Počítačové vyhledávání genů a funkčních oblastí na DNA

Hledání genů

- Geny tvoří **obsahovou složku** genomu
  - Variabilní délka
  - Jedinečné sekvence
  - Mnohdy složené z exonů a intronů
  - Geny pro funkční RNA

- Jakým způsobem vyhledávat geny?

Přístupy pro hledání genů

- 1. Metody založené na hledání podobností s již popsanými geny
- 2. Metody srovnávací genomiky
  - Srovnání více dokončených genomů
- 3. Využití algoritmů a statistických metod pro analýzu sekvence
  - Hledání signálů
- Integrované přístupy

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
  - 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%)
  - Struktura intron/exon
  - statistické modely frekvencí nukleotidů
  - sledování závislostí přítomných ve struktuře kodonů
  - Obsah GC
  - Přesnost dosahuje cca 50 %
  - Problémy: mnoho!

Metody založené na podobnosti

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

1. Metody založené na hledání podobností s již popsanými geny

- Databáze
  - Proteiny
  - cDNA
  - EST
- Nástroje pro párové přiložení sekvencí umožňující analýzu genů
  - **Hledání genů na základě podobnosti sekvencí proteinů**
    - **blastx**
    - **tblastn**
    - **fastX**

2. Srovnávací genomika

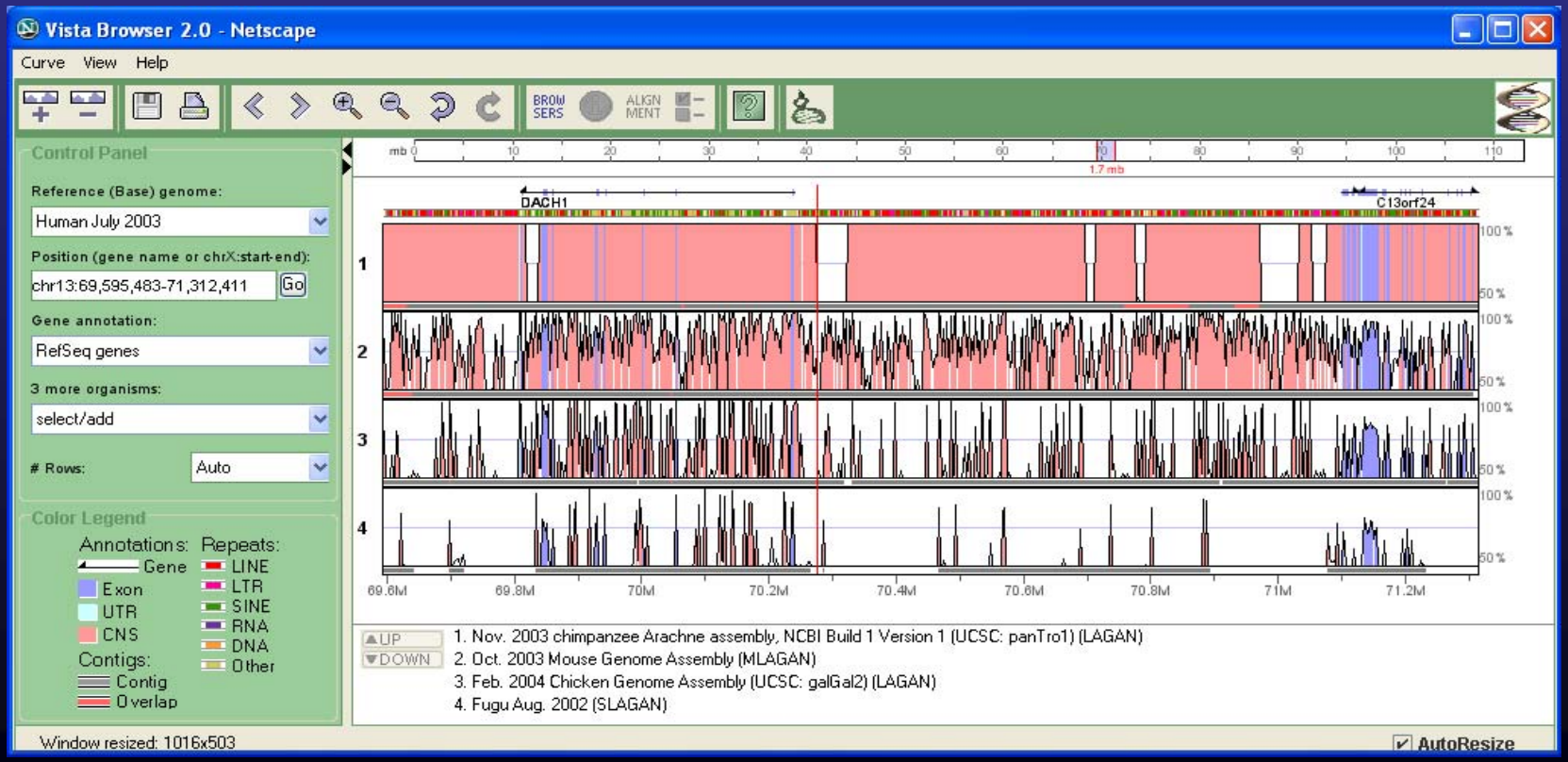
- Založené na předpokladu, že kódující sekvence jsou více konzervativní než nekódující
- Dva přístupy:
  - intra-genomický (genové rodiny)
  - inter-genomický (mezi druhy)
- Mnohonásobné přiložení homologických oblastí
  - exony
  - regulační oblasti
- Obtížné stanovení limitů podobnosti a optimální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

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

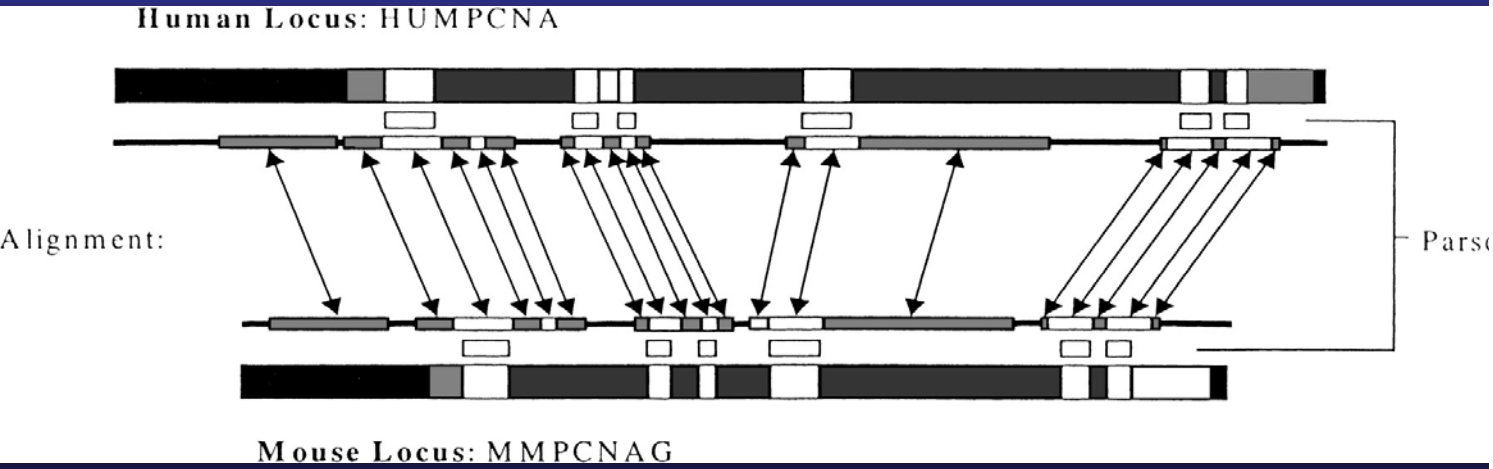


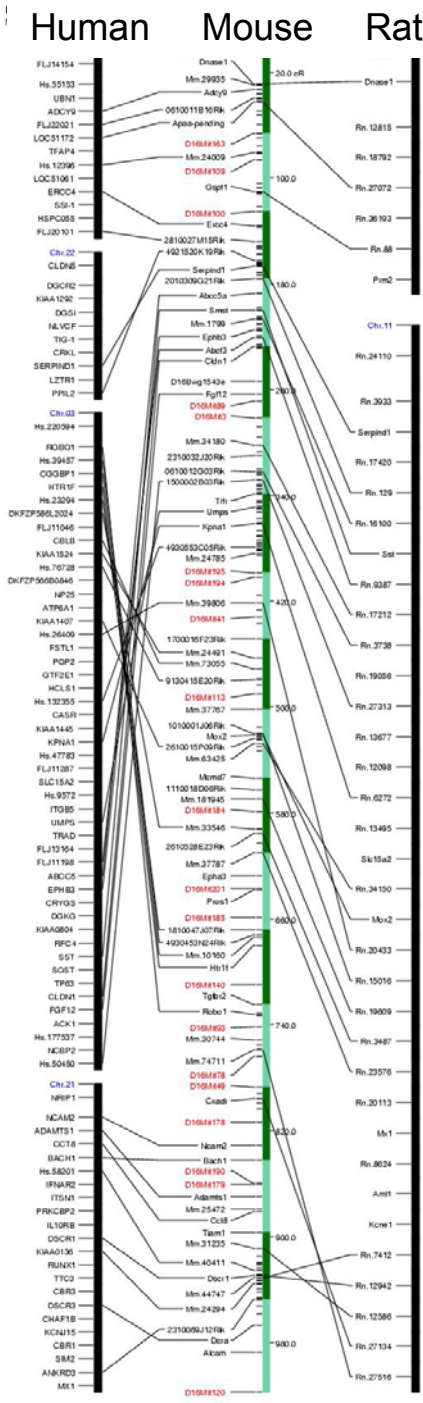
- šimpanz
- myš
- kuře
- Fugu

VISTA Genome Browser http://genome.lbl.gov/vista/index.shtml

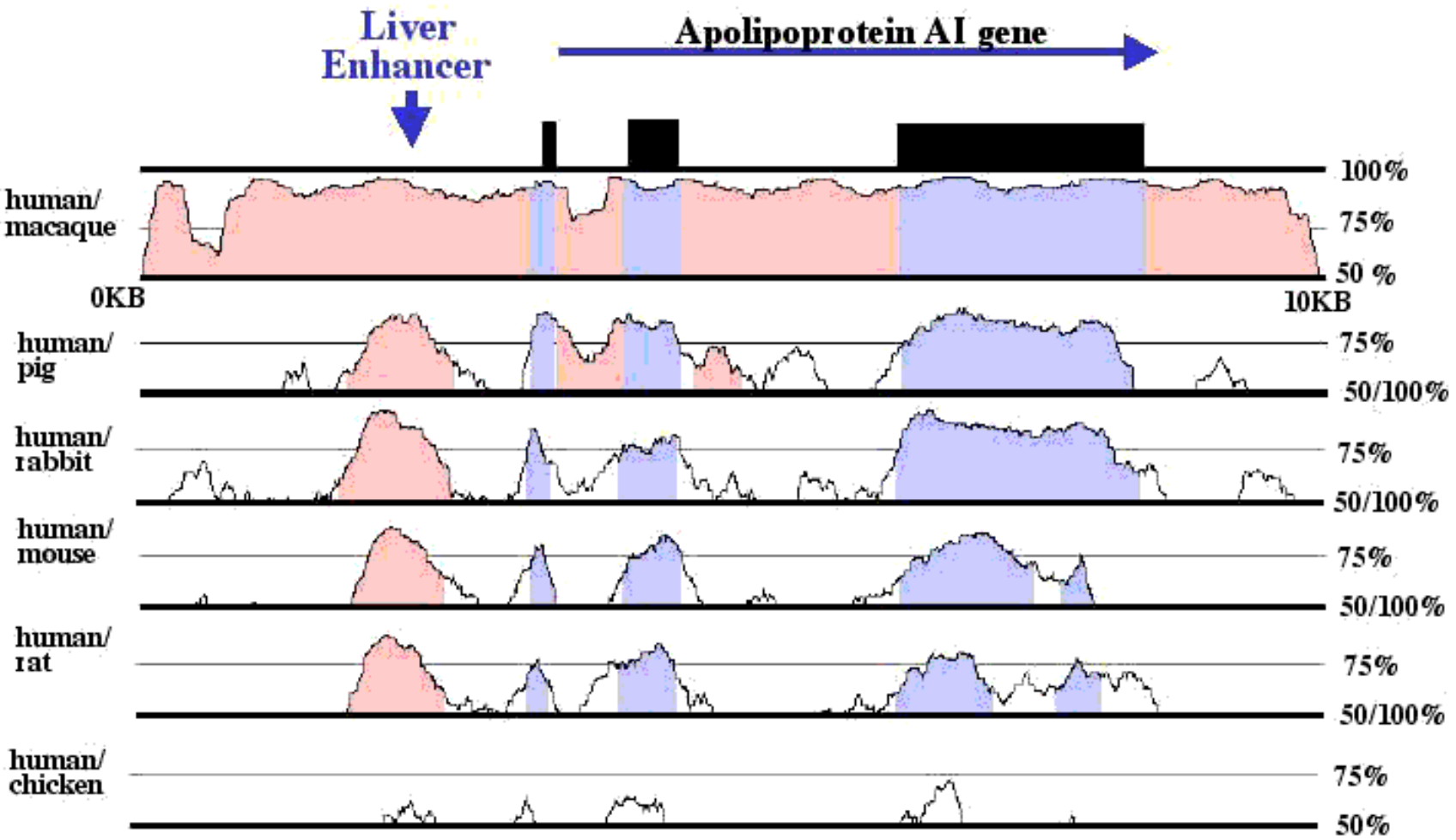
Příklad srovnání lokusů a chromozómů

Charakterizace rozdílů umožňuje odhalit mechanismy změn

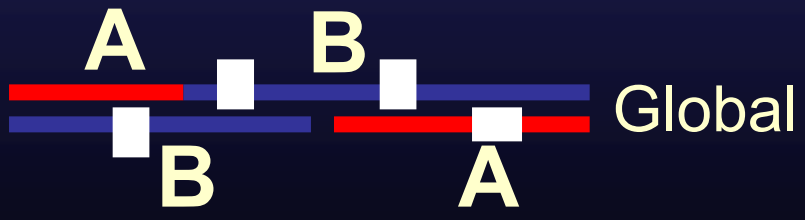


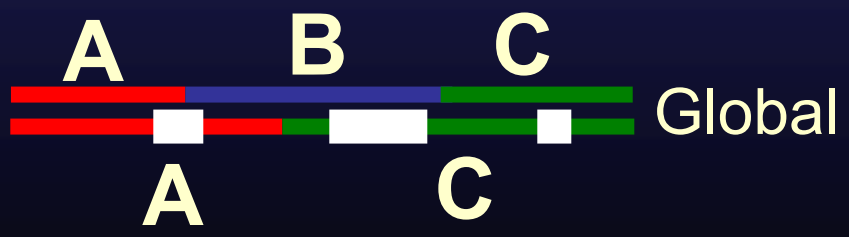


Multi-Species Comparative Analysis



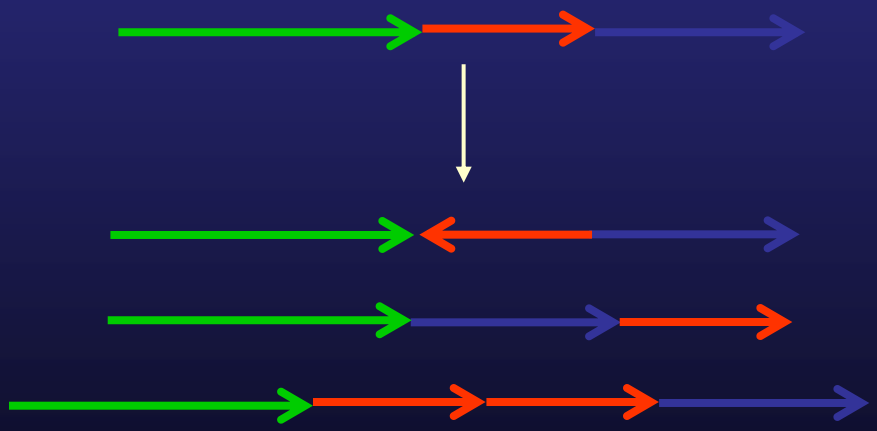
Lokální versus globalizované přiřazení





Problém globálního přiřazení

Nalezení nejefektivnější transformace jedné sekvence do druhé vyžaduje využití nových přístupů



- Bodové změny, delece
- Inverze
- Translokace
- Duplikace
- Kombinace uvedených změn

Základní zdroje a přístupy

- Databáze
  - NCBI: Genomy, Geny, Proteiny, SNPs, ESTs, Taxonomie, atd.
  - TIGR: databáze genomových center
- Analytický software
  - Databázové dotazy (nalezení podobných sekvencí), algoritmy pro přiřazení, shluková analýza, vyhledávání repetic, predikce genů
- Algoritmy pro dlouhá globální přiřazení
  - algoritmy pro lokální přiřazení s rozšířeným vkládáním mezer – citlivé, ale málo specifické pro dlouhé sekvence
    - BLASTZ
    - BLAT
  - algoritmy pro globální přiřazení
    - 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 odpovídajících si 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 přiřazením dlouhých sekvencí (BLAT)
  - rat – mouse
  - rat - human

Multi-LAGAN (MLAGAN)

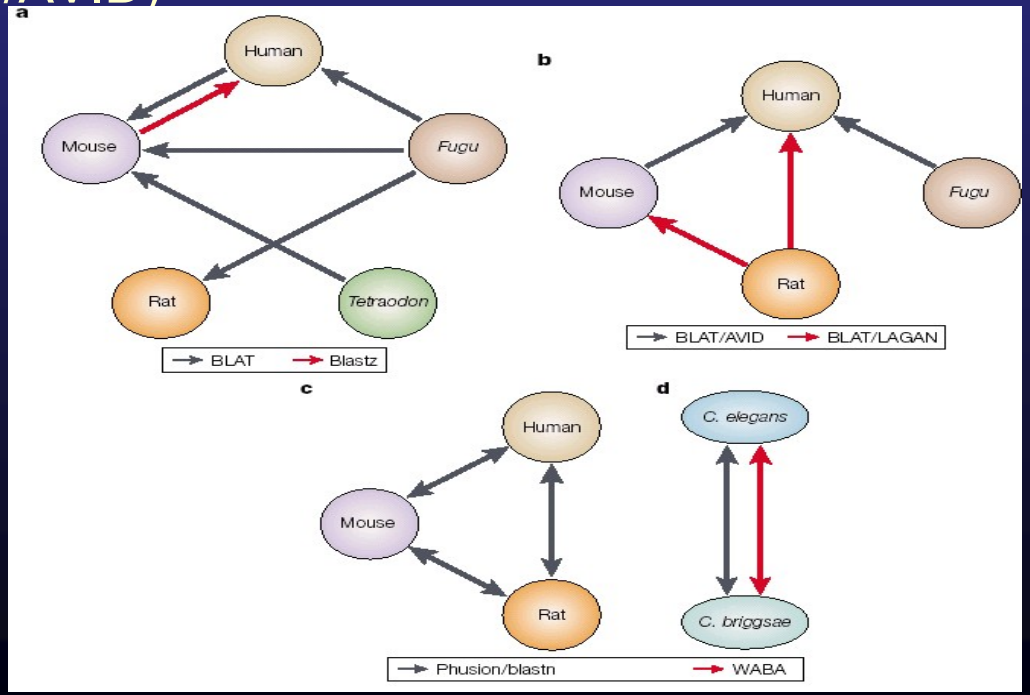
- V porovnání s LAGAN provádí navíc mnohonásobná globální přiřazení
- Nejprve provede přiřazení více příbuzných genomů a následně přiřazuje genomy více fylogeneticky vzdálené
- Umožňuje konstrukci fylogenetických stromů na základě globálního přiřazení genomů

Shuffle-LAGAN (S-LAGAN)

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

Precomputed alignments

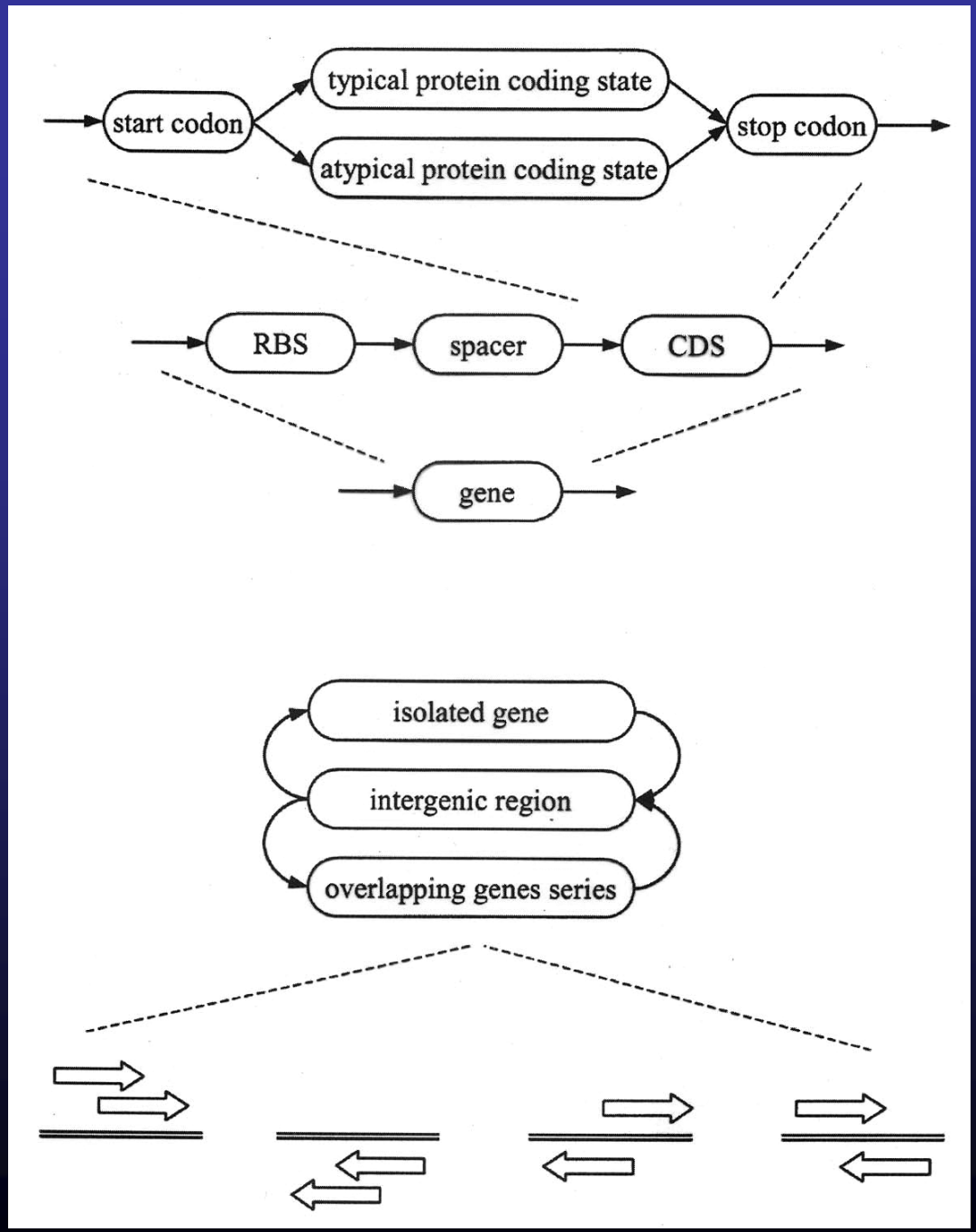
- U významných skupin organismů jsou k dispozici rozsáhlá mezidruhová srovnání
  - UC Santa Cruz/PennState (translated BLAT or BLASTZ)
  - Berkeley Genome Pipeline (BLAT/AVID)
  - Ensembl (Phusion/Blastn)
  - Vista (LAGAN/SLAGAN/AVID)



3. Predikce překládané oblasti na základě hledání signálů

- Hledání otevřených čtecích rámců doplněné hledáním konzervativních signálů v transkripčních jednotkách

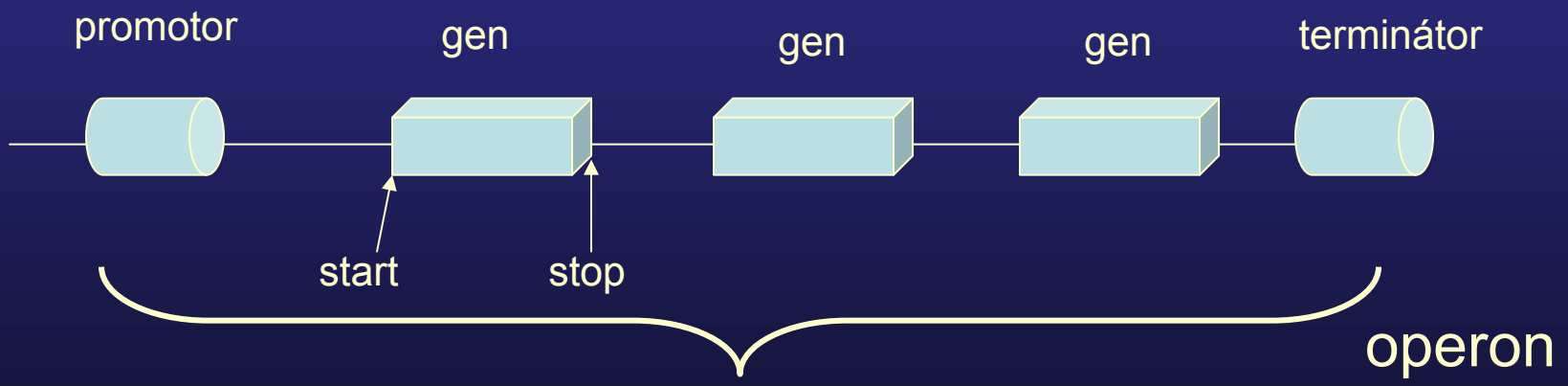
- **ORF Finder (Open Reading Frame Finder)**
  - http://www.ncbi.nlm.nih.gov/gorf/gorf.html


start codon
typical protein coding state
atypical protein coding state
stop codon
RBS
spacer
CDS
gene
isolated gene
intergenic region
overlapping genes series

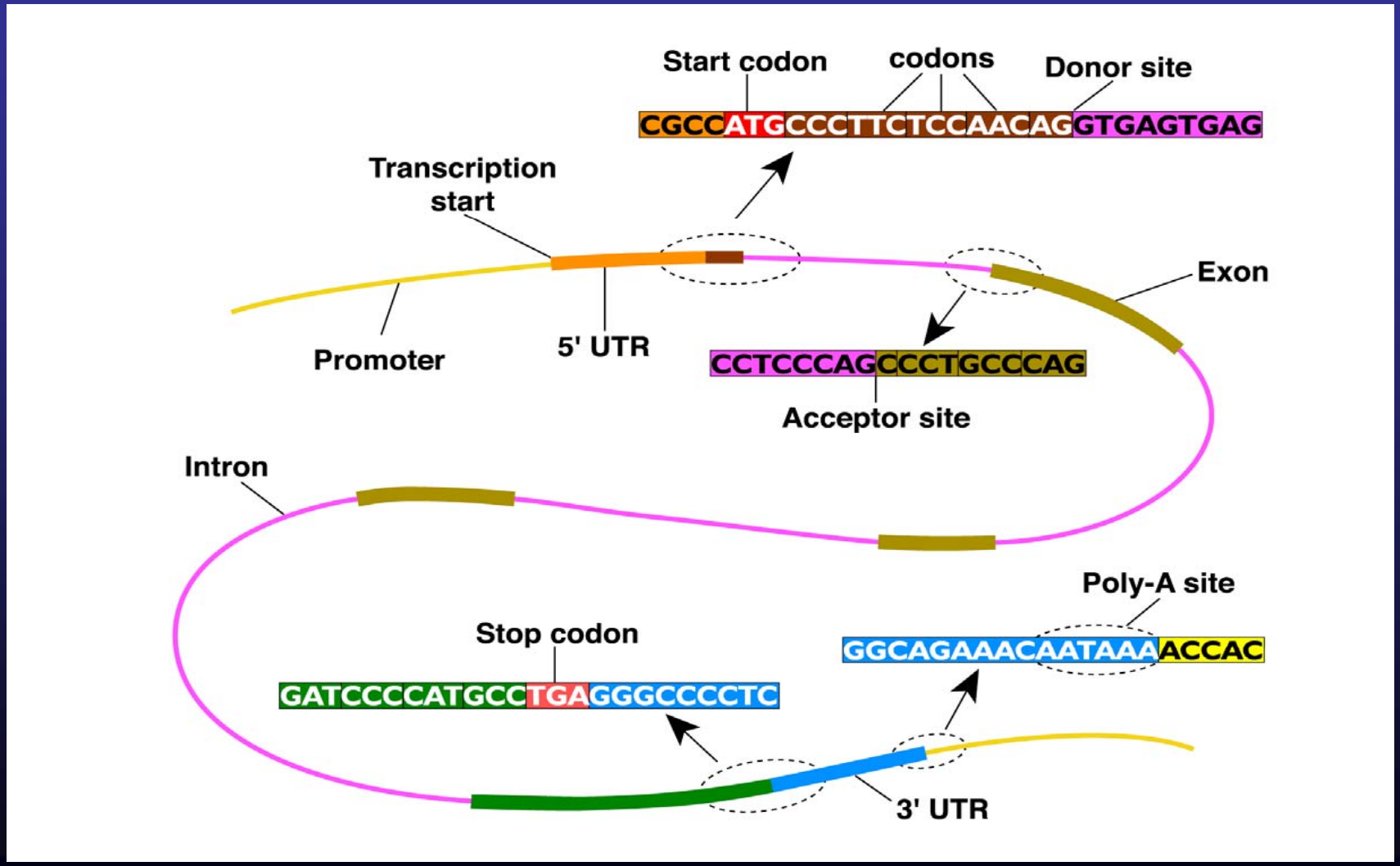
Výpočetní přístupy
*Klíčové jsou **signály** pro odhalení genů*

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

Struktura prokaryotické transkripční jednotky



Signály – senzory ve struktuře eukaryotického genu



Signály v jednoduchém strukturním genu

fem gene

```
  1 ATATGGTCAGTGCATATAAAATTTGTTATCATTAGAGTAATTAAAGGTCATTTAATAACTTTTGGAATCA  70
 71 ATTGGAGGTTCTCATATGTTATCTTTTAGTCAAAATAGAAGTCATAGCTTAGAACAATCTTTAAAAGAAG 140
141 GATATTCACAAATGGCTGATTTAAATCTCTCCCTAGCGAACGAAGCTTTTCCGATAGAGTGTGAAGCATG 210
211 CGATTGCAACGAAACATATTTATCTTCTAATTCAACGAATGAATCATTAGACGAGGAGATGTTTATTTAG 280
281 CAGATTTATCACCAGTACAGGGATCTGAACAAGGGGGAGTCAGACCTGTAGTCATAATTCAAAATGATAC 350
351 TGGTAATAAATATAGTCCTACAGTTATTGTTGCGGCAATAACTGGTAGGATTAATAAAGCGAAAATACCG 420
421 ACACATGTAGAGATTGAAAAGAAAAAGTATAAGTTGGATAAAGACTCAGTTATATTATTAGAACAAATTC 490
491 GTACACTTGATAAAAAACGATTGAAAGAAAAACTGACGTACTTATCCGATGATAAAATGAAAGAAGTAGA 560
561 TAATGCACTAATGATTAGTTTAGGGCTGAATGCAGTAGCTCACCAGAAAAATTAGGCGTCTATTATATGT 630
631 ATTTTTCAGAGATAAATAAAATATTGATATAAAAGACAATAACTTTATAATAATTATAACTATTTCTAAA 700
701 TTCTGTACGAAGAATTTTCTTATAAACAAAGATTTTAGCAAATACCAGTTATGATATTCATATTTTTTAT 770
771 TATAAAAGGATGTCTTAAGTTTTTTAGGCTTTAGGTATTCCATCCTAAAGTTTTTTTTAGCTTAAAAGTA 840
841 TCATCTACAGCAAAATTGCAAACGACAAAATTGATAAGTGCAATTAAATAAATGTTAGTAAGTGAATCAT 910
911 AATTATCCTTGCTTAAGCATTTGCTTTGTAAGGGAAGTGAGGAGGCAACTAATCG 965
```

rsbU gene

putative promotor

putative RBS

start

stop

terminator

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 seřazení příslušné subsekvence našeho zájmu

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

konsensus sequence TATAAT

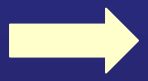
konsensus (IUPAC) TATRNT

- 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
TATAAT
TATAAT
GATACT
TATGAT
TATGTT



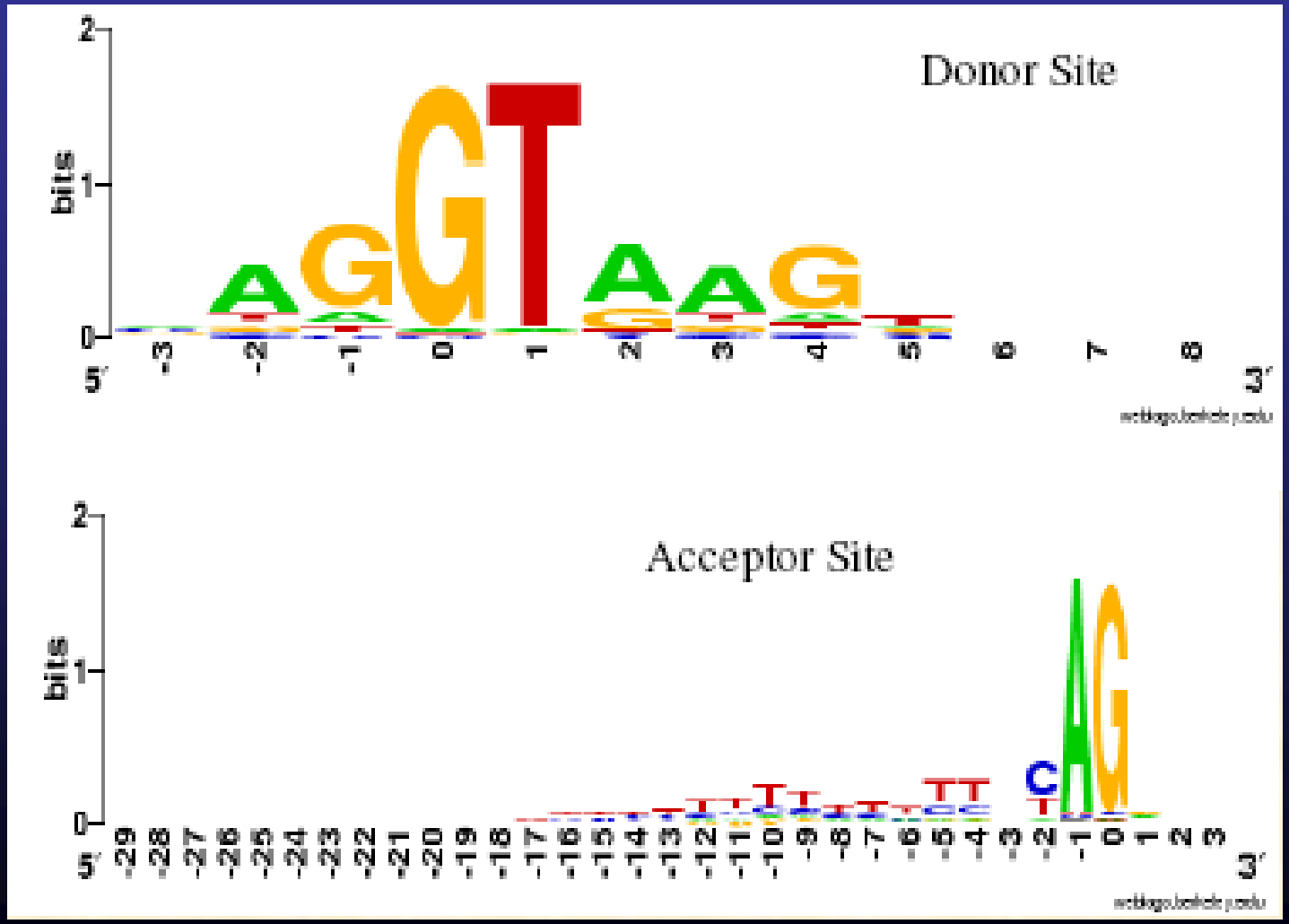
	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

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

Příklad signálu

RBS (vazebné místo pro ribozóm)

Příklad signálu: místo sestřihu (myš)



Analýza sekvence predikovaného genu

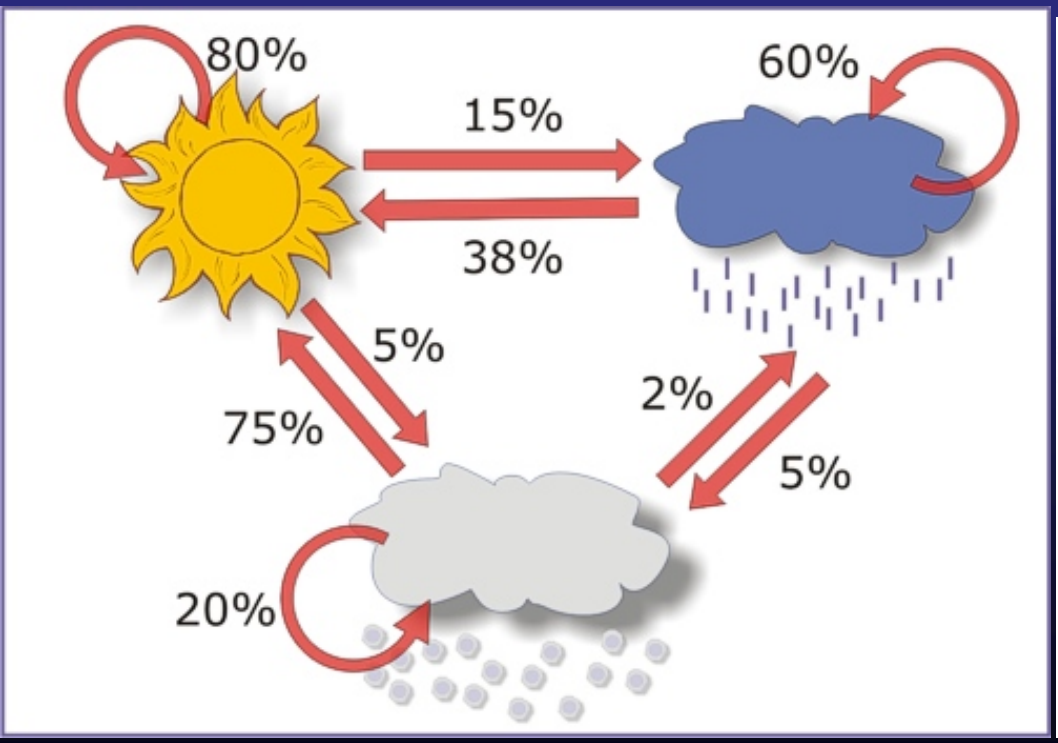
- Důležité je posouzení charakteru sekvence
  - délka
  - obsah GC
  - statistické modely modely frekvencí nukleotidů
  - frekvence využití kodonů

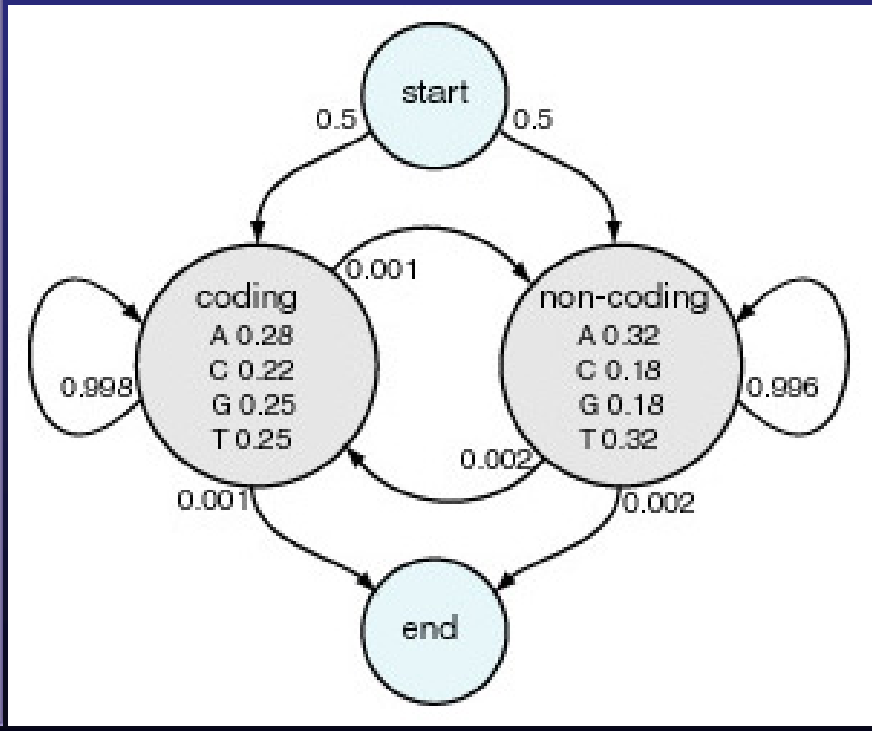
The Human Codon Usage Table

Gly	GGG	17.08	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.94	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.86	0.58	Arg	CGC	10.82	0.19
Glu	GAG	58.82	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.06	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.48	0.58	His	CAC	14.00	0.59
Val	GTG	28.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

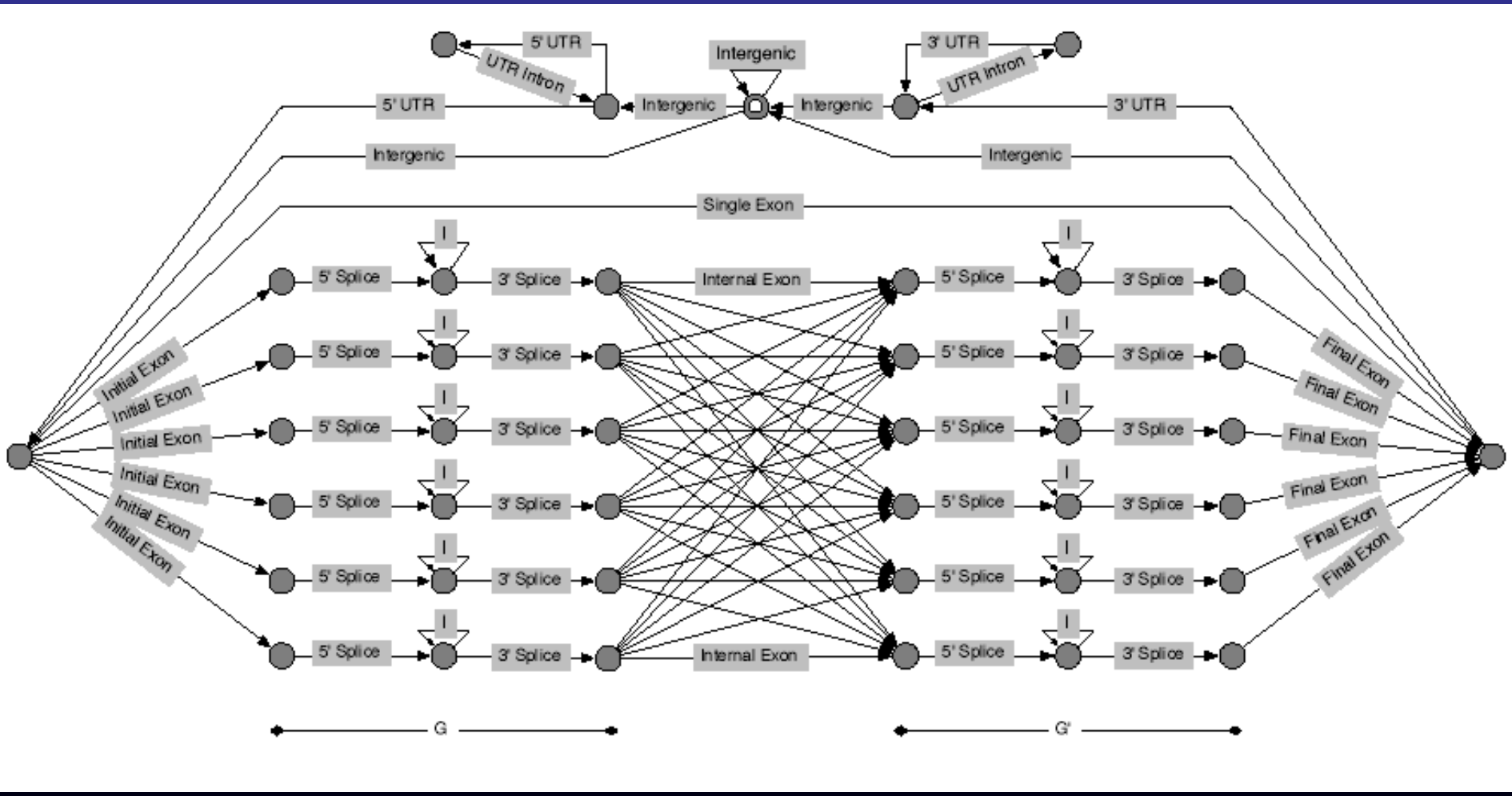
Markovovy modely

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

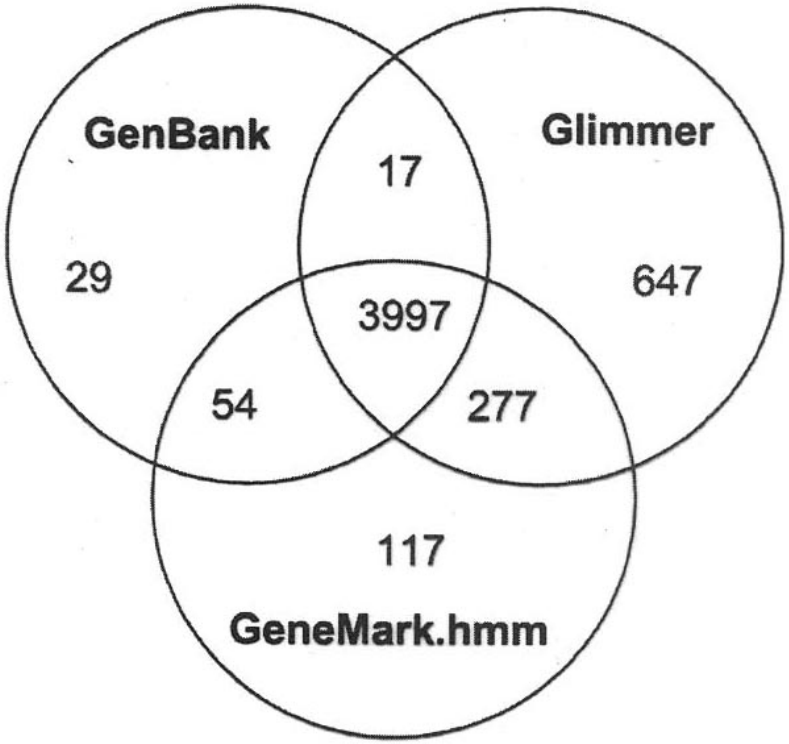




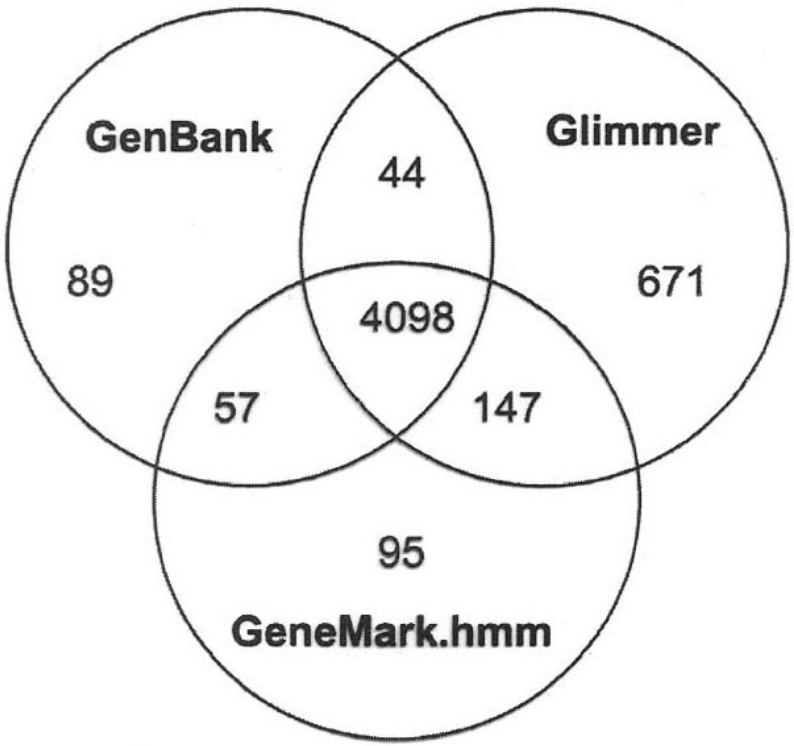
Příklad komplexního algoritmu se skrytými Markovovy modely (HMM)



Srovnání různých přístupů pro vyhledávání genů

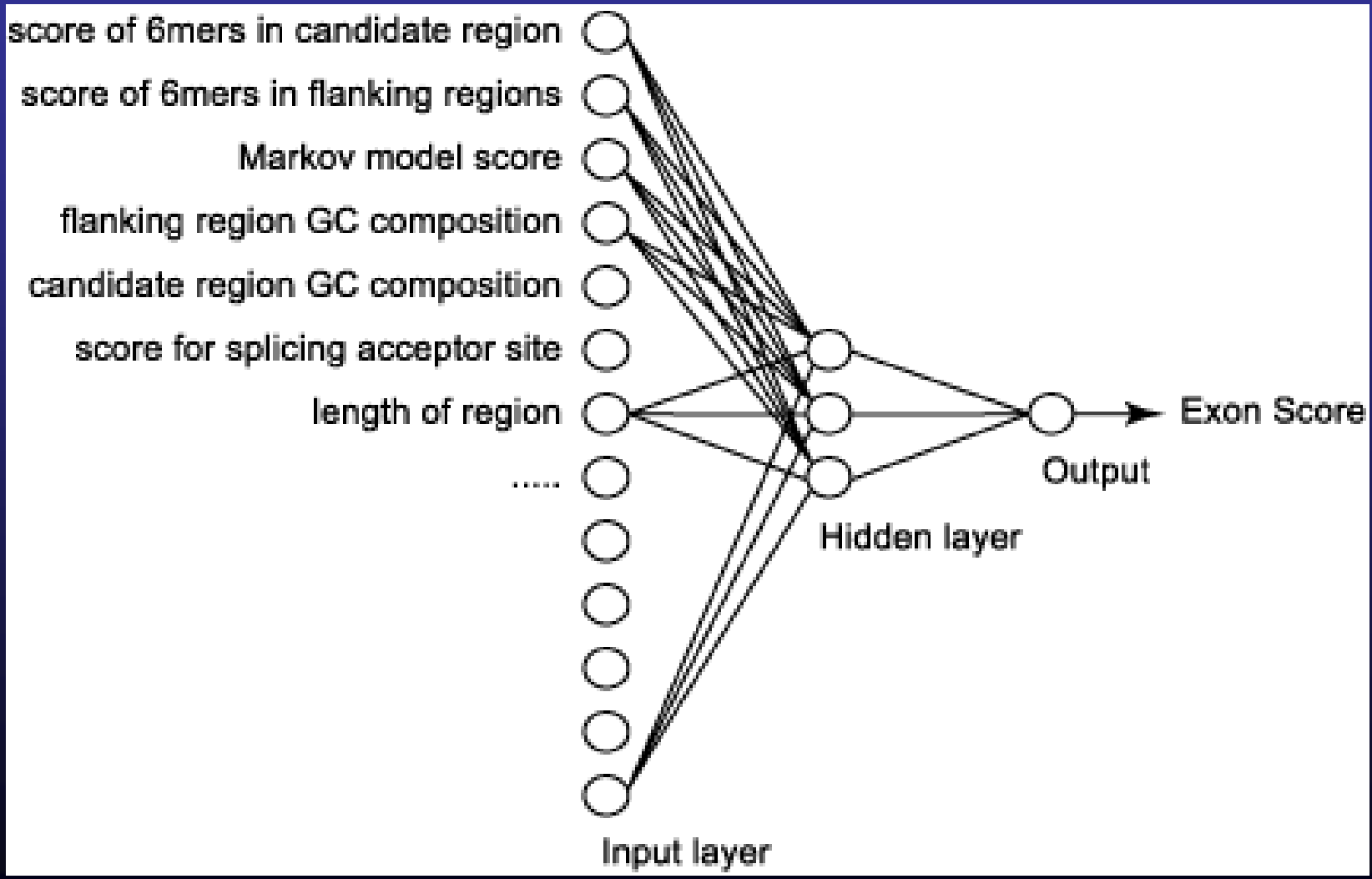


B. subtilis



E. coli

Predikce eukaryotických genů: GRAIL II: využívá neuronové sítě



Doporučený software pro cvičení – GenMark

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