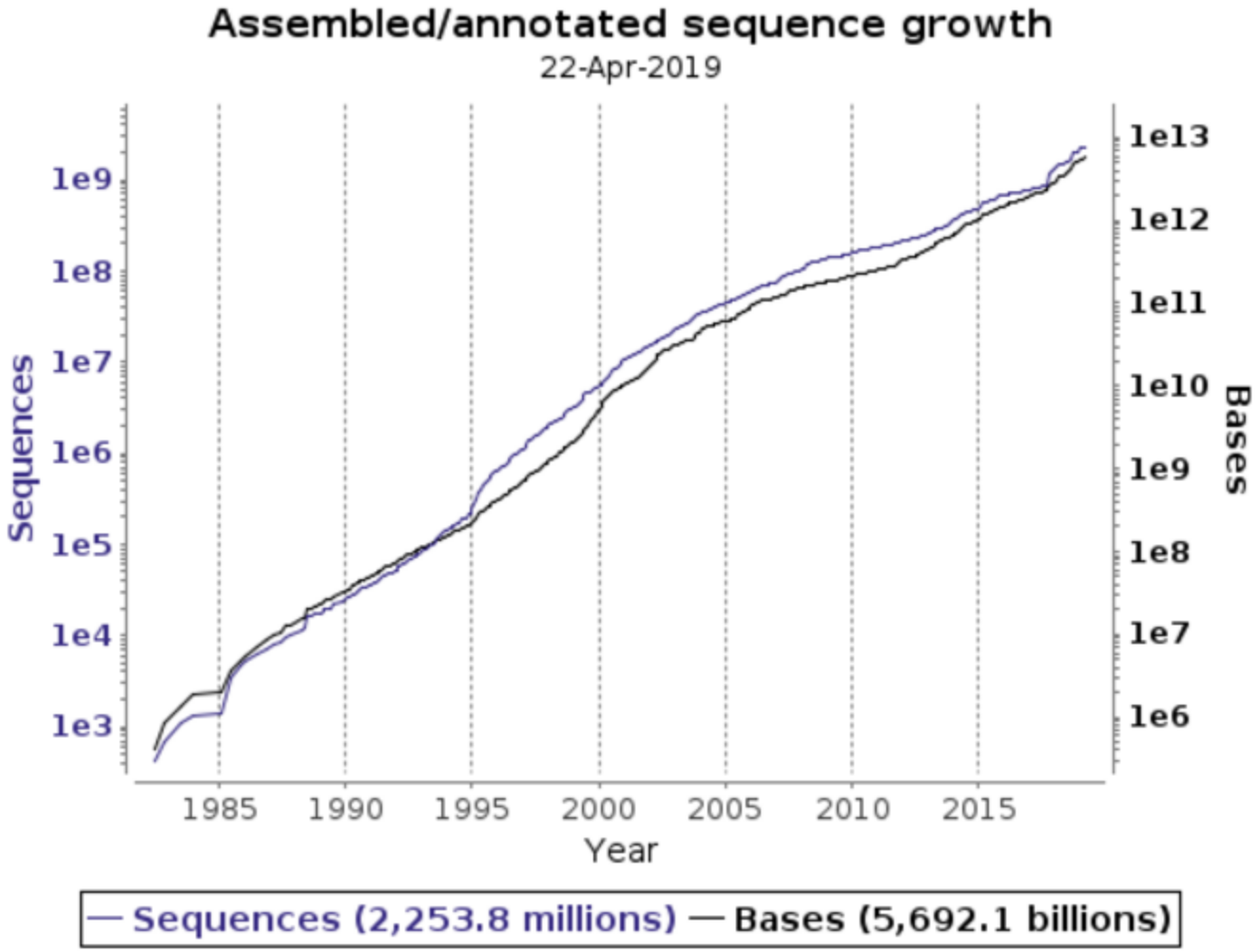
Analýza sekvenčních dat v molekulární biologii

- **Bioinformatika** je disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie
- Termín bioinformatika se objevil poprvé až v roce 1991
- Představuje spojení technologií z oblastí
  - molekulární biologie
  - informačních technologií
- Bioinformatika zahrnuje
  - studium
  - praktické uchovávání
  - vyhledávání
  - zobrazování
  - manipulaci
  - a modelování biologických dat
- Potřeba pracovat s velice obsáhlými databázemi si vyžádala vývoj výpočetních nástrojů umožňujících analýzu dat a stanovení jejich vzájemných vztahů.
- Vývoj vysoce výkonných technologií umožňujících získání molekulárně biologických dat přispěl k jejich dramatickému nárůstu a tím současně zvýšil obtížnost jejich zkoumání a hodnocení ve vztahu k biologickým otázkám.

Trend nárůstu množství dat v bioinformatických databázích

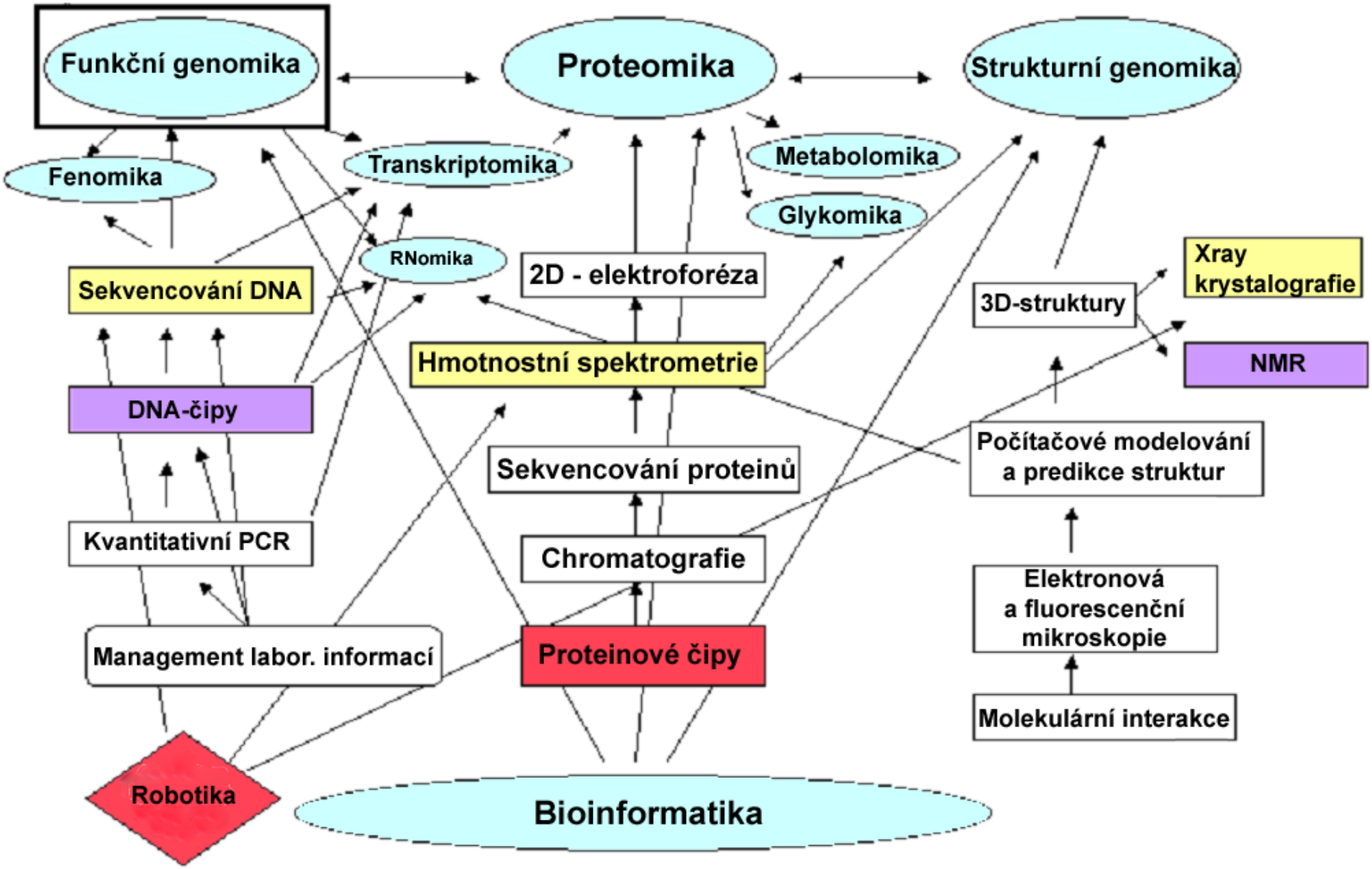
Assembled/annotated sequence growth



Základní zdroje a aplikace bioinformatiky

Výpočetní základy	**Zdroje dat**	**Aplikace bioinformatiky**
Algoritmy	Obecně dostupné databáze	Získávání dat
Grafika, vizualizace		Nástroje pro přístup k databázím
Zpracování signálu		Mapování a srovnávání genomů
Architektura hardwaru		Sekvenční přiložení, assembly
Informační teorie		Identifikace genů
Správa databází		Funkční identifikace proteinů
Statistika		Molekulární evoluce
Simulace		Molekulární modelování
Umělá inteligence	Zpracování laboratorních dat	Predikce struktur
Zpracování obrazu		Srovnávání struktur
Robotika		Stanovení makromolekulárních struktur
Softwarové inženýrství		Vývoj léčiv na základě struktur

„...omiky“ v molekulární biologii



Nejdůležitější instituce zabývající se shromažďováním biomedicínských informací

- K nejdůležitějším institucím zabývajícím se, správou dat a vývojem nástrojů pro jejich analýzu a poskytováním informací patří:
  - **Evropský institut pro bioinformatiku (EBI)** se sídlem v Hinxtonu v UK (http://www.ebi.ac.uk/),
  - **Národní centrum pro biotechnologické informace (NCBI)** založené původně v rámci Národní lékařské knihovny (NLM) v USA (http://www.ncbi.nlm.nih.gov/),
  - **Centrum pro informační biologii (CIB)** založené jako oddělení Národního genetického institutu (NIG) v Mishimě, Japonsko (http://www.cib.nig.ac.jp/).
- V současné době je prostřednictvím Internetu dostupných přibližně 550 databází zabývajících se shromažďováním bioinformací.
  - Jejich přehled a popis je každoročně publikován ve specializovaném, volně dostupném čísle časopisu Nucleic Acids Research.

Nejdůležitější databáze sekvencí nukleových kyselin a proteinů

- V každém ze tří hlavních bioinformatických center je spravována **genomová databáze** sekvencí nukleových kyselin a odpovídajících, z nich přeložených proteinů.
  - **EMBL Nucleotide Sequence Database / European Nucleotide Archive** (v rámci institutu EBI) – 1980
  - **GenBank** (v rámci institutu NCBI) – 1982
  - **DDBJ** (The DNA Data Bank of Japan) - 1984
- Tři samostatné báze vznikly v důsledku potřeby rychlé dostupnosti databáze sekvencí na jednotlivých kontinentech v době, kdy ještě nebyly rozvinuté vysokorychlostní komunikační sítě.

Sdílení dat ve třech základních databázích

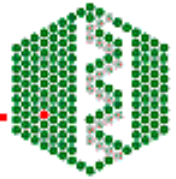
V každém z bioinformatických center jsou dostupné jednoduché nástroje pro manipulaci s daty

GenBank: *http://www.ncbi.nlm.nih.gov/*

National Center for Biotechnology Information (NCBI)

EMBL: *http://www.ebi.ac.uk*

EMBL
European Bioinformatics Institute



DDBJ: *http://www.ddbj.nig.ac.jp/*

National Institute of Genetics (NIG)

ExPASy: *http://tw.expasy.org*

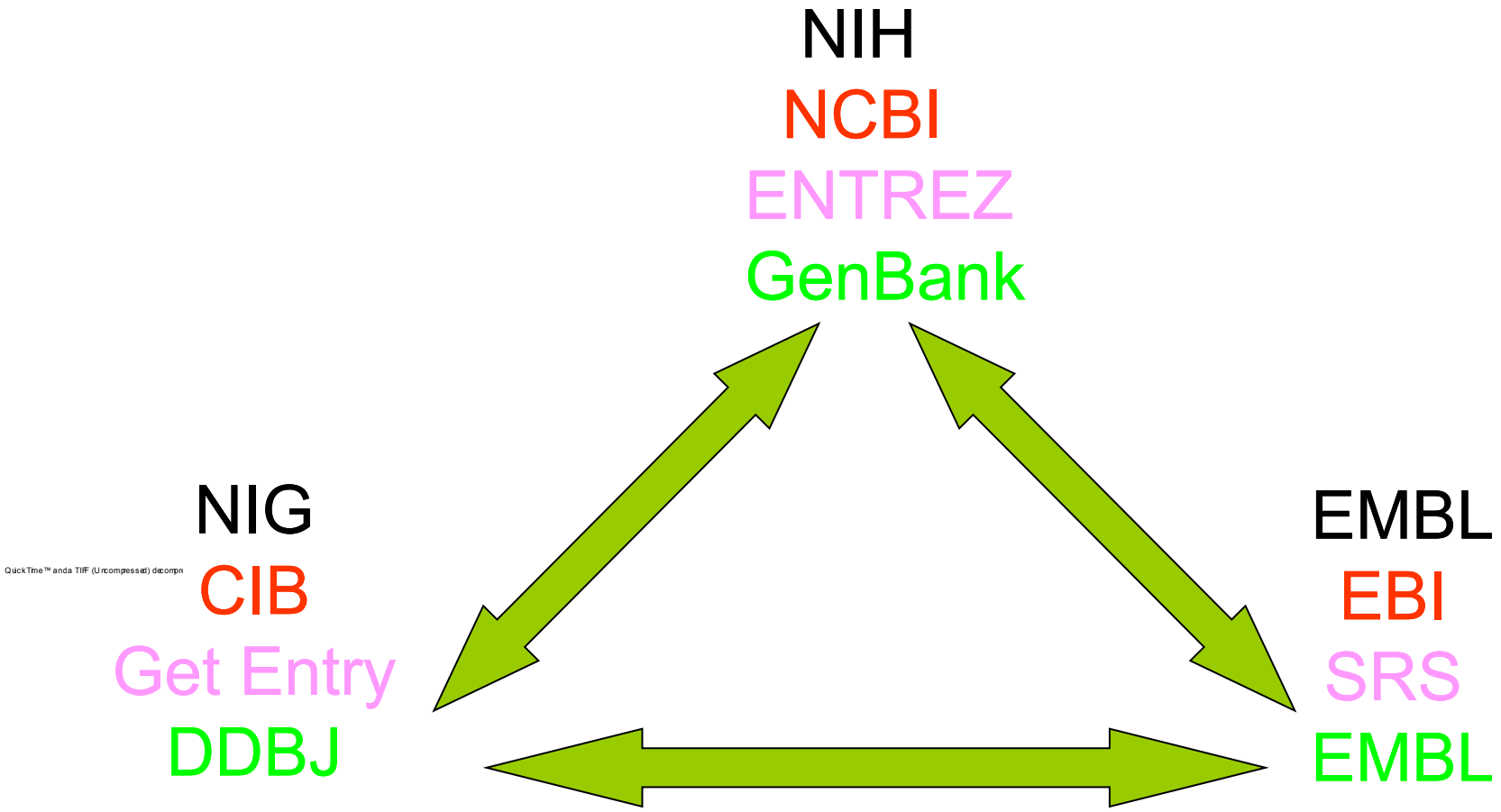
Expert Protein Analysis System

Databáze sekvencí proteinů

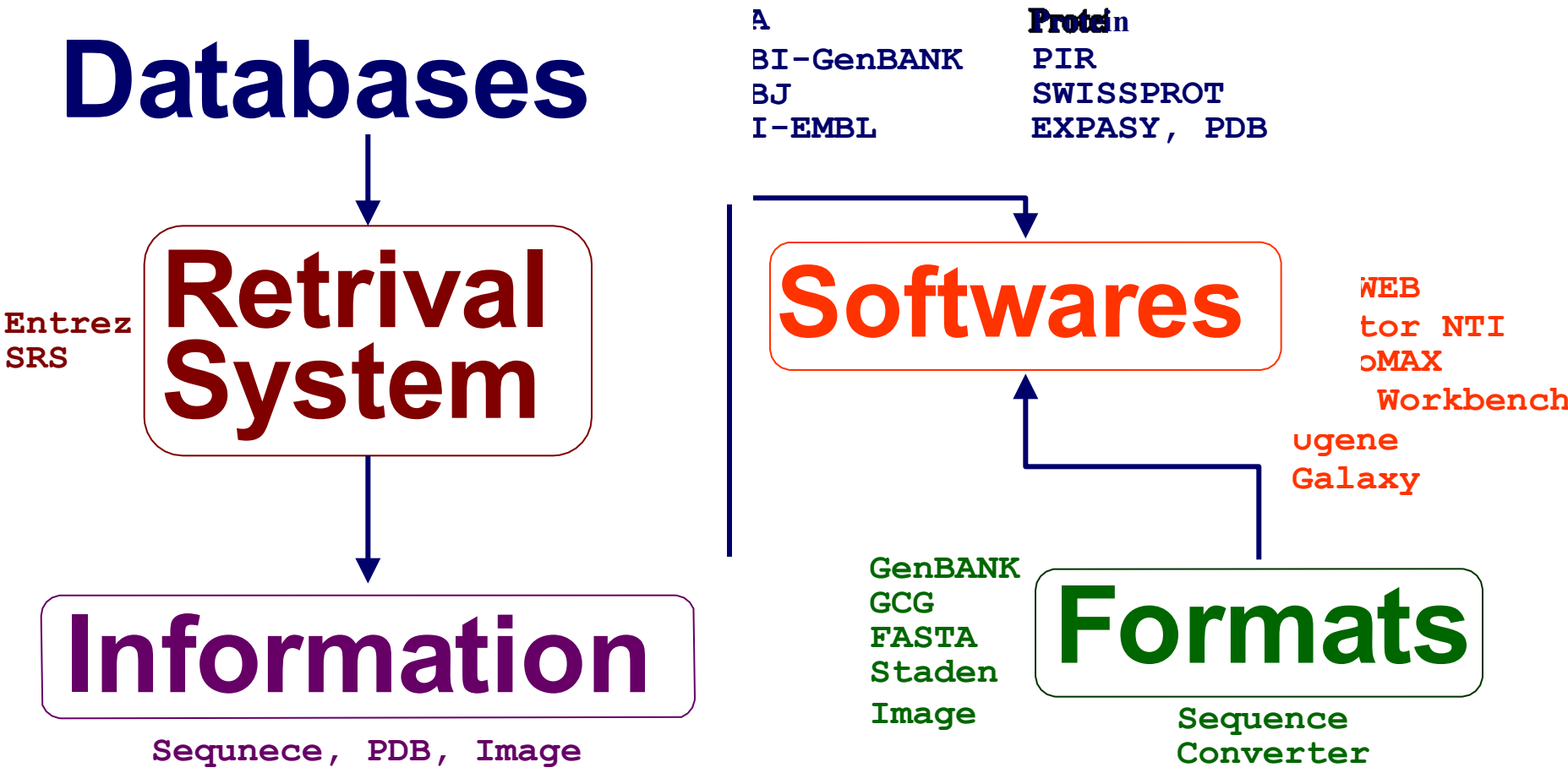
- Sekvence proteinů, u nichž byly experimentálně stanoveny jejich aminokyselinové sekvence, charakterizovány jednotlivé proteinové domény a stanovena jejich funkce jsou ukládány v databázi **SWISS-PROT** založené na Univerzitě v Ženevě v roce 1986.
- Databázi spravuje Švýcarský institut pro bioinformatiku (SIB), který se podílí na vytváření sítě propojených databází sekvencí.
- Kompletní databázi sekvencí proteinů obsahuje SWISS-PROT spolu s doplňkem označeným **TrEMBL**, který obsahuje automaticky doplňované překlady kódujících oblastí z databáze sekvencí nukleových kyselin EMBL.

Mezinárodní spolupráce sekvenčních databází

- Databáze sdílejí stejná data



Získání dat a manipulace se sekvencemi



- Ve sféře biotechnologií a medicíny je důležitou stránkou bioinformatiky přístup k publikované vědecké literatuře a také k patentovým archivům.
  - Jednou z největších databází na světě je **MEDLINE (PubMed)**, obrovský archiv odkazů z biologických a biomedicínských odborných časopisů pokrývající období od roku 1965 do současnosti a v poskytující kromě abstraktů také odkazy na celé texty článků u jednotlivých vydavatelů.

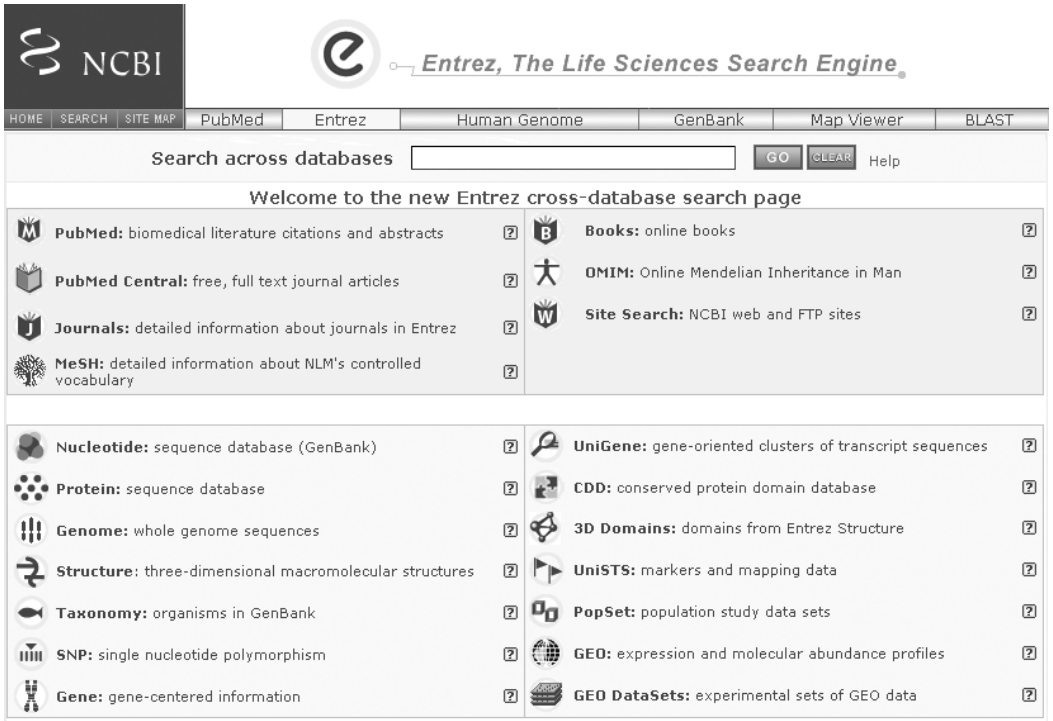
Textové vyhledávání v databázích

- Množství důležitých molekulárně-biologických dat se zvyšuje tak rychle, že je nezbytné mít k dispozici prostředky, pomocí kterých můžeme k těmto datům snadno přistupovat.
- Existují **tři prostředky** na získávání informací, které umožňují vyhledávání v molekulárně biologických databázích.
- Tyto prostředky jsou vstupním bodem do mnoha integrovaných databází a každý z nich byl vyvinut v jednom ze tří hlavních center pro bioinformatiku.
- Navzájem se liší v databázích, které mohou prohledávat, ve vazbách, které vytvářejí mezi jednotlivými databázemi a ve vazbách vztahujících se k dalším informacím

Entrez http://www.ncbi.nlm.nih.gov/

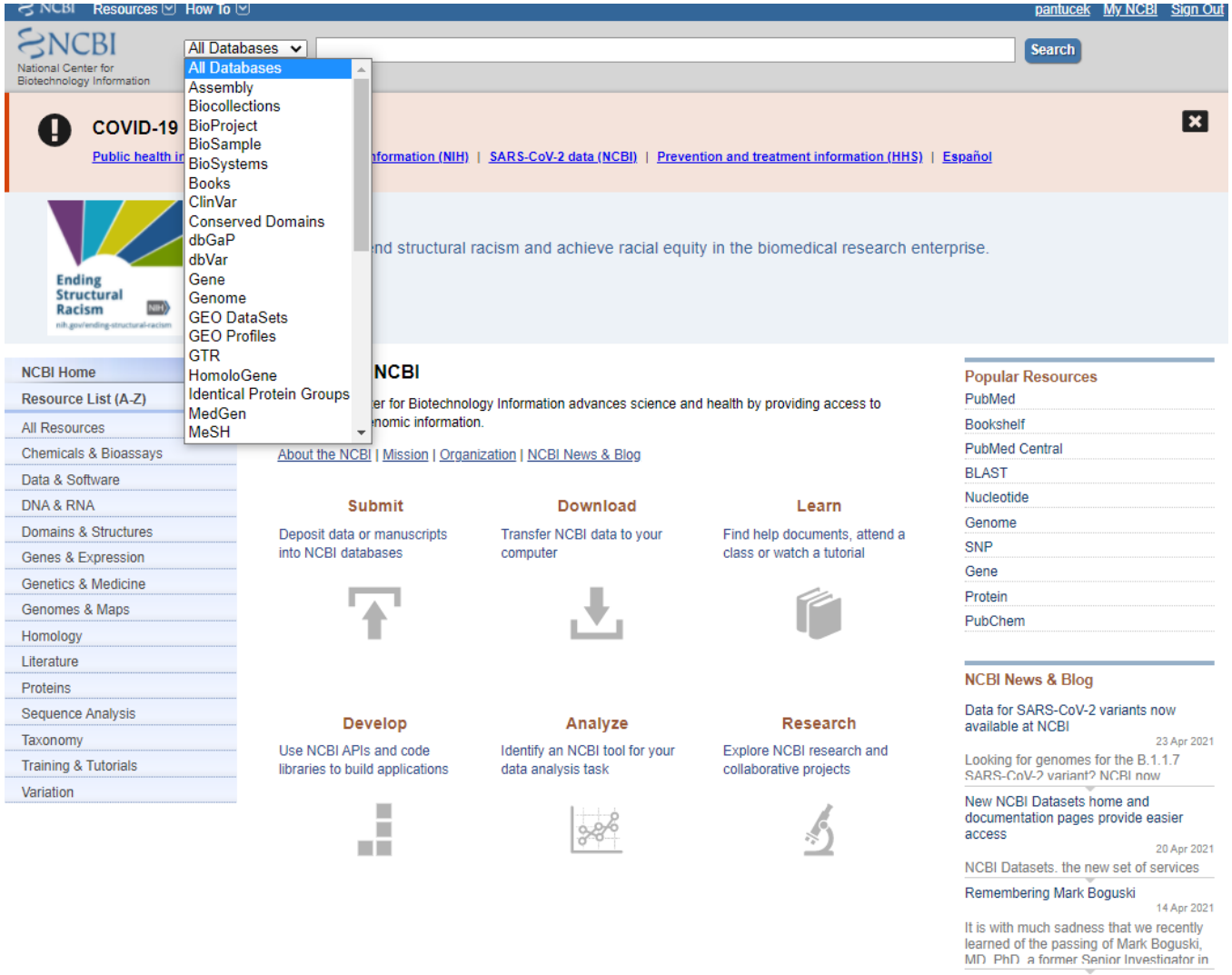
- **Entrez** je vyhledávací systém pro molekulárně biologické databáze vyvinutý v NCBI
- Je vstupním bodem pro průzkum 45 různých integrovaných databází z nichž řada je virtuálních.

- K nejvýznamnějším databázím patří
  - databáze PubMed, umožňující přístup k literární databázi MEDLINE
  - databáze sekvencí nukleových kyselin a proteinů
  - databáze 3-D struktur MMDB (Molecular Modeling Database)
  - skupina databází genomů
  - taxonomická databáze usnadňující získávání sekvencí na základě taxonomických skupin
- Ze tří vyhledávacích prostředků je Entrez uživatelsky nejpřijatelnější



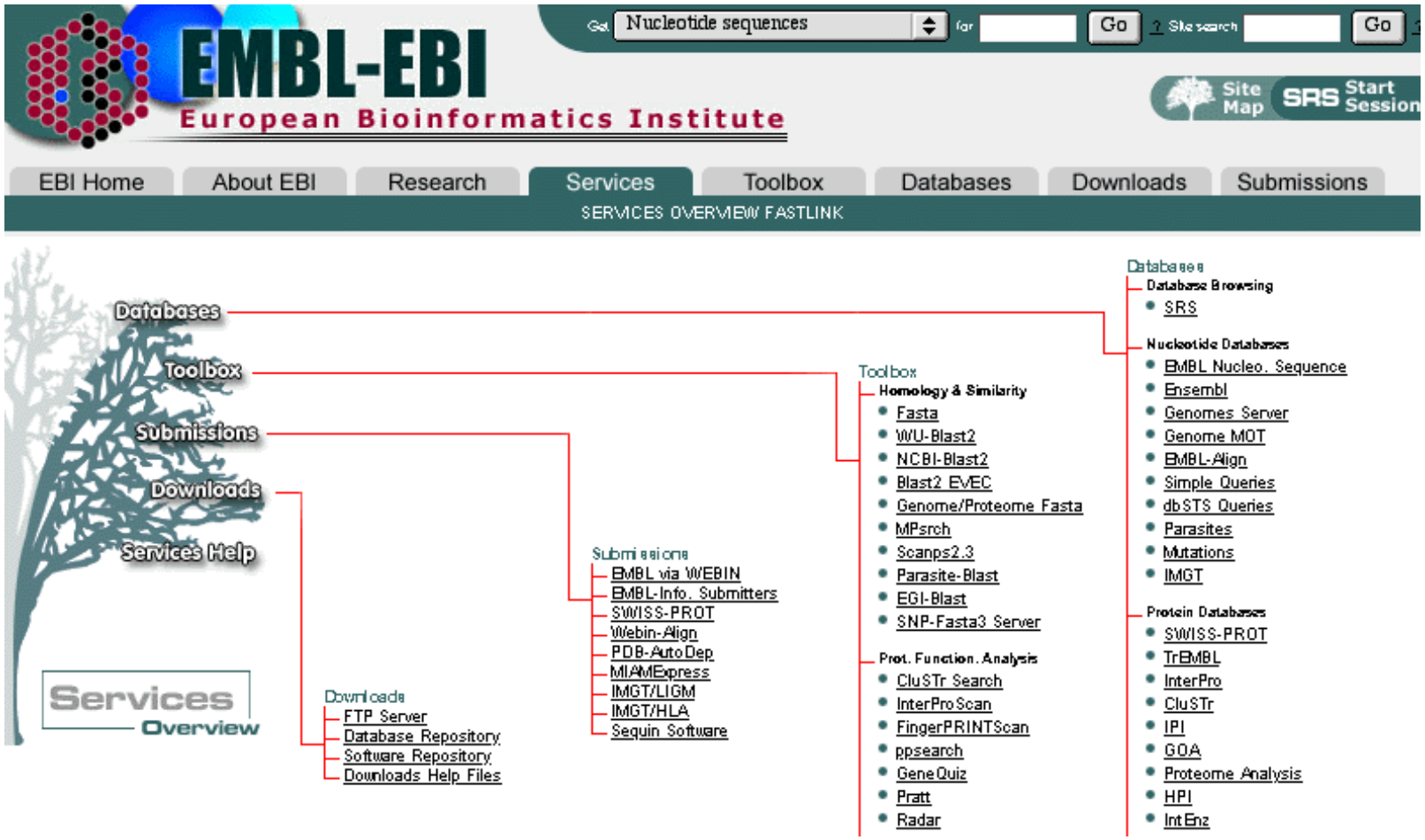
Entrez Molecular Sequence Database System

NCBI http://www.ncbi.nlm.nih.gov/



Sequence Retrieval System (SRS)

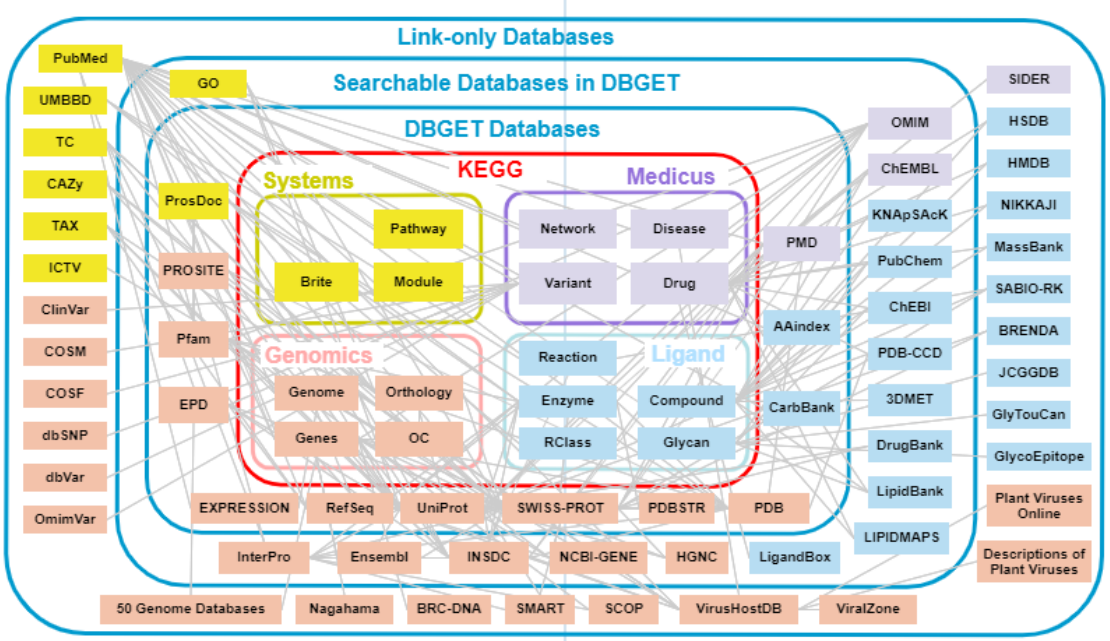
EBI http://www.ebi.ac.uk/



DBGET/Link DB

http://www.genome.ad.jp/dbget

- **DBGET/Link DB** je integrovaný systém pro získávání dat z databází vyvinutý v Institutu pro chemický výzkum na Univerzitě Kyoto v Japonsku
- Unikátní je propojení na databázi KEGG (Kyoto Encyclopedia of Genes and Genomes), což je databáze regulačních a metabolických drah u organizmů ze známým genomem.



Jak se data dostanou do databází?

- Předání dat prostřednictvím WWW portálu
  - BankIt (GenBank)
    - http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank
  - Submission Portal
    - https://submit.ncbi.nlm.nih.gov/
  - WebIn (EMBL/European Nucleotide Archive)
    - http://www.ebi.ac.uk/ena/submit
  - Sakura (DDBJ)
    - http://www.ddbj.nig.ac.jp/sub/websub-e.html
- Samostatná aplikace pro PC
  - Sequin
    - http://www.ncbi.nlm.nih.gov/Sequin/download/seq_download.html
  - pro delší sekvence (genomy)
  - fylogenetické, populační nebo mutační studie obsahující sekvenční přiložení
- Tbl2asn – batch submissin
  - command-line program for MAC a Unix
  - automatizuje vytvoření záznamu sekvence
  - určený pro celé genomy, EST, STS a zaslání velkých dávek sekvencí

Zápis sekvence

- **Sekvence** – zápis posloupnosti jednoznačných znaků odpovídajících jednotlivým zbytkům (monomerům), které se nacházejí v odpovídající posloupnosti v dané makromolekule
  - **DNA nebo RNA od 5‘-konce k 3‘-konci**
    - 5′CAAACGTCGTCTA3′
  - **protein od N-konce k C-konci**
    - (NH_2-)MKRLSALGPGGLTRR(–COOH)
- **používají se jednopísmenové kódy dle pravidel IUPAC**

Standardní kódy pro sekvence nukleových kyselin podle IUB/IUPAC

A	adenosin
C	cytidin
G	guanidin
T	thymidin
U	uridin
R	G/A (puRin)
Y	T/C (pYrimidin)
K	G/T (nukleosid s Keto skupinou)
M	A/C (nukleosid s aMino skupinou)
S	G/C (silná = Strong vazba)
W	A/T (slabá = Weak vazba)
B	G/T/C (not A)
D	G/A/T (not C)
H	A/C/T (not G)
V	G/C/A (not T)
N	A/G/C/T (jakýkoli)
-	mezera (gap) neurčené délky

Využití zápisu s degenerovanými nukleotidy

```
TACGGT
TATAAT
TATAAT
GATACT
TATGAT
TATATT
```

Konsenzní sekven `TATAAT`

Degenerovaná sek `KAYRNT`

Standardní kódy pro sekvence aminokyselin podle IUB/IUPAC

A	alanin
B	kys. asparagová nebo asparagin
C	cystein
D	kys. asparagová
E	kys. glutamová
F	fenylalanin
G	glycin
H	histidin
I	isoleucin
K	lysin
L	leucin
M	metionin
N	asparagin
P	prolin
Q	glutamin
R	arginin
S	serin
T	treonin
U	selenocystein
V	valin
W	tryptofan
Y	tyrosin
Z	kys. glutamová nebo glutamin
X	jakákoli aminokyselina
*****	translační stop (terminační kodon)
-	mezera (gap) neurčené délky

FASTA FORMAT

Může obsahovat více sekvencí

Začíná specifickým záhlavím „**>**“, za kterým následuje definice

Příklad:

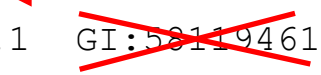
```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAG

>LinB_protein
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGR
LIACDLIGMGDSDKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVHDWGSALG
FDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGEELVLQDNVFVE
QVLPGLILRPLSEAEMAAYREPFLAAGEARRPTLSWPRQIPIAGTPADVVAIARDYA
GWLSESPIPKLFINAEPGALTTGRMRDFCRTWPNQTEITVAGAHFIQEDSPDEIGAA
IAAFVRRLRPA
```

Použití: univerzální formát pro zápis sekvence vhodný jako vstupní data pro většinu software.

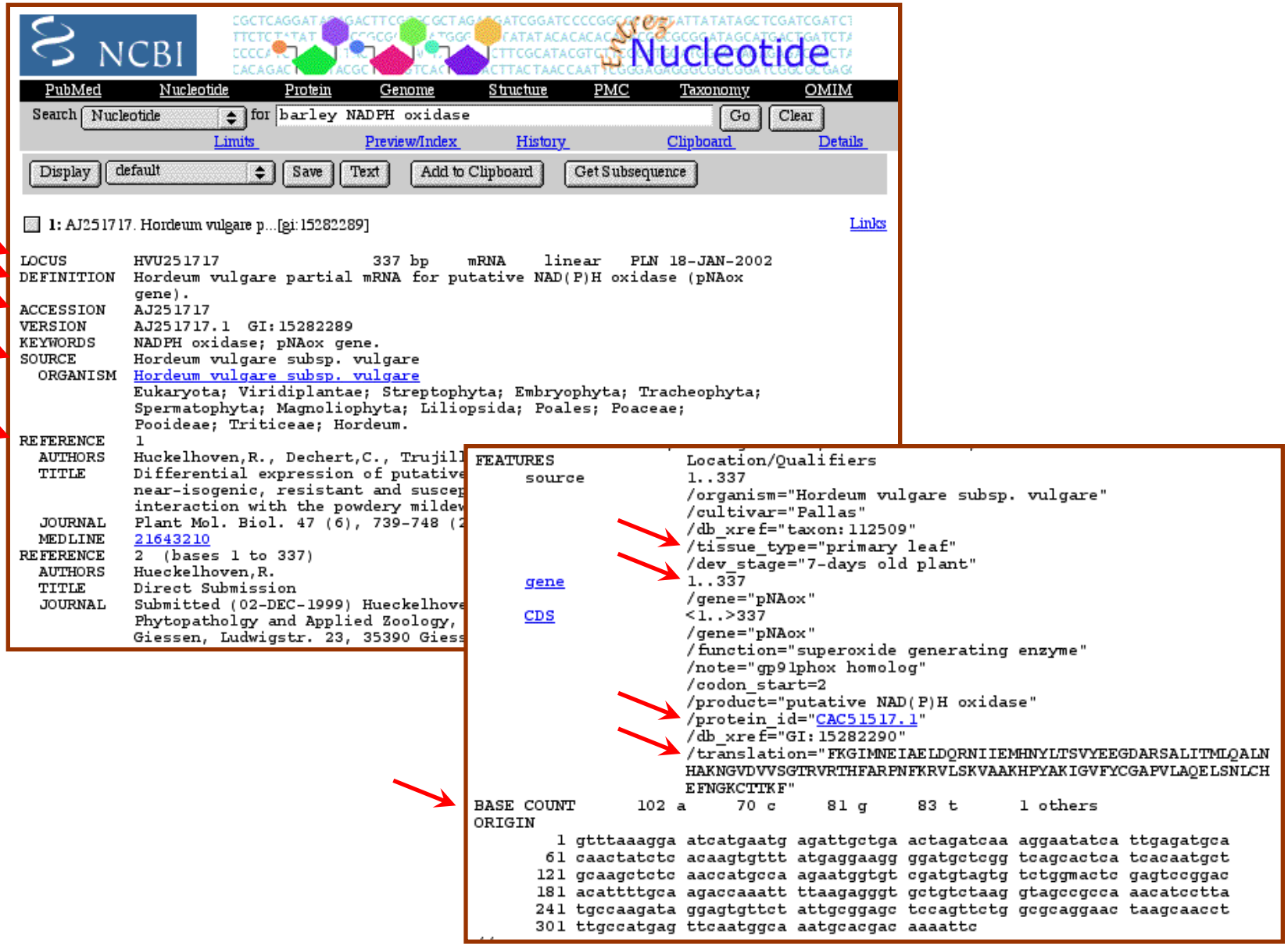
Identifikace záznamu v primárních sekvenčních databázích

- GenBank
- EMBL-Bank (European Nucleotide Archive, ENA)
- DDBJ
- **Přístupový kód (Accession Number)**
- **číslo GI (GenBank Identifier)**

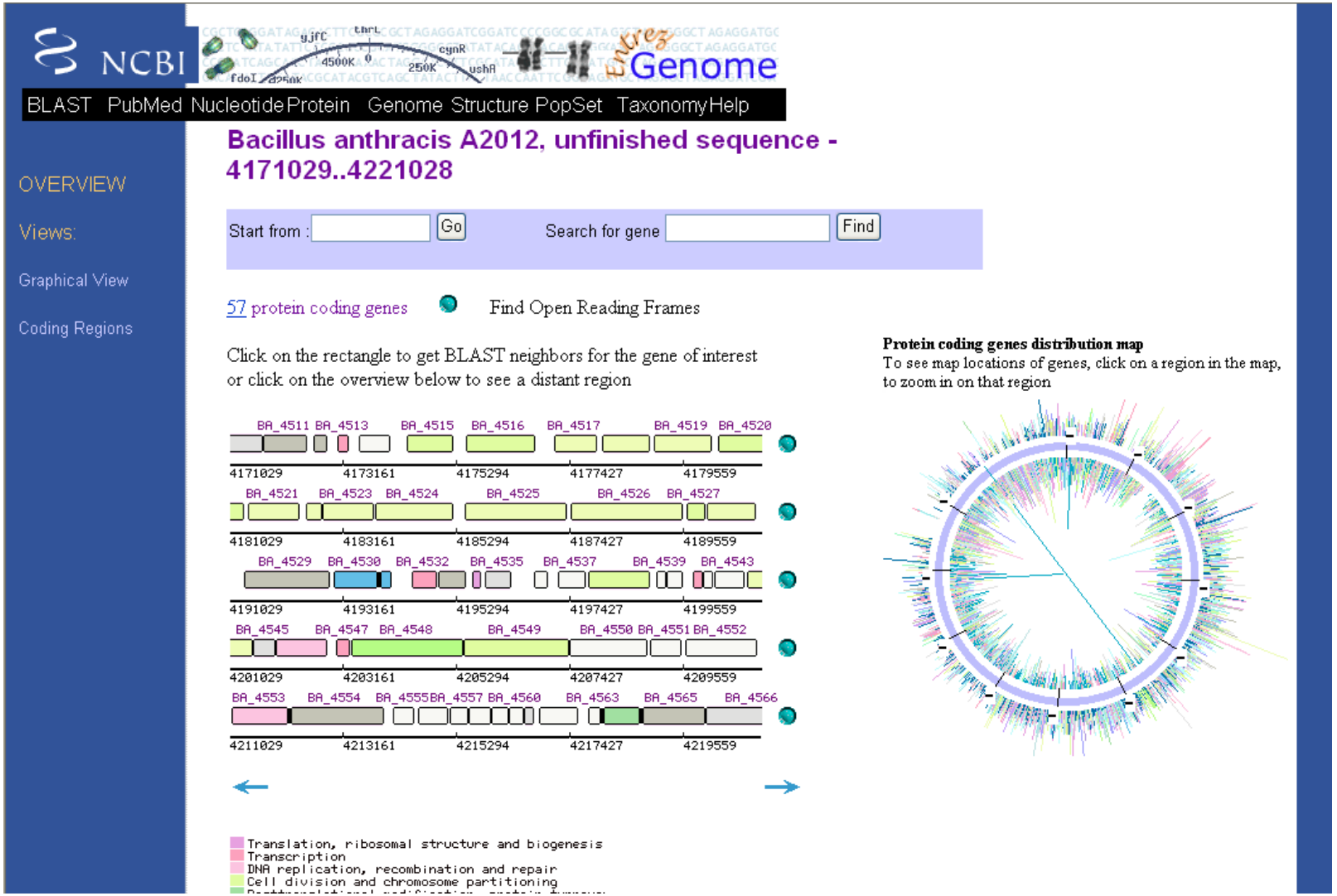
```
LOCUS       AY870395                 553 bp    DNA     linear   BCT 30-JAN-2005
DEFINITION  Macrococcus brunensis strain CCM 4811 60 kDa chaperonin (cpn60)
            gene, partial cds.
ACCESSION   AY870395
VERSION     AY870395.1  GI:58119461
```



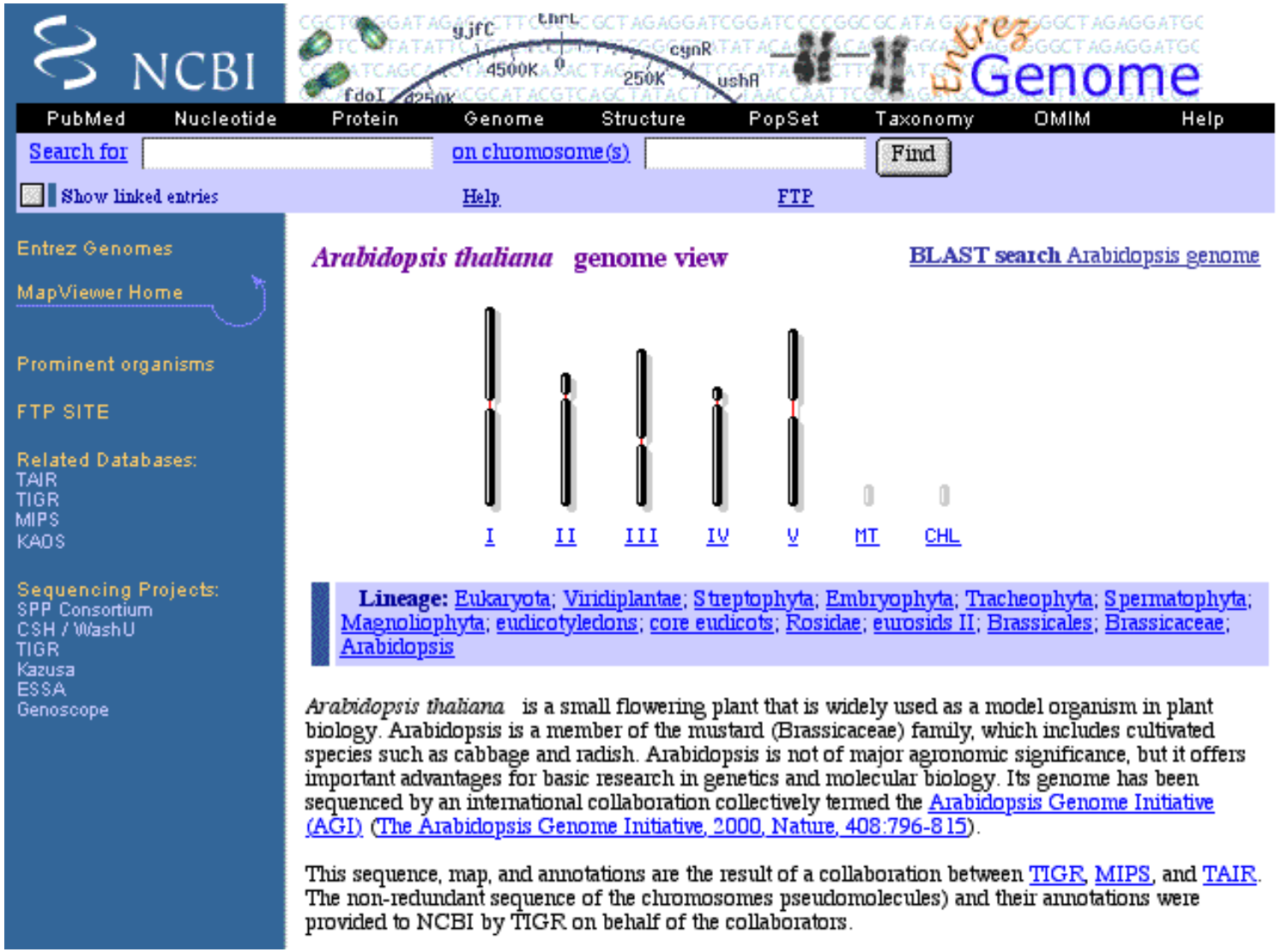
- Struktura zápisu sekvence ve formátu GenBank
- http://www.ncbi.nlm.nih.gov/Genbank/



Genomové databáze v NCBI – prokaryota

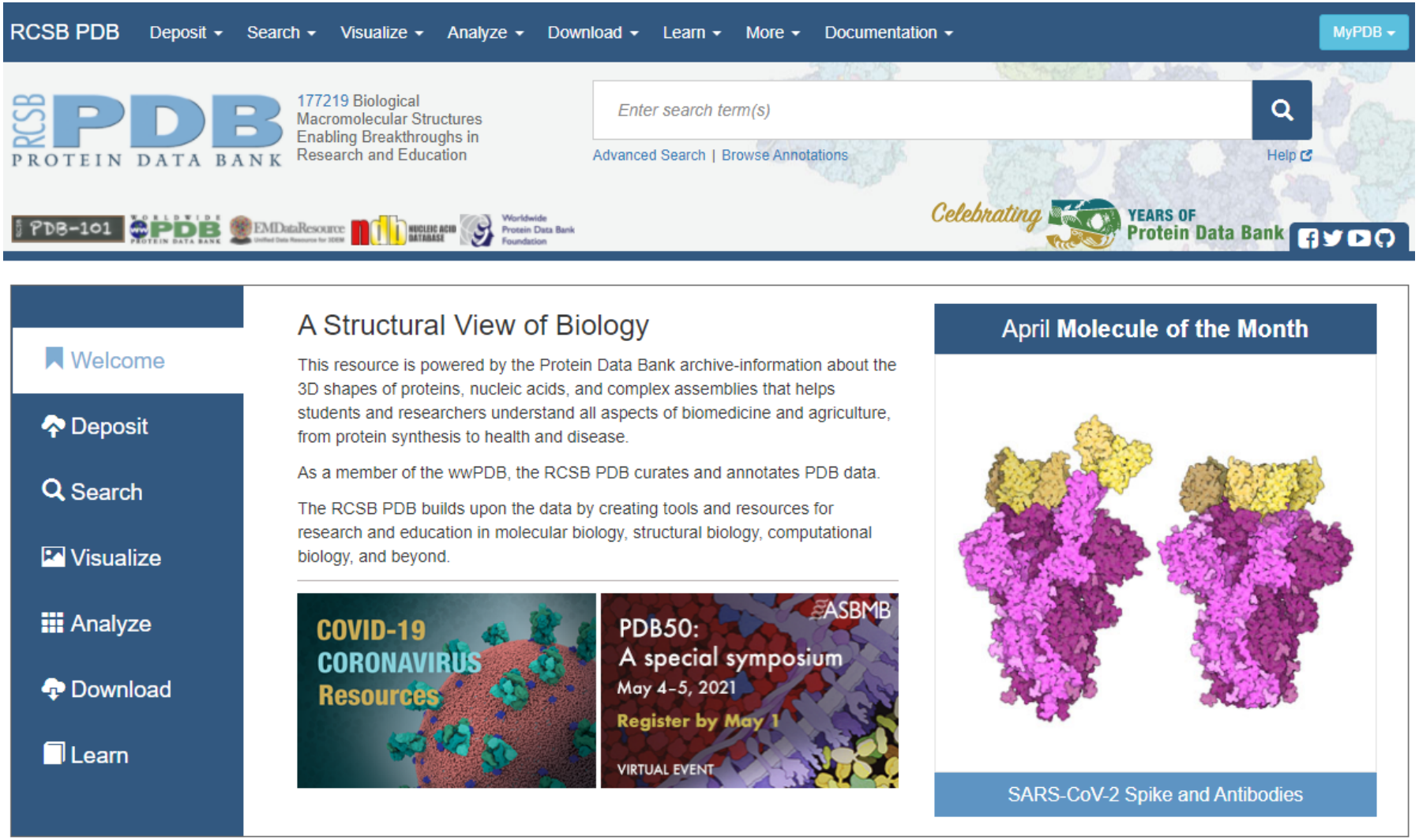


Genomové databáze v NCBI - eukaryota



Důležitou databází spojenou s proteiny je **PDB** (The Protein Databank), která se zabývá archivací a analýzou 3-D **proteinových struktur**.

- PDB http://www.rcsb.org/



Stanovení podobnosti sekvencí

- Textové vyhledávání sekvencí v databázích (podle klíčových slov)
  - Neefektivní - chybí anotace řady sekvencí
  - Nejednotná nomenklatura genů
  - Řada nesouvisejících výsledků
- Prohledávání databází podle podobnosti sekvencí

  - Výpočet lokálního/párového přiložení (alignment) = uspořádání do 2 pod sebou ležících řádků tak, aby identické zbytky ležely pod sebou
  - Výpočet mnohonásobného přiložení (multiple alignment) pro 3 a více sekvencí

Význam sekvenčního přiložení

Použití	Princip
Stanovení podobnosti	Identifikace stejných (podobných) zbytků na základě přiložení
Hledání v databázích	Identifikace podobných sekvencí, charakterizace genů
Identifikace vzorů	Stanovení konzervovaných vzorů, profilů a identifikace funkčních oblastí a domén
Predikce, extrapolace	Klasifikace necharakterizovaných sekvencí do rodin / skupin
Fylogenetická analýza	Rekonstrukce evoluce z ortologních sekvencí
Predikce struktury	Kvalitní přiložení umožňují predikci sekundární struktury jak u RNA, tak proteinů
Sestavení celogenomových sekvencí (assembly)	Využívá techniky přiložení pro vytváření kontigů ze sekvenačních dat
Analýza oligonukleotidů pro PCR	Design primerů a sond, posouzení sekundárních struktur

Terminologie použitá pro srovnávání sekvencí

- **Identita sekvencí** (Sequence identity), podíl identických aminokyselinových nebo nukleotidových zbytků ve stejné pozici
- **Podobnost sekvencí** (Sequence similarity/positivity), podíl identických plus substituovaných zbytků s podobnými chemickými vlastnostmi.
- **Homologie sekvencí** (Sequence homology), termín použitelný pouze u evolučně příbuzných sekvencí, např. stanovení ANI (average nucleotide identity) z celogenomových sekvencí nebo data z DNA-DNA hybridizací

Nástroje pro vyhledávání lokálních podobností sekvencí

Sady programů zahrnujících algoritmy pro vyhledávání podobnosti v dostupných databázích sekvencí bez ohledu na to zdali dotazovaná sekvence je **DNA** nebo **protein**.

- BLAST
- Altschul et al., 1990
- dostupný na serveru NCBI

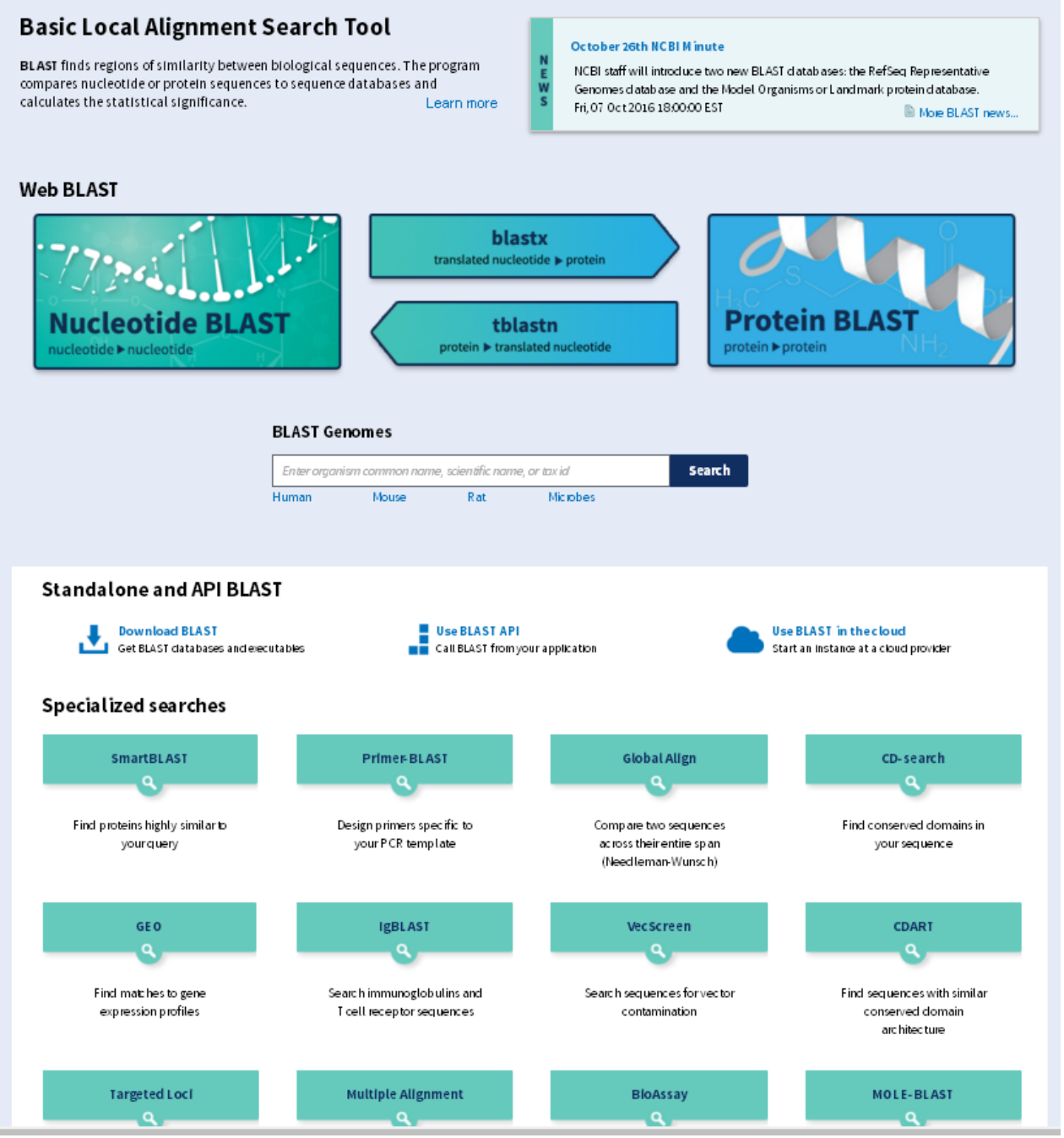
- FASTA
- Lipman a Pearson 1985
- dostupný na serveru EBI

Co je to BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Hledání lokálních podobností
  - Heuristický přístup založený na **Smith-Watermanově** algoritmu
  - Vyhledá nejoptimálnější **přiložení sekvencí**
  - Poskytuje data o statistické významnosti
  - Zobrazuje vzájemně párové přiložení sekvencí
  - Lokalizuje oblasti sekvencí s vysokou podobností a umožňuje zobrazení jejich primární struktury a funkce

Co je to BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Hledání lokálních podobností
  - Heuristický přístup založený na **Smith-Watermanově** algoritmu
  - Vyhledá nejoptimálnější **přiložení sekvencí**
  - Poskytuje data o statistické významnosti
  - Zobrazuje vzájemně párové přiložení sekvencí
  - Lokalizuje oblasti sekvencí s vysokou podobností a umožňuje zobrazení jejich primární struktury a funkce



Výchozí stránka BLAST

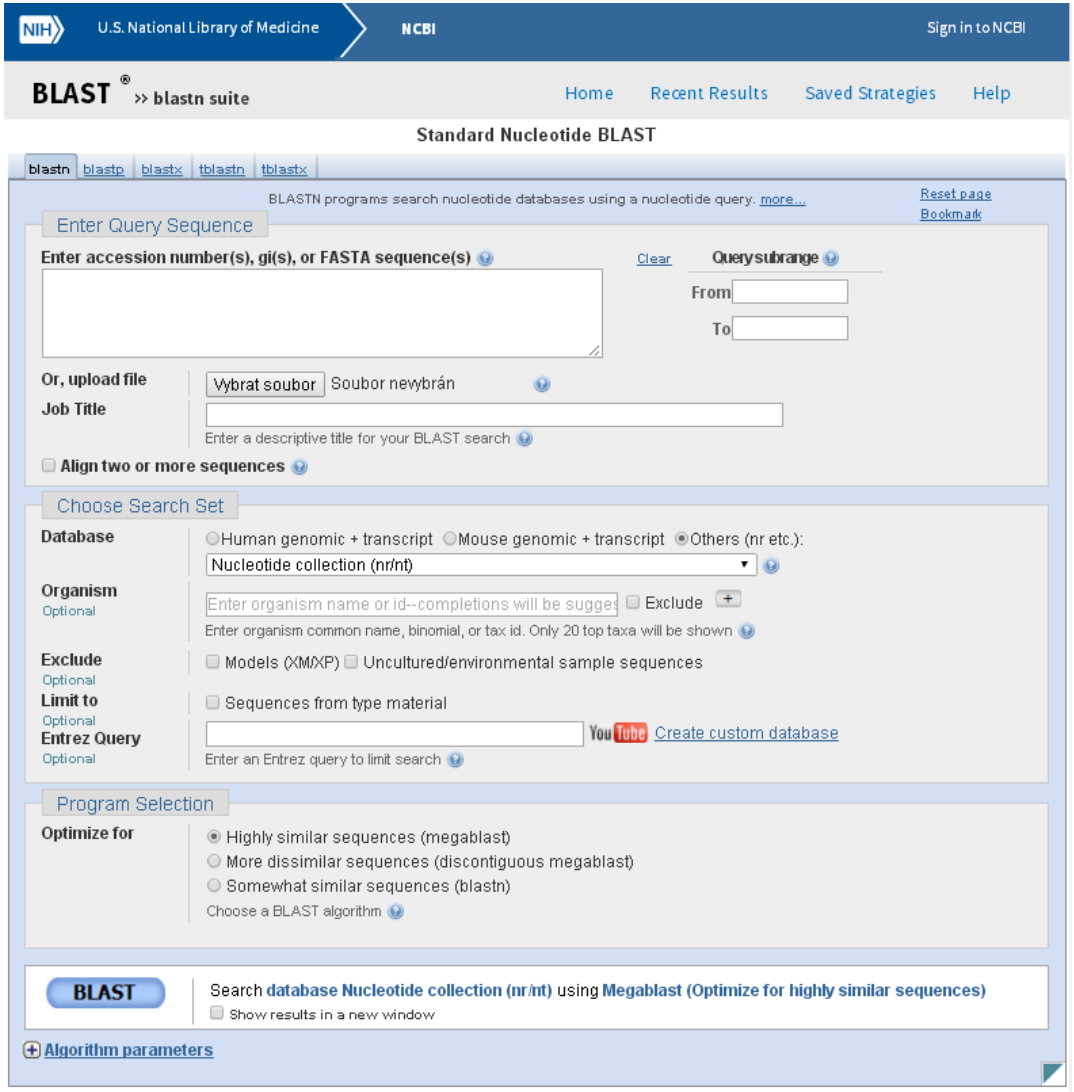
http://www.ncbi.nlm.nih.gov/BLAST

Basic BLAST – výběr programů

Využití jednotlivých programů BLAST

Program	Dotaz	Databáze	Úroveň srovnání	Použití
blastn	DNA	DNA	DNA	Hledání identických sekvencí DNA
blasp	Protein	Protein	Protein	Hledání podobných proteinů
blastx	DNA	Protein	Protein	Hledání genů a podobných proteinů na DNA
tblastn	Protein	DNA	Protein	Hledání genů u necharakterizovaných DNA
tblastx	DNA	DNA	Protein	Studium struktury genů

Uživatelské rozhraní BLAST



- Home Tab: Odkaz na úvodní stránku
- Recent Results Tab: Odkaz na výsledky, které jste získali za posledních 36 hodin
- Saved Strategies Tab: Vyplněné vstupní formuláře pro hledání, které jste uložili do *MyNCBI*
- Help Tab: Katalog s dokumentací a nápovědou

Jak používat BLAST?

- http://www.ncbi.nlm.nih.gov/BLAST

1. Vybrat příslušný BLAST-program (blastn, blastp, blastx, tblastn, tblastx, specializované varianty algoritmů)
2. Vložit sekvenci (DNA nebo protein nebo Accession number)
3. Vybrat databázi, která má být prohledána
4. Upřesnit nastavení parametrů algoritmu
5. Odeslat požadavek na vyhledání

Jak BLAST pracuje?

- Proces zahrnuje 3 kroky
  1. Příprava dotazu
     - rozseká zkoumanou sekvenci na krátké úseky a sestaví z nich vhodnou tabulku
  2. Vyhledává shody v databázi
  3. Rozšiřuje vyhledávání v oblasti nalezených shod, tak aby byla splněna zadaná kritéria

Jak BLAST pracuje?

- Proces zahrnuje 3 kroky
  1. Příprava dotazu
     - rozseká zkoumanou sekvenci na krátké úseky a sestaví z nich vhodnou tabulku
  2. Vyhledává shody v databázi
  3. Rozšiřuje vyhledávání v oblasti nalezených shod, tak aby byla splněna zadaná kritéria

Tvorba dotazu pro nukleotidové sekvence

Dotaz: **GTACTGGACATGGACCCTACAGGAA**

~~GTACTGGACAT~~

Word size = 11

TACTGGACATG

minimální velikost = 7

blastn default = 11

megablast default = 28 (16 – 256)

tabulka se všemi slovy dotazu

ACTGGACATGG

CTGGACATGGA

TGGACATGGAC

GGACATGGACC

GACATGGACCC

ACATGGACCCT

. . . . . .

přiložení sekvencí, které BLAST **může** nalézt

```
1 AATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAG
  ||||| |||||||||||||||| || ||||||||||||||
1 AATGGAAAAGACTACTGGATCATCAAAAACTCCTGGGGAG
```

sekvence obsahují definovanou shodu slova

přiložení sekvencí, které BLAST **nemůže** nalézt

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
     |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

Tvorba dotazu pro proteinové sekvence

Dotaz: **GTQITVEDLFYNIATRRKALKN**

GTQ Word size = 3

Velikost slova může být 2, 3 (default = 6)

TQI

tabulka se všemi slovy dotazu

QIT

Sousedící slova

ITV ⟶ LTV, MTV, ISV, LSV, etc.

TVE

VED

EDL

DLF

. . .

Minimální požadavek pro shodu

ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT přesná shoda slova

1 nalezená shoda

- Nucleotidový BLAST vyžaduje jednu přesnou shodu
- Proteinový BLAST vyžaduje dvě sousedící shody v úseku 40 aa

GTQITVEDLFYNI

SEI **YYN** sousedící slova

2 nalezené shody

Hodnocení výsledků přiložení

- K posouzení významnosti shody nalezených úseků se používá numerická hodnota označovaná jako **skóre sekvenčního přiložení** (***S***)
  - Hrubé skóre (Raw score)
    - Suma skóre pro identity plus substituce minus penalizace mezer
  - Normalizované skóre (Normalised score)
    - Nezávislé na systému, umožňuje srovnání různých přiložení

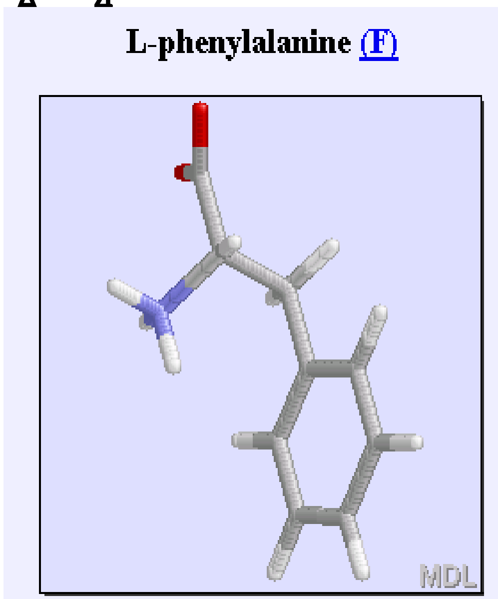
Typy matic pro výpočet skóre

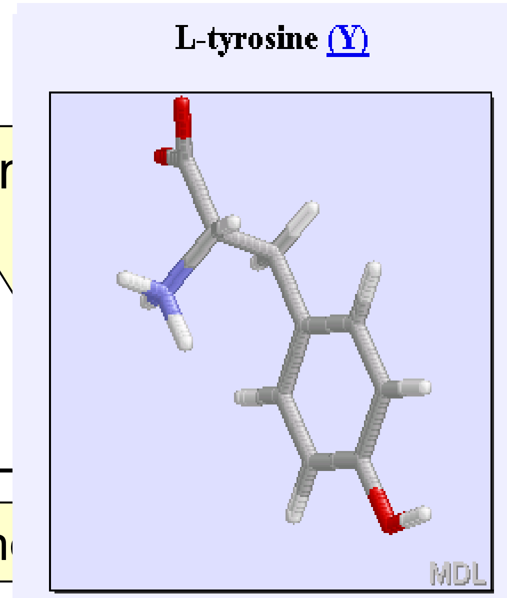
- Matice identity
  - Především pro nukleotidové sekvence
  - Neschopné transformovat na jiné zbytky
  - Pro přiložení velmi podobných sekvencí
- Matice podobnosti
  - Používané u proteinových sekvencí
  - Vyjadřují biochemické/biologické vlastnosti aminokyselin
  - Vyšší účinnost při srovnávání sekvencí

Matice BLOSUM

- **BL**ocks **S**ubstitution **M**atrix
- Matice BLOSUM jsou sestaveny na základě **analýzy mnohonásobných přiložení evolučně příbuzných proteinů v databázi BLOCKS**
- BLOSUM-x používá analýzu pouze těch proteinů, které mají **alespoň x %** identitu
  - BLOSUM45, BLOSUM50, BLOSUM62, BLOSUM80
- Matice BLOSUM jsou vhodné pro hledání v databázích
- Změny probíhající během dlouhodobé evoluce nejsou často vhodné pro výpočty a sledování malých recentních změn

Příklad matice BLOSUM62





Běžné ar… …ší významnost

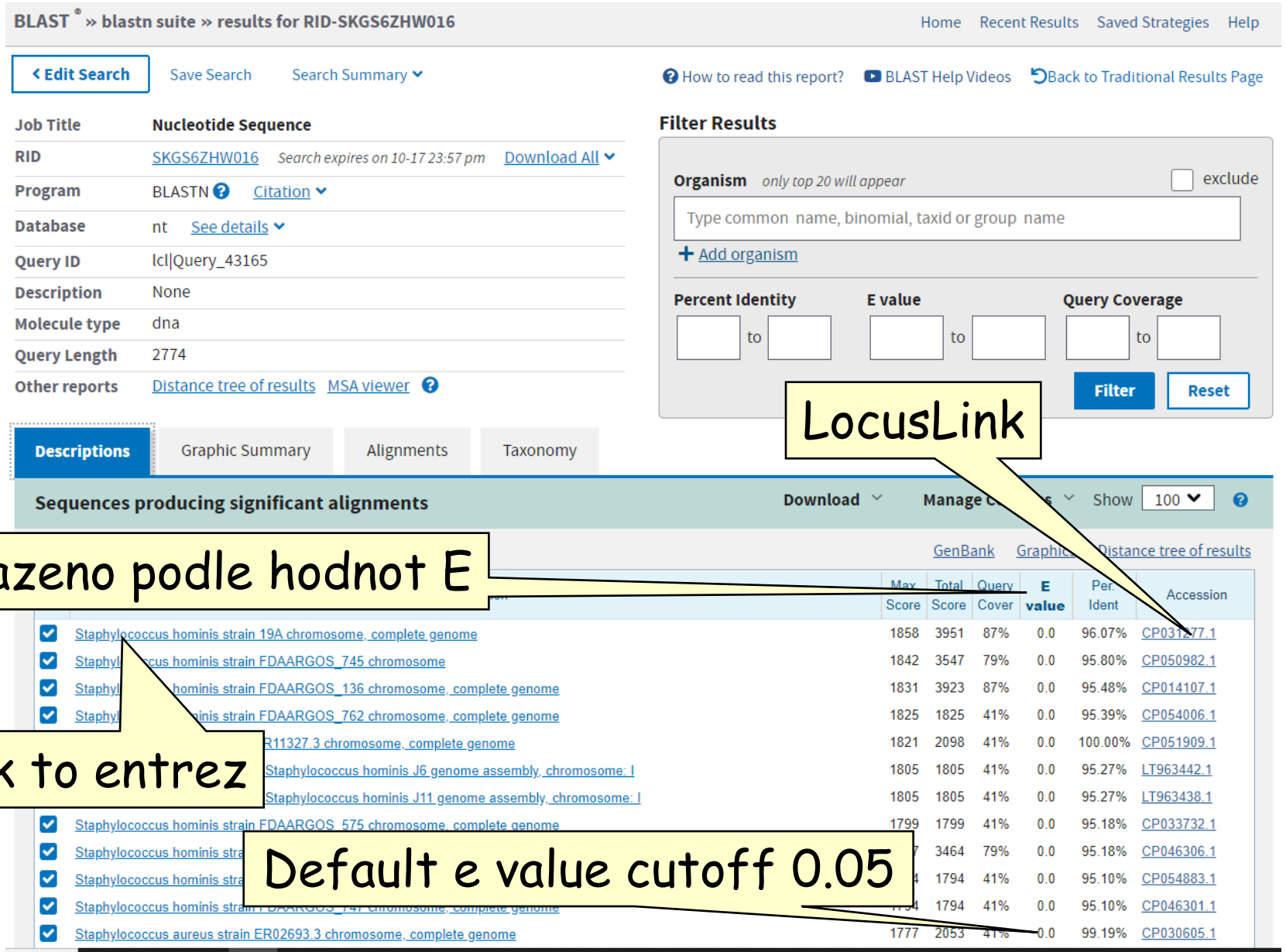
Vzácn… …šší významnost

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
							5														
							-2	6													
							0	-2													
							-3	-4	-												
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2		-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1				-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
													-1	-2	-1	1	5				
													-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2		-1	-2	-2	0	-3	-1	4	
X	0														-2	0	0	-2	-1	-1	-1

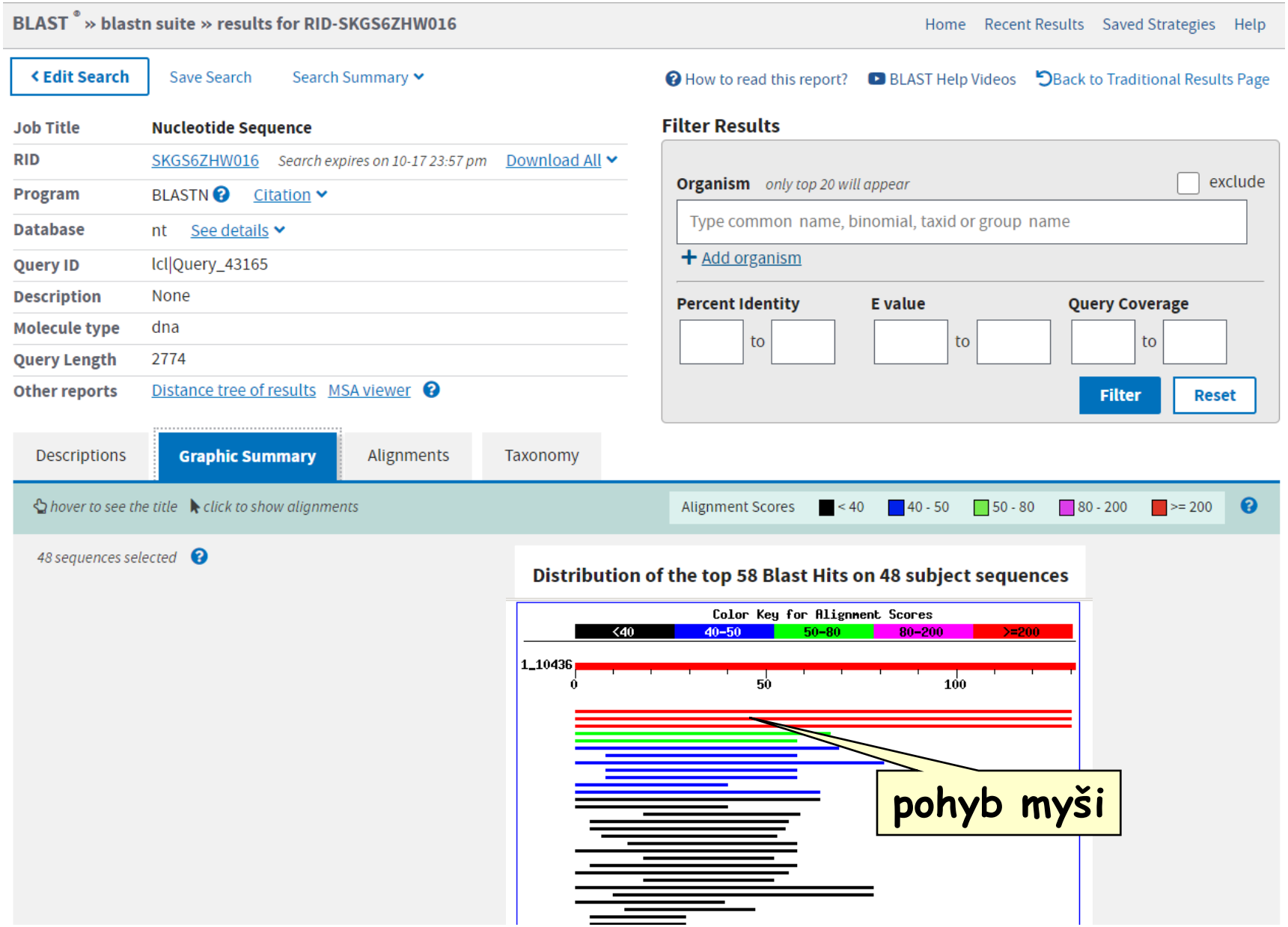
Negativní pro málo pravděpodobné substituce

Pozitivní pro více pravděpodobné substituce

BLAST –výstup, výsledky



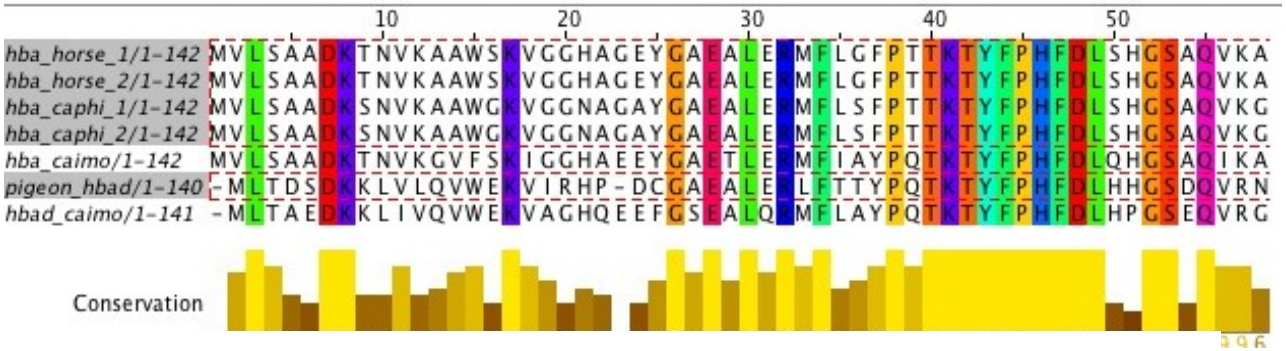
BLAST –výstup, výsledky

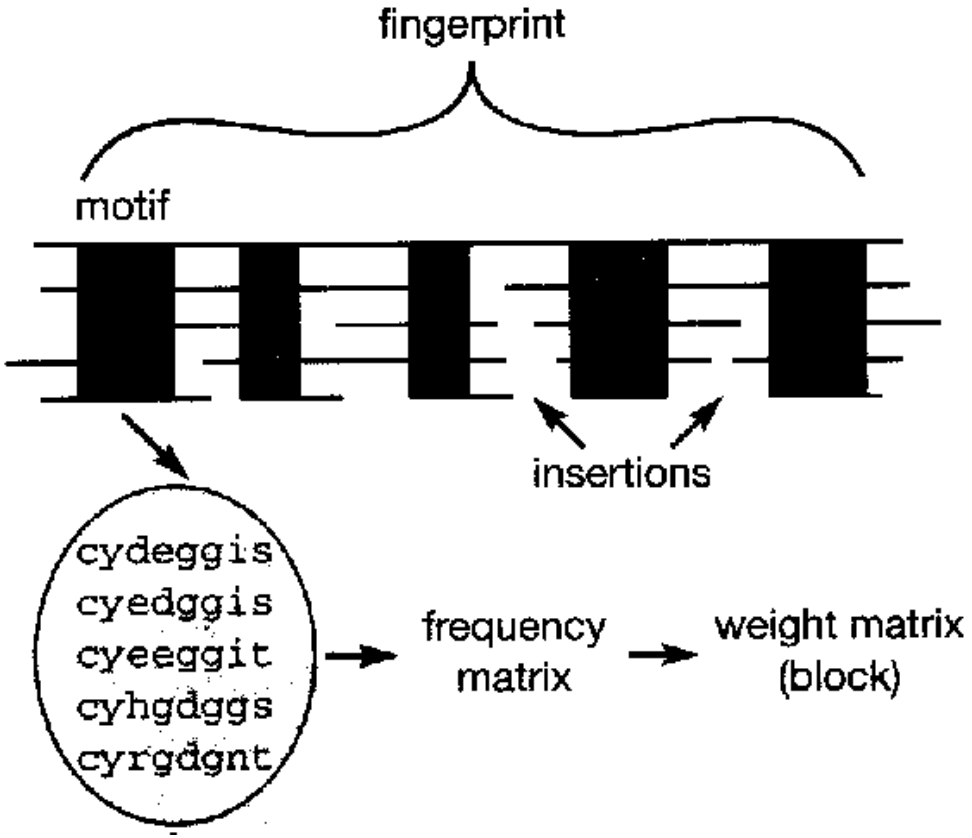


Lokální versus mnohonásobné srovnání

- Dosud jsme srovnávali pouze **dvě sekvence navzájem**
- Podobnosti mezi dvěma sekvencemi se stávají významnými, pokud se vyskytují i u dalších sekvencí
- Mnohonásobné přiložení sekvencí je srovnání tří a více sekvencí nukleových kyselin nebo proteinů s mezerami vloženými do sekvencí tak, že úseky sekvencí s úplnou nebo částečnou homologií jsou seřazeny nad sebou ve stejném sloupci
- Může identifikovat podobnosti a identifikovat **konzervativní motivy**, které nejsme schopni identifikovat lokálním srovnáním
- Využití mnohonásobných sekvenčních přiložení:
  - Analýza struktury genů (identifikace konzervativních domén, konzervativních regulačních oblastí)
  - Analýza příbuznosti, konstrukce fylogenetických stromů z ortologních genů
  - Assembly sekvenačních dat
  - Identifikace konzervativních nebo jedinečných cílů pro diagnostiku, identifikační databáze
  - Klasifikace proteinů, databáze sekvenčních motivů, domén (PROSITE, Pfam, PRINTS, ProDom, SMART, Blocks) integrované databáze (InterPro, CDD search)

Příklad analýzy mnohonásobného přiložení





Přínos 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?
  - Funkce – Jakou funkci mají kódované proteiny?
  - Jakých procesů se účastní – V jakých metabolických drahách?
  - Regulace – Oblasti důležité pro expresi genů
  - Evidence – Experimentální důkaz genu / omiky (omics)
    - 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) a jiné regulační 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 rámec je využíván
- Jaké jsou koordináty genu
- Kde jsou hranice exonů a intronů
- Kde se nacházejí regulační sekvence
- Jaká je modulární struktura genomů

Sekvenování RNA pak umožňuje popsat expresi genů a její regulaci

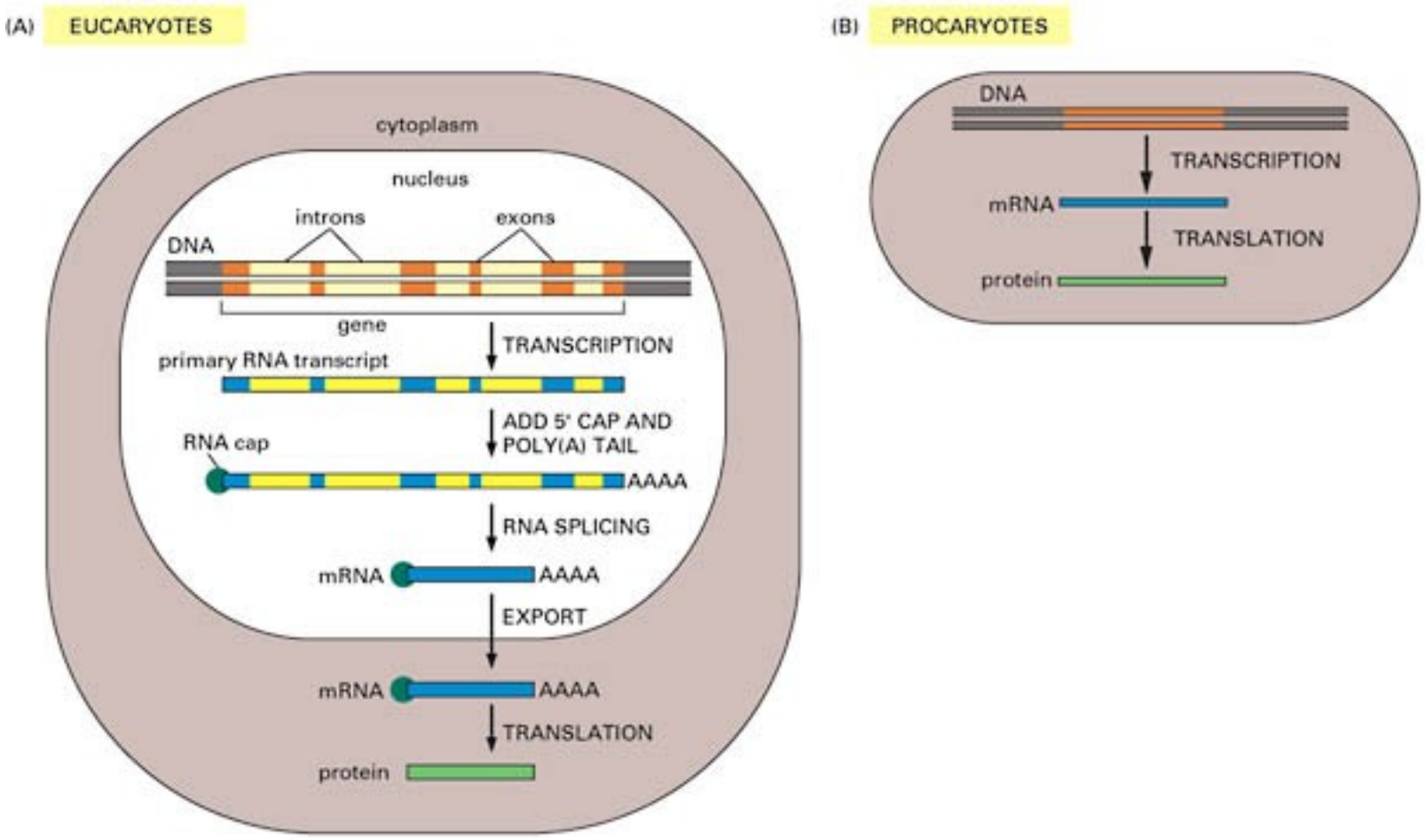
Přístupy pro hledání genů

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

Příklady velikostí genomů

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

Prokaryotický versus eukaryotický gen



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%)
  - Konzervovanost 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

3. Predikce kódující oblasti na základě hledání (*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
  - Analýza frekvencí

Vyhledání otevřených čtecích rámců (http://www.ncbi.nlm.nih.gov/projects/gorf/)

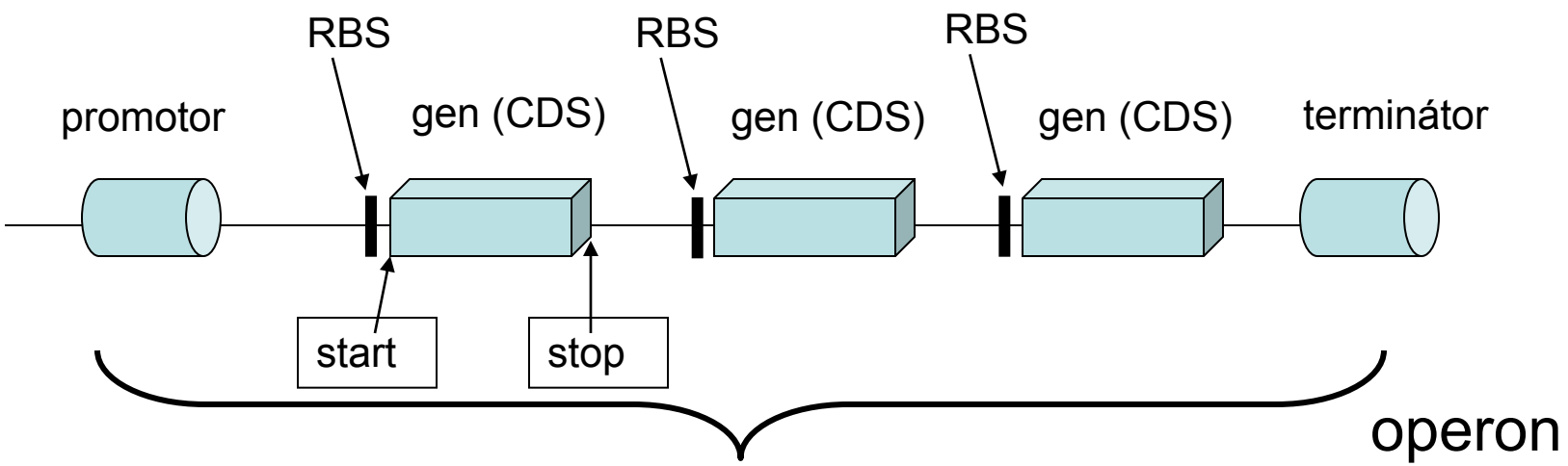
3. Predikce kódující oblasti na základě hledání (*ab initio*)

- 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
  - Analýza frekvencí

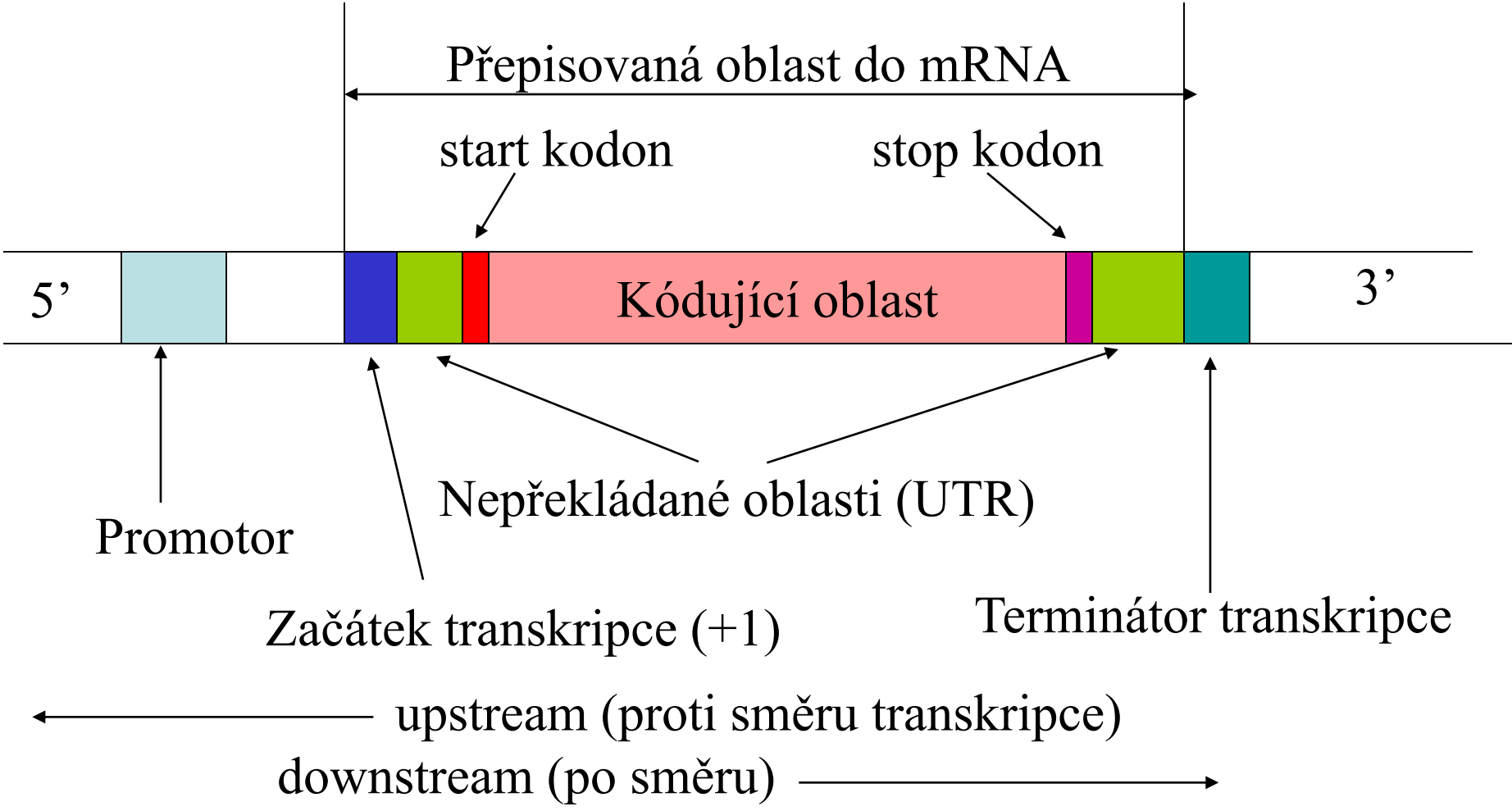
Klíčové signály pro odhalení genů

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

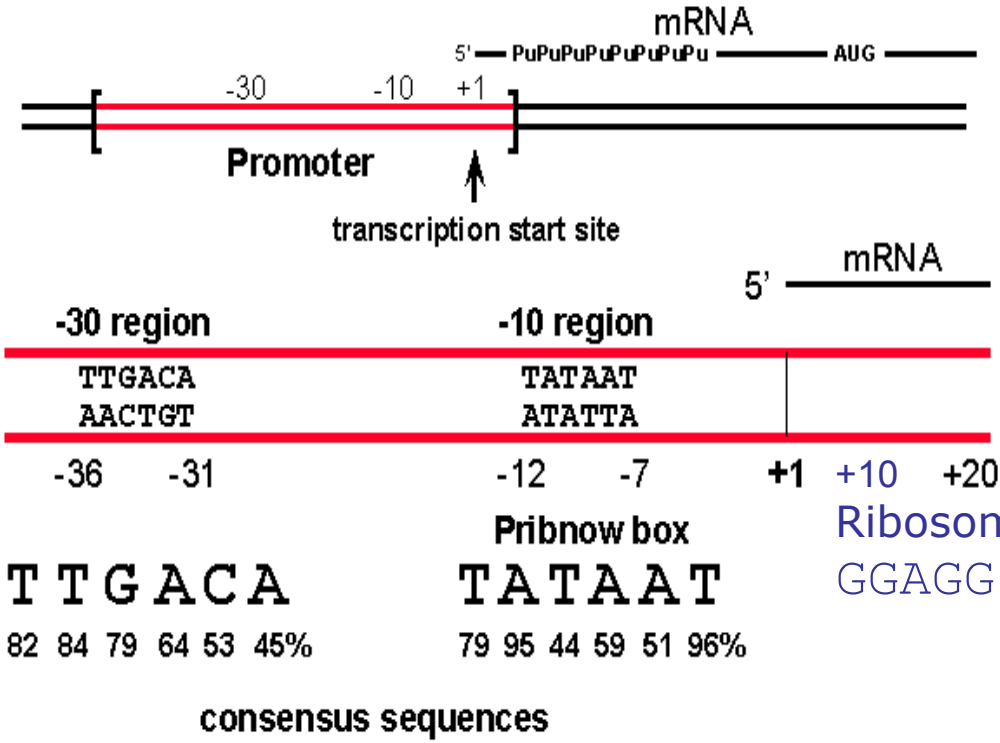
Struktura prokaryotické transkripční jednotky



Struktura prokaryotického genu



Konzervativní struktury v promotoru prokaryot

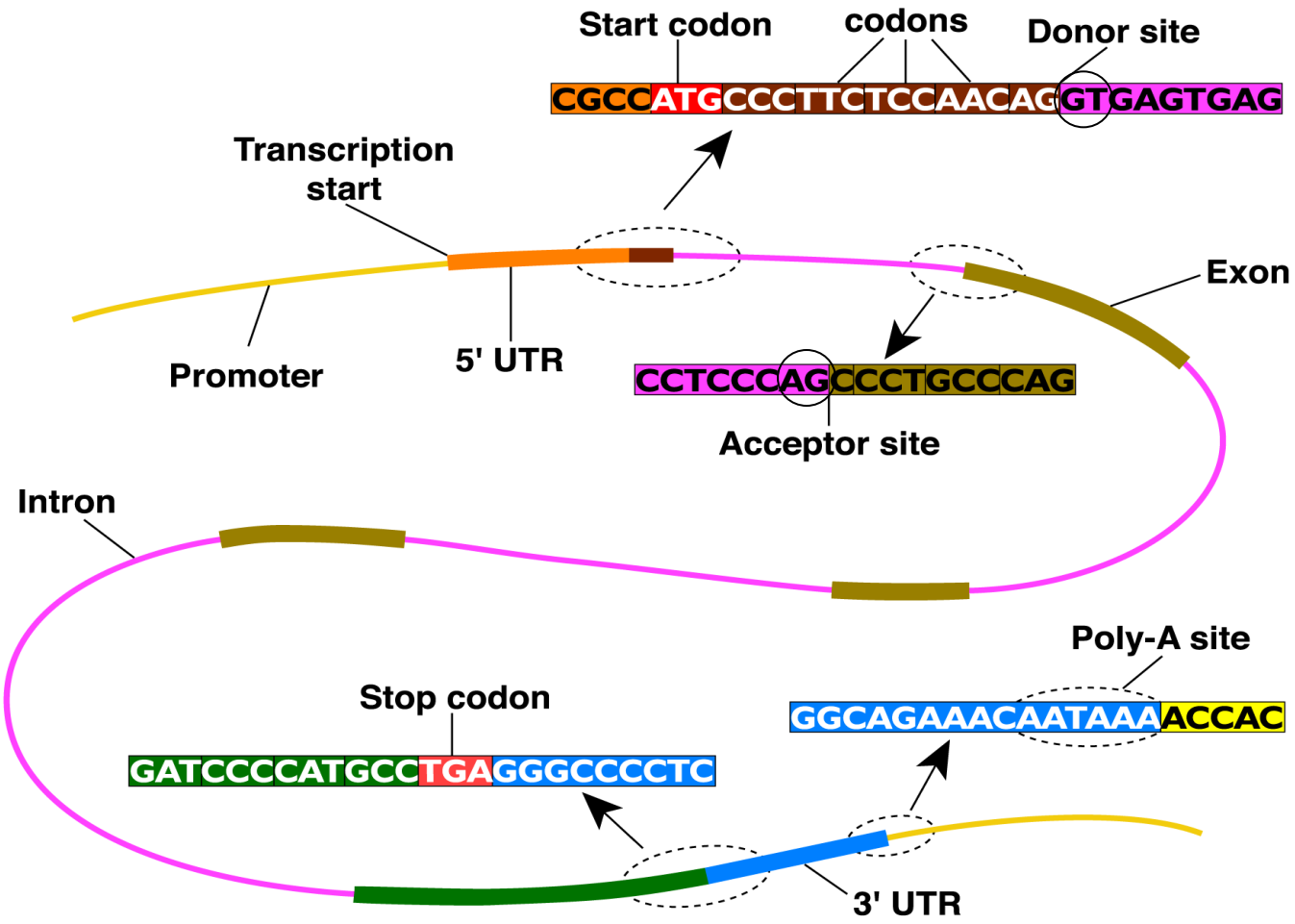


Ribosomal binding site (RBS)

GGAGG ATG

protein

Signály – senzory ve struktuře eukaryotického genu



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

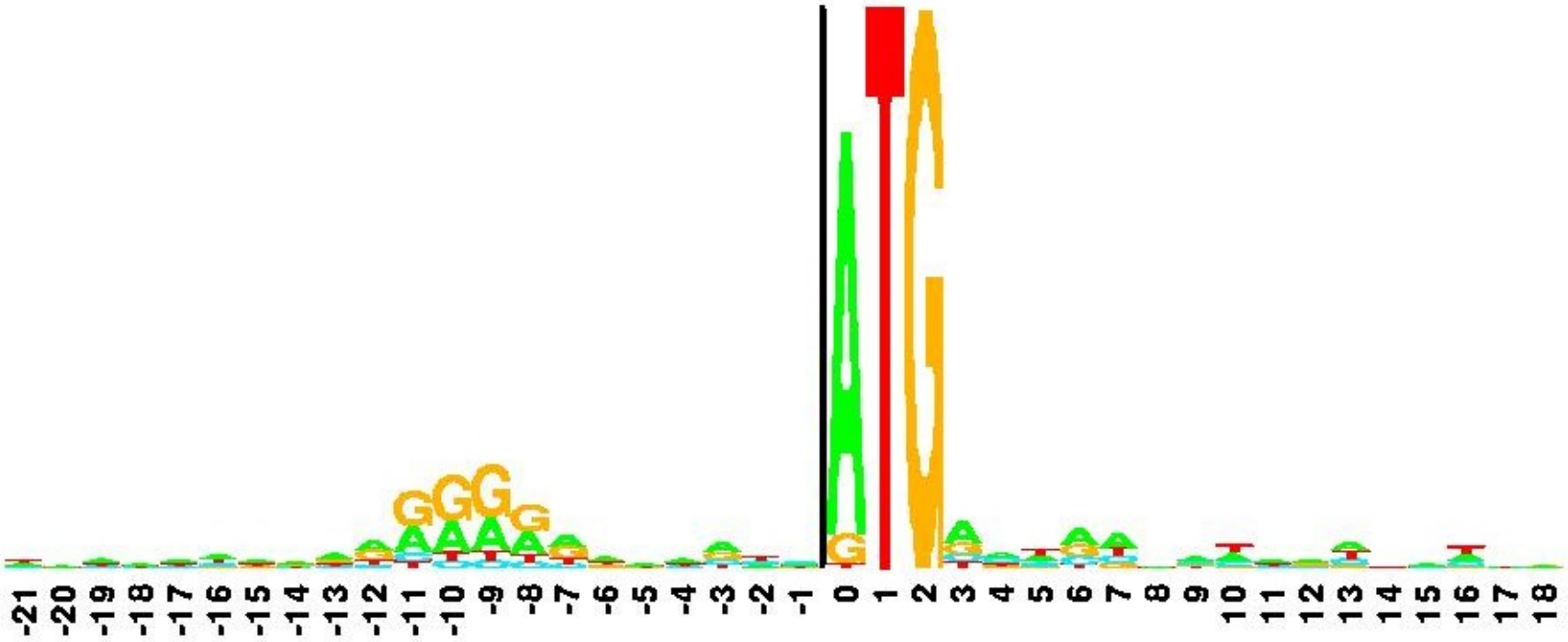
- Vyjadřuje frekvenci každé báze v každé pozici příslušné sekvence

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

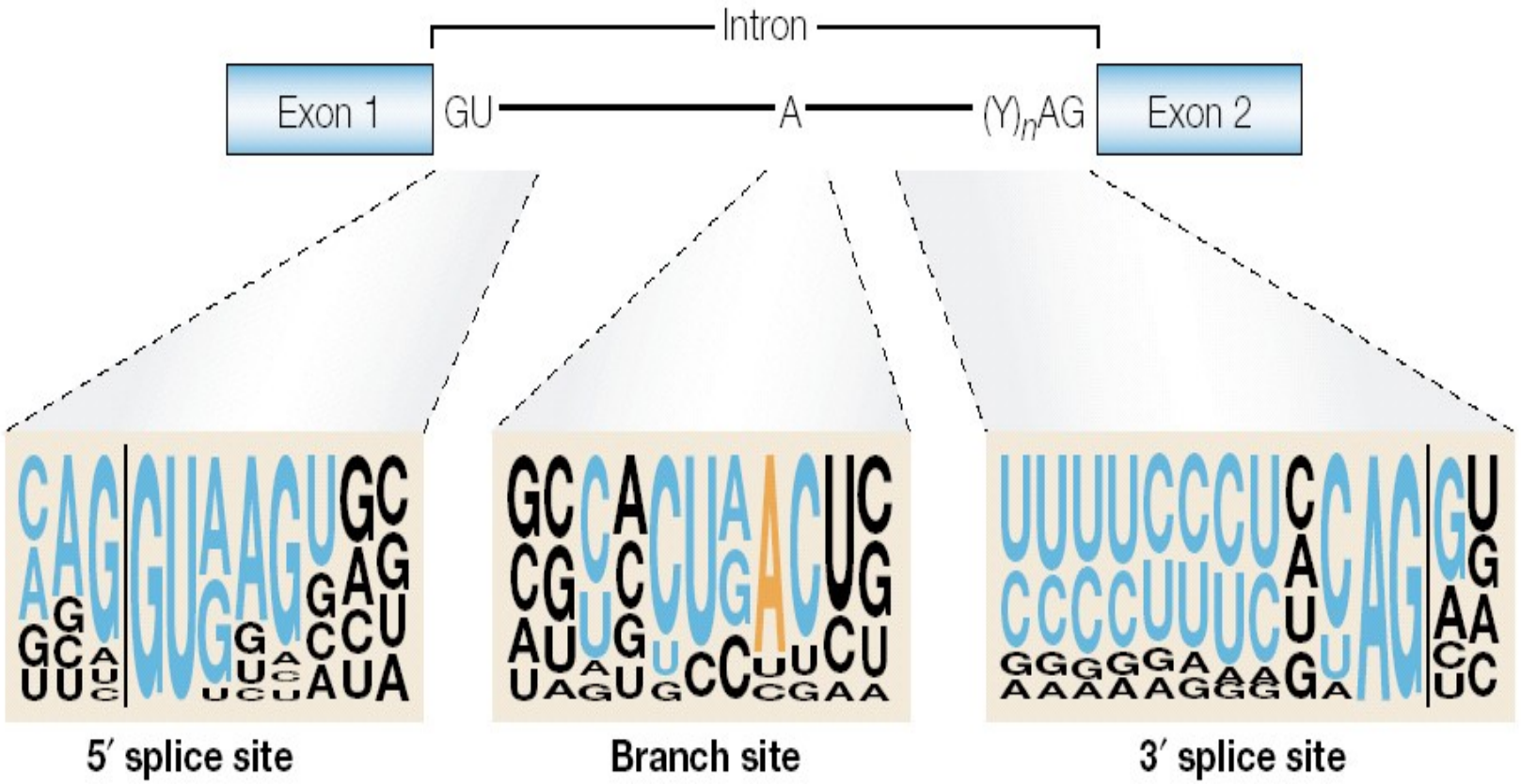
	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

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

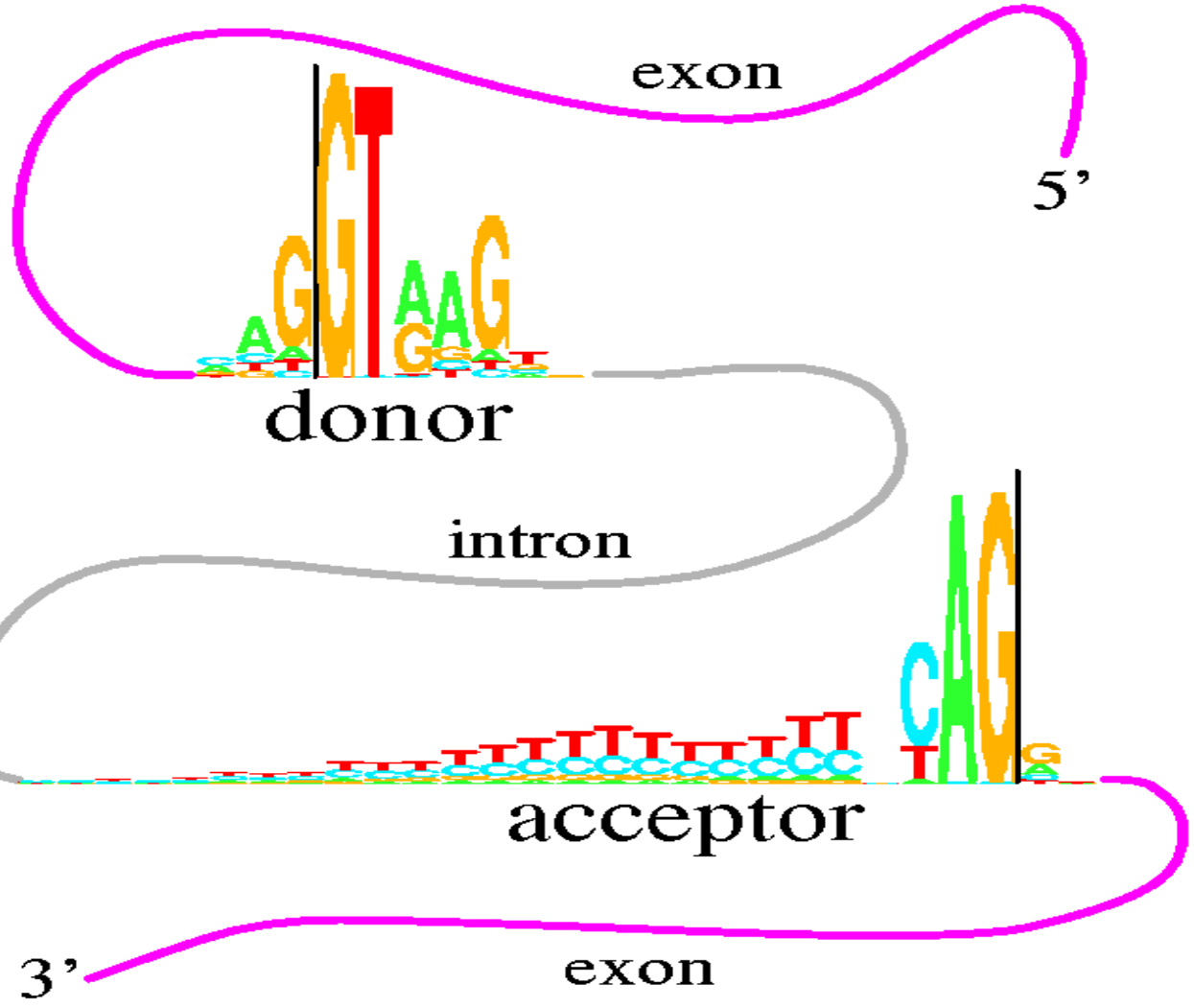
Vazebné místo pro ribozóm (RBS) a iniciační kodon ATG u *E. coli*



Predikce míst sestřihu



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)
  - periodicita nukleotidů

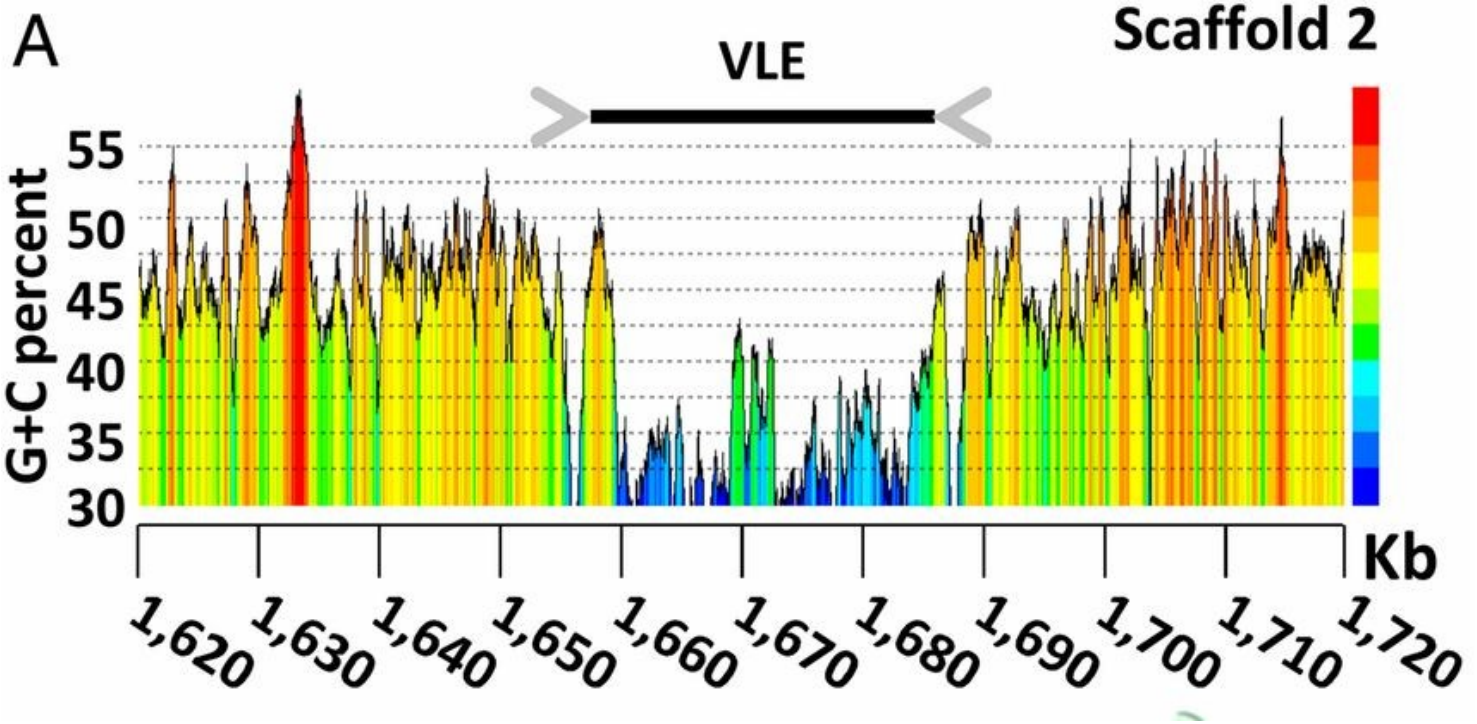
Testování exonů – využití kodonů

AA	codon	/1000	frac
Ser	TCG	4.31	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	6.33	0.11
Pro	CCA	17.10	0.28
Pro	CCT	18.31	0.30
Pro	CCC	18.42	0.31

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

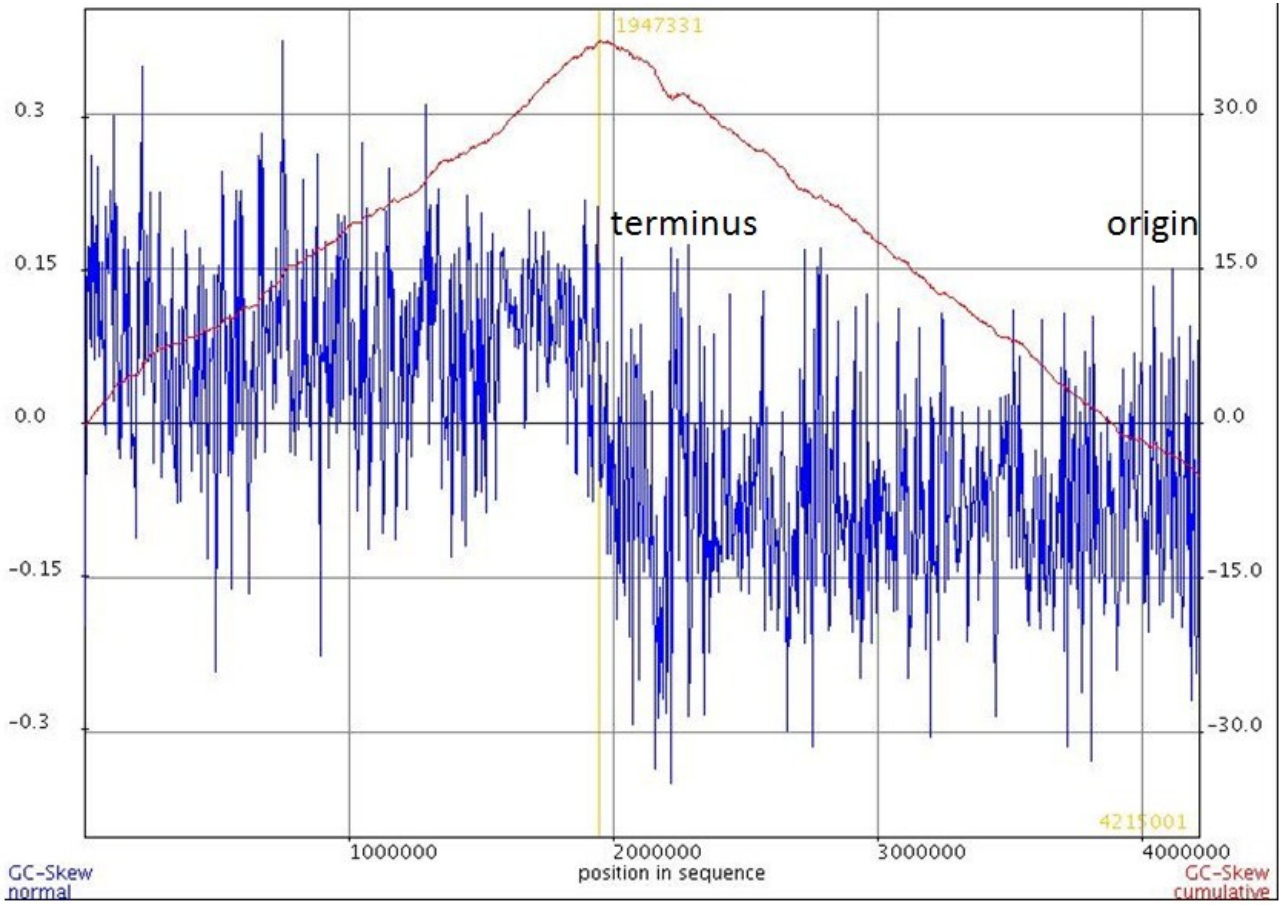
Codon usage database: http://www.kazusa.or.jp/codon/

Obsah G+C – příklad využití pro identifikaci mobilního elementu



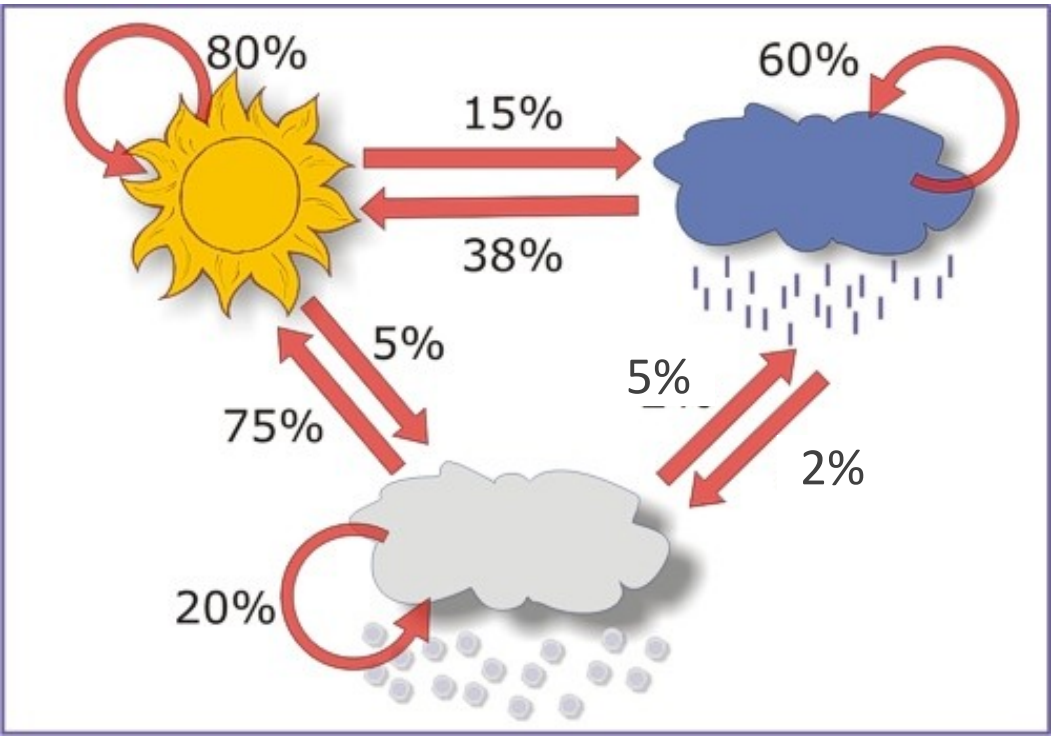
Odlišný obsah G+C indikuje horizontální přenos

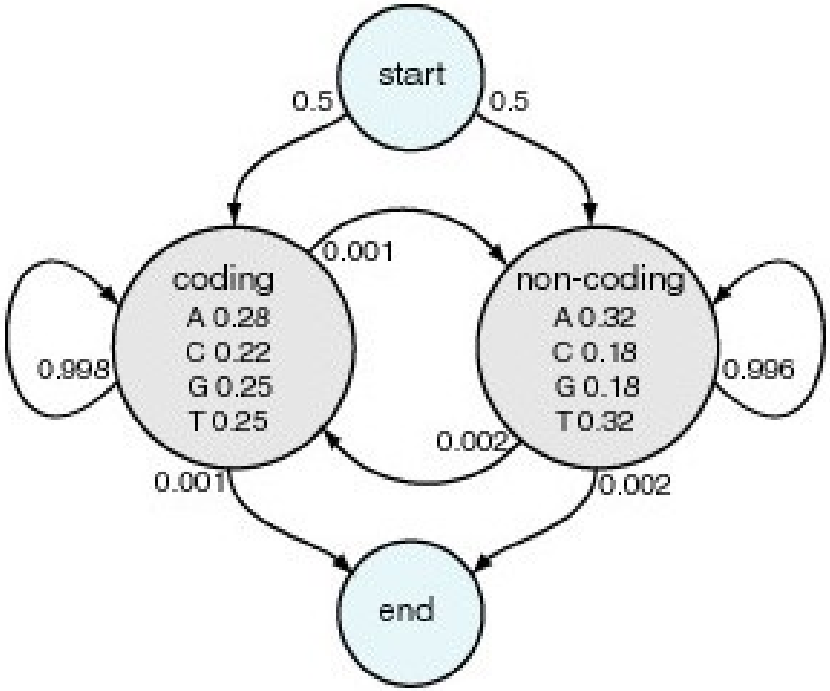
GC skew – příklad využití pro identifikaci počátku replikace



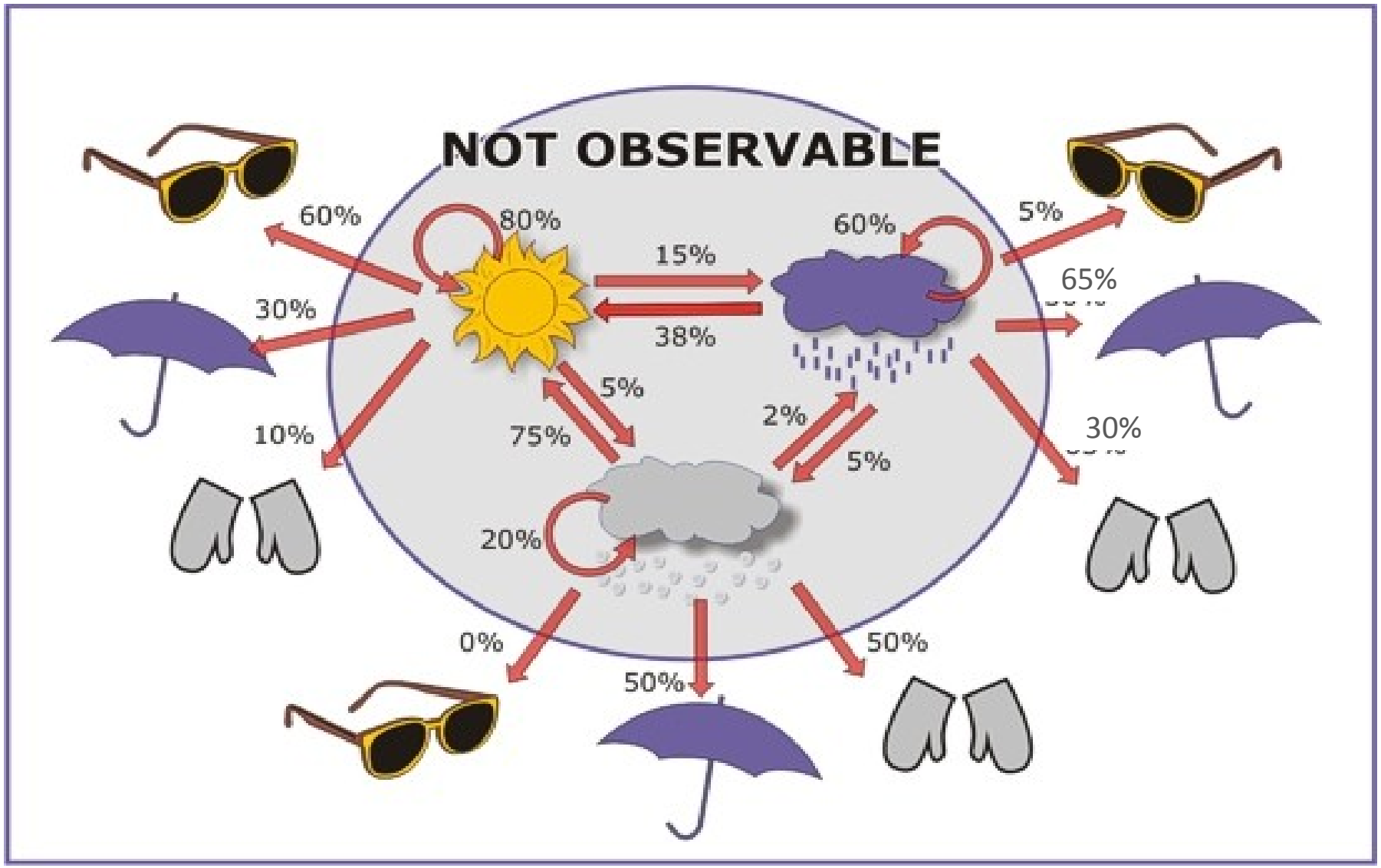
Markovovy modely

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

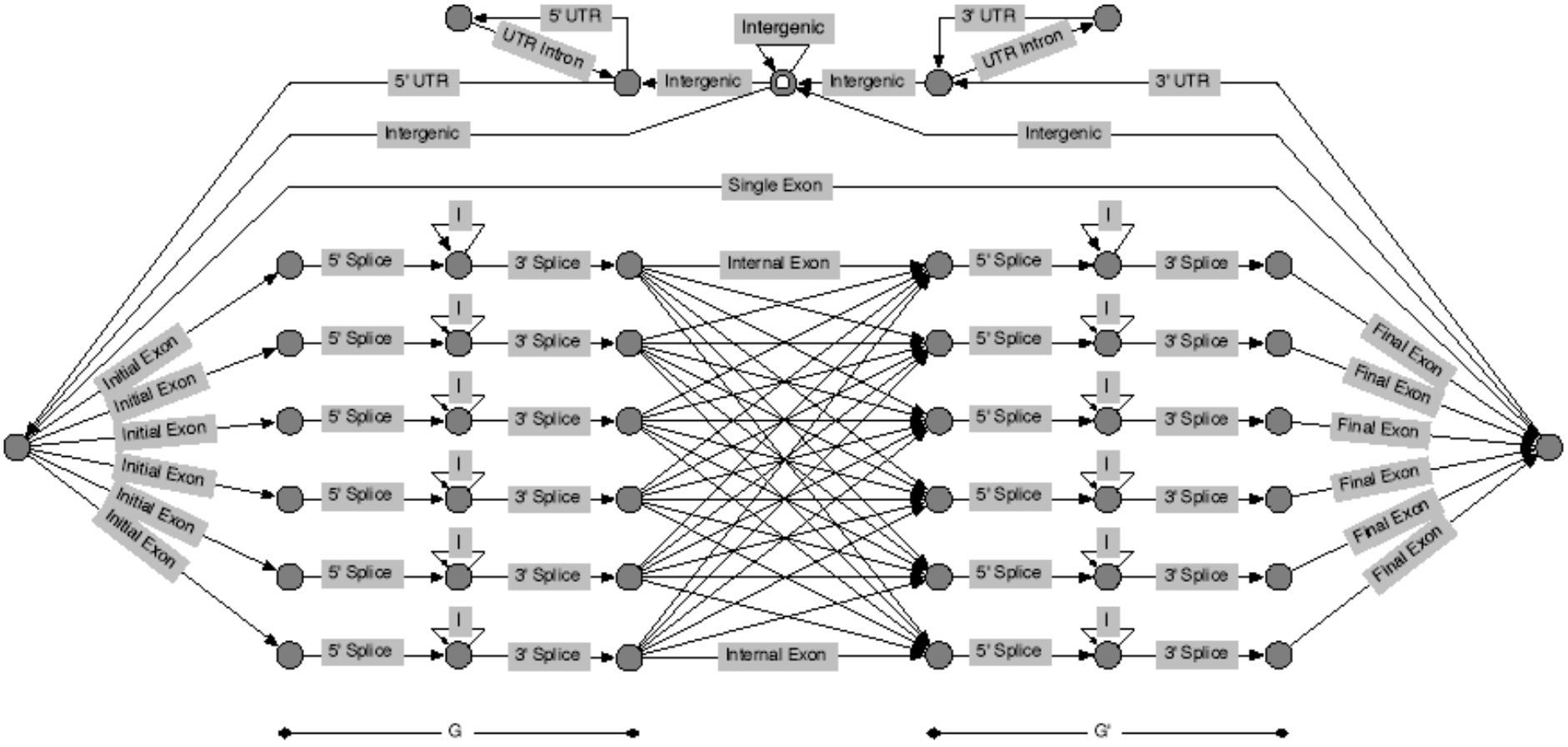




Hidden Markov Models (HMM)



Příklad komplexního algoritmu se skrytými Markovovy modely (HMM)



RAST Rapid Annotation using Subsystem Technology version 2.0

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org. http://rast.theseed.org/FIG/rast.cgi

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	View closest neighbors
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

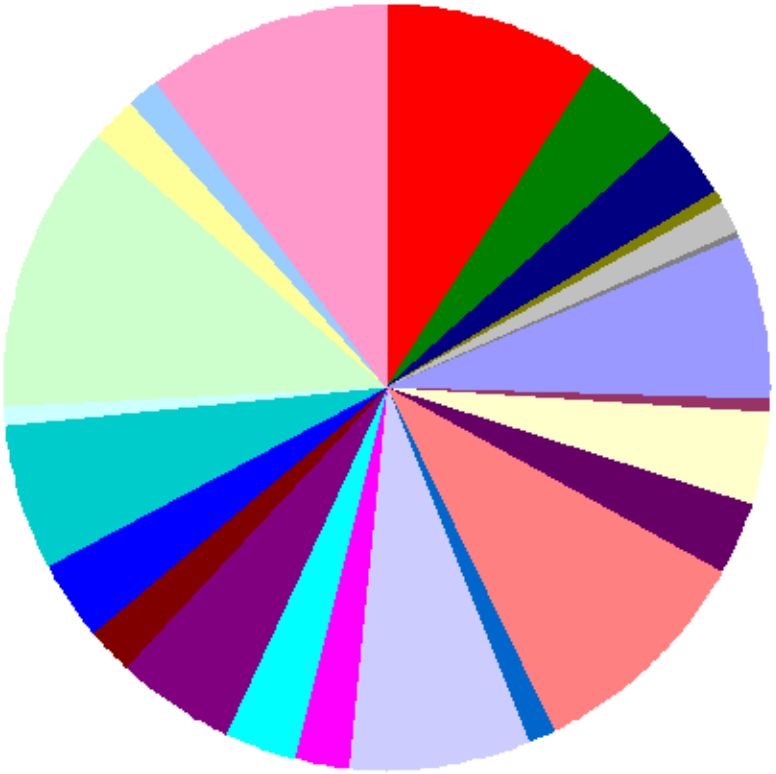
- RAST rozdělí anotované geny do jednotlivých funkčních kategorií, ty zahrnují další podkategorie – např. geny zapojené do jednotlivých metabolických drah

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