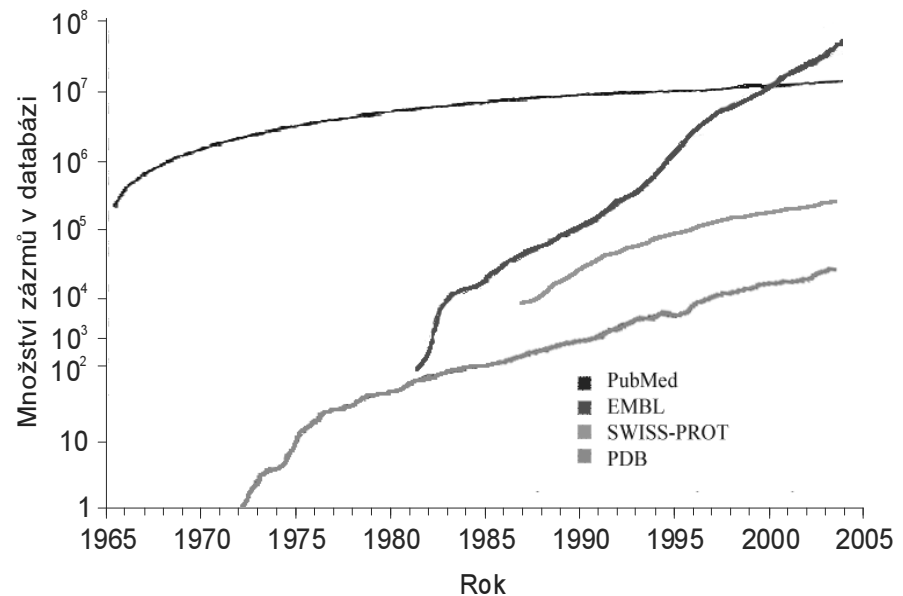


Bioinformatika je nová disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie

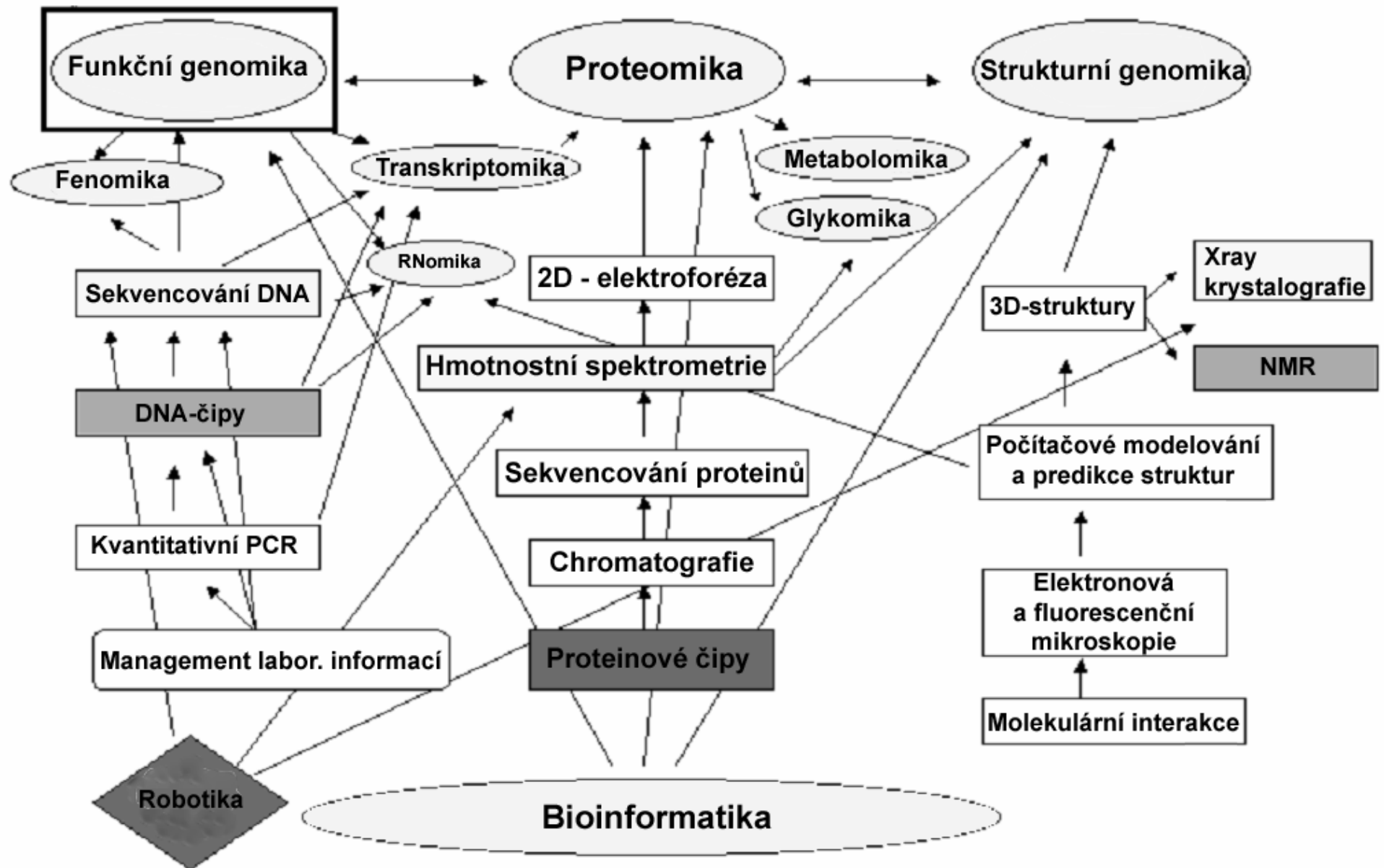
- Bioinformatika zahrnuje
 - studium
 - praktické uchovávání
 - vyhledávání
 - zobrazování
 - manipulaci
 - a modelování biologických dat
- Vývoj vysoce výkonných technologií umožňujících získání molekulárně biologických dat přispěl k jejich dramatickému nárůstu a tím současně zvýšil obtížnost jejich zkoumání a hodnocení ve vztahu k biologickým otázkám.



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		Seřazení sekvencí
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

Současné biotechnologické nástroje



- 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

- Primárním cílem těchto analýz je objasnění informačního obsahu biomolekul a porozumění, jak bioinformace přímo ovlivňují vývoj a funkce u živých organizmů.
 - **Hledání v databázích**
 - **Srovnávání sekvencí nukleových kyselin a proteinů**
 - **Hledání genů**
 - **Funkční genomika**
 - **Klasifikace proteinů**
 - **Fylogenetické studie**

Výuková stránka

<http://orion.sci.muni.cz/kgmb/bioinformat/>

Informační zdroje pro molekulární biologii

Odkazy pro výuku v přednášce BI5000 Úvod do bioinformatiky (zpracoval R.Pantlůček)

[Pravidla pro práci v učebně COMPB](#)

- [Úvod](#)
- [Literární informační zdroje pro biomedicínské obory](#)
- [Databázové zdroje pro molekulární biologii \(přehled\)](#)
- [Příklady formátů sekvencí](#)
- [Obecný princip sekvencování \[PDF\]](#)
- [Hledání v databázích sekvencí: nástroje pro lokální seřazení sekvencí](#)
- [Nástroje mnohonásobné a globální seřazení sekvencí DNA](#)
- [Fylogenetická analýza](#)
- [Návrh PCR-primerů a oligonukleotidů](#)
- [Restrikční analýza](#)
- [Počítačové vyhledávání genů](#)
- [Analýza sekundární struktury RNA](#)
- [Nástroje pro analýzu sekvencí proteinů](#)
- [Odkazy na komplexní sady nástrojů pro bioinformatiku](#)
- [Analýza metabolických drah](#)

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** (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.

- Všechny tři genomové databáze jsou volně dostupné a přijímají data získaná v genomových centrech nebo na odborných pracovištích zabývajících se sekvencováním nukleových kyselin.
- V současné době si jednotlivé databáze předávají získaná data, takže databanky GenBank/EMBL/DDBJ prakticky sdílejí stejná data v jakoukoli dobu.
- V současné době databáze EMBL obsahuje xxxxxxxxx sekvencí a xxxxxxxxxxx nukleotidových bází pocházejících celkem od více než 60 000 různých organismů nebo virů.
- Nové sekvence nukleových kyselin se do databází vkládají pomocí speciálního WWW formuláře nazvaného BankIt pro databázi GenBank, WebIn pro databázi EMBL nebo Sakura pro databázi DDBJ.

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

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM

Search Nucleotide for barley NADPH oxidase Go Clear

Limits Preview/Index History Clipboard Details

Display default Save Text Add to Clipboard Get Subsequence

1: AJ251717. Hordeum vulgare p... [gi:15282289] Links

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
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Pooideae; Triticeae; Hordeum.

REFERENCE 1

AUTHORS Huckelhoven,R., Dechert,C., Trujillo

TITLE Differential expression of putative near-isogenic, resistant and susceptible interaction with the powdery mildew

JOURNAL Plant Mol. Biol. 47 (6), 739-748 (2005)

MEDLINE 21643210

REFERENCE 2 (bases 1 to 337)

AUTHORS Hueckelhoven,R.

TITLE Direct Submission

JOURNAL Submitted (02-DEC-1999) Hueckelhoven, Phytopatholgy and Applied Zoology, Giessen, Ludwigstr. 23, 35390 Giessen

FEATURES 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 HAKNGVDVVSQTRVTRTHFARPNFKRVLSKVAAKHPYAKIGVFYCGAPVLAQELSNLCH EFNGKCTTKF"

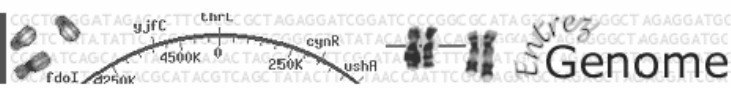
BASE COUNT 102 a 70 c 81 g 83 t 1 others

ORIGIN

1 gtttaaagga atcatgaatg agattgctga actagatcaa aggaatatca ttgagatgca
61 caactatctc acaagtgttt atgaggaagg ggatgctcgg tcagcactca tcacaatgct
121 gcaagctctc aaccatgcc aagaatggtg cgatgtagtg totggmactc gagtccggac
181 acattttgca agaccaaatt ttaagagggt gctgtctaa gtagccgcc aacatcotta
241 tgccaagata ggagtgttet attgcccagc tccagttctg ggcaggaac taagcaacct
301 ttgccatgag ttcaatggca aatgcacgac aaaattc

Genomové databáze v NCBI – prokaryota





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

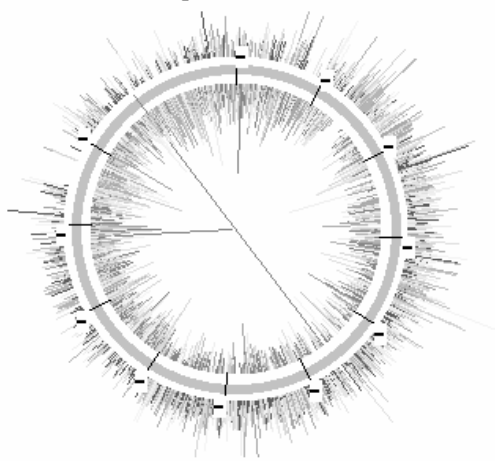
Bacillus anthracis A2012, unfinished sequence - 4171029..4221028

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

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

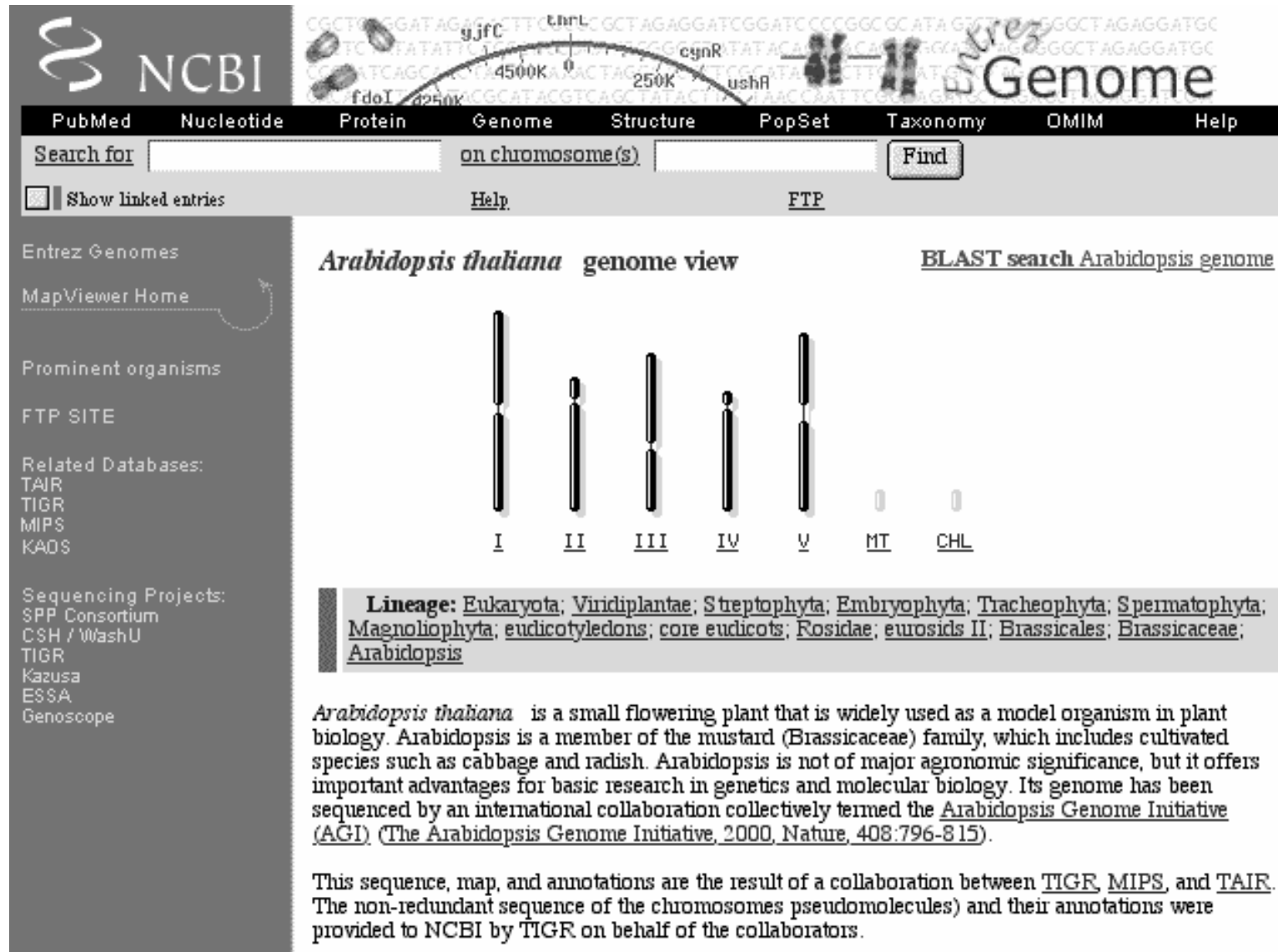


BA_4511	BA_4513	BA_4515	BA_4516	BA_4517	BA_4519	BA_4520
4171029	4173161	4175294	4177427	4179559		
BA_4521	BA_4523	BA_4524	BA_4525	BA_4526	BA_4527	
4181029	4183161	4185294	4187427	4189559		
BA_4529	BA_4530	BA_4532	BA_4535	BA_4537	BA_4539	BA_4543
4191029	4193161	4195294	4197427	4199559		
BA_4545	BA_4547	BA_4548	BA_4549	BA_4550	BA_4551	BA_4552
4201029	4203161	4205294	4207427	4209559		
BA_4553	BA_4554	BA_4555	BA_4557	BA_4560	BA_4563	BA_4565
4211029	4213161	4215294	4217427	4219559		

←
→

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

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

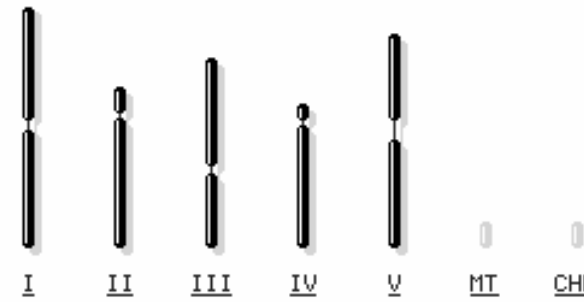
Prominent organisms

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](#); [euosids II](#); [Brassicales](#); [Brassicaceae](#); [Arabidopsis](#)

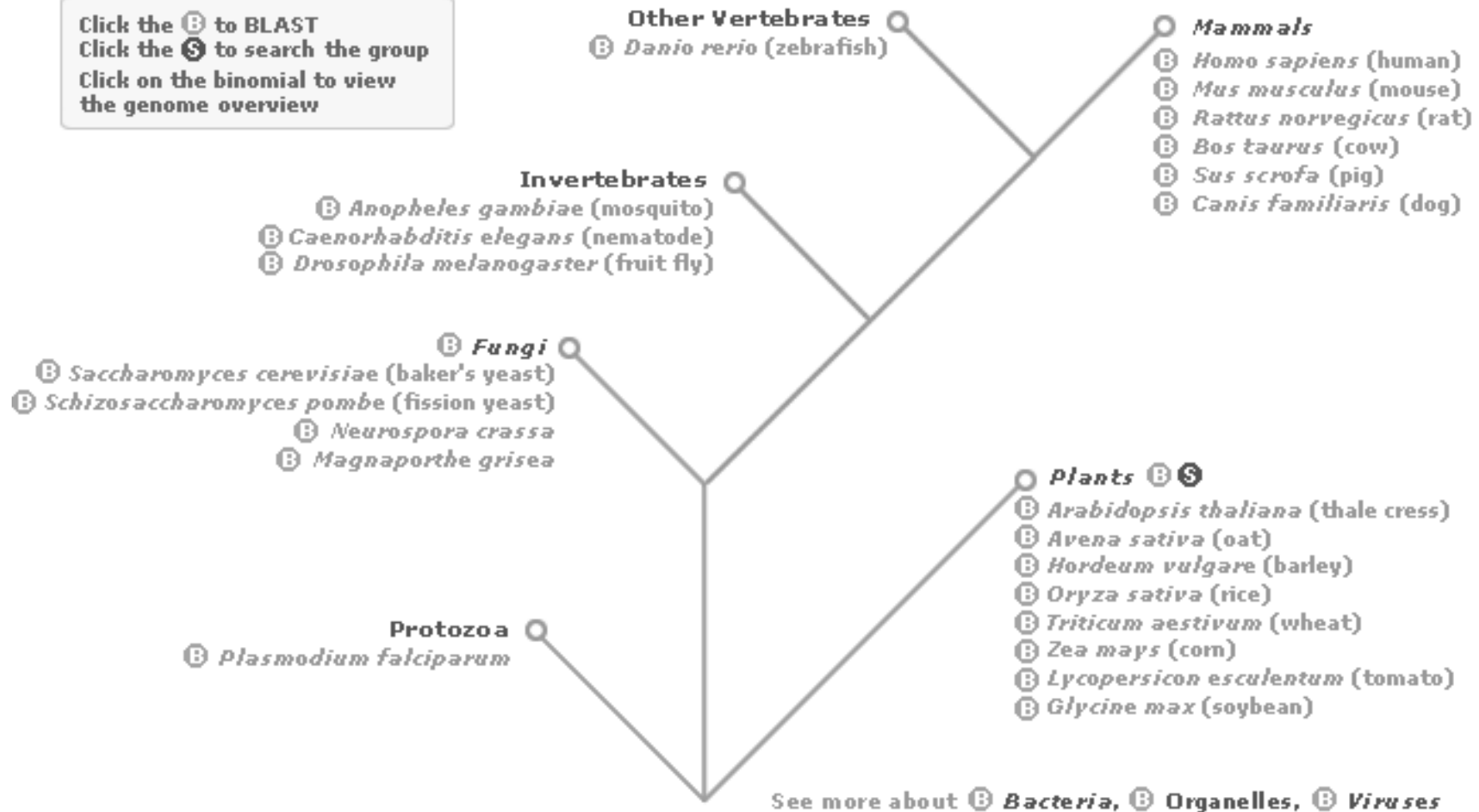
Arabidopsis thaliana is a small flowering plant that is widely used as a model organism in plant biology. *Arabidopsis* is a member of the mustard (*Brassicaceae*) family, which includes cultivated species such as cabbage and radish. *Arabidopsis* is not of major agronomic significance, but it offers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

 out
 zoom
 in

Cyto



Drosophila melanogaster Map View

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

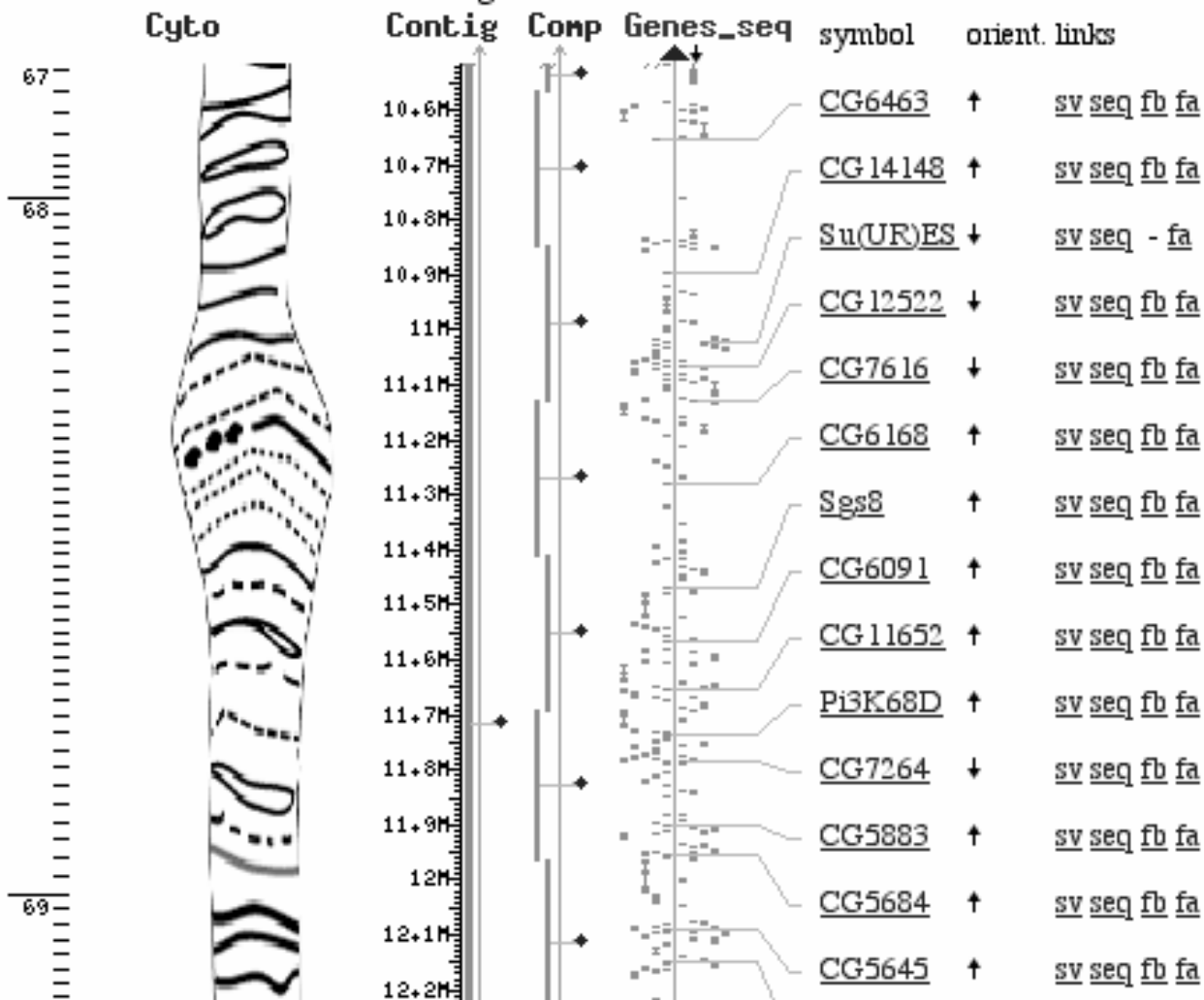
Master Map: Genes On Sequence

Maps & Options

Total Genes On Chromosome: 2617

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

Genes Labeled: 20 Total Genes in Region: 286

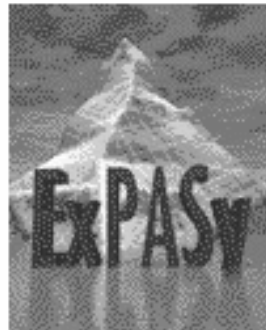


Databáze sekvencí proteinů

- 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

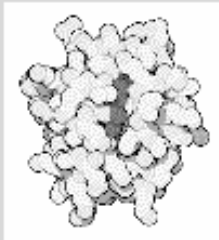
Důležitou databází spojenou s proteiny je PDB (The Protein Databank), která se zabývá archivací a analýzou 3-D **proteinových struktur**.

- 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



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

[RCSB Home](#) [Contact Us](#) [Help](#)

[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](#)

[pdb-J 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](#)

+ - - - - +



MMDB Structure Summary

PubMed BLAST Structure Taxonomy

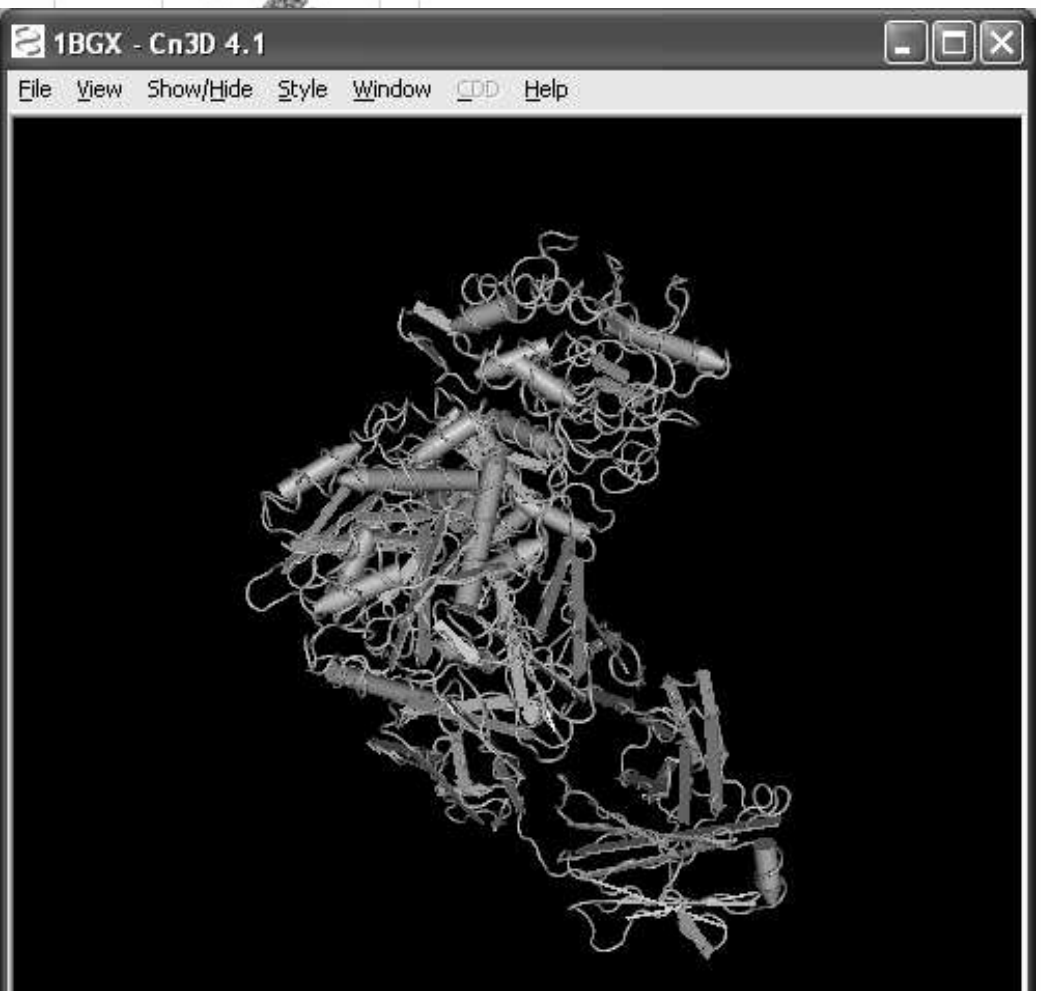
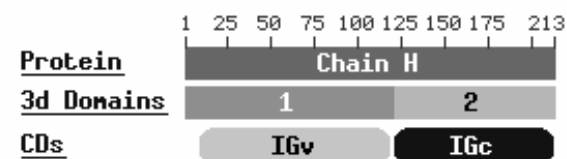
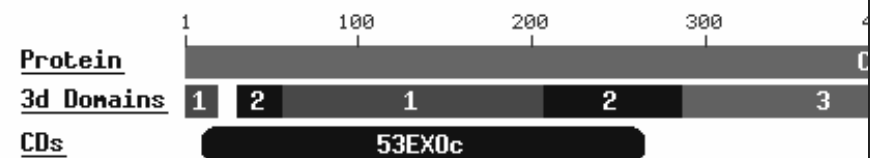
Description: Taq Polymerase In Complex With Tp7, An Inhibitor

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 s
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 e
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 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 nejpříjemnější

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 the text "Search across databases" and buttons for GO, CLEAR, and Help. The main content area is titled "Welcome to the new Entrez cross-database search page" and lists 20 different databases in a two-column grid. Each database entry includes an icon, the database name, a brief description, and a help icon.

PubMed: biomedical literature citations and abstracts	Books: online books
PubMed Central: free, full text journal articles	OMIM: Online Mendelian Inheritance in Man
Journals: detailed information about journals in Entrez	Site Search: NCBI web and FTP sites
MeSH: detailed information about NLM's controlled vocabulary	
Nucleotide: sequence database (GenBank)	UniGene: gene-oriented clusters of transcript sequences
Protein: sequence database	CDD: conserved protein domain database
Genome: whole genome sequences	3D Domains: domains from Entrez Structure
Structure: three-dimensional macromolecular structures	UniSTS: markers and mapping data
Taxonomy: organisms in GenBank	PopSet: population study data sets
SNP: single nucleotide polymorphism	GEO: expression and molecular abundance profiles
Gene: gene-centered information	GEO DataSets: experimental sets of GEO data

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

- **SRS** je homogenní rozhraní pro přístup k více než 160 molekulárně biologickým databázím vyvinuté v EBI
- 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



Reset

Quick Search

Search Options

1. Select the **databanks** you want to search

2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

[Standard Query Form](#)

[Extended Query Form](#)

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:

[Browse Entries](#)

Tips

► bookmark this [link](#) to return to your project
► [Linking to SRS?](#)
- Please read this [document](#) for important information regarding linking to our SRS server.

BookMarkLets

[About BookmarkLets](#)

- [Protein Seq](#)
- [DNA/RNA Seq](#)
- [Structures](#)

Available Databanks

Expand all Collapse all

Show databanks tooltips:

Literature, Bibliography and Reference Databases

Nucleotide sequence databases

- all EMBL EMBL (Release) EMBL (Updates) EMBL (WGS)
 EMBL (TPA) EMBL (Contig) REFSEQ ENSEMBL HUMAN
 ENSEMBL MOUSE ENSEMBL FLY ENSEMBL FISH IMGTHLA
 IMGT/LIGM-DB PATENT DNA LiveLists

UniProt Universal Protein Resource

- all UniProt UniParc UniRef100 UniRef90 UniRef50
 UniProt/Swiss-Prot UniProt/TrEMBL

Other protein sequence databases

- all IPI REFSEQ EPO Proteins JPO Proteins
 USPTO Proteins MHCBN BCIPEP SWISSCHANGE

Deprecated Protein Databases

- all Swall (SPTR) TrEMBL (Updates) RemTrEMBL PIR

Nucleotide related databases

- all CPGISLAND ENSEMBLCPG EPD HGNC
 HSAGENES LOCUSLINK MOUSE2HUMAN REBASE
 TFCLASS TFCCELL TFFACTOR TFGENE
 TFMATRIX TFSITE UNIGENE UNILIB
 UTR UTRSITE EMBLALIGN

Protein function databases

- all BLOCKS INTERPRO IPRMATCHES
 IPRMATCHES_ENSEMBL NICEDOM PEP (ORFs)
 PFAMA PFAMB PFAMSEED
 PFAMHMMFS PFAMHMMLS PRINTS
 PRODOM PROSITE PROSITEDOC

Protein structure databases

- all DSSP FSSP HSSP PDB PDBFINDER
 RESID

Enzymes, reactions and metabolic pathway databases

- all EMP ENZYME LCOMPOUND LENZYME LREACTION
 MPW PATHWAY UCOMPOUND UENZYME UIMAGEMAP
 UPATHWAY UREACTION

Mutation and SNP databases

Gene ontology resources

Mapping databases

Other databases

User owned databases

Application result databases

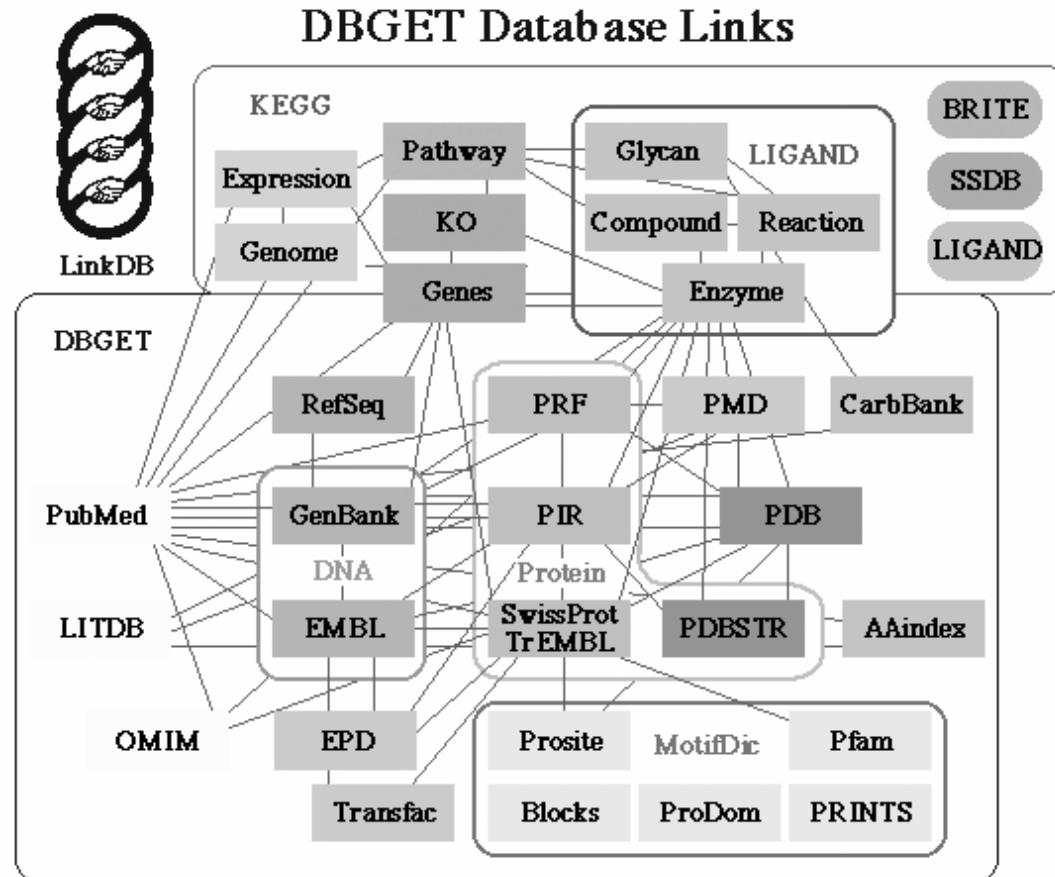
EMBOSS result databases

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.

DBGET/Link DB



DBGET

Integrated database retrieval system

Basic Search [DBGET Links Diagram | IDEAS Interface]

Category	Database name			Original site		
	Compound name	Full name	Abbreviation			
Nucleic acid sequences		refseq	rs	NCBI		
	dna	genbank	gb	NCBI		
		embl	emb	EBI		
		swissprot	sp	EXPASY		
Protein sequences	protein	trembl	tr	EBI		
		trembl_new	trnew			
		pir	pir	NBRF		
		prf	prf	PRF		
		pdbstr	str	RCSB		
		pdb	pdb			
		epd	epd	ISREC		
3D structures	transfac	tfactor	tf	GBF		
		tfsite	ts			
		motifdic	prosite	ps	EXPASY	
	blocks		bl	FHCRC		
	prodom		pd	INRA		
	prints		pr	UMBER		
	pfam		pf	Wash. U / Sanger		
	Compounds and reactions		ligand	compound	cpd	Kyoto
				glycan	gl	
		reaction		rn		
enzyme		ec				
KEGG pathways		pathway	path	Kyoto		
KEGG orthology		ko	ko			
KEGG gene catalogs	genes	eukaryotes eubacteria archaea	See KEGG for individual organisms			
KEGG organisms		genome	gn	Kyoto		
Gene expressions		expression	exp			
Virus gene catalogs		vgenes	vg	NCBI		
Organelle gene catalogs		ogenes	og			
Genetic diseases		omim	mim	NCBI		
Protein mutants		pmd	pmd	DBDJ		
Amino acid indices	aaindex	aaindex1	aa1	Kyoto		
		aaindex2	aa2			
Carbohydrate structures		carbbank	ccsd	Telkyo U / U Georgia		
Protein/peptide literature		litdb	lit	PRF		
Motif literature		prosdoc	pd	EXPASY		
Biomedical literature		medline	ui	NCBI		

LinkDB Search

- LinkDB

Full Text Search

- STAG: Searching Texts in All over the GenomeNet - JAIST, Hokuriku

Last updated: September 22, 2003

[Feedback form | GenomeNet | KEGG | BLAST | FASTA | MOTIF]

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

Řada variant programu BLAST

NCBI

PubMed Entrez **BLAST** OMIM Taxonomy Structure

NEW 15 Nov 2004 Download the [BLAST poster](#) from [SC2004!](#)

About BLAST <ul style="list-style-type: none">• News• Mailing list• References• NCBI Contributors	Nucleotide <ul style="list-style-type: none">• Quickly search for highly similar sequences (megablast)• Quickly search for divergent sequences (discontiguous megablast)• Nucleotide-nucleotide BLAST (blastn)• Search for short, nearly exact matches• Search trace archives with megablast or discontiguous megablast	Protein <ul style="list-style-type: none">• Protein-protein BLAST (blastp)• PHI- and PSI-BLAST• Search for short, nearly exact matches• Search the conserved domain database (rpsblast)• Search by domain architecture (cdart)
BLAST Services <ul style="list-style-type: none">• FAQs• Program selection guide• Web service interface	Translated <ul style="list-style-type: none">• Translated query vs. protein database (blastx)• Protein query vs. translated database (tblastn)• Translated query vs. translated database (tblastx)	Genomes <ul style="list-style-type: none">• Chicken, cow, pig, dog, sheep, cat• Environmental samples• Human, mouse, rat• Fugu rubripes, zebrafish• Insects, nematodes, plants, fungi, malaria• Microbial genomes, other eukaryotic genomes
BLAST Software <ul style="list-style-type: none">• Databases• Documentation• Errata• Executables• Source code	Special <ul style="list-style-type: none">• Search for gene expression data (GEO BLAST)• Align two sequences (bl2seq)• Screen for vector contamination (VecScreen)• Immunoglobulin BLAST (IgBlast)• SNP BLAST <small>NEW</small>	Meta <ul style="list-style-type: none">• Retrieve results by RID

Support

- Contact us

Jak používat BLAST?

- <http://www.ncbi.nlm.nih.gov/BLAST>
1. Vybrat příslušný BLAST-program (blastn, blastp, blastx, tblastn, tblastx)
 2. Vybrat databázi, která má být prohledána
 3. Vložit sekvenci (DNA nebo protein)
 4. Odeslat požadavek na vyhledání

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

Volba programu, jestliže Vaše sekvence je NUKLEOTIDOVÁ

Délka	Databáze	Účel vyhledávání	BLAST Program
20 bp nebo delší	DNA	Identifikace dotazované sekvence	<u>MEGABLAST</u> <u>Standard BLAST</u> (blastn)
		Vyhledání podobných sekvencí jako dotazovaná	<u>Standard BLAST</u> (blastn)
		Vyhledání podobných proteinů k překladu dotazované sekvence v přeložených databázích DNA	<u>Translated BLAST</u> (tblastx)
	Protein	Vyhledání podobných proteinů k překladu dotazované sekvence v databázích proteinů	<u>Translated BLAST</u> (blastx)
7 - 20 bp	DNA	Vyhledání vazebných míst primerů nebo krátkých motivů	<u>Search for short, nearly exact matches</u>

Volba programu, jestliže Vaše sekvence je PROTEIN

Délka	Databáze	Účel vyhledávání	BLAST program
15 aminokyselinových zbytků nebo delší	Protein	Identifikace dotazované sekvence nebo vyhledání sekvencí podobných proteinů	<u>Standard Protein BLAST</u> (blastp)
		Vyhledání členů proteinové rodiny nebo tvorba vlastní pozičně-specifické matrice skóre	<u>PSI-BLAST</u>
		Vyhledání proteinů podobných dotazovanému v okolí určitého vzoru	<u>PHI-BLAST</u>
	Konzervativní domény	Vyhledání konzervativních domén v dotazované sekvenci	<u>CD-search</u> (RPS-BLAST)
	Konzervativní domény	Vyhledání konzervativních domén v dotazované sekvenci a identifikace ostatních proteinů s podobnou architekturou domén	<u>Conserved Domain Architecture Retrieval Tool</u> (CDART)
	DNA	Vyhledání podobných proteinů v přeložených databázích DNA	<u>Translated BLAST</u> (tblastn)
5-15 zbytků	Protein	Hledání peptidových motivů	<u>Search for short, nearly exact matches</u>

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
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

Seřazení sekvencí, které BLAST může nalézt

```
1 AATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAG
  ||||| ||||||||||||||||| || |||||||||||||
1 AATGGAAAAGACTACTGGATCATCAAAAACCTCCTGGGGAG
```


BLAST - Možnosti nastavení

Options for advanced blasting

Limit by entrez query or select from:

Composition-based statistics

Choose filter Low complexity Mask for lookup table only Mask lower case

Expect ←

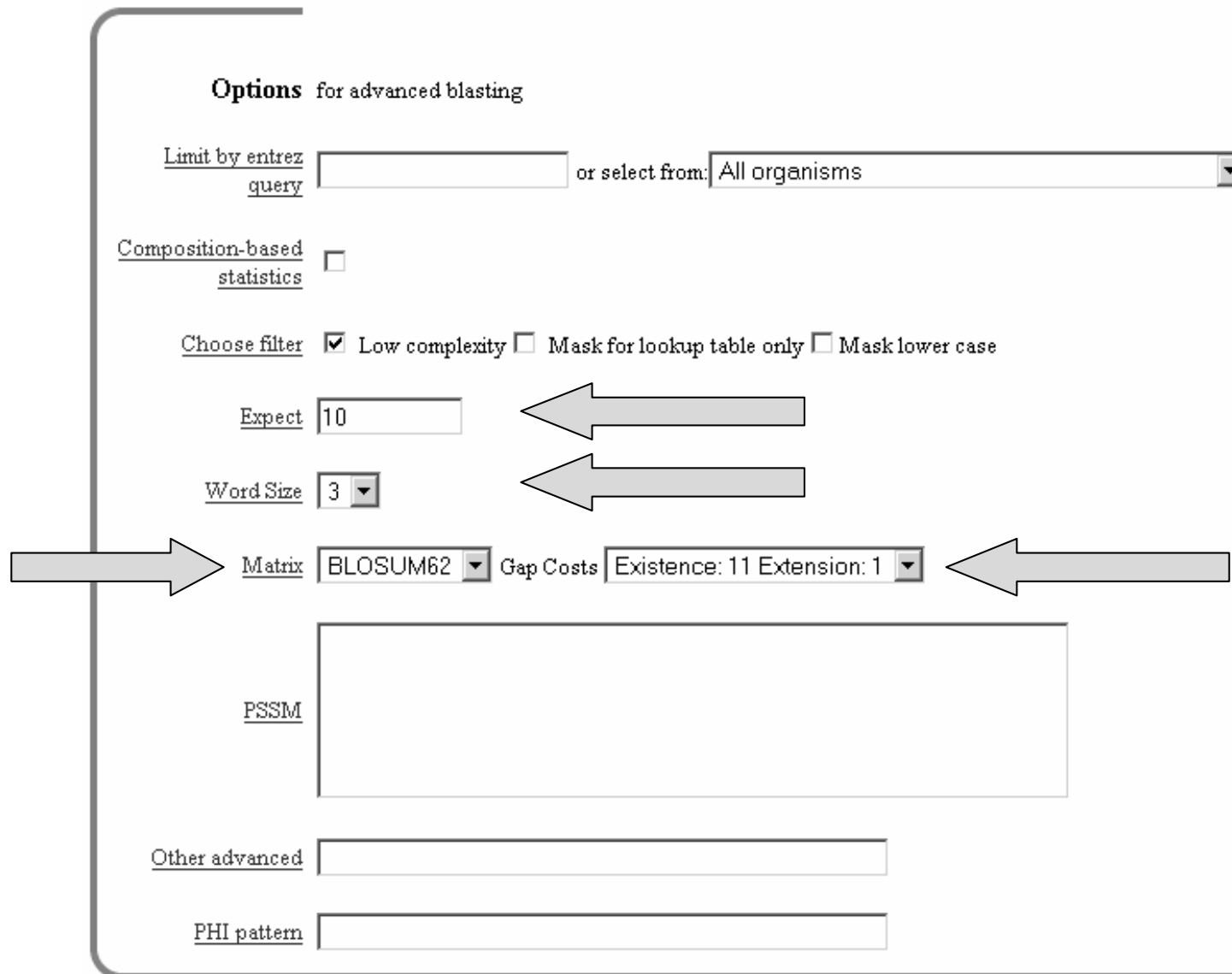
Word Size ←

Matrix Gap Costs ←

PSSM

Other advanced

PHI pattern

A screenshot of the BLAST 'Options for advanced blasting' section. The interface includes several input fields and checkboxes. A large grey arrow on the left points towards the 'Matrix' dropdown menu. Two grey arrows on the right point towards the 'Expect' and 'Word Size' input fields. Another grey arrow on the right points towards the 'Gap Costs' dropdown menu. The 'Expect' field contains the value '10', 'Word Size' contains '3', and 'Matrix' is set to 'BLOSUM62'. The 'Gap Costs' field shows 'Existence: 11 Extension: 1'. Other fields like 'Limit by entrez query', 'Composition-based statistics', 'Choose filter', 'PSSM', 'Other advanced', and 'PHI pattern' are also visible but not highlighted with arrows.

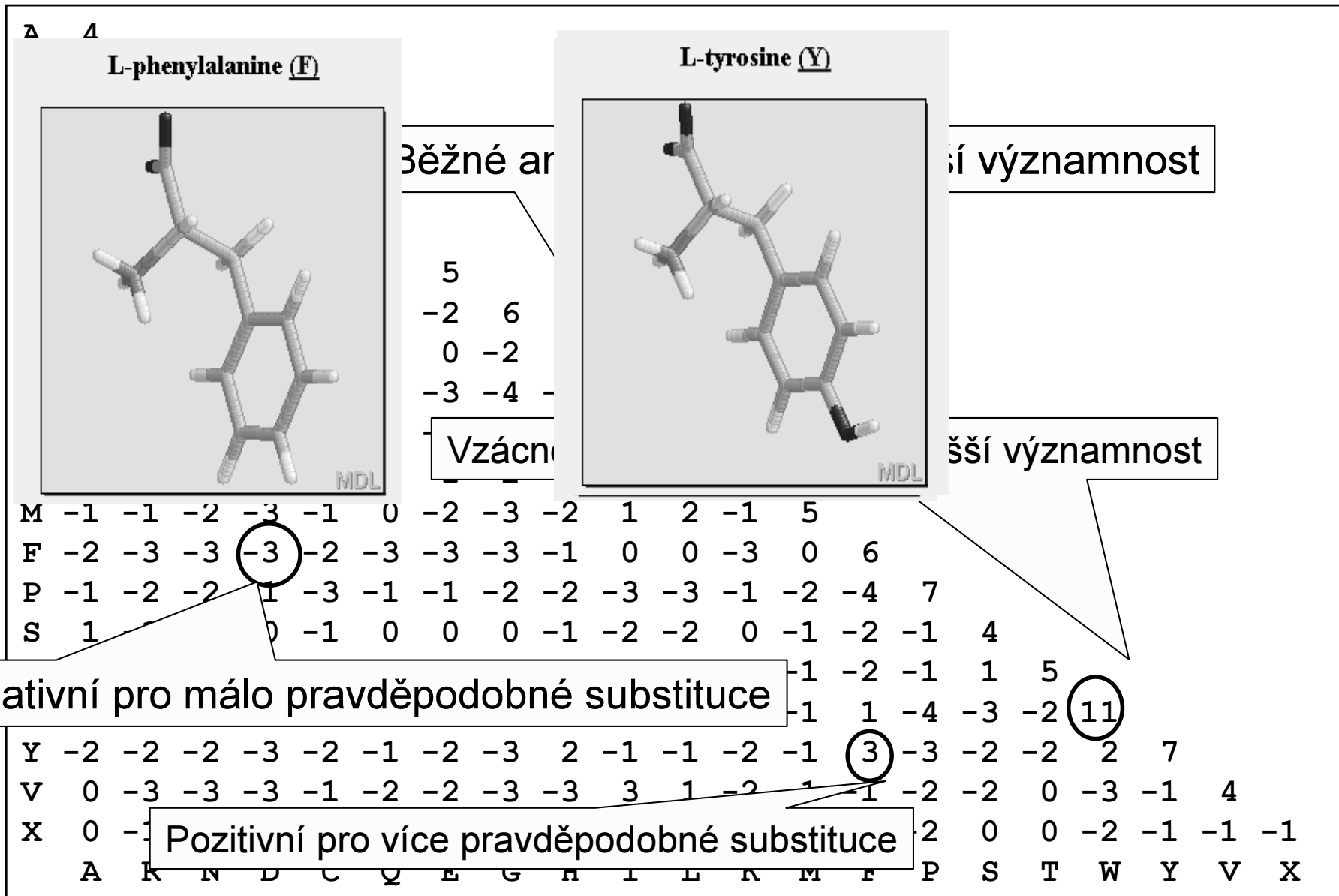
Substituční Matice

- Co je substituční matice?
 - Kompletní sada skóre pro všechny kombinace párů zbytků se nazývá substituční matice
 - 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 matice 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.

Příklad matice BLOSUM62



Různé typy substitučních matic

- Matice identity
 - Především pro nukleotidové sekvence
 - Neschopné transformovat na jiné zbytky
 - Pro seřazení velmi podobných sekvencí
 - Vypadá následovně

Matrice PAM

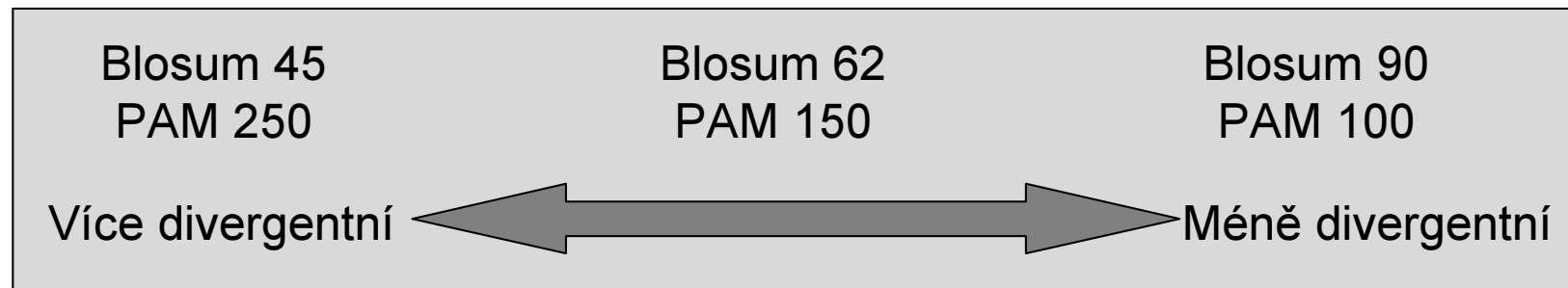
- PAM
 - Percent Accepted Mutation
 - založeny na konceptu akceptovatelných bodových mutací za 10^8 let v globálních mnohonásobných seřazeních blízce příbuzných proteinů
 - Stanoveny na základě výpočtů u blízce příbuzných proteinů
 - PAM1 reprezentuje 1% změn
 - $PAM250 = (PAM1)^{250}$

Matrice BLOSUM






- **Blocks Substitution Matrix**
- Změny probíhající během dlouhodobé evoluce nejsou často vhodné pro výpočty a sledování malých recentních změn
- Matice BLOSUM jsou sestaveny na základě analýzy mnohonásobných seřazení evolučně příbuzných proteinů v databázi BLOCKS
- BLOSUM-x používá analýzu pouze těch proteinů, které mají alespoň x % identitu

PAM versus BLOSUM

- PAM Matice (Percent Accepted Mutation)
 - Odvozené z pozorování; malé množství seřazených dat
 - vhodné pro evoluční modely
 - Všechny výpočty prováděny s PAM1
 - PAM250 je nejpoužívanější
- BLOSUM (BLOck SUBstitution Matrices)
 - Odvozené z pozorování; velké množství vysoce konzervovaných sekvencí (BLOCKS)
 - Každá matice odvozená samostatně podle definované procentuální identity
 - BLOSUM62 – výchozí matice pro BLAST



PAM versus BLOSUM

- PAM100  Blosum90
- PAM120  Blosum80
- PAM160  Blosum60
- PAM200  Blosum52
- PAM250  Blosum45

Obecné závěry

- Klíčovým elementem vyhodnocujícím výsledky srovnání aminokyselinových sekvencí je substituční matice
- Různé matice jsou přizpůsobené pro detekci podobností u sekvencí, které se vyznačují různým stupněm divergence
- BLOSUM je vhodnější pro lokální srovnání
 - BLOSUM-62 je optimální pro detekci nízkých podobností proteinů
 - BLOSUM-45 je vhodnější pro detekci nízkých podobností u dlouhých sekvencí

Významnost shody

- K posouzení významnosti shody nalezených úseků se používá numerická hodnota označovaná jako **skóre seřazení sekvencí (S)**
- Popisuje celkovou kvalitu seřazení sekvencí na základě porovnání pravděpodobnosti výskytu nalezených segmentů o určité sekvenci podobnosti s pravděpodobností, že se taková podobnost vyskytne mezi dvěma náhodnými sekvencemi
- Vyšší číslo odpovídá vyšší podobnosti
- Ekvivalentem skóre S je **hodnota E** („Expectation value“), která vyjadřuje počet různých seřazení sekvencí se skórem shodným nebo vyšším než je hodnota S, jejíž výskyt je očekáván při náhodném vyhledávání v databázi.

$$E = mn 2^{-S}$$

- Potom platí, že čím je hodnota E nižší, tím je skóre významnější.

BLAST - Možnosti nastavení

Options for advanced blasting

Limit by entrez query or select from:

Composition-based statistics

Choose filter Low complexity Mask for lookup table only Mask lower case

Expect

Word Size

Matrix Gap Costs ←

PSSM

Other advanced

PHI pattern

BLAST - Možnosti nastavení

Options for advanced blasting

Limit by entrez query or select from:

Composition-based statistics

Choose filter Low complexity Mask for lookup table only Mask lower case

Expect ←

Word Size

Matrix Gap Costs

PSSM

Other advanced


PHI pattern

BLAST – Výstup (Výsledky)

Skládají se ze 4 částí

- 1) úvod, který informuje o tom kde bylo vyhledání provedeno a jaké databáze byly použity
- 2) seznam sekvencí v databázi, obsahující segmenty podobných sekvencí, jejichž skóre je alespoň tak vysoké jako zadané parametry
- 3) seřazení podobných sekvencí s vysokým skóre
- 4) kompletní seznam parametrů použitých pro vyhledání.

Proteinový BLAST

 **NCBI** *protein-protein* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

Search

```
>Mutated in Colon Cancer
IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILER
VQQHIESKLLGSNSSRMYFTQTLPLGLAGPSGEMVKSTTSLTSSSTSGSS
DKVYAHQMVRTDSREQKLD AFLQPLSKPLSS
```

Set subsequence From: To:


Choose database

Do CD-Search

Now: **BLAST!** or

Protein database

BLAST – stránka pro formátování


 **NCBI** *formatting* **BLAST**

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = Mutated in Colon Cancer (131 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

or

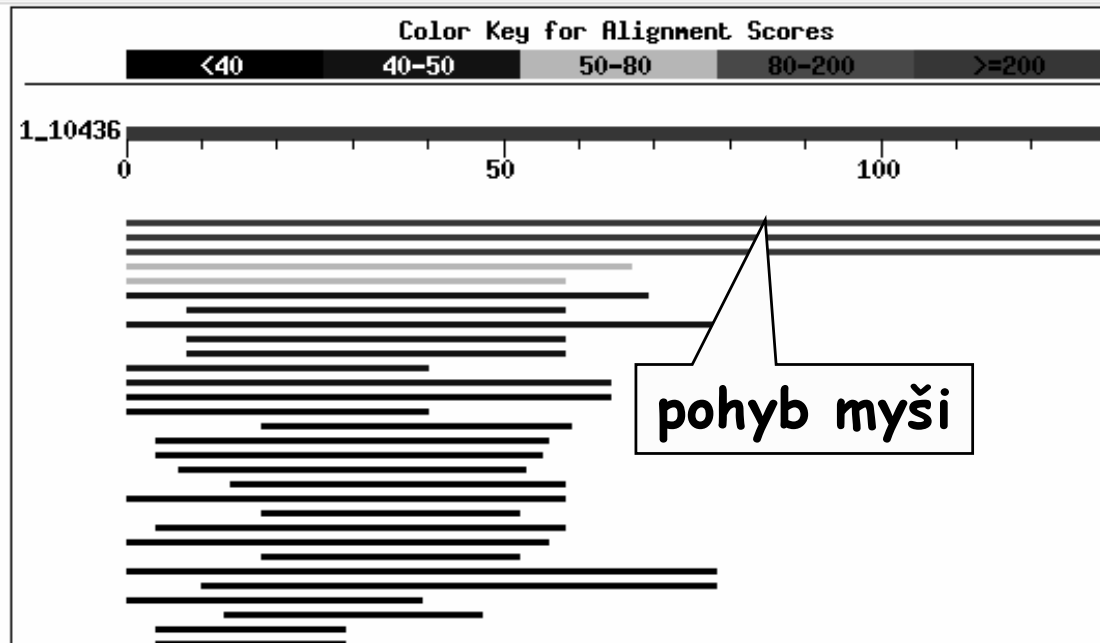
The results are estimated to be ready in 36 seconds but may be done sooner.

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 Output: Descriptions

Sequences producing significant alignments	seřazeno podle hodnot E	Score	E Value
gi 730028 sp P40692 MLH1 HUMAN	DNA mismatch repair protein ...	233	8e-62
gi 13878583 sp Q9JK91 MLH1 MOUSE	DNA mismatch repair protein ...	214	4e-56
gi 13878571 sp P97679 MLH1 RAT	DNA mismatch repair protein ...	212	1e-55
gi 1709056 sp P38920 MLH1 YEAST	MUTL protein homolog 1 (DNA...	72	7e-13
gi 1171080 sp P44494 MUTL HAEIN	DNA mismatch repair protein...	54	7e-08
gi 13431695 sp P57886 MUTL PASMU	DNA mismatch repair protei...	50	1e-06
gi 18929224 sp P40925 MUTL THEMA	DNA mismatch repair protein...	48	4e-06
gi 127553 sp P14161 MUTL BACHD	DNA mismatch repair protei...	46	1e-05
gi 127553 sp P14161 MUTL ECOLI	DNA mismatch repair protei...	44	5e-05
gi 127553 sp P14161 MUTL SALTY	DNA mismatch repair protei...	44	7e-05
gi 6225738 sp Q9ZC88 MUTL RICPR	DNA mismatch repair protei...	40	7e-04
gi 14194944 sp Q9PJG5 MUTL CHLMU	DNA mismatch repair protei...	40	0.001
gi 8928218 sp O84579 MUTL CHLTR	DNA mismatch repair protein...	39	0.001
gi 20043258 sp Q9KV13 MUTL VIBCH	DNA mismatch repair protei...	39	0.002
gi 13631230 sp Q9RP66 MUTL CAUCR	DNA mismatch repair protei...	39	0.002
gi 8928214 sp O51229 MUTL BORBU	DNA mismatch repair protein...	39	0.002
gi 1709188 sp P49850 MUTL BACSU	DNA mismatch repair protein...	38	0.005
gi 8039787 sp O83325 MUTL TREPA	DNA mismatch repair protein...	36	0.013
gi 19856116 sp P14160 HEXP	DNA mismatch repair protei...	36	0.020
gi 3914082 sp P70754 MUTL	DNA mismatch repair protei...	35	0.020
gi 11386926 sp P57633 MUTL	DNA mismatch repair protei...	35	0.026
gi 8928240 sp Q9Z794 MUTL CHLPN	DNA mismatch repair protei...	35	0.026
gi 1709684 sp P54280 PMS1 SCHPO	DNA mismatch repair protei...	33	0.16
gi 3914081 sp O67518 MUTL AQUAE	DNA mismatch repair protei...	32	0.24
gi 1709685 sp P54278 PMS2 HUMAN	PMS1 protein homolog 2 (DNA...	32	0.24
gi 1709686 sp P54279 PMS2 MOUSE	PMS1 PROTEIN HOMOLOG 2 (DNA...	32	0.24
gi 8928222 sp P73349 MUTL SYNY3	DNA mismatch repair protein...	31	0.60
gi 1709683 sp P54277 PMS1 HUMAN	PMS1 protein homolog 1 (DNA...	30	0.85
gi 126232 sp P02239 LGB1 LUPLU	Leghemoglobin I	30	1.2
gi 126238 sp P02240 LGB2 LUPLU	Leghemoglobin II	28	4.1

link to entrez

seřazeno podle hodnot E

4 X 10⁻⁵⁶

LocusLink

Default e value cutoff 10

Bacterial mismatch repair proteins

BLAST – výstup seřazení sekvencí

```
>gi|127552|sp|P23367|MUTL_ECOLI DNA mismatch repair protein mutL  
Length = 615
```

```
Score = 44.3 bits (103), Expect = 5e-05
```

```
Identities = 25/59 (42%), Positives = 33/59 (55%), Gaps = 8/59 (13%)
```

```
Query: 9 LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 59  
L + P L LEI P VDVNVHP KHEV F +H+ + +L +QQ +E+ L  
Sbjct: 280 LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRVLVHDFIYQGVLSVLQQQLEETPL 338
```

BLAST – výstup filtrování sekvencí

```
>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair protein Mlh1 1)
      Length = 756
```

```
Score = 233 bits (593), Expect = 8e-62
Identities = 117/131 (89%), Positives = 117/131 (89%)
```

```
Query: 1 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60
        IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct: 276 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335
```

```
Query: 61 GSNSSRMYFTQTL PGLAGPSGEMVXXXXXXXXXXXXXXXXXXXXK VYAHQMVRTDSREQK LDA 120
        GSNSSRMYFTQTL PGLAGPSGEMV                               DKVYAHQMVRTDSREQK LDA
Sbjct: 336 GSNSSRMYFTQTL PGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQK LDA 395
```

```
Query: 121 FLQPLSKPLSS 131
        FLQPLSKPLSS
Sbjct: 396 FLQPLSKPLSS 406
```

sekvence s nízkou komplexitou

BLAST – výstup, přehled parametrů

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
Posted date: Oct 8, 2004 12:07 AM
Number of letters in database: 697,528,283
Number of sequences in database: 2,075,303

Lambda	K	H
0.316	0.133	0.382

Gapped

Lambda	K	H
0.267	0.0410	0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 158,300,126
Number of Sequences: 2075303
Number of extensions: 6358952
Number of successful extensions: 16758
Number of sequences better than 10.0: 54
Number of HSP's better than 10.0 without gapping: 40
Number of HSP's successfully gapped in prelim test: 14
Number of HSP's that attempted gapping in prelim test: 16640
Number of HSP's gapped (non-prelim): 73
length of query: 616
length of database: 697,528,283
effective HSP length: 132
effective length of query: 484
effective length of database: 423,588,287
effective search space: 205016730908
effective search space used: 205016730908
T: 11
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.6 bits)
S2: 77 (34.3 bits)

Mnohonásobné seřazení sekvencí (multiple alignment)

Definice: Mnohonásobné seřazení 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.

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 seřazení může identifikovat podobnosti a identifikovat konzervativní motivy, které nejsme schopni identifikovat lokálním srovnáním

Důvody provedení mnohonásobného seřazení

- Organizace dat a manipulace s daty týkajícími se podobných sekvencí
- Dedukce fylogeneze
- Vyhledání konzervativních míst nebo oblastí
- Vyhledání variabilních oblastí
- Odhalení změn ve struktuře genů

Algoritmus:

mnohonásobné seřazení = hledání optimální cesty
více konzervativních sloupců = lepší seřazení

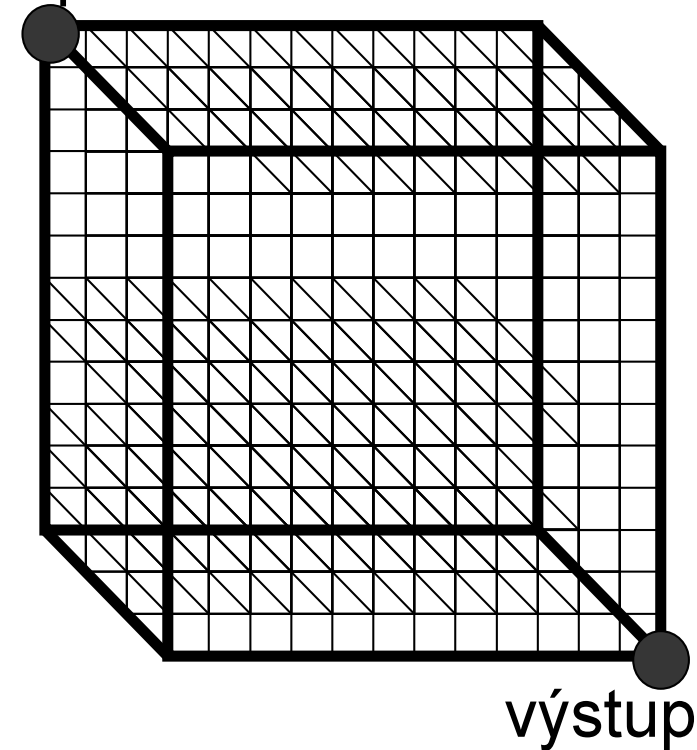
0	1	1	2	3	4
	A	--	T	G	C
0	1	2	3	3	4
	A	A	T	--	C
0	0	1	2	3	4
	--	A	T	G	C

x koordináta

y koordináta

z koordináta

vstup

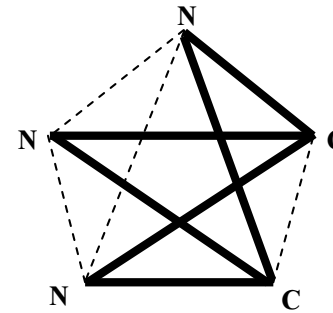
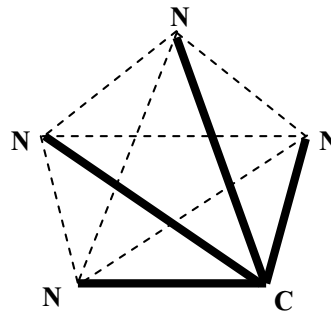
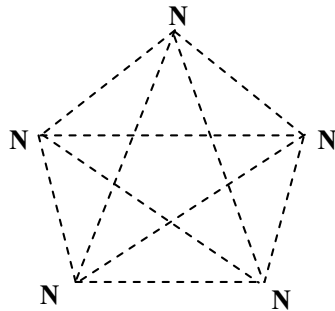


- Výsledná cesta v 3-rozměrném(x,y,z) prostoru:

$(0,0,0) \rightarrow (1,1,0) \rightarrow (1,2,1) \rightarrow (2,3,2) \rightarrow (3,3,3) \rightarrow (4,4,4)$

Výpočet skóre seřazení

Sequence	Column A	Column B	Column C
1N.....N.....N
2N.....N.....N
3N.....N.....N
4N.....N.....C
5N.....C.....C



No. of N-N matched pairs (each scores 6):

10

6

4

No. of N-C matched pairs (each scores -3):

0

4

6

Vytvoření konsenzní sekvence

- Nejjednodušší forma:
Jedna sekvence, která reprezentuje výskyt nejběžnějších zbytků v každé pozici

Y	D	D	G	A	V	-	E	A	L
Y	D	G	G	-	-	-	E	A	L
F	E	G	G	I	L	V	E	A	L
F	D	-	G	I	L	V	Q	A	V
Y	E	G	G	A	V	V	Q	A	L
Y	D	G	G	A/I	V/L	V	E	A	L

Vytvoření profilu

-	A	G	G	C	T	A	T	C	A	C	C	T	G
T	A	G	-	C	T	A	C	C	A	-	-	-	G
C	A	G	-	C	T	A	C	C	A	-	-	-	G
C	A	G	-	C	T	A	T	C	A	C	-	G	G
C	A	G	-	C	T	A	T	C	G	C	-	G	G

A		1				1			.8				
C	.6			1			.4	1		.6	.2		
G			1	.2					.2			.4	1
T	.2				1	.6						.2	
-	.2			.8						.4	.8	.4	

ClustalW

- Obecně používaným programem pro mnohonásobné seřazení sekvencí je Clustal W (Higgins et al., 1994), který počítá optimální shodu mezi sekvencemi a umožňuje i grafické znázornění jejich podobnosti formou kladogramu nebo fylogenetického stromu.
- Proces zahrnuje 3 kroky:
 - 1.) Konstrukce všech párových seřazení
 - 2.) Výpočet vodícího stromu
 - 3.) Progresivní seřazení dle vodícího stromu

ClustalW: krok 1

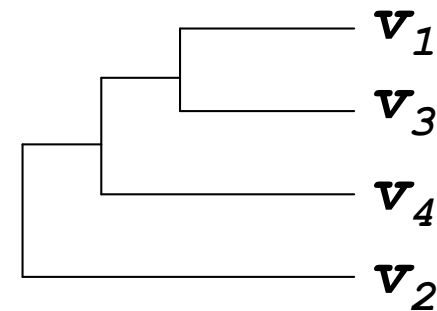
- Seřazení všech párů sekvencí
- Výpočet matice podobností (hodnoty procentuální identity)

	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–

ClustalW: krok 2

- Z matice podobností vypočísá shlukovou analýzou vodící strom
 - Používá statistickou metodu Neighbor-joining
 - Strom hrubě odráží evoluční souvislosti

	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–



$V_{1,3}$ = alignment (v_1, v_3)
 $V_{1,3,4}$ = alignment $((v_{1,3}), v_4)$
 $V_{1,2,3,4}$ = alignment $((v_{1,3,4}), v_2)$

ClustalW: krok 3

- Začíná seřazením 2 nejpodobnějších sekvencí
- Sleduje vodící strom a přidává další nejpodobnější sekvenci
- Podle potřeby vkládá mezery

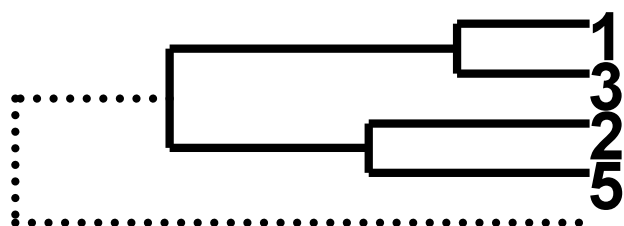
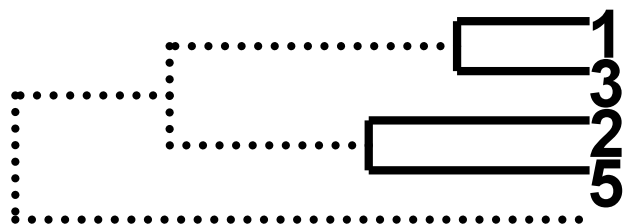
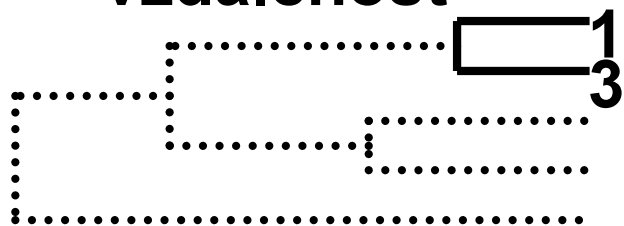
```
FOS_RAT      PEEMSVTS-LDLTGGLPEATTPESSEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE   PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKISINVELKAEPFD
FOS_CHICK   SEELAAATALDLG-----APSPAAAEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE  PGPGLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----LPGFQ
FOSB_HUMAN  PGPGLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----LPGFQ
.           . : ** . :.. *:. * * . * **:
```



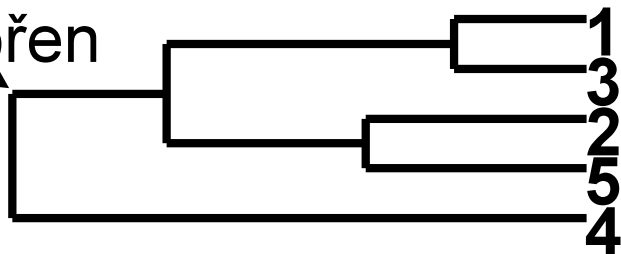
Hvězdičky a tečky označují stupeň konzervovanosti sekvencí

Princip progresivního seřazení

← vzdálenost →



kořen



Problém přesnosti

Při progresivním seřazení se mohou kumulovat chyby.

“Once a gap, always a gap”

Feng & Doolittle, 1987

Prakticky prováděné kroky

- Získání sekvencí (databáze, sekvencování)
- Manipulace se sekvencemi (změna formátu, orientační párové seřazení)
- Výběr vzájemně odpovídajících úseků
- Mnohonásobné seřazení
- Následné fylogenetické analýzy

Lokální (párové) seřazení

- BLAST 2 Sequences (NCBI)
<http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>
- LALIGN local alignment program is available at several servers:
<http://www2.igh.cnrs.fr/bin/lalign-guess.cgi>
http://www.ch.embnet.org/software/LALIGN_form.html
- LFASTA uses FASTA for local alignment of 2 sequences:
<http://pbil.univ-lyon1.fr/lfasta.html>

Netscape: Blast 2 Sequences

Back Forward Reload Home Search Netscape Images Print Security Shop Stop

Location: <http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html> What's Related

[NCBI](#) [Entrez](#) [BLAST 2 sequences](#) [BLAST](#) [Example](#) [Help](#)

BLAST 2 SEQUENCES

This tool produces the alignment of two given sequences using BLAST engine for local alignment. The stand-alone executable for blasting two sequences (bl2seq) can be retrieved from NCBI ftp site
Reference: Tatiana A. Tatusova, Thomas L. Madden (1999), "Blast 2 sequences - a new tool for comparing protein and nucleotide sequences", FEMS Microbiol Lett. 174:247-250

Program Matrix

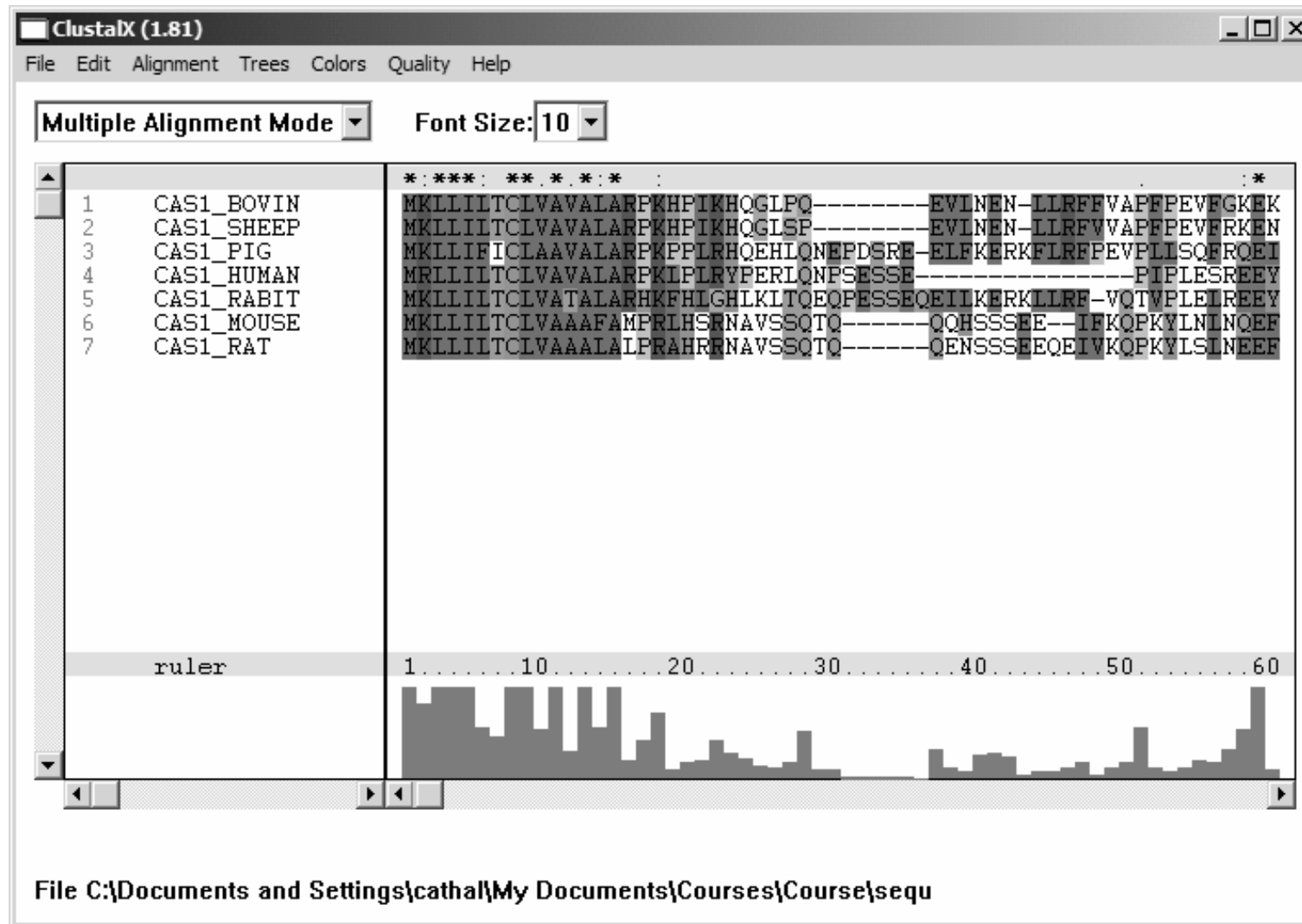
Parameters used in BLASTN program only:
Reward for a match: Penalty for a mismatch:
 Use Mega BLAST Strand option

Open gap and extension gap penalties
gap x_dropoff expect word size Filter

Sequence 1 Enter accession or GI or download from file
or sequence in FASTA format from: to:

Sequence 2 Enter accession or GI or download from file
or sequence in FASTA format from: to:

Software pro mnohonásobné seřazení



WIN_XP

<ftp://ftp.ebi.ac.uk/pub/software/dos/clustalx/clustalx1.83.XP.zip>

WIN_9x

<ftp://ftp.ebi.ac.uk/pub/software/dos/clustalx/clustalx1.83.zip>

UNIX

<ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/clustalw1.83.UNIX.tar.gz>

Webové stránky

- **CLUSTALW** <http://www.ebi.ac.uk/clustalw/>

Match-Box

http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.shtml

- **MUSCA** <http://cbcsrv.watson.ibm.com/Tmsa.html>

- **T-Coffee** <http://www.ch.embnet.org/software/TCoffee.html>

- **MULTALIN** <http://www.toulouse.inra.fr/multalin.html>

- **Dialign** <http://bibiserv.techfak.uni-bielefeld.de/dialign/>

Editory mnohonásobných seřazení: GeneDoc

GeneDoc - [LIPID8]

File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help

C S G

C Q P E S H I L D G M U

Humlbpa : M---MGALARALPS-ILLALLLTSTPEALGA-NPGLVARITDKGLQYAAQEGLLALQSELLRITLPDFTGDLRI PHVGRGRYEF : 79
Rablpb : M---MGTWARALLGSTLLSLLAAAPGALGT-NPGLITRITDKGLEAAREGLLALQRKLLLEVTL PDSGDFRIKHF GRAQYKF : 80
Ratlbp : M---MKSATGPLLP-TLLGLLLLSI PRTQGV-NPAMVVTRITDKGLEAAREGLLSLORELYKITL PDFSGDFKIKAVGRGQYEF : 79
Humcetp : M---MLAATVLT---LALLGNAHACSKGTS H-EAGIVCRITKPA LLVLNHETAKVIQTAFQRASY PDITGEKAMMLLGQVKYGL : 77
Maccetp : M---MLAATVLT---LALLGNVHACSKGTS H-KAGIVCRITKPA LLVLNHETAKVIQSAFQRANY PNITGEKAMMLLGQVKYGL : 77
Rabcetp : -----ACPKGASY-EAGIVCRITKPA LLVLNHETAKVVQTAFQRAGY PDVSGERAVMLLGRVKYGL : 60
Humbpi : MRENMARGPCNAPRWVSLMVLVAIGTAVTAAVNPVVVRI SOKGLDYASOQGTAA LQKELKRIKI PDYSDFSFKIKHLGKGHYSE : 84
Bovbpi : M---MARGPDTARRWATLVVLAALGTAVTTT-NPGLVARITDKGLDYACQGVLT LQKELKRIKI PNFSGNFKIKYLGKQYSE : 80

m m l g66 RI3 L 2 6Q P1 g 6 G Y

Humlbpa : HSLNIHSCCELLHSALRPPVPGQGLSLSISDSSIRVQGRWK---VRKSEFKLQGSFDVSVKGISISVWLLLGSES- SGRPTGYCLS : 159
Rablpb : YSLKIPRFELLRGTLRPLPGQGLSLDISDAYIHVRGSKW---VRKAEFLRLKNSFDLYVKGLTISVHLVVGSES- SGRPTVTTSS : 160
Ratlbp : HSLEIQSCQLRGS SLKPLPGRGLSLSISDSSISVRGKWK---VRRSEVKLHGSEFDLDVKSVTISVDLLLGVDP- SERPTVTASG : 159
Humcetp : HNIQI SHLSIASSQVELVEAKSIDVSIQNVSVVFKGTLKYGYTTAWWLGIDQSIDFEID- SAIDLQINTQLTCDSSGRVRTDAPD : 160
Maccetp : HNIQI SHLSIASSRVELVEAKSIDVSIQNVSVVFKGTLKYGYTTAWWLGIDQSVDFEID- SAIDLQINTQLTCDSSGRVRTDAPD : 160
Rabcetp : HNLQI SHLSIASSQVELVDAKTIDVAIQNVSVVFKGTLNYSYTSAWWLGINQSVDFEID- SAIDLQINTELTCDAGSVRTNAPD : 143
Humbpi : YSMDIRFQLPSSQISMPVNVGLKFSISMANIKISGKWK---AQKRELFKMSGNFDLSIEGMSI SADLKLGSNPTS GSKPTITCSS : 165
Bovbpi : FSMVIQGFNLPNSQIRPLPDKGLDLSIRDASIKIRGKWK---ARKNEIKLGGNFDLSVEGISILAGLNLGYDPASGHSTVTCSS : 161

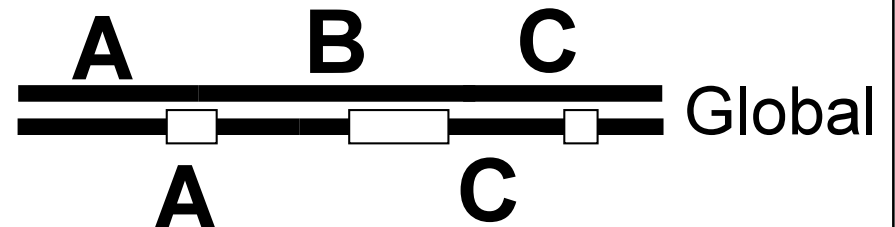
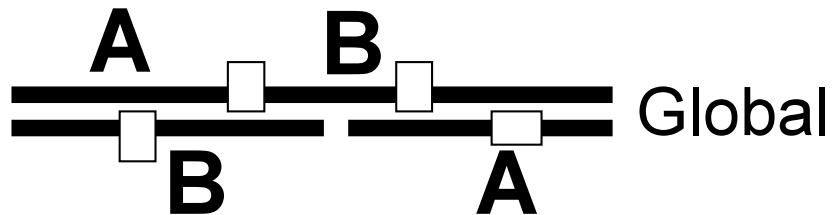
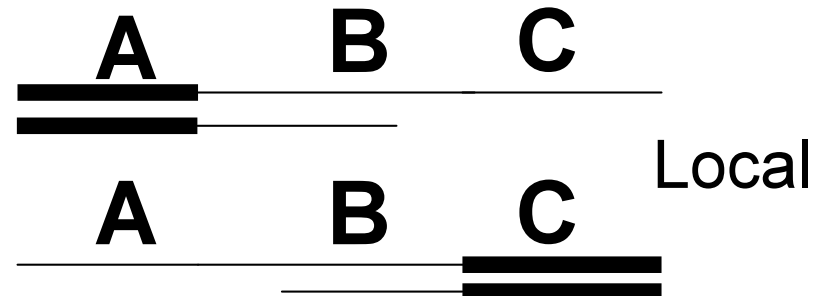
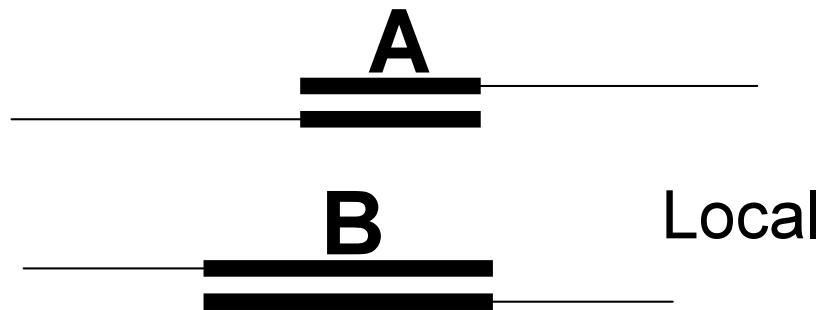
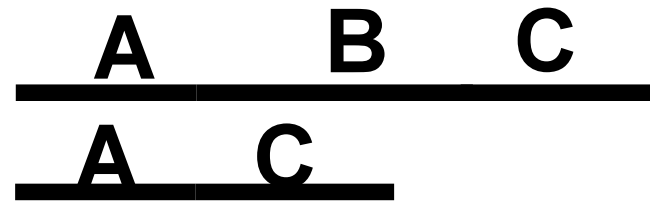
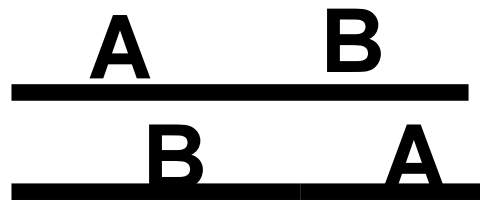
6 I 6 s 6 6 6 sI 1 s6 G k 6 s D 6 I 6 sg

Humlbpa : CSSDIADVEVDMSGD--SGWLLNLFHNQIESKFEQKVLERSICEMIQKSVSSDLQPYLQTL PVTTEIDSFADIDYSLVEAPRATA : 241
Rablpb : CSSDIQNVELDIEGD--LEELLNLLQSQIDARLREVLESKICRQIEEAVTAHLQPYLQTL PVTQIDSFAGIDYSLMEAPRATA : 242
Ratlbp : CSNRIRDLELHVSGN--VGWLLNLFHNQIESKLFQKVLERSICEMIQKSVTSDLQPYLQTL PVTADIDTILGIDYSLVAAPQAKA : 241
Humcetp : CYLSFHKLLLHLQGEREPGWIKQLETFNFI SFTLKLVLKQIQCKEI-NVISNIMADVFQTRAASILSDGDI GVDISLTGDEVITA : 243
Maccetp : CYLSFHKLLLHLQGEREPGWIKQLETFNFI SFTLKLVLKQIQCKEI-NIISNIMADVFQTRAASILSDGDI GVDISLTGDEIITA : 243
Rabcetp : CYLAFHKLLLHLQGEREPGWIKQLETFNFI SFTLKLVLKQIQCKEI-NTISNIMADVFQTRAASILSDGDI GVDISVTGAPVITA : 226
Humbpi : CSSHINSVHVHISKSK-VGWLIIQLFHKKIESALRNKMNSQVCEKVTNSVSSKLPQYFQTL PVMTKIDSVAGINYGLVAPRATA : 248
Bovbpi : CSSGINTVRIHISGSS-LGWLIIQLERKRIESL LQKSMTRKICEVVTSTVSSKLPQYFQTL PVTTKLDKAVGVDYSLVAPRATA : 244

C 6 6h6 g gw6 Lf I 1 6 6C 6 63 6 5 QT D g61 s6 P tA

For Help, press F1 Column: 1 NUM

Globalizované lokální seřazení



Formát sekvencí – multi FASTA

```
>S.nepalensis
AATACATGCAAGTCGAGCGAACAGATAAAGGAGCTTGCTCCTTTGACGTTAG
CGCGCGACGGGTGAGTAACACGTGGGTAACTACCTATAAGACTGGAATAACTCCGGGAAACCGGGGCTA
ATGCCGGATAATATTTAGAACCGCATGGTCTAAAAGTGAAAGATGGTTTTGCTATCACTTATAGATGGAC
CGCGCCGTATTAGCTAGTTGGTGGGGTAATGGCTTACCAAGGCAACGATACGTAGCCGACCTGAGAGGG
TGATCGCCACACTGGAAGTGAACACGGTCCAGACTCTACGGGAGGCAGCAGTAGGGAACTCTCCGCA
ATGGCGAAAGCCTGACGGAGCAACCGCCGCTGAGTGATGAAGGCTTCCGGATCGTAAAACCTGTATT
AGGGAAGAACAAATGTGTAAGTAACTGTGCACGCTTGACGGTACCTAATCAGAAAGCCACGGCTAACTA
CGTCCAGCAGCCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATATTGGCGTAAAGCGCGCTA
GGCGGTYTCTTAAGTCTGATGTGAAAGCCACCGCTCAACCGTGGAGGGTCATTGAAACTGGGAACT
TGAGTGCAGAAGANGAAAGTGAATTCC
>S.cohnii.Lepidoptera
AATACATGCAAGTCGAGCGAACAGATAAAGGAGCTTGCTCCTTTGACGTTAGCGCGGACGGGTGAGTA
ACACGTGGGTAACTACCTATAAGACTGGAATAACTCCGGGAAACCGGGGCTAATGCCGGATAATATTTA
GAACCGCATGGTCTAAAAGTGAAAGATGGTTTTGCTATCACTTATAGATGGACCCGCGCTATTAGCTA
GTTGGTGGGTAATGGCTCACCAAGGCAACGATACGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGA
ACTGAGACACGGTCCAGACTCTACGGGAGGCAGCAGTAGGGAACTCTCCGCAATGGCGAAAGCCTGAC
GGAGCAACCGCCGCTGAGTGATGAAGGCTTCCGATCGTAAAACCTGTATTAGGGAAGAACAATGTG
TAAGTAACTGTGCACGCTTGACGGTACCTAATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCG
TAATACGTAGTGGCAAGCGTTATCCGGAATATTGGCGTAAAGCGCGCTAGCCGGTTCTTAAGTCT
GATGTGAAAGCCACCGCTCAACCGTGGAGGGTCATTGAAACTGGGAACTTGAAGTGCAGAAGAGGAAA
GTGAATTCC
>S.cohnii.cohnii
AATACATGCAAGTCGAGCGAACAGATAAAGGAGCTTGCTCCTTTGAC
GTTAGCGCGGACGGGTGAGTAACACGTGGGTAACTACCTATAAGACTGGAATAACTCCGGGAAACCGG
GGCTAATGCCGGATAACATTTAGAACCGCATGGTCTAAAAGTGAAAGATGGTTTTGCTATCACTTATAGA
TGGACCCGCGCTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAACGATACGTAGCCGACCTGA
GAGGGTGATCGGCCACACTGGAAGTGAACACGGTCCAGACTCTACGGGAGGCAGCAGTAGGGAACTCT
CCGCAATGGCGAAAGCCTGACGGAGCAACCGCCGCTGAGTGATGAAGGCTTCCGATCGTAAAACCTG
TTATTAGGGAAGAACAATGTGTAAGTAACTATGCACGCTTGACGGTACCTAATCAGAAAGCCACGGCT
AACTACGTGCCAGCAGCCCGGTAACTACGTAGGTGGCAAGCGTTATCCGGAATATTGGCGTAAAGCGC
CGGTAGCCGGTTCTTAAGTCTGATGTGAAAGCCACGGCTCAACCGTGGAGGGTCATTGAAACTGGGA
AACTTGAGTGCAGAAGAGGAAAGTGAATTCC
>S.cohnii.urealyt
AATACATGCAAGTCGAGCGAACAGATAA
GGAGCTTGCTCCTTTGACGTTAGCGCGGACGGGTGAGTAACACGTGGGTAACTACCTATAAGACTGGA
ATAACTCCGGGAAACCGGGCTAATGCCGGATAACATTTAGAACCGCATGGTCTAAAAGTGAAAGATGGT
TTTGCTATCACTTATAGATGGACCCGCGCTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAAC
GATACGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGAAGTGAACACGGTCCAGACTCTACGGGAG
GCAGCAGTAGGGAACTCTCCGCAATGGCGAAAGCCTGACGGAGCAACCGCCGCTGAGTGATGAAGGCT
TCGGATCGTAAAACCTGTATTAGGGAAGAACAATGTGTAAGTAACTGTGCACGCTTGACGGTACCT
AATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCCGGTAACTACGTAGGTGGCAAGCGTTATCCGGAA
TTATTGGCGTAAAGCGCGCTAGCCGGTTCTTAAGTCTGATGTGAAAGCCACGGCTCAACCGTGGAG
GGTCATTGGAACCTGGGAACTTGAAGTGCAGAAGAGGAAAGTGAATTCC
>S.xylosus.type
AATACATGCAAGTCGAGCGAACAGATAAAGGAGCTTGCTCCTTTGAA
GTTAGCGCGGACGGGTGAGTAACACGTGGGTAACTACCTATAAGACTGGGATAACTCCGGGAAACCGG
AGCTAATACCGGATAACATTTAGAACCGCATGGTCTAAAAGTGAAAGATGGTTTTGCTATCACTTATAGA
TGGACCCGCGCTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAACGATACGTAGCCGACCTGA
GAGGGTGATCGGCCACACTGGAAGTGAACACGGTCCAGACTCTACGGGAGGCAGCAGTAGGGAACTCT
CCGCAATGGCGAAAGCCTGACGGAGCAACCGCCGCTGAGTGATGAAGGTTCCGGATCGTAAAACCTG
TTATTAGGGAAGAACAATGTGTAAGTAACTGTGCACATCTGACGGTACCTAATCAGAAAGCCACGGCT
AACTACGTGCCAGCAGCCCGGTAACTACGTAGGTGGCAAGCGTTATCCGGAATATTGGCGTAAAGCGC
CGGTAGCCGGTTCTTAAGTCTGATGTGAAAGCCACGGCTCAACCGTGGAGGGTCATTGAAACTGGGA
AACTTGAGTGCAGAAGAGGAAAGTGAATTCC
```