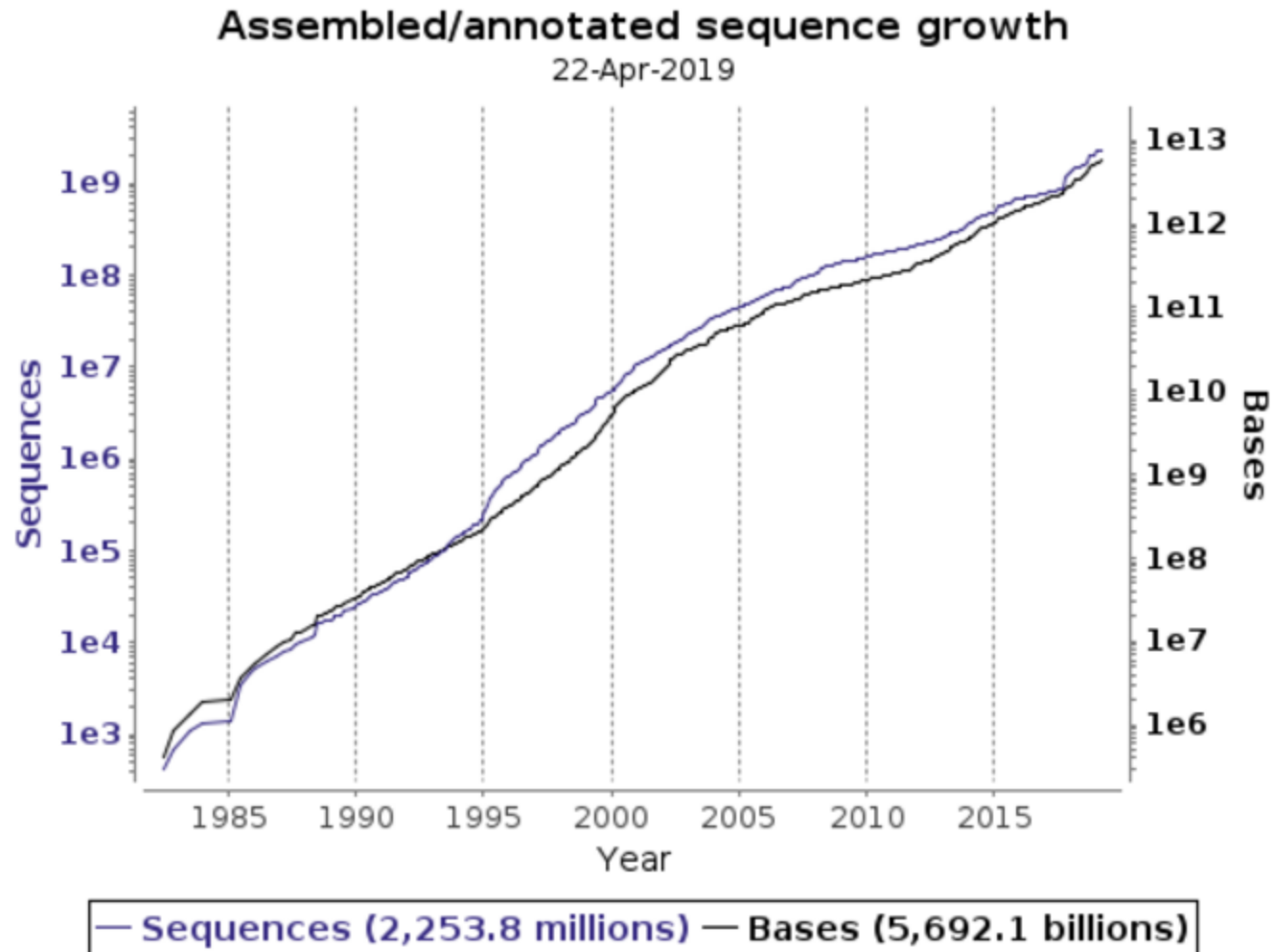


Bioinformatika je disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie

- 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

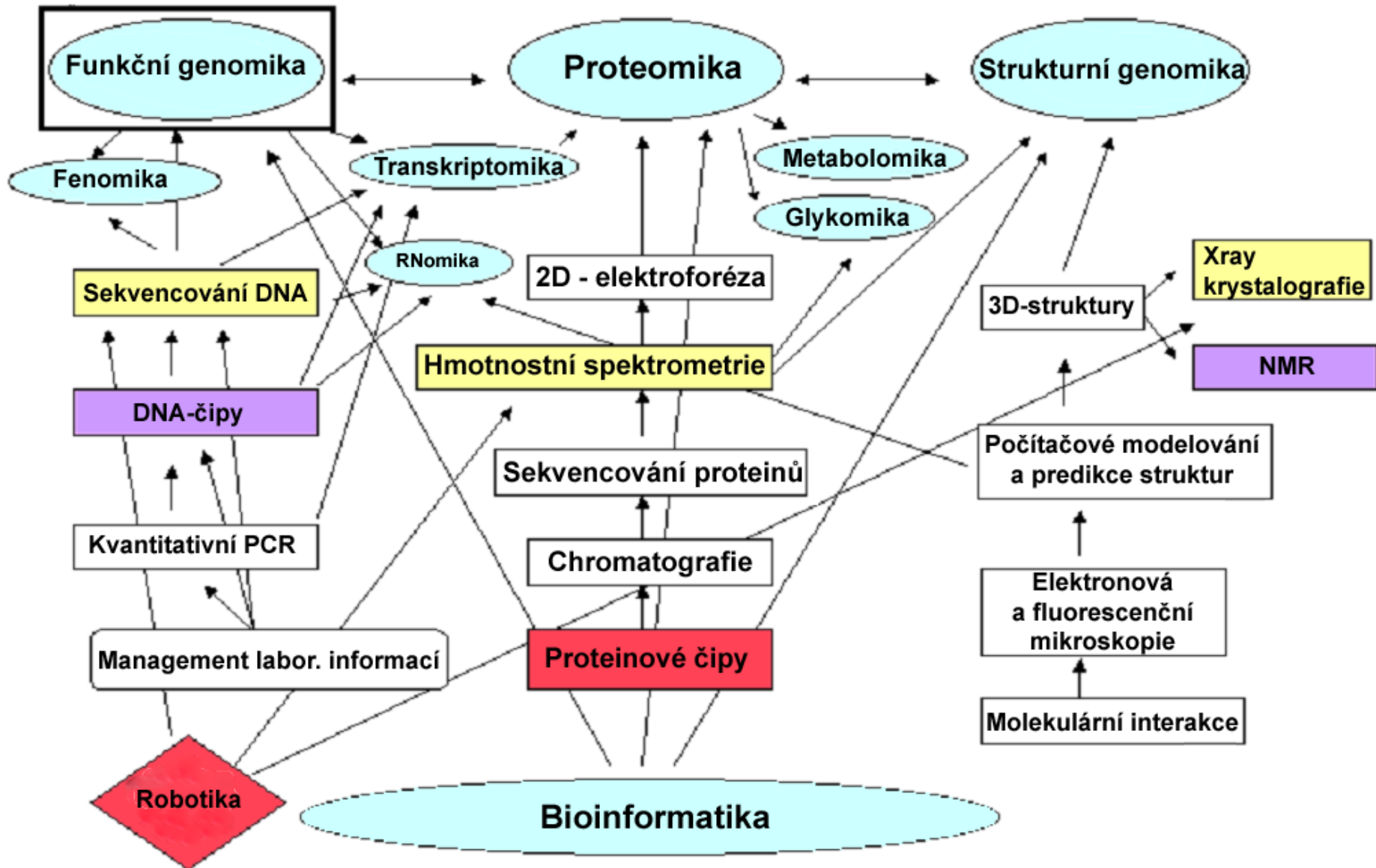
Assembled/annotated sequence growth



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



- Mezi hlavní oblasti zájmu bioinformatiky patří studium širokého rozmezí biologických dat, zejména
 - sekvencí nukleových kyselin
 - sekvencí proteinů
 - genů a genových map
 - expresních profilů
 - organizace genomů
 - interakce proteinů
 - mechanismy fyziologických funkcí

Nejdůležitější instituce zabývající se shromažďováním biomedicínských informací

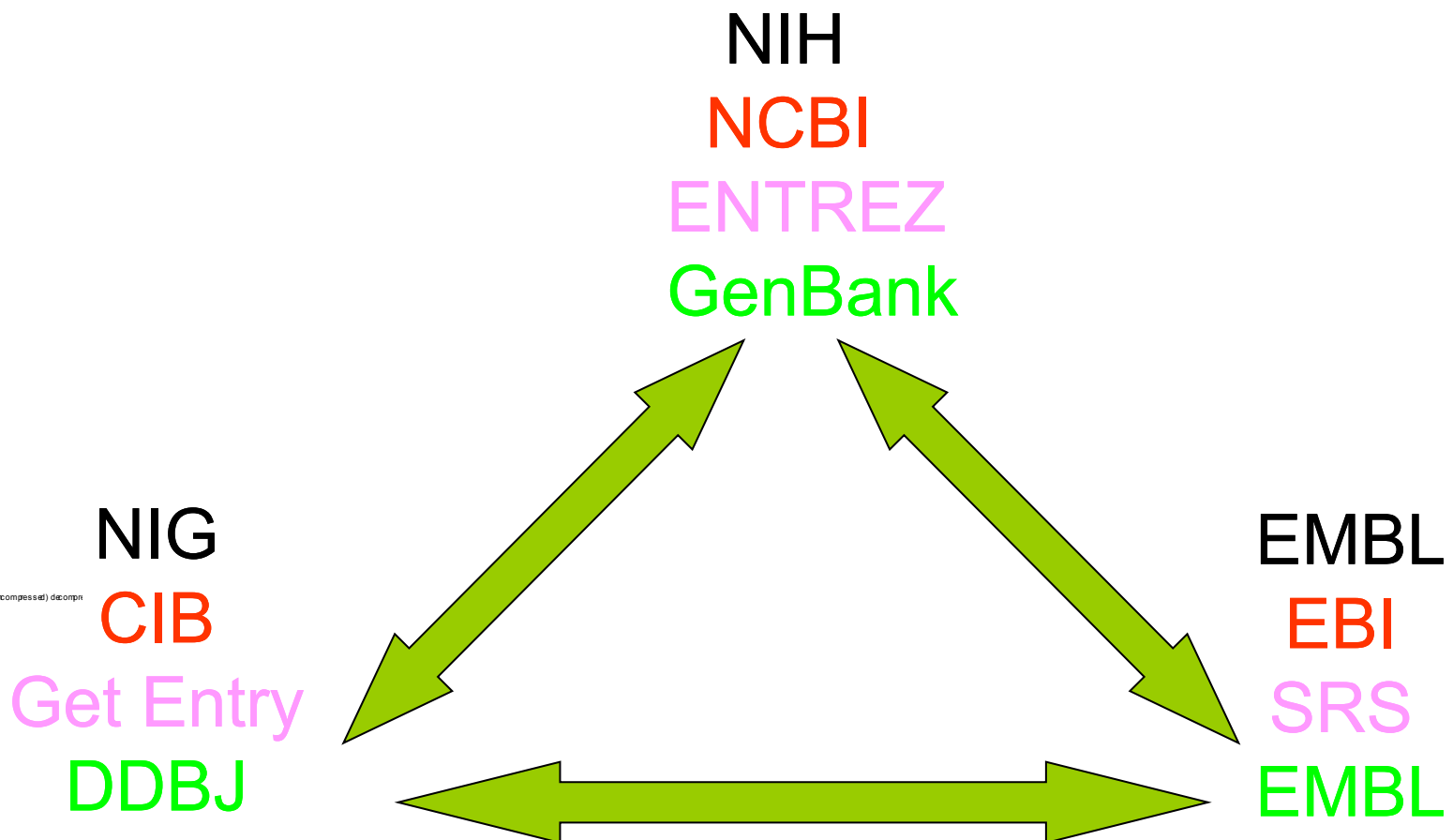
- V současné době je prostřednictvím Internetu dostupných přibližně 550 databází zabývajících se shromažďováním bioinformací.
 - Jejich přehled a popis je každoročně publikován ve specializovaném, volně dostupném čísle časopisu [Nucleic Acids Research](#).
- K nejdůležitějším institucím zabývajícím se, 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/>).

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.

Mezinárodní spolupráce sekvenčních databází

- Databáze sdílejí stejná data



- 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.

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

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     Macrocooccus 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]'. To the left of the screenshot, several red arrows point to the 'DEFINITION', 'ACCESSION', 'VERSION', 'KEYWORDS', 'SOURCE', 'ORGANISM', 'REFERENCE', and 'FEATURES' sections of the entry.

This block shows a detailed view of the 'FEATURES' section of the GenBank entry. It lists various features with their locations and qualifiers. Red arrows point to specific features: 'source', 'gene', 'CDS', and the 'translation' line. The 'translation' line shows the amino acid sequence: HAKNGVDVVS...EFNGKCTTKF. Below the features is the 'BASE COUNT' section, which shows the counts for each base (a, c, g, t) and others. The 'ORIGIN' section shows the nucleotide sequence in 60-base blocks.

FEATURE	Location/Qualifiers
source	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 /gene="pNAox"
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="FKGIMNEIAELDQRNIEMHNYLTSVYEEGDARSALITMLQALN HAKNGVDVVS...EFNGKCTTKF"
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 agaattggtg cgatgtagtg tctggmactc gagtccggac 181 acatthttgca agaccaaatt ttaagagggt gctgtctaag gtagccgcca aacatcctta 241 tgccaagata ggagtgttct attgogggagc tccagttctg gccgagggaac taagcaacct 301 ttgccatgag ttcaatggca aatgcaagcag aaaatc

Genomové databáze v NCBI – prokaryota





Genome

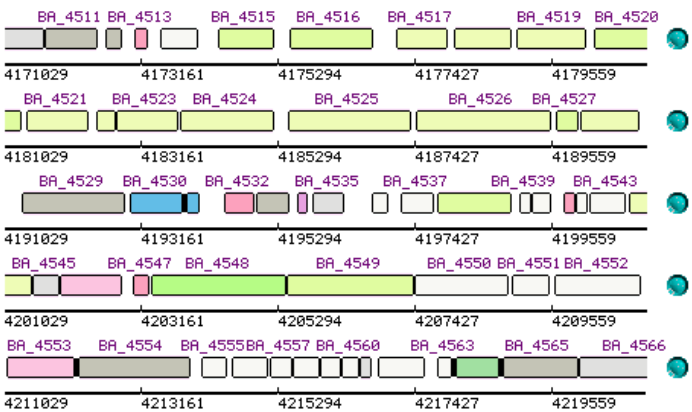
[BLAST](#) [PubMed](#) [Nucleotide](#) [Protein](#) [Genome](#) [Structure](#) [PopSet](#) [Taxonomy](#) [Help](#)

Bacillus anthracis A2012, unfinished sequence - 4171029..4221028

Start from : Search for gene

[57 protein coding genes](#) ● [Find Open Reading Frames](#)

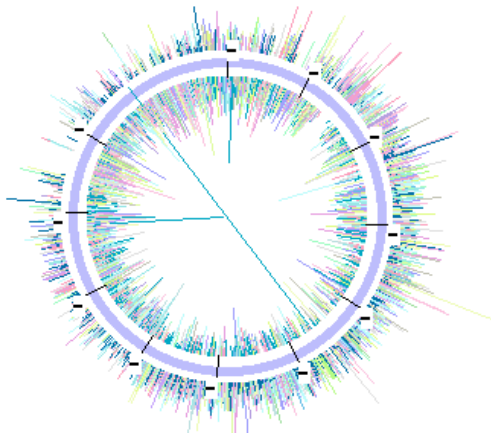
Click on the rectangle to get BLAST neighbors for the gene of interest or click on the overview below to see a distant region



← →

- Translation, ribosomal structure and biogenesis
- Transcription
- DNA replication, recombination and repair
- Cell division and chromosome partitioning
- Posttranslational modification, protein turnover, and

Protein coding genes distribution map
To see map locations of genes, click on a region in the map, to zoom in on that region



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](#)

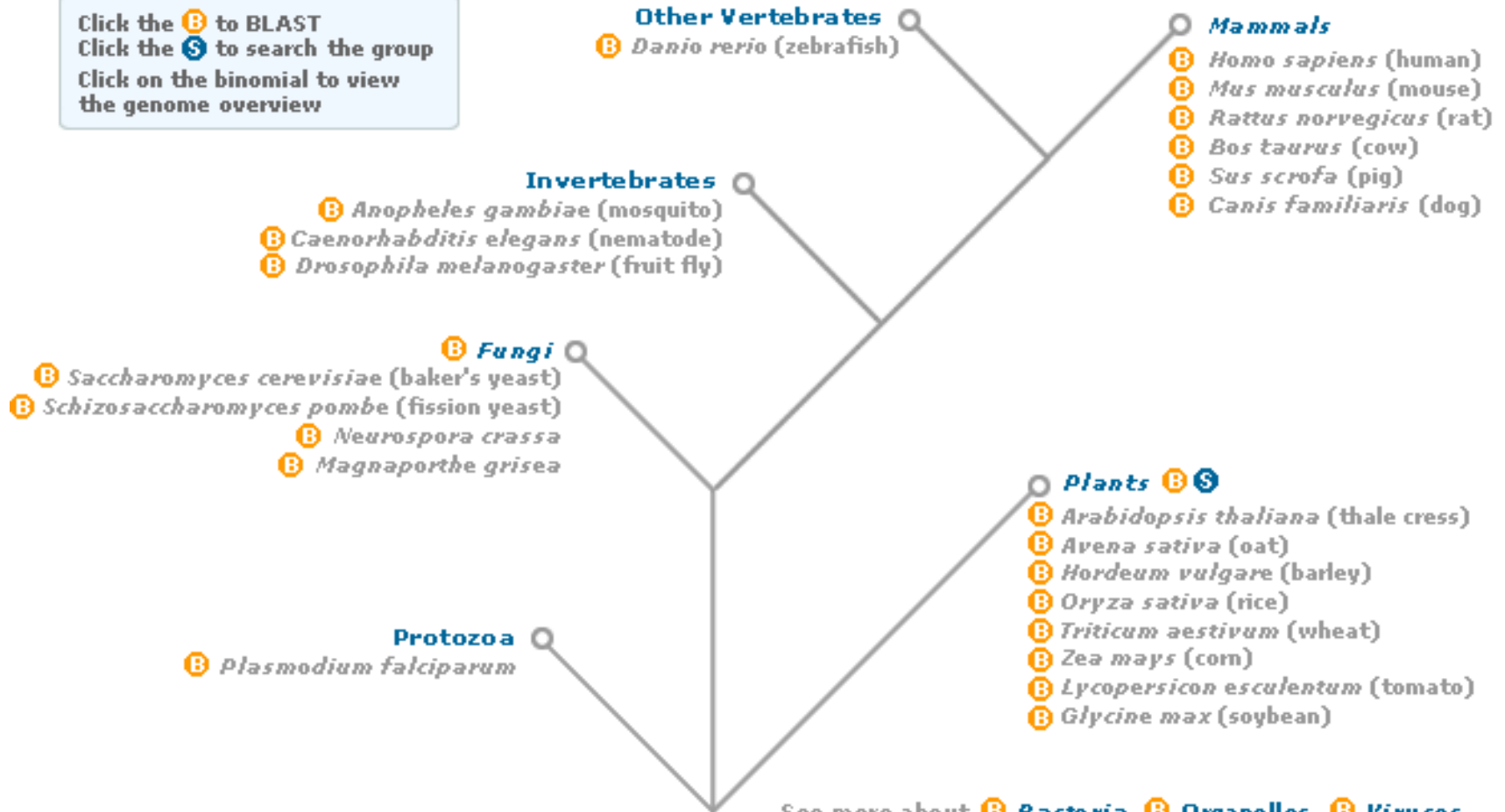
Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish. Arabidopsis is not of major agronomic significance, but it offers important 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.

Gemonové mapy - MapView

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

Click the **B** to BLAST
Click the **S** to search the group
Click on the binomial to view
the genome overview



[MapViewer Home](#)

[Map Viewer Help](#)
[Drosophila Maps Help](#)
[FTP](#)

[Data As Table View](#)

Maps&Options

Compress Map

Region Shown:

Go



Cyto



Drosophila melanogaster Map View

Chromosome: [X](#) [2L](#) [2R](#) [[3L](#)] [3R](#) [4](#)

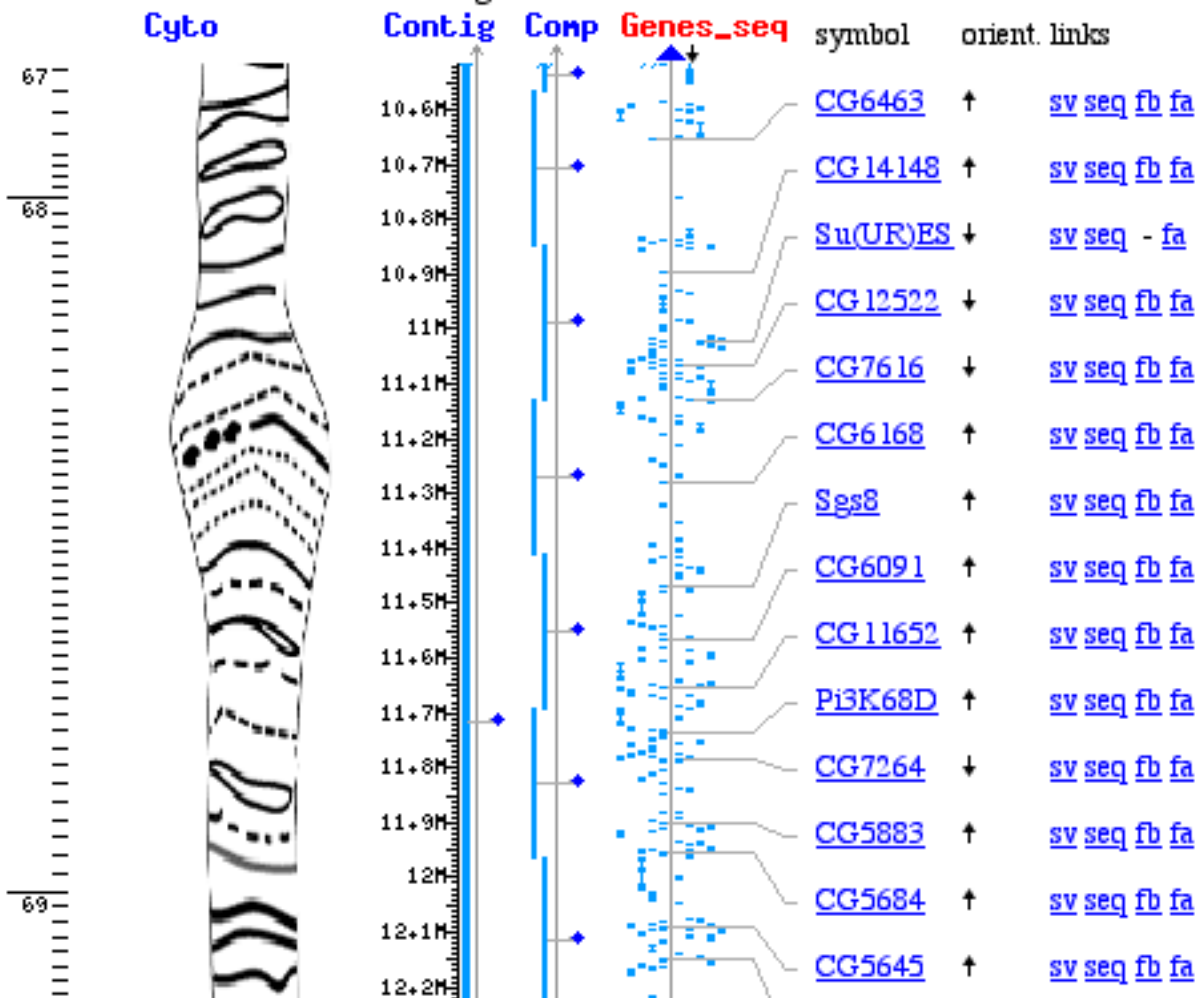
Master Map: Genes On Sequence

Total Genes On Chromosome: 2617

Region Displayed: 10M-12M bp [Download/View Sequence/Evidence](#)

Genes Labeled: 20 Total Genes in Region: 286

Maps & Options



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.

- EXPASY <http://www.expasy.ch>

Site Map	Search ExpASY	Contact us
Hosted by CBR Canada Mirror sites: Bolivia China Korea Switzerland Taiwan USA		



ExPASy Molecular Biology Server

The ExPASy (Expert Protein Analysis System) [proteomics](#) server of the [Swiss Institute of Bioinformatics](#) (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE ([Disclaimer](#) / [Reference](#)).

[\[Announcements\]](#) [\[Job opening\]](#) [\[Mirror Sites\]](#)

Databases	Tools and software packages
<ul style="list-style-type: none"> ● SWISS-PROT and TrEMBL - Protein knowledgebase ● PROSITE - Protein families and domains ● SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis ● ENZYME - Enzyme nomenclature ● SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules ● SWISS-MODEL Repository - Automatically generated protein models ● CD40Lbase - CD40 ligand defects ● SeqAnalRef - Sequence analysis bibliographic references ● Links to many other molecular biology databases 	<ul style="list-style-type: none"> ● Proteomics and sequence analysis tools <ul style="list-style-type: none"> ○ Proteomics [PeptIdent, PeptideMass, ...] ○ DNA -> Protein [Translate] ○ Similarity searches [BLAST] ○ Pattern and profile searches [ScanProsite] ○ Post-translational modification and topology prediction ○ Primary structure analysis [ProtParam, pI/MW, ProtScale] ○ Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer] ○ Alignment [T-COFFEE, SIM] ○ Biological text analysis ● Melanie 3 - Software for 2-D PAGE analysis ● Roche Applied Science's Biochemical Pathways

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

- PDB <http://www.rcsb.org/pdb/>

[DEPOSIT data](#)
[DOWNLOAD files](#)
[browse LINKS](#)
[BETA TEST new features](#)
[BETA mmCIF files](#)

Current Holdings

[19623 Structures](#)
[Last Update: 30-Dec-2002](#)
[PDB Statistics](#)



[Molecule of the Month: Cytochrome c](#)

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the National Institute of Standards and Technology -- three members of the [Research Collaboratory for Structural Bioinformatics \(RCSB\)](#). The PDB is supported by funds from the [National Science Foundation](#), the [Department of Energy](#), and two units of the National Institutes of Health: the

PROTEIN DATA BANK



Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

[Did you find what you wanted?](#)

[ABOUT PDB](#) | [DATA UNIFORMITY](#) | [RECENT FEATURES](#) | [USER GUIDES](#) |
[FILE FORMATS](#) | [EDUCATION](#) | [STRUCTURAL GENOMICS](#) | [PUBLICATIONS](#) |
[SOFTWARE](#)

Search the Archive



Enter a [PDB ID](#) or keyword

[Query Tutorial](#)

- query by PDB id only match exact word
 remove sequence homologues

[SearchLite](#) keyword search form with examples
[SearchFields](#) customizable search form
[Status Search](#) find entries awaiting release

News

[Complete News](#)
[Newsletter](#)

[pdb4 Archive](#)
[Subscribe](#)

23-Dec-2002

[Happy Holidays from the PDB!](#) The PDB staff wish to extend our [best wishes](#) to the community for a happy holiday season and a wonderful new year!



PDB Mirrors

Please bookmark a mirror site

[San Diego Supercomputer Center*](#)

[Rutgers University*](#)

[National Institute of Standards and Technology*](#)

[Cambridge Crystallographic Data Centre, UK](#)

[National University of Singapore](#)

[Osaka University, Japan](#)

[Universidade Federal de Minas Gerais, Brazil](#)

[Max Delbrück Center for Molecular Medicine, Germany](#)

[OTHER SITES](#)

+ - - - -

PubMed BLAST Structure Taxonomy

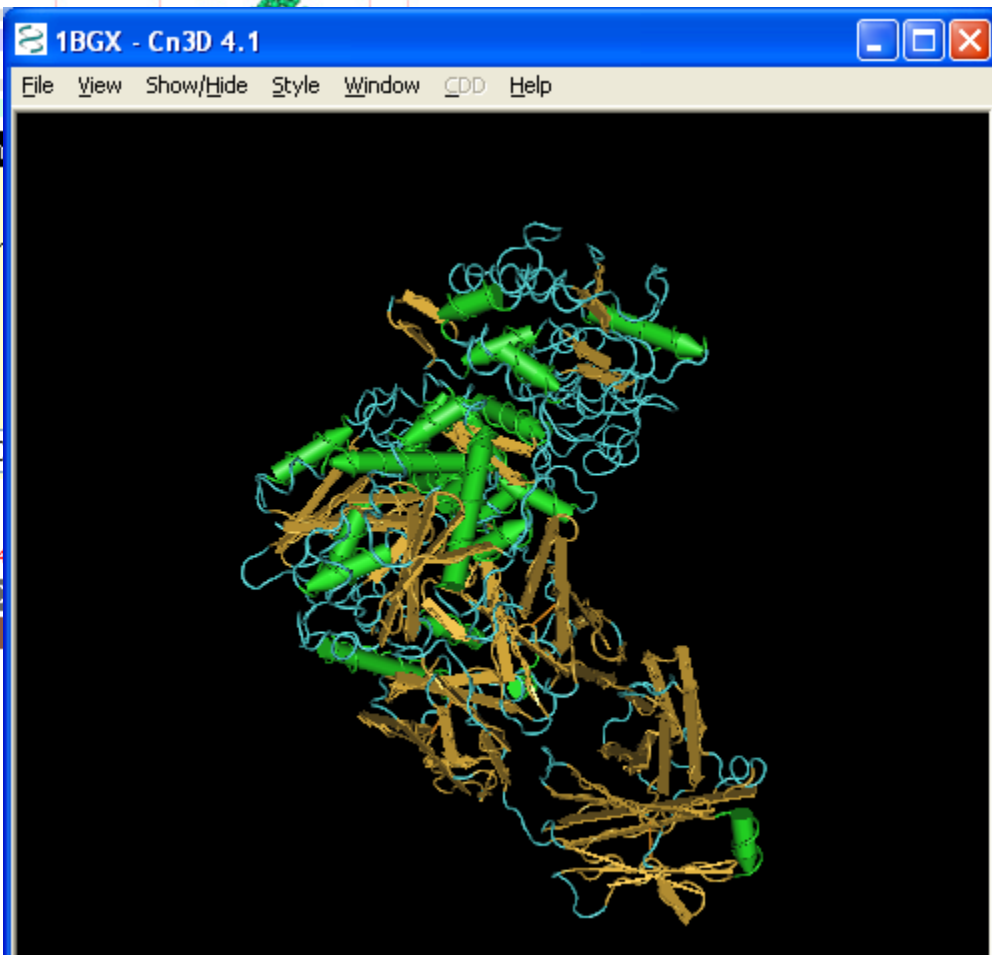
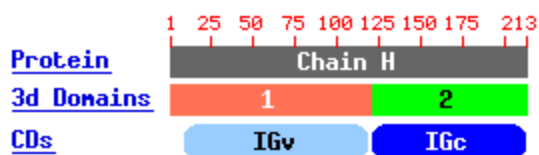
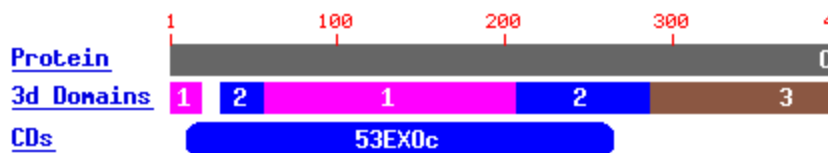
Description: Taq Polymerase In Complex With Tp7, An Inhibitor

Deposition: R.Murali, D.J.Sharkey, J.L.Daiss & H.M.Krishna

Taxonomy: T [Thermus aquaticus](#); H, L [Mus musculus](#)

Reference: [PubMed](#) **MMDB:** [8845](#) **PDB:** [1BGX](#)

View 3D Structure of Best Model with Cn3D



1BGX - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

P	1BGX_T	m r g m l p l f e p k g r v l l v d g h h l a y r t f h a l k g l t t s r g e p v q a v y g f a k s l l k a l k e d g d a v i v v f d a k a p s f r h e a y g g y k a g
3	1BGX_H	e v q l q e s g p g l v k p y q s l s l s c t v t g y s i t s d y a w n w i r q f p g n k l e w m g y i t y s g t t d y n p s l k s r i s i t r d t s k n q f f l q l n s
C	1BGX_L	d i q m t q s p a i m s a s p g e k v t m t c s a s s s v s y m y w y q q k p g s s p r l l i y d s t n l a s g v p v r f s g s g s g t s y s l t i s r m e a e d a a t y

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/>

- **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 left is the NCBI logo. To its right is the Entrez logo and the text "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. Below the search bar is a welcome message: "Welcome to the new Entrez cross-database search page". The main content area is a grid of database icons and descriptions, each with a help icon (question mark in a circle). The databases listed include PubMed, PubMed Central, Journals, MeSH, Nucleotide, Protein, Genome, Structure, Taxonomy, SNP, Gene, Books, OMIM, Site Search, UniGene, CDD, 3D Domains, UniSTS, PopSet, GEO, and GEO DataSets.

Entrez Molecular Sequence Database System

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

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search for

SITE MAP
Guide to NCBI resources

About NCBI
The science behind our resources. An introduction for researchers, educators and the public.

GenBank
Sequence submission support and software

Molecular databases
Sequences, structures and taxonomy

Literature databases
PubMed, OMIM, Books and PubMed Central

Genomic biology
The human genome, whole genomes and related resources

What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

Hot Spots

- ▶ Cancer genome anatomy project
- ▶ Clusters of orthologous groups
- ▶ Coffee Break
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human/mouse homology maps
- ▶ LocusLink

Mouse Genome
Resources: explore tools for manipulating the mouse genome.

Try these: Map Viewer Sequencing Progress Human-Mouse Homology

BLink and get results fast!

Sequence Retrieval System (SRS)

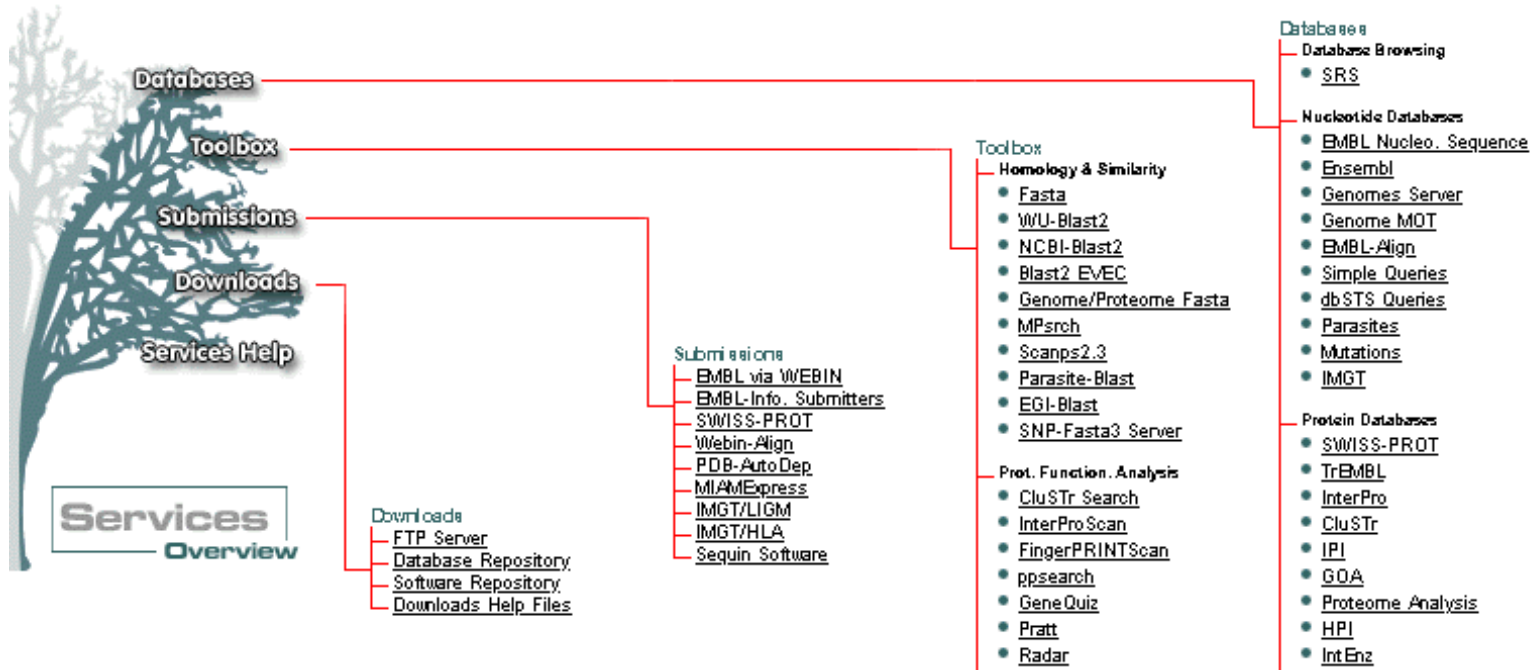
EBI <http://www.ebi.ac.uk/>

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



SRS

<http://srs.ebi.ac.uk/>

- Na serveru EBI
- SRS je homogenní rozhraní pro přístup k více než 160 molekulárně databázím
- Typy databází zahrnují
 - sekvence a z nich odvozená data
 - metabolické dráhy
 - transkripční faktory
 - 3-D struktury
 - Genomy
 - Mapování
 - Mutace
 - jednonukleotidové polymorfizmy
 - výsledky získané pomocí analytických nástrojů
- Webové rozhraní umožňuje provádět před vyhledáváním výběr z jednotlivých databází a poskytuje alternativní formuláře pro zadávání vyhledávacích dotazů.
- Na Internetu běží několik verzí SRS a každá z nich obsahuje jinou sadu databází a analytických nástrojů.

SRS Release 7.1.1 Copyright © 1997-2003 ICRF BioScience AG. All Rights Reserved. Terms of Use Feedback

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
- Poskytuje přístup do databází, které mohou být dotazovány samostatně.
- Jako výsledek DBGET prezentuje kromě seznamu vyhledaných záznamů také přehled vazeb na související informace ve všech integrovaných databázích.
- Další ojedinělou vlastností je propojení na databázi KEGG (Kyoto Encyclopedia of Genes and Genomes), což je databáze regulačních a metabolických drah u organismů ze známým genomem.
- V porovnání se SRS a Entrez je však DBGET jednodušší a omezenější vyhledávací prostředek.

Posuzování podobnosti sekvencí

Nástroje pro vyhledávání lokálních
podobností sekvencí

Postup stanovení podobnosti

- textové vyhledávání příbuzných sekvencí v databázích
- prohledávání databází podle podobnosti sekvencí
- výpočet lokálního přiřazení (alignment)
= uspořádání do 2 pod sebou ležících řádků tak, aby identické zbytky ležely pod sebou

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**.

Využívají heuristickou analýzu pro identifikaci krátkých homologických subsekvencí bez mezer s následným rozšiřováním vyhledávání v okolí subsekvencí s cílem získat lokálně seřazené sekvence, do nichž mohou být vloženy mezery

- BLAST
- Altschul et al., [1990](#)
- dostupný na serveru NCBI
- FASTA
- Lipman a Pearson [1985](#)
- dostupný na serveru EBI

Co je to BLAST?

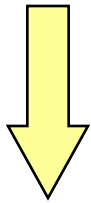
- **Basic Local Alignment Search Tool**
 - Hledání lokálních podobností
 - Heuristický přístup založený na Smith-Watermanově algoritmu
 - Vyhledá neoptimálnější **seřazení sekvencí**
 - Poskytuje data o statistické významnosti
 - Zobrazuje vzájemně seřazené sekvence
 - Lokalizuje oblasti sekvencí s vysokou podobností a umožňuje zobrazení jejich primární struktury a funkce
 - Literatura: [Nucleic Acids Res.](#) 2004 Jul 1;32(Web Server issue):W20-5.

OPEN ACCESS
OXFORD JOURNALS

FREE full text article
in PubMed Central

BLAST: at the core of a powerful and diverse set of sequence analysis tools.

[McGinnis S, Madden TL.](#)



Výchozí stránka BLAST

BLAST Basic Local Alignment Search Tool

My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)

News

[New Human and Mouse pre-indexed databases](#)
Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.
2007-09-04 10:55:00
[More BLAST news...](#)

Tip of the Day

Using Genomic BLAST

Genomic BLAST pages are helpful because they allow the genomic context of a BLAST search to be displayed in the Map Viewer. For example, discontinuous (cross-species) MegaBLAST against the human RefSeq transcript for albumin (NM_000477) can be used to identify the homolog in the rat genome.

[More tips...](#)

<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 homologních proteinů
<u>blastx</u>	DNA	Protein	Protein	Hledání genů a homologních proteinů na DNA
<u>tblastn</u>	Protein	DNA	Protein	Hledání genů u necharakterizovaných DNA
<u>tblastx</u>	DNA	DNA	Protein	Studium struktury genů

Jak BLAST pracuje?

- Proces zahrnuje 3 kroky
 1. Příprava dotazu
 - rozseká 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

Slova pro nukleotidové sekvence

Dotaz: **GTACTGGACATGGACCCTACAGGAA**

~~GTACTGGACAT~~

Velikost slova = 11

minimální velikost = 7

TACTGGACATG

blastn default = 11

tabulka se všemi slovy dotazu
ACTGGACATGG megablast default = 28

CTGGACATGGA

TGGACATGGAC

GGACATGGACC

GACATGGACCC

ACATGGACCCT

.....

Slova pro proteinové sekvence

Dotaz: **GTQITVEDLIFYNIATRRKALKN**

GTQ
Velikost = 3

Velikost slova může být 2 nebo 3 (default = 3)

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**

GTQITVEDLFIYNI

SEI

YIN

sousedící slova

2 nalezené shody

Substituční Matice

- Co je substituční matice?
 - Kompletní sada skóre pro všechny kombinace párů zbytků se nazývá substituční matice
 - Uplatňuje se při srovnání sekvencí proteinů
 - Stanovuje frekvenci při které každý možný zbytek v sekvencích může být změněn za kterýkoli jiný zbytek během času (evoluce)
 - Např., hydrofobní zbytek má vyšší pravděpodobnost zachování v příslušné pozici sekvence než jiný.
 - Každá matrice je určena pro určitý typ vyhledávání – JE TŘEBA VĚDĚT CO HLEDÁME!

Substituční Matice

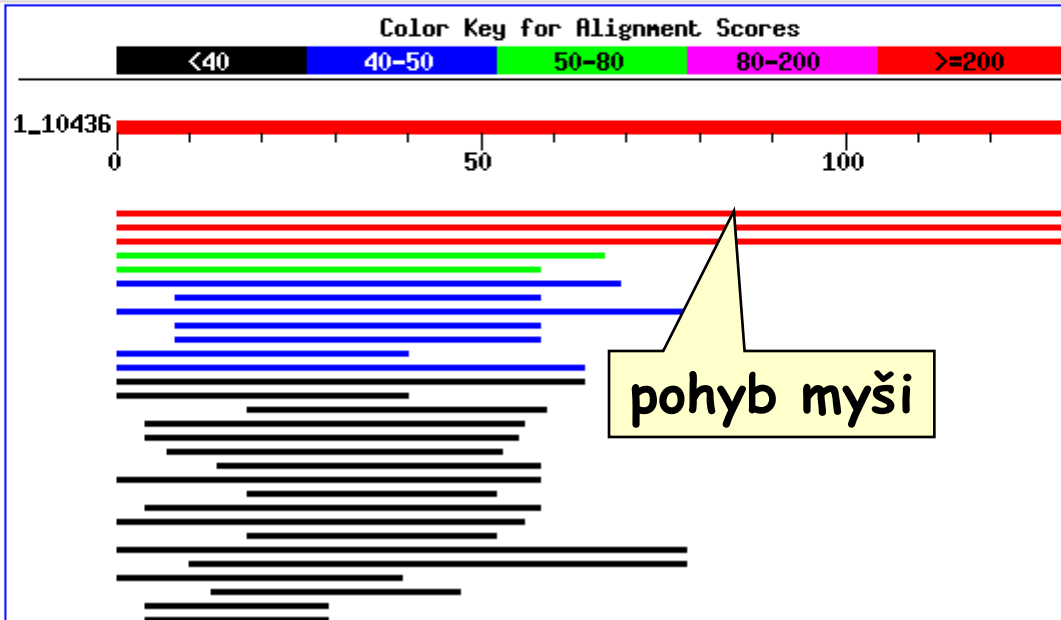
- Proč používat substituční matice?
 1. Stanovit pravděpodobnou homologii dvou sekvencí.
 2. Substituce, které jsou více pravděpodobné získají vyšší skóre
 3. Substituce, které jsou méně pravděpodobné obdrží nižší skóre.

BLAST – grafický výstup

[Taxonomy reports](#)

Distribution of 30 Blast Hits on the Query Sequence

P40692 DNA mismatch repair protein Mlh1 (MutL protein homolog 1..S= 233 E=8e-62



BLAST – příklad výstupu u DNA



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

Příklad analýzy mnohonásobného příložení

The screenshot displays the SeqLab Main Window on mendel. The main window shows a list of sequences and a sequence alignment. The alignment is displayed in a grid format with columns representing positions (10-110) and rows representing sequences (ef1a_EIMBO, ef11_human, ef11_crigr, ef11_mouse, ef10_xenla, ef1a_chick, ef12_human, ef12_mouse, ef1a_rhyam, ef1a_oryla, ef1a_brare, ef13_xenla, ef12_xenla, ef1a_bommo, ef12_drome, ef1a_helvi, ef1a_spojr, ef1a_artsa, ef11_drome). The alignment is color-coded by features. A dialog box titled "Sequence Features" is open, showing a list of features and a detailed view of a selected feature.

SeqLab Main Window on mendel

File Edit Functions Options Windows Help

List: /users/thompson/seqlab/working.list

Mode: Editor Display: Features Coloring 1:1

CUT COPY PASTE PROTECT INFO GROUP Insert Wrap Invert

EF1A_EIMBO GTSQADVALLVVPADQGGFEGAFSKEGQTRHALLAFTLVGKQMIVGI**NKMD**ATTPDKYSETR
ef11_human MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef11_crigr MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef11_mouse MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef10_xenla MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef1a_chick MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef12_human MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef12_mouse MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef1a_rhyam **SF**KYAWVL**DKLKAERERGITIDIAL**LWKFETAKYYVTI**IDAPGH**RDFIKNMITGTSQADCAVLI
ef1a_oryla MGKEKTHINIVVI**GHVDSGKS**TSTGHLIYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef1a_brare MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef13_xenla MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef12_xenla MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef1a_bommo MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef12_drome MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef1a_helvi **HVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKLKAERERGITIDIAL**LW
ef1a_spojr **HVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKLKAERERGITIDIAL**LW
ef1a_artsa **GKEK**THINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKL**
ef11_drome MGKEKTHINIVVI**GHVDSGKS**TTTGHLYKCGGID**KRTIEKFEKEAAEMGKGSF**YAWVL**DKLKAERERGITIDIAL**LWKFETAKYYVTI**IDAPGH**RDFIKNMITGTSQADCAVLI

pos:17 col:17 ef11_human -->

Sequence Features

Edit Add Raise Delete

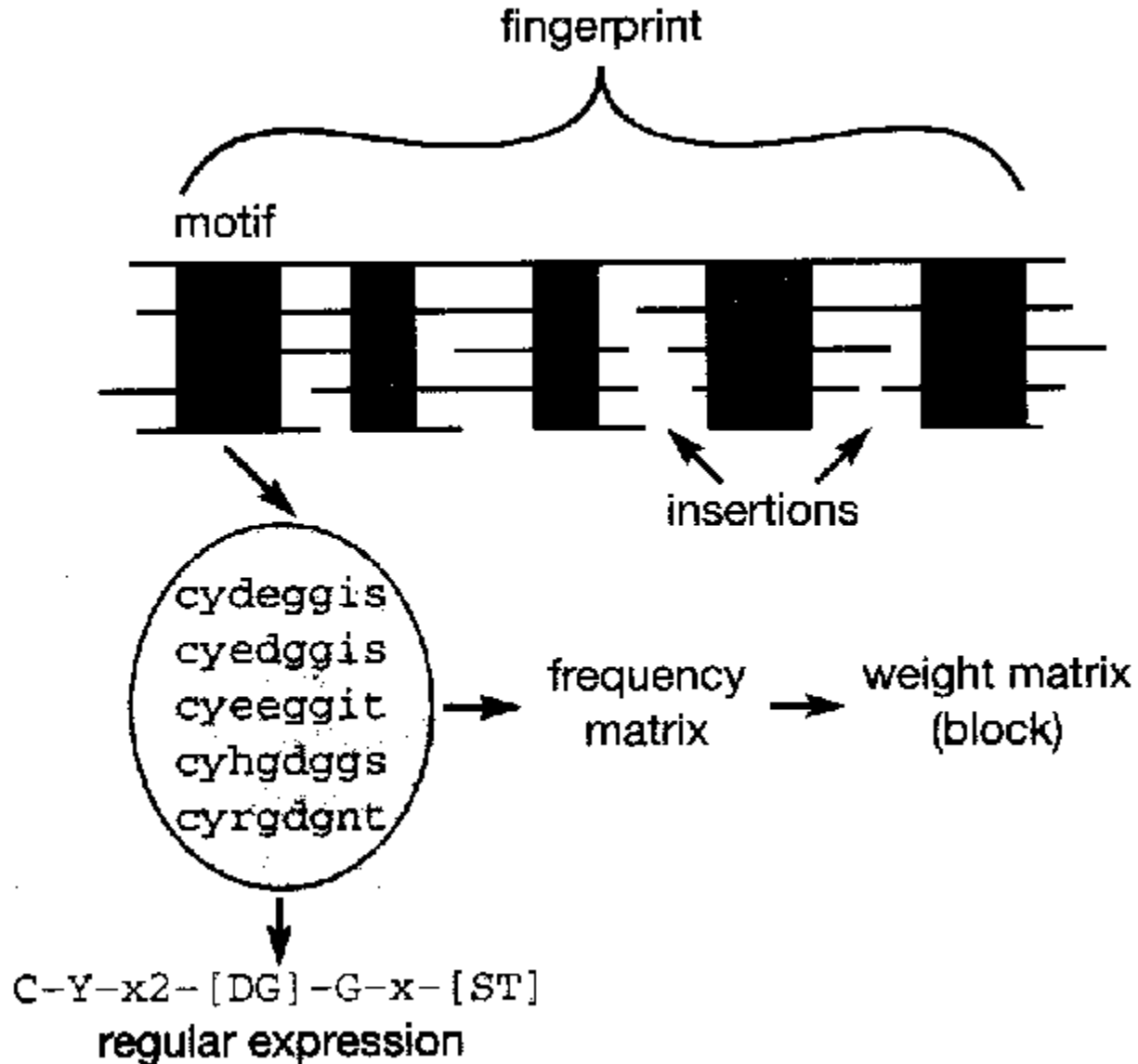
Show: Features at cursor

1	NP BIND	(14-21)
13	Motifs Match	(14-21)

FeatureStart 14
FeatureEnd 21
MotifName Atp_Stp_A

Close Help

Identifikace konzervativních motivů



Klasifikační databáze proteinů

- PROSITE
 - Pfam
 - PRINTS
 - ProDom
 - SMART
 - Blocks
 - InterPro
- Databáze sekvenčních motivů představují značně roztráštěný soubor zdrojů
 - Asi 30 databází
 - Částečně se překrývají, ale nejsou navzájem propojeny
 - Integrované vyhledávání ve více databázích umožňuje např. InterPro Scan

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

Vyhledání otevřených čtecích rámců

(<http://www.ncbi.nlm.nih.gov/projects/gorf/>)

The image shows the NCBI ORF Finder web interface. On the left is a dark blue sidebar with the NCBI logo and navigation links: PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. Below these are links for NCBI, Tools for data mining, GenBank sequence submission support and software, and FTP site for downloading data and software. The main content area has a title 'ORF Finder (Open Reading Frame Finder)' and a description of the tool. It includes a search form with a text input for 'GI or ACCESSION', 'OrfFind' and 'Clear' buttons, and a text area for 'FASTA format'. Below the text area are 'FROM:' and 'TO:' input fields, and a 'Genetic codes' dropdown menu currently 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

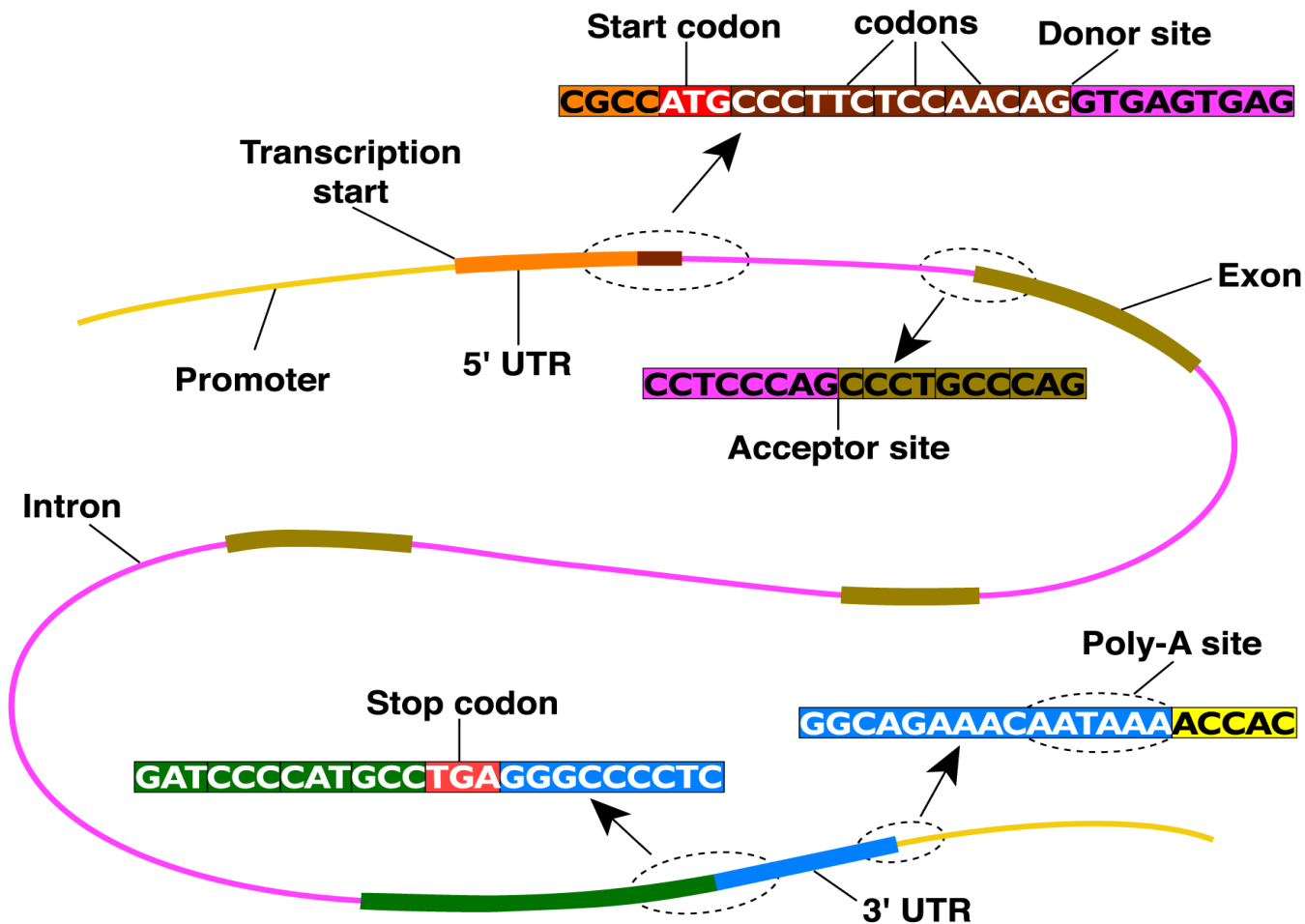
or sequence in FASTA format

FROM: **TO:**

[Genetic codes](#)

1 Standard

Signály – senzory ve struktuře eukaryotického genu



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!