

LOSCHMIDT
LABORATORIES



Úvod do bioinformatiky a bioinformatické databáze



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

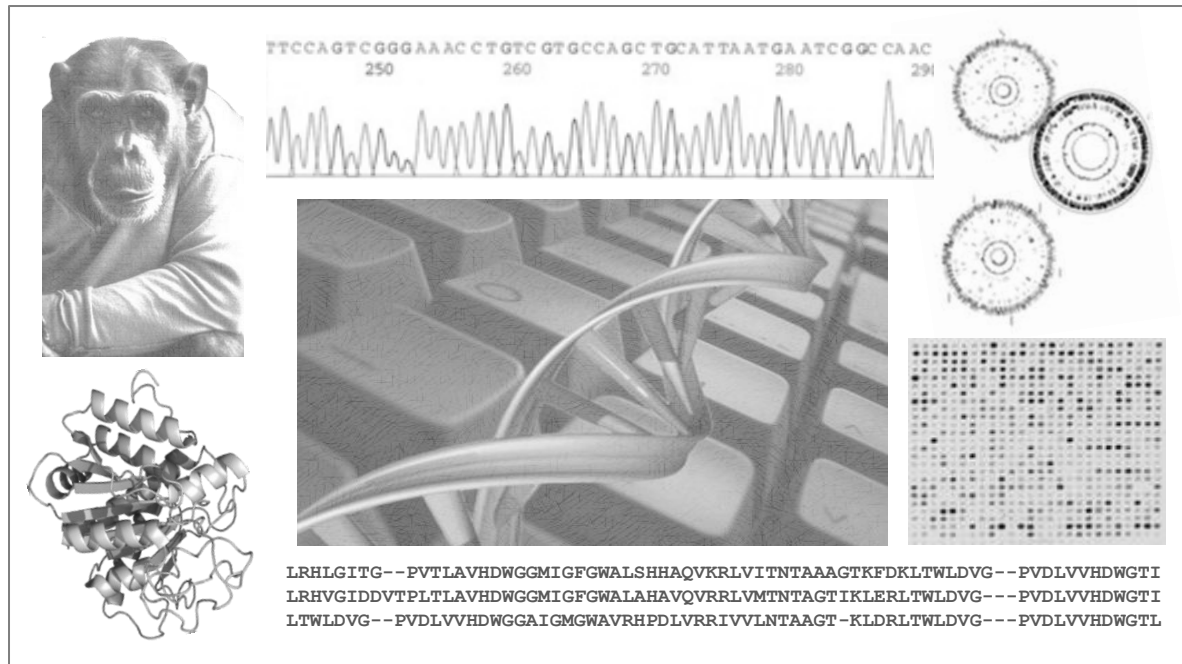
- ❑ Organizační informace
- ❑ Studijní literatura
- ❑ Bioinformatika
- ❑ Historie bioinformatiky
- ❑ Bioinformatické instituce
- ❑ Bioinformatické databáze
- ❑ Prohledávání databází

- Kolokvium
 - Písemný test
 - Celkem 25 otázek s jednou i více správnými odpověďmi
 - Minimálně 17 správných odpovědí

- Bi5000 Bioinformatika I – nukleové kyseliny
- Bi9060 Bioinformatika II – proteiny
- Bi9061 Bioinformatika – cvičení

- ❑ Xiong, J. **Essential Bioinformatics**. Cambridge University Press, New York, 2006.
- ❑ Claverie, J., and Notredame, C. **Bioinformatics for Dummies** ^{2 ed.}. Wiley Publishing, Hoboken, 2006
- ❑ Cvrčková, F. **Úvod do praktické bioinformatiky**. Academia, Praha 2006.
- ❑ Misener, S., Krawetz S.A. **Bioinformatics: methods and protocols**. Humana Press, Totowa, New Jersey 2000.
- ❑ Attwood, T.K., Parry-Smith, D.J. **Introduction to bioinformatics**. Longman, Essex, 1999.
- ❑ Baxevanis, A.D., Ouellette, F.B.F. **Bioinformatics: a practical guide to the analysis of genes and proteins**. Wiley-Interscience, New York 1998.

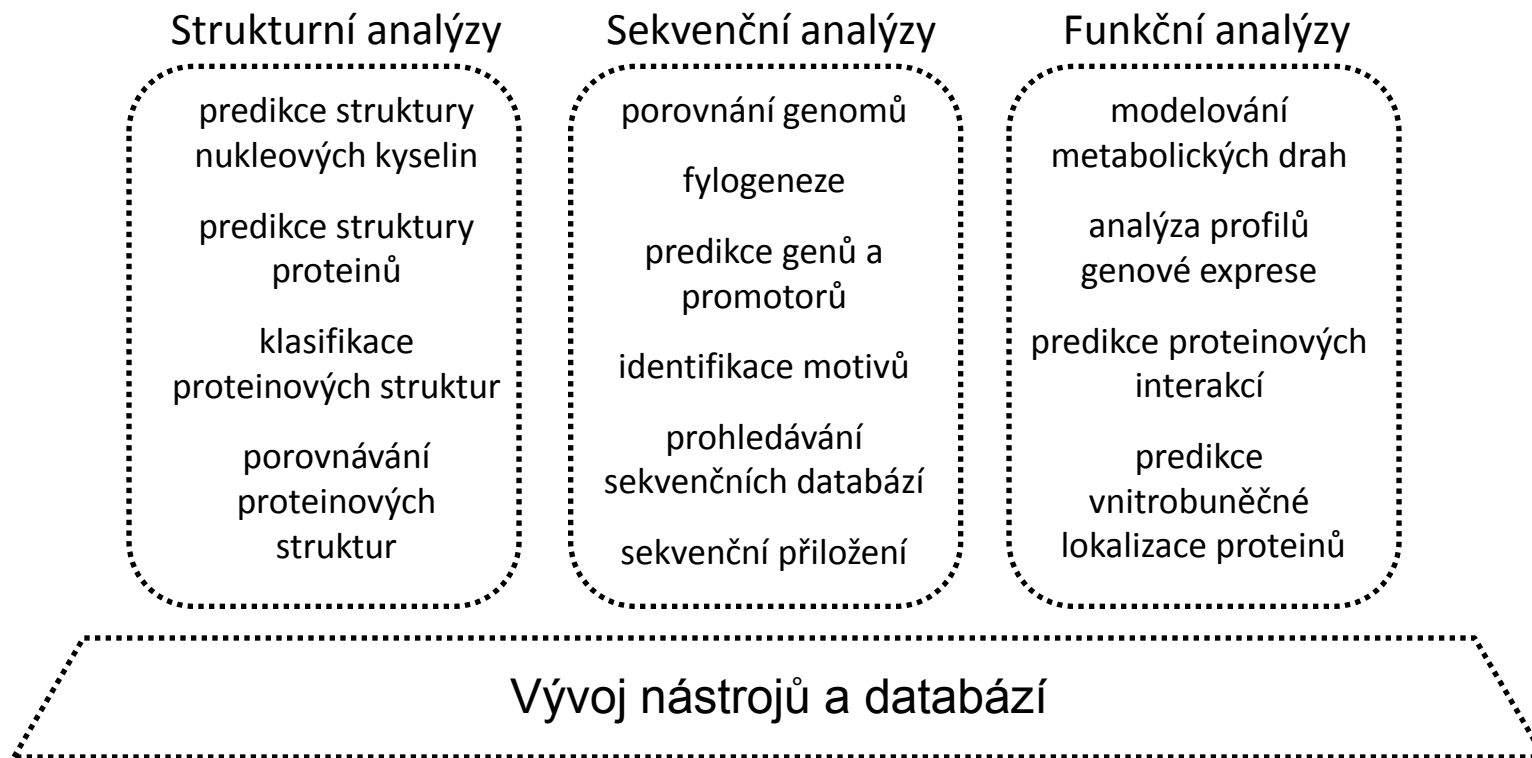
- Informační technologie sloužící k **získávání, uchování, analýze a distribuci informací** týkajících se biomakromolekul



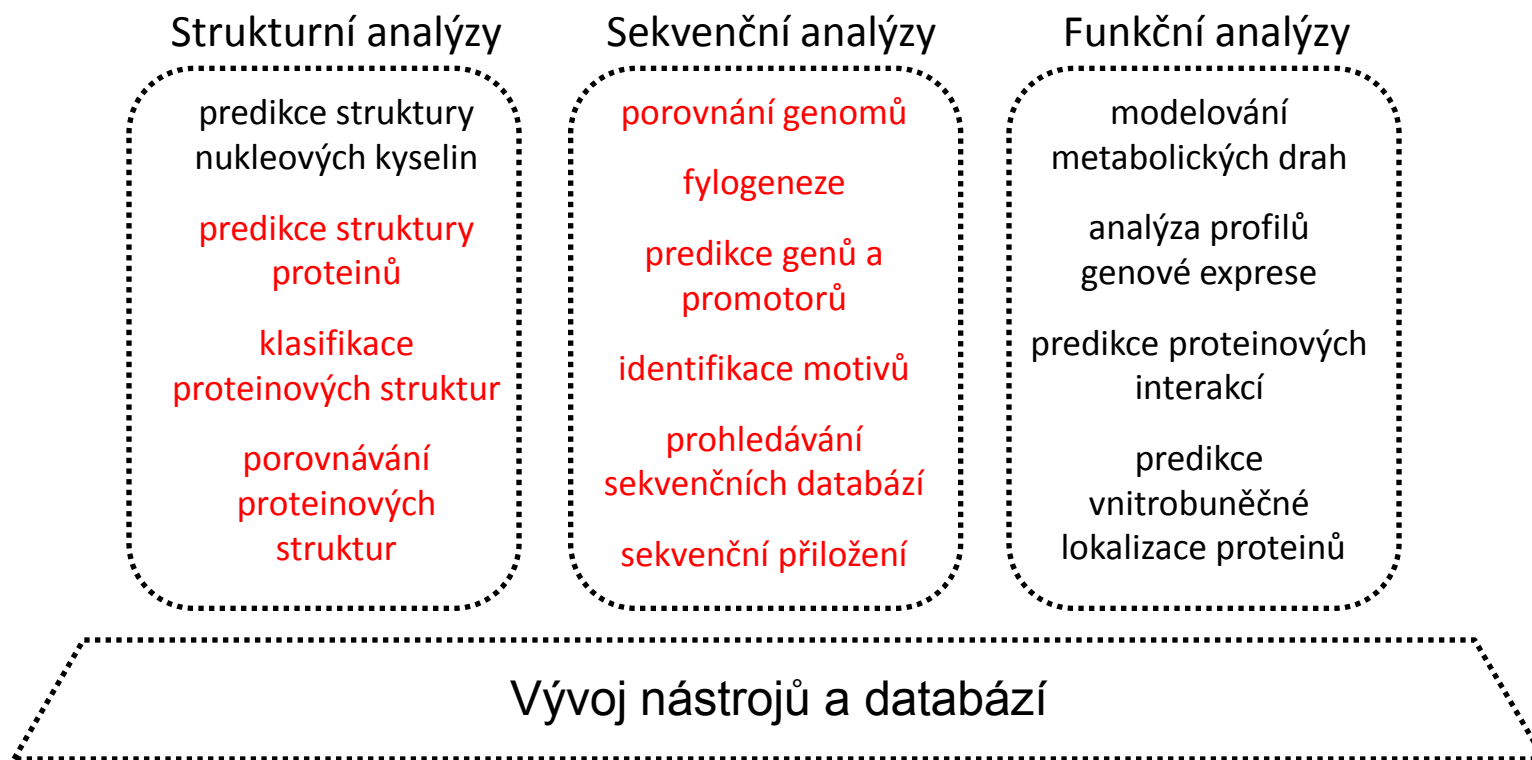
- Informační technologie sloužící k **získávání, uchovávání, analýze a distribuci informací** týkajících se biomakromolekul

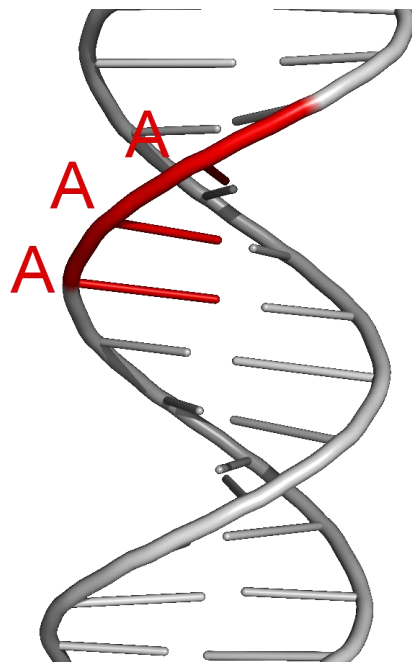
Vývoj nástrojů a databází

- Informační technologie sloužící k získávání, uchovávání, analýze a distribuci informací týkajících se biomakromolekul

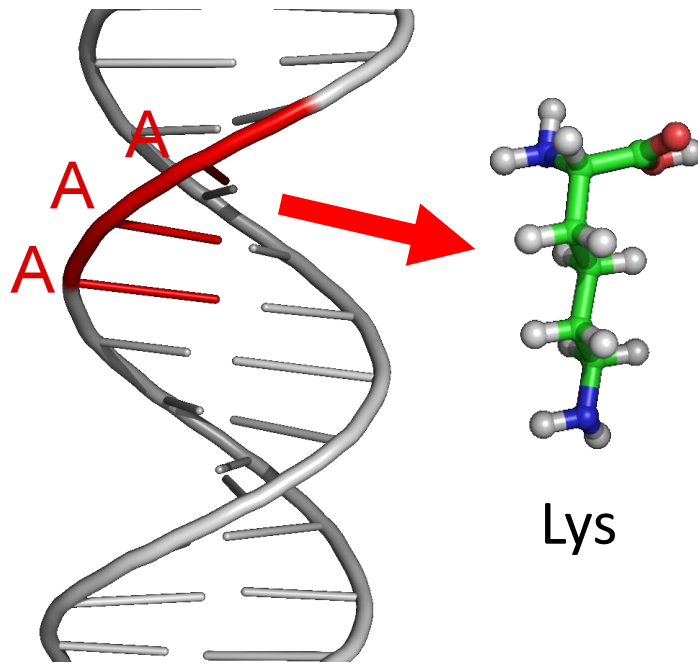


- Informační technologie sloužící k získávání, uchovávání, analýze a distribuci informací týkajících se biomakromolekul

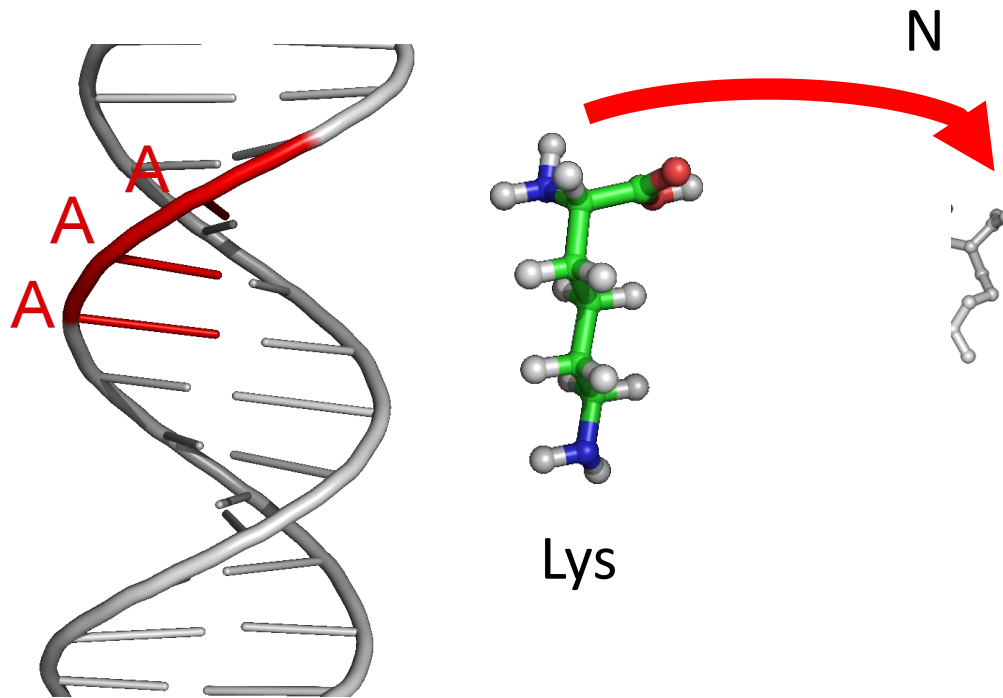




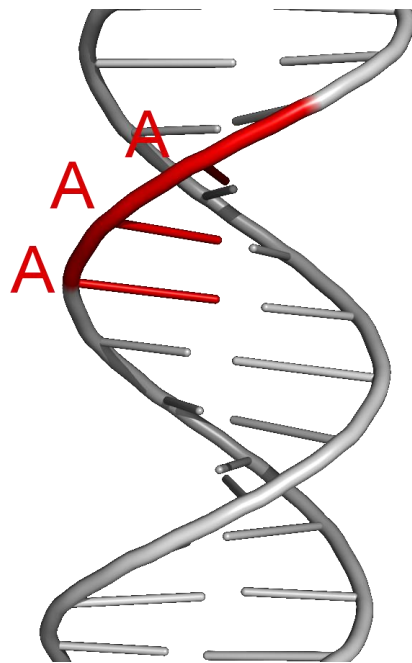
Nukleové kyseliny



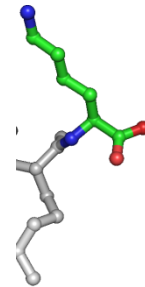
Nukleové kyseliny



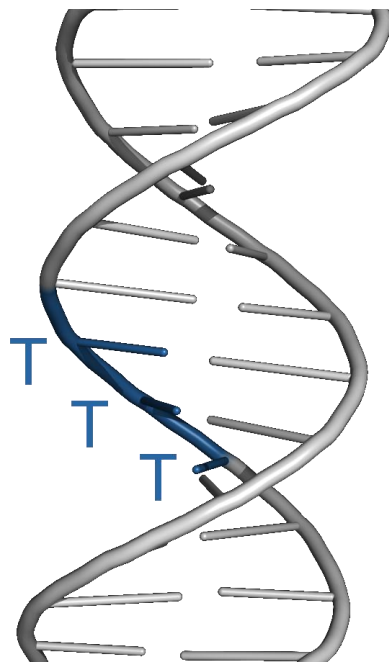
Nukleové kyseliny



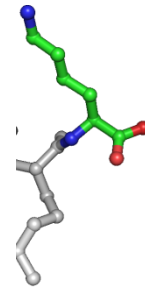
N-Lys



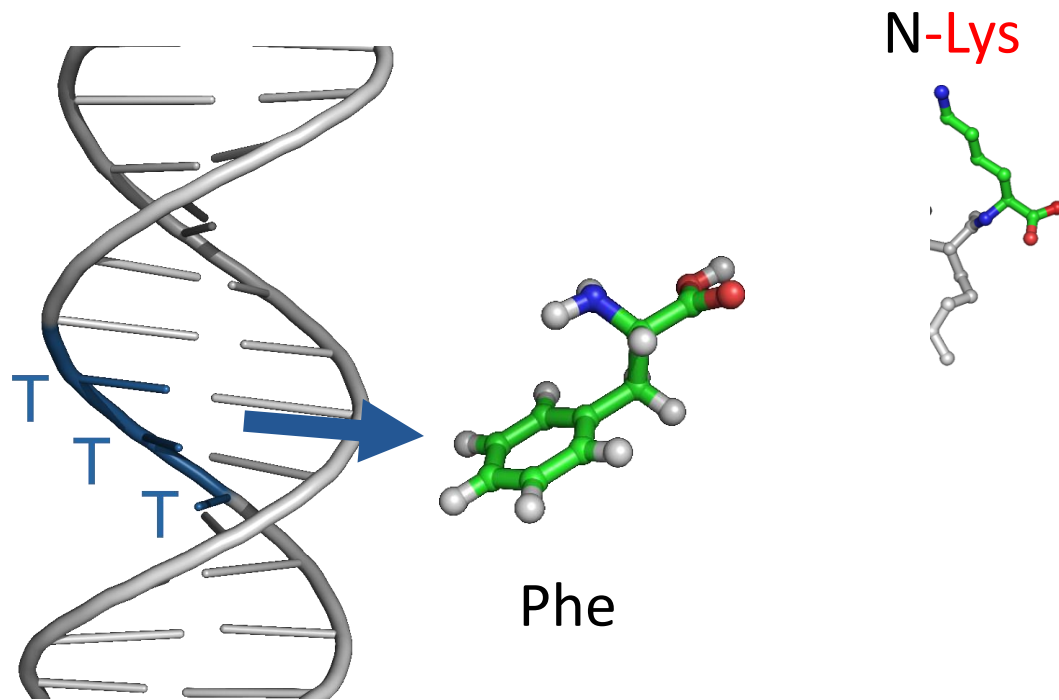
Nukleové kyseliny



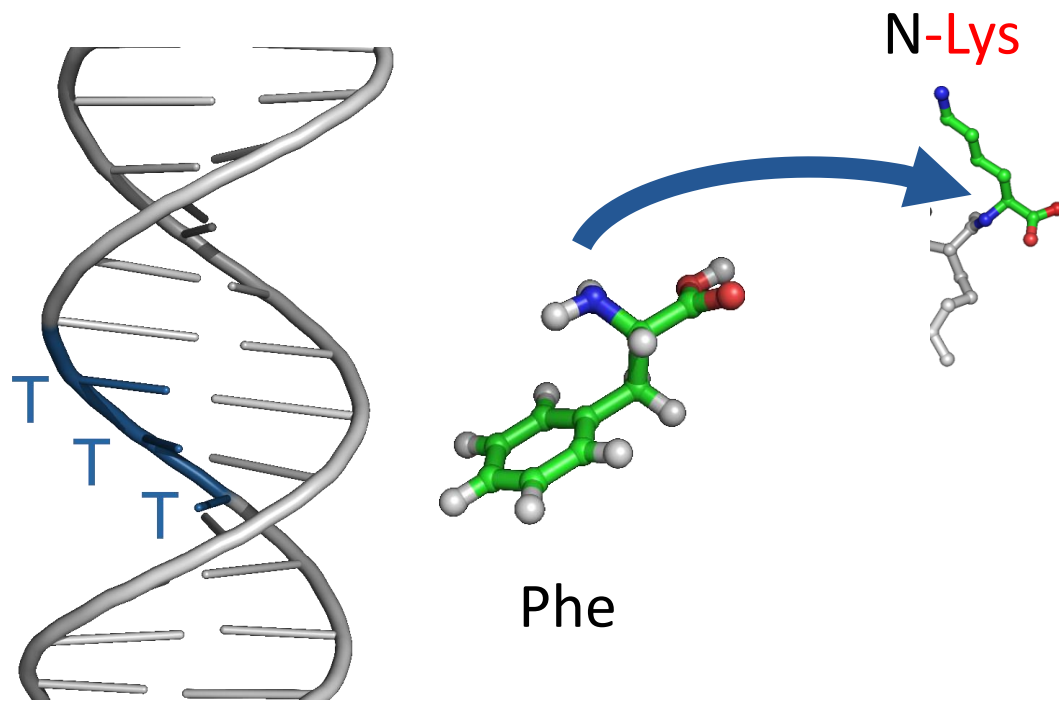
N-Lys



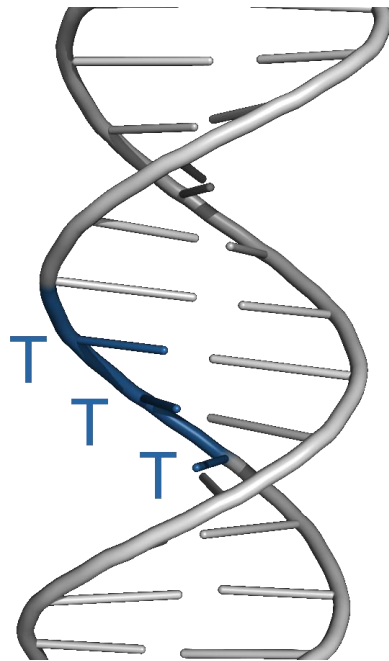
Nukleové kyseliny



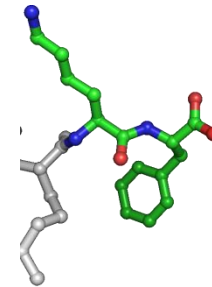
Nukleové kyseliny



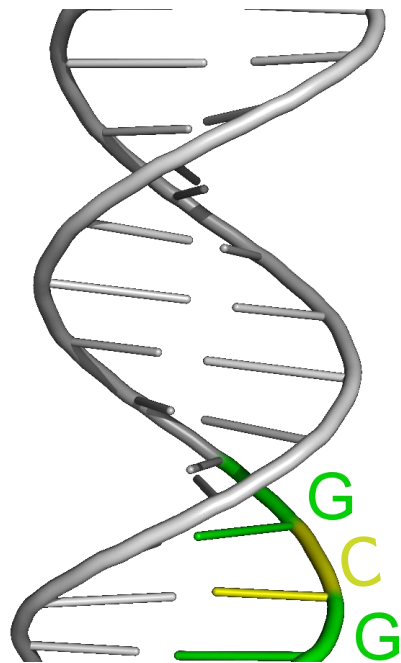
Nukleové kyseliny



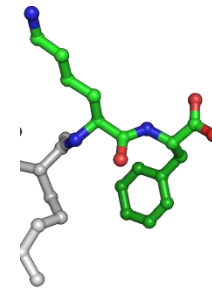
N-Lys-Phe



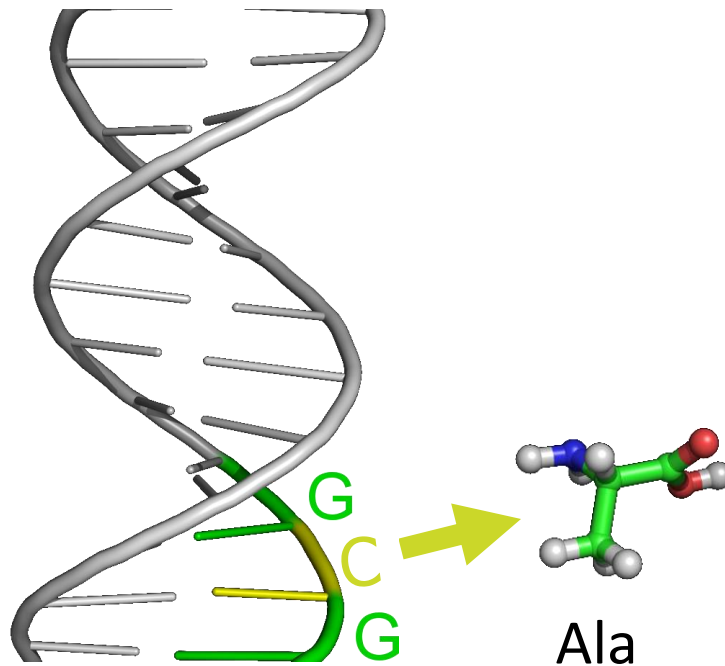
Nukleové kyseliny



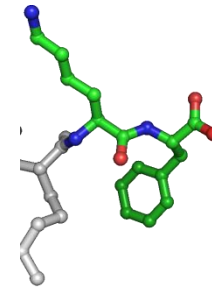
N-Lys-Phe



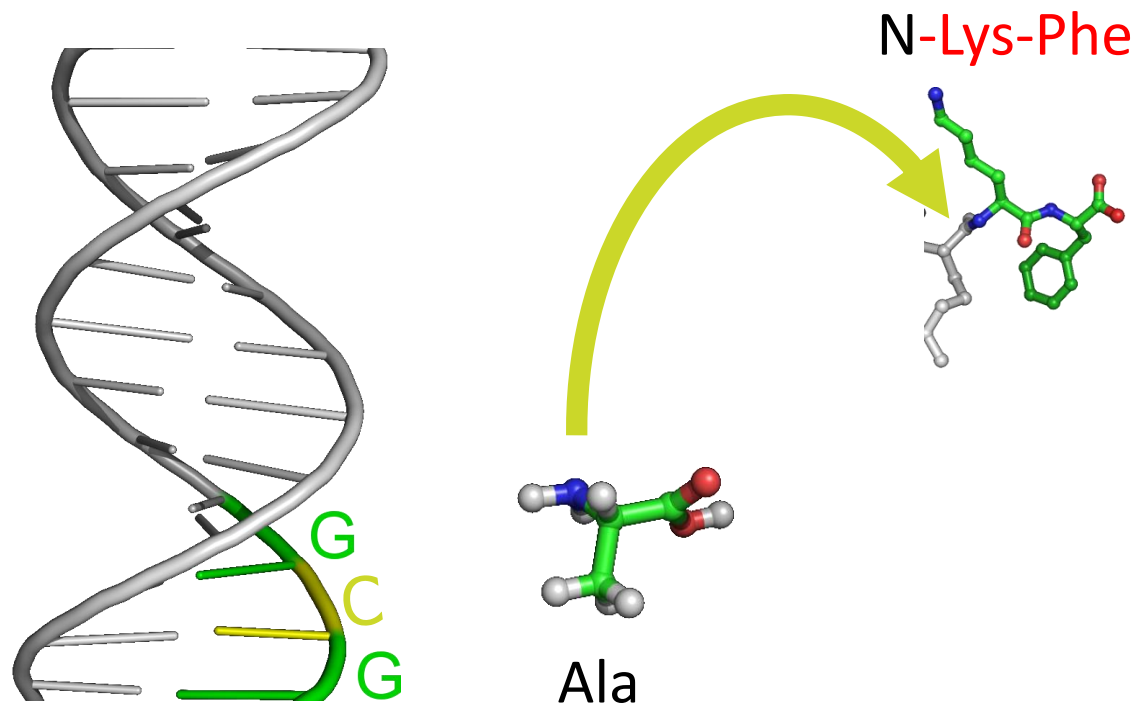
Nukleové kyseliny



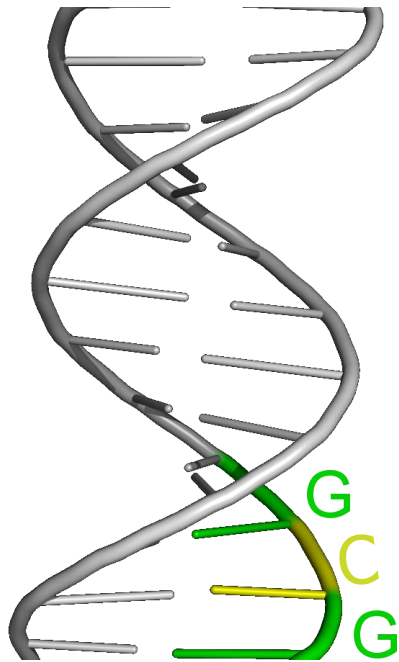
N-Lys-Phe



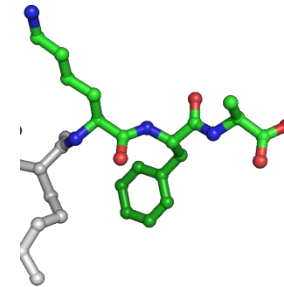
Nukleové kyseliny



Nukleové kyseliny

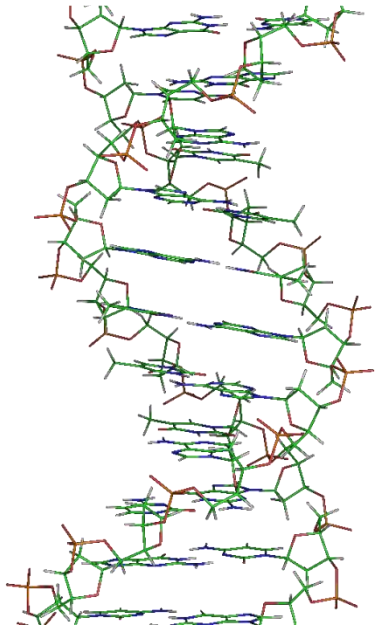


N-Lys-Phe-Ala

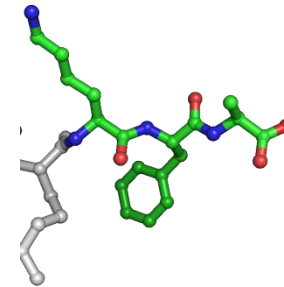


Nukleové kyseliny

5'-NCG-AAA-TTT-GCG-3'

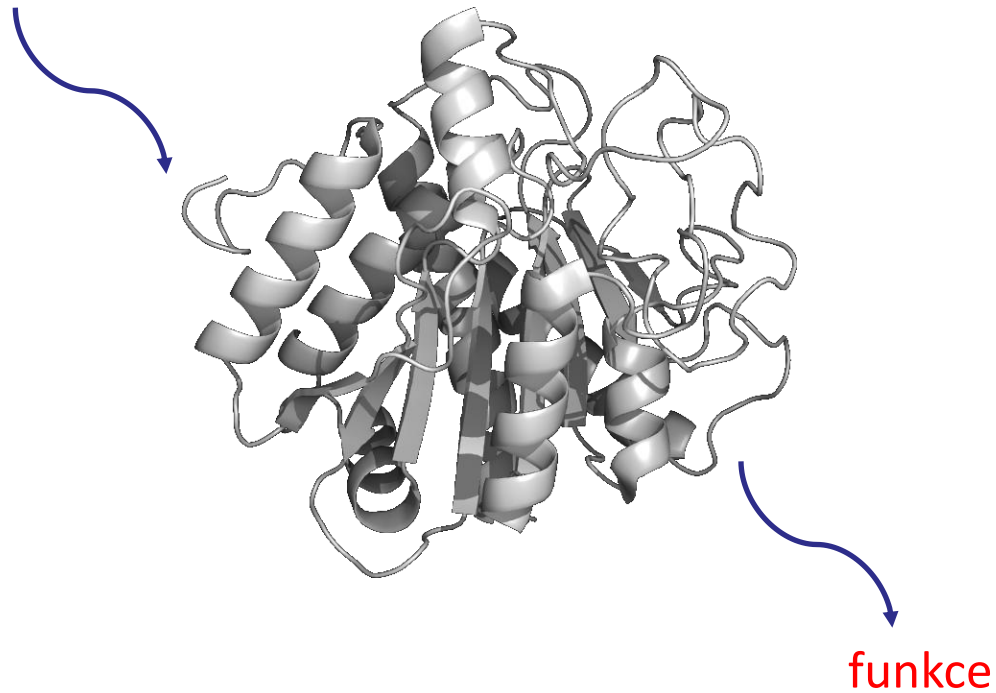


N-Lys-Phe-Ala



Proteiny

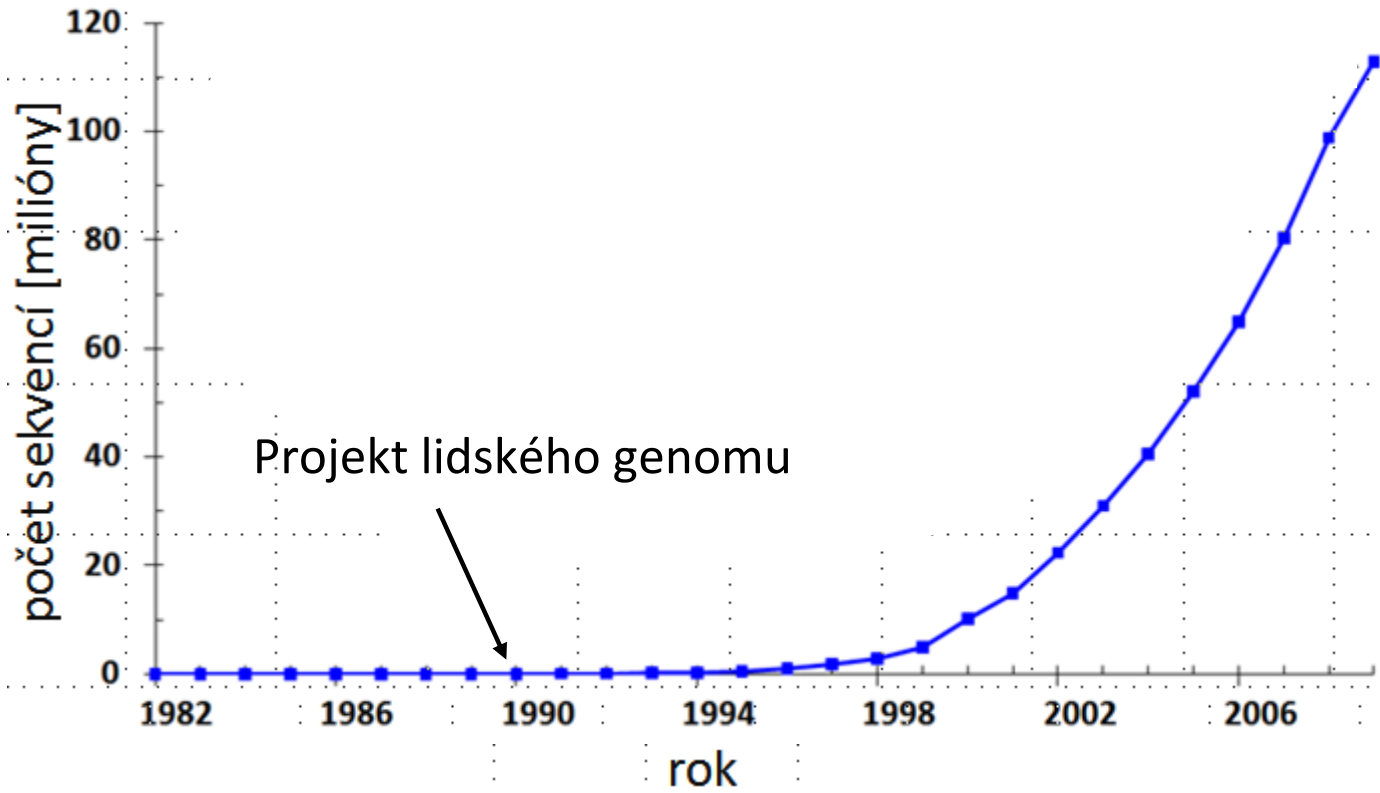
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIMPHCA
GLGRLIACDLIGMGDSKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVV
HDWGSALGFDWARRHRERVQGIAYMEAIAMPIEWADFPEQDRDLFQAFRS
QAGEELVLQD



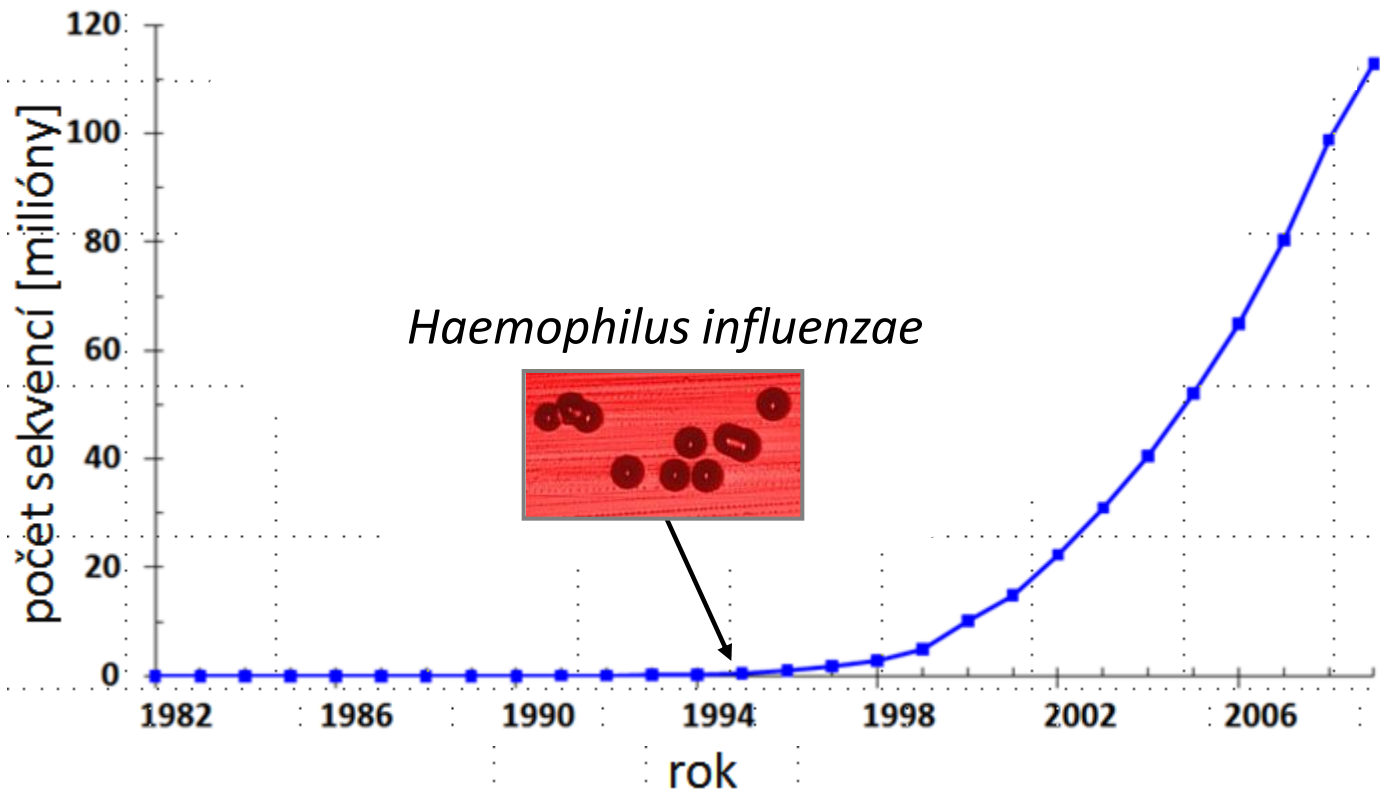
□ Revoluce



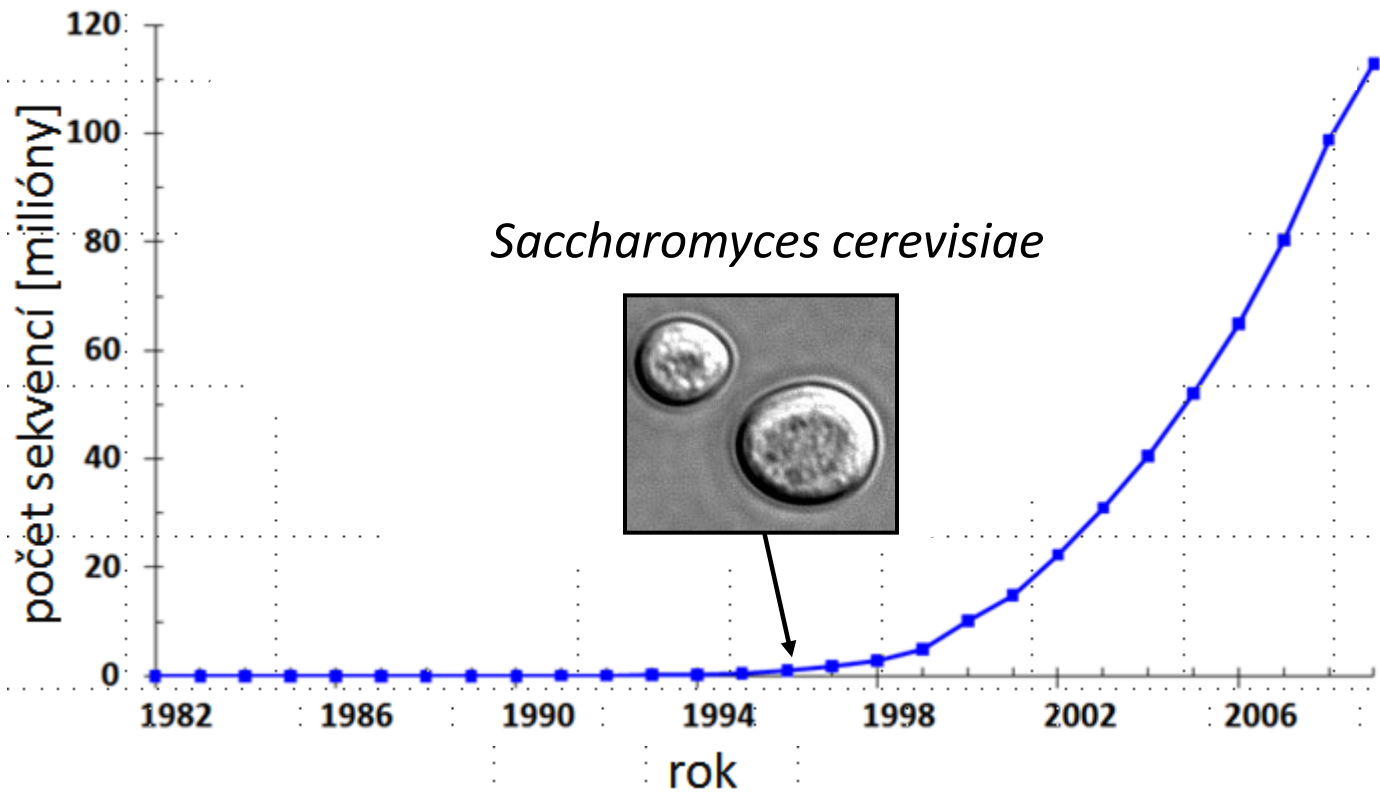
Historie bioinformatiky



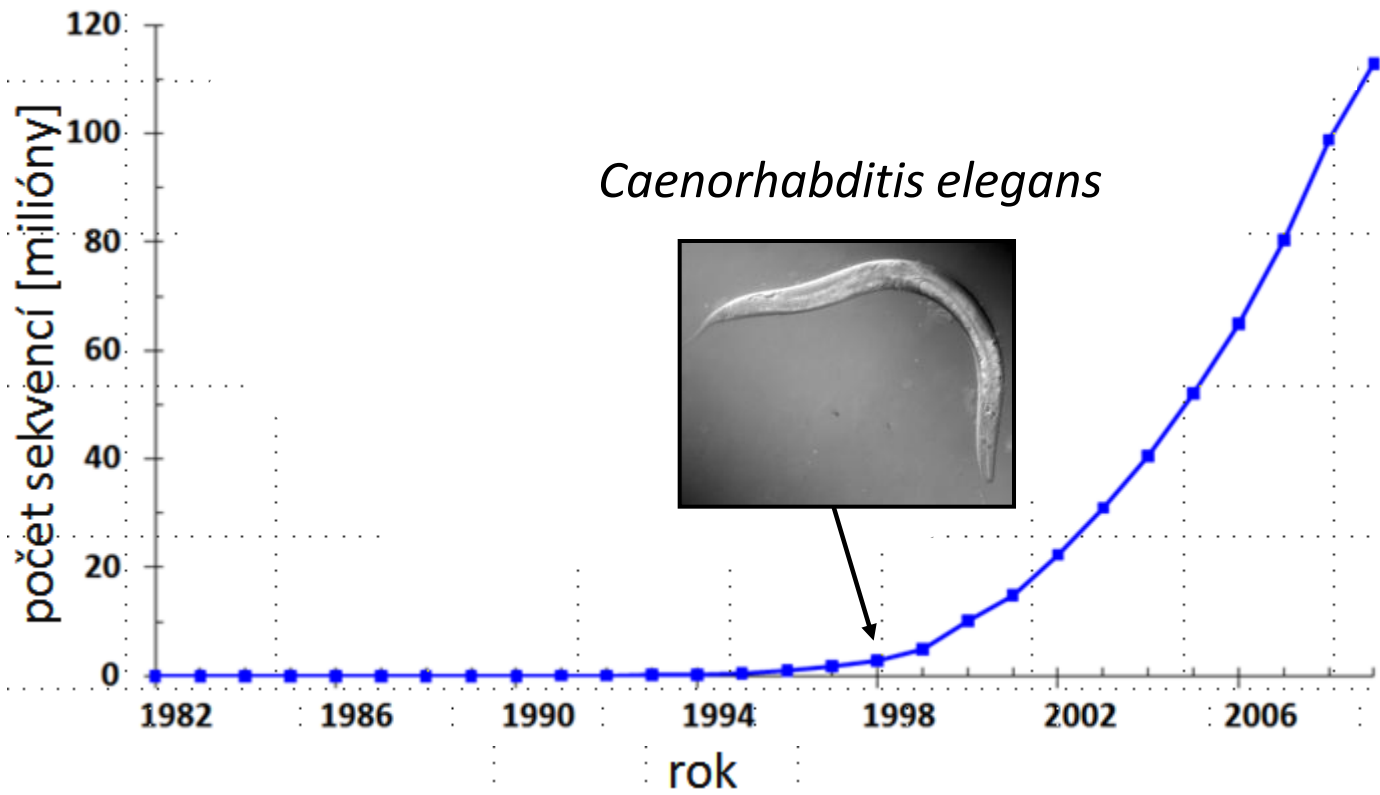
Historie bioinformatiky



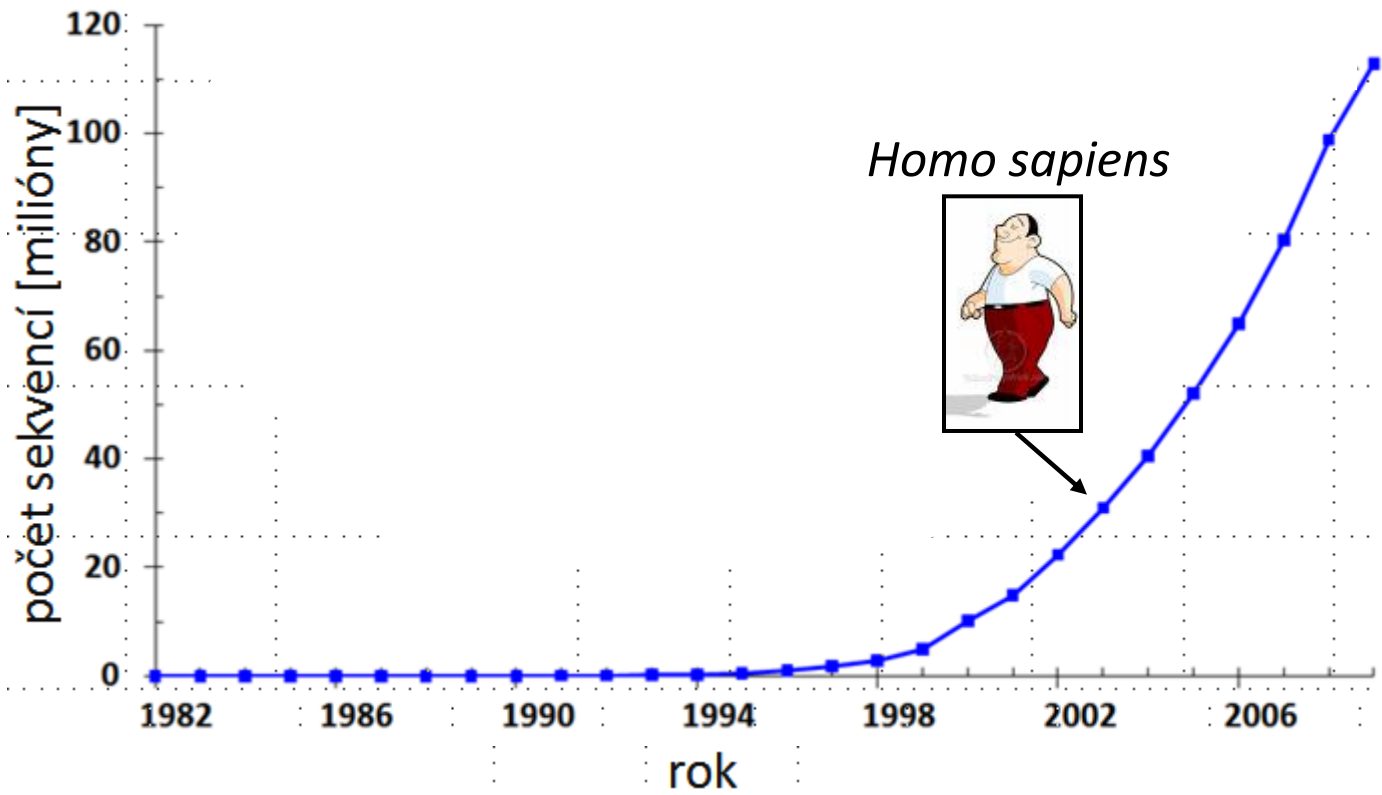
Historie bioinformatiky



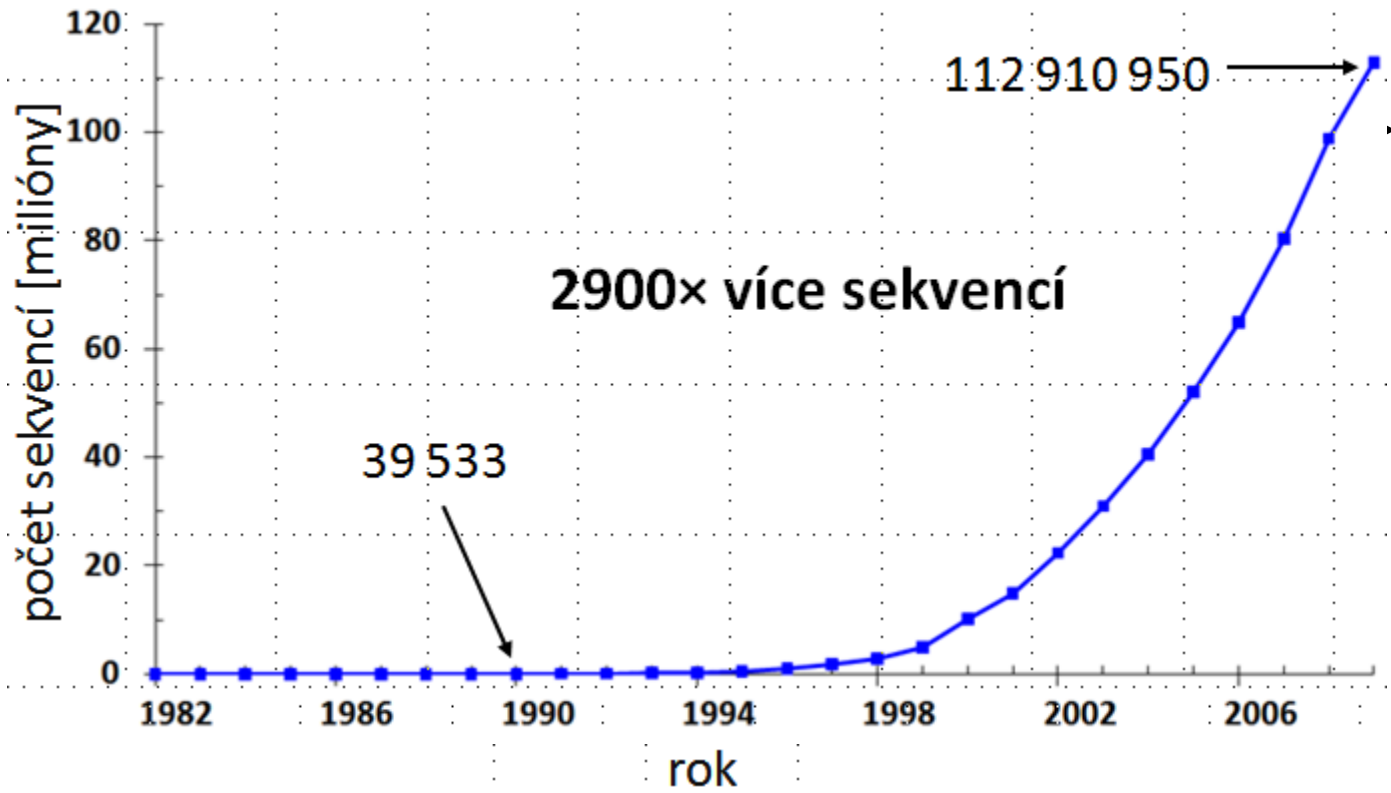
Historie bioinformatiky



Historie bioinformatiky



Historie bioinformatiky



Bioinformatické instituce

- ❑ National Center for Biotechnology Information (NCBI)
- ❑ European Bioinformatics Institute (EBI)



- National Center for Biotechnology Information (NCBI)
 - oddělení National Library of Medicine při National Institutes of Health v USA
 - Poskytuje Databáze GenBank, PubMed, OMIM, Genome dbSNP, ...
 - Informace dostupné přes vyhledávací **system Entrez**



□ National Center for Biotechnology Information (NCBI)

The screenshot displays the NCBI Entrez homepage. At the top left is the NCBI logo. To its right is the Entrez logo and the text "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with tabs for "HOME", "SEARCH", "SITE MAP", "PubMed", "All Databases", "Human Genome", "GenBank", "Map Viewer", and "BLAST". A search bar is located below the navigation bar, with the text "Search across databases" and buttons for "GO", "Clear", and "Help".

Below the search bar is a heading: "Welcome to the Entrez cross-database search page". This is followed by a grid of database links, each with an icon and a description:

- PubMed:** biomedical literature citations and abstracts
- PubMed Central:** free, full text journal articles
- Site Search:** NCBI web and FTP sites
- Books:** online books
- OMIM:** online Mendelian Inheritance in Man
- OMIA:** online Mendelian Inheritance in Animals
- Nucleotide:** Core subset of nucleotide sequence records
- EST:** Expressed Sequence Tag records
- GSS:** Genome Survey Sequence records
- Protein:** sequence database
- Genome:** whole genome sequences
- Structure:** three-dimensional macromolecular structures
- Taxonomy:** organisms in GenBank
- SNP:** single nucleotide polymorphism
- dbGaP:** genotype and phenotype
- UniGene:** gene-oriented clusters of transcript sequences
- CDD:** conserved protein domain database
- 3D Domains:** domains from Entrez Structure
- UniSTS:** markers and mapping data
- PopSet:** population study data sets
- GEO Profiles:** expression and molecular abundance profiles
- GEO DataSets:** experimental sets of GEO data

- European Bioinformatics Institute (EBI)
 - Součást European Molecular Biology Laboratory (EMBL), Wellcome Trust Genome Campus ve Velké Británii
 - Poskytuje databáze EMBL-Bank, UniProt, Ensembl, InterPro, ...
 - Informace dostupné přes vyhledávací **system SRS**



Bioinformatické inštituce

□ European Bioinformatics Institute (EBI)

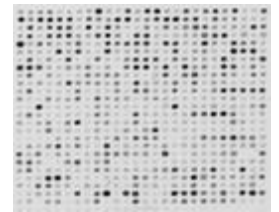
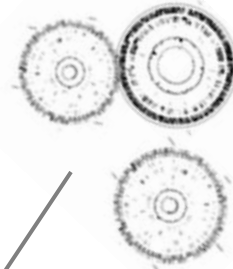
The screenshot displays the EMBL-EBI website interface. At the top, there is a search bar with the text "Nucleotide sequences" and a "Go" button. The main navigation menu includes "EBI Home", "About EBI", "Research", "Services", "Toolbox", "Databases", "Downloads", and "Submissions". Below the navigation menu, there is a "SERVICES OVERVIEW FASTLINK" section. A tree diagram on the left side of the page shows the following structure:

- Databases
 - Database Browsing
 - SRS
 - Nucleotide Databases
 - EMBL Nucleo. Sequence
 - Ensembl
 - Genomes Server
 - Genome MOT
 - EMBL-Align
 - Simple Queries
 - dbSTS Queries
 - Parasites
 - Mutations
 - IMG
 - Protein Databases
 - SWISS-PROT
 - TrEMBL
 - InterPro
 - CluSTR
 - IPi
 - GOA
 - Proteome Analysis
 - HPI
 - IntEnz
- Toolbox
 - Homology & Similarity
 - Fasta
 - WU-Blast2
 - NCBI-Blast2
 - Blast2_EVEC
 - Genome/Proteome Fasta
 - MPsrch
 - Scanps2_3
 - Parasite-Blast
 - EG1-Blast
 - SNP-Fasta3 Server
 - Prot. Function. Analysis
 - CluSTR Search
 - InterProScan
 - FingerPRINTScan
 - ppsearch
 - GeneQuiz
 - Pratt
 - Radar
- Submissions
 - EMBL via WEBIN
 - EMBL-Info. Submitters
 - SWISS-PROT
 - Webin-Align
 - PDB-AutoDep
 - MIAMEpress
 - IMG/HLA
 - Sequin Software
- Downloads
 - FTP Server
 - Database Repository
 - Software Repository
 - Downloads Help Files
- Services Help

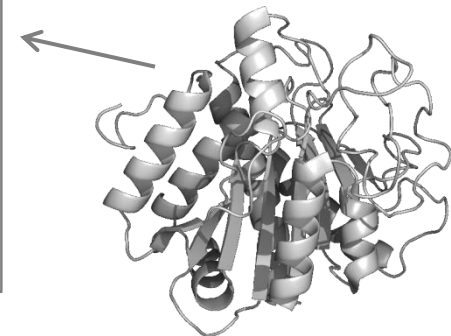
Bioinformatické databáze

- ❑ Bibliografie
- ❑ Nukleotidové sekvence
- ❑ Proteinové sekvence
- ❑ Proteinové struktury
- ❑ Genomy

```
LRHLGITGPVTLAVHDWGGMIGFGWALSHHAQVKRLVI  
TNTAAAGTKFDKLTWLDVGPVDLVVHDWGTISRMEEGT  
WYKLIRTTVWHQAIVLAEIGTVVCKTQENPA
```

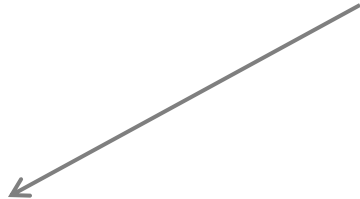


Count	Database	Description	Search Options
44	PubMed	biomedical literature citations and abstracts	none
79	PubMed Central	free, full text journal articles	none
none	Site Search	NCBI web and FTP sites	none
45	Nucleotide	sequence database (includes GenBank)	none
39	Protein	sequence database	none
4	Genome	whole genome sequences	12
12	Structure	three-dimensional macromolecular structures	none
none	Books	online books	none
none	OMIM	online Mendelian Inheritance in Man	none
none	OMIA	Online Mendelian Inheritance in Animals	none
none	UniGene	gene-oriented clusters of transcript sequences	none
none	CDD	conserved protein domain database	none
12	3D Domains	domains from Entrez Structure	none
none	UniSTS	markers and mapping data	none



Bibliografické databáze

- ❑ PubMed
- ❑ Web of Science



The screenshot shows the PubMed website interface. At the top, it says 'NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health www.pubmed.gov'. Below this, there are navigation tabs for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'OMIM', 'PMC', and 'Journals'. A search bar contains 'PubMed' and 'for'. There are buttons for 'Go', 'Clear', and 'Advanced Search'. Below the search bar, there are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. On the left side, there is a sidebar with links for 'About Entrez', 'Text Version', 'Entrez PubMed', 'Overview', 'Help | FAQ', 'Tutorials', 'New/Noteworthy', 'E-Utilities', 'PubMed Services', 'Journals Database', 'MeSH Database', 'Single Citation Matcher', 'Batch Citation Matcher', 'Clinical Queries', 'Special Queries', and 'LinkOut'. The main content area has a heading 'To get started with PubMed, enter one or more search terms.' and 'Search terms may be topics, authors or journals.' Below this, there is a highlighted box with the text 'Set up an automated PubMed update in fewer than five minutes.' and a list of three steps: 1. Create a My NCBI account, 2. Save your search, 3. Your PubMed updates can be e-mailed directly to you. At the bottom, it says 'PubMed is a service of the U.S. National Library of Medicine that includes over 17 million citations'.

The screenshot shows the ISI Web of Knowledge website interface. At the top, it says 'ISI Web of Knowledge SM Take the next step'. Below this, there are navigation tabs for 'Web of Science' and 'Additional Resources'. A search bar contains 'Search' and 'Cited Reference Search'. There are buttons for 'Advanced Search', 'Search History', and 'Marked List (0)'. Below the search bar, there are three search fields: 'Search for:' with a dropdown menu for 'Topic', 'AND' with a dropdown menu for 'Author', and 'AND' with a dropdown menu for 'Publication Name'. Each field has a search button. Below the search fields, there are buttons for 'Search' and 'Clear'.



□ PubMed

- Provozováno National Library of Medicine
- Obsahuje více než 22 milionů citací **biomedicínské literatury**
- Integruje MEDLINE, časopisy z oblasti živých věd a online knihy
- Prohledávání možné přes **Entrez** nebo **DBGET**
- Obsahuje kromě **abstraktů** odkazy na **plné texty** dostupné přes PubMed Central nebo stránky nakladatelství



□ PubMed



U.S. National Library of Medicine
National Institutes of Health

[Limits](#) [Advanced search](#) [Help](#)

[Display Settings:](#) Abstract [Send to:](#)



[Nat Chem Biol.](#) 2009 Oct;5(10):727-33. Epub 2009 Aug 23.

Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate.

[Pavlova M](#), [Klvana M](#), [Prokop Z](#), [Chaloupkova R](#), [Banas P](#), [Otyepka M](#), [Wade RC](#), [Tsuda M](#), [Nagata Y](#), [Damborsky J](#).
Loschmidt Laboratories, Institute of Experimental Biology and National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic.

Abstract

Engineering enzymes to degrade anthropogenic compounds efficiently is challenging. We obtained *Rhodococcus rhodochrous* haloalkane dehalogenase mutants with up to 32-fold higher activity than wild type toward the toxic, recalcitrant anthropogenic compound 1,2,3-trichloropropane (TCP) using a new strategy. We identified key residues in access tunnels connecting the buried active site with bulk solvent by rational design and randomized them by directed evolution. The most active mutant has large aromatic residues at two out of three randomized positions and two positions modified by site-directed mutagenesis. These changes apparently enhance activity with TCP by decreasing accessibility of the active site for water molecules, thereby promoting activated complex formation. Kinetic analyses confirmed that the mutations improved carbon-halogen bond cleavage and shifted the rate-limiting step to the release of products. Engineering access tunnels by combining computer-assisted protein design with directed evolution may be a valuable strategy for refining catalytic properties of enzymes with buried active sites.

PMID: 19701186 [PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances, Secondary Source ID

LinkOut - more resources

Full Text Sources:
[Nature Publishing Group](#)

Related citations

Biodegradation of 1,2,3-trichloropropane through directed evolution an [Appl Environ Microbiol. 2002]

Pathways and mechanisms for product release in the engineered haloalkane dehalo [J Mol Biol. 2009]

Mechanism of enhanced conversion of 1,2,3-trichloropropane b [J Comput Aided Mol Des. 2006]

Review Evolving haloalkane dehalogenases. [Curr Opin Chem Biol. 2004]

Review Alpha/Beta-hydrolase fold enzymes: structures, functions [Curr Protein Pept Sci. 2000]

[See reviews...](#)
[See all...](#)

All links from this record

[Related Citations](#)

[Compound \(MeSH Keyword\)](#)

[Compound \(Publisher\)](#)

[Substance \(MeSH Keyword\)](#)

[Substance \(Publisher\)](#)

□ Web of Science

- Komerční databáze
- Součást ISI Web of Knowledge
- Používá se pro zjištění **citovanosti** a **impaktního** faktoru časopisů

Bibliografické databáze

□ Web of Science

The screenshot shows a Web of Science article page. The header includes 'ISI Web of Knowledge SM' and navigation options like 'Search', 'Cited Reference Search', 'Structure Search', 'Advanced Search', 'Search History', and 'Marked List (0)'. The article title is 'Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate'. It lists authors: Pavlova M, Klvana M, Prokop Z, Chaloupkova R, Banas P, Otyepka M, Wade RC, Tsuda M, Nagata Y, and Damborsky J. The source is 'NATURE CHEMICAL BIOLOGY', Volume 5, Issue 10, Pages 727-733, published in OCT 2009. The abstract discusses engineering enzymes to degrade anthropogenic compounds. A 'Cited by: 6' section lists citing articles, including one by Kourist R et al. in 'Engineering CHEMBIOCHEM' and another by Stsiapanava A et al. in 'ACTA CRYSTALLOGRAPHICA SECTION D-BIOLOGICAL CRYSTALLOGRAPHY'. A 'Create Citation Alert' button is visible.

ISI Web of KnowledgeSM

Web of Science Additional Resources

Search Cited Reference Search Structure Search Advanced Search Search History Marked List (0)

Web of Science® – with Conference Proceedings

<< Back to results list Record 1 of 1 Record from Web of Science®

Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate

SFX NCBI Print E-mail Add to Marked List Save to EndNote Web Save to EndNote, RefMan, ProCite more options

Author(s): Pavlova M (Pavlova, Martina)^{1,2}, Klvana M (Klvana, Martin)^{1,2}, Prokop Z (Prokop, Zbynek)^{1,2}, Chaloupkova R (Chaloupkova, Radka)^{1,2}, Banas P (Banas, Pavel)^{3,4}, Otyepka M (Otyepka, Michal)^{3,4}, Wade RC (Wade, Rebecca C.)⁵, Tsuda M (Tsuda, Masataka)⁶, Nagata Y (Nagata, Yuji)⁶, Damborsky J (Damborsky, Jiri)^{1,2}

Source: NATURE CHEMICAL BIOLOGY **Volume:** 5 **Issue:** 10 **Pages:** 727-733 **Published:** OCT 2009

Times Cited: 6 **References:** 50 [Citation Map](#)

Abstract: Engineering enzymes to degrade anthropogenic compounds efficiently is challenging. We obtained Rhodococcus rhodochrous haloalkane dehalogenase mutants with up to 32-fold higher activity than wild type toward the toxic, recalcitrant anthropogenic compound 1,2,3-trichloropropane (TCP) using a new strategy. We identified key residues in access tunnels connecting the buried active site with bulk solvent by rational design and randomized them by directed evolution. The most active mutant has large aromatic residues at two out of three randomized positions and two positions modified by site-directed mutagenesis. These changes apparently enhance activity with TCP by decreasing accessibility of the active site for water molecules, thereby promoting activated complex formation. Kinetic analyses confirmed that the mutations improved carbon-halogen bond cleavage and shifted the rate-limiting step to the release of products. Engineering access tunnels by combining computer-assisted protein design with directed evolution may be a valuable strategy for refining catalytic properties of enzymes with buried active sites.

Document Type: Article

Language: English

KeyWords Plus: SPHINGOMONAS-PAUCIMOBILIS UT26; HALOALKANE DEHALOGENASE; DIRECTED EVOLUTION; CYTOCHROME P450S; HETEROLOGOUS EXPRESSION; XENOBIOTIC COMPOUNDS; CATALYTIC MECHANISM; ESCHERICHIA-COLI; ENZYME; SPECIFICITY

Reprint Address: Damborsky, J (reprint author), Masaryk Univ, Fac Sci, Loschmidt Labs, Inst Expt Biol, CS-61137 Brno, Czech Republic

Cited by: 6

This article has been cited 6 times (from Web of Science).

Kourist R, Jochens H, Bartsch S, et al. *The alpha/beta-Hydrolase Fold 3DM Database (ABHDB) as a Tool for Protein Engineering* CHEMBIOCHEM 11 12 1635-1643 AUG 16 2010

Stsiapanava A, Dohnalek J, Gavira JA, et al. *Atomic resolution studies of haloalkane dehalogenases DhaA04, DhaA14 and DhaA15 with engineered access tunnels* ACTA CRYSTALLOGRAPHICA SECTION D-BIOLOGICAL CRYSTALLOGRAPHY 66 962-969 Part 9 SEP 2010

Brouk M, Derry NL, Shainsky J, et al. *The influence of key residues in the tunnel entrance and the active site on activity and selectivity of toluene-4-monooxygenase* JOURNAL OF MOLECULAR CATALYSIS B-ENZYMATIC 66 1-2 72-80 SEP 2010

[[view all 6 citing articles](#)]

[Create Citation Alert](#)

Related Records:

Databáze nukleotidových sekvencí



- ❑ GenBank
 - ❑ EMBL-Bank
 - ❑ DDBJ
-
- ❑ Anotované kolekce veřejně dostupných nukleotidových sekvencí
 - ❑ Data získaná z genomových center a odborných pracovišť
 - ❑ **Každodenní** vzájemná **synchronizace** nových a aktualizovaných dat
 - ❑ **“Accession number”** - jedinečný identifikátor záznamu, ve všech třech databázích

Databáze nukleotidových sekvencí

□ GenBank

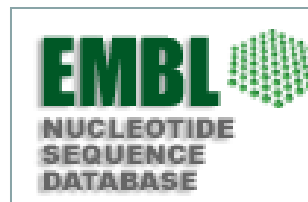
- Založena v roce 1982, provozována NCBI
- Přístupná prostřednictvím vyhledávacího systému **Entrez** nebo systému **DBGET**
- Obsahuje více než 156.000.000 sekvencí (2012)
- Nové sekvence možné vložit pomocí **BankIt** nebo **Sequin**



Databáze nukleotidových sekvencí

□ EMBL-Bank

- Založena v roce 1980, Provozována EBI
- Přístupná prostřednictvím vyhledávacího systému **SRS** či **DBGET**
- Obsahuje více než 247.000.000 sekvencí (2012)
- Nové sekvence možné vložit pomocí **Webin** nebo **Sequin**



Databáze nukleotidových sekvencí

- DNA Data Bank of Japan (DDBJ)
 - Založena v roce 1984, provozována National Institute of Genetics
 - Obsahuje více než 153.000.000 sekvencí (2012)
 - Nové sekvence možné vložit pomocí **Sakura** nebo **Sequin**



Ukázka záznamu v GenBank

□ Hlavička

- Základní informace o záznamu
- Lokus, definice, přístupový kód, klíčová slova, organizmus, reference,..

X.autotrophicus haloalkane dehalogenase (dh1A) gene, complete cds

[Comment](#) [Features](#) [Sequence](#)

```
LOCUS          XAADHLA                      3041 bp    DNA    linear    BCT 15-FEB-1996
DEFINITION     X.autotrophicus haloalkane dehalogenase (dh1A) gene, complete cds.
ACCESSION     M26950
VERSION       M26950.1  GI:155347
KEYWORDS      haloalkane dehalogenase.
SOURCE        Xanthobacter autotrophicus
  ORGANISM    Xanthobacter autotrophicus
              Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;
              Xanthobacteraceae; Xanthobacter.
REFERENCE     1 (bases 1 to 3041)
  AUTHORS     Janssen,D.B., Pries,F., van der Ploeg,J., Kazemier,B., Terpstra,P.
              and Witholt,B.
  TITLE       Cloning of 1,2-dichloroethane degradation genes of Xanthobacter
              autotrophicus GJ10 and expression and sequencing of the dh1A gene
  JOURNAL     J. Bacteriol. 171 (12), 6791-6799 (1989)
  PUBMED     2687254
COMMENT       Draft entry and computer readable copy of sequence [1] kindly
              provided by D.B.Janssen, 11-AUG-1989.
```

Ukázka záznamu v GenBank

□ Charakteristiky

- Popis jednotlivých oblastí genu
- Promotor, RBS (ribozóm vazebné místo), CDS (kódující sekvence), ...

```
gene          918..1931
              /gene="dh1A"
promoter      918..946
              /gene="dh1A"
              /note="putative"
promoter      945..974
              /gene="dh1A"
              /note="putative"
RBS           986..998
              /gene="dh1A"
CDS           999..1931
              /gene="dh1A"
              /codon_start=1
              /transl_table=11
              /product="haloalkane dehalogenase"
              /protein_id="AAA88691.1"
              /db_xref="GI:155348"
              /translation="MINAIRTPDQRFNSLDQYPFSPNYLDDLPGYPLRAHYLDEGNS
DAEDVFLCLHGEPTWSYLYRKMIPVFAESGARVIAPDFFGFGKSDKPVDEEDYTFEFH
RNFLALIERLDRNITLVVQDWGGFLGLTLPMDPSRFKRLIIMNACLMTDPVTQPA
FSAFVTQPADGFTAWKYDLVTPSDLRLDQFMKRWAPTLEAEASAYAAPPDTSYQAG
```

Ukázka záznamu v GenBank



□ Charakteristiky

FEATURES	Location/Qualifiers
source	1..3041 /organism="Xanthobacter autotrophicus" /mol_type="genomic DNA" /strain="GJ10" /db_xref="taxon:280"
CDS	complement(316..924) /note="ORF 1; putative" /codon_start=1 /transl_table=11 /product="unknown protein" /protein_id="AAA88690.1" /db_xref="GI:1197026" /translation="MSTFFEPENGMKQNAKTERILDVALELLETEGEFGLTMRQVATQ ADMSLSNVQYYFKSEDLVVAMADRYFQRCCLTMAEHPPLSAGRDRQHAQLRALRELL GHGLEISEMCRIFREYWAIAATRNETHVHGKLSYYRDLAEVMAEKLAPLASSEKALAVA VSLVIPYVEGYSVTAIAMPESIDTISETLTNVVLEQLRISNS"
gene	918..1931 /gene="dh1A"
promoter	918..946 /gene="dh1A" /note="putative"
promoter	945..974 /gene="dh1A" /note="putative"
RBS	986..998 /gene="dh1A"
CDS	999..1931 /gene="dh1A" /codon_start=1 /transl_table=11 /product="haloalkane dehalogenase" /protein_id="AAA88691.1" /db_xref="GI:155348" /translation="MINAIRTPDQRFNSLDQYPPSPNYLDDLPGYPLRAHYLDEGNS DAEDVFLCLHGEPWWSYLYRKMIPVFAESGARVIAPDFFGKSDKPVDEEDYTFEFH RNFLLALIERLDRNITLVVQDWGGFLGLTLPMDPSRFKRLIIMNACLMTDPVTQPA FSAFVTQPADGFTAWKYDLVTPSDLRLDQFMKRWAPTLTEAEASAYAAPFPDTSYQAG VRKFKMVAQRDQACIDISTEAFVQNDWNGQTFMAIGMKDKLLGPDVMPMKALIN GCPEPLEIADAGHFVQEFGEQVAREALKHFAETE"

Ukázka záznamu v GenBank

□ Sekvence

```
ORIGIN      1 bp upstream of BamHI site.
      1 atgataaatg caattcgcac cccggaccaa cgcttcagca atctcgatca gtatccgttc
     61 agccccaact acctggacga cctccccggc taccgggat  tgcgggcaca ctacctcgac
    121 gagggcaatt ctgacgctga agacgttttt ctctgccttc atggcgagcc cacctggagt
    181 tacctgtatc gcaagatgat cccggtatth gctgaatcag ggcacagagt tattgcgcca
    241 gacttttttg gattcggaaa atccgacaag ccagtagacg aagaagacta caccttcgaa
    301 tttcaccgca acttcctgct tgcactaatc gaacggcttg acttgcgcaa cattacgctg
    361 gtcgttcagg actggggcgg atttttgggg ctgaccttac cgatggccga cccttcccgc
    421 ttcaagcgcc tgatcatcat gaacgcctgc ttgatgaccg acccggtcac ccagcctgcg
    481 tttagcgcct ttgtcaccca gcctgcggat ggctttaccg cctggaaata cgatctggtt
    541 acgccatcag acctgcgcct tgaccagttc atgaagcgtt gggcgcccac actgaccgaa
    601 gctgaggcct ccgcgtatgc tgcgcctttc cctgacactt cctatcaggc tgggtgtacgc
    661 aagtttccca agatggtcgc gcaacgcgac caggcctgca tcgacatttc aaccgaagcg
    721 atttcgttct ggcagaacga ctggaatggc cagaccttca tggccattgg catgaaagac
    781 aaattgctgg gaccggacgt catgtatcct atgaaggcgc tcattaatgg ctgcccggaa
    841 ccctcgaaa tagcggacgc tggccatttc gtacaggagt ttggcgagca agtggctcgc
    901 gaggcctga  aacactttgc cgagacagaa tag
```

//

Databáze proteinových sekvencí

- UniProtKB
- nr Protein Database



Databáze proteinových sekvencí

□ UniProtKB

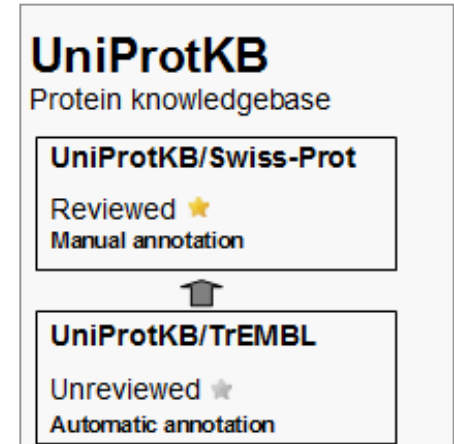
- Spolupráce EBI, Swiss Institute of Bioinformatics a Protein Information Resource
- Centrální úložiště proteinových sekvencí a funkčních informací
- **Kvalitní anotace** - informace o funkci proteinu a jednotlivých aminokyselin, experimentální informace, biologické ontologie, klasifikace, odkazy do dalších databází
- **Indikace kvality** anotace (**manuální** vs. **automatická**)



Databáze proteinových sekvencí

□ UniProtKB/Swiss-Prot

- Vysoká kvalita manuálních anotací
- 😊 Manuální anotace – **spolehlivé informace**
- 😞 500.000 sekvencí (srpen 2012)



□ UniProtKB/TrEMBL

- Sekvence konceptuální translací kódujících sekvencí EMBL-Bank
- 😞 Automatická anotace – **nižší kvalita, možnosti chyb**
- 😊 24.000.000 sekvencí (srpen 2012)



Ukázka záznamu v UniProtKB

- ❑ Názvy a zdroj proteinu
- ❑ Vlastnosti proteinu

Names and origin		Hide Top
Protein names	<i>Recommended name:</i> Haloalkane dehalogenase EC=3.8.1.5 <i>Alternative name(s):</i> 1,3,4,6-tetrachloro-1,4-cyclohexadiene hydrolase 1,4-TCDN chlorohydrolase	
Gene names	Name: linB	
Organism	Pseudomonas paucimobilis (Sphingomonas paucimobilis)	
Taxonomic identifier	13689 [NCBI]	
Taxonomic lineage	Bacteria › Proteobacteria › Alphaproteobacteria › Sphingomonadales › Sphingomonadaceae › Sphingomonas	
Protein attributes		Hide Top
Sequence length	296 AA.	
Sequence status	Complete.	
Sequence processing	The displayed sequence is further processed into a mature form.	
Protein existence	Evidence at protein level.	

Ukázka záznamu v UniProtKB

□ Obecná anotace

General annotation (Comments) Hide Top	
Function	Catalyzes hydrolytic cleavage of carbon-halogen bonds in halogenated aliphatic compounds, leading to the formation of the corresponding primary alcohols, halide ions and protons. Has a broad substrate specificity since not only monochloroalkanes (C3 to C10) but also dichloroalkanes (> C3), bromoalkanes, and chlorinated aliphatic alcohols were good substrates. Shows almost no activity with 1,2-dichloroethane, but very high activity with the brominated analog. Is involved in the degradation of the important environmental pollutant gamma-hexachlorocyclohexane (lindane) as it also catalyzes conversion of 1,3,4,6-tetrachloro-1,4-cyclohexadiene (1,4-TCDN) to 2,5-dichloro-2,5-cyclohexadiene-1,4-diol (2,5-DDOL) via the intermediate 2,4,5-trichloro-2,5-cyclohexadiene-1-ol (2,4,5-DNOL). HAMAP MF_01231
Catalytic activity	1-haloalkane + H ₂ O = a primary alcohol + halide. HAMAP MF_01231 1,4-TCDN + 2 H ₂ O = 2,5-DDOL + 2 chloride. HAMAP MF_01231
Enzyme regulation	Competitively inhibited by the key pollutants 1,2-dichloroethane (1,2-DCE) and 1,2-dichloropropane (1,2-DCP). HAMAP MF_01231
Pathway	Xenobiotic degradation; gamma-hexachlorocyclohexane degradation. HAMAP MF_01231
Subunit structure	Monomer. HAMAP MF_01231
Subcellular location	Periplasm. Ref.4
Induction	Constitutively expressed. HAMAP MF_01231
Miscellaneous	Is not N-terminally processed during export, so it may be secreted into the periplasmic space via a hitherto unknown mechanism. HAMAP MF_01231
Sequence similarities	Belongs to the haloalkane dehalogenase family, Type 2 subfamily.
Biophysicochemical properties	pH dependence: Optimum pH is 8.2. HAMAP MF_01231

Ukázka záznamu v UniProtKB

□ Ontologie

Ontologies	
Keywords	
Biological process	Detoxification
Cellular component	Periplasm
Molecular function	Hydrolase
Technical term	3D-structure Direct protein sequencing
Gene Ontology (GO)	
Biological process	response to toxin <small>Inferred from electronic annotation. Source: UniProtKB-KW</small>
Cellular component	periplasmic space <small>Inferred from electronic annotation. Source: UniProtKB-SubCell</small>
Molecular function	haloalkane dehalogenase activity <small>Inferred from electronic annotation. Source: HAMAP</small>
Complete GO annotation...	











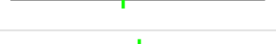




Ukázka záznamu v UniProtKB

□ Anotace sekvence

Sequence annotation (Features)						
	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
<input checked="" type="checkbox"/>	Initiator methionine	1	1	Removed (Ref.4) (Ref.1)		
<input type="checkbox"/>	Chain	2 – 296	295	Haloalkane dehalogenase (HAMAP MF_01231)		PRO_0000216778
Sites						
<input checked="" type="checkbox"/>	Active site	108	1	Nucleophile (HAMAP MF_01231)		
<input checked="" type="checkbox"/>	Active site	132	1	Proton donor (HAMAP MF_01231)		
<input checked="" type="checkbox"/>	Active site	272	1	Proton acceptor (HAMAP MF_01231)		
<input checked="" type="checkbox"/>	Binding site	38	1	Halide (HAMAP MF_01231)		
<input checked="" type="checkbox"/>	Binding site	109	1	Halide (HAMAP MF_01231)		
Natural variations						
<input checked="" type="checkbox"/>	Natural variant	81	1	A → T in strain: B90.		
<input checked="" type="checkbox"/>	Natural variant	112	1	A → V in strain: B90.		
<input checked="" type="checkbox"/>	Natural variant	134 – 135	2	IA → VT in strain: B90.		
<input checked="" type="checkbox"/>	Natural variant	138	1	I → L in strain: B90.		
<input checked="" type="checkbox"/>	Natural variant	247	1	A → H in strain: B90.		
<input checked="" type="checkbox"/>	Natural variant	253	1	M → I in strain: B90.		

Ukázka záznamu v UniProtKB

□ Anotace sekvence

Natural variations					
■	Natural variant	81	1	A → T in strain: B90.	
■	Natural variant	112	1	A → V in strain: B90.	
■	Natural variant	134 – 135	2	IA → VT in strain: B90.	
■	Natural variant	138	1	I → L in strain: B90.	
■	Natural variant	247	1	A → H in strain: B90.	
■	Natural variant	253	1	M → I in strain: B90.	
Experimental info					
■	Mutagenesis	38	1	N → D, E, F or Q: Loss of activity. Ref.7	
■	Mutagenesis	108	1	D → A: Loss of activity. Ref.6	
■	Mutagenesis	108	1	D → N: 58% of wild-type activity. Ref.6	
■	Mutagenesis	109	1	W → L: Loss of activity. Ref.7	
■	Mutagenesis	132	1	E → Q: Loss of activity. Ref.6	
■	Mutagenesis	151	1	F → L, W or Y: Increase in activity. Ref.7	
■	Mutagenesis	169	1	F → L: 31% of wild-type activity. Ref.7	
■	Mutagenesis	244	1	E → Q: 38% of wild-type activity. Ref.6	
■	Mutagenesis	272	1	H → A: Loss of activity. Ref.6	

Ukázka záznamu v UniProtKB

□ Sekvence

Sequences

Sequence	Length	Mass (Da)	Tools
<input type="checkbox"/> P51698-1 [UniParc]. FASTA	296	33,108	<input type="text" value="Blast"/> <input type="button" value="go"/>

Last modified January 23, 2007. Version 4.
Checksum: 6EEE011B157DBAE1

```
      10      20      30      40      50      60
MSLGAKPFGE KKFIEIKGRR MAYIDEGTGD PILFQHGMP T SSYLWRNIMP HCAGLGRLIA
      70      80      90     100     110     120
CDLIGHGDS D KLDPSGPERY AYAHRDYLD ALWEALDLGD RVVLVVHDWG SALGFDWARR
     130     140     150     160     170     180
HRERVQGIAY MEAIAMPIEW ADFPEQDRDL FOAFRSQAGE ELVLQDNV FV EQVLPGLILR
     190     200     210     220     230     240
PLSEAEMAAY REPFLAAGEA RRPTLSWPRQ IPIAGTPADV VAIARDYAGW LSESPIPKLF
     250     260     270     280     290
INAEPGALTT GRMRDFCRTW PNQTEITVAG AHFIQEDSPD EIGAAIAAFV RRLRPA
```

[« Hide](#)

Ukázka záznamu v UniProtKB

□ Reference

References		Hid
[1]	<p>"Cloning and sequencing of a dehalogenase gene encoding an enzyme with hydrolase activity involved in the degradation of gamma-hexachlorocyclohexane in <i>Pseudomonas paucimobilis</i>." Nagata Y., Nariya T., Ohtomo R., Fukuda M., Yano K., Takagi M. J. Bacteriol. 175:6403-6410(1993) [PubMed: 7691794] [Abstract] <u>Cited for:</u> NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 2-16. <u>Strain:</u> UT26.</p>	
[2]	<p>Nagata Y., Nariya T., Ohtomo R., Fukuda M., Yano K., Takagi M. Submitted (MAR-1999) to the EMBL/GenBank/DDBJ databases <u>Cited for:</u> SEQUENCE REVISION.</p>	
[3]	<p>"Cloning and characterization of lin genes responsible for the degradation of hexachlorocyclohexane isomers by <i>Sphingomonas paucimobilis</i> strain B90." Kumari R., Subudhi S., Suar M., Dhingra G., Raina V., Dogra C., Lal S., van der Meer J.R., Holliger C., Lal R. Appl. Environ. Microbiol. 68:6021-6028(2002) [PubMed: 12450824] [Abstract] <u>Cited for:</u> NUCLEOTIDE SEQUENCE [GENOMIC DNA]. <u>Strain:</u> B90.</p>	
[4]	<p>"Two different types of dehalogenases, LinA and LinB, involved in gamma-hexachlorocyclohexane degradation in <i>Sphingomonas paucimobilis</i> UT26 are localized in the periplasmic space without molecular processing." Nagata Y., Futamura A., Miyauchi K., Takagi M. J. Bacteriol. 181:5409-5413(1999) [PubMed: 10464214] [Abstract] <u>Cited for:</u> PROTEIN SEQUENCE OF 2-10, SUBCELLULAR LOCATION.</p>	
[5]	<p>"Purification and characterization of a haloalkane dehalogenase of a new substrate class from a gamma-hexachlorocyclohexane-degrading bacterium, <i>Sphingomonas paucimobilis</i> UT26." Nagata Y., Miyauchi K., Damborsky J., Manova K., Ansorgova A., Takagi M. Appl. Environ. Microbiol. 63:3707-3710(1997) [PubMed: 9293022] [Abstract] <u>Cited for:</u> CHARACTERIZATION. <u>Strain:</u> UT26.</p>	

Databáze proteinových sekvencí

□ nr Protein Database

- Databáze proteinových sekvencí NCBI
- Kolekce sekvencí získaných konceptuální translací kódujících oblastí GenBank/EMBL-Bank/DDBJ a dále sekvencí z UniProtKB, PRF a RCSB PDB
- ☹ většinou automatická anotace – **nižší kvalita, možnost chyb**
- ☹ chybí indikace původu anotace
- ☺ více než 20.000.000 sekvencí (2012)

Formáty sekvencí

- Prostá sekvence

```
DQLTEEQIAEFKEAFSLFDK
```

Formáty sekvencí

- Prostá sekvence
- GenBank

```
LOCUS      AAU03518      237 bp      DNA                PLN      04-FEB-1995
DEFINITION Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
           rRNA and 5.8S rRNA genes, partial sequence.
ACCESSION  U03518
BASE COUNT      41 a      77 c      67 g      52 t
ORIGIN
      1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc
     61 tattgtaccg tgttgcttcg gcgggcccgc cgcttgctcg ccgcccgggg ggcgcctctg
    121 cccccggggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
    181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc
//
```

Formáty sekvencí

- Prostá sekvence
- GenBank
- EMBL

```
ID   AA03518      standard; DNA; FUN; 237 BP.
XX
AC   U03518;
XX
DE   Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
DE   rRNA and 5.8S rRNA genes, partial sequence.
XX
SQ   Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other;
aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc          60
tattgtaccc tgttgcttcg gcgggcccgc cgcttgctcg ccgccggggg ggcgcctctg          120
ccccccgggc ccgtgccgcg cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc          180
tgagttgatt gaatgcaatc agttaaactt ttcaacaatg gatctcttgg ttccggc          237
//
```

Formáty sekvencí



- ❑ Prostá sekvence
- ❑ GenBank
- ❑ EMBL
- ❑ **FASTA**

➔ `>gi|155348|gb|AAA88691.1| haloalkane dehalogenase
MINAIRTPDQRFNSLDQYPFSPNYLDDLPGY PGLRAHYLDEGNSDAEDVF
AESGARVIAPDFFGFGKSDKPVDEEDYTFEFHRNFLLALIERLDRNITL
FKRLIIMNACLMTDPVTQPAESAFVTQPADGFTAWKYDLVTPSDLRLDQF
PDTSYQAGVRKFKPMVAQRDQACIDISTEAISEFWQNDWNGQTFMAIGMKD
PLEIADAGHFVQEFGEQVAREALKHFAETE`

Nejčastěji používaný formát

Databáze proteinových struktur

- Worldwide Protein Data Bank (wwPDB)
 - Světový depositář proteinových struktur, obsahuje rovněž struktury nukleových kyselin a biomolekulárních komplexů
 - Research Collaboratory for Structural Bioinformatics (RCSB PDB), Protein Data Bank Europe (PDBe), Protein Data Bank Japan (PDBj), Biological Magnetic Resonance Data Bank (BioMagResBank)
 - Obsahuje více než 84.000 struktur (2012)
 - Struktury získané rentgenovou krystalografií (88%) a nukleární magnetickou rezonancí (11 %)



Databáze genomů

- ❑ Entrez Genome
- ❑ Ensembl
- ❑ Genomes OnLine Database GOLD

- ❑ Informace o zdrojovém organizmu
- ❑ Nukleotidové a proteinové sekvence
- ❑ Geny v kontextu genomu
- ❑ Anotace a analýza genomů

Ukázka záznamu v Entrez Genome

☐ Přehled prokaryotických genomů

Overview [8348] Eukaryotes [2215] Prokaryotes [14202] Viruses [3212]

First Previous **Shown: 1 - 100 out of 1309 items** Next Last Download selected records

Organism/Name	BioProject	Group	SubGroup	Size (Mb)	GC%	Chromosomes		WGS	Scaffolds	Gene	Protein	Release Date	Modify Date	Status
		Actinobacte ▾	— All Actinc ▾			RefSeq	INSDC							All ▾
Acaricomes phytoseiuli DSM 14247	PRJNA174970	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data
Acidimicrobium ferrooxidans DSM 10331	PRJNA59215 PRJNA29525	Actinobacteria	Actinobacteria	2.16	68.30	NC_013124.1	CP001631.1	-	-	2089	1964	2009/08/18	2012/01/30	Complete
Acidothermus cellulolyticus 11B	PRJNA58501 PRJNA16097	Actinobacteria	Actinobacteria	2.44	66.90	NC_008578.1	CP000481.1	-	-	2217	2157	2006/11/09	2012/01/24	Complete
Actinoaloteichus spitiensis RMV-1378	PRJNA76807	Actinobacteria	Actinobacteria	5.71	72.40	-	-	AGVX01	-	-	-	2011/12/20	2012/05/31	Scaffolds or contigs
Actinobaculum sp. oral taxon 183 str. F0552	PRJNA173932	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data
Actinobaculum massiliae ACS-171-V-CoI2	PRJNA52091	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	SRA or Traces
Actinobaculum schaalii	PRJNA52093	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data

Ukázka záznamu v Entrez Genome

☐ Přehled prokaryotických genomů

Overview [8348] Eukaryotes [2215] Prokaryotes [14202] Viruses [3212]

First Previous **Shown: 1 - 100 out of 1309 items** Next Last Download selected records

Organism/Name	BioProject	Group	SubGroup	Size (Mb)	GC%	Chromosomes		WGS	Scaffolds	Gene	Protein	Release Date	Modify Date	Status
		Actinobacte ▾	— All Actinc ▾			RefSeq	INSDC							All ▾
Acaricomes phytoseiuli DSM 14247	PRJNA174970	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data
Acidimicrobium ferrooxidans DSM 10331	PRJNA59215 PRJNA29525	Actinobacteria	Actinobacteria	2.16	68.30	NC_013124.1	CP001631.1	-	-	2089	1964	2009/08/18	2012/01/30	Complete
Acidothermus cellulolyticus 11B	PRJNA58501 PRJNA16097	Actinobacteria	Actinobacteria	2.44	66.90	NC_008578.1	CP000481.1	-	-	2217	2157	2006/11/09	2012/01/24	Complete
Actinoalloteichus spitiensis RMV-1378	PRJNA76807	Actinobacteria	Actinobacteria	5.71	72.40	-	-	AGVX01	-	-	-	2011/12/20	2012/05/31	Scaffolds or contigs
Actinobaculum sp. oral taxon 183 str. F0552	PRJNA173932	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data
Actinobaculum massiliae ACS-171-V-CoI2	PRJNA52091	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	SRA or Traces
Actinobaculum schaalii	PRJNA52093	Actinobacteria	Actinobacteria	-	-	-	-	-	-	-	-	-	-	No data

Ukázka záznamu v Entrez Genome

□ Informace o genomu

[Organism Overview](#) ; [Genome Project Report](#) ; [Genome Annotation Report](#)

Acidothermus cellulolyticus 11B

Thermotolerant cellulolytic organism

Lineage: [Bacteria\[3351\]](#); [Actinobacteria\[547\]](#); [Actinobacteria\[547\]](#); [Actinobacteridae\[502\]](#); [Actinomycetales\[485\]](#); [Frankineae\[11\]](#); [Acidothermaceae\[1\]](#); [Acidothermus\[1\]](#); [Acidothermus cellulolyticus\[1\]](#); [Acidothermus cellulolyticus 11B\[0\]](#)

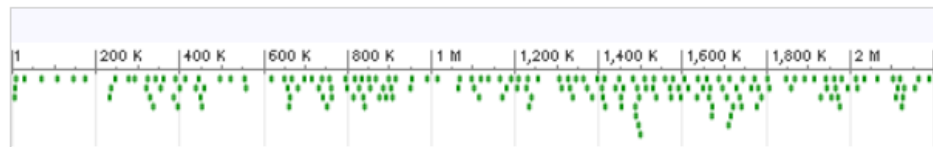
Acidothermus cellulolyticus strain 11B. This strain (11B; ATCC 43068) is the type strain for the species. The genome sequence from this organism will provide information on the regulation and production of potentially useful enzymes.

▾ Genome Sequencing Projects

◆ Chromosomes [1] ◆ Scaffolds or contigs [0] ◆ SRA or Traces [0] ◆ No data [0]

Organism	BioProject	Assembly	Status	Chrs	Size (Mb)	GC%	Gene	Protein
Acidothermus cellulolyticus 11B	PRJNA58501, PRJNA16097	ASM1502v1	◆	1	2.44	66.9	2,217	2,157

▾ Genome Region



[Go to nucleotide](#) [Graphics](#) [FASTA](#) [GenBank](#)

Ukázka záznamu v Entrez Genome

□ Informace o genomu

[Organism Overview](#) ; [Genome Project Report](#) ; [Genome Annotation Report](#)

Acidothermus cellulolyticus 11B

Thermotolerant cellulolytic organism

Lineage: [Bacteria\[3351\]](#); [Actinobacteria\[547\]](#); [Actinobacteria\[547\]](#); [Actinobacteridae\[502\]](#); [Actinomycetales\[485\]](#); [Frankineae\[11\]](#); [Acidothermaceae\[1\]](#); [Acidothermus\[1\]](#); [Acidothermus cellulolyticus\[1\]](#); [Acidothermus cellulolyticus 11B\[0\]](#)

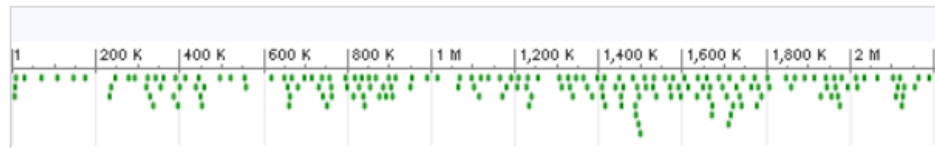
Acidothermus cellulolyticus strain 11B. This strain (11B; ATCC 43068) is the type strain for the species. The genome sequence from this organism will provide information on the regulation and production of potentially useful enzymes.

▾ Genome Sequencing Projects

◆ Chromosomes [1] ◆ Scaffolds or contigs [0] ◆ SRA or Traces [0] ◆ No data [0]

Organism	BioProject	Assembly	Status	Chrs	Size (Mb)	GC%	Gene	Protein
Acidothermus cellulolyticus 11B	PRJNA58501, PRJNA16097	ASM1502v1	◆	1	2.44	66.9	2,217	2,157

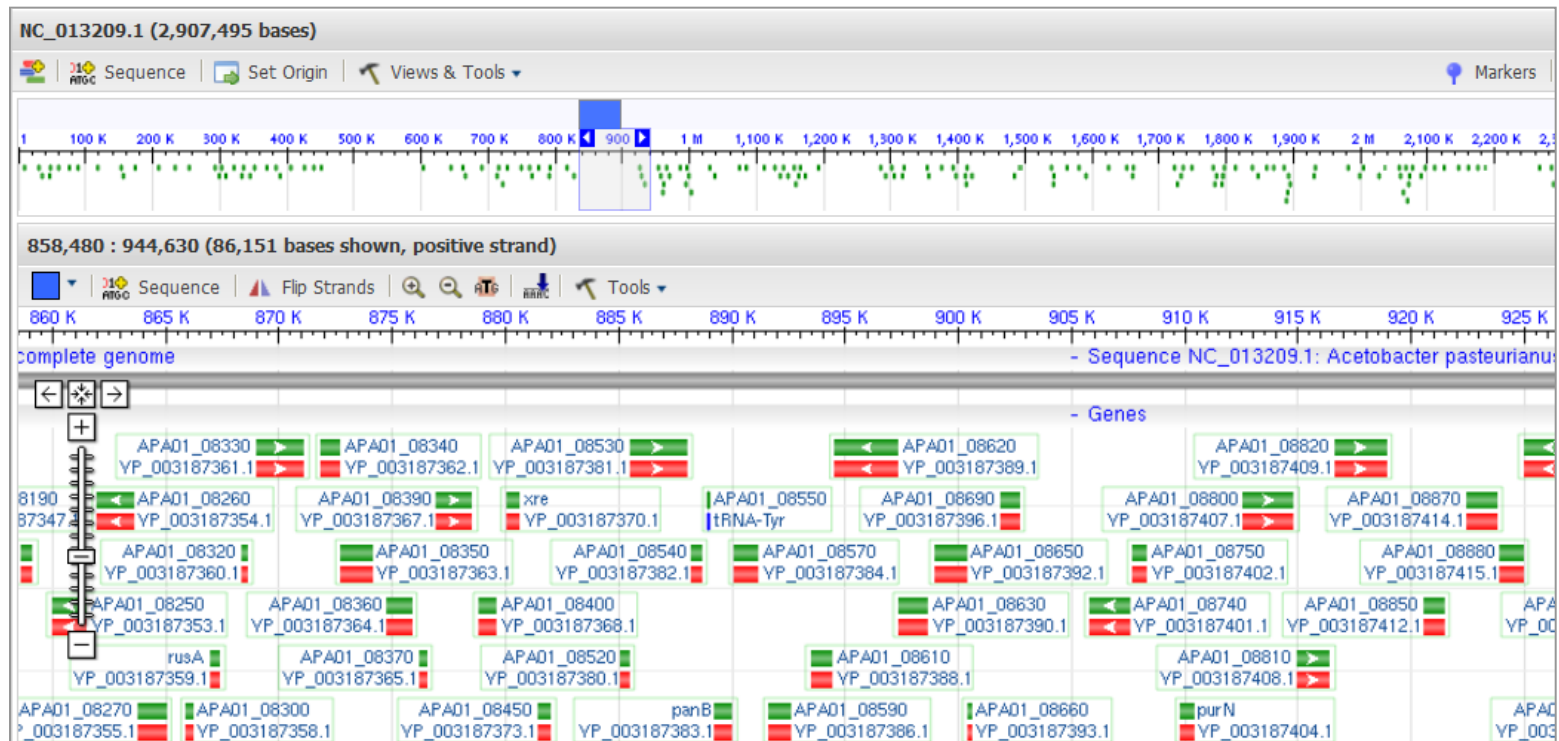
▾ Genome Region



Go to nucleotide [Graphics](#) [FASTA](#) [GenBank](#)

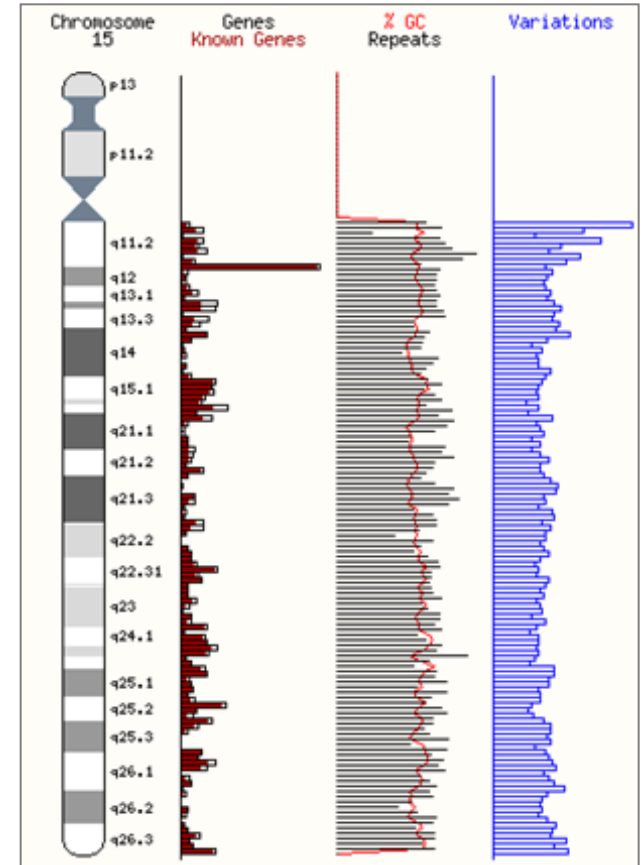
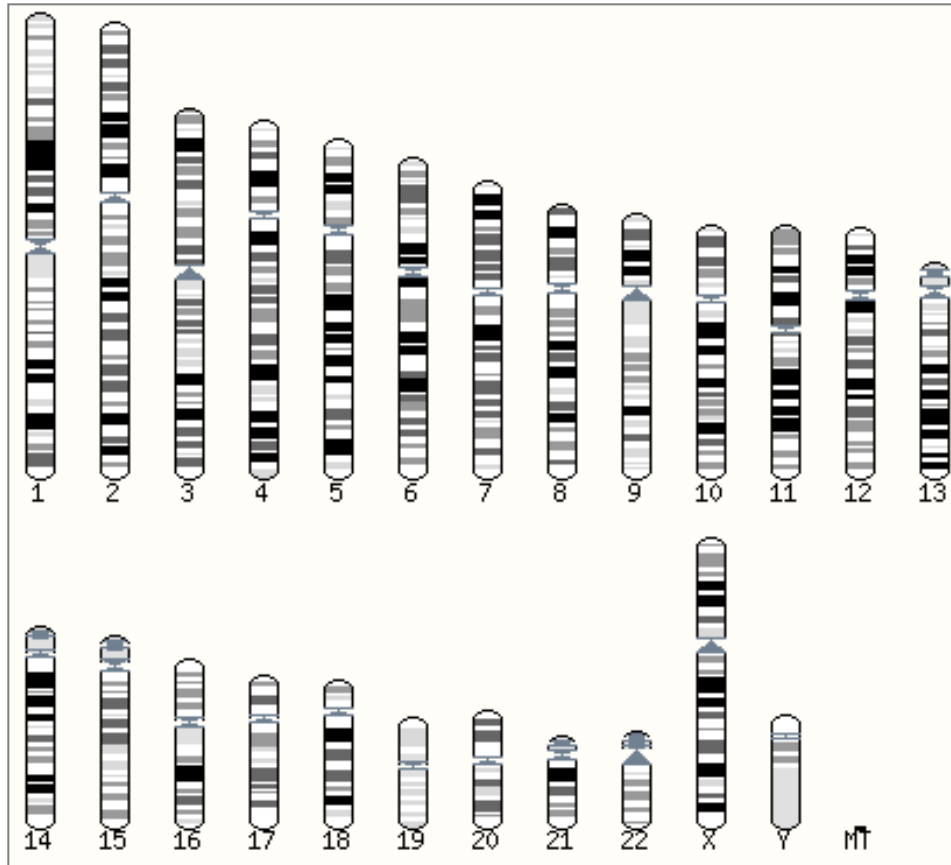
Ukázka záznamu v Entrez Genome

- Grafické znázornění genomu



Ukázka záznamu v Ensembl

- Anotace eukaryotických genomů



Ukázka záznamu v GOLD

Seznam dokončených a probíhajících genomových projektů

Complete Published Genome Projects: 1375												
A Archaeal: 94 B Bacterial: 1148 E Eukaryal: 133												
<< first < prev 1 2 3 4 5 next > last >> 100												
GOLD ID	ORGANISM	DOMAIN	INFORMATION	SIZE	CHROM #	PLASM #	GC %	DATA	SEQUENCING CENTER	GENOME DATABASE	PUBLICATION	CONTACT
Gc01325	Saccharomyces cerevisiae S288c	E	FUNGI-ASCOMYCOTA Taxonomy Entrez		16				Saccharomyces Genome Database			
Gc01378	Ferrimonas balearica PAT, DSM 9799	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez GEBA	4279 Kb 3947 orfs	1		50%	CP002209	DOE Joint Genome Institute DSMZ	IMG-GEBA	Unpublished 2010-09-24	Wovke T
Gc01374	Vulcanisaeta distributa IC-017, DSM 14429	A	CRENARCHAEOTA-THERMOPROTEI Taxonomy Entrez Isolation GEBA	2374 Kb 2592 orfs MAP	1		45.4%	NC_014537	DOE Joint Genome Institute DSMZ	NCBI	Unpublished 2010-09-22	Bruce D
Gc01375	Halomonas elongata DSM 2581	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez	4061 Kb 3556 orfs MAP	1		63%	NC_014532	Max-Planck Institute	NCBI	Environmental Microbiology in press 2010-09-17	Kunte,H.J
Gc01372	Methanoplanus petrolearius SEBR 4847, DSM 11571	A	EURYARCHAEOTA-METHANOMICROBIA Taxonomy Entrez Isolation GEBA	2843 Kb 2881 orfs MAP	1		50%	NC_014507	DOE Joint Genome Institute DSMZ	NCBI	Unpublished 2010-09-17	Bruce D
Gc01373	Sulfurimonas autotrophica OK10, DSM 16294	B	PROTEOBACTERIA-EPSILON Taxonomy Entrez Isolation	2153 Kb 2220 orfs	1		35.2%	CP002205	DOE Joint Genome Institute DSMZ	IMG-GEBA	2010-09-15	Bruce D
Gc01376	Spirochaeta thermophila DSM 6192	B	SPIROCHAETES Taxonomy Entrez	2472 Kb	1		52%	CP001698	Goettingen Genomics Laboratory		Unpublished 2010-09-10	Liebl,W
Gc01377	Dickeya dadantii 3937	B	PROTEOBACTERIA-GAMMA Taxonomy Entrez Plant Pathogen Article	4922 Kb	1			CP002038	J. Craig Venter Institute Univ of Wisconsin	Univ of Wisconsin	Unpublished 2010-09-10	Perna N

PROJECT TYPE DISTRIBUTION

SEQUENCING STATUS DISTRIBUTION

PHYLOGENETIC DISTRIBUTION

Problémy bioinformatických databází

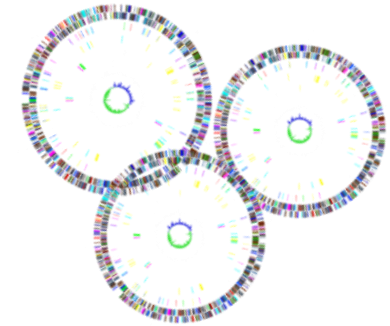


- ❑ Vysoká **redundance** dat
- ❑ **Chyby** v sekvencích
- ❑ **Chyby** a nepřesnosti v anotacích
- ❑ Propagace **chyb** během automatických anotací

Prohledávání databází

- ❑ Textové prohledávání
- ❑ Sekvenční prohledávání

vložení dat



prohledání databáze

NCBI Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases GO CLEAR Help

44		PubMed: biomedical literature citations and abstracts	?	none		Books: online books
79		PubMed Central: free, full text journal articles	?	none		OMIM: online Mendelian Inheritance
none		Site Search: NCBI web and FTP sites	?	none		OMIA: Online Mendelian Inheritance
45		Nucleotide: sequence database (includes GenBank)	?	none		UniGene: gene-oriented clusters of sequences
39		Protein: sequence database	?	none		CDD: conserved protein domain data
4		Genome: whole genome sequences	?	12		3D Domains: domains from Entrez
12		Structure: three-dimensional macromolecular structures	?	none		UniSTS: markers and mapping data

1: [ABI93216](#). Report: LinB [Xanthomonas...[gi:115291795]]
>gi|115291795|gb|ABI93216.1| LinB [Xanthomonas sp. ICHL2]
MILGANAFCENKFTIEIKGRFMYIDEFTGDPILFQHGNETS SYLWRNIMPHCA GLGRLIACDLI GMGDS D
KLDPS GPERYTYAERHRYLDALMEALDLGDRVLAWHDQGSVLGFTDARRRHEEQQG IAYMEAVTMPLD
ADFFEQERDLFQAFPS QAGEELVLDQDFVEQLPLGLILRPLSEAEYREPLAA GEARRFTL 30PRQ
IP IAGTPADVAIARDVA GGLSESPIPKLFINAEPGLITGRIEDFCRTMPNQTETITVAGAFI QEDSPD
EIG&AIA&FVRE

2: [AAR05978](#). Report: LinB [Spingomonas...[gi:37963683]]
>gi|37963683|gb|AAR05978.1| LinB [Spingomonas paucimobilis]
MSLGAHPFCENKFTIEIKGRFMYIDEFTGDPILFQHGNETS SYLWRNIMPHCA GLGRLIACDLI GMGDS D
KLDPS GPERYTYAERHRYLDALMEALDLGDRVLAWHDQGSVLGFTDARRRHEEQQG IAYMEAVTMPLD
ADFFEQERDLFQAFPS QAGEELVLDQDFVEQLPLGLILRPLSEAEYREPLAA GEARRFTL 30PRQ
IP IAGTPADVAIARDVA GGLSESPIPKLFINAEPGLITGRIEDFCRTMPNQTETITVAGAFI QEDSPD
EIG&AIA&FVRE

Textové prohledávání databází



- ❑ Entrez
- ❑ SRS
- ❑ DBGET


- ❑ **Integrují** data z různých databází
- ❑ Umožňují prohledání mnoha databází **současně**
- ❑ Umožňují formulaci dotazů pomocí **logických operátorů**

Textové prohledávání databází















□ Entrez

- Vyhledávací systém pro databáze **NCBI**
- Integruje data ze 40 různých databází, **pouze NCBI**

NCBI  *Entrez, The Life Sciences Search Engine*

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Help

44		PubMed: biomedical literature citations and abstracts	<input data-bbox="927 921 956 942" type="button" value="?"/>	none		Books: online books	<input data-bbox="1593 921 1622 942" type="button" value="?"/>
79		PubMed Central: free, full text journal articles	<input data-bbox="927 978 956 999" type="button" value="?"/>	none		OMIM: online Mendelian Inheritance in Man	<input data-bbox="1593 978 1622 999" type="button" value="?"/>
none		Site Search: NCBI web and FTP sites	<input data-bbox="927 1035 956 1056" type="button" value="?"/>	none		OMIA: Online Mendelian Inheritance in Animals	<input data-bbox="1593 1035 1622 1056" type="button" value="?"/>
45		Nucleotide: sequence database (includes GenBank)	<input data-bbox="927 1120 956 1142" type="button" value="?"/>	none		UniGene: gene-oriented clusters of transcript sequences	<input data-bbox="1593 1120 1622 1142" type="button" value="?"/>
39		Protein: sequence database	<input data-bbox="927 1178 956 1199" type="button" value="?"/>	none		CDD: conserved protein domain database	<input data-bbox="1593 1178 1622 1199" type="button" value="?"/>
4		Genome: whole genome sequences	<input data-bbox="927 1235 956 1256" type="button" value="?"/>	12		3D Domains: domains from Entrez Structure	<input data-bbox="1593 1235 1622 1256" type="button" value="?"/>

Textové prohledávání databází



□ SRS

- Vyhledávací systém pro databáze **EBI**
- Umožňuje prohledávat i databáze **jiných** institucí, databáze **uživatelů** či databáze **výsledků** vybraných výpočetních nástrojů

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks

SRS

[Start a Permanent Project](#)

Quick Text Search Search Tips

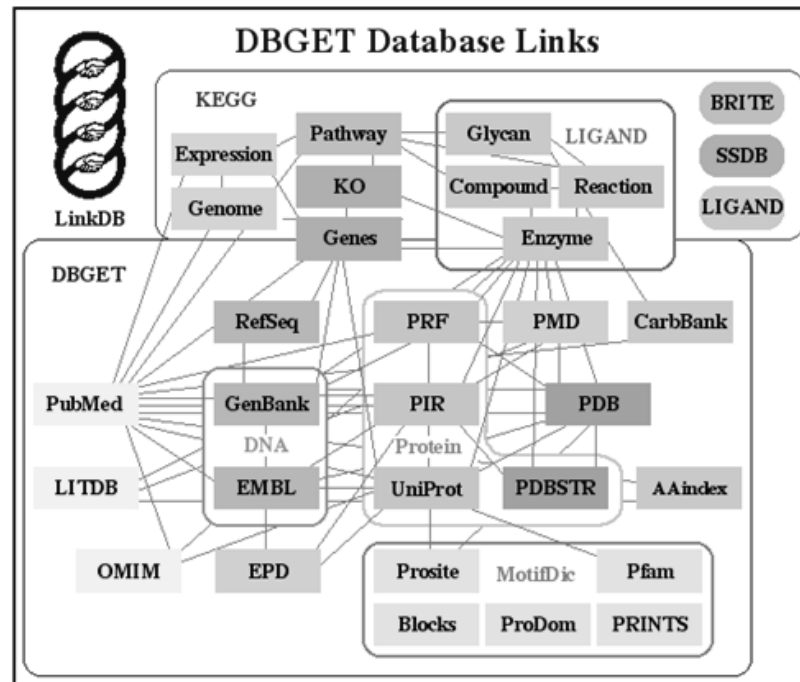
Find : Nucleotides matching : Enter Text Here

Textové prohledávání databází



□ DBGET

- Vyhledávací systém pro databáze **LinkDB**
- Umožňuje mj. prohledávat databázi **metabolických drah KEGG**

















Ukázka textového prohledávání

□ Vyhledávání na základě klíčových slov

Search across databases [Help](#)

■ - Result counts displayed in gray indicate one or more terms not found

1258	 PubMed: biomedical literature citations and abstracts	30	 Books: online books
312	 PubMed Central: free, full text journal articles	703	 OMIM: online Mendelian Inheritance in Man
4	 Site Search: NCBI web and FTP sites	none	 OMIA: online Mendelian Inheritance in Animals
152	 Nucleotide: Core subset of nucleotide sequence records	none	 dbGaP: genotype and phenotype
1	 EST: Expressed Sequence Tag records	1	 UniGene: gene-oriented clusters of transcript sequences
12	 GSS: Genome Survey Sequence records	none	 CDD: conserved protein domain database
96	 Protein: sequence database	none	 3D Domains: domains from Entrez Structure

Ukázka sekvenčního prohledávání

□ Vyhledávání na základě sekvenční podobnosti

```
>gb|AAT70109.1 CurN [Lyngbya majuscula]  
Length=341
```

```
Score = 303 bits (777), Expect = 8e-81, Method: Composition-based stats.  
Identities = 148/297 (49%), Positives = 188/297 (63%), Gaps = 8/297 (2%)
```

```
Query 2 SEIGTGFPPDPHYVEVLGERMHYVDVGRDGTPLVFLHGNPTSSYLWRNIIPHV-APSHR 60  
I + FPF VEV G + YVD G G PVLFLHGNPTSSYLWRNIIPV A +R  
Sbjct 41 LPISSEFPFAKRTVEVEGATIAYVDEG--SGQPVLFLHGNPTSSYLWRNIIPYVVAAGYR 98  
  
Query 61 CIAPDLIGMGKSDKPDLDYFFDDHVRYLDAFIEALGLEEVVLVIHDWGSALGFHWAKRNP 120  
+APDLIGMG S KPD++Y DHV Y+D FI+ALGL+++VLVIHDWGS +G A+ NP  
Sbjct 99 AVAPDLIGMGDSAKPDIEYRLQDHVAYMDGFIDALGLDDMLVVIHDWGSVIGMRHARLNP 158  
  
Query 121 ERVKGCIACMEFIRPI----PTWDEWPEFARETFQAFRTADVGRELIIDQNAFIEGVLPK- 175  
+RV +A ME + P P+++ F+ RTADVG ++++D N F+E +LP+  
Sbjct 159 DRVAAVAEMEALVPPALPMPYSYAMGPQLGLPLFRDLRTADVGEKMLVDGNFFVETILPEM 218  
  
Query 176 CVVRPLTEVEMDHYREPFLLKPVDPREPLWRFNEIPIAGEPANIVALVEAYMNWLHQSPVP 235  
VVR L+E EM YR PF R P ++P E+PI GEPA A V WL SP+P  
Sbjct 219 GVVRSLSEAEEMAAYRAPFPTRQSRPLTLQWPREVPIGGEPAFAEAEVLKNGEWLMA SPIP 278  
  
Query 236 KLLFWGTPGVLIPPAEAAARLAESLPNCKTVDIGPGLHYLQEDNPDLIGSEIARWLPG 292  
KLLF PG L P L+E++PN + +G G H+LQED+P LIG IA WL  
Sbjct 279 KLLFHAEIPGALAPKPVVDYLSENVNLEVRVFGAGTHFLQEDHPHLIGQGIADWLR 335
```

Sequences producing significant alignments:

		Score (Bits)	E Value
sp P59336 DHAA_RHOD	Haloalkane dehalogenase >pdb 1BN6 A Chai...	429	1e-118
sp POA3G2 DHAA_RHORH	Haloalkane dehalogenase >sp POA3G3 DHAA_...	424	3e-117
pdb 1CQW A	Chain A, Nai CocrySTALLISED With Haloalkane Dehalo...	424	4e-117
sp Q9ZER0 DHAA_MYCSX	Haloalkane dehalogenase >emb CAA10076.1 ...	422	1e-116
gb AAV70825.1	HT2 [Expression vector pHT2]	415	1e-114
ref YP_001675030.1	alpha/beta hydrolase fold [Shewanella hal...	320	8e-86
ref YP_734675.1	alpha/beta hydrolase fold [Shewanella sp. MR...	318	3e-85
ref YP_001473250.1	alpha/beta hydrolase fold [Shewanella sed...	317	6e-85
ref ZP_01736514.1	alpha/beta hydrolase [Marinobacter sp. ELB...	317	6e-85
ref YP_738656.1	alpha/beta hydrolase fold [Shewanella sp. MR...	316	9e-85
ref YP_001502590.1	alpha/beta hydrolase fold [Shewanella pea...	316	9e-85
ref NP_717353.1	hydrolase, alpha/beta hydrolase fold family ...	315	2e-84
ref YP_750057.1	alpha/beta hydrolase fold [Shewanella frigid...	315	2e-84
ref YP_268879.1	hydrolase, alpha/beta hydrolase fold family ...	315	2e-84
ref YP_001761524.1	alpha/beta hydrolase fold [Shewanella woo...	315	3e-84
ref ZP_01841154.1	alpha/beta hydrolase fold [Shewanella bait...	315	3e-84
ref YP_870347.1	alpha/beta hydrolase fold [Shewanella sp. AN...	314	4e-84
ref YP_129676.1	putative haloalkane dehalogenase [Photobacte...	314	7e-84
ref ZP_01221858.1	putative haloalkane dehalogenase [Photobac...	313	8e-84
ref YP_001365757.1	alpha/beta hydrolase fold [Shewanella bal...	313	9e-84
ref YP_562379.1	alpha/beta hydrolase fold [Shewanella denitr...	313	9e-84
ref ZP_01897865.1	putative haloalkane dehalogenase [Moritell...	313	1e-83
ref YP_001049934.1	alpha/beta hydrolase fold [Shewanella bal...	313	1e-83
ref YP_943362.1	alpha/beta hydrolase fold [Psychromonas ingr...	313	1e-83
ref YP_001182970.1	alpha/beta hydrolase fold [Shewanella put...	312	2e-83
ref YP_001554014.1	alpha/beta hydrolase fold [Shewanella bal...	312	2e-83
ref ZP_01706252.1	alpha/beta hydrolase fold [Shewanella putr...	310	7e-83
ref YP_964030.1	alpha/beta hydrolase fold [Shewanella sp. W3...	310	9e-83
ref YP_510562.1	haloalkane dehalogenase [Jannaschia sp. CCS1...	308	3e-82
ref ZP_01216824.1	hydrolase, alpha/beta hydrolase fold famil...	307	8e-82
ref YP_001093840.1	alpha/beta hydrolase fold [Shewanella loi...	306	1e-81
ref NP_106032.1	haloalkane dehalogenase [Mesorhizobium loti ...	303	8e-81
gb AAT70109.1	CurN [Lyngbya majuscula]	303	8e-81
ref ZP_01055470.1	haloalkane dehalogenase [Roseobacter sp. M...	303	1e-80
ref ZP_01617455.1	haloalkane dehalogenase [marine gamma prot...	302	2e-80
ref ZP_01592200.1	alpha/beta hydrolase fold [Geobacter lovle...	300	7e-80
ref ZP_01911259.1	alpha/beta hydrolase [Plesiocystis pacific...	300	9e-80
ref YP_001230772.1	alpha/beta hydrolase fold [Geobacter uran...	300	9e-80

Problémy prohledávání databází



□ **Textové** vyhledávání

- ☹ chybné, nepřesné či obecné anotace
- ☹ synonyma
- ☹ velký počet falešně pozitivních a falešně negativních výsledků

□ **Sekvenční** vyhledávání

- ☹ podmínka evoluční příbuznosti
- ☹ větší nároky na uživatele
- ☹ potenciálně falešně pozitivní i falešně negativní výsledky

□ Vhodné je oba přístupy **kombinovat**

Reference

- ❑ Claverie, J-M., & Notredame, C. (2006) **Bioinformatics for Dummies** (2nd ed.) Wiley Publishing, Hoboken, p. 436.
- ❑ Xiong, J. (2006) **Essential Bioinformatics**, Cambridge University Press, New York, p. 352.

- ❑ **ENTREZ tutorial:** <http://www.ncbi.nlm.nih.gov/Entrez/tutor.html>
- ❑ **SRS documentation:** <http://srs.ebi.ac.uk/srs/doc/index.html>
- ❑ **NCBI handbook:** <http://www.ncbi.nlm.nih.gov/books/NBK21101/>
- ❑ **UniProtKB manual:** <http://www.uniprot.org/manual/>

- ❑ **NCBI:** <http://www.ncbi.nlm.nih.gov/>
- ❑ **EBI:** <http://www.ebi.ac.uk/>
- ❑ **Pubmed:** <http://www.ncbi.nlm.nih.gov/pubmed>
- ❑ **Web of Science:** <http://apps.isiknowledge.com>

Reference

- ❑ **GenBank:** <http://www.ncbi.nlm.nih.gov/genbank/>
- ❑ **EMBL-EBI:** <http://www.ebi.ac.uk/embl/>
- ❑ **DDBJ:** <http://www.ddbj.nig.ac.jp/>
- ❑ **UniProt:** <http://www.uniprot.org/>
- ❑ **nrdb:** <http://www.ncbi.nlm.nih.gov/protein/>
- ❑ **wwPDB:** <http://www.wwpdb.org/>
- ❑ **Entrez Genome:** <http://www.ncbi.nlm.nih.gov/genome>
- ❑ **Ensembl:** <http://www.ensembl.org>
- ❑ **GOLD:** <http://www.genomesonline.org/>
- ❑ **Entrez:** <http://www.ncbi.nlm.nih.gov/sites/gquery>
- ❑ **SRS:** <http://srs.ebi.ac.uk>
- ❑ **DBGET:** <http://www.genome.jp/dbget/>