlot](#) | [BLAST](#) | [FTP](#) | [Contact us](#)

### Microbial Genome Annotation Tools

GLIMMER is a system for finding genes in microbial DNA, especially the genomes of bacteria and archaea. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions.

- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER, *Nucleic Acids Research* 27:23 (1999), 4636-4641.
- Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models, *Nucleic Acids Research* 26:2 (1998), 544-548.

Download [GLIMMER](#) from the Center for Bioinformatics and Computational Biology.

<b>Genomes</b>
<a href="#">Genome Projects</a>
<a href="#">Prokaryotic Projects</a>
<a href="#">Microbial Genomes</a>
<a href="#">Home</a>
<a href="#">Complete Genomes</a>
<a href="#">Draft Assemblies</a>
<a href="#">Registered</a>
<a href="#">Entrez Genome</a>
<b>Submit a Genome</b>
<a href="#">Sequin</a>
<a href="#">Submission Guide</a>
<a href="#">Register a Project</a>
<a href="#">Submit a Genome</a>
<a href="#">Submit Traces</a>
<b>Tools</b>
<b>Resources</b>
<a href="#">Sequencing Centers</a>
<a href="#">Collaborators</a>
<b>Statistics</b>

Upload your sequence from file:

Or copy/paste your sequence FASTA here:



## RAST (Rapid Annotation using Subsystem Technology)

- Anotace na základě vlastní pipeline
- Využívá integrovaný přístup včetně NCBI databáze (BLAST)
- Klasifikace genů do subsystémů a identifikace metabolických drah podle KEGG: Kyoto Encyclopedia of Genes and Genomes (<https://www.genome.jp/kegg/>)
- Příklad parametrů anotovaného genomu:

### Organism Overview for *Massilia sp. CCM 8692* (6666666.478097)

Genome	Massilia sp. CCM 8692
Domain	Bacteria
Taxonomy	Bacteria; Massilia sp. CCM 8692
Neighbors	<a href="#">View closest neighbors</a>
Size	7,576,397
GC Content	63.8
N50	191842
L50	11
Number of Contigs (with PEGs)	141
Number of Subsystems	475
Number of Coding Sequences	6982
Number of RNAs	104

For each genome we offer a wide set of information to browse, compare and download.

[Browse](#) [Compare](#) [Download](#) [Annotate](#)

Browse through the features of [Massilia sp. CCM 8692](#) both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click [here](#) to get to the Genome Browser

- RAST rozdělí anotované geny do jednotlivých funkčních kategorií, ty zahrnují další podkategorie – např. geny zapojené do jednotlivých metabolických drah

## Subsystem Information

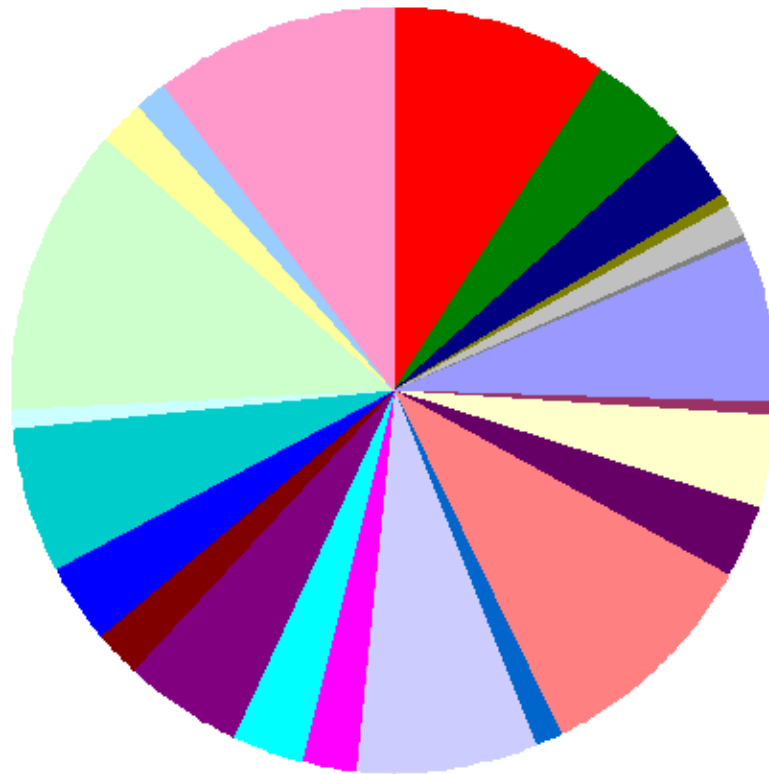
Subsystem Statistics

Features in Subsystems

### Subsystem Coverage



### Subsystem Category Distribution

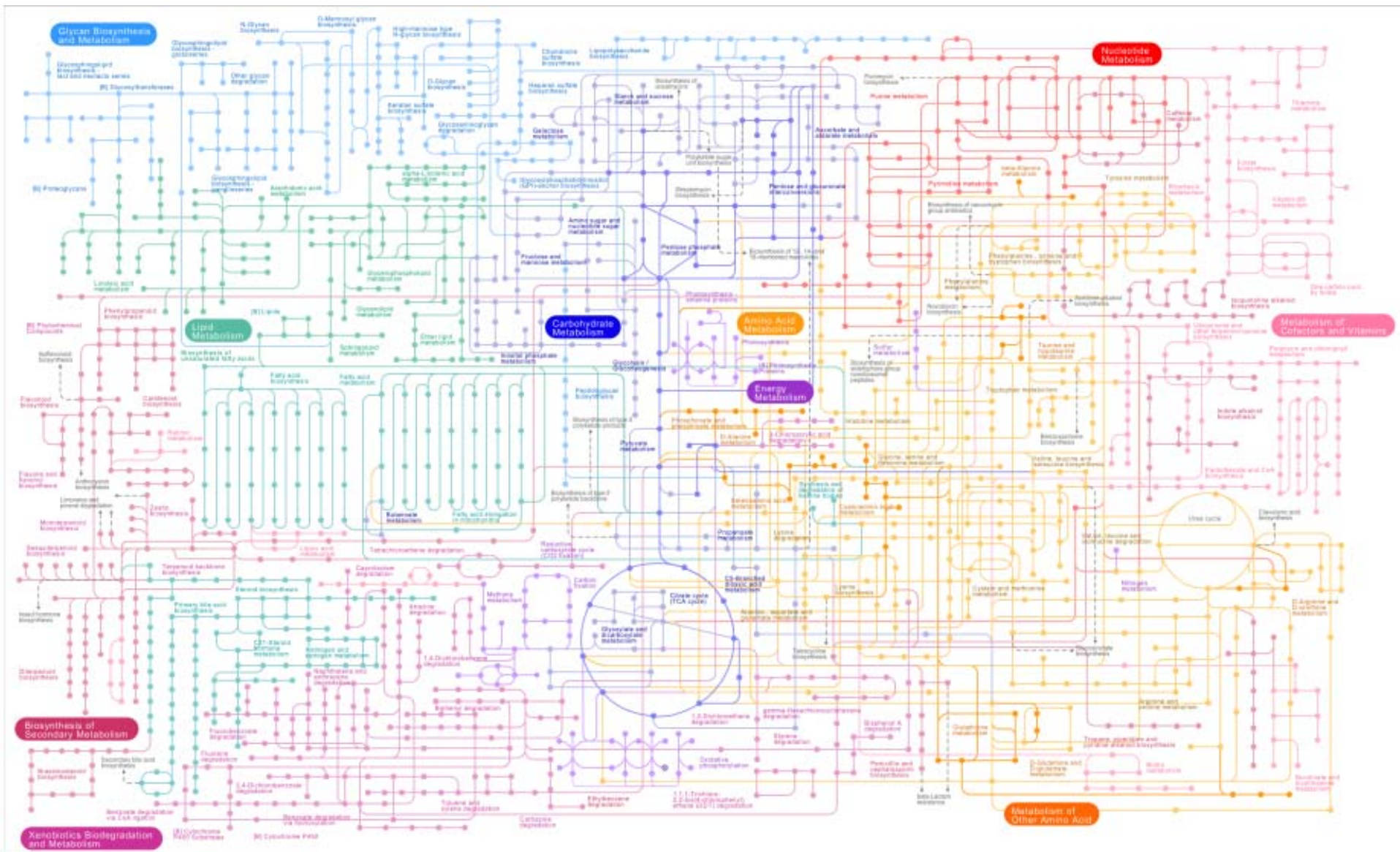


### Subsystem Feature Counts

- ☒ Cofactors, Vitamins, Prosthetic Groups, Pigments (327)
- ☒ Cell Wall and Capsule (145)
- ☒ Virulence, Disease and Defense (116)
- ☒ Potassium metabolism (19)
- ☒ Photosynthesis (0)
- ☒ Miscellaneous (45)
- ☒ Phages, Prophages, Transposable elements, Plasmids (13)
- ☒ Membrane Transport (248)
- ☒ Iron acquisition and metabolism (16)
- ☒ RNA Metabolism (135)
- ☒ Nucleosides and Nucleotides (108)
- ☒ Protein Metabolism (343)
- ☒ Cell Division and Cell Cycle (39)
- ☒ Motility and Chemotaxis (280)
- ☒ Regulation and Cell signaling (71)
- ☒ Secondary Metabolism (5)
- ☒ DNA Metabolism (116)
- ☒ Fatty Acids, Lipids, and Isoprenoids (178)
- ☒ Nitrogen Metabolism (63)
- ☒ Dormancy and Sporulation (4)
- ☒ Respiration (118)
- ☒ Stress Response (222)
- ☒ Metabolism of Aromatic Compounds (24)
- ☒ Amino Acids and Derivatives (428)
- ☒ Sulfur Metabolism (69)
- ☒ Phosphorus Metabolism (57)
- ☒ Carbohydrates (346)

KEGG mapa všech metabolických drah nalezených u daného organismu dle automatické anotace RASTem.

- možnost sledovat jednotlivé metabolické dráhy



# Přehledové tabulky proteinů/genů zapojených do metabolické dráhy – identifikace na základě podobnosti v BLASTu

## Subsystem: Nitrosative stress

*This subsystem's description is:*

For more information, please check out the description and the additional notes tabs, below

Diagram		Functional Roles		Subsystem Spreadsheet		Additional Notes	
Group Alias	Abbrev.	Functional Role	Reactions	Scenario Reactions	GO	Literature	
all ▼							
	HMP	Flavoheмоprotein	-	-	-	none	
	HCP	Hydroxylamine reductase (EC 1.7.-.-)	-	-	<a href="#">GO:0005737</a> , <a href="#">GO:0006118</a> , <a href="#">GO:0016491</a>	none	
*HCPadd	HCR	NADH oxidoreductase hcr (EC 1.-.-.-)	-	-	<a href="#">GO:0005506</a> , <a href="#">GO:0006118</a> , <a href="#">GO:0009055</a> , <a href="#">GO:0016491</a>	none	
*HCPadd	FrdX	Ferredoxin 3 fused to uncharacterized domain	-	-	-	none	
	NorV	Anaerobic nitric oxide reductase flavorubredoxin	-	-	<a href="#">GO:0006118</a> , <a href="#">GO:0009055</a> , <a href="#">GO:0010181</a> , <a href="#">GO:0016491</a> , <a href="#">GO:0046872</a>	none	
	NorW	Nitric oxide reductase FIRd-NAD(+) reductase (EC 1.18.1.-)	-	-	<a href="#">GO:0006118</a> , <a href="#">GO:0015036</a>	none	
	qNor	Nitric-oxide reductase (EC 1.7.99.7), quinol-dependent	-	-	<a href="#">GO:0016966</a>	none	
	DnrN	Nitric oxide-dependent regulator DnrN or NorA	-	-	-	none	
	NnrS	NnrS protein involved in response to NO	-	-	-	none	
	NsrR	Nitrite-sensitive transcriptional repressor NsrR	-	-	<a href="#">GO:0005554</a>	none	
	NorR	Anaerobic nitric oxide reductase transcription regulator NorR	-	-	-	none	
	HcpR	Hcp transcriptional regulator HcpR (Crp/Fnr family)	-	-	-	none	



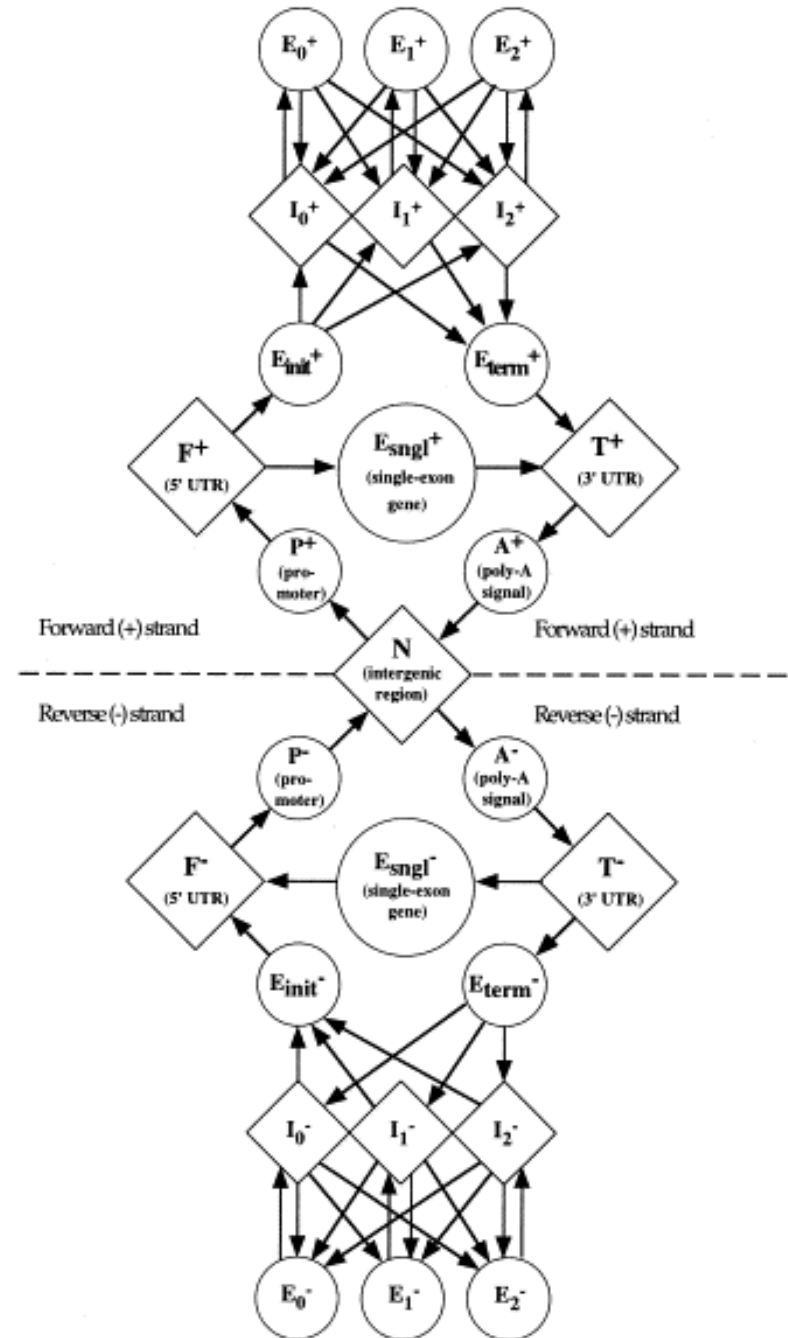


# Vyhledávače eukaryotických genů

- **Genie** používá informace ze známých genů a odhaduje, které oblasti genomu pravděpodobně obsahují nové geny
- **Fgenes** je vhodný pro hledání exonů a stanovení struktury genů
- **Genscan** přený vyhledávač využívá integrované přístupy
- **Veil** komplexní vyhledávač exonů a intronů a míst sestřihu, HMM modely

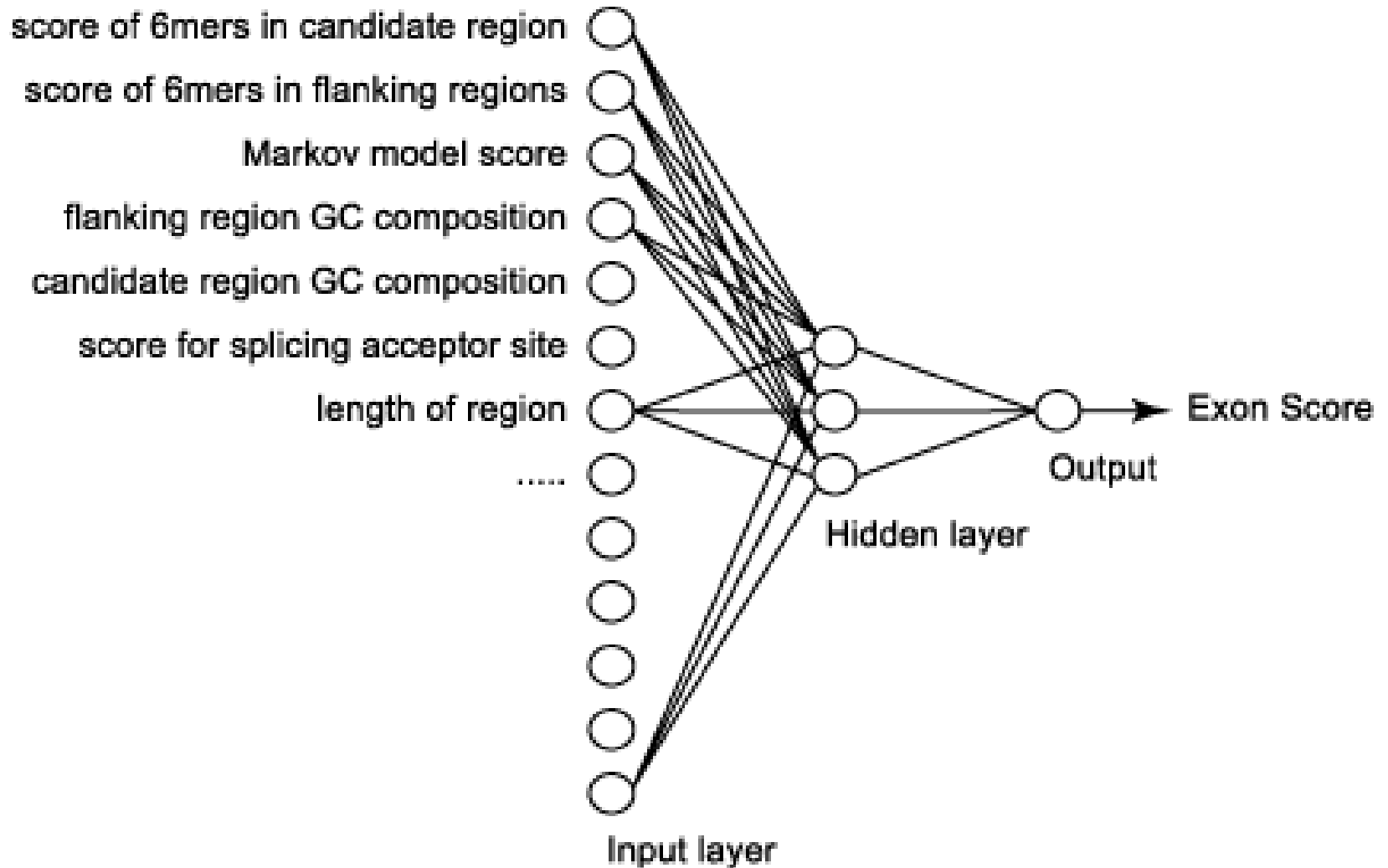
# Genscan

- přesný vyhledávač eukaryotických genů - využívá integrované přístupy



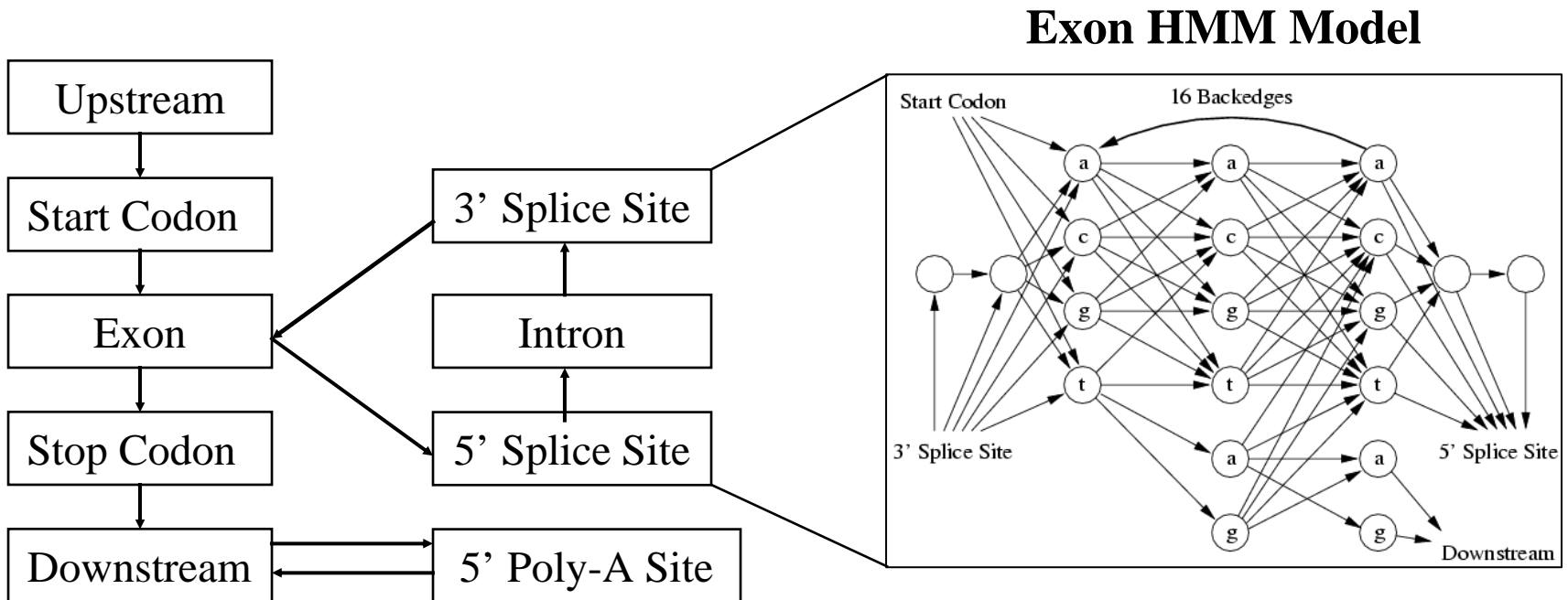
# Predikce eukaryotických genů

## GRAIL II: využívá neuronové sítě



# Veil komplexní vyhledávač exonů a intronů a míst sestřihu, HMM modely

- Obsahuje 9 skrytých modulů, z nichž každý je komplexní Markovův model



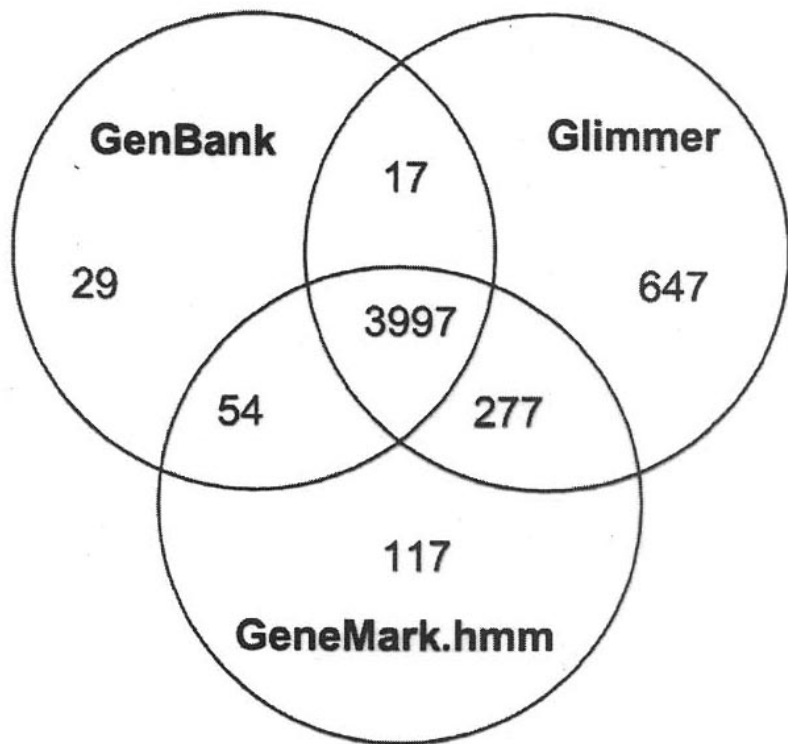
# Nomenklatura používaná při anotacích genomů

- **Known Gene** – Predikovaný gen shodující se v celé délce se známým experimentálně dokázaným genem.
- **Putative Gene** – Predikovaný gen obsahující region homologický s konzervovaným regionem známého genu. *Also referred to as “like” or “similar to”.*
- **Unknown Gene** – Predikovaný gen vykazující shodu s genem nebo EST, jejichž funkci neznáme.
- **Hypothetical Gene** – Predikovaný gen nevykazující významnou podobnost k žádnému známému genu nebo EST.

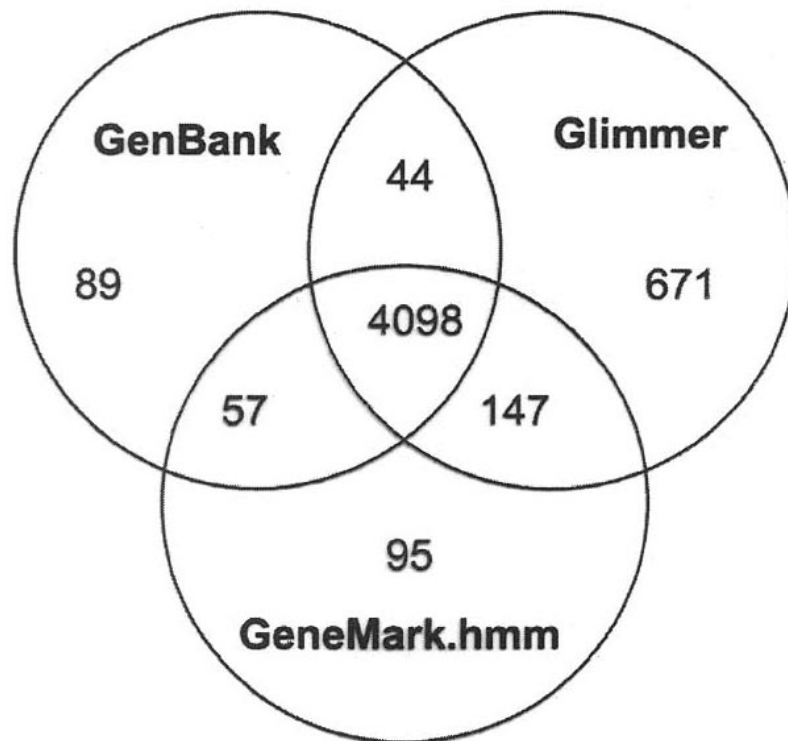
# Evaluace vyhledávačů genů

- Citlivost versus specificita
- Musí být optimalizovány / „trénovány“ pro specifický organizmus
  
- Citlivost
  - Kolik genů bylo nalezeno?
- Specificita
  - Kolik predikovaných genů představuje skutečné geny?

# Srovnání různých přístupů pro vyhledávání genů



*B. subtilis*



*E. coli*

# Užitečné nástroje

- Vyhledávače ORF
  - NCBI: <http://www.ncbi.nih.gov/gorf/gorf.html>
- Predikce promotoru
  - CSHL: <http://rulai.cshl.org/software/index1.htm>
  - BDGP: [fruitfly.org/seq\\_tools/promoter.html](http://fruitfly.org/seq_tools/promoter.html)
  - ICG: [TATA-Box predictor](#)
- Predikce polyA signálu
  - CSHL: [argon.cshl.org/tabaska/polyadq\\_form.html](http://argon.cshl.org/tabaska/polyadq_form.html)
- Predikce míst sestřihu
  - BDGP: [http://www.fruitfly.org/seq\\_tools/splice.html](http://www.fruitfly.org/seq_tools/splice.html)
- Identifikace start/stop kodonu
  - DNALC: [Translator/ORF-Finder](#)
  - BCM: [Searchlauncher](#)