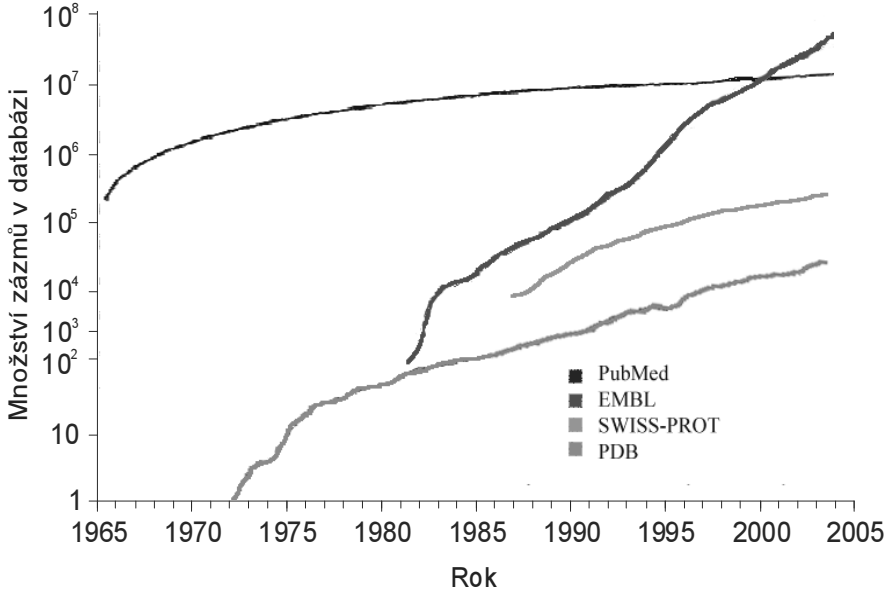
Bioinformatika je nová disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie

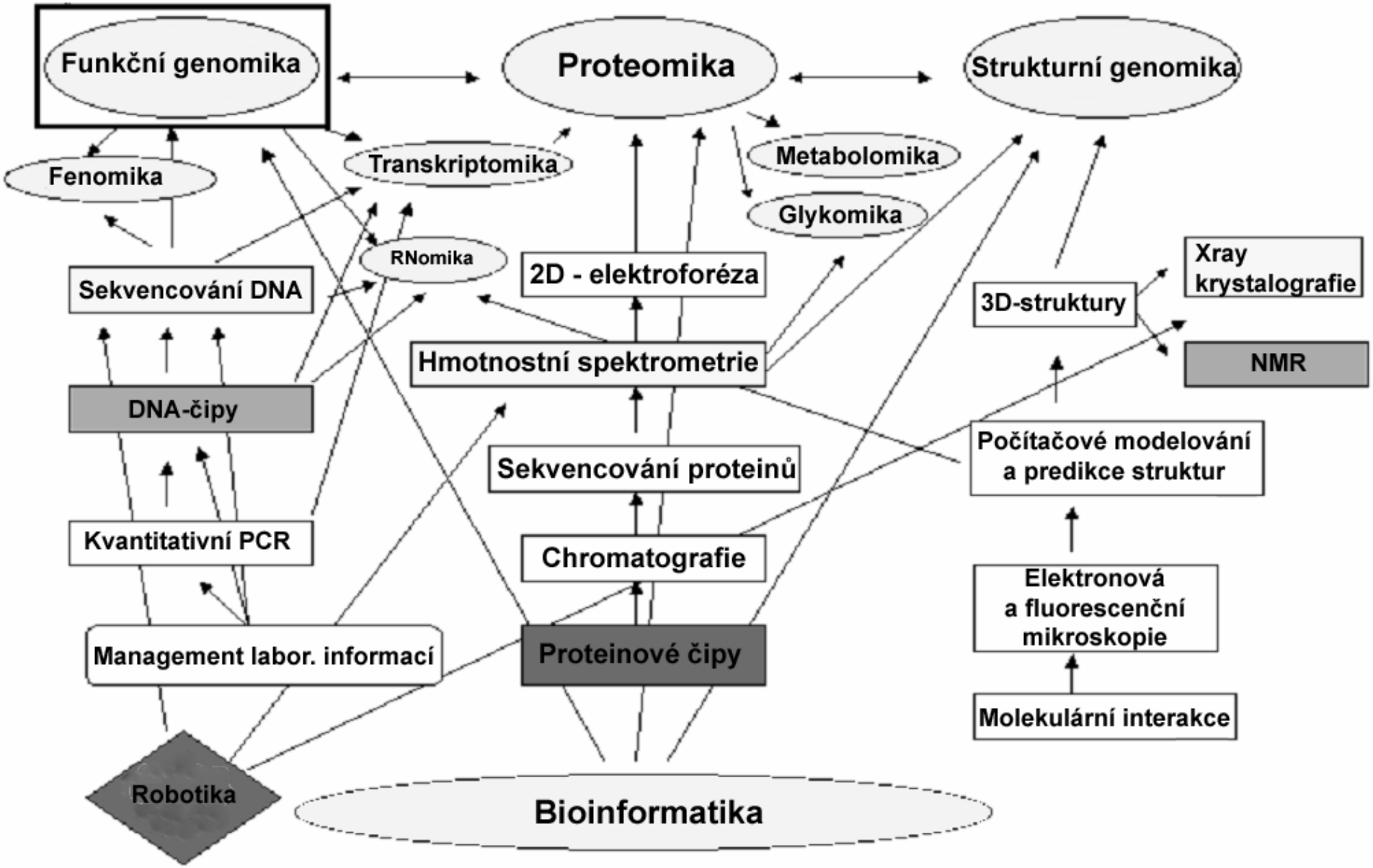
- Bioinformatika zahrnuje
  - studium
  - praktické uchovávání
  - vyhledávání
  - zobrazování
  - manipulaci
  - a modelování biologických dat
- Vývoj vysoce výkonných technologií umožňujících získání molekulárně biologických dat přispěl k jejich dramatickému nárůstu a tím současně zvýšil obtížnost 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	Zpracování laboratorních dat	Predikce struktur
Zpracování obrazu		Srovnávání struktur
Robotika		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
  - mechanizmy fyziologických funkcí

- Primárním cílem těchto analýz je objasnění informačního obsahu biomaromolekul 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 předmětu Bi5000 Úvod do bioinformatiky (zpracoval R.Pantůček)

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

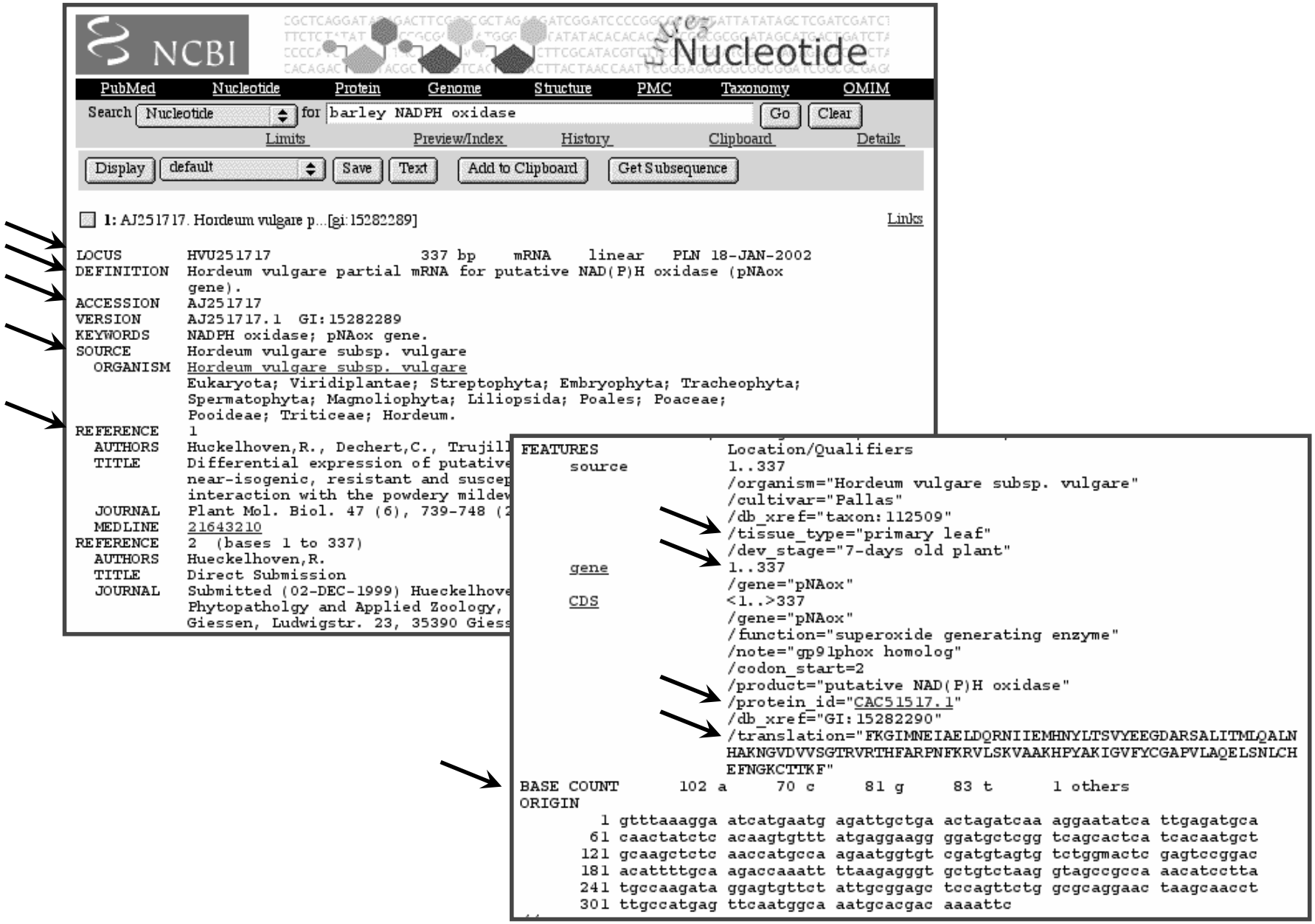
- V současné době je prostřednictvím Internetu dostupných přibližně 550 databází zabývajících se shromažďováním bioinformací.
  - Jejich přehled a popis je každoročně publikován ve specializovaném, volně dostupném čísle časopisu Nucleic Acids Research.
- K nejdůležitějším institucím zabývajícím se, správou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

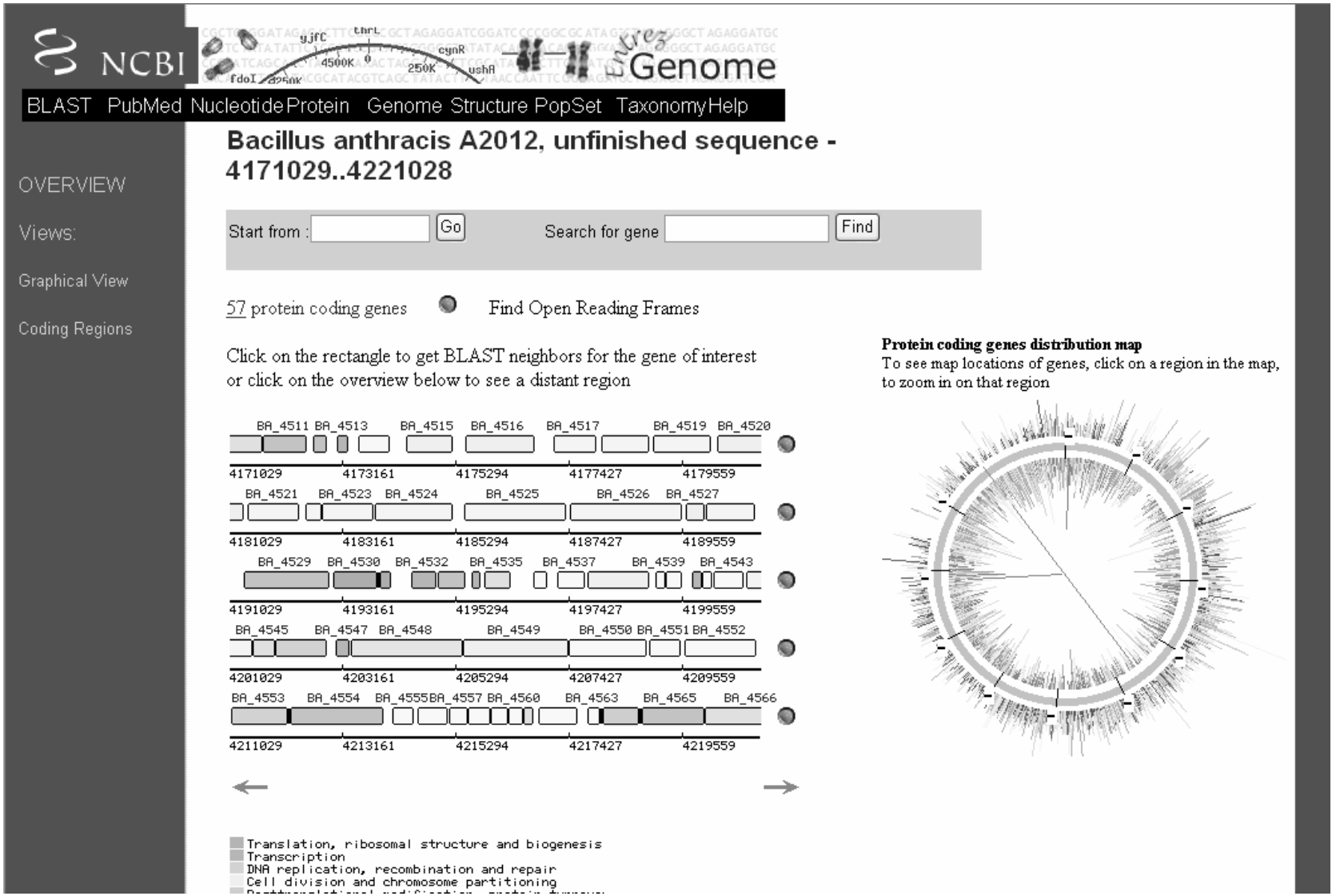
- V každém ze tří hlavních bioinformatických center je spravována **genomová databáze** sekvencí nukleových kyselin a odpovídajících, z nich přeložených proteinů.
  - **EMBL Nucleotide Sequence Database** (v rámci institutu EBI) – 1980
  - **GenBank** (v rámci institutu NCBI) – 1982
  - **DDBJ** (The 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 skvencováním 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 xxxxxxxxx nukleotidových bází pocházejících celkem od více než 60 000 různých organizmů 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.

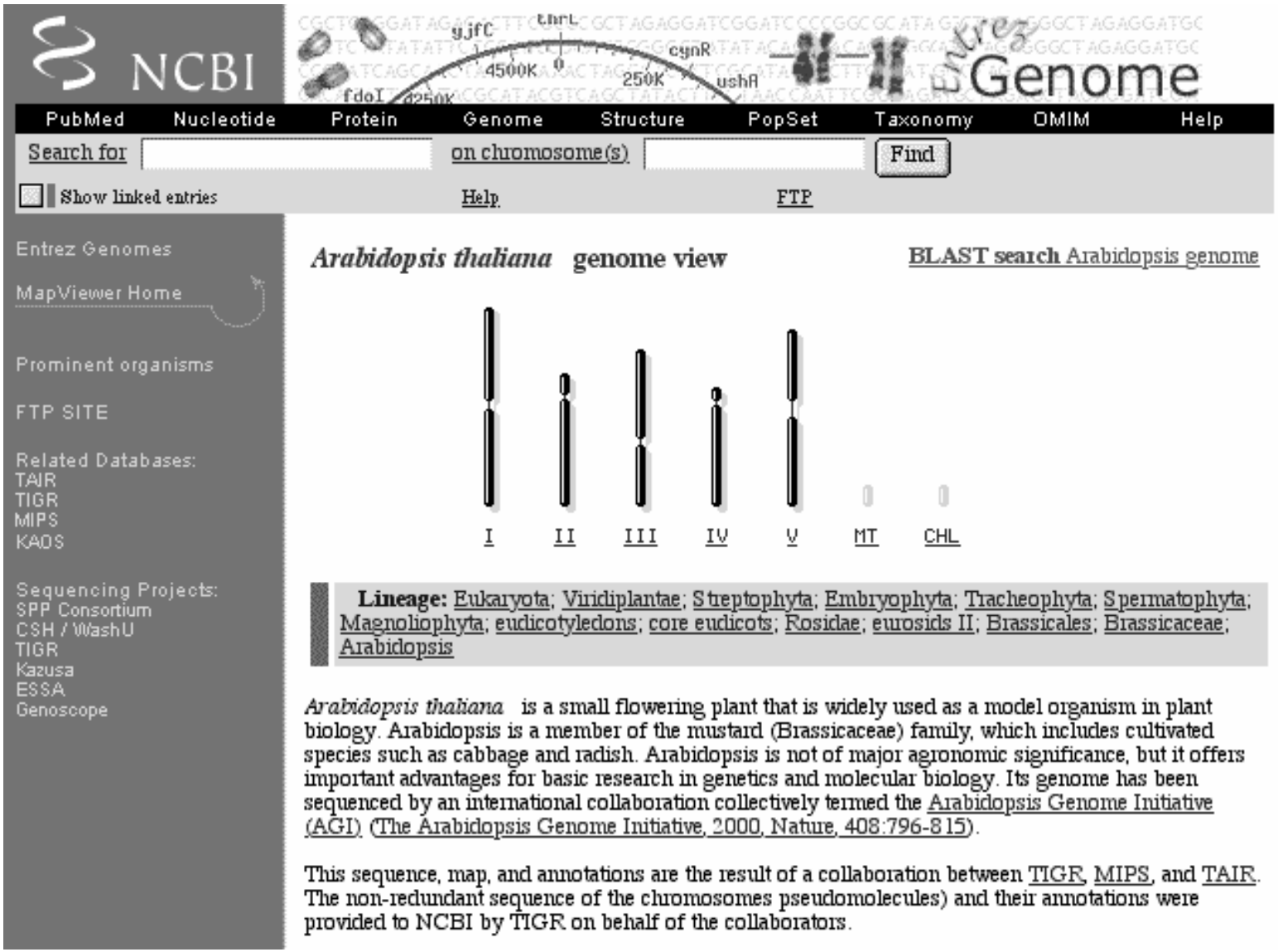
- GenBank http://www.ncbi.nlm.nih.gov/Genbank/



Genomové databáze v NCBI – prokaryota

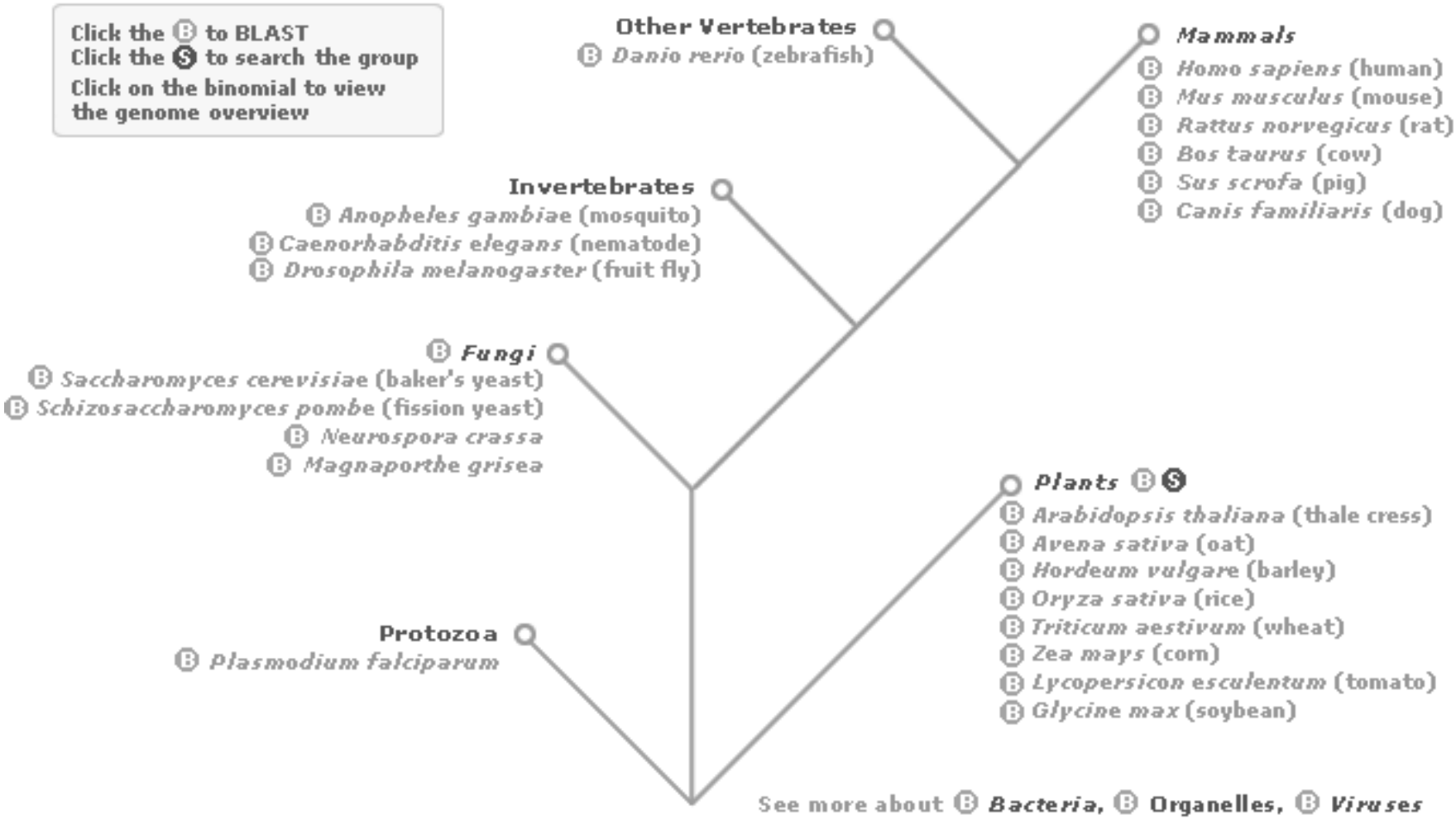


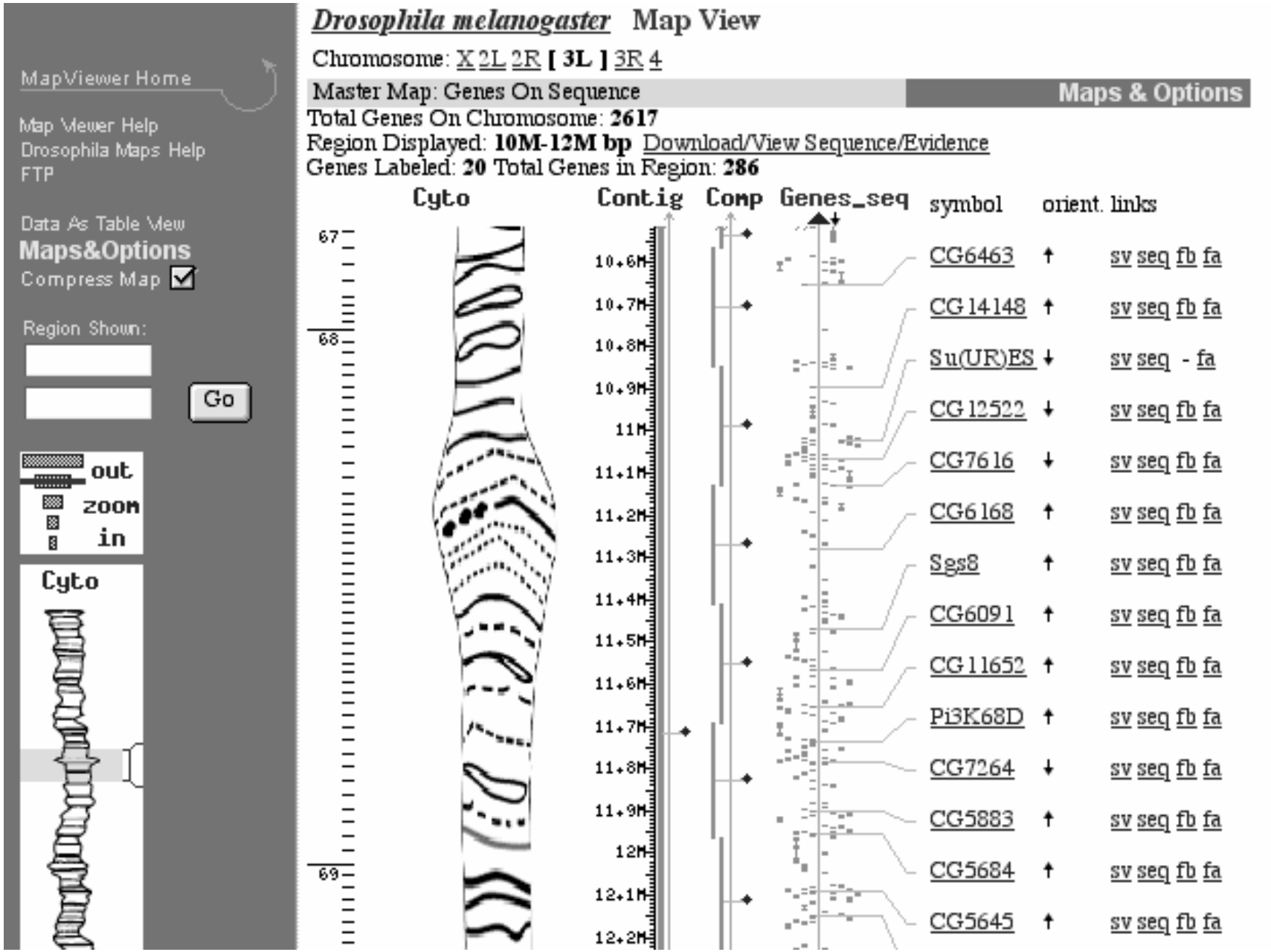
Genomové databáze v NCBI - eukaryota



Gemonové mapy - MapView

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




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
Cyto
Contig
Comp
Genes_seq
symbol
orient.
links
67
68
69
10+6M
10+7M
10+8M
10+9M
11M
11+1M
11+2M
11+3M
11+4M
11+5M
11+6M
11+7M
11+8M
11+9M
12M
12+1M
12+2M
CG6463 sv seq fb fa
CG14148 sv seq fb fa
Su(UR)ES sv seq - fa
CG12522 sv seq fb fa
CG7616 sv seq fb fa
CG6168 sv seq fb fa
Sgs8 sv seq fb fa
CG6091 sv seq fb fa
CG11652 sv seq fb fa
Pi3K68D sv seq fb fa
CG7264 sv seq fb fa
CG5883 sv seq fb fa
CG5684 sv seq fb fa
CG5645 sv seq fb fa

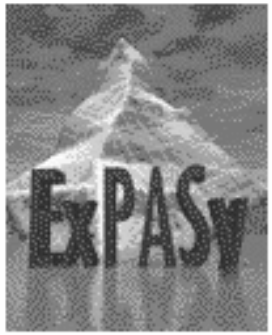
Databáze sekvencí proteinů

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

- 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 (**Ex**pert **P**rotein **A**nalysis **Sy**stem) 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).

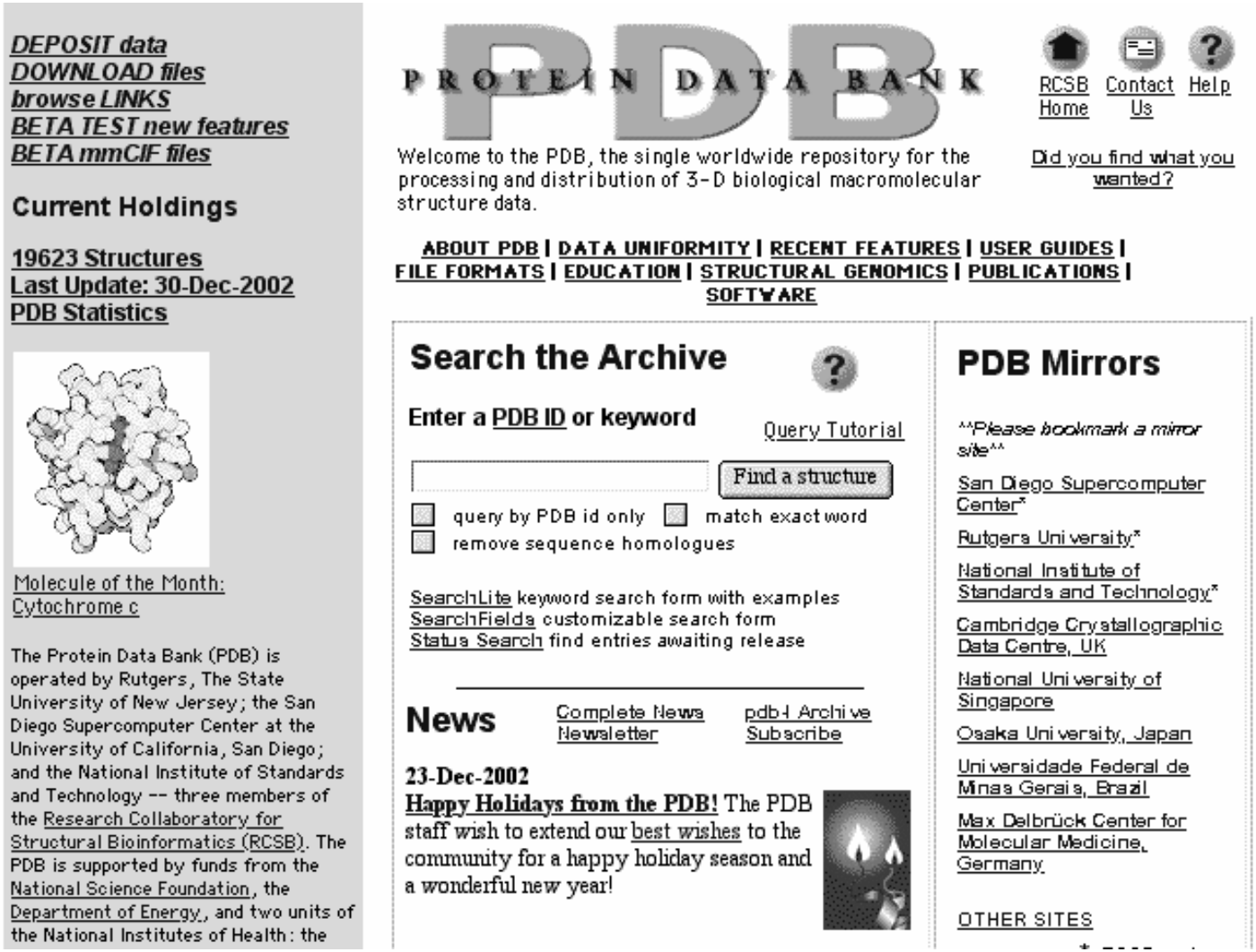
[Announcements] [Job opening] [Mirror Sites]

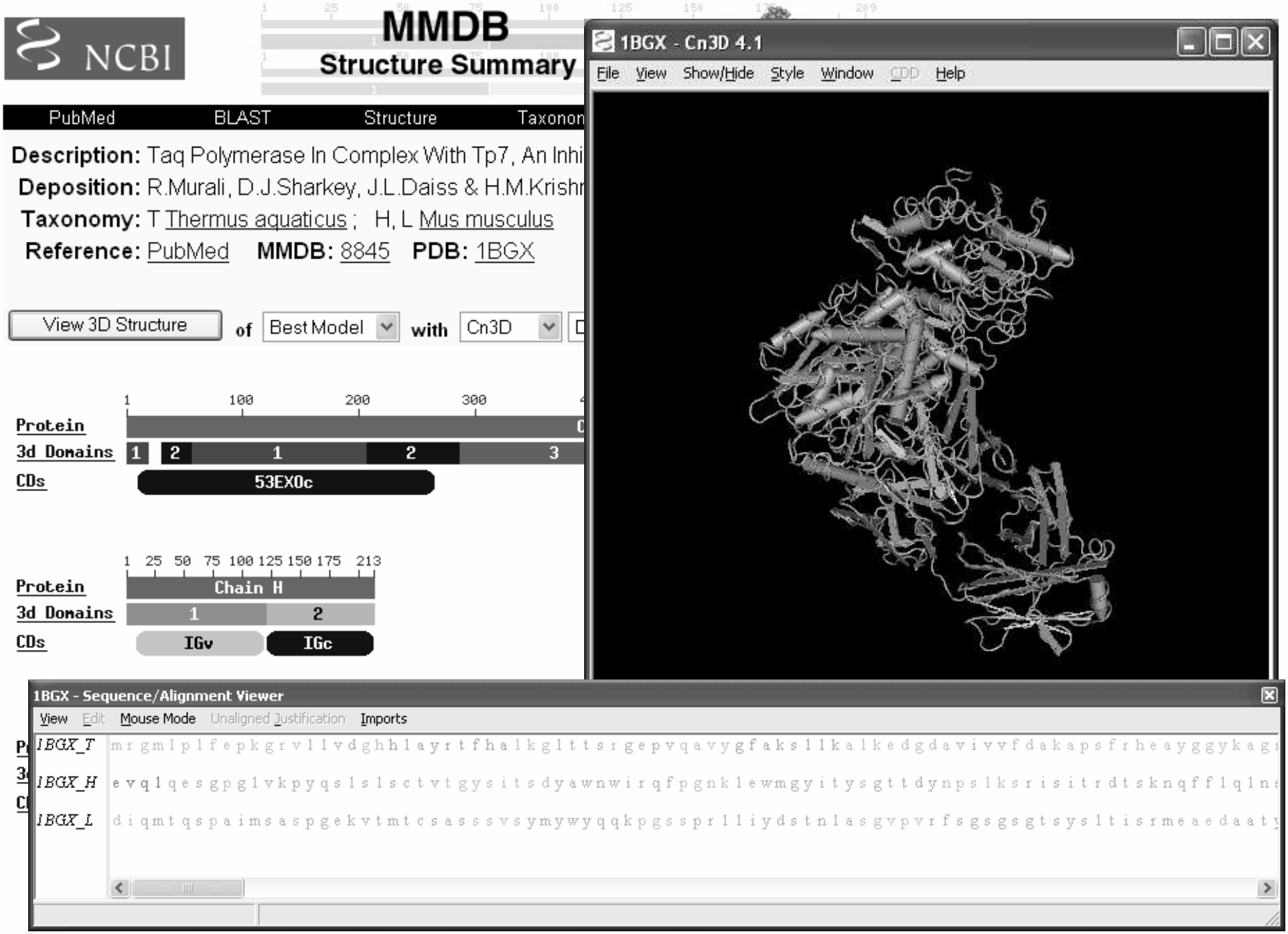
Databases	Tools and software packages

• **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**
	• **Proteomics and sequence analysis tools**
	  ○ 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/




NCBI
MMDB
Structure Summary
PubMed
BLAST
Structure
Description: Taq Polymerase In Complex With Tp7, An Inhi
Deposition: R.Murali, D.J.Sharkey, J.L.Daiss & H.M.Krishr
Taxonomy: T Thermus aquaticus ; H, L Mus musculus
Reference: PubMed MMDB: 8845 PDB: 1BGX
View 3D Structure
of
Best Model
with
Cn3D
Protein
3d Domains
CDs
53EXOc
Chain H
IGv
IGc
1BGX - Cn3D 4.1
File View Show/Hide Style Window CDD Help
1BGX - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
1BGX_T
1BGX_H
1BGX_L

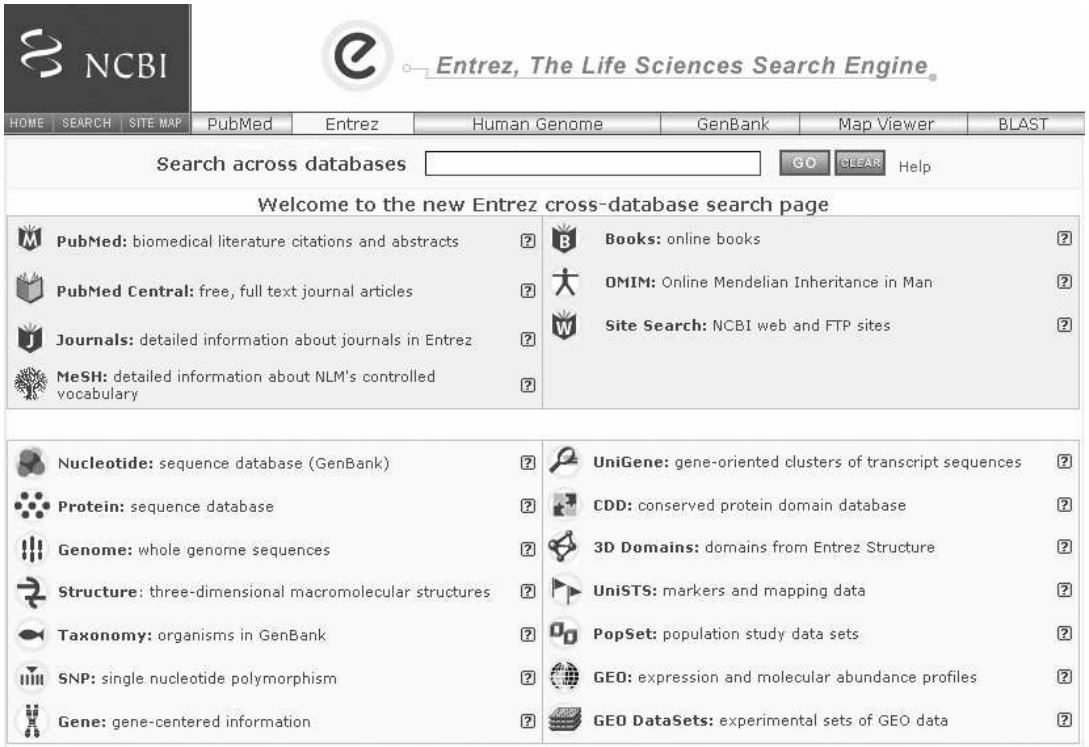
Textové vyhledávání v databázích

- 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 vztahujcích 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.

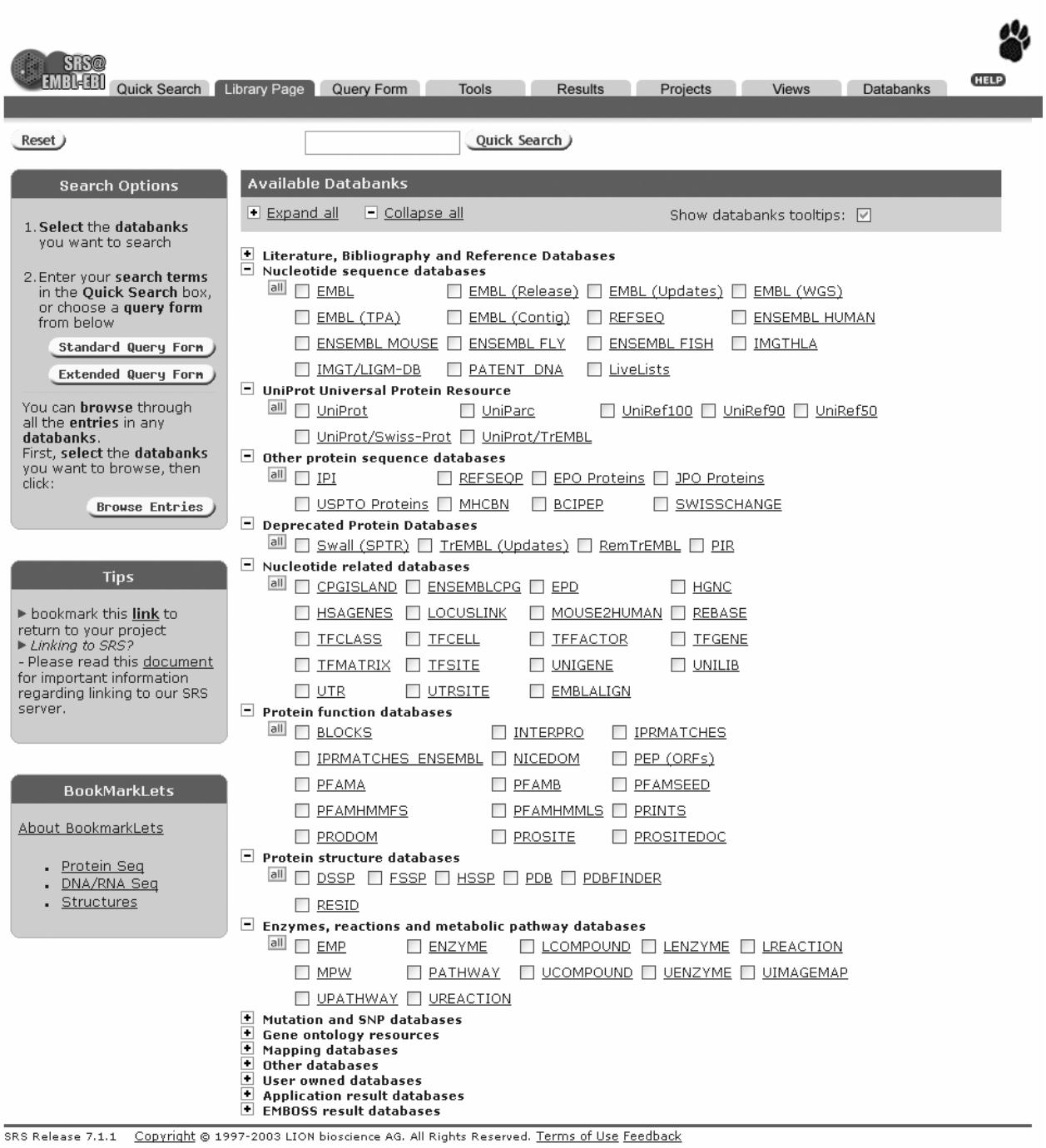
- K nejvýznamnějším databázím patří
  - databáze PubMed, umožňující přístup k literární databázi MEDLINE
  - databáze sekvencí nukleových kyselin a proteinů
  - databáze 3-D struktur MMDB (Molecular Modeling Database)
  - skupina databází genomů
  - taxonomická databáze usnadňující získávání sekvencí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



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

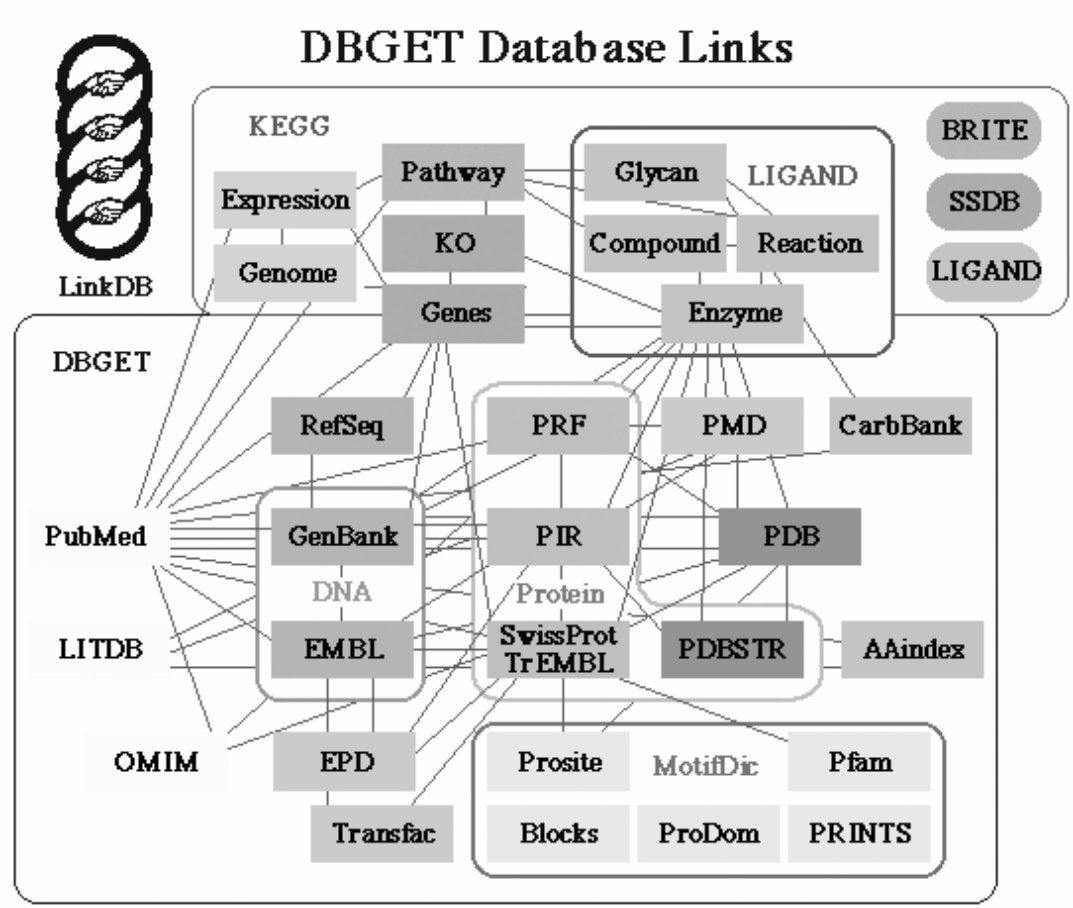

SRS@ EMBL-EBI
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
HELP
Reset
Quick Search
Search Options
1. Select the databanks you want to search
2. Enter your search terms in the Quick Search box, or choose a query form from below
Standard Query Form
Extended Query Form
You 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 REFSEQP 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 TFCELL 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
SRS Release 7.1.1 Copyright © 1997-2003 LION 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 organizmů ze známým genomem.
- V porovnání se SRS a Entrez je však DBGET jednodušší a omezenější vyhledávací prostředek.

DBGET/Link DB



GenomeNet Japan

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	dna		refseq	rs	NCBI
			genbank	gb	NCBI
			embl	emb	EBI
Protein sequences	protein	sptrembl	swissprot	sp	ExPASy
			trembl	tr	EBI
			trembl_new	trnew	
			pir	pir	NBRF
			prf	prf	PRF
			pdbstr	str	RCSB
3D structures			pdb	pdb	
Sequence motifs			epd	epd	ISREC
	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	See: KEGG for individual organisms		
		eubacteria			
		archaea			
KEGG organisms			genome	gn	
Gene expressions			expression	exp	
Virus gene catalogs			vgenes	vg	NCBI
Organelle gene catalogs			ogenes	og	
Genetic diseases			omim	mim	NCBI
Protein mutants			pmd	pmd	DDBJ
Amino acid indices	aaindex		aaindex1	aax1	Kyoto
			aaindex2	aax2	
Carbohydrate structures			carbbank	ccsd	Teikyo 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

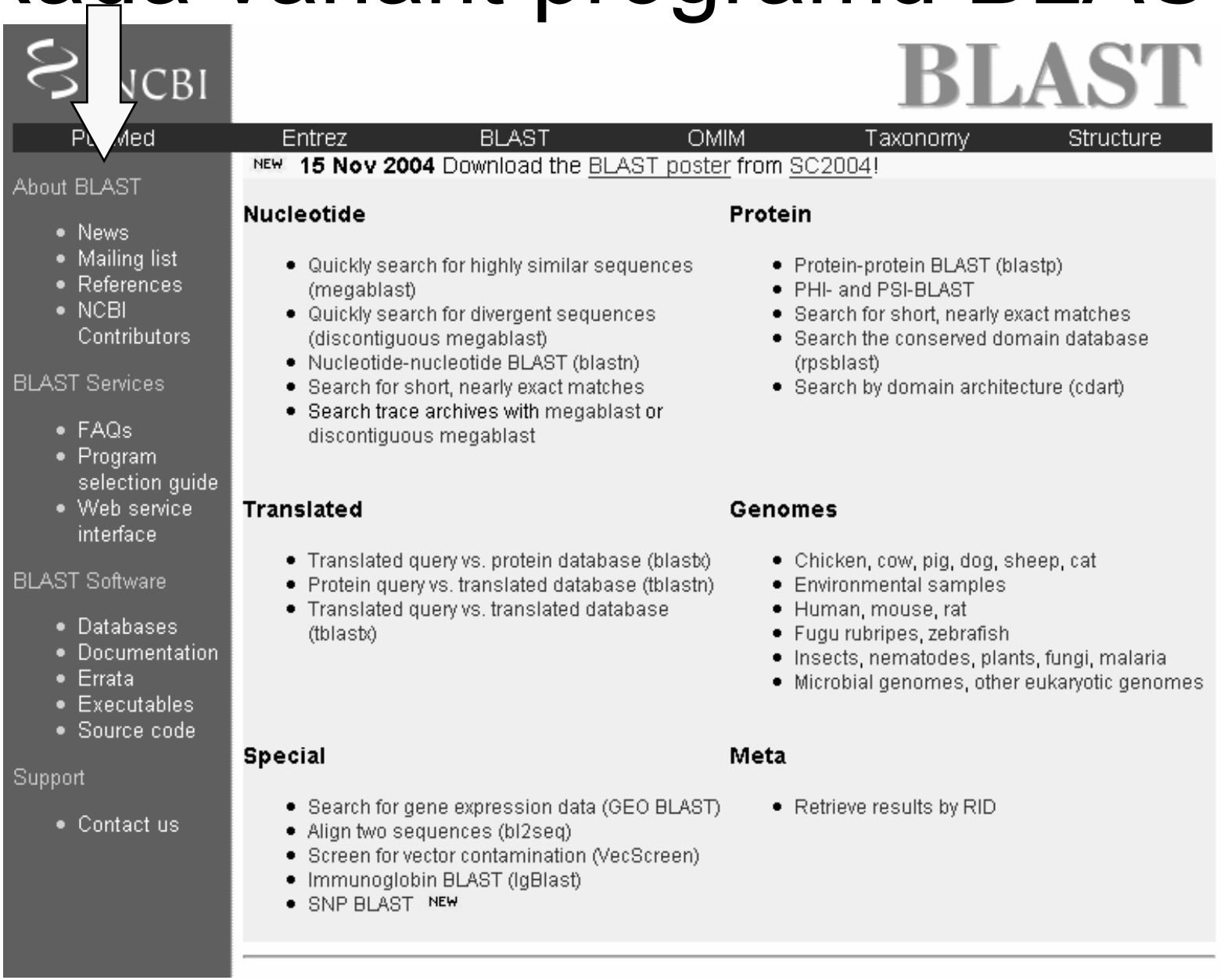
- BLAST
- Altschul et al., 1990
- dostupný na serveru NCBI

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

Co je to BLAST?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Hledání lokálních podobností
  - Heuristický přístup založený na Smith-Watermanově algoritmu
  - Vyhledá nejoptimálnější **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



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
blastn	DNA	DNA	DNA	Hledání identických sekvencí DNA
blasp	Protein	Protein	Protein	Hledání homologních proteinů
blastx	DNA	Protein	Protein	Hledání genů a homologních proteinů na DNA
tblastn	Protein	DNA	Protein	Hledání genů u necharakterizovaných DNA
tblastx	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	MEGABLAST Standard BLAST (blastn)
		Vyhledání podobných sekvencí jako dotazovaná	Standard BLAST (blastn)
		Vyhledání podobných proteinů k překladu dotazované sekvence v přeložených databázích DNA	Translated BLAST (tblastx)
	Protein	**Vyhledání podobných proteinů k překladu dotazované sekvence v databázích proteinů**	Translated BLAST (blastx)
7 - 20 bp	DNA	Vyhledání vazebných míst primerů nebo krátkých motivů	Search for short, nearly exact matches

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

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

TACTGGACATG

ACTGGACATGG

CTGGACATGGA

TGGACATGGAC

GGACATGGACC

GACATGGACCC

ACATGGACCCT

. . . . . .

minimální velikost = 7

blastn default = 11

megablast default = 28

tabulka se všemi slovy dotazu

Slova pro proteinové sekvence

Dotaz: **GTQITVEDLFYNIATRRKALKN**

GTQ Velikost = 3

TQI

QIT

ITV → LTV, MTV, ISV, LSV, etc.

TVE

VED

EDL

DLF

. . .

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

tabulka se všemi slovy dotazu

Sousedící slova

Minimální požadavek pro shodu

ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT přesná shoda slova

1 nalezená shoda

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

GTQITVEDLFYNI

SEI **YYN** sousedící slova

2 nalezené shody

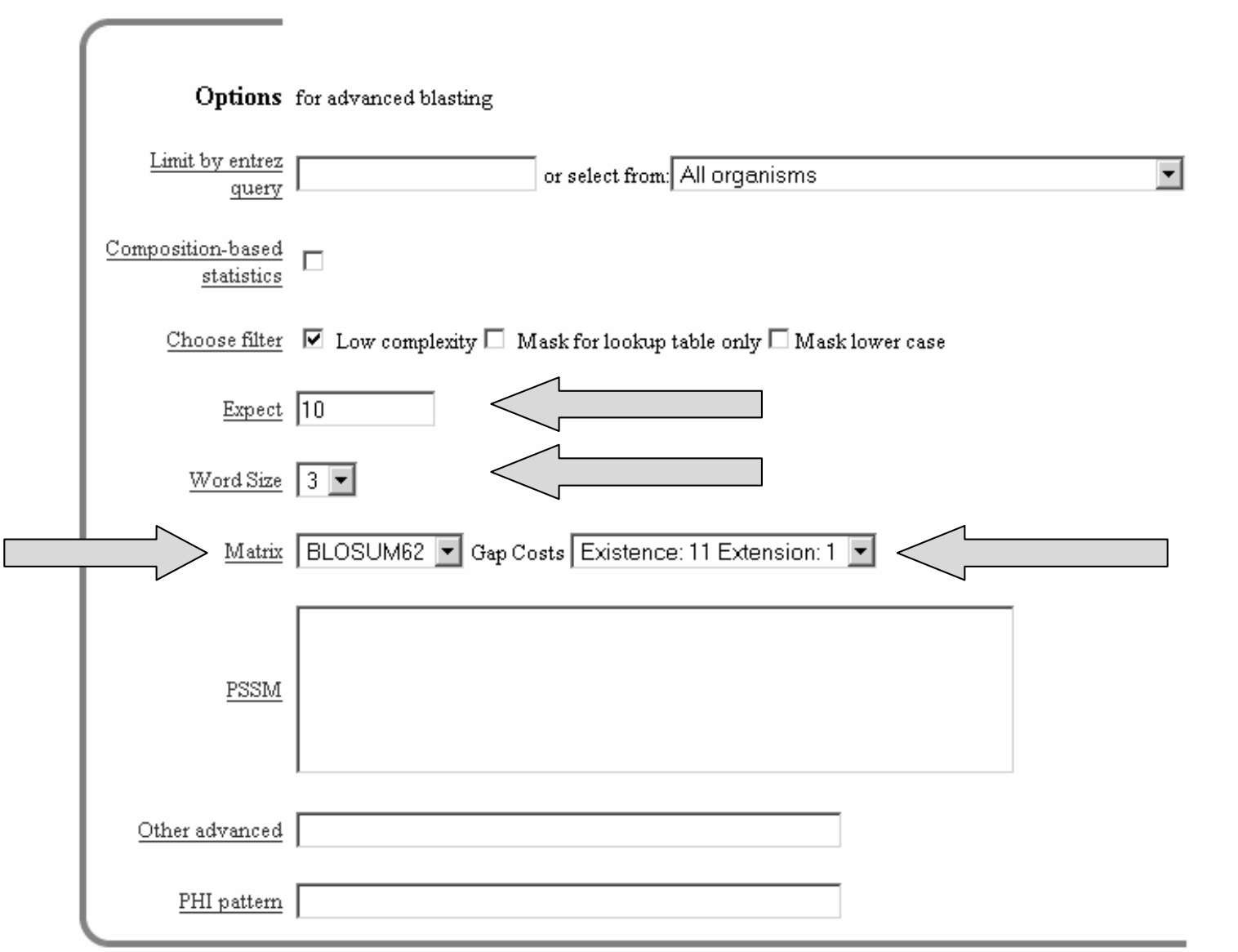
Seřazení sekvencí, které BLAST **může** nalézt

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

Seřazení sekvencí, které BLAST **nemůže** nalézt

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

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

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

BLAST - Možnosti nastavení



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 pravděpodobnot zachování v příslušné pozici sekvence než jiný.
  - Každá matrice je určená pro určitý typ vyhledávání – JE TŘEBA VĚDĚT CO HLEDÁME!

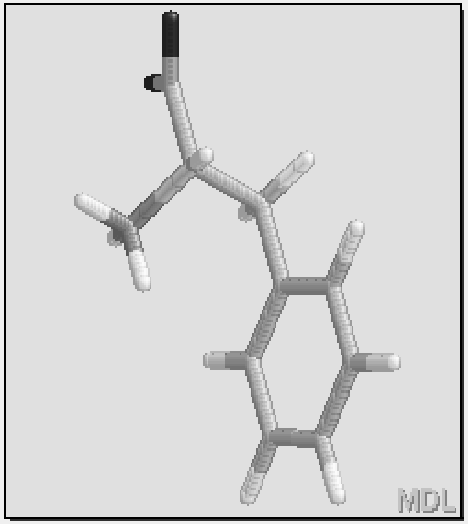
Substituční Matice

- Proč používat substituční matice?

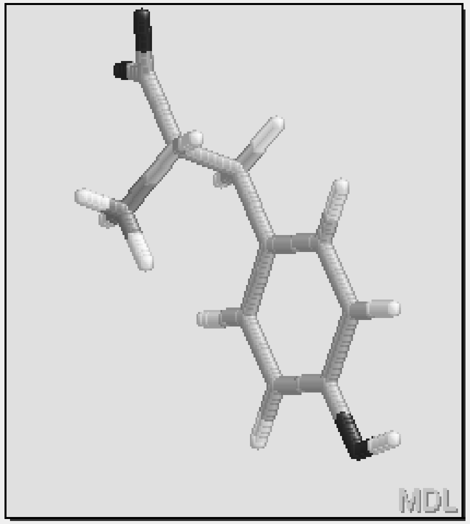
1. Stanovit pravděpodobnou homologii dvou sekvencí.
2. Substituce, které jsou více pravděpodobné získají vyšší skóre
3. Substituce, které jsou méně pravděpodobné obdrží nižší skóre.

Příklad matice BLOSUM62

L-phenylalanine (F)



L-tyrosine (Y)



Běžné a… …ší významnost

Vzácn… …šší významnost

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

Negativní pro málo pravděpodobné substituce

Pozitivní pro více pravděpodobné substituce

Různé typy substitučních matic

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

Matice identity

	A	G	C	T
A	+1	−3	−3	−3
G	−3	+1	−3	−3
C	−3	−3	+1	−3
T	−3	−3	−3	+1

```
CAGGTAGCAAGCTTGCATGTCA
||  ||||||||||||  |||||     skóre = 19-9 = 10
CACGTAGCAAGCTTG-GTGTCA
```

Matice PAM

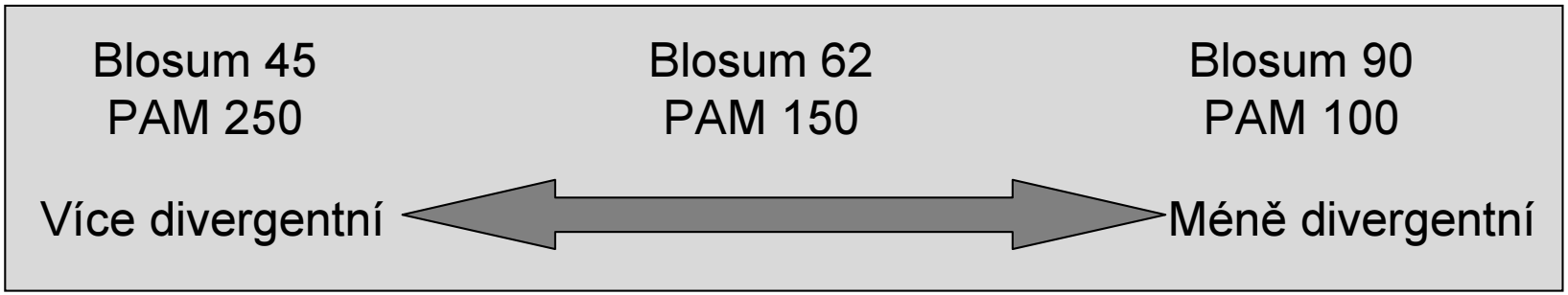
- PAM
  - **P**ercent **A**ccepted **M**utation
  - založeny na konceptu akceptovatelných bodových mutací za 10^8 let v globálních mnohonásobných seřazeních blízce příbuzných proteinů
  - Stanoveny na základě výpočtů u blízce příbuzných proteinů
  - PAM1 reprezentuje 1% změn
  - PAM250 = $(PAM1)^{250}$

Matice BLOSUM

- **BL**ocks **S**ubstitution **M**atrix
- 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

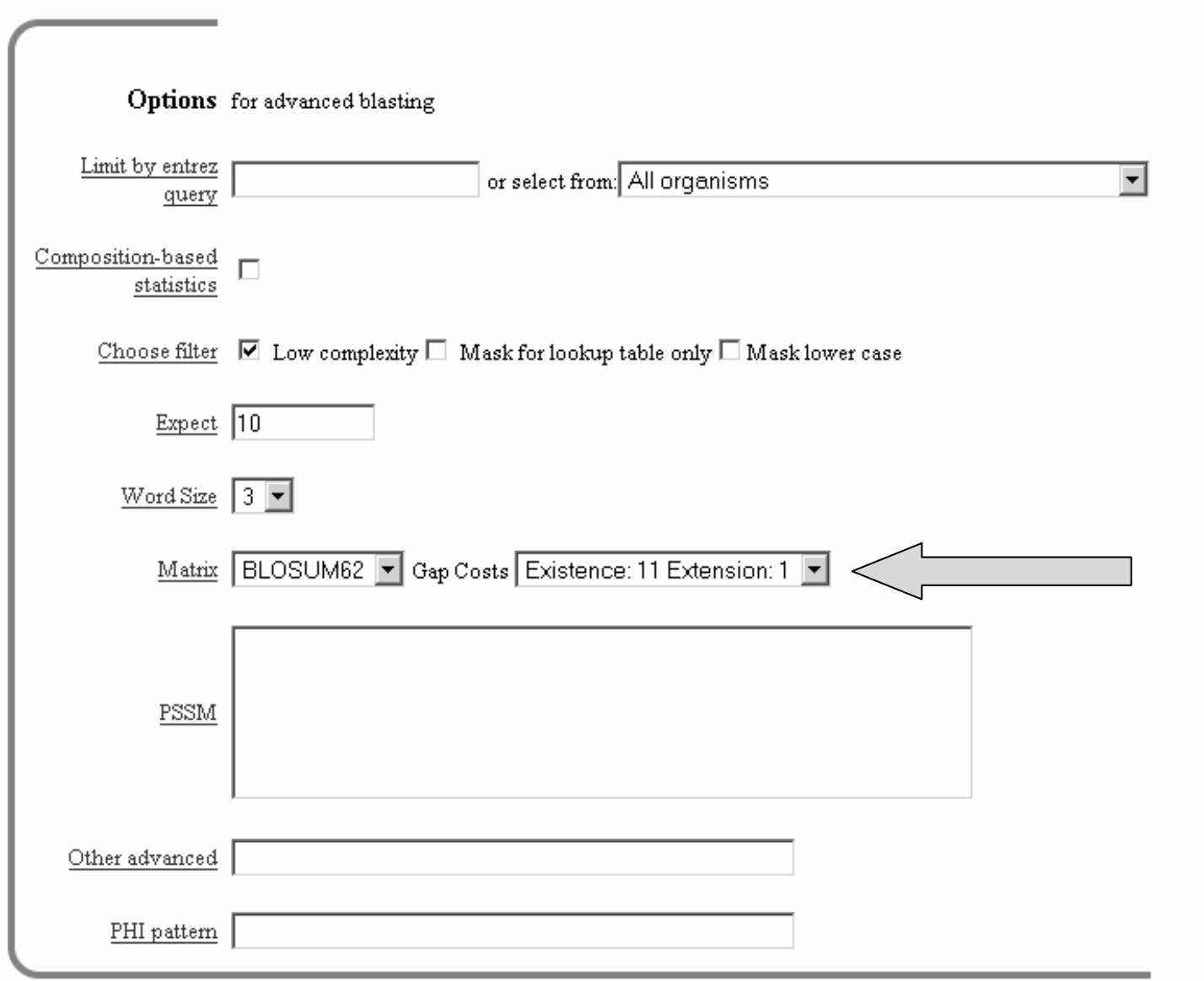
Významnost shody

- 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 sekvenční 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.

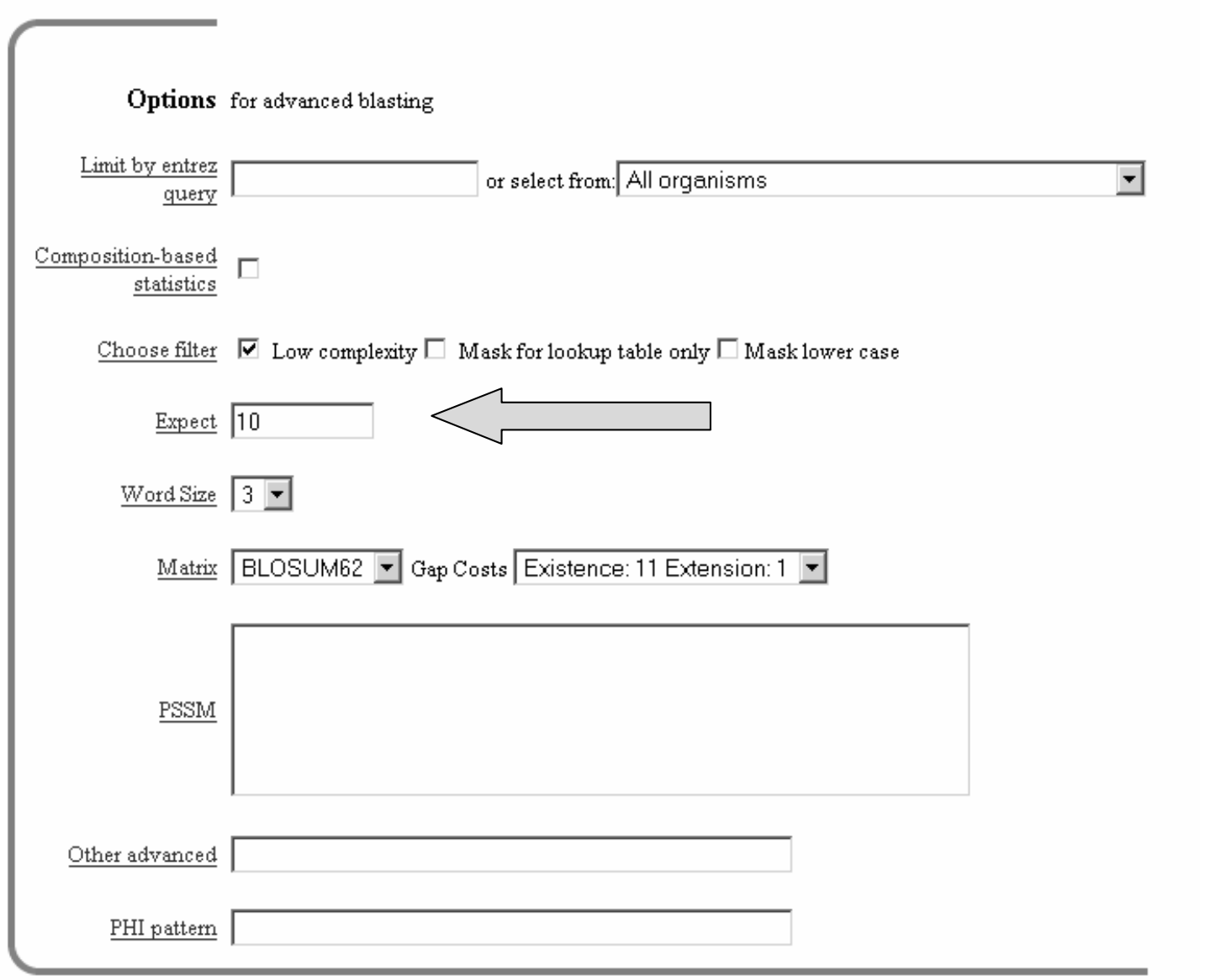
$$E = mn\,2^{-S}$$

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

BLAST - Možnosti nastavení



BLAST - Možnosti nastavení



BLAST – Výstup (Výsledky)

Skládají se ze 4 částí

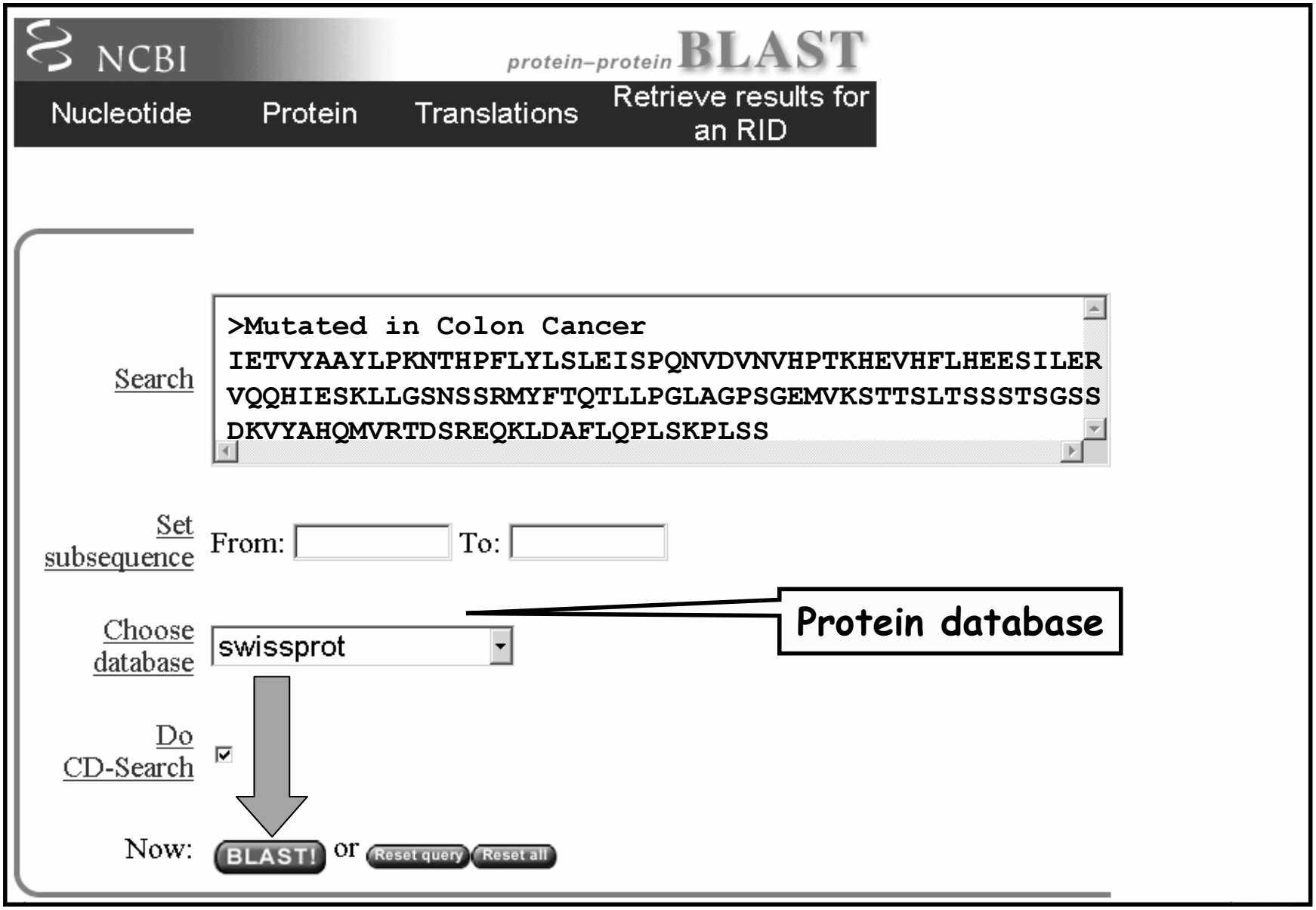
1) úvud, který informuje o tom kde bylo vyhledání provedeno a jaké databáze byly použity

2) seznam sekvencí v databázi, obsahující segmenty podobných sekvencí, jejichž skóre je alespoň tak vysoké jako zadané parametry

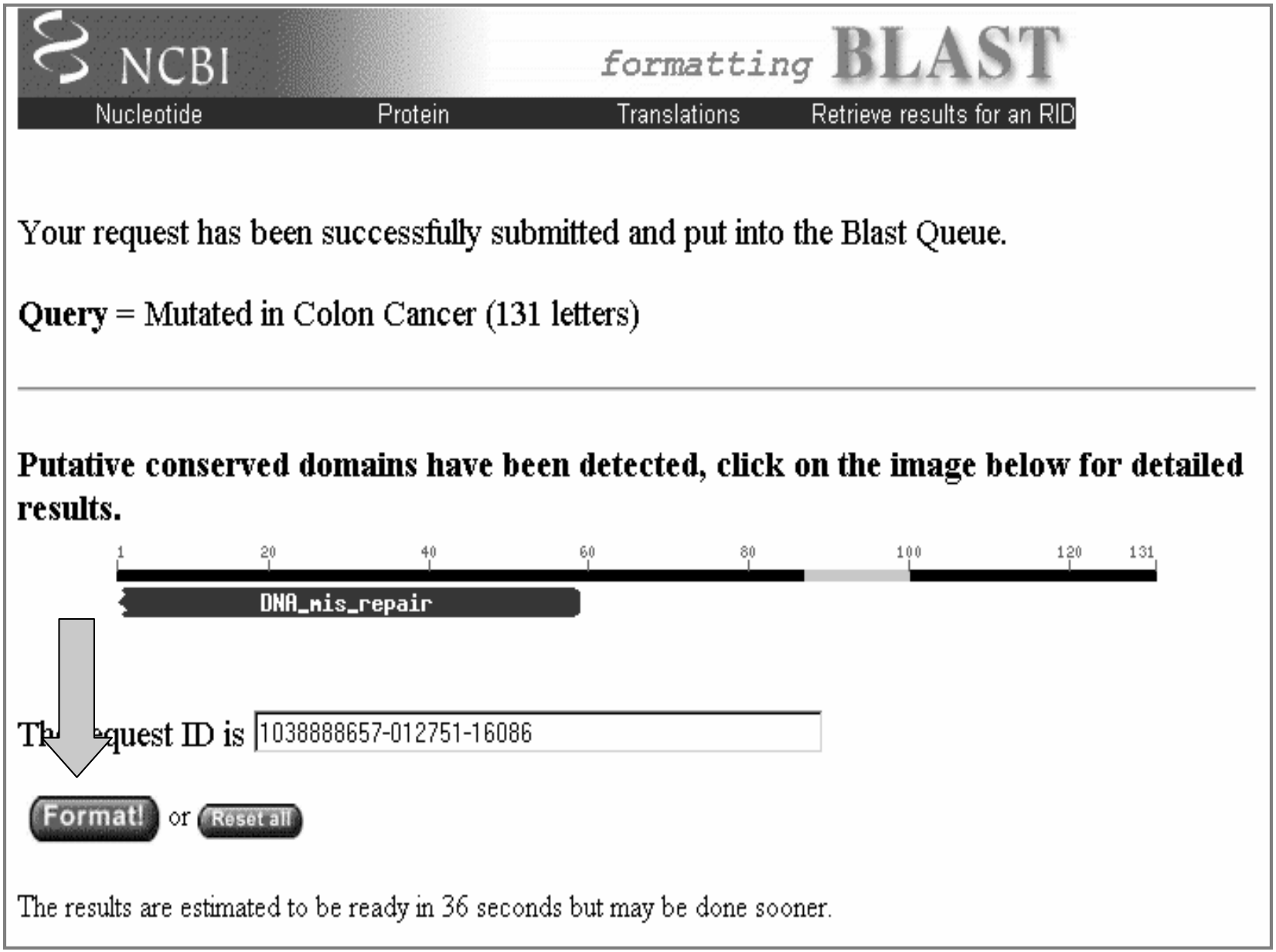
3) seřazení podobných sekvencí s vysokým skóre

4) kompletní seznam parametrů použitých pro vyhledání.

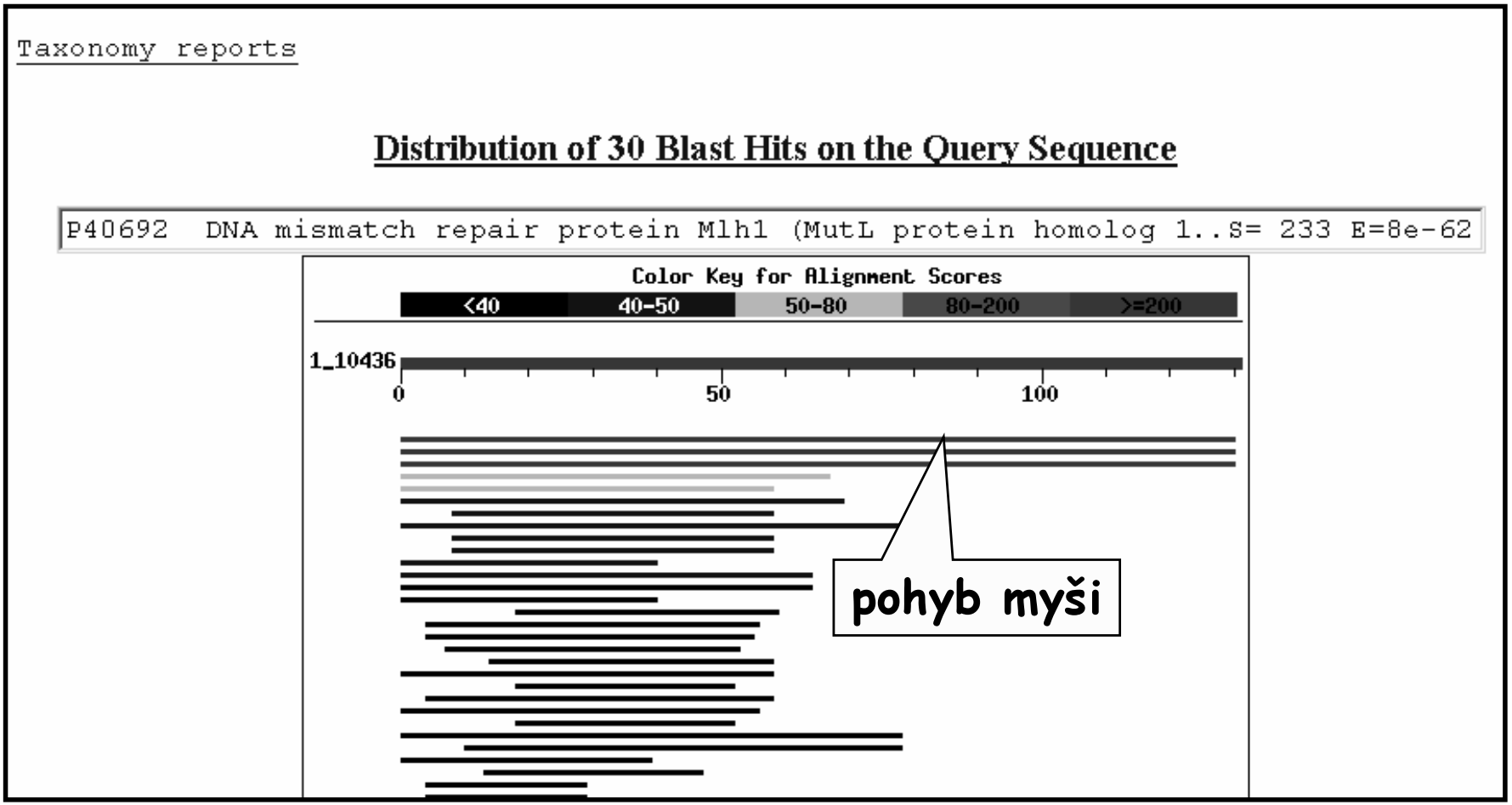
Proteinový BLAST



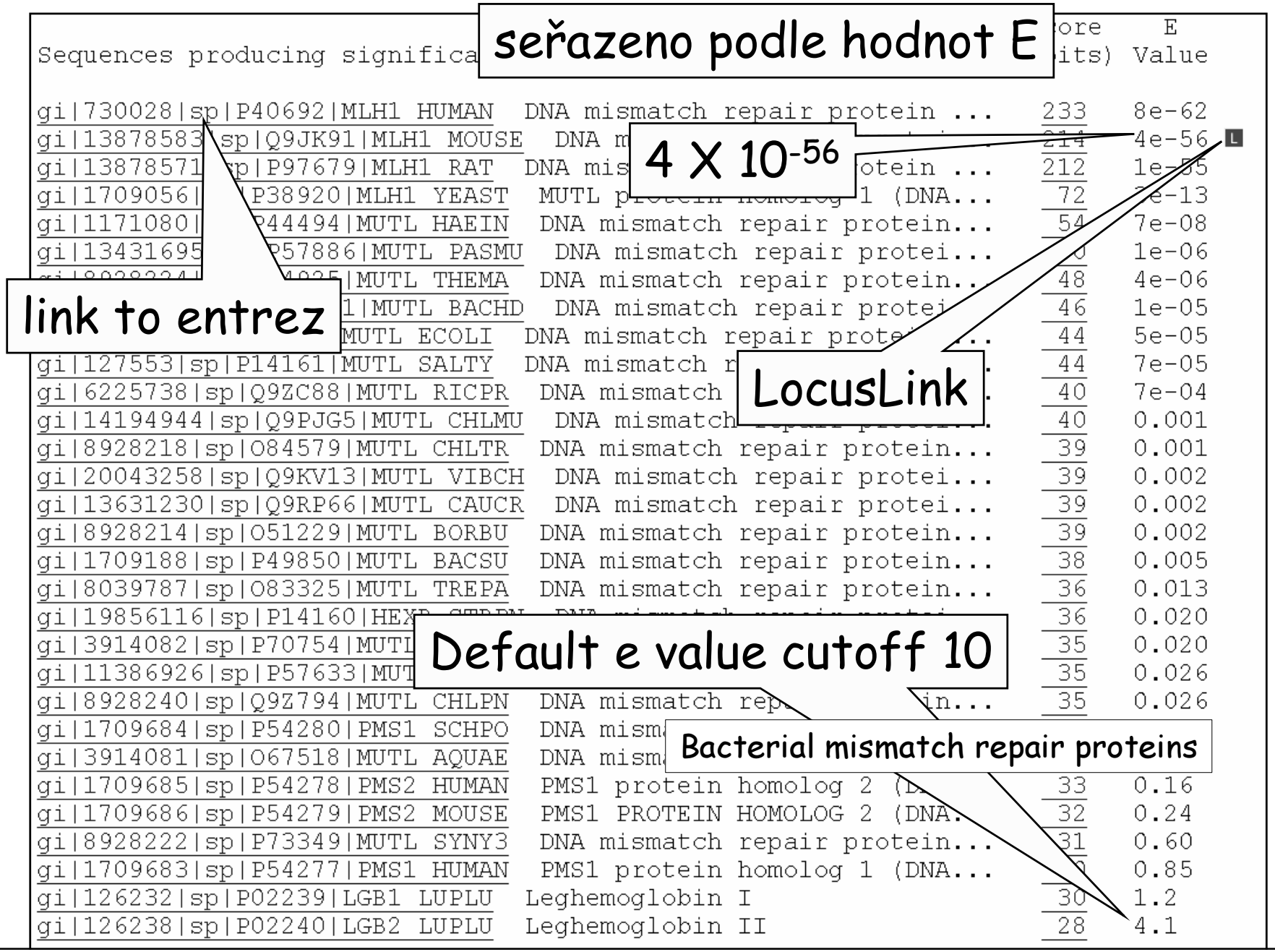
BLAST – stránka pro formátování



BLAST – grafický výstup



BLAST Output: Descriptions



BLAST – výstup seřazení sekvencí

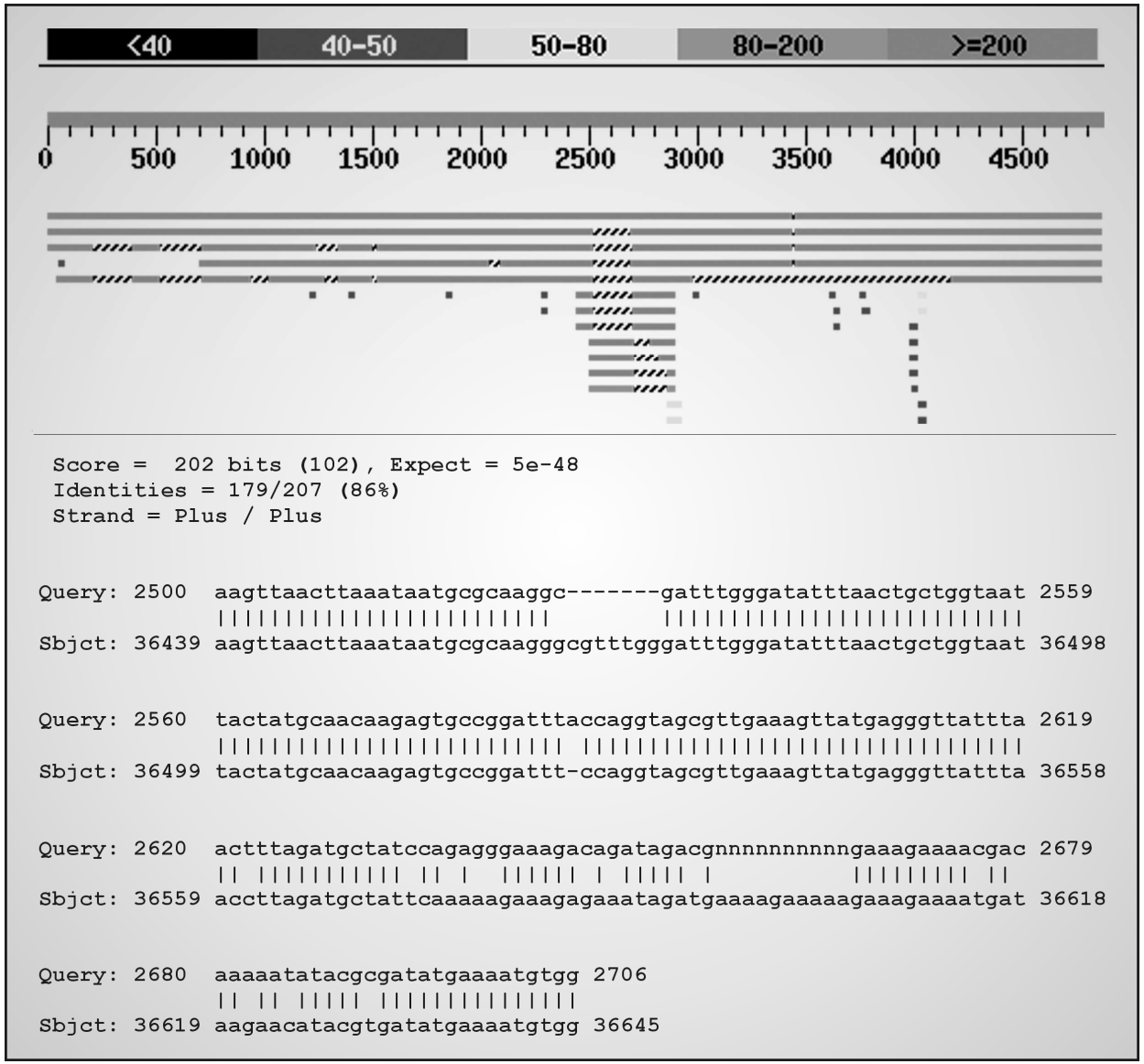
```
>gi|127552|sp|P23367|MUTL_ECOLI  DNA mismatch repair protein mutL
          Length = 615

 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 LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQQQLETPL 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  GSNSSRMYFTQTLLPGLAGPSGEMVKXXXXXXXXXXXXXXDKVYAHQMVRTDSREQKLDA 120
           GSNSSRMYFTQTLLPGLAGPSGEMVK              DKVYAHQMVRTDSREQKLDA
Sbjct: 336 GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA 395

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

sekvence s nízkou komplexitou

BLAST – příklad výstupu u DNA




```
 Score =  202 bits (102), Expect = 5e-48
 Identities = 179/207 (86%)
 Strand = Plus / Plus


Query: 2500  aagttaacttaaataatgcgcaaggc-------gatttgggatatttaactgctggtaat 2559
              ||||||||||||||||||||||||        ||||||||||||||||||||||||||||
Sbjct: 36439 aagttaacttaaataatgcgcaagggcgtttgggatttgggatatttaactgctggtaat 36498


Query: 2560  tactatgcaacaagagtgccggatttaccaggtagcgttgaaagttatgagggttattta 2619
              |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct: 36499 tactatgcaacaagagtgccggattt-ccaggtagcgttgaaagttatgagggttattta 36558


Query: 2620  actttagatgctatccagagggaaagacagatagacgnnnnnnnnnngaaagaaaacgac 2679
              || |||||||||| || |  |||||| | ||||| |          ||||||||| ||
Sbjct: 36559 accttagatgctattcaaaaagaaagagaaatagatgaaaagaaaaagaaagaaaatgat 36618


Query: 2680  aaaaatatacgcgatatgaaaatgtgg 2706
              || || ||||| ||||||||||||||
Sbjct: 36619 aagaacatacgtgatatgaaaatgtgg 36645
```

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 stuktuře genů

Algoritmus:

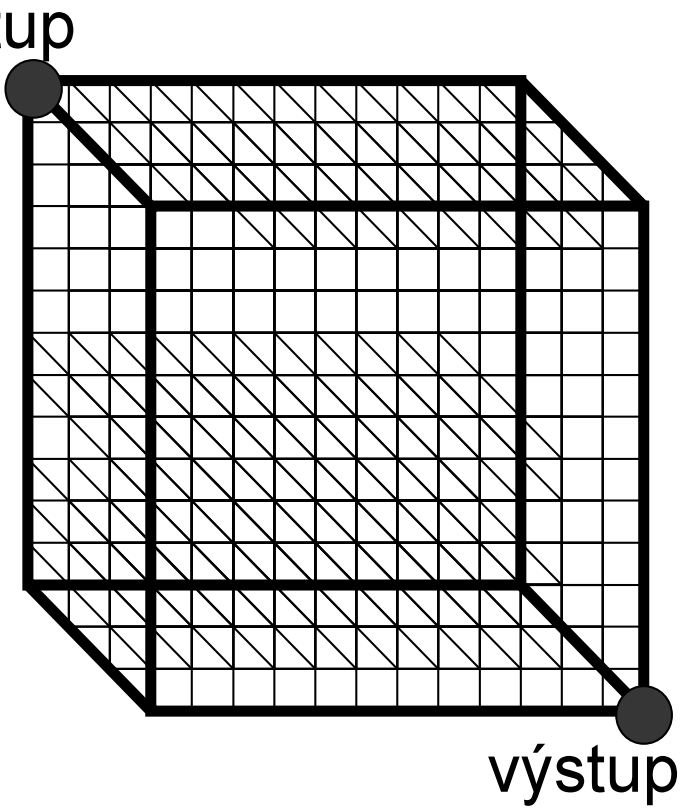
mnohonásobné seřazení = hledání optimální cesty
více konzervativních sloupců = lepší seřazení

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

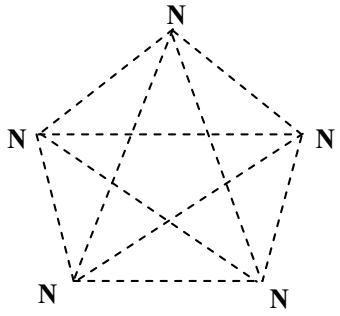


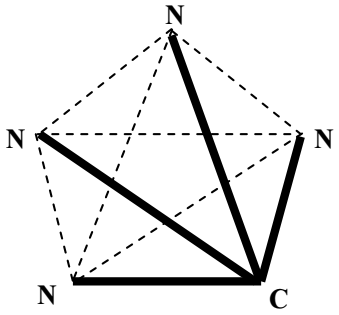
- Výsledná cesta v 3-rozměrném*(x,y,z)* prostoru:

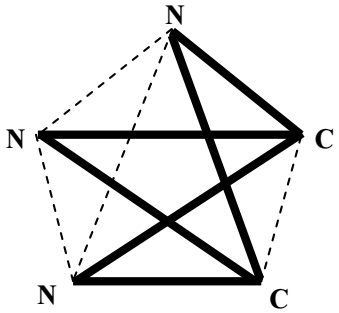
$(0,0,0) \rightarrow (1,1,0) \rightarrow (1,2,1) \rightarrow (2,3,2) \rightarrow (3,3,3) \rightarrow (4,4,4)$

Výpočet skóre seřazení

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

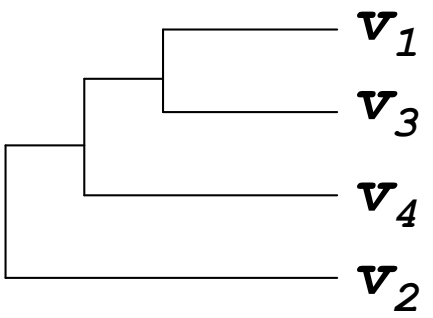
- Seřazení všech párů sekvencí
- Výpočet matice podobností (hodnoty procentuální identity)

	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$((v_{1,3,4}), v_2)$

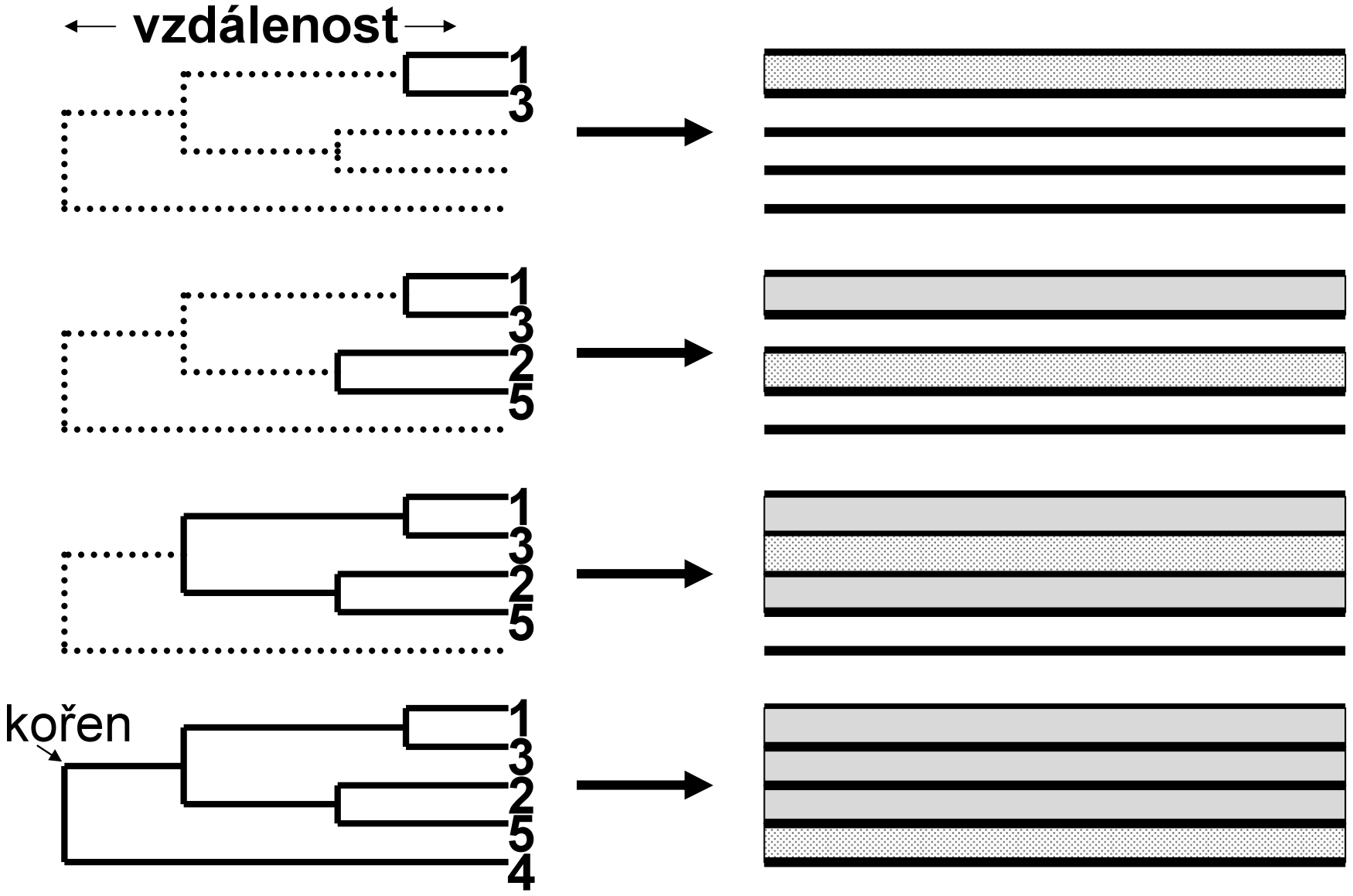
ClustalW: krok 3

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

```
FOS_RAT          PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE        PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD
FOS_CHICK        SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE       PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----------------LPFQ
FOSB_HUMAN       PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----------------LPFQ
                 .   . :   ** .     :..  *:.*   *   . *                   **:
```

Hvězdičky a tečky označují stupeň konzervovanosti sekvencí

Princip progresivního seřazení



Problém přesnosti

Při progresivním seřazení se mohou kumulovat chyby.

“Once a gap, always a gap”

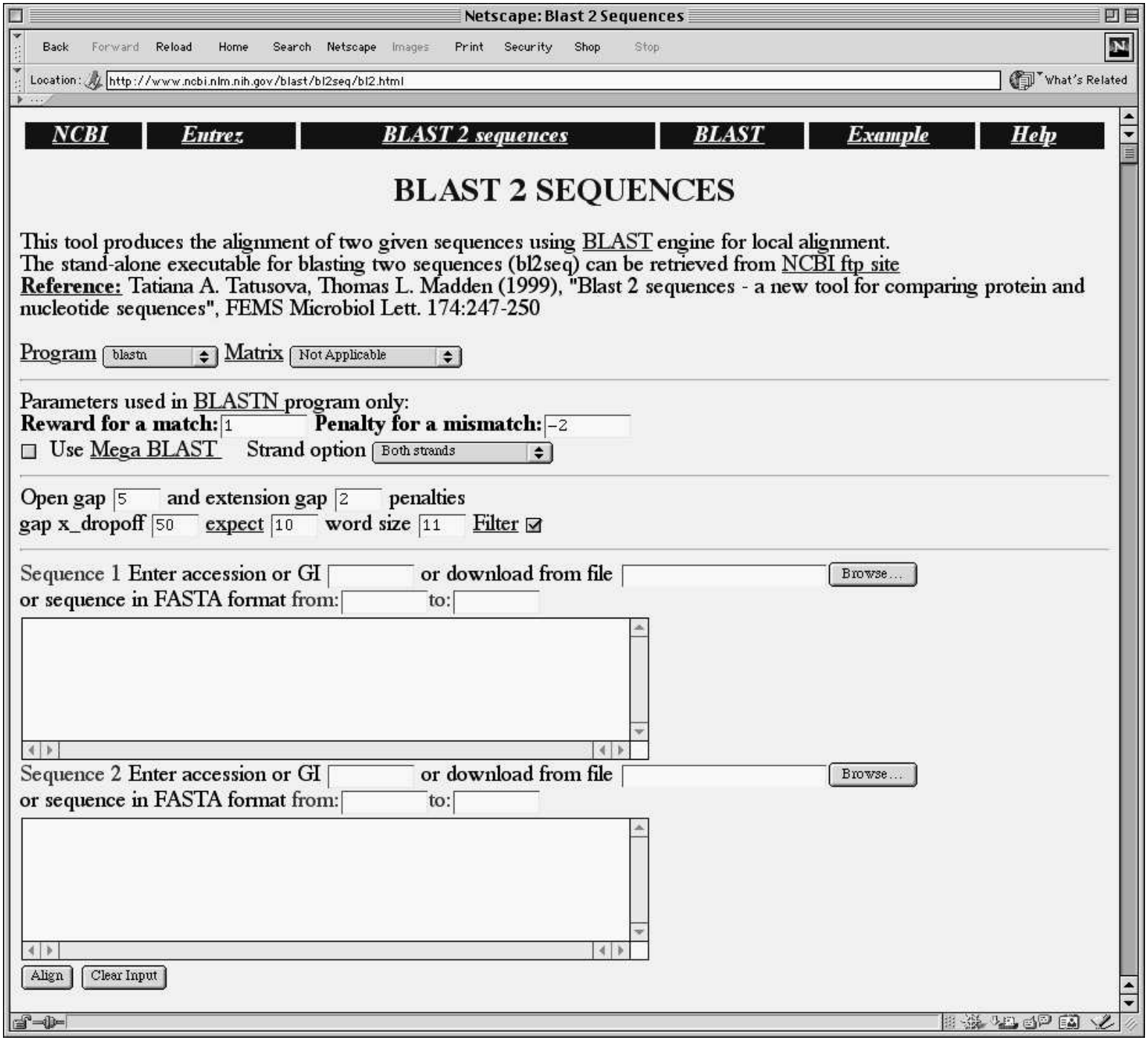
Feng & Doolittle, 1987

Prakticky prováděné kroky

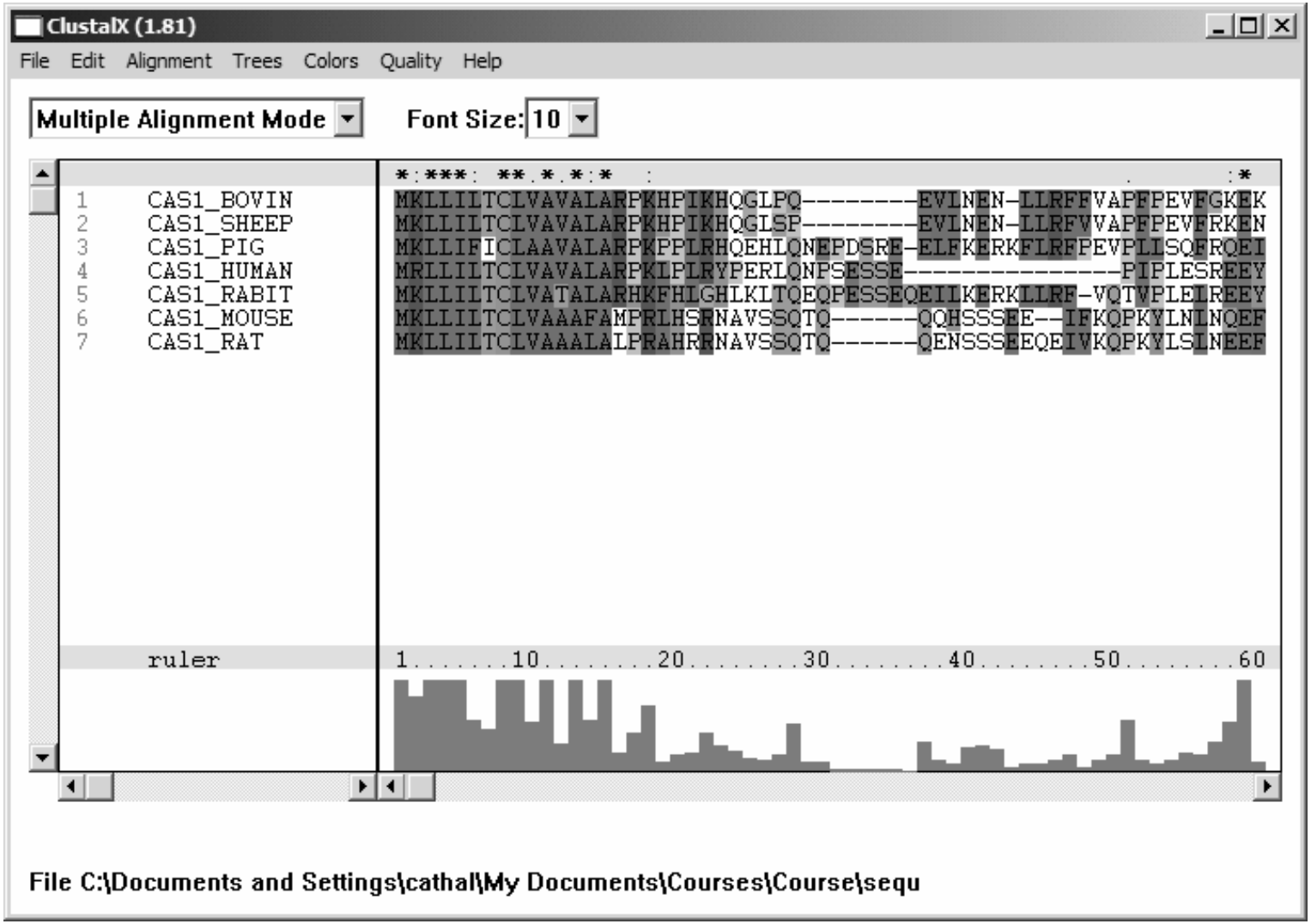
- Získání sekvencí (databáze, sekvencování)
- Manipulace se sekvencemi (změna formátu, orientační párové seřazení)
- Výběr vzájemně odpovídajících úseků
- Mnohonásobné seřazení
- Následné fylogenetické analýzy

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 blastn Matrix Not Applicable
Parameters used in BLASTN program only:
Reward for a match: 1 Penalty for a mismatch: -2
Use Mega BLAST Strand option Both strands
Open gap 5 and extension gap 2 penalties
gap x_dropoff 50 expect 10 word size 11 Filter
Sequence 1 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: to:
Sequence 2 Enter accession or GI or download from file Browse...
or sequence in FASTA format from: to:
Align
Clear Input

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

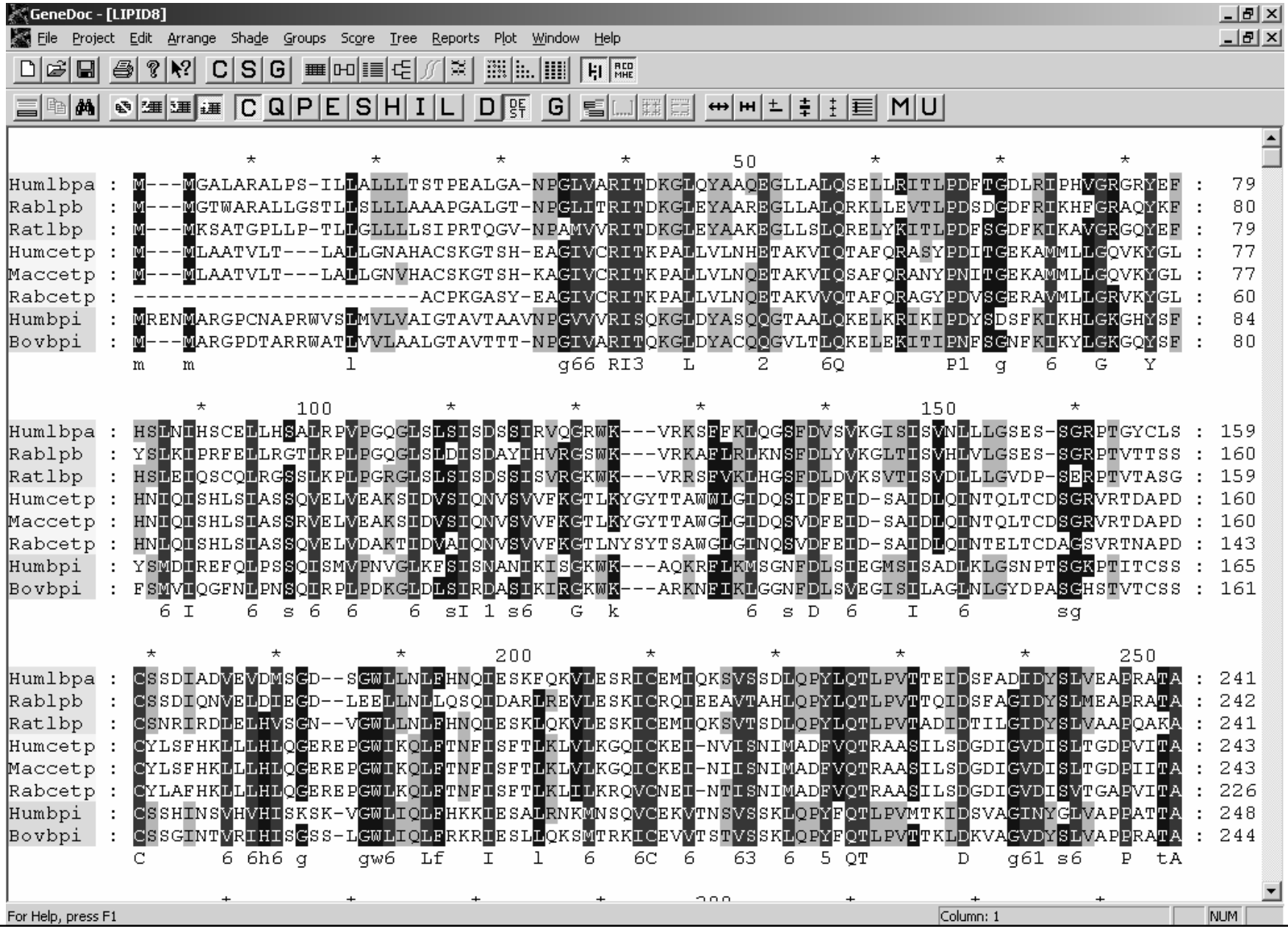
- **CLUSTALW** http://www.ebi.ac.uk/clustalw/

  Match-Box

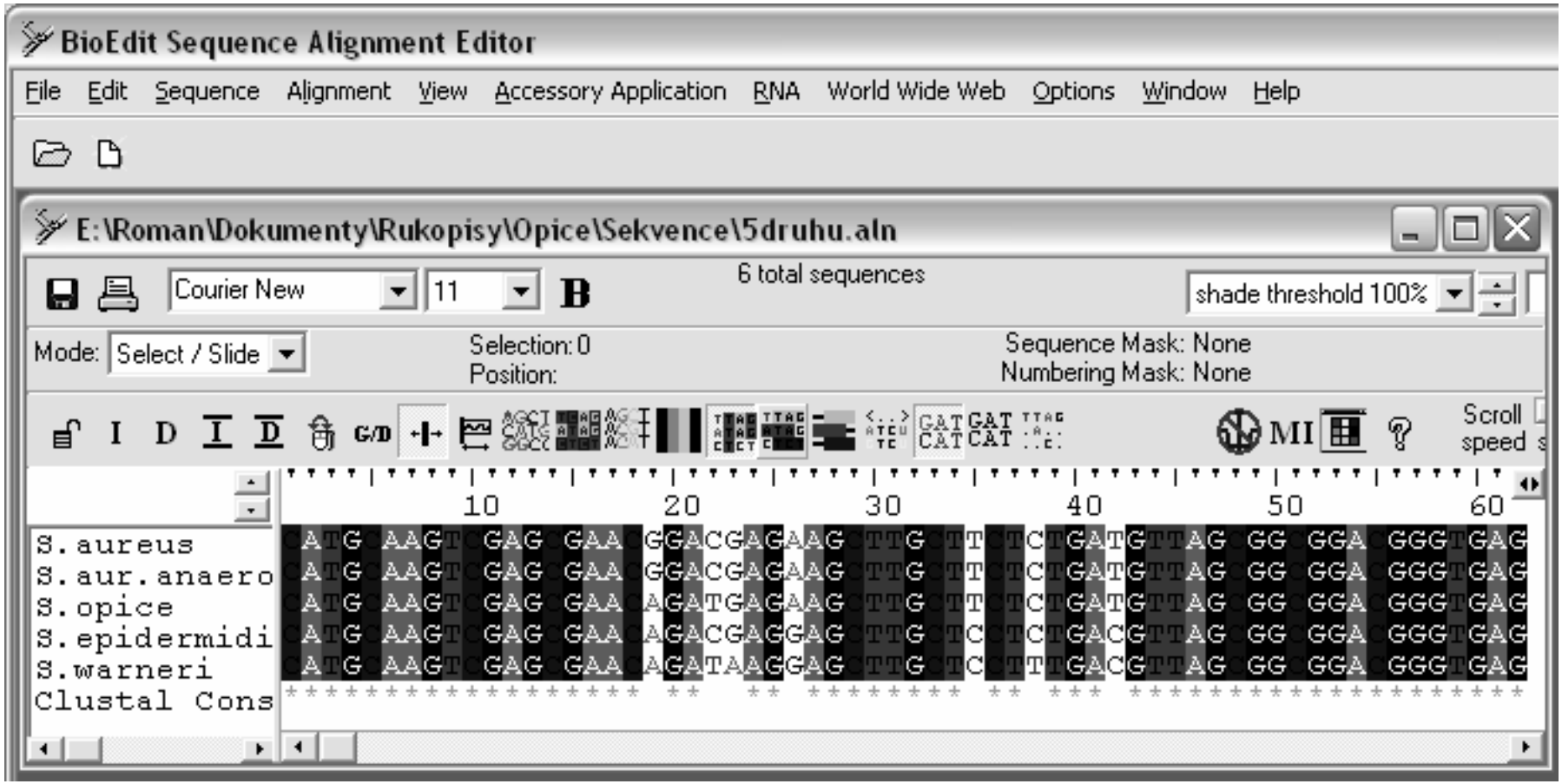
  http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.shtml

- **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

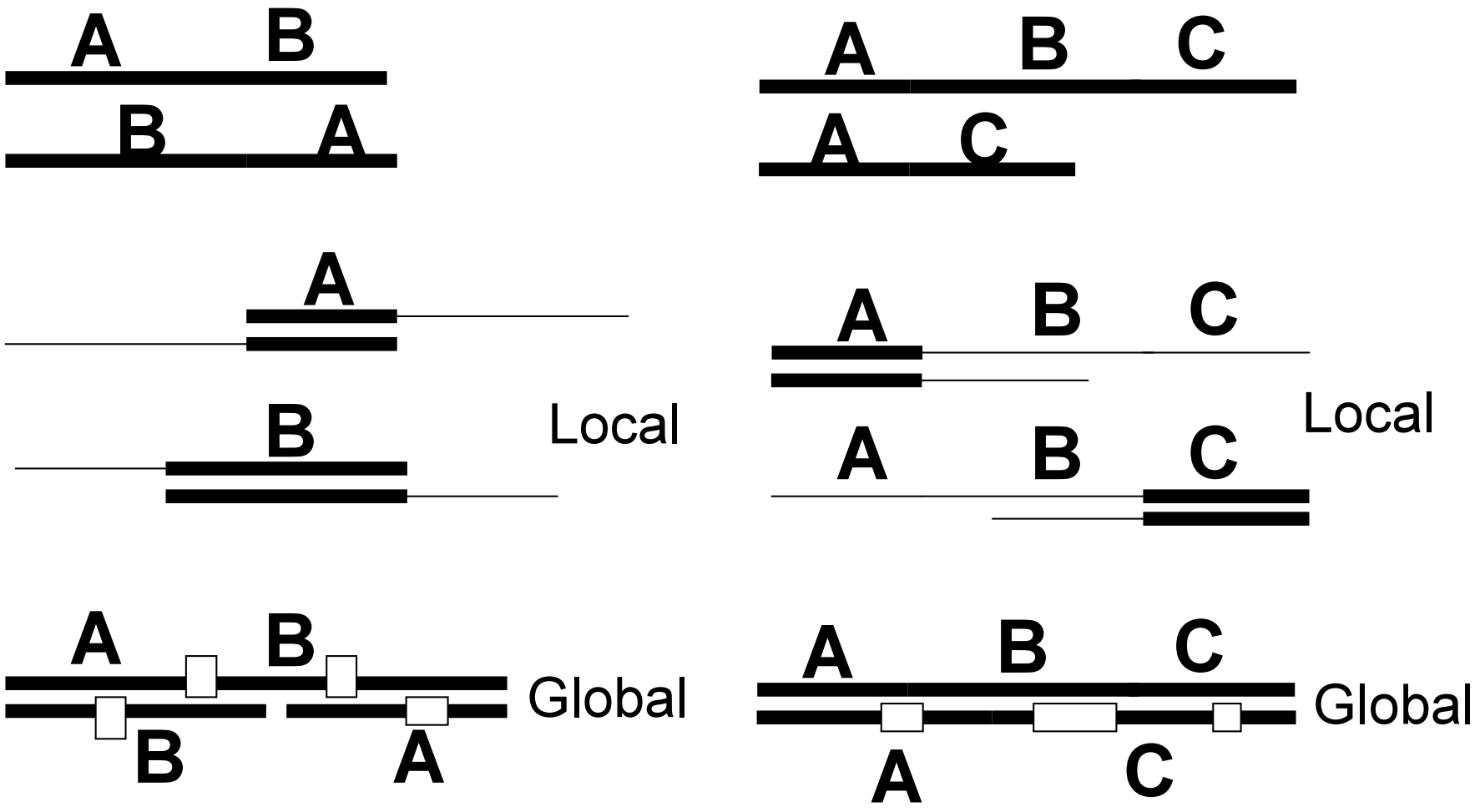


Editory mnohonásobných seřazení: BioEdit (North Carolina State University)



http://www.mbio.ncsu.edu/BioEdit/page2.html

Globalizované lokální seřazení



Formát sekvencí – multi FASTA

```
>S.nepalensis
AATACATGCAAGTCGAGCGAACAGATAAGGAGCTTGCTCCTTTGACGTTAG
CGGCGGACGGGTGAGTAACACGTGGGTAACCTACCTATAAGACTGGAATAACTCCGGGAAACCGGGGCTA
ATGCCGGATAATATTTAGAACCGCATGGTTCTAAAGTGAAAGATGGTTTTGCTATCACTTATAGATGGAC
CCGCGCCGTATTAGCTAGTTGGTGGGGTAATGGCTTACCAAGGCAACGATACGTAGCCGACCTGAGAGGG
TGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCCGCA
ATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTCTTCGGATCGTAAAACTCTGTTATT
AGGGAAGAACAAATGTGTAAGTAACTGTGCACGTCTTGACGGTACCTAATCAGAAAGCCACGGCTAACTA
CGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGTA
GGCGGTYTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGAAACT
TGAGTGCAGAAGANGAAAGTGGAATTCC
>S.cohnii.Lepidoptera
AATACATGCAAGTCGAGCGAACAGATAAGGAGCTTGCTCCTTTGACGTTAGCGGCGGACGGGTGAGTA
ACACGTGGGTAACCTACCTATAAGACTGGAATAACTCCGGGAAACCGGGGCTAATGCCGGATAATATTTA
GAACCGCATGGTTCTAAAGTGAAAGATGGTTTTGCTATCACTTATAGATGGACCCGCGCCGTATTAGCTA
GTTGGTGGGGTAATGGCTCACCAAGGCAACGATACGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGA
ACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGAC
GGAGCAACGCCGCGTGAGTGATGAAGGTCTTCGGATCGTAAAACTCTGTTATTAGGGAAGAACAAATGTG
TAAGTAACTGTGCACGTCTTGACGGTACCTAATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGG
TAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGTAGGCGGTTTCTTAAGTCT
GATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGAAACTTGAGTGCAGAAGAGGAAA
GTGGAATTCC
>S.cohnii.cohnii
AATACATGCAAGTCGAGCGAACAGATAAGGAGCTTGCTCCTTTGAC
GTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTACCTATAAGACTGGAATAACTCCGGGAAACCGG
GGCTAATGCCGGATAACATTTAGAACCGCATGGTTCTAAAGTGAAAGATGGTTTTGCTATCACTTATAGA
TGGACCCGCGCCGTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAACGATACGTAGCCGACCTGA
GAGGGTGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTT
CCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTCTTCGGATCGTAAAACTCTG
TTATTAGGGAAGAACAAATGTGTAAGTAACTATGCACGTCTTGACGGTACCTAATCAGAAAGCCACGGCT
AACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGC
GCGTAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGA
AACTTGAGTGCAGAAGAGGAAAGTGGAATTCC
>S.cohnii.urealyt
AATACATGCAAGTCGAGCGAACAGATAA
GGAGCTTGCTCCTTTGACGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTACCTATAAGACTGGA
ATAACTCCGGGAAACCGGGGCTAATGCCGGATAACATTTAGAACCGCATGGTTCTAAAGTGAAAGATGGT
TTTGCTATCACTTATAGATGGACCCGCGCCGTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAAC
GATACGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAG
GCAGCAGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTCT
TCGGATCGTAAAACTCTGTTATTAGGGAAGAACAAATGTGTAAGTAACTGTGCACGTCTTGACGGTACCT
AATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAA
TTATTGGGCGTAAAGCGCGCGTAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAG
GGTCATTGGAAACTGGGAAACTTGAGTGCAGAAGAGGAAAGTGGAATTCC
>S.xylosus.type
AATACATGCAAGTCGAGCGAACAGATAAGGAGCTTGCTCCTTTGAA
GTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTACCTATAAGACTGGGATAACTTCGGGAAACCGG
AGCTAATACCGGATAACATTTAGAACCGCATGGTTCTAAAGTGAAAGATGGTTTTGCTATCACTTATAGA
TGGACCCGCGCCGTATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCGACGATACGTAGCCGACCTGA
GAGGGTGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTT
CCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGAAGGGTTTCGGCTCGTAAAACTCTG
TTATTAGGGAAGAACAAATGTGTAAGTAACTGTGCACATCTTGACGGTACCTAATCAGAAAGCCACGGCT
AACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGC
GCGTAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGA
AACTTGAGTGCAGAAGAGGAAAGTGGAATTCC
```