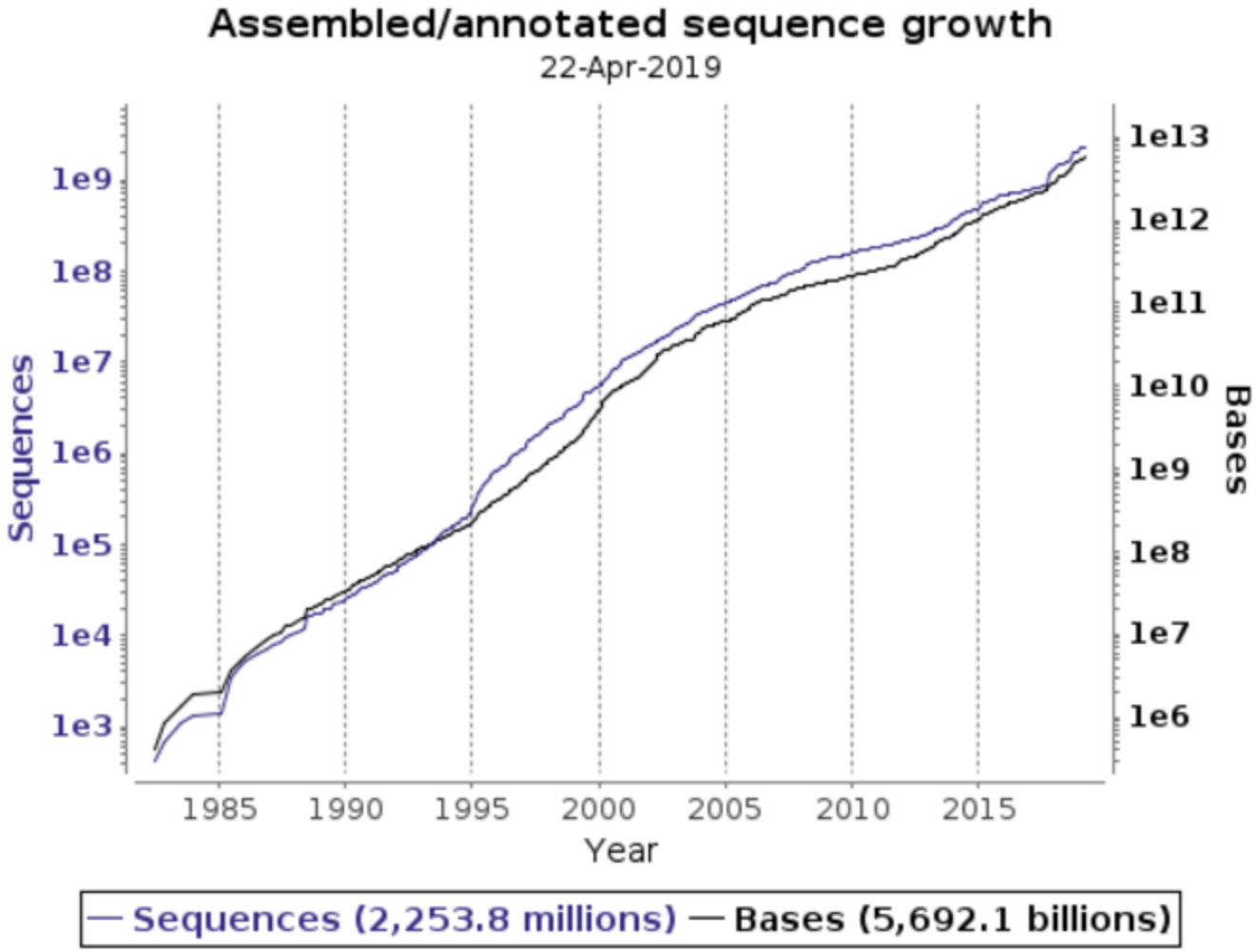
Bioinformatika je disciplína na rozhraní počítačových věd, informačních technologií, matematiky a biologie

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

Trend nárůstu množství dat v bioinformatických databázích

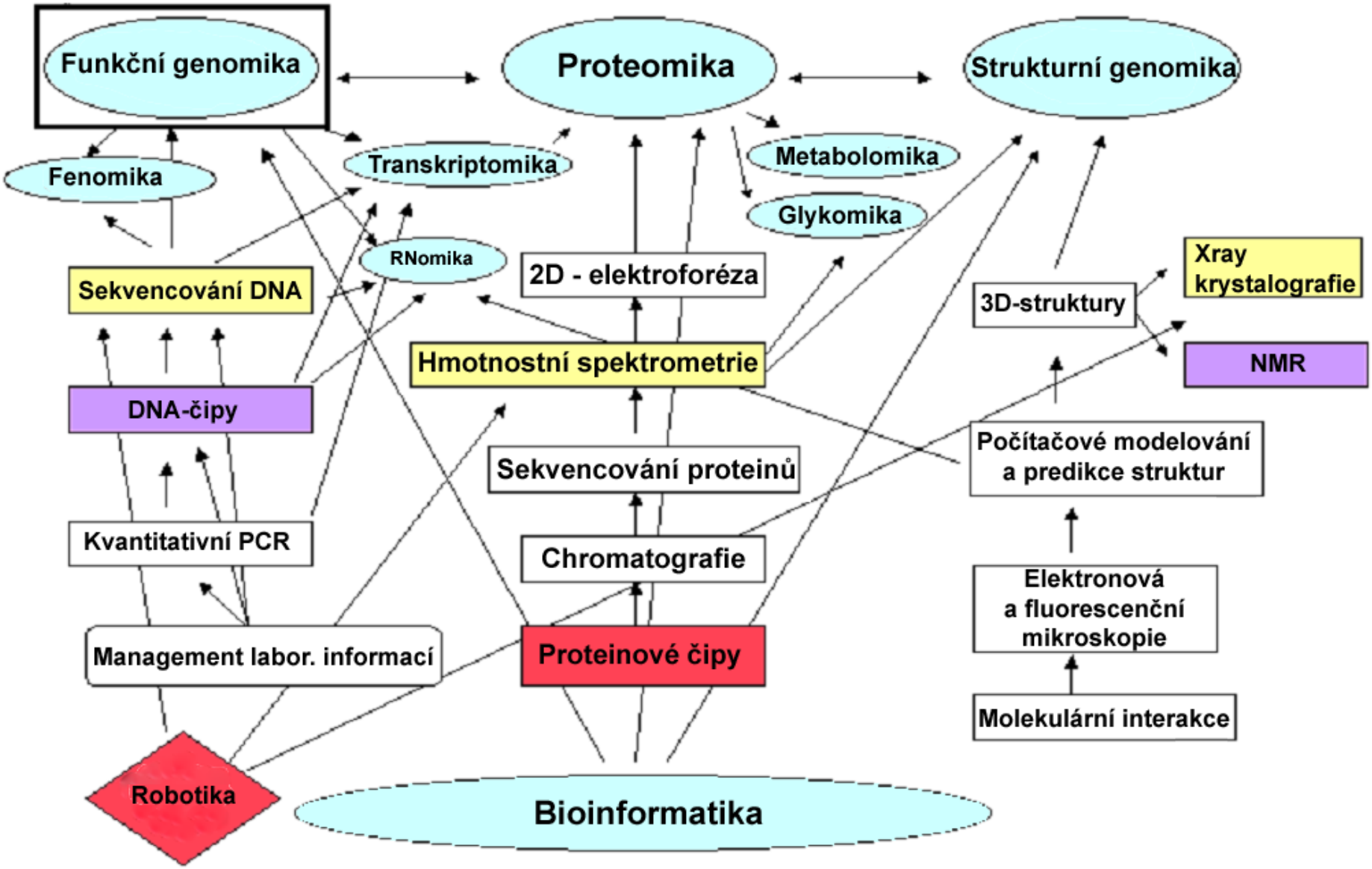
Assembled/annotated sequence growth



Základní zdroje a aplikace bioinformatiky

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

„...omiky“ v molekulární biologii



- 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

Nejdůležitější instituce zabývající se shromažďováním biomedicínských informací

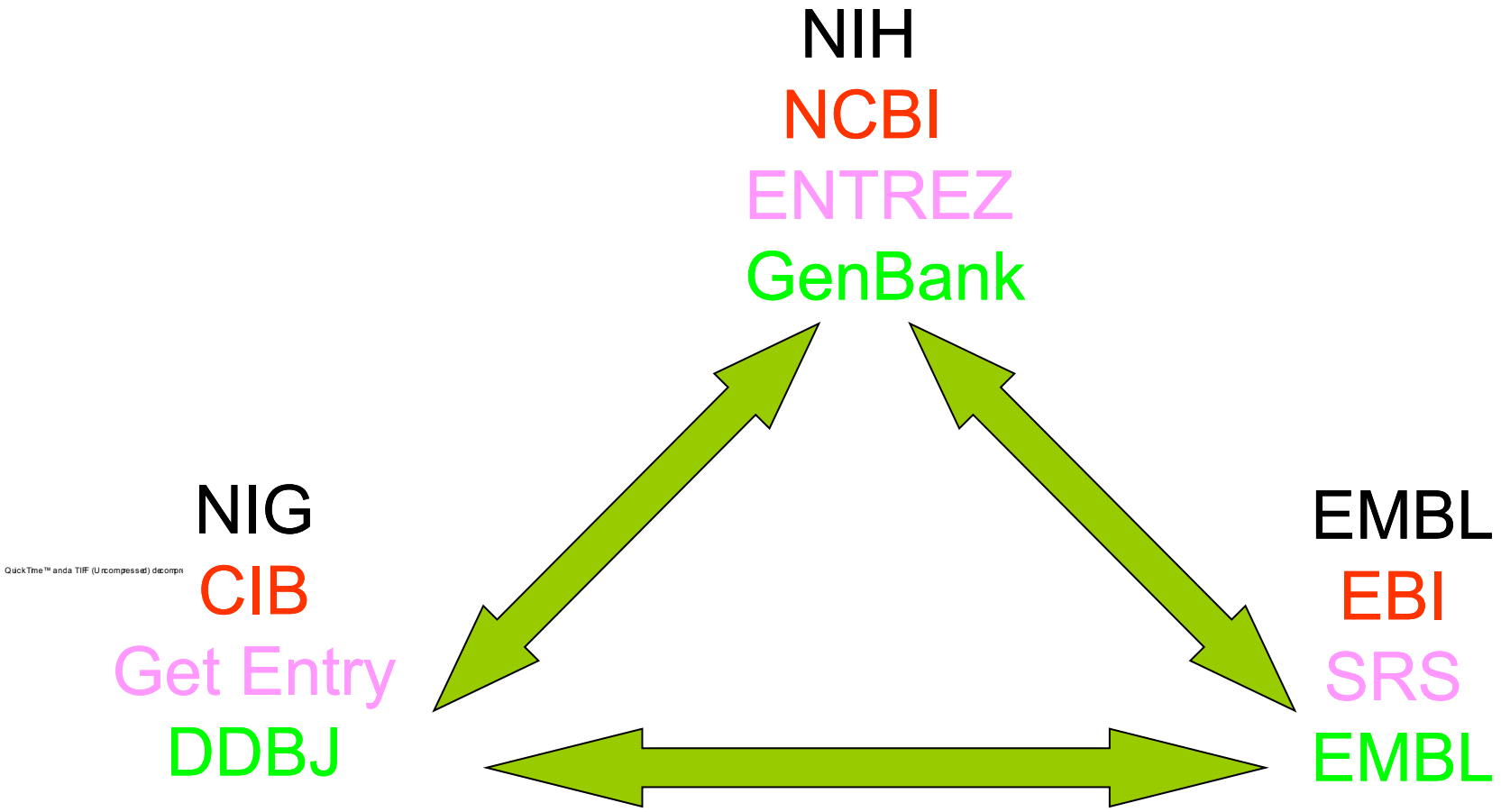
- V současné době je prostřednictvím Internetu dostupných přibližně 550 databází zabývajících se shromažďováním bioinformací.
  - Jejich přehled a popis je každoročně publikován ve specializovaném, volně dostupném čísle časopisu Nucleic Acids Research.
- K nejdůležitějším institucím zabývajícím se, správou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

Nejdůležitější databáze sekvencí nukleových kyselin a proteinů

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

Mezinárodní spolupráce sekvenčních databází

- Databáze sdílejí stejná data



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

Jak se data dostanou do databází?

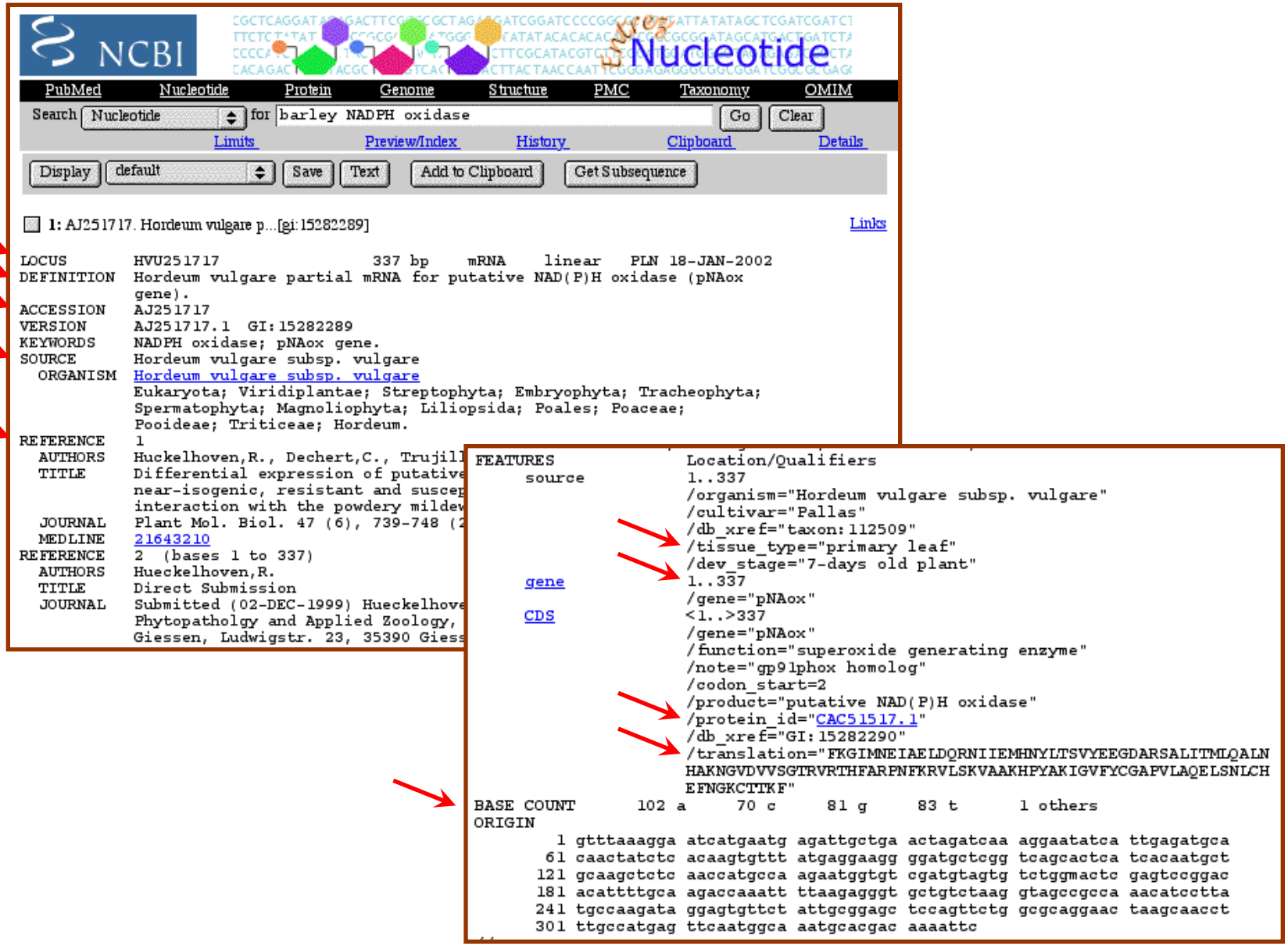
- Předání dat prostřednictvím WWW portálu
  - BankIt (GenBank)
    - http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank
  - Submission Portal
    - https://submit.ncbi.nlm.nih.gov/
  - WebIn (EMBL/European Nucleotide Archive)
    - http://www.ebi.ac.uk/ena/submit
  - Sakura (DDBJ)
    - http://www.ddbj.nig.ac.jp/sub/websub-e.html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

Identifikace záznamu v primárních sekvenčních databázích

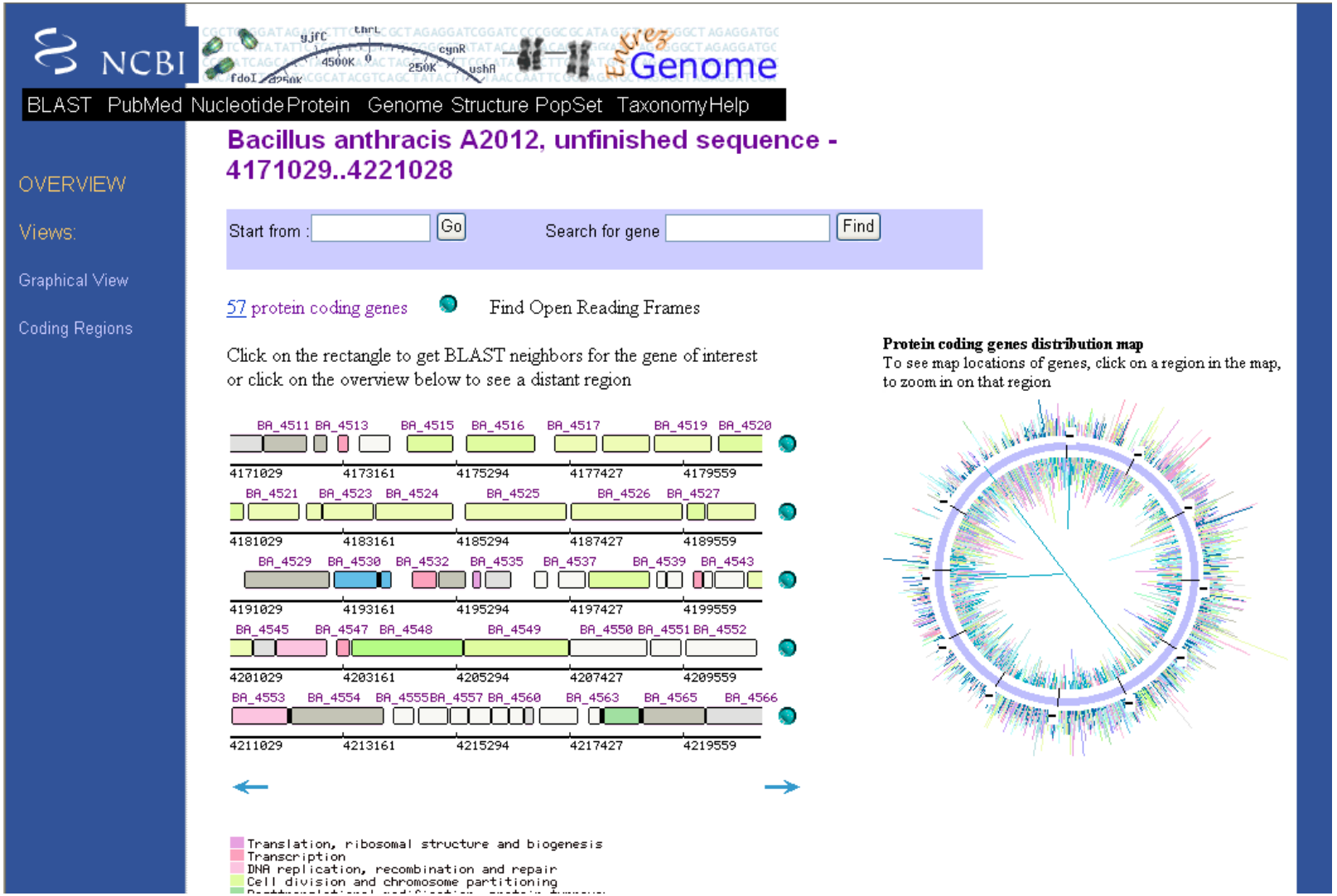
- GenBank
- EMBL-Bank (European Nucleotide Archive, ENA)
- DDBJ
- **Přístupový kód (Accession Number)**
- **číslo GI (GenBank Identifier)**

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

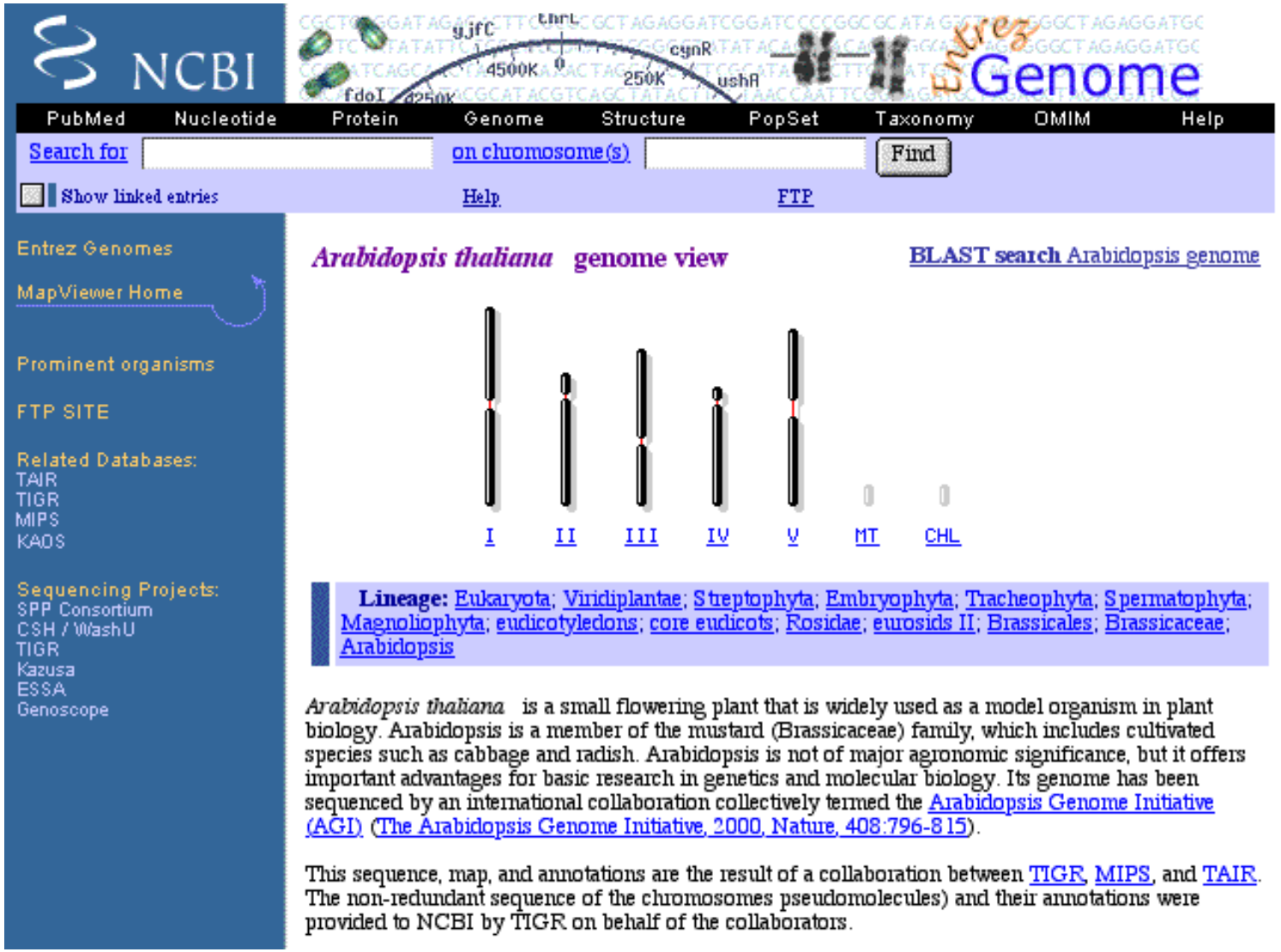
- Struktura zápisu sekvence ve formátu GenBank
- http://www.ncbi.nlm.nih.gov/Genbank/



Genomové databáze v NCBI – prokaryota

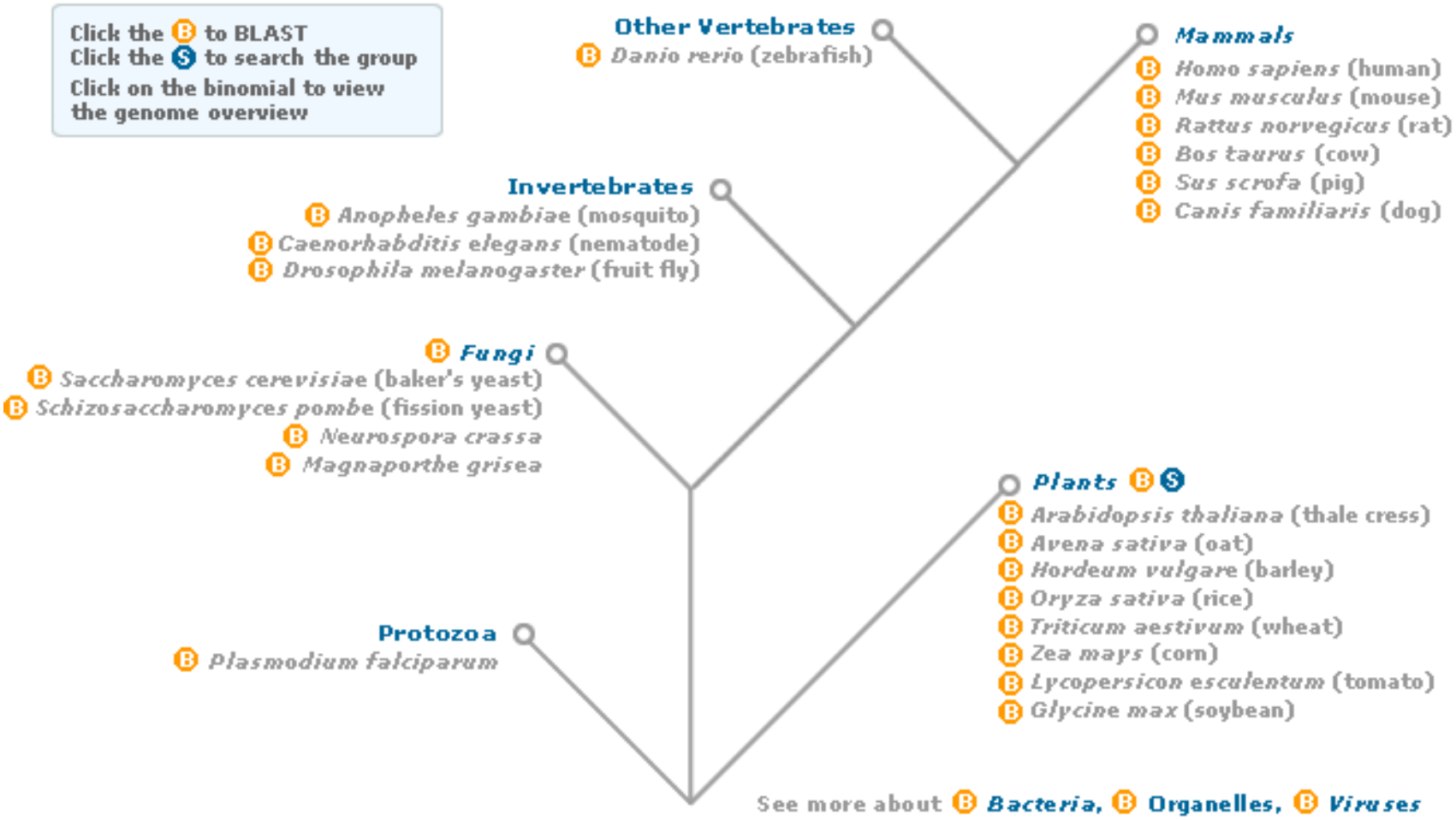


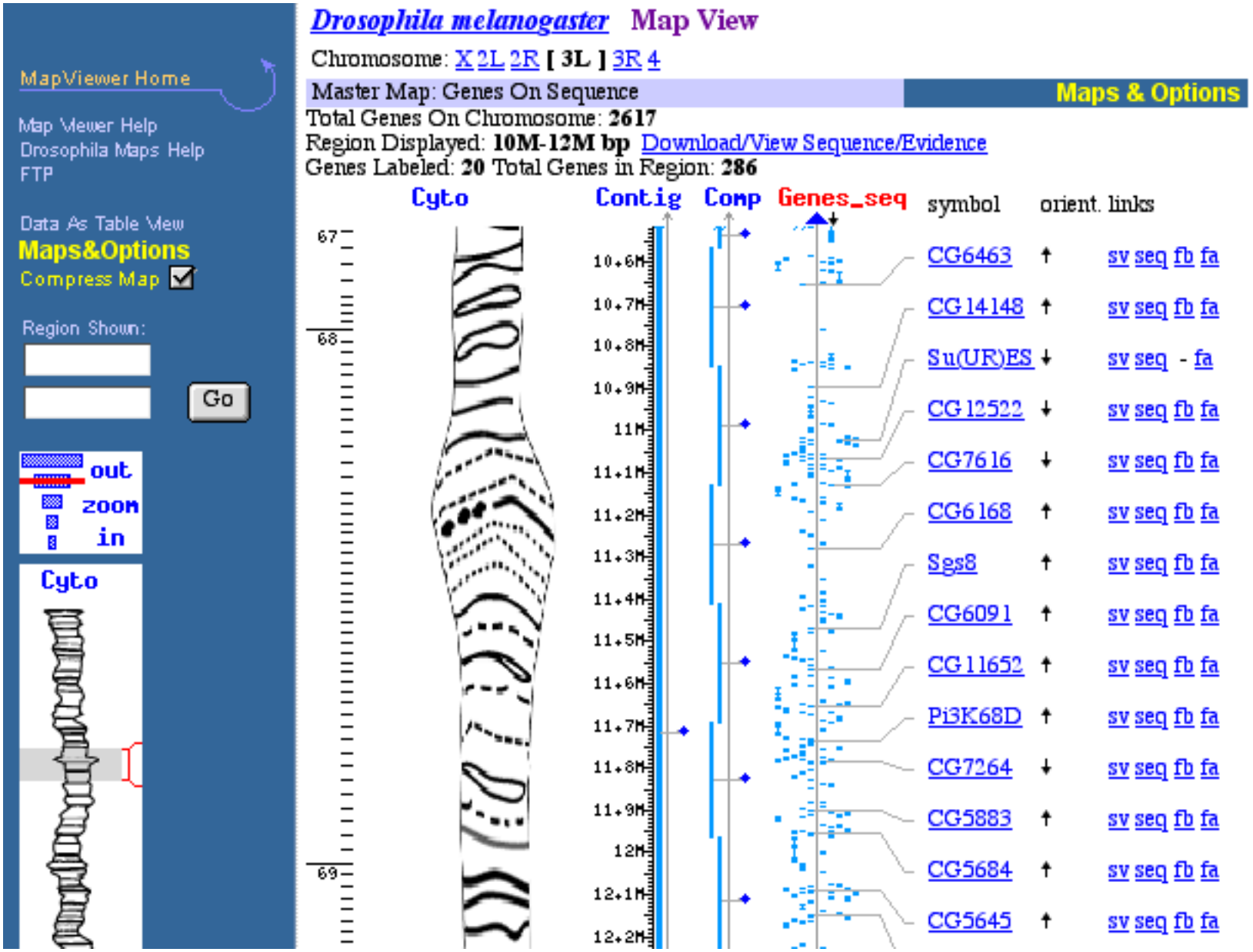
Genomové databáze v NCBI - eukaryota



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
Ses8 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

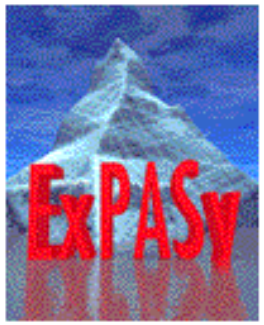
Databáze sekvencí proteinů

- 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).

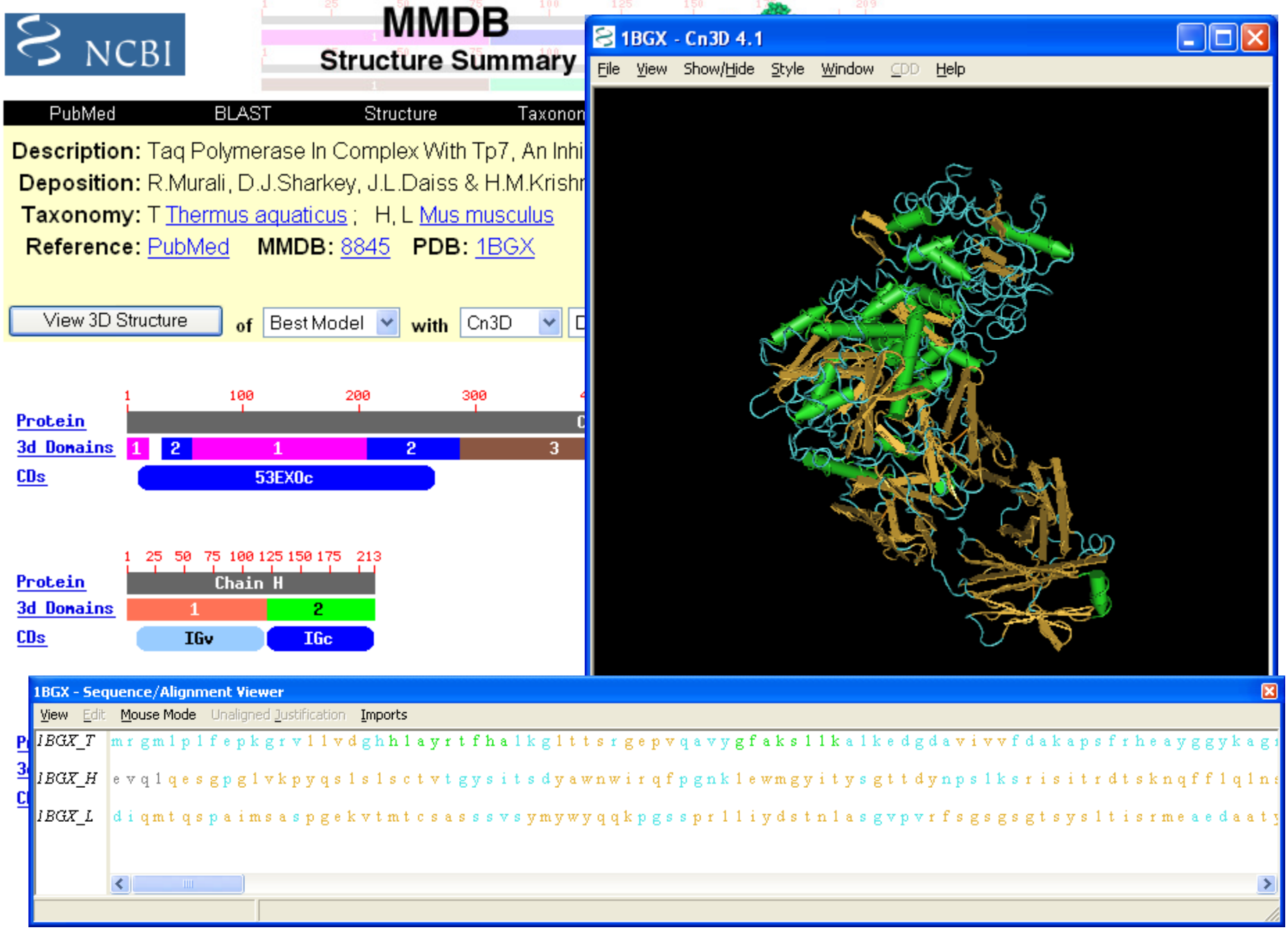
[Announcements] [Job opening] [Mirror Sites]

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
Taxonon
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

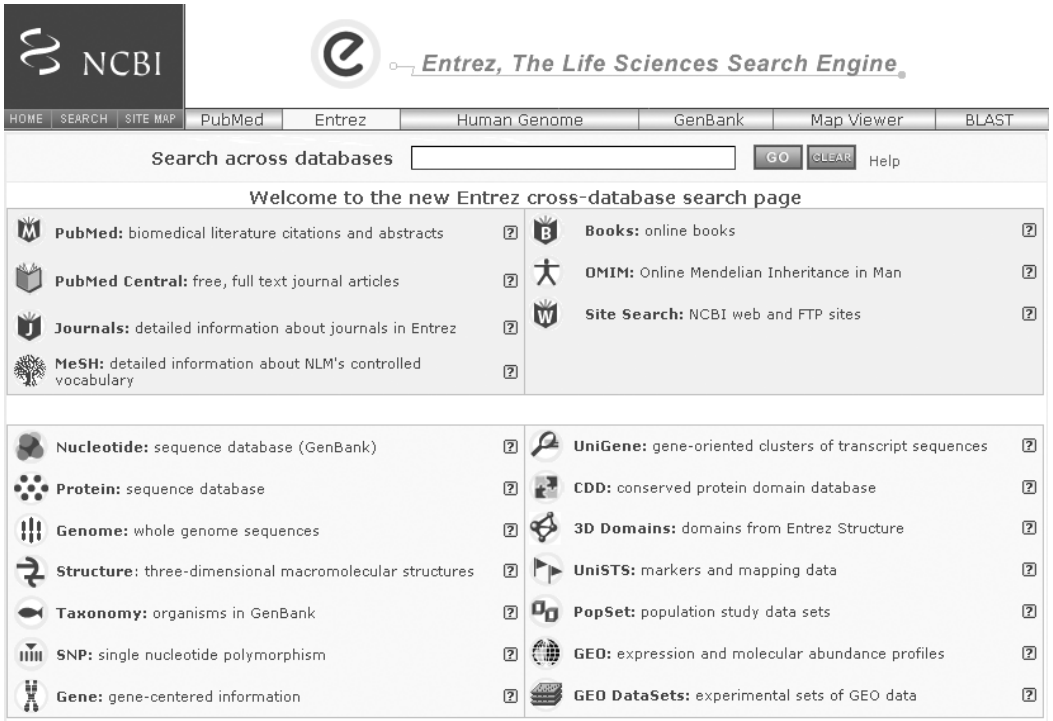
Textové vyhledávání v databázích

- 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/

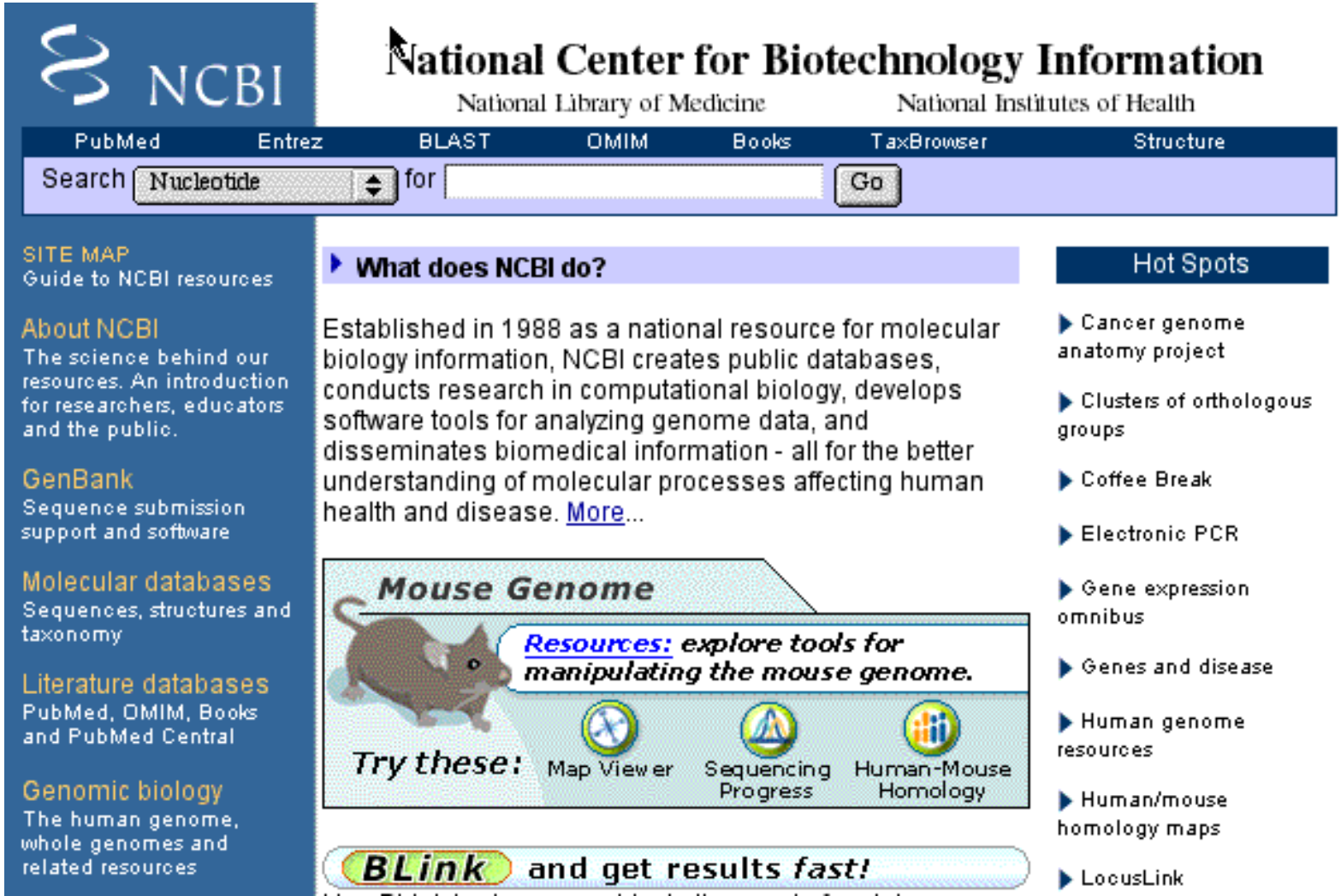
- **Entrez** je vyhledávací systém pro molekulárně biologické databáze vyvinutý v NCBI
- Je vstupním bodem pro průzkum 45 různých integrovaných databází z nichž řada je virtuálních.

- 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



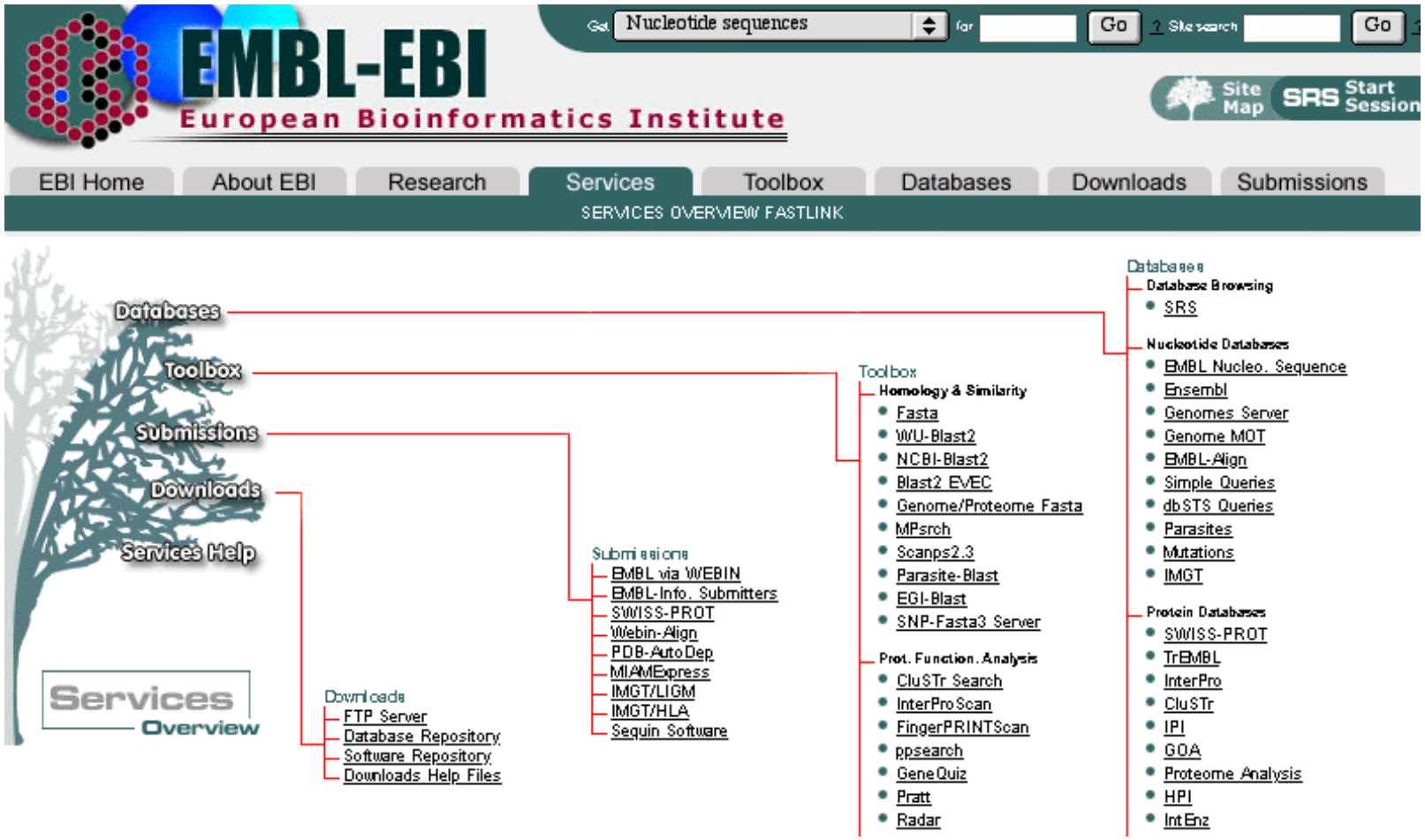
Entrez Molecular Sequence Database System

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

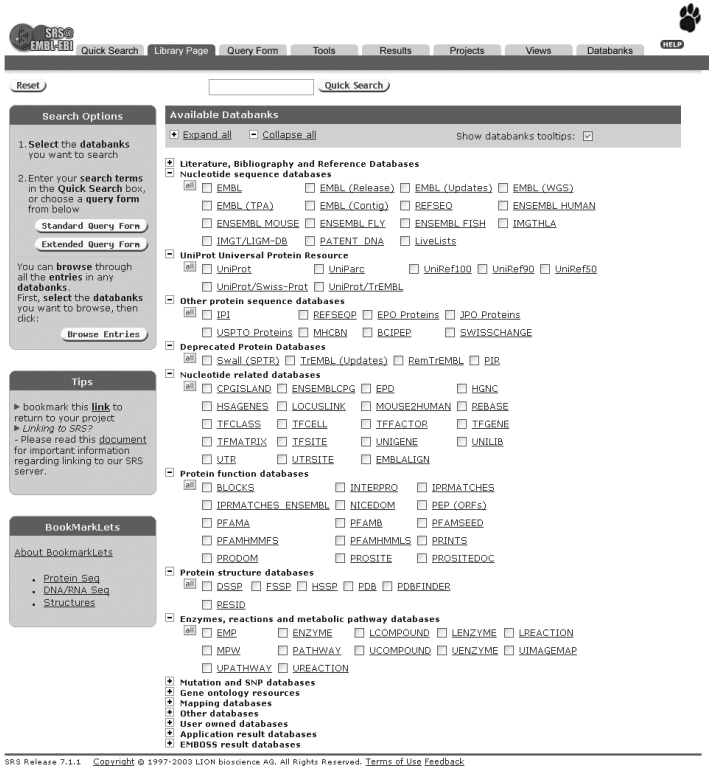


Sequence Retrieval System (SRS)

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



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



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

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.

Posuzování podobnosti sekvencí

Nástroje pro vyhledávání lokálních podobností sekvencí

Postup stanovení podobnosti

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

Nástroje pro vyhledávání lokálních podobností sekvencí

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

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

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

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

Co je to BLAST?

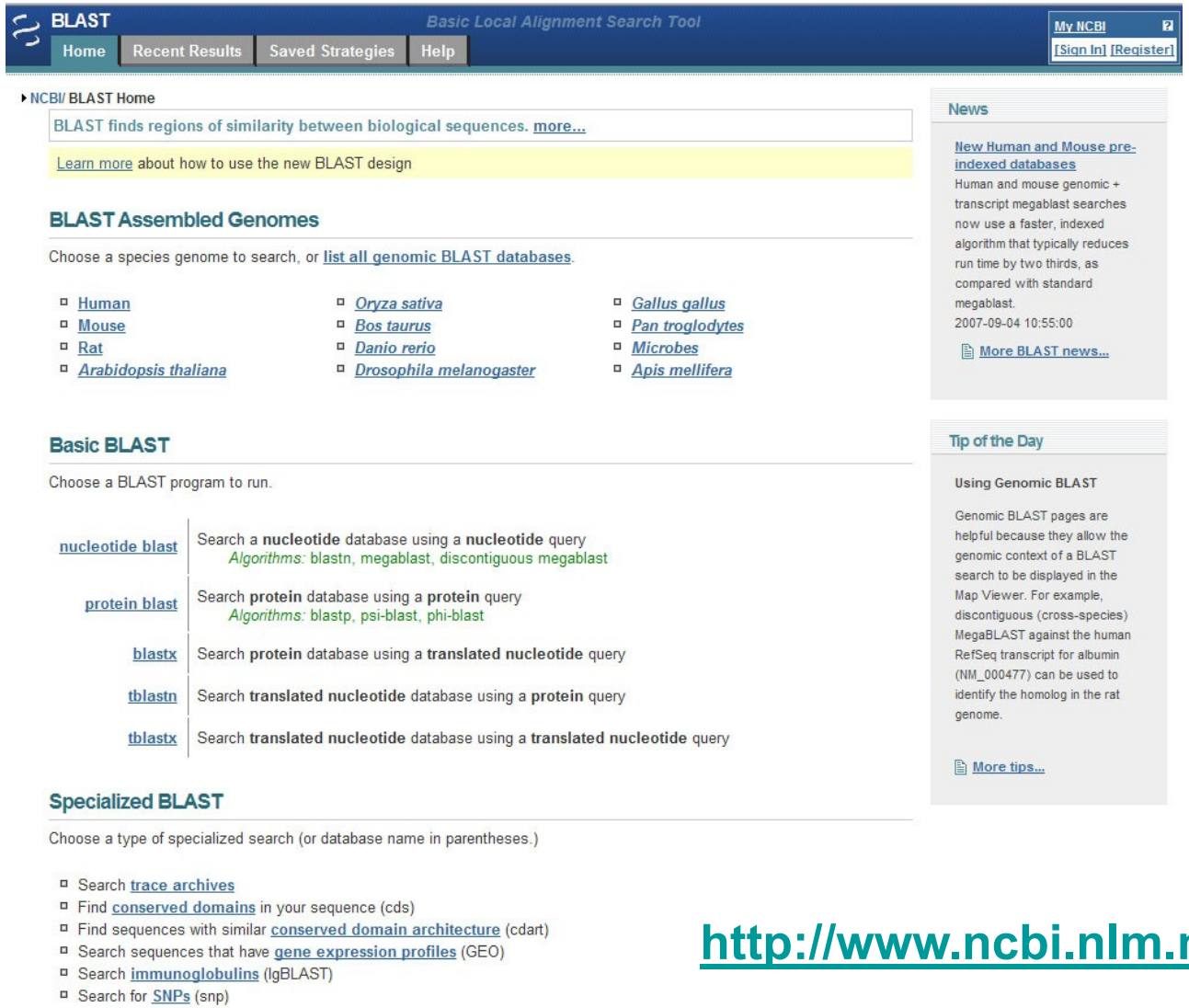
- **B**asic **L**ocal **A**lignment **S**earch **T**ool
  - Hledání lokálních podobností
  - Heuristický přístup založený na Smith-Watermanově algoritmu
  - Vyhledá nejoptimálnější **seřazení sekvencí**
  - Poskytuje data o statistické významnosti
  - Zobrazuje vzájemně seřazené sekvence
  - Lokalizuje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

Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W20-5.

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

McGinnis S, Madden TL.

Výchozí stránka BLAST



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

Basic BLAST – výběr programů

Využití jednotlivých programů BLAST

Program	Dotaz	Databáze	Úroveň srovnání	Použití
blastn	DNA	DNA	DNA	Hledání identických sekvencí DNA
blasp	Protein	Protein	Protein	Hledání 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

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

Substituční Matice

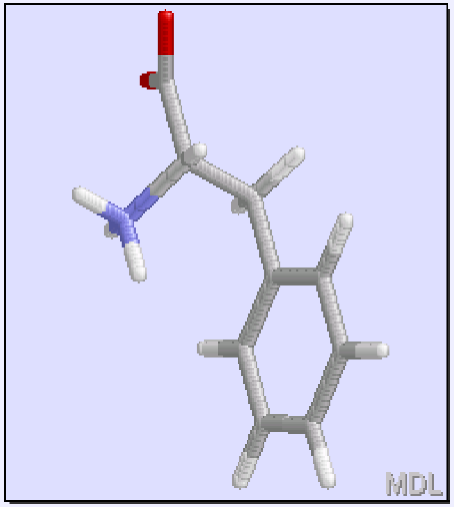
- Co je substituční matice?
  - Kompletní sada skóre pro všechny kombinace párů zbytků se nazývá substituční matice
  - Uplatňuje se při srovnání sekvencí proteinů
  - Stanovuje frekvenci při které každý možný zbytek v sekvencích může být změněn za kterýkoli jiný zbytek během času (evoluce)
  - Např.,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

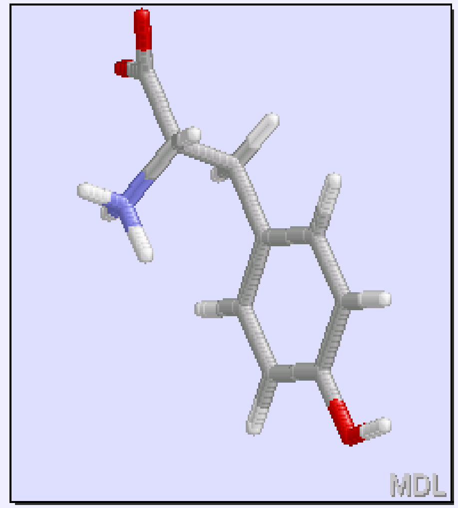
Substituční Matice

- 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





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	-2	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

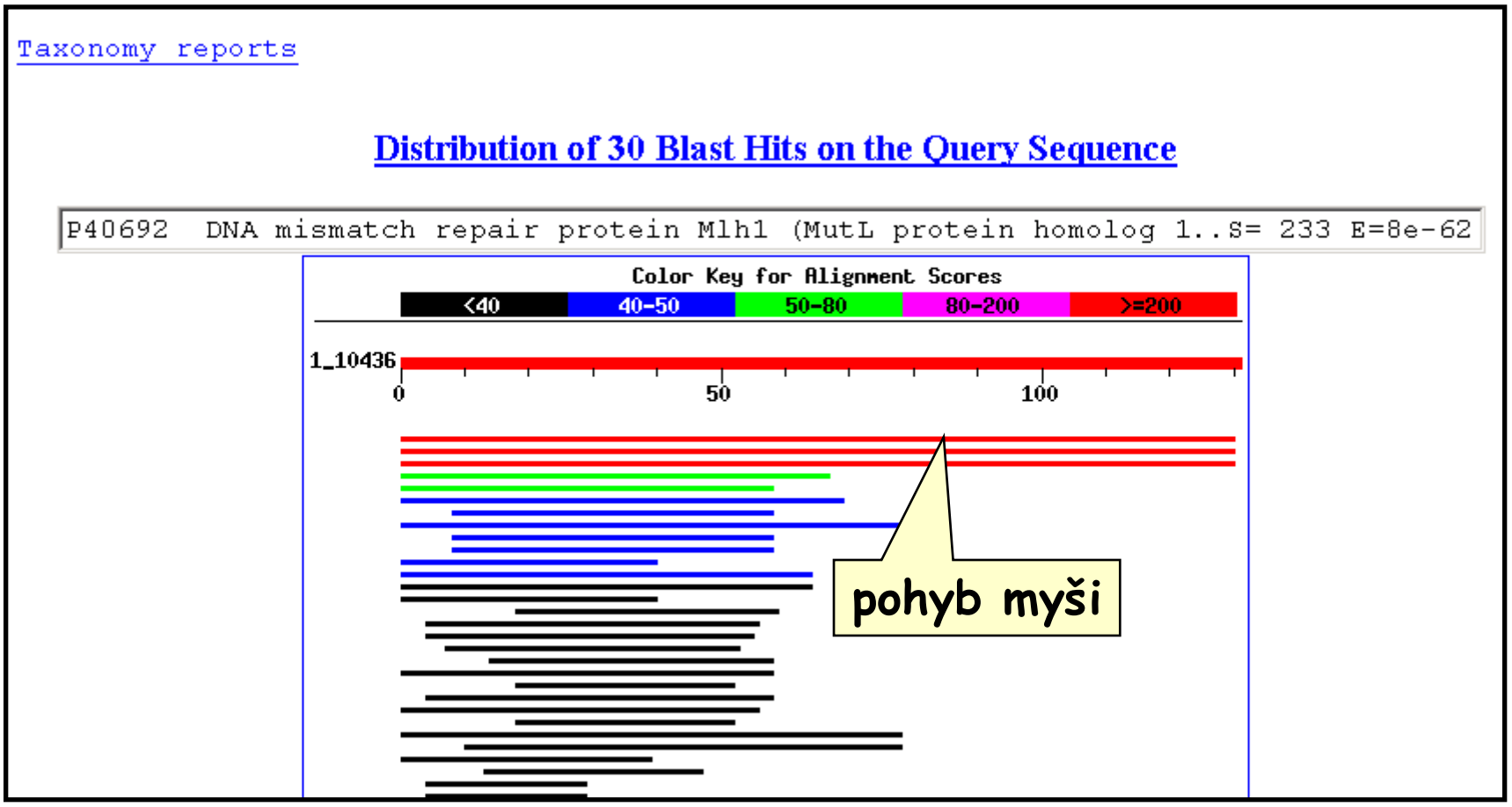
Běžné ar… …ší významnost

Vzácn… …šší významnost

Negativní pro málo pravděpodobné substituce

Pozitivní pro více pravděpodobné substituce

BLAST – grafický výstup



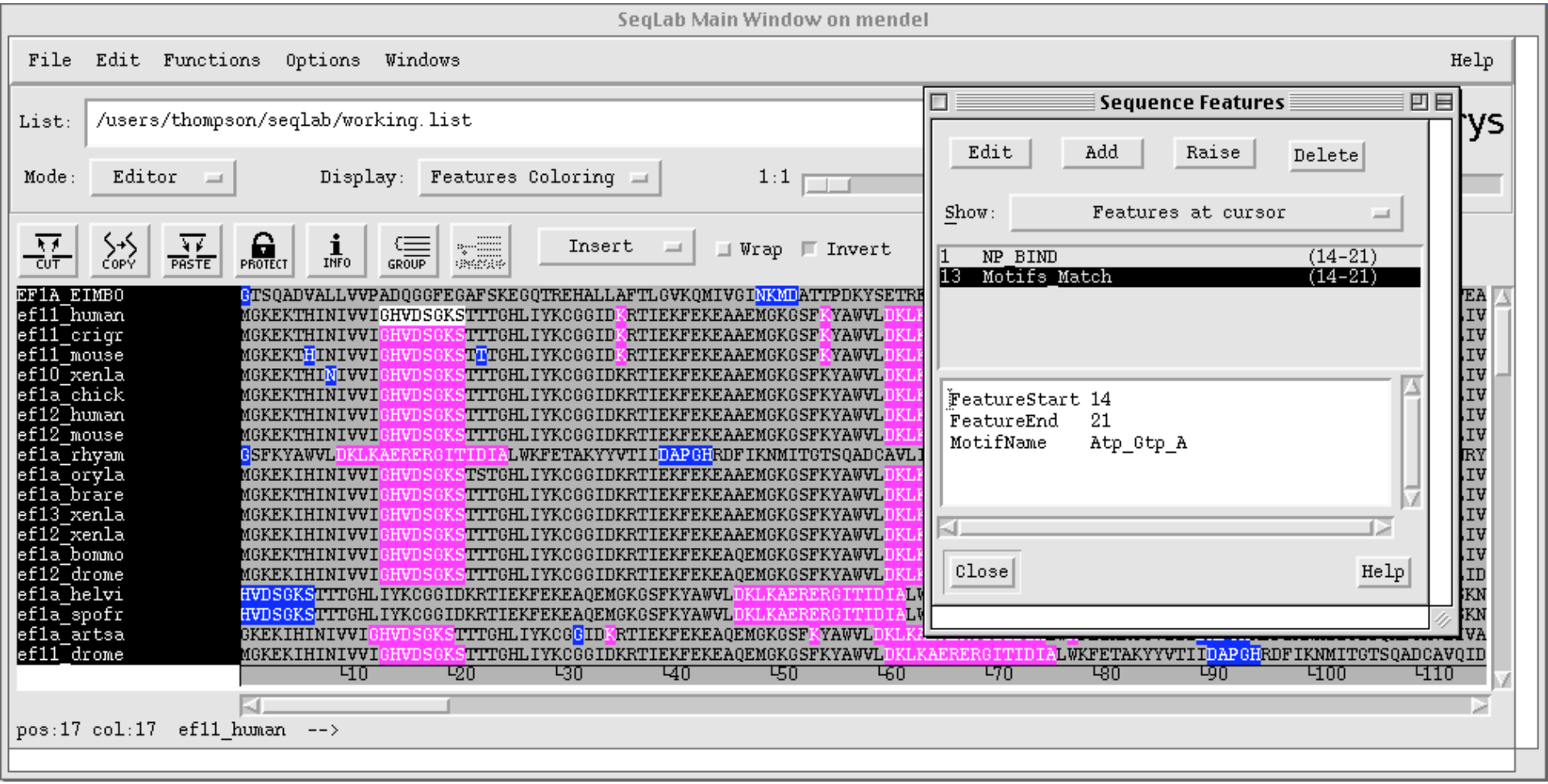
BLAST – příklad výstupu u DNA


```
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
```

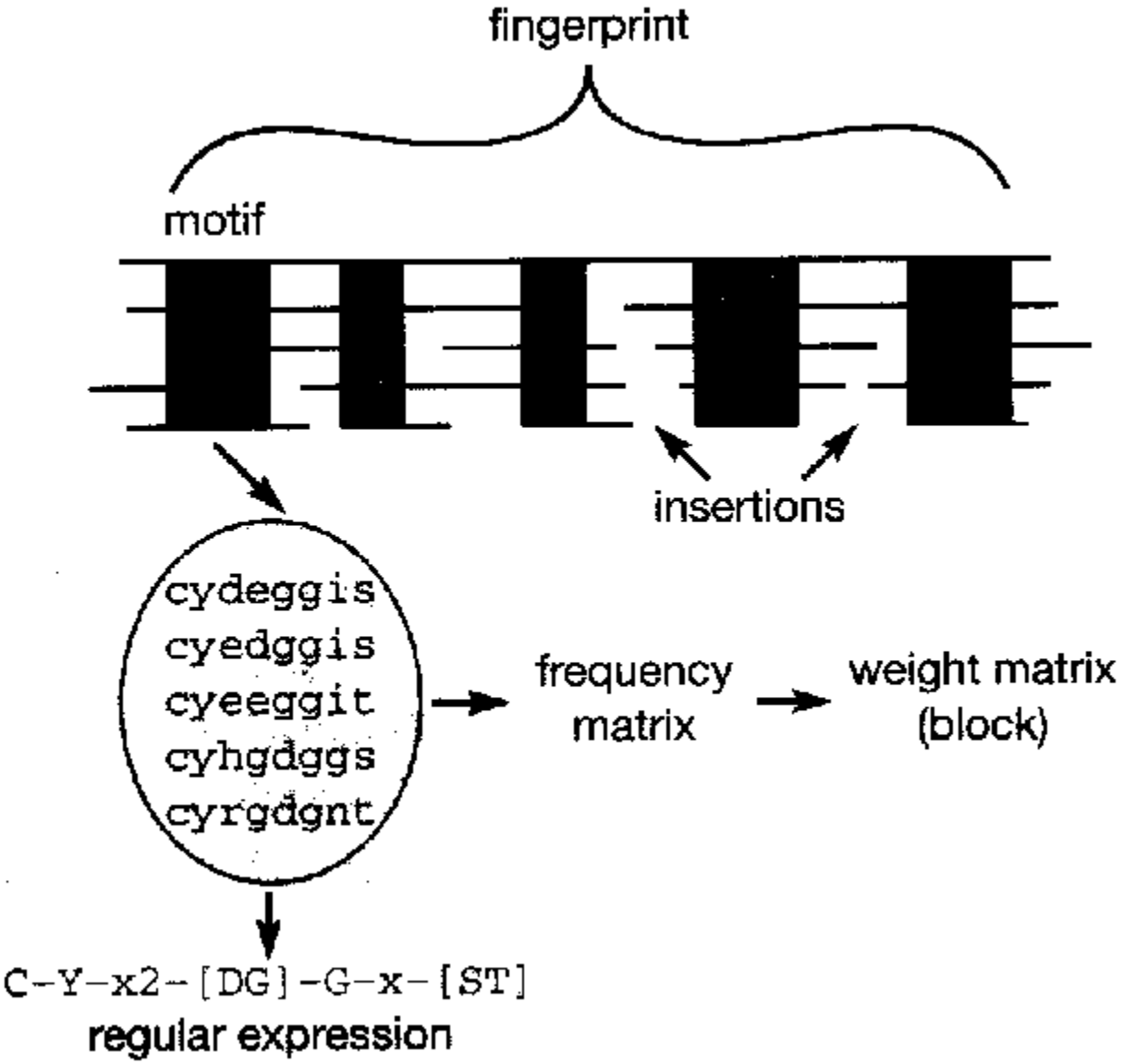
Lokální versus mnohonásobné srovnání

- Dosud jsme srovnávali pouze **dvě sekvence navzájem**
- Podobnosti mezi dvěma sekvencemi se stávají významnými, pokud se vyskytují i u dalších sekvencí
- Mnohonásobné přiložení sekvencí je srovnání tří a více sekvencí nukleových kyselin nebo proteinů s mezerami vloženými do sekvencí tak, že úseky sekvencí s úplnou nebo částečnou homologií jsou seřazeny nad sebou ve stejném sloupci
- Může identifikovat podobnosti a identifikovat **konzervativní motivy**, které nejsme schopni identifikovat lokálním srovnáním

Příklad analýzy mnohonásobného přiložení



Identifikace konzervativních motivů



Klasifikační databáze proteinů

- PROSITE
- Pfam
- PRINTS
- ProDom
- SMART
- Blocks

- InterPro

- Databáze sekvenčních motivů představují značně roztříštěný soubor zdrojů
  - Asi 30 databází
- Částečně se překrývají, ale nejsou navzájem propojeny

- Integrované vyhledávání ve více databázích umožňuje např. InterPro Scan

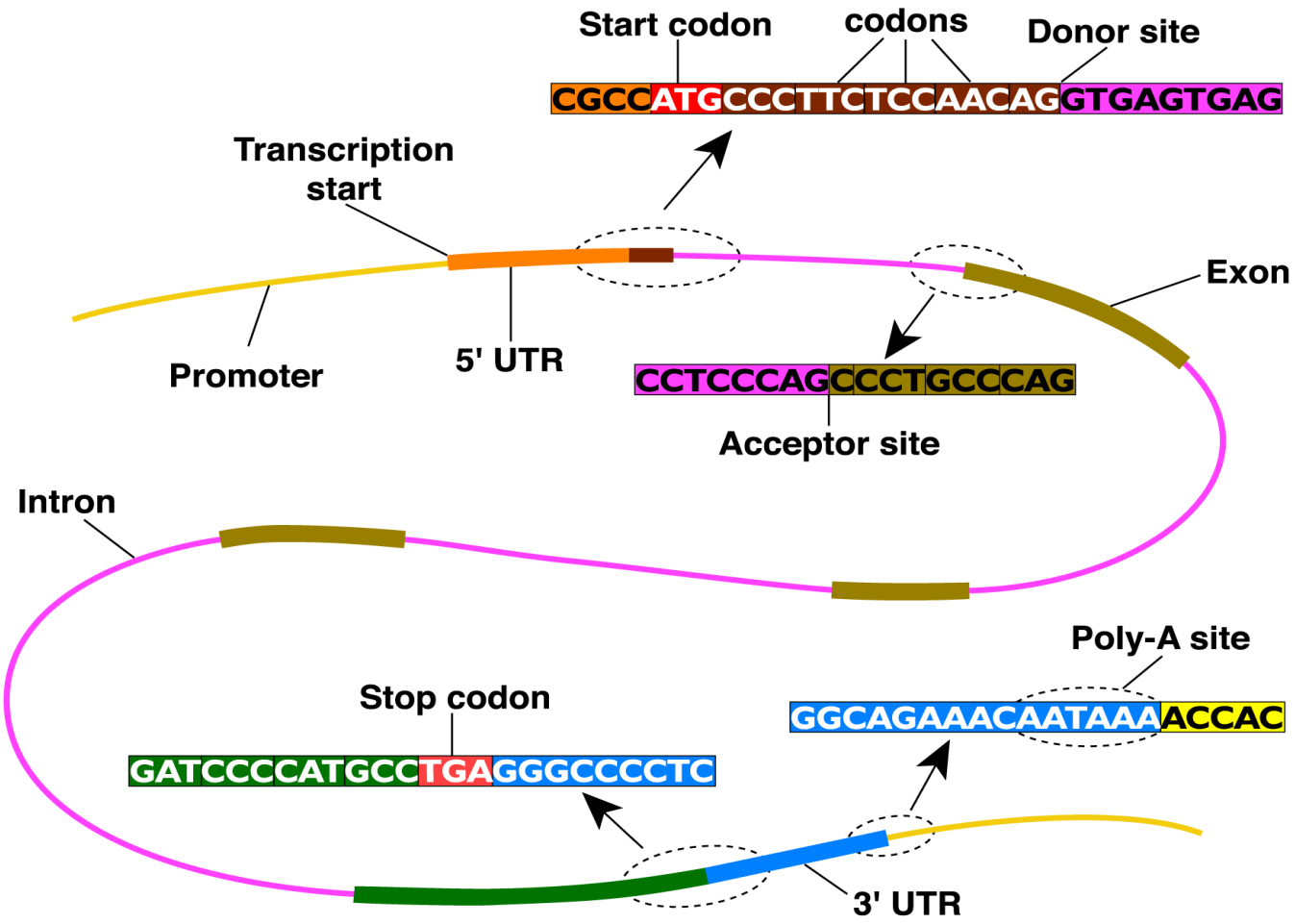
Hledání genů

- Geny tvoří **obsahovou složku** genomu
  - Variabilní délka
  - Jedinečné sekvence
  - Mnohdy složené z exonů a intronů
  - Geny pro funkční RNA

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

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

Signály – senzory ve struktuře eukaryotického genu



Prokaryotický versus eukaryotický gen **vyžadují odlišné přístupy**

- Prokaryota
  - malé genomy 0.5 – $10 \cdot 10^6$ bp
  - Vysoká hustota kódujících sekvencí (>90%)
  - Žádné introny (vyjímky Archea, fágy)
  - hledání otevřených čtecích rámců
  - doplněno např. hledáním signálů pro vazebná místa ribozómu
  - Úspěšnost cca 99 %
  - Problémy: překrývající se ORFs, krátké geny, místa TSS a promotory
- Eukaryota
  - Velké genomy 10^7 – 10^{10} bp
  - Nízká hustota kódujících sekvencí (<50%)
  - Struktura intron/exon
  - statistické modely frekvencí nukleotidů
  - sledování závislostí přítomných ve struktuře kodonů
  - Obsah GC
  - Přesnost dosahuje cca 50 %
  - Problémy: mnoho!