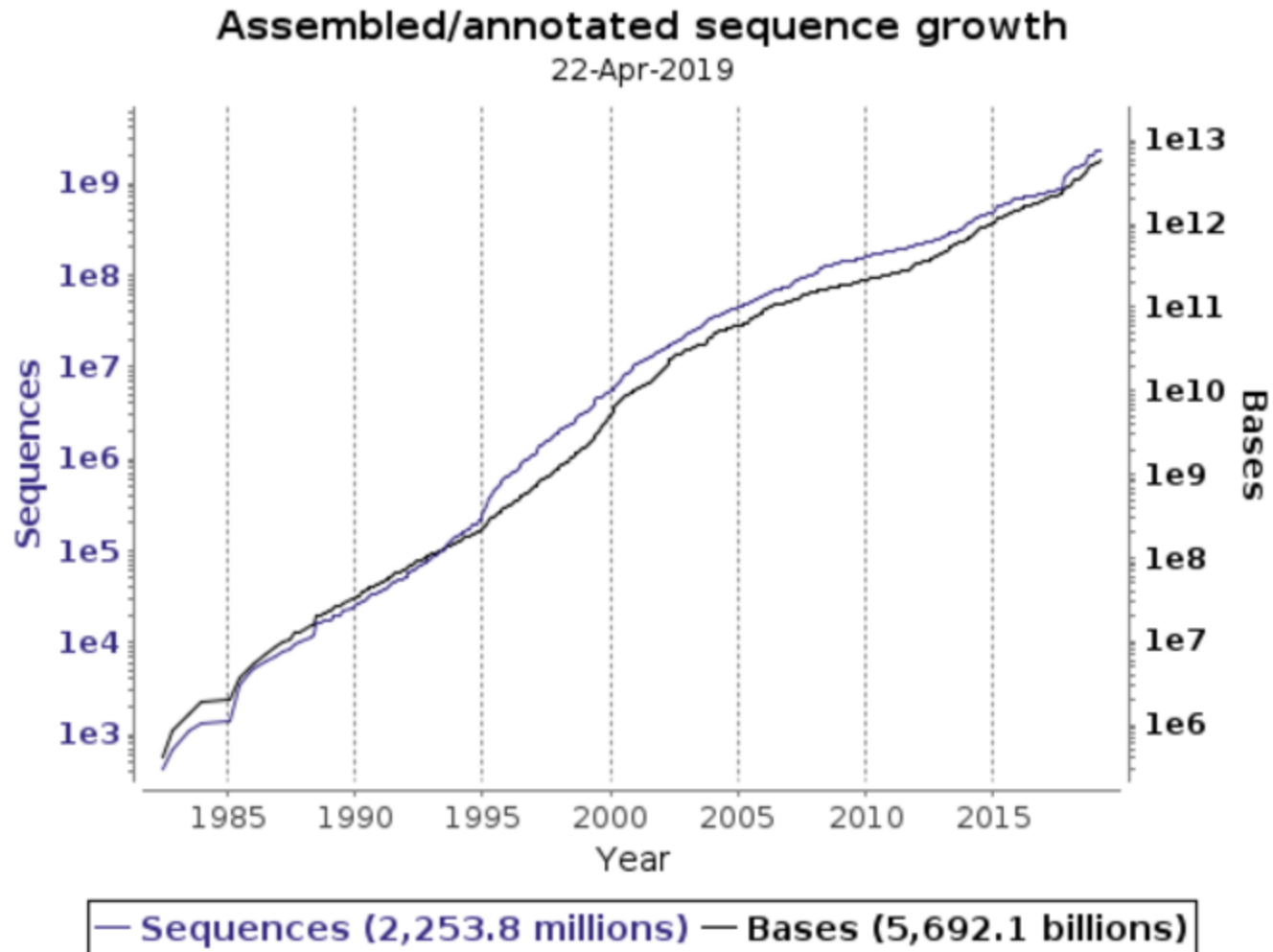


Analýza sekvenčních dat v molekulární biologii

- **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

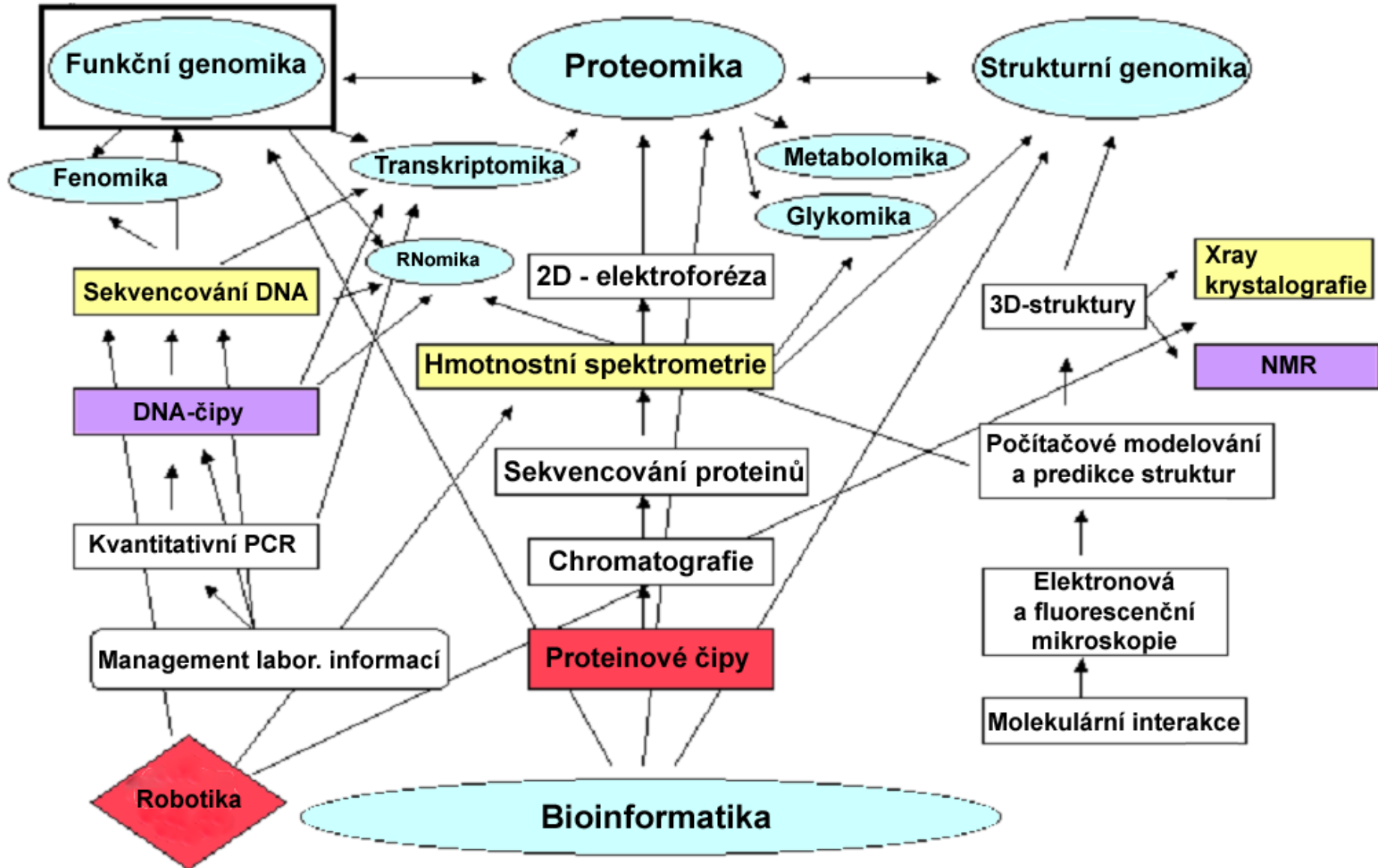
Assembled/annotated sequence growth



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 příložen, assembly |
| Informační teorie | | Identifikace genů |
| Správa databází | | Funkční identifikace proteinů |
| Statistika | | Molekulární evoluce |
| Simulace | | Molekulární modelování |
| Umělá inteligence | | Predikce struktur |
| Zpracování obrazu | | Srovnávání struktur |
| Robotika | Zpracování laboratorních dat | 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

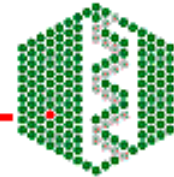


GenBank: <http://www.ncbi.nlm.nih.gov/>

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>

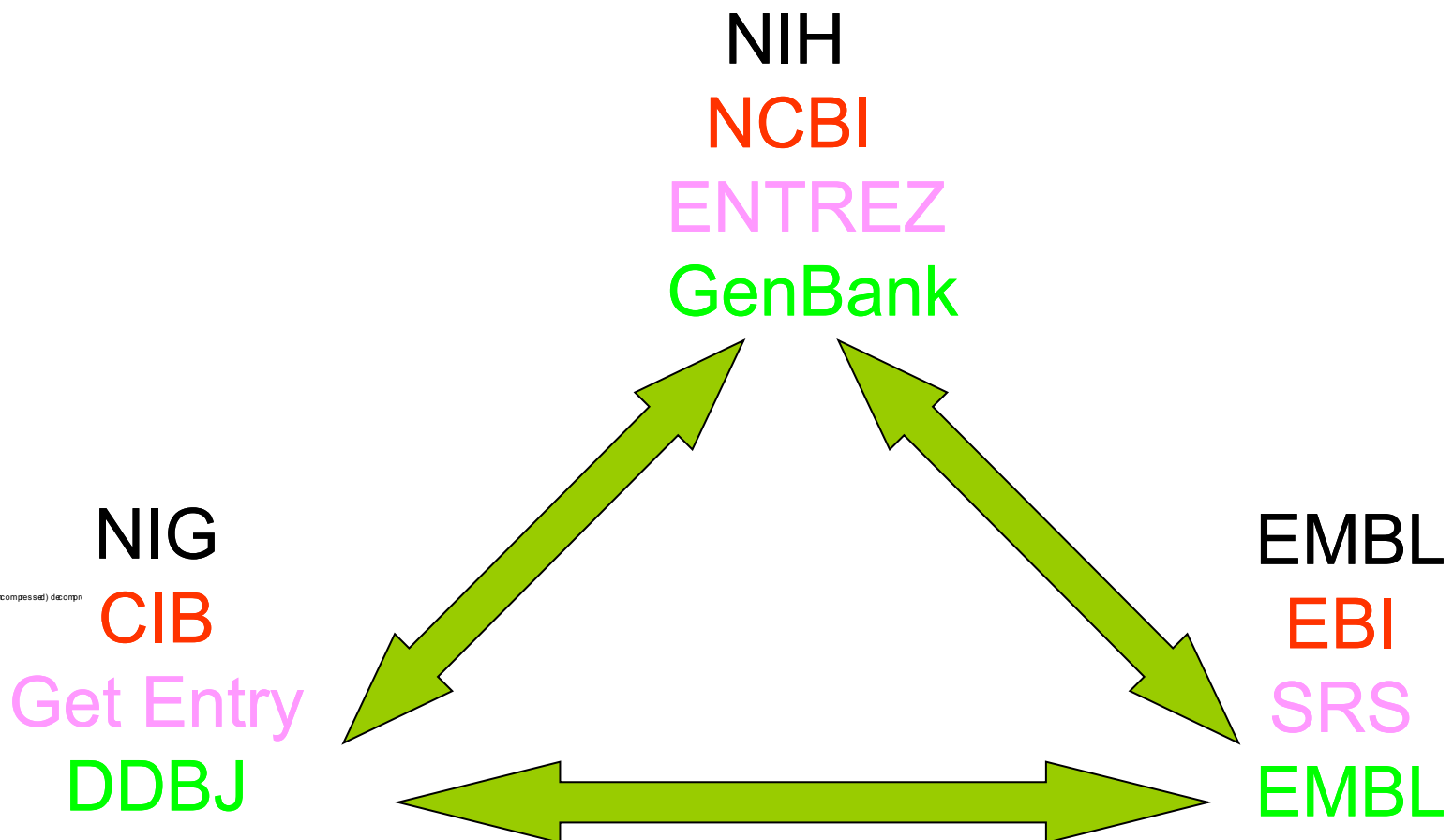
Expert Protein Analysis System

Databáze sekvencí proteinů

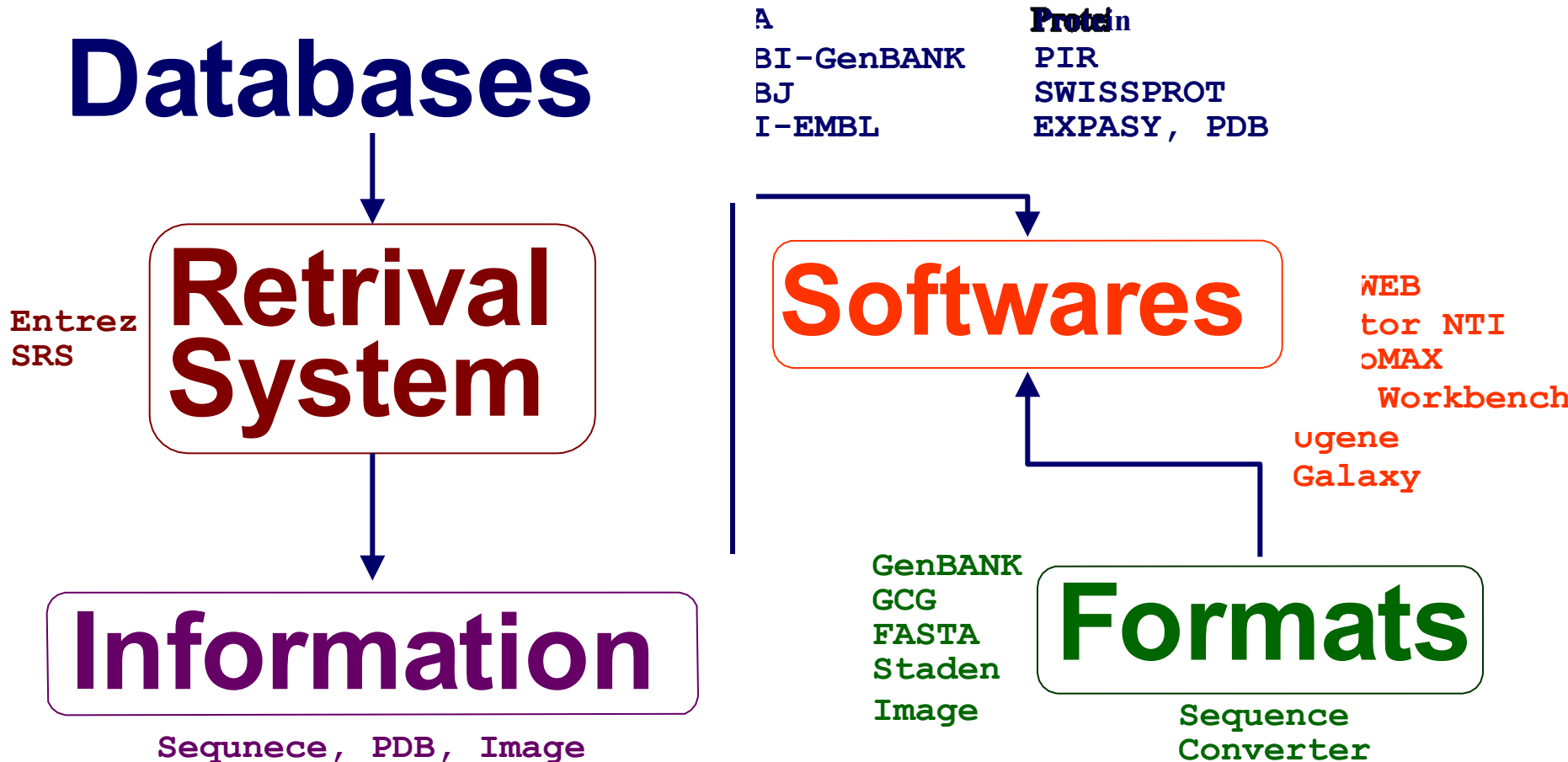
- Sekvence proteinů, u nichž byly experimentálně stanoveny jejich aminokyselinové sekvence, charakterizovány jednotlivé proteinové domény a stanovená jejich funkce jsou ukládány v databázi **SWISS-PROT** založené na Univerzitě v Ženevě v roce 1986.
- Databázi spravuje Švýcarský institut pro bioinformatiku (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

The screenshot shows the NCBI Entrez search engine interface. At the top, there is the NCBI logo and the Entrez logo with the tagline "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, Entrez, Human Genome, GenBank, Map Viewer, and BLAST. A search bar is present with a "GO" button and a "CLEAR" button. The main content area is titled "Welcome to the new Entrez cross-database search page" and lists 24 different databases, each with a small icon and a question mark icon for help. The databases listed are: PubMed, PubMed Central, Journals, MeSH, Books, OMIM, Site Search, Nucleotide, Protein, Genome, Structure, Taxonomy, SNP, Gene, UniGene, CDD, 3D Domains, UniSTS, PopSet, GEO, and GEO DataSets.

Entrez Molecular Sequence Database System

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

The screenshot displays the NCBI website interface. At the top, there is a navigation bar with 'Resources' and 'How To' menus, a search bar, and user options like 'panlucsek', 'My NCBI', and 'Sign Out'. A dropdown menu for 'All Databases' is open, listing various databases such as Assembly, Biocollections, BioProject, BioSample, BioSystems, Books, ClinVar, Conserved Domains, dbGaP, dbVar, Gene, Genome, GEO DataSets, GEO Profiles, GTR, HomoloGene, Identical Protein Groups, MedGen, and MeSH. The main content area features a 'COVID-19' banner, a 'Public health' link, and a 'SARS-CoV-2 data (NCBI)' link. Below this, there are sections for 'Submit', 'Download', 'Learn', 'Develop', 'Analyze', and 'Research', each with an icon and a brief description. On the right side, there are sections for 'Popular Resources' (PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI News & Blog' with several news items dated in April 2021.

Sequence Retrieval System (SRS)

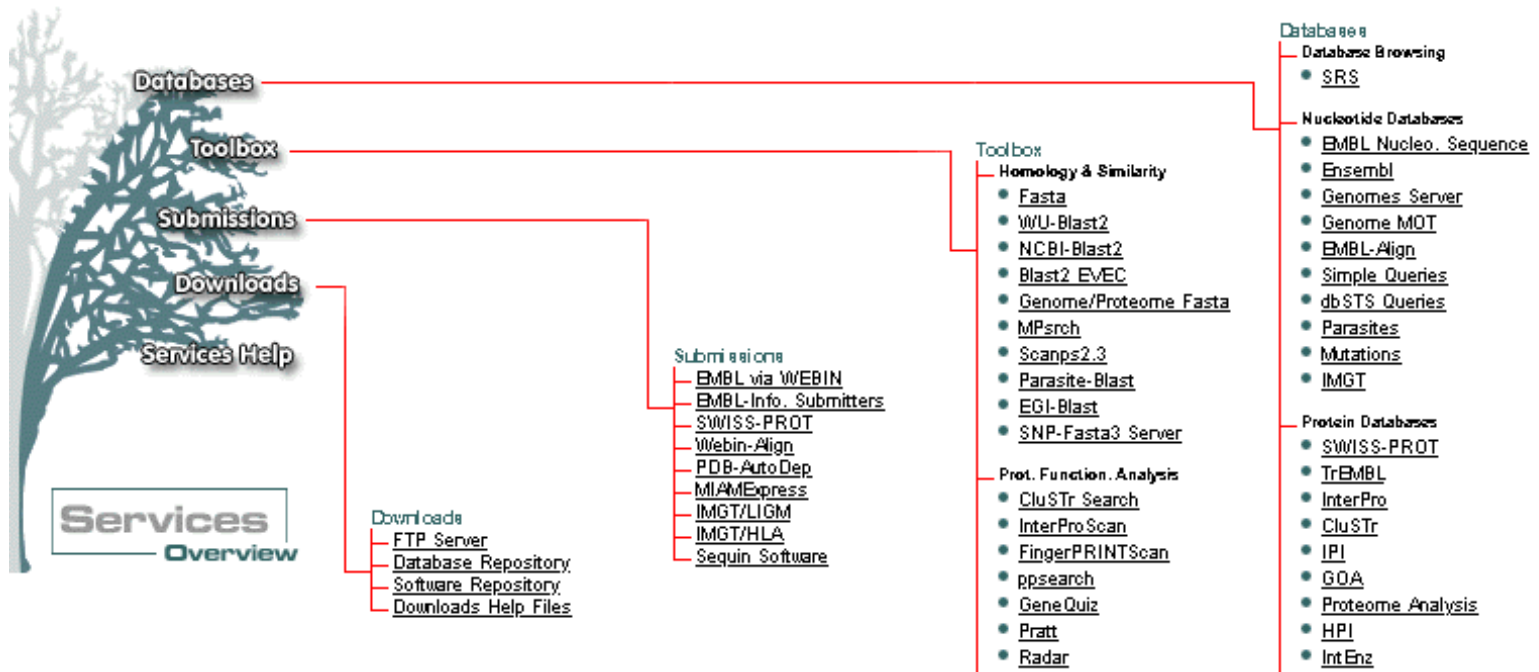
EBI <http://www.ebi.ac.uk/>

EMBL-EBI
European Bioinformatics Institute

Nucleotide sequences for [] Go Site search [] Go

EBI Home About EBI Research **Services** Toolbox Databases Downloads Submissions

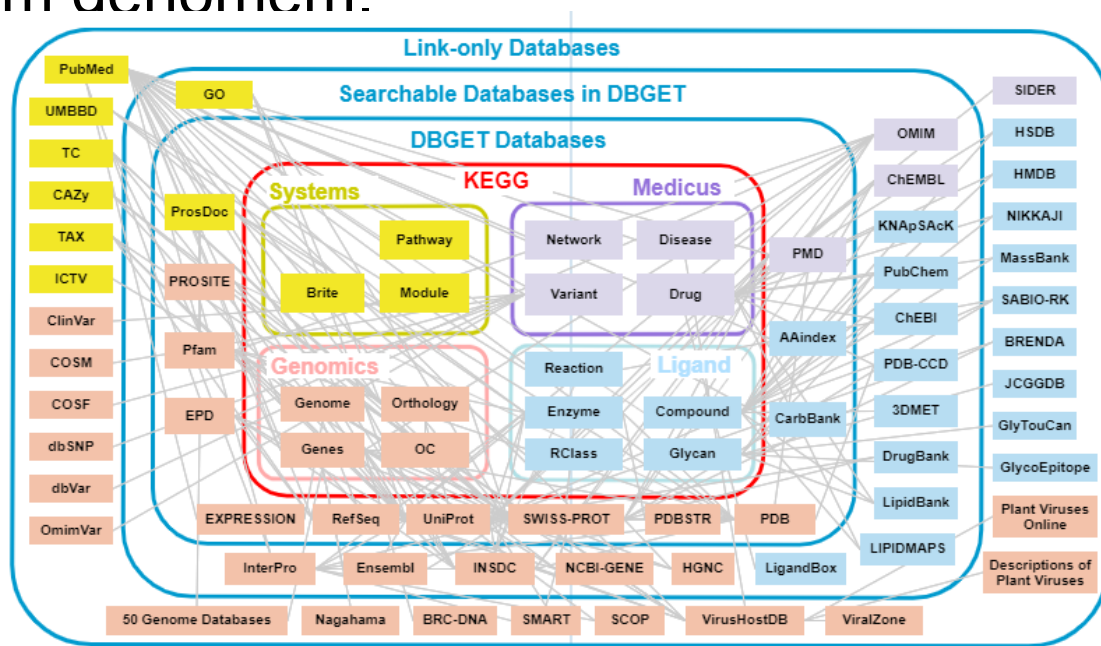
SERVICES OVERVIEW FASTLINK



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 organismů 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 přílohy
- Tbl2asn – batch submission
 - 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**
 - ◆ (NH₂-) 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 (pu <u>R</u> in) |
| Y | T/C (p <u>Y</u> rimidin) |
| K | G/T (nukleosid s <u>K</u> eto skupinou) |
| M | A/C (nukleosid s a <u>M</u> ino skupinou) |
| S | G/C (silná = <u>S</u> trong vazba) |
| W | A/T (slabá = <u>W</u> weak vazba) |
| <hr/> | |
| 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
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAG
```

```
>LinB_protein
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCAGLGR
LIACDLIGMGDSKLDPSGPERYAYA EHRDYLDALWEALDLGDRVVLVVDWGSALG
FDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRSQAGEELVLQDNVFE
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/>

The screenshot shows the NCBI GenBank search interface. At the top, there's a search bar with 'Nucleotide' selected and 'barley NADPH oxidase' entered. Below the search bar are navigation tabs for PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and OMIM. A search button 'Go' and a 'Clear' button are present. Below the search bar are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. A 'Display' dropdown is set to 'default', and there are buttons for 'Save', 'Text', 'Add to Clipboard', and 'Get Subsequence'. The search results show one entry: '1: AJ251717. Hordeum vulgare p...[gi:15282289]'. The entry details include:

- LOCUS: HVU251717 337 bp mRNA linear PLN 18-JAN-2002
- DEFINITION: Hordeum vulgare partial mRNA for putative NAD(P)H oxidase (pNAox gene).
- ACCESSION: AJ251717
- VERSION: AJ251717.1 GI:15282289
- KEYWORDS: NADPH oxidase; pNAox gene.
- SOURCE: Hordeum vulgare subsp. vulgare
- ORGANISM: [Hordeum vulgare subsp. vulgare](#)
- REFERENCE: 1
- AUTHORS: Huckelhoven, R., Dechert, C., Trujillo
- TITLE: Differential expression of putative near-isogenic, resistant and suscep interaction with the powdery mildew
- JOURNAL: Plant Mol. Biol. 47 (6), 739-748 (2006)
- MEDLINE: [21643210](#)
- REFERENCE: 2 (bases 1 to 337)
- AUTHORS: Hueckelhoven, R.
- TITLE: Direct Submission
- JOURNAL: Submitted (02-DEC-1999) Hueckelhoven, R. Phytopathology and Applied Zoology, Giessen, Ludwigstr. 23, 35390 Giessen

This block shows a detailed view of the GenBank entry features and base count. The features section includes:

- source**: Location/Qualifiers
 - 1..337
 - /organism="Hordeum vulgare subsp. vulgare"
 - /cultivar="Pallas"
 - /db_xref="taxon:112509"
 - /tissue_type="primary leaf"
 - /dev_stage="7-days old plant"
- gene**: 1..337
- CDS**:
 - <1..>337
 - /gene="pNAox"
 - /function="superoxide generating enzyme"
 - /note="gp9lphox homolog"
 - /codon_start=2
 - /product="putative NAD(P)H oxidase"
 - /protein_id="[CAC51517.1](#)"
 - /db_xref="GI:15282290"
 - /translation="FKGIMNEIAELDQRNIIEMHNYLTSVYEEGDARSALITMLQALN HAKNGVDVVSQTRVTRTHFARPNFKRVLKSKVAAKHPYAKIGVFYCGAPVLAQELSNLCH EFNKGCTTKF"

 The base count section shows:

| BASE COUNT | 102 a | 70 c | 81 g | 83 t | 1 others | | |
|------------|-------|-------------|-------------|-------------|-------------|-------------|------------|
| ORIGIN | 1 | gtttaaagga | atcatgaatg | agattgctga | actagatcaa | aggaatatca | ttgagatgca |
| | 61 | caactatctc | acaagtgttt | atgaggaagg | ggatgctcgg | tcagcactca | tcacaatgct |
| | 121 | gcaagctctc | aaccatgccca | agaatggtgt | cgatgtagtg | tctggmactc | gagtcgggac |
| | 181 | acatthttgca | agaccaaatt | ttaagagggt | gctgtctaag | gtagccgcca | aacatcctta |
| | 241 | tgccaagata | ggagtgttct | attgogggagc | tccagtctctg | gccgagggaac | taagcaacct |
| | 301 | ttgccatgag | ttcaatggca | aatgcacgac | aaaatc | | |

Genomové databáze v NCBI - eukaryota

NCBI Entrez Genomes

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Help

Search for on chromosome(s) Find

Show linked entries Help FTP

Entrez Genomes
MapViewer Home

Prominent organisms

FTP SITE

Related Databases:
TAIR
TIGR
MIPS
KAOS

Sequencing Projects:
SPP Consortium
CSH / WashU
TIGR
Kazusa
ESSA
Genoscope

Arabidopsis thaliana genome view [BLAST search Arabidopsis genome](#)

I II III IV V MT CHL

Lineage: [Eukaryota](#); [Viridiplantae](#); [Streptophyta](#); [Embryophyta](#); [Tracheophyta](#); [Spermatophyta](#); [Magnoliophyta](#); [eudicotyledons](#); [core eudicots](#); [Rosidae](#); [eurosids II](#); [Brassicales](#); [Brassicaceae](#); [Arabidopsis](#)

Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish. Arabidopsis is not of major agronomic significance, but it offers important advantages for basic research in genetics and molecular biology. Its genome has been sequenced by an international collaboration collectively termed the [Arabidopsis Genome Initiative \(AGI\)](#) ([The Arabidopsis Genome Initiative, 2000, Nature, 408:796-815](#)).

This sequence, map, and annotations are the result of a collaboration between [TIGR](#), [MIPS](#), and [TAIR](#). The non-redundant sequence of the chromosomes (pseudomolecules) and their annotations were provided to NCBI by TIGR on behalf of the collaborators.

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/>

The screenshot shows the top navigation bar of the RCSB PDB website. It includes a dark blue header with white text for navigation: 'RCSB PDB', 'Deposit', 'Search', 'Visualize', 'Analyze', 'Download', 'Learn', 'More', and 'Documentation'. A 'MyPDB' button is on the right. Below the navigation bar is a light blue banner with the PDB logo on the left, which states '177219 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education'. A search bar with the placeholder 'Enter search term(s)' and a magnifying glass icon is in the center. Below the search bar are links for 'Advanced Search' and 'Browse Annotations', and a 'Help' link with an external icon. At the bottom of the banner are logos for 'PDB-101', 'EMDataResource', 'Nucleic Acid Database', and 'Worldwide Protein Data Bank Foundation'. On the right side of the banner, there is a 'Celebrating' graphic for '25 YEARS OF Protein Data Bank' and social media icons for Facebook, Twitter, YouTube, and LinkedIn.

A vertical navigation sidebar on the left side of the page with a dark blue background and white text. It contains the following items from top to bottom: 'Welcome' with a bookmark icon, 'Deposit' with a plus icon, 'Search' with a magnifying glass icon, 'Visualize' with a camera icon, 'Analyze' with a grid icon, 'Download' with a plus icon, and 'Learn' with a book icon.

A Structural View of Biology

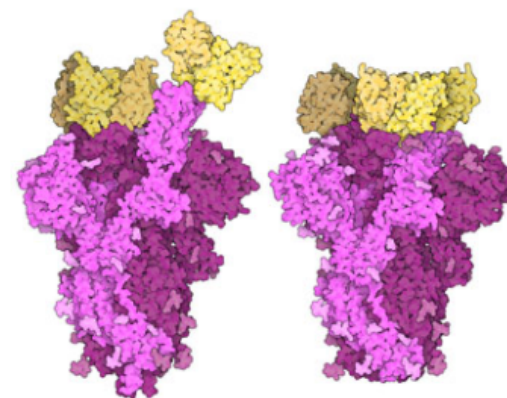
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.


Two promotional banners are shown side-by-side. The left banner features a 3D model of a coronavirus spike protein and text: 'COVID-19 CORONAVIRUS Resources'. The right banner features a 3D model of a protein structure and text: 'PDB50: A special symposium May 4-5, 2021 Register by May 1 VIRTUAL EVENT' with the ASBMB logo.

April Molecule of the Month



SARS-CoV-2 Spike and Antibodies

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?

- **Basic Local Alignment Search Tool**
- Hledání lokálních podobností
- Heuristický přístup založený na **Smith-Watermanově** algoritmu
- Vyhledá neoptimálnější **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?

- **Basic Local Alignment Search Tool**
 - Hledání lokálních podobností
 - Heuristický přístup založený na **Smith-Watermanově** algoritmu
 - Vyhledá neoptimálnější **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

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST [More BLAST news...](#)



Výchozí stránka BLAST

Web BLAST

Nucleotide BLAST
nucleotide → nucleotide

blastx
translated nucleotide → protein

tblastn
protein → translated nucleotide

Protein BLAST
protein → protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an Instance at a cloud provider

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

GEO

Find matches to gene expression profiles

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Targeted Loci

Multiple Alignment

BioAssay

MOLE-BLAST

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



Basic BLAST – výběr programů

Využití jednotlivých programů BLAST

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

Uživatelské rozhraní BLAST

The screenshot displays the NCBI BLAST web interface. At the top, there is a navigation bar with the NIH logo, 'U.S. National Library of Medicine', the NCBI logo, and a 'Sign in to NCBI' link. Below this is a secondary navigation bar with 'BLAST' and '>> blastn suite' on the left, and 'Home', 'Recent Results', 'Saved Strategies', and 'Help' on the right. The main heading is 'Standard Nucleotide BLAST'. Below the heading are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. A sub-header reads 'BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)' with 'Reset page' and 'Bookmark' links. The 'Enter Query Sequence' section includes a text input for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' input fields. Below this is an 'Or, upload file' section with a file selection button and a 'Job Title' input field. A checkbox for 'Align two or more sequences' is also present. The 'Choose Search Set' section has a 'Database' dropdown set to 'Nucleotide collection (nr/nt)', an 'Organism' input field with an 'Exclude' checkbox, and an 'Exclude' section with checkboxes for 'Models (XMP)' and 'Uncultured/environmental sample sequences'. There is also a 'Limit to' section with a checkbox for 'Sequences from type material' and an 'Entrez Query' input field. The 'Program Selection' section has an 'Optimize for' section with radio buttons for 'Highly similar sequences (megablast)', 'More dissimilar sequences (discontiguous megablast)', and 'Somewhat similar sequences (blastn)', along with a 'Choose a BLAST algorithm' dropdown. At the bottom, there is a 'BLAST' button and a summary of the search: 'Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)'. A checkbox for 'Show results in a new window' is also visible. A link for 'Algorithm parameters' is at the bottom left.

- [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

minimální velikost = 7

TACTGGACATG

blastn default = 11

tabulka se všemi slovy dotazu

ACTGGACATGG

megablast default = 28
(16 - 256)

CTGGACATGGA

TGGACATGGAC

GGACATGGACC

GACATGGACCC

ACATGGACCCT

.....

přiložení sekvencí, které BLAST může nalézt

```
1 AATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAG
  ||||| ||||||||||||||||| || |||||||||||||
1 AATGGAAAAGACTACTGGATCATCAAAACTCCTGGGGAG
```

sekvence obsahují definovanou shodu slova

Tvorba dotazu pro proteinové sekvence



Dotaz: **GTQITVEDLIFYNIATRRKALKN**

GTQ Word size = 3

TQI

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

QIT

Sousedící slova

ITV → LTV, MTV, ISV, LSV, etc.

TVE

VED

EDL

DLF

...

tabulka se všemi slovy dotazu

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**

GTQITVEDLFIYNI

SEI

YIN

sousedící slova

2 nalezené shody

Hodnocení výsledků příložení

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



Typy matic pro výpočet skóre

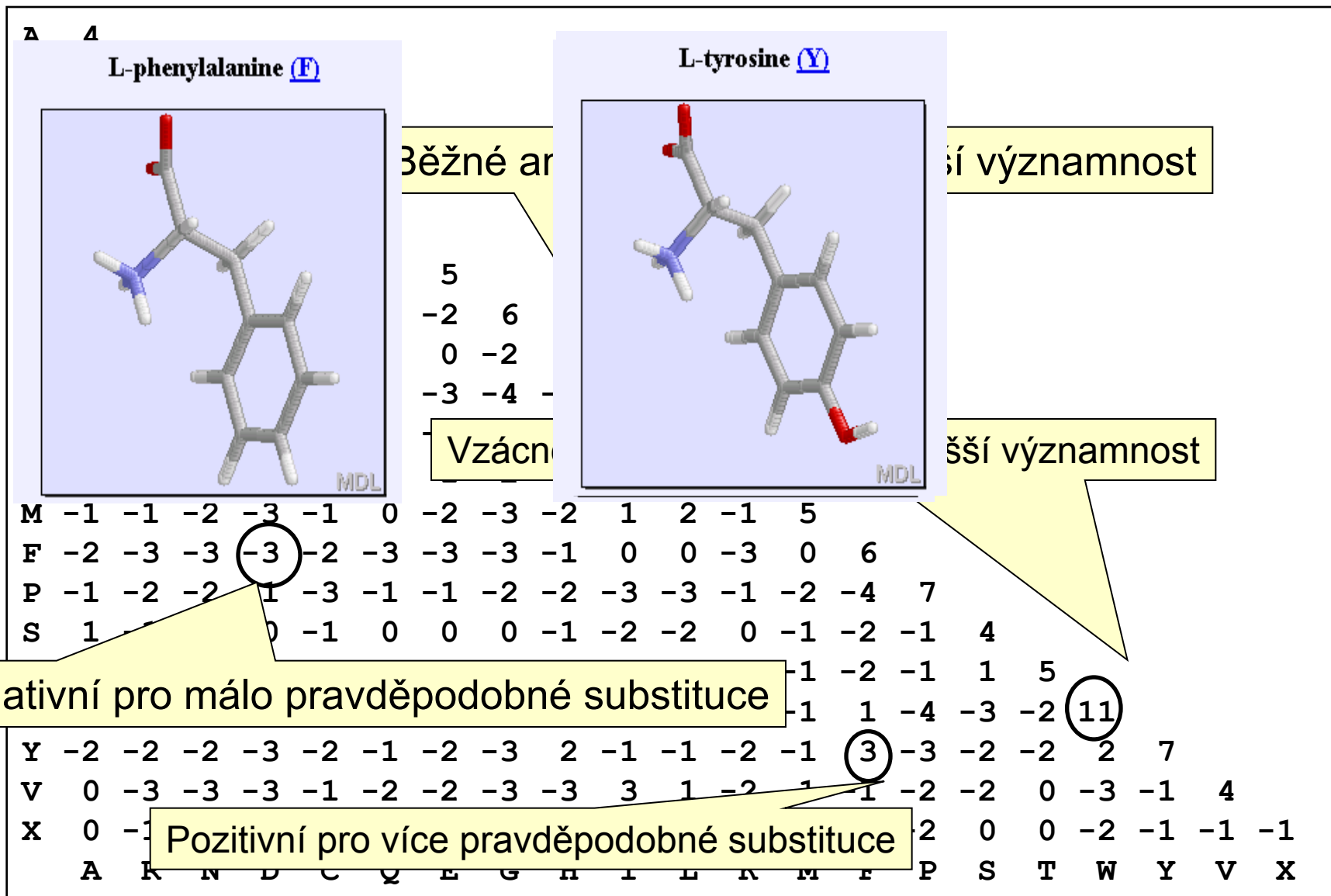
- Matice identity
 - Především pro nukleotidové sekvence
 - Neschopné transformovat na jiné zbytky
 - Pro přiložení velmi podobných sekvencí
- Matice podobnosti
 - Používané u proteinových sekvencí
 - Vyjadřují biochemické/biologické vlastnosti aminokyselin
 - Vyšší účinnost při srovnávání sekvencí



Matrice BLOSUM

- **Blocks Substitution Matrix**
- Matice BLOSUM jsou sestaveny na základě **analýzy mnohonásobných příložení 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



BLAST –výstup, výsledky



[← Edit Search](#) [Save Search](#) [Search Summary](#) ▾

Job Title Nucleotide Sequence

RID [SKGS6ZHW016](#) Search expires on 10-17 23:57 pm [Download All](#) ▾

Program BLASTN [?](#) [Citation](#) ▾

Database nt [See details](#) ▾

Query ID lcl|Query_43165

Description None

Molecule type dna

Query Length 2774

Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

[?](#) How to read this report? [BLAST Help Videos](#) [↶ Back to Traditional Results Page](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to **E value** to **Query Coverage** to

[Filter](#) [Reset](#)

LocusLink

Descriptions [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments [Download](#) ▾ [Manage](#) ▾ [Show](#) 100 ▾ [?](#)

seřazeno podle hodnot E

link to entrez

Default e value cutoff 0.05

| | Max Score | Total Score | Query Cover | E value | Per Ident | Accession |
|--|-----------|-------------|-------------|---------|-----------|----------------------------|
| <input checked="" type="checkbox"/> Staphylococcus hominis strain 19A chromosome, complete genome | 1858 | 3951 | 87% | 0.0 | 96.07% | CP031277.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_745 chromosome | 1842 | 3547 | 79% | 0.0 | 95.80% | CP050982.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_136 chromosome, complete genome | 1831 | 3923 | 87% | 0.0 | 95.48% | CP014107.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_762 chromosome, complete genome | 1825 | 1825 | 41% | 0.0 | 95.39% | CP054006.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_747 chromosome, complete genome | 1821 | 2098 | 41% | 0.0 | 100.00% | CP051909.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis J6 genome assembly, chromosome 1 | 1805 | 1805 | 41% | 0.0 | 95.27% | LT963442.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis J11 genome assembly, chromosome 1 | 1805 | 1805 | 41% | 0.0 | 95.27% | LT963438.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_575 chromosome, complete genome | 1799 | 1799 | 41% | 0.0 | 95.18% | CP033732.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_746 chromosome, complete genome | 1797 | 3464 | 79% | 0.0 | 95.18% | CP046306.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_744 chromosome, complete genome | 1794 | 1794 | 41% | 0.0 | 95.10% | CP054883.1 |
| <input checked="" type="checkbox"/> Staphylococcus hominis strain FDAARGOS_747 chromosome, complete genome | 1794 | 1794 | 41% | 0.0 | 95.10% | CP046301.1 |
| <input checked="" type="checkbox"/> Staphylococcus aureus strain ER02693.3 chromosome, complete genome | 1777 | 2053 | 41% | 0.0 | 99.19% | CP030605.1 |

BLAST –výstup, výsledky

BLAST[®] » blastn suite » results for RID-SKGS6ZHW016

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[< Edit Search](#)

[Save Search](#)

[Search Summary](#) ▾

[? How to read this report?](#)

[▶ BLAST Help Videos](#)

[↶ Back to Traditional Results Page](#)

| | |
|----------------------|--|
| Job Title | Nucleotide Sequence |
| RID | SKGS6ZHW016 <small>Search expires on 10-17 23:57 pm</small> Download All ▾ |
| Program | BLASTN ? Citation ▾ |
| Database | nt See details ▾ |
| Query ID | lcl Query_43165 |
| Description | None |
| Molecule type | dna |
| Query Length | 2774 |
| Other reports | Distance tree of results MSA viewer ? |

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+](#) [Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

Filter

Reset

[Descriptions](#)

[Graphic Summary](#)

[Alignments](#)

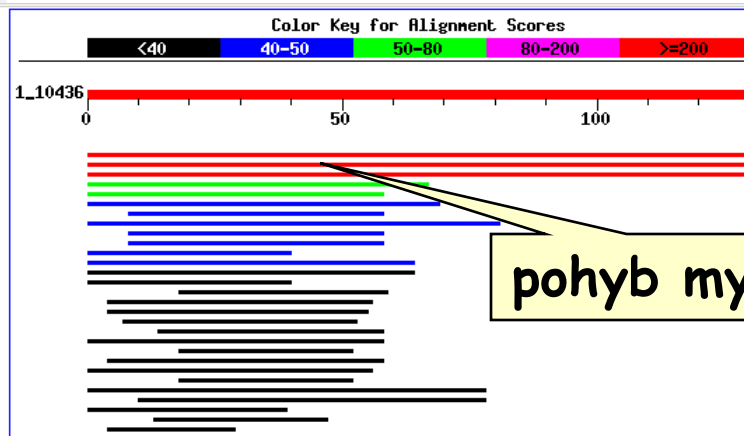
[Taxonomy](#)

[hover to see the title](#) [click to show alignments](#)

Alignment Scores < 40 40 - 50 50 - 80 80 - 200 >= 200 [?](#)

48 sequences selected [?](#)

Distribution of the top 58 Blast Hits on 48 subject sequences

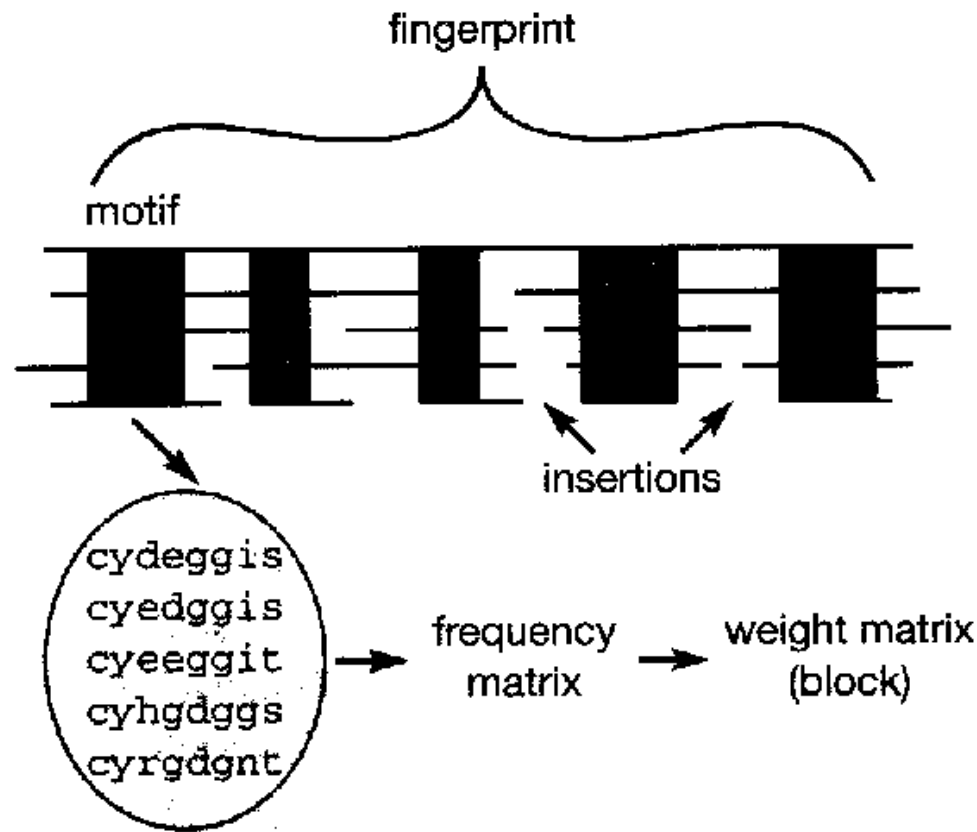
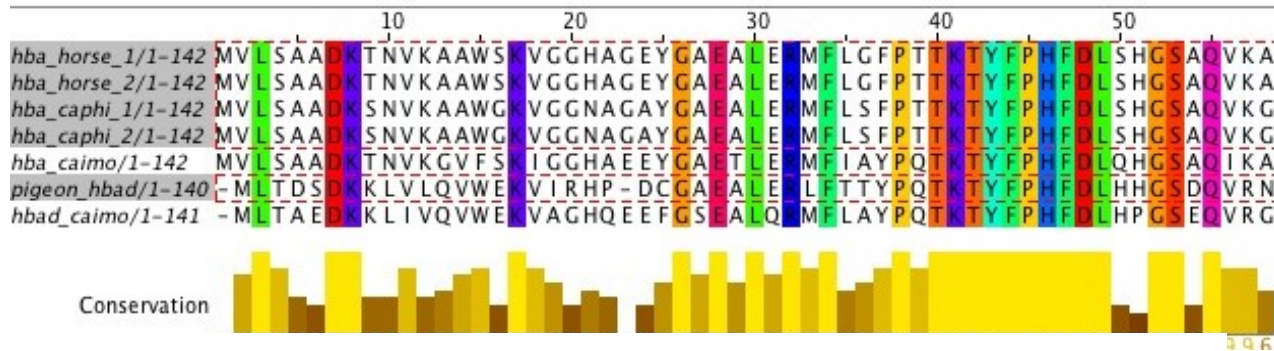


pohyb myši

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 rámeček 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 popsányými 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

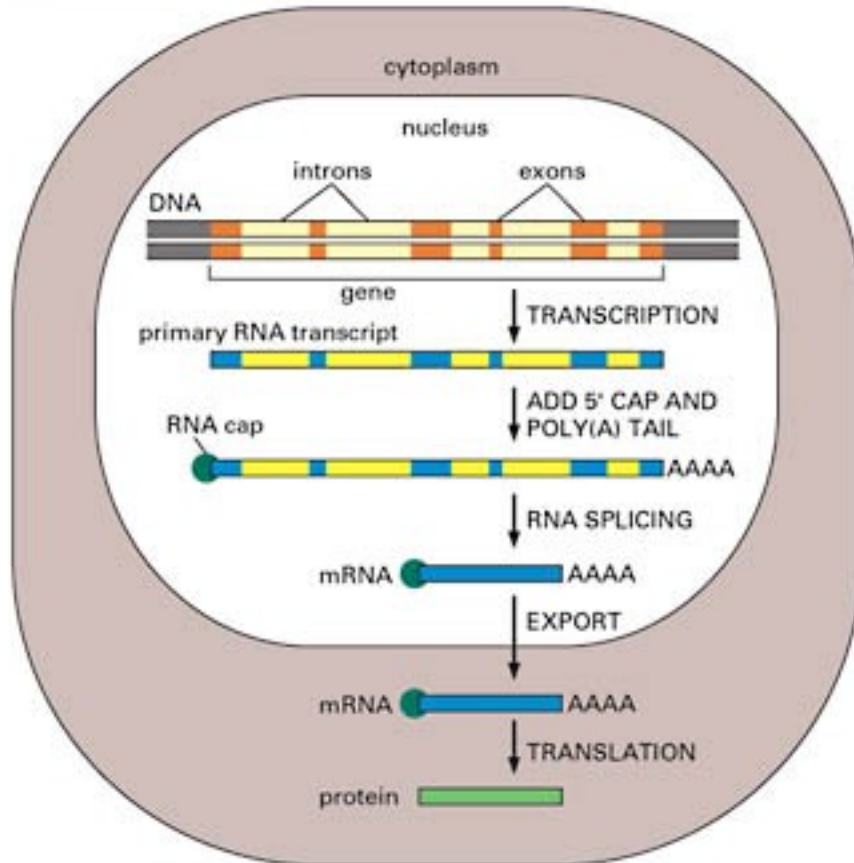
Příklady velikostí genomů

| Druh | Velikost | Genů | Genů na Mb |
|------------------------|----------|--------|------------|
| <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 |

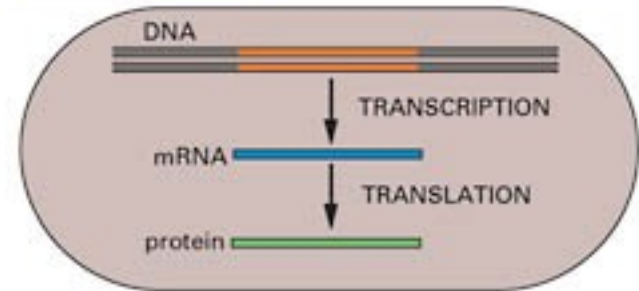


Prokaryotický versus eukaryotický gen

(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%)
 - 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/>)

The screenshot shows the NCBI ORF Finder web interface. The top navigation bar includes links for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is titled "ORF Finder (Open Reading Frame Finder)" and contains two paragraphs of text describing the tool's function and usage. Below the text are input fields for "Enter GI or ACCESSION" and "or sequence in FASTA format", along with "OrfFind" and "Clear" buttons. At the bottom, there are "FROM:" and "TO:" input fields and a "Genetic codes" dropdown menu set to "1 Standard".

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI

Tools
for data mining

GenBank
sequence submission support and software

FTP site
download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.

This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION

or sequence in FASTA format

FROM: **TO:**

[Genetic codes](#)

1 Standard

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



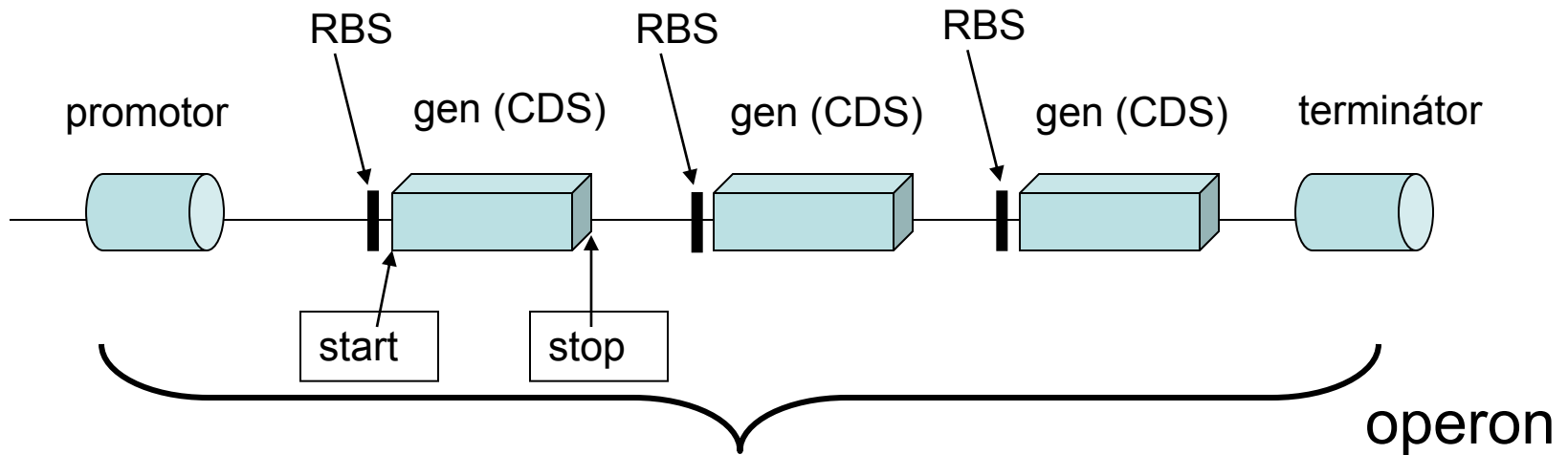
- Využívá pouze sekvenční data a výpočetní přístupy integrující analýzu sekvence a detekci signálů
- Prokaryota
 - Hledání otevřených čtecích rámců doplněné hledáním konzervativních signálů v transkripčních jednotkách
 - **ORF Finder (Open Reading Frame Finder)**
<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>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



Klíčové signály pro odhalení genů

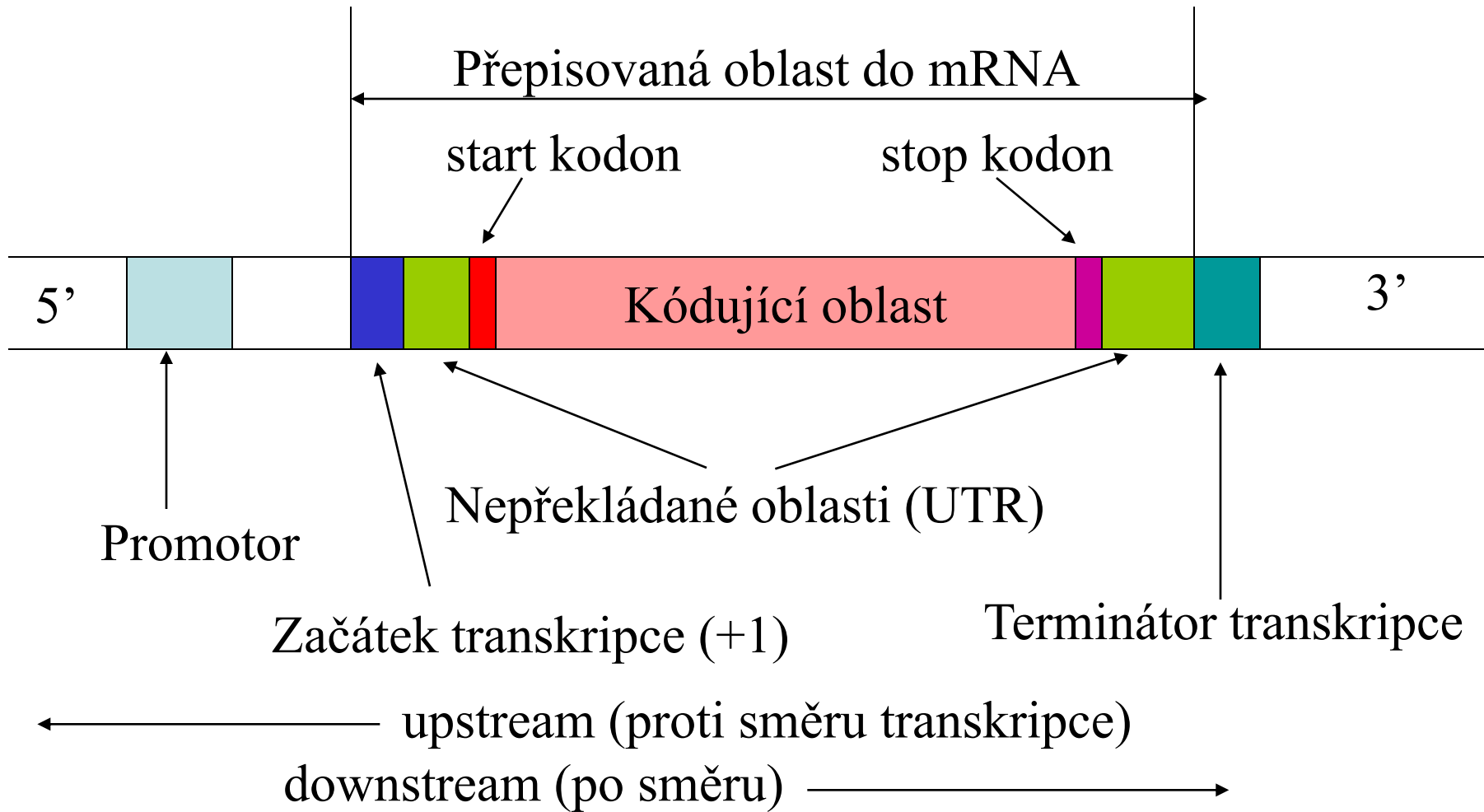
- iniciační a terminační kodony
- promotory
- vazebná místa pro ribozómy (RBS)
- místa sestřihu
- terminátory transkripce
- polyadenylační místa
- vazebná místa pro transkripční faktory

Struktura prokaryotické transkripční jednotky



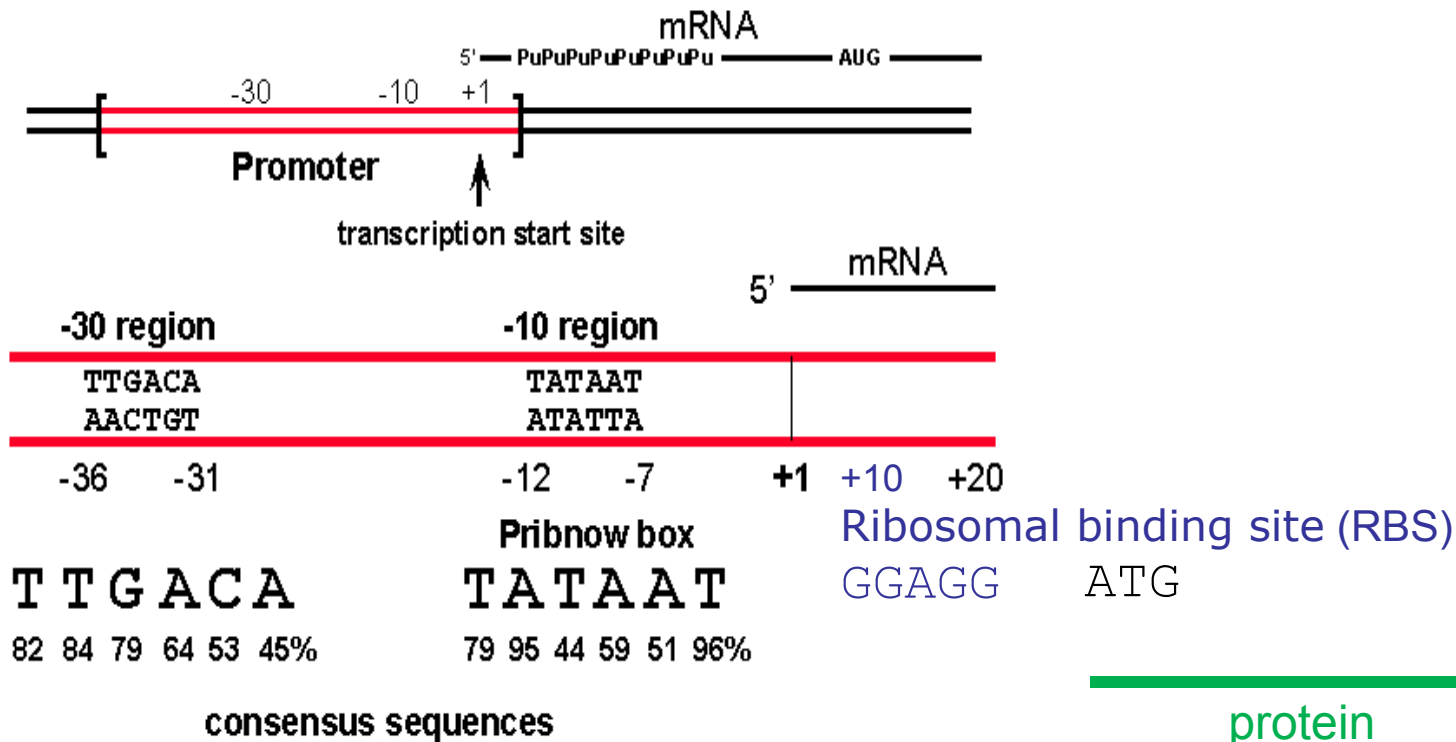


Struktura prokaryotického genu



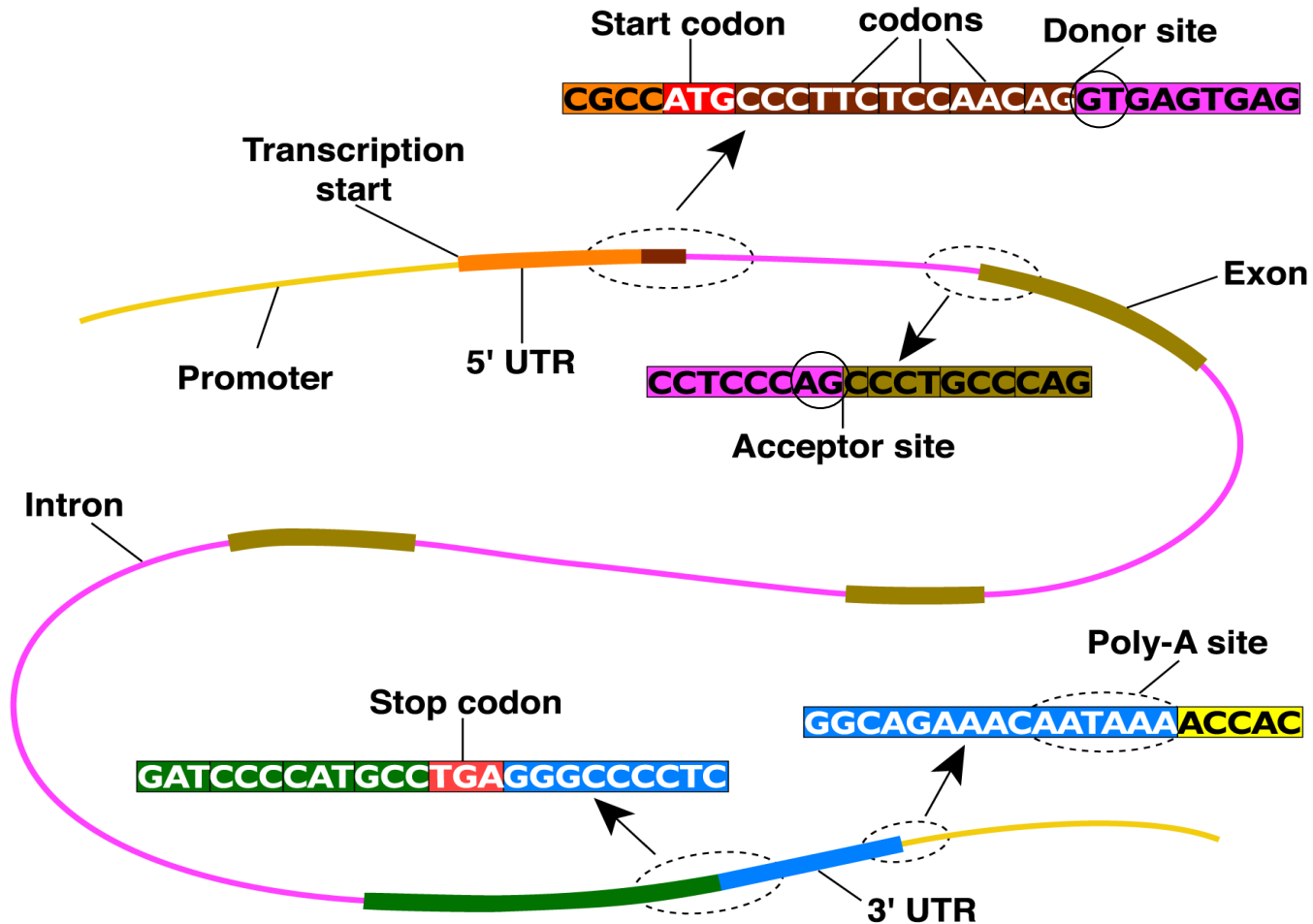


Konzervativní struktury v promotoru prokaryot





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

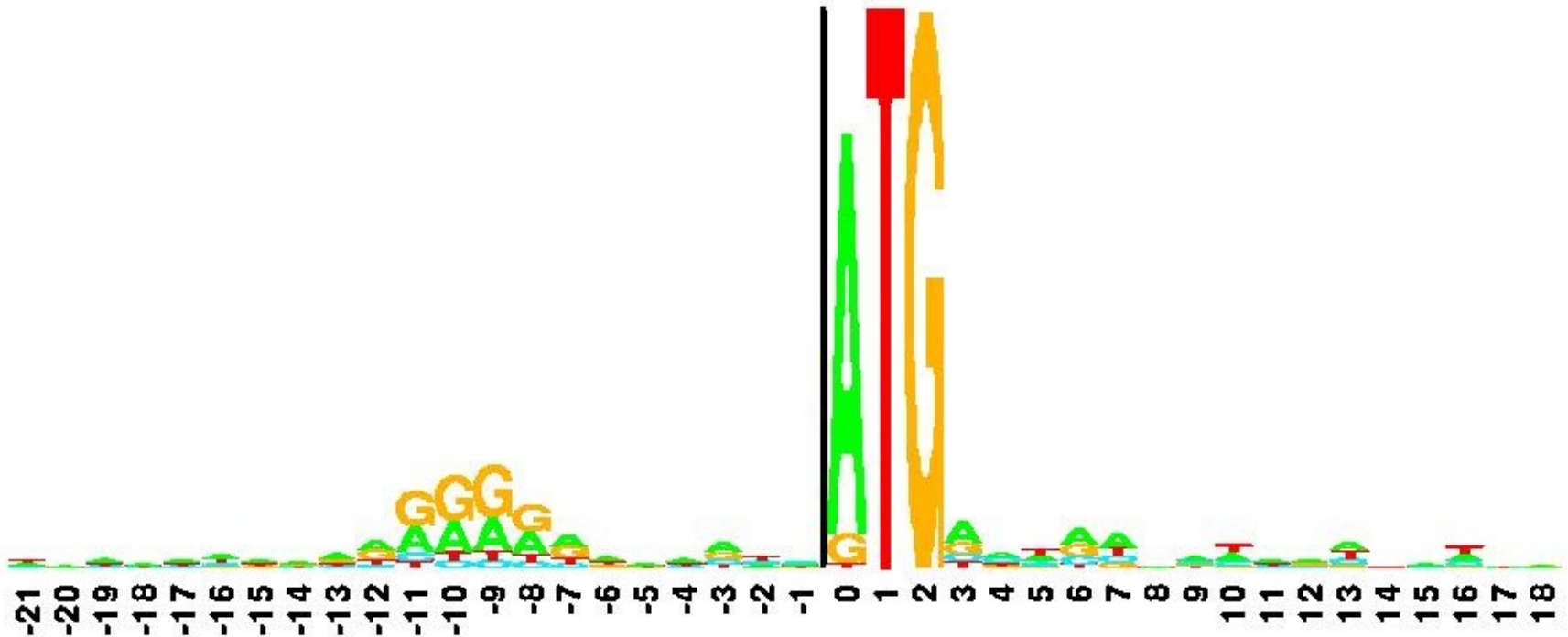
Příklad poziční vážené matice

- Vyjadřuje frekvenci každé báze v každé pozici příslušné sekvence

| | | | | | | | |
|--------|---|---|---|---|---|---|---|
| TACGAT | | 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 | | | | | | | |

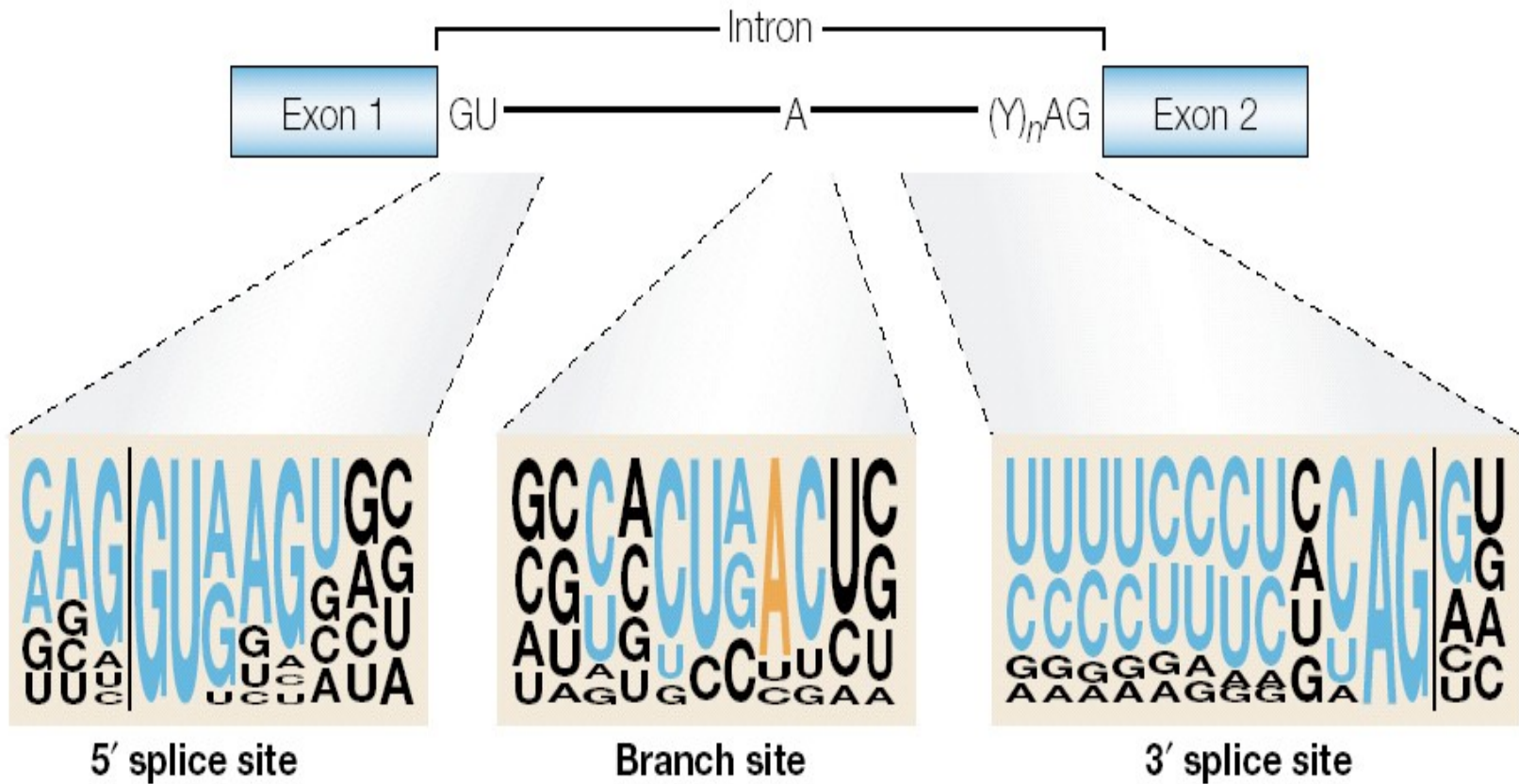
- Skóre každého předpokládaného místa je vyjádřeno součtem hodnot z matice (převáděno na pravděpodobnosti)
- Nevýhody:
 - Je vyžadována hraniční hodnota
 - Předpokládá nezávislost sousedících bází

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

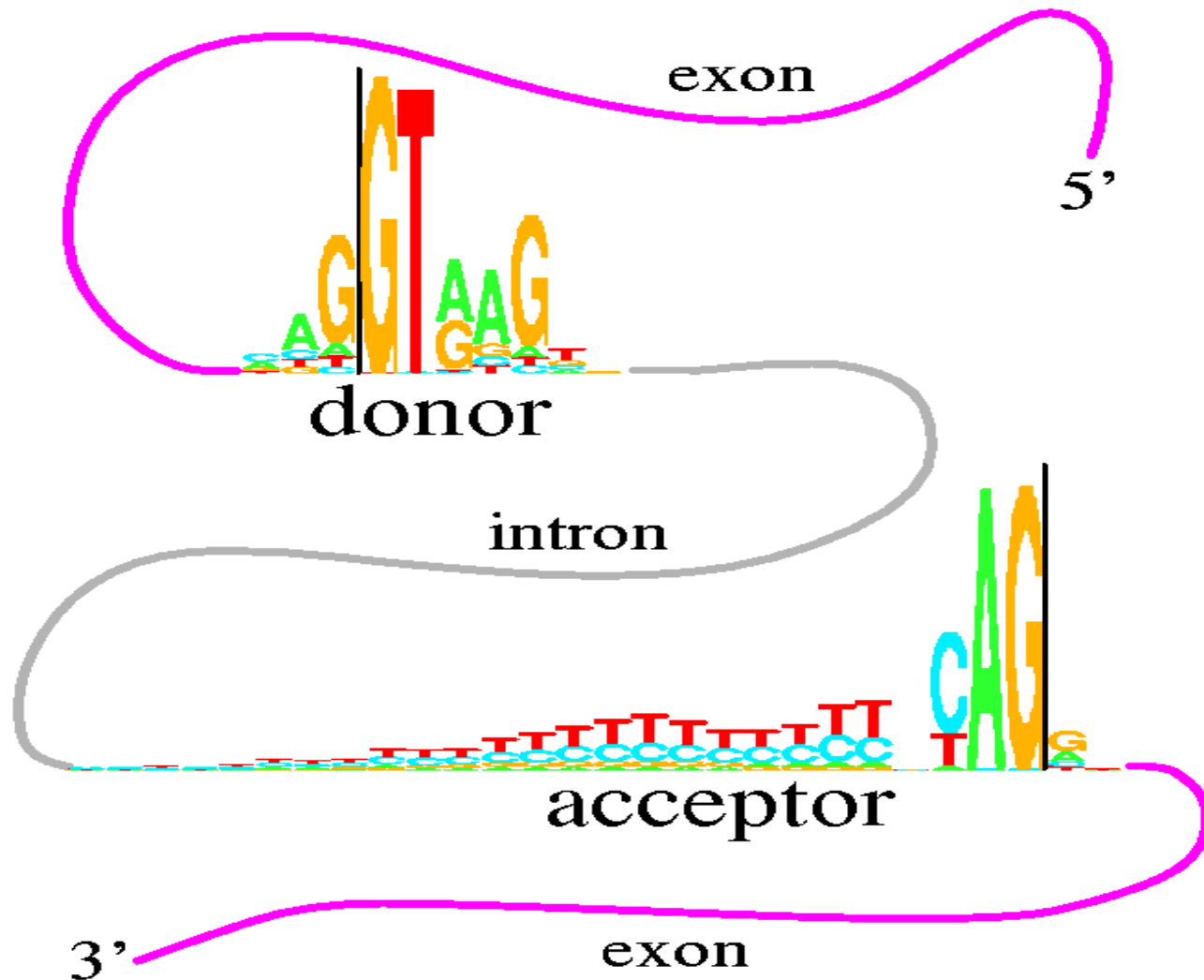




Predikce míst sestřihu



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)
 - periodicitu nukleotidů

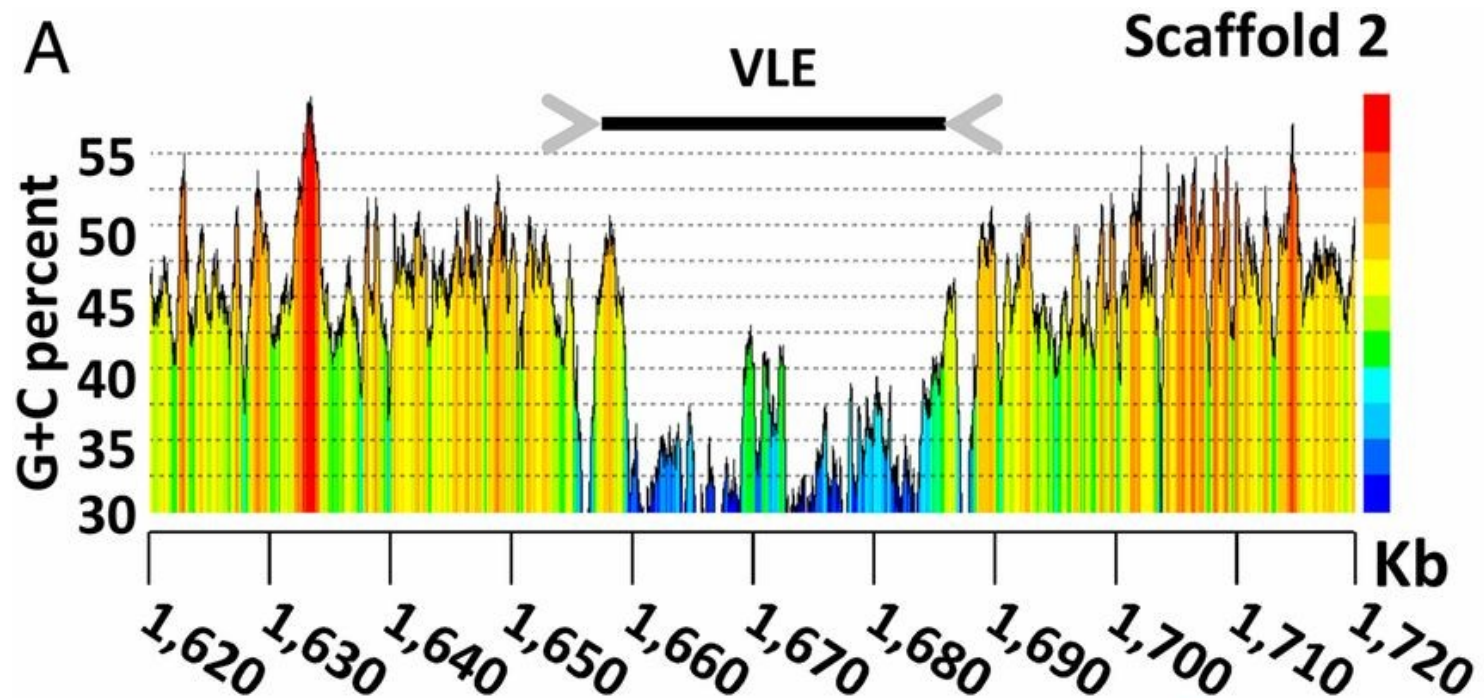
Testování exonů – využití kodonů

| 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 |

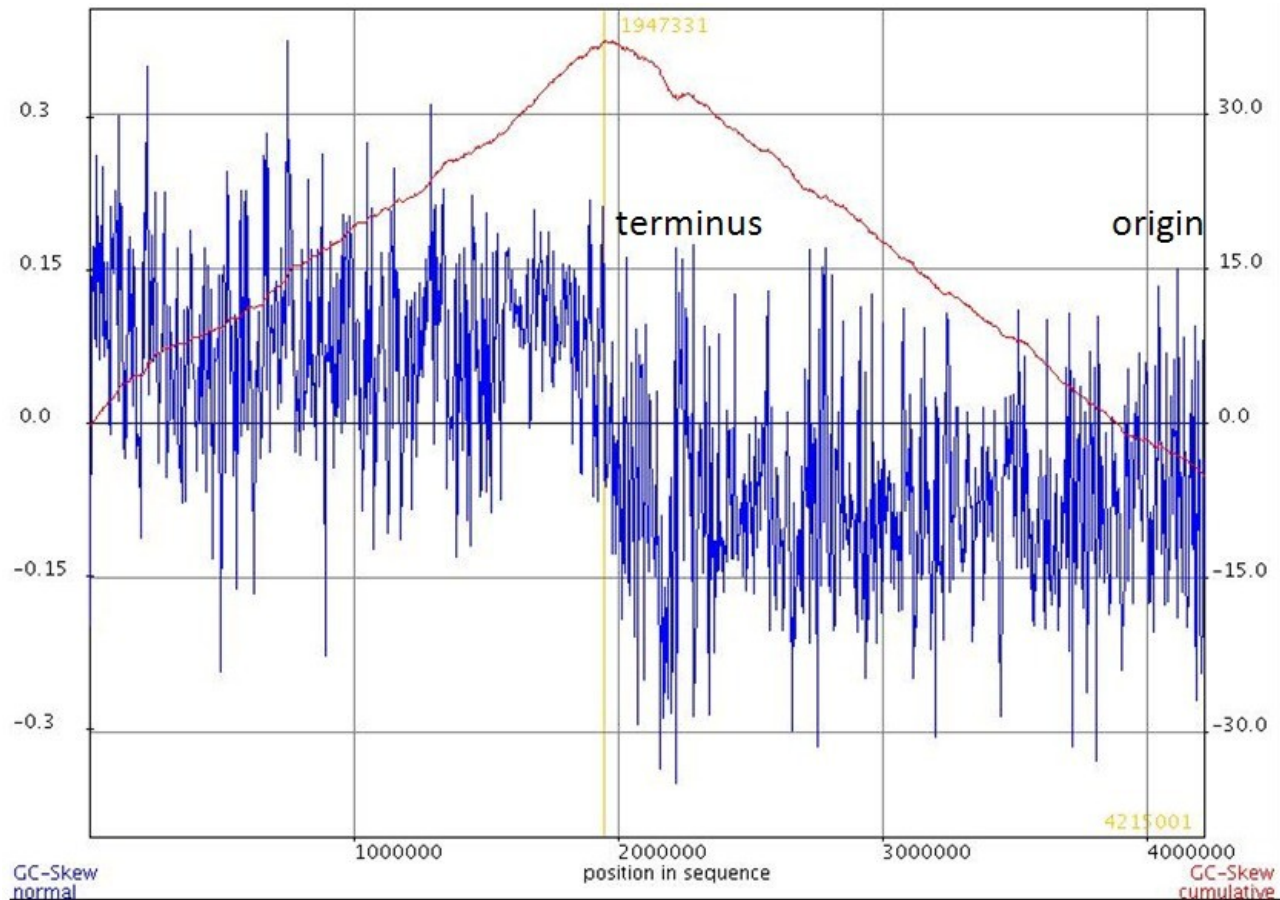
Codon usage database: <http://www.kazusa.or.jp/codon/>

Obsah G+C – příklad využití pro identifikaci mobilního elementu



Odlišný obsah G+C indikuje horizontální přenos

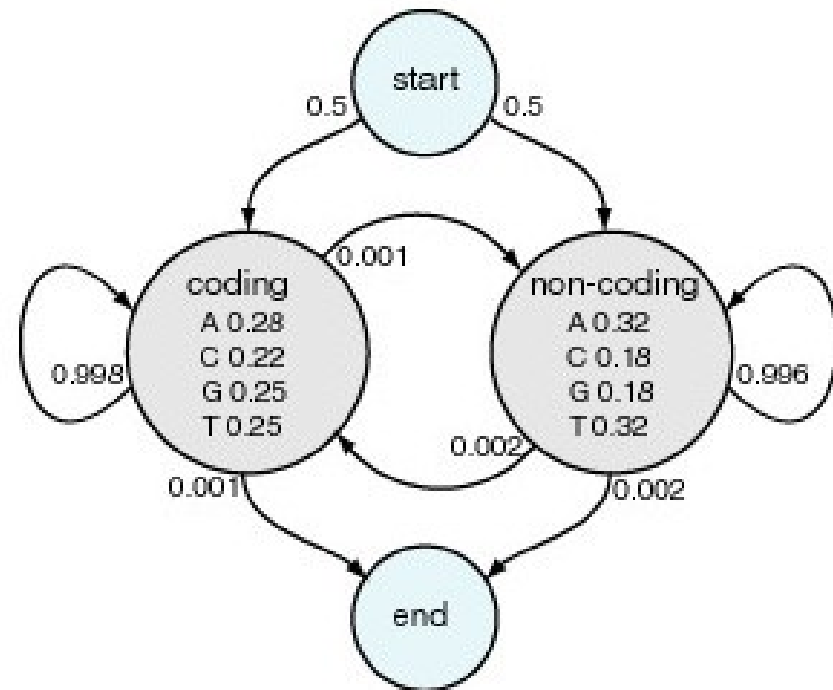
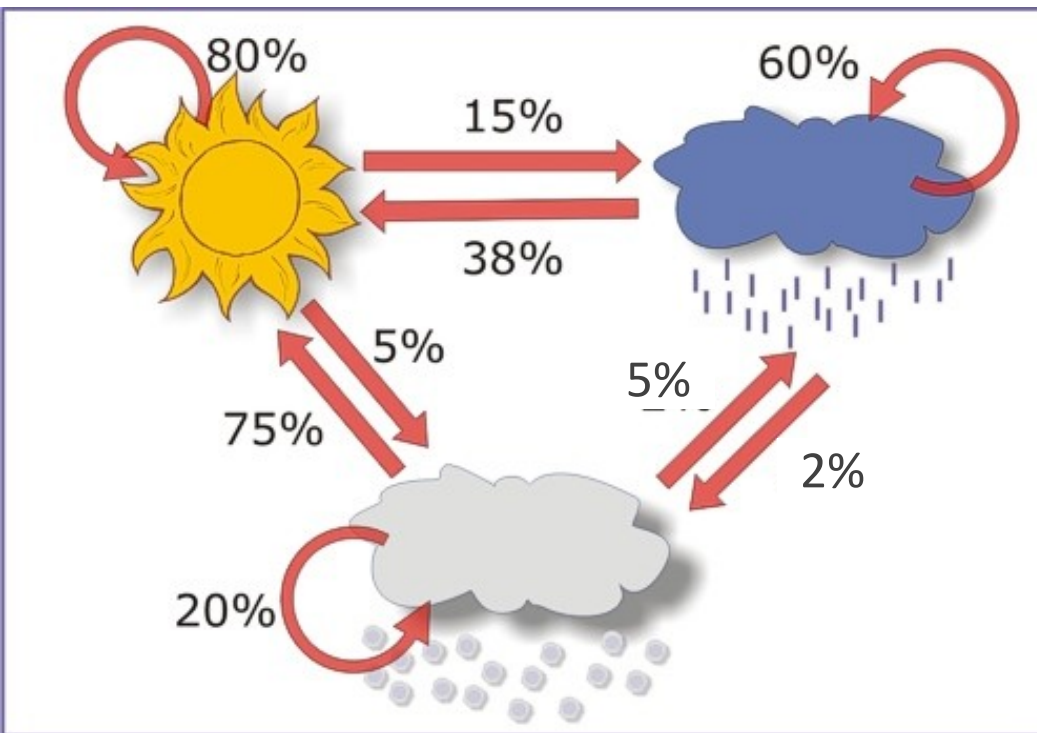
GC skew – příklad využití pro identifikaci počátku replikace



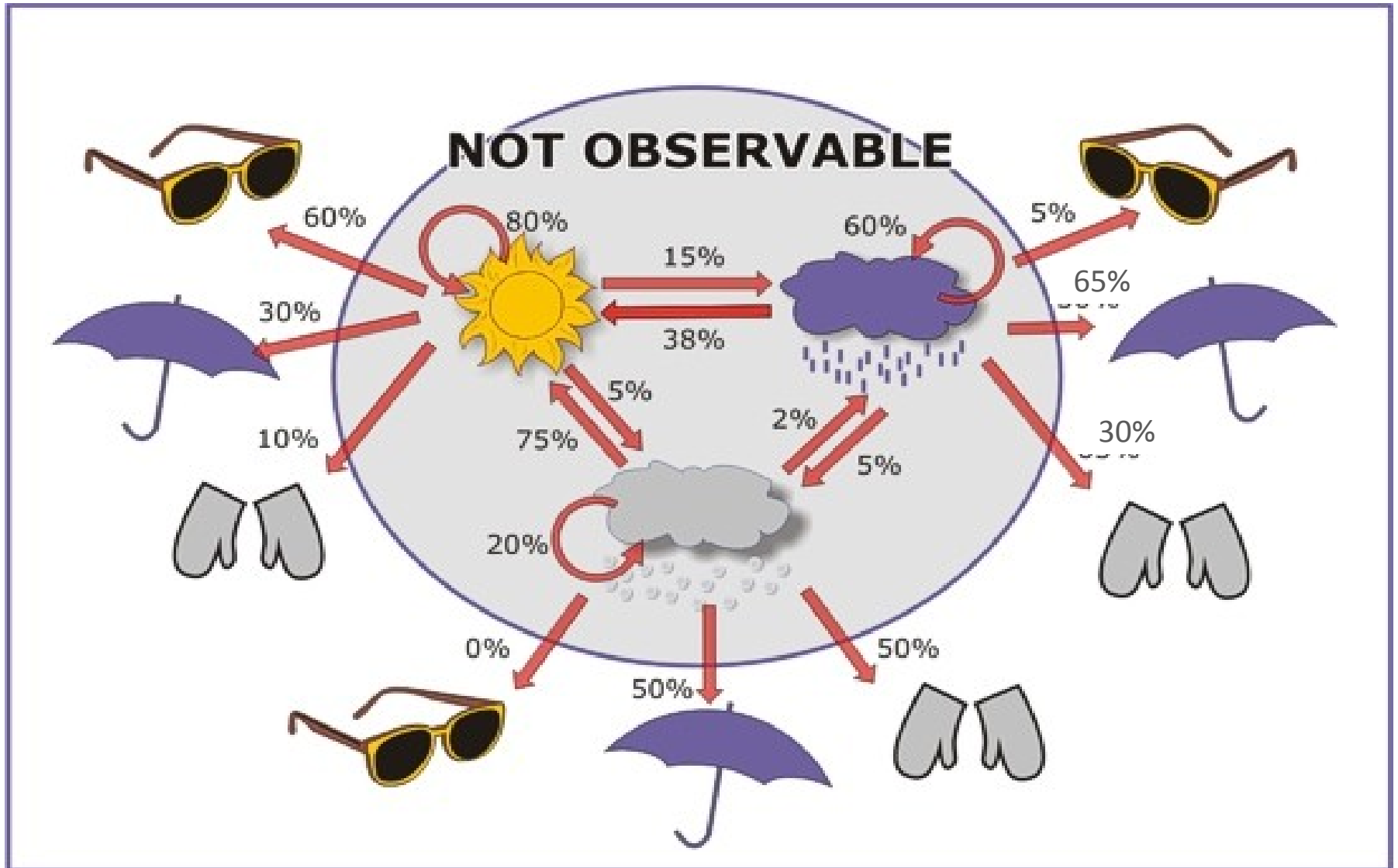


Markovovy modely

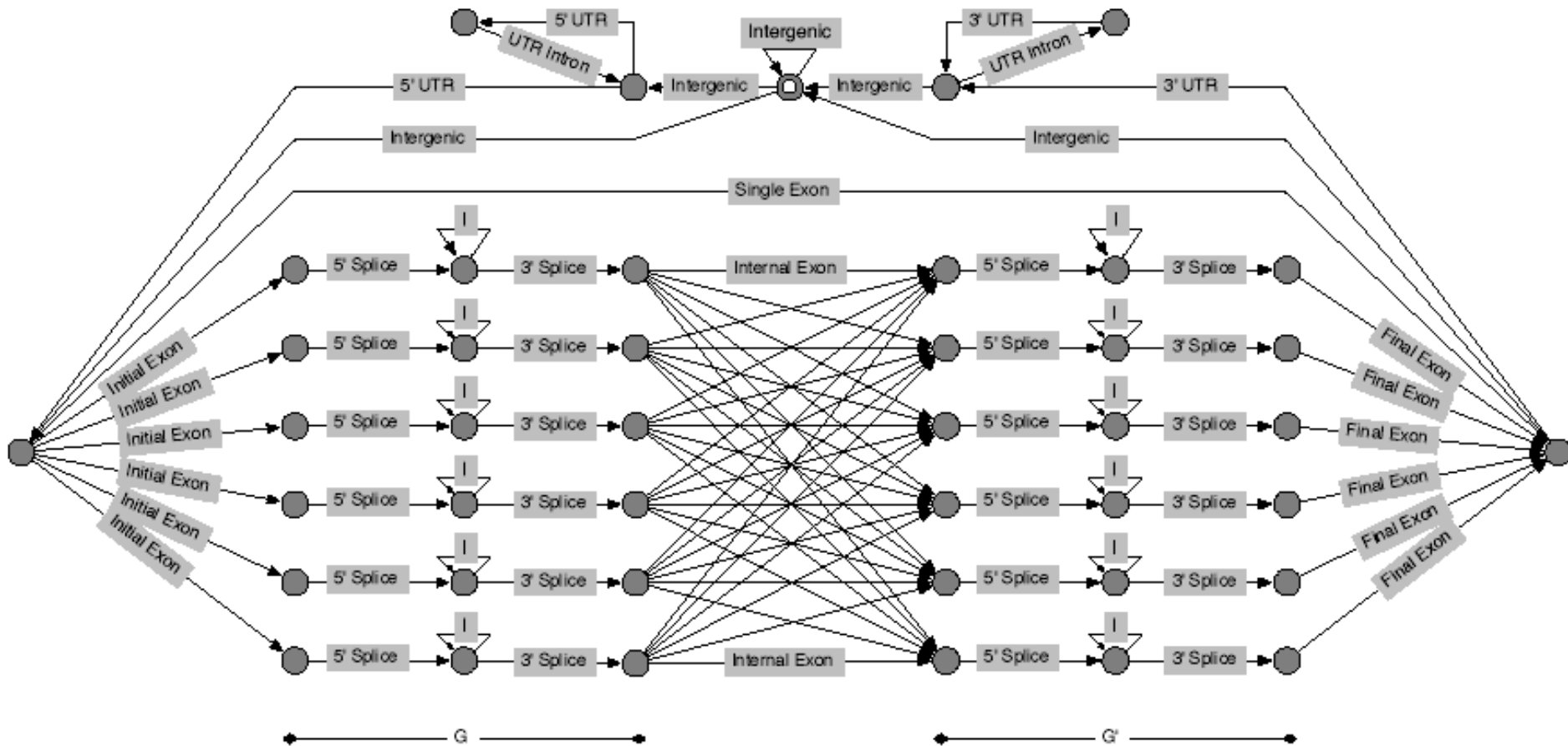
- Nejčastěji používané statistické modely pro hledání genů
- Vyjadřují pravděpodobnost sekvenčních událostí



Hidden Markov Models (HMM)



Příklad komplexního algoritmu se skrytými Markovovými modely (HMM)





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

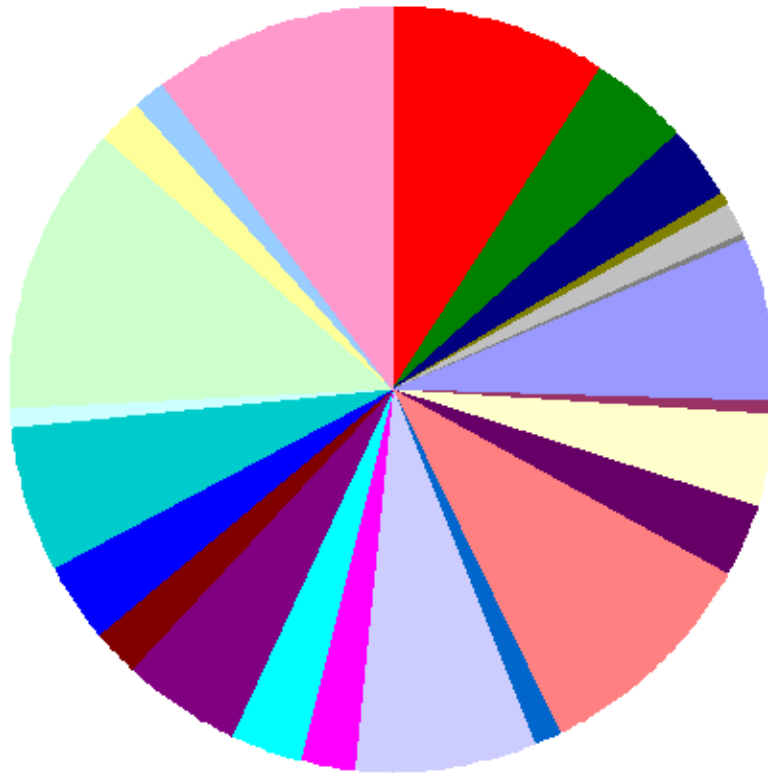
Subsystem Statistics

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

