

Blok 3

Funkce proteinů

C3211 Aplikovaná bioinformatika
Přednášející: Josef Houser



Funkce proteinů

- **Enzymy** – 6 hlavních tříd
- **Strukturní proteiny** – keratin, kolagen
- **Transportní proteiny** – albumin, hemoglobin
- **Obranné proteiny** – protilátky
- **Regulátory a receptory** – hormony, transkripční faktory, rhodopsin
- ...



Klasifikace enzymů

Dle IUBMB: <http://www.chem.qmul.ac.uk/iubmb/enzyme/>

	Třída	Charakteristika	Počet Podtříd
1	Oxidoreduktasy	Katalyzují různé redoxní reakce – přenos vodíku, kyslíku, elektronů (obvykle s využitím koenzymů, např. NADH, NADPH, FADH ₂ nebo hemu)	23
2	Transferasy	Katalyzují přenos skupin: amino-, metyl-, acyl-, glykosyl-, fosforyl-	10
3	Hydrolasy	Katalyzují hydrolytické štěpení vazeb mezi atomem uhlíku a jinými atomy (spotřeba vody H ₂ O)	13
4	Lyasy	Katalyzují adiční reakci na dvojnou vazbu nebo eliminační reakci mezi 2 atomy uhlíku za vzniku dvojné vazby	7
5	Isomerasy	Katalyzují racemizaci optických izomerů nebo vytvoření polohových izomerů	6
6	Ligasy	Katalyzují tvorbu vazeb mezi uhlíkem a jinými atomy spojenou se štěpením ATP	6

Transportní proteiny

Dle TCDB (transporter classification database):

<http://www.chem.qmul.ac.uk/iubmb/mtp/>

	Třída
1	Póry a kanály
2	Přenašeče řízené elektrochemickým potenciálem
3	Přenašeče řízené chemickou reakcí
4	Skupinové přenašeče
5	Transmembránové elektronové přenašeče
6	Nepřiřazeno
7	Nepřiřazeno
8	Accessory factors involved in transport
9	Nedostatečně charakterizované transportní systémy

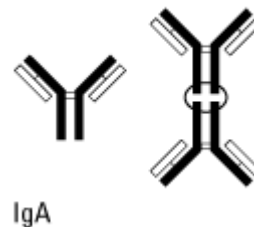
Protilátky

	Subtypů (člověk)	Charakteristika
IgA	2	Monomer/dimer, přítomna ve slinách, slzách (15%)
IgD	1	Monomer, funkce neznámá (0,2%)
IgE	1	Monomer, obrana proti parazitům, význam pro alergické reakce (0,002%)
IgG	4	Monomer, hlavní lidská protilátka v sekundární imunitní odpovědi (75%)
IgM	1	Pentamer, hlavní protilátka v primární imunitní odpovědi (10%)

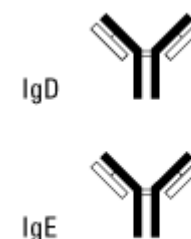
IgG



IgM



IgA



IgE

Určení funkce proteinu

- **Experimentální**

- Izolace proteinu s konkrétní funkcí
- Stanovení funkce u konkrétního proteinu



- **Predikce** – na základě podobnosti

- Lokalizace
- Multiple sequence alignment (BLAST, Pfam)
- Struktura molekuly (ProFunc, Catalytic site atlas)
- Textové hledání v publikacích (STRING)



Určení funkce proteinu

- Nezavrhujte jednoduchá řešení

Jakou funkci má tento protein?



Taq DNA polymerase



Predikce funkce proteinu

Nutno znát **sekvenci**

- Databáze
- Sekvenace

Lépe znát **strukturu** (2D, 3D)

- Databáze
- Určení 2D struktury viz. předchozí blok
- Určení 3D struktury viz. následující blok

Databáze strukturních a funkčních motivů

- Neanotované, nerevidované – „slepé“
přebírání dat
- Anotované, revidované – probíhá kontrola
vkládaných dat
- Obsahují různé informace – sekvenční,
strukturní, odkazy na experimentální data,...
- Slouží jako zdroj informací pro nadstavbové
programy.

Databáze strukturních a funkčních motivů

Často navzájem provázané. Např.:

- **UniProtKB** – kombinovaná proteinová databáze, vč. biologických dat
- **Pfam** – odvozená z UniProtKB
- **KEGG** – složená databáze obsahující systémové, genomické a chemické informace
- **CDD** – proteinové domény a další data
- ...



Kombinace několika databází

Vyhledávání pomocí klíčových slov i pomocí sekvence

A screenshot of the UniProt search interface. At the top, there are navigation tabs: Search, Blast, Align, Retrieve, and ID Mapping. Below these, there is a "Search in" dropdown menu currently set to "Protein Knowledgebase (UniProtKB)". To the right of this is a "Query" input field. Further right are three buttons: "Search", "Advanced Search »", and "Clear".

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- [Text search](#)
- [Sequence similarity searches \(BLAST\)](#)
- [Sequence alignments](#)
- [Batch retrieval](#)

NEWS

UniProt release 2013_02 - Feb 6, 2013

The smoke's devils | Cross-references to mycoCLAP

- › [Statistics for UniProtKB:](#)
 - [Swiss-Prot](#) · [TrEMBL](#)
 - › [Forthcoming changes](#)
 - › [News archives](#)

Follow @uniprot 521 followers

SITE TOUR



Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT


Úloha

- Vyhledejte homologní proteiny k následující sekvenci pomocí Blastu na serveru **UniProt**. Pokuste se na základě výsledku určit funkci tohoto proteinu.

```
SHLSQPWPITCFADRPTPRRSSPDASGQTMHSVFVVHVPYPVVFLKPAH  
LTPQWYRHPIPVNPVVRQPHLPVLYPAPNAGHTPAHSRQGDAALQPLF  
SVPQTVNPTGPVIHGDVAKQKPDTGQSWALNPYCTENWRRILRISRNS  
HGQRMPLTTLLQKTSGRNATLITKNSDQNTTTSIVSESSMTISACCHSAIL  
RNN
```

Graphical overview

Color code for identity 0-100% =



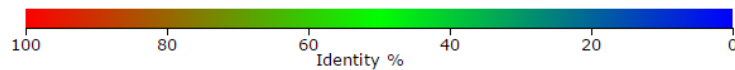
Accession	Entry name	0Query hit199	0Match hit (sqrt scale)3105	Name (Organism)
<input type="checkbox"/> Query				
2013031826IMB16GAX				
<input type="checkbox"/> Q8CLU5	Q8CLU5_YERPE			Uncharacterized protein (Yersinia pestis)
<input type="checkbox"/> J2E0Y8	J2E0Y8_KLEPN			Uncharacterized protein (Klebsiella pneumoniae subsp. pneumoni...)
<input type="checkbox"/> I1BJU5	I1BJU5_RHIO9			Uncharacterized protein (Rhizopus delemar (strain RA 99-880 / ...))
<input type="checkbox"/> G6HG35	G6HG35_9ACTO			Putative uncharacterized protein (Frankia sp. CN3)
<input type="checkbox"/> F9VC24	F9VC24_LACGL			Putative uncharacterized protein (Lactococcus garvieae (strain Lg2))
<input type="checkbox"/> F9V737	F9V737_LACGT			Putative uncharacterized protein (Lactococcus garvieae (strain ATCC 491...))
<input type="checkbox"/> B4NC70	B4NC70_DROWI			GK25804 (Drosophila willistoni)
<input type="checkbox"/> I1RMT3	I1RMT3_GIBZE			Uncharacterized protein (Gibberella zeae (strain PH-1 / ATCC M...))
<input type="checkbox"/> E7FCT7	E7FCT7_DANRE			Uncharacterized protein (Danio rerio)
<input type="checkbox"/> Q9RS36	Q9RS36_DEIRA			Cell wall glycyl-glycine endopeptidas... (Deinococcus radiodurans (strain ATCC ...))
<input type="checkbox"/> G7L8Y9	G7L8Y9_MEDTR			Extensin-like protein (Medicago truncatula)
<input type="checkbox"/> A9XEB8	A9XEB8_SACOF			27kD gamma canein (Saccharum officinarum)
<input type="checkbox"/> G2R972	G2R972_THITE			Putative uncharacterized protein (Thielavia terrestris (strain ATCC 380...))
<input type="checkbox"/> I8SZ67	I8SZ67_9LACT			Uncharacterized protein (Lactococcus garvieae IPLA 31405)
<input type="checkbox"/> K0SQJ7	K0SQJ7_THAOC			Uncharacterized protein (Thalassiosira oceanica)
<input type="checkbox"/> G7JXY3	G7JXY3_MEDTR			Putative uncharacterized protein (Medicago truncatula)
<input type="checkbox"/> A7S8E1	A7S8E1_NEMVE			Predicted protein (Nematostella vectensis)
<input type="checkbox"/> Q2GXL2	Q2GXL2_CHAGB			Putative uncharacterized protein (Chaetomium globosum (strain ATCC 6205...))
<input type="checkbox"/> Q3ZB05	Q3ZB05_MOUSE			Acr protein (Mus musculus)
<input type="checkbox"/> Q3ZB06	Q3ZB06_MOUSE			Acr protein (Mus musculus)
<input type="checkbox"/> P23578	ACRO_MOUSE			Acrosin (Mus musculus)
<input type="checkbox"/> Q6PAA4	Q6PAA4_XENLA			MGC68472 protein (Xenopus laevis)
<input type="checkbox"/> Q3KQB9	Q3KQB9_XENLA			MGC68472 protein (Xenopus laevis)
<input type="checkbox"/> A7XYK4	A7XYK4_TETNG			Jxc1-B (Tetraodon nigroviridis)
<input type="checkbox"/> H3C1M2	H3C1M2_TETNG			Uncharacterized protein (Tetraodon nigroviridis)
<input type="checkbox"/> H3D438	H3D438_TETNG			Uncharacterized protein (Tetraodon nigroviridis)

UniProt - New style

(BLAST) finds regions of local similarity between sequences, which can be used to infer relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

[Help](#) [Blast help video](#) [Other tutorials and videos](#) [Downloads](#)



[Edit and resubmit](#) Order by: **Score**

Overview

[Show all 78](#)

Entry	Protein names	Match hit	Identity
Q8CLU5	Uncharacterized protein (Yersinia pestis)		100.0%
J2E0Y8	Uncharacterized protein (Klebsiella pneumoniae subsp...)		66.0%
I1BJU5	Uncharacterized protein (Rhizopus delemar (strain RA ..)		26.7%
M1B905	Uncharacterized protein (Solanum tuberosum)		32.4%
M1B906	Uncharacterized protein (Solanum tuberosum)		32.4%
A0A0K6D906	Uncharacterized protein (Solanum tuberosum)		31.5%

Alignments

[BLAST](#) [Align](#) [Download](#) [Add to basket](#) [Columns](#)

◀ 1 to 25 of 78 ▶ Show 25 ▼

Entry	Alignment overview	Info	Status
<input type="checkbox"/>	Query: B2016032256586W0MKZ		
<input type="checkbox"/>	Q8CLU5 Q8CLU5_YERPE - Uncharacterized protein Yersinia pestis - View alignment	E-value: 510E-150 Score: 1,083 Ident.: 100.0%	

CDD (conserved domain database)

<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

- Doména – část proteinu s vlastní aktivitou nebo strukturní funkcí (více v bloku o 3D a 4D struktuře proteinů)
- Domény často obsahují sekvenční motiv, který můžeme nalézt u více proteinů s určitou funkcí – **konzervované domény**

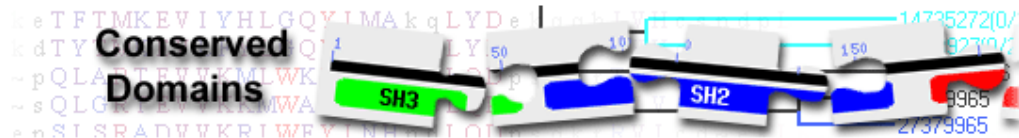
CDD zdroje

Abbreviation	Database Name	Description
SMART	Simple Modular Architecture Research Tool	SMART is a web tool for the identification and annotation of protein domains, and provides a platform for the comparative study of complex domain architectures in genes and proteins. SMART is maintained by Chris Ponting, Peer Bork and colleagues, mainly at the EMBL Heidelberg. CDD contains a large fraction of the SMART collection.
Pfam	Protein families	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. Pfam is maintained by Alex Bateman and colleagues, mainly at the Wellcome Trust Sanger Institute. CDD contains a large fraction of the Pfam collection.
COGs	Clusters of Orthologous Groups of proteins	COGs is an NCBI-curated protein classification resource. Sequence alignments corresponding to COGs are created automatically from constituent sequences and have not been validated manually when imported into CDD.
TIGRFAM	The Institute for Genomic Research's database of protein families	TIGRFAM , a research project of the J. Craig Venter Institute, is a collection of manually curated protein families from The Institute for Genomic Research and consists of hidden Markov models (HMMs), multiple sequence alignments, Gene Ontology (GO) terminology, cross-references to related models in TIGRFAM and other databases, and pointers to literature.
PRK	Protein K(c)lusters	Protein Clusters is an NCBI collection of related protein sequences (clusters) consisting of Reference Sequence proteins encoded by complete prokaryotic and chloroplast plasmids and genomes. It includes both curated and non-curated (automatically generated) clusters.

CD search - NCBI

<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

Možnost hledání záznamů v CDD dle klíčového slova nebo identifikace konzervované domény v zadané sekvenci (CDS)



HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems

Search for Conserved Domains within a protein or coding nucleotide sequence

NEW! Use **Batch CD-search** to submit multiple query proteins at once!

Enter **protein** or **nucleotide** query as accession, gi, or sequence in **FASTA format** ?

Submit Reset

OPTIONS

Search against database ? : CDD v3.08 - 43334 PSSMs ▾

Expect Value ? threshold: 0.01 ▾

Apply low-complexity filter ?

Force live search ?

Maximum number of hits ? 500

Result mode Concise ? Full ?

Retrieve previous CD-search result

Request ID: Retrieve ?

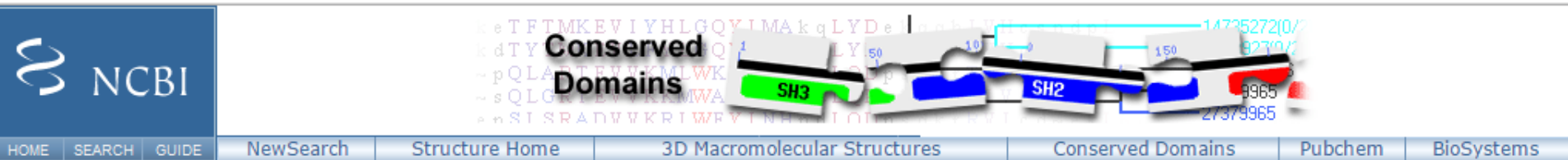
Úloha

- Vyhledejte pomocí nástroje CD search (NCBI) konzervované domény následujícího proteinu.

Sekvence:

PEVRSSTQSEGMSQWMGKILSIRGAGLIIGVFGLCALIAATSVTLPEEQLLIVAFVFCVVIFFIVGHKPSRRSQIFLEVLVSLGLVSLRYLTWRLT
ETLSFDTWLQGLLGTMLLVAELYALMMLFLSYFQTIAPLHRAPLPLPPNPDEWPTVDIFVPTYNEELSIVRLTVLGSGLIDWPPEKVRVHIL
DDGRRPEFAAFAAECGANYIARPTNEHAKAGNLNYAIGHTDGDYILIFDCDHVPTRAFLQLTMGWMVEDPKIALMQTPHHFYSPDPF
QRNLSAGYRTPPEGNLFYGVVQDGNDFWDATFFCGSCAILRRTAIEQIGGFATQTVTEDAHTALKMQRLGWSTAYLRIPLAGGLATERLI
LHIGQRVRWARGMLQIFRIDNPLFGRGLSWGQRLCYLSAMTSFLFAVPRVIFLSSPLAFLFFGQNIIAASPLALLAYAIPHMFHAVGTASKI
NKGWRYSEFWSEVYETTMALFLVRVTIVTLLSPSRGKFNVTDKGGLLEKGYFDLGAVYPNIILGLIMFGGLARGVYELSFHGLDQIAERAYL
LNSAWAMLSLIIILAAIAVGRETQQKRNSHRIPATIPVEVANADGSIIVTGVTEDLSMGGAAVKMSWPAKLSGPTPVYIRTVLDGEELILPA
RIIRAGNGRGIFIWTIDNLQQEFSVIRLVFGRADAWVDWGNKYADRPLLSLMDMVLSVKGLFRSSGDIVHRSSPTKPSAGNALSDDTNN
PSRKERV LKGTVKMVSLLALLTFASSAQAASAPRAVAAKAPAHQPEASDLPLPALLPATSGAAQAGSGDAGADGPGSPTGQPLAADSA
DALVENAENTS DTATVHNYTLKDLGAAGSITMRGLAPLQGIEFGIPSDQLVTSARLVLSGSMSPNLRPETNSVTMTLNEQYIGTLRPDPA
HPTFGPMSFEINPIFFVSGNRLNFNFASGSKGCS DITNDTLWATISQNSQLQITTI ALPPRRLSRLPQPFYDKNVRQHVTVPMVLAQTYD
PQILKSAGILASWFGKQTD FLGVTFPVSSTIPQSGNAILIGVADELPTS FGRPQVNGPAVLELPNPS DANATILVVTGRDRDEVITASKGIAF
ASAPLPTDSHMDVAPVDIAPRKPN DAPSFIAMDH PVRFGDLVTASKLQGTGFTSGVLSVPFRIPPDLYTWRNRPYKMQRFRSPAGEA
KDVEKSRLDVGINEVYLHSYPLRETHGLIGAVLQGVGLARPASGMQVHDLVPPWTVFGQDQLNFYFDAMPLARGICQSGAANNAF
HLGLDPDSTIDFSRAHHIAQMPNLAYMATVGFPTTYADLSQTAVVLPEHPNAATVGAYL DLMGFMGAATWYPVAGVDIVSADHVSD
VADRNLLVISTLATS GEIAPLLSRSSYEADGHLRTVSHASALDNAIKAVDDPLTAFRDRDSK PQDQDVTPLTGGVGAMIEAESPLTAGRTVL
ALLSSDGAGLNNLLQMLGERKKQANIQGD LVVAHGEDLSSYRTSPVYTIGTLPLWLWPDWYMHNRPV RVLVGLLGCILIVSVLARALA
RHAARRFKQLEDERRKS

- Conserved domain search (CDS)

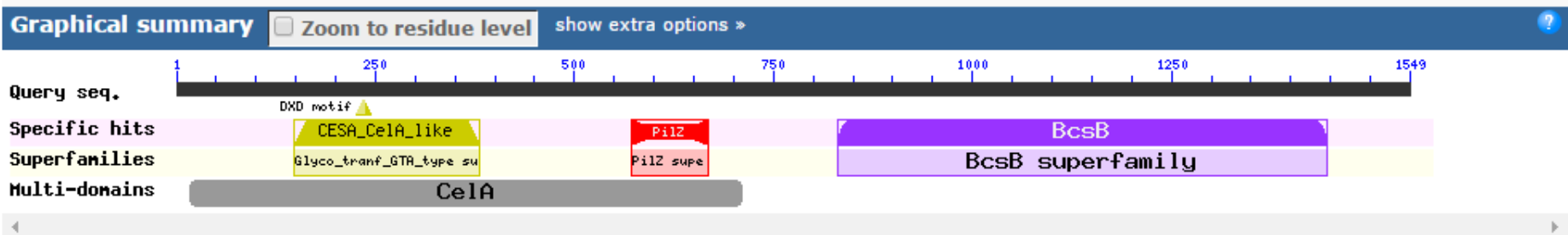


Conserved Domains

Conserved domains on [lcl|seqsig_PEVRS_d698b6fa98438c3287899de7ead8da83] View Concise Results ?

Local query sequence

Graphical summary Zoom to residue level show extra options ?



Query seq. 1 250 500 750 1000 1250 1549

Specific hits: CESA_CeIA_like, PiIZ, BcsB

Superfamilies: Glyco_tranf_GTA_type su, PiIZ supe, BcsB superfamily





Multi-domains: CeIA

Search for similar domain architectures ? Refine search ?

List of domain hits ?

+	Name	Accession	Description	Interval	E-value
[+]	CESA_CeIA_like	cd06421	CESA_CeIA_like are involved in the elongation of the glucan chain of cellulose; Family of ...	148-380	5.75e-118
[+]	BcsB	pfam03170	Bacterial cellulose synthase subunit; This family includes bacterial proteins involved in ...	830-1446	0e+00
[+]	PiIZ	pfam07238	PiIZ domain; PiIZ is a c-di-GMP binding domain which is found C terminal to pfam07317. ...	571-669	2.16e-11
[+]	CeIA	TIGR03030	cellulose synthase catalytic subunit (UDP-forming); Cellulose synthase catalyzes the beta-1,4 ...	18-710	0e+00

References:

-  Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", **Nucleic Acids Res.**43(D)222-6.
-  Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", **Nucleic Acids Res.**39(D)225-9.
-  Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", **Nucleic Acids Res.**37(D)205-10.
-  Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", **Nucleic Acids Res.**32(W)327-331.

CDART

(Conserved domain architecture retrieval tool)

<http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi>

- Nevyhledává pouze izolované domény, ale zohledňuje jejich kombinace a vzájemná umístění v jednom proteinovém řetězci.



CONSERVED Domain Architecture Retrieval Tool



HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains Pubchem BioSystems Help

Launch a new search

Enter query protein sequence ?

Submit Reset

Úloha


- V následujícím proteinu byla zjištěna kombinace rhodanasové a ankyrinové domény. Zjistěte, zda je tento případ unikátní a v jakých jiných kombinacích se tyto domény v přírodě vyskytují. Použijte aplikaci **CDART**.

Sekvence:

MNTRSFHRIDVHKARELLQRPDTVLLDCRHPSDFRAGHIAGASPLGDYNADDHVLNIAKHRPVLIYCYHG
NASQMRAQLFADFGFAEVYSLDGGYEAWRKVHTPANSQLTEALQCWLMAQEFPAADIHARTRDGVTP
MRAAGEGDPARVAELLAAGADPHQRNNDGNQALWFACVSENLDTLDLLVAVGAHLNHQNDNGATCL
MYAASA GKTAVVERLLAFGADRSLLSLDDFTALDMAANLECLNLLRETPRRIKAVT


Conserved domain architecture retrieval tool (CDART)

[Query] Icl|local_MNTRSFHRID (Local query sequence)




Total architectures: 736

Rhodanese-like protein with Ankyrin
 taxonomy span: Proteobacteria
 Similarity score: 2
 Total nr sequences: 23

Filter your results: Apply 

[Query] Icl|local_MNTRSFHRID (Local query sequence)

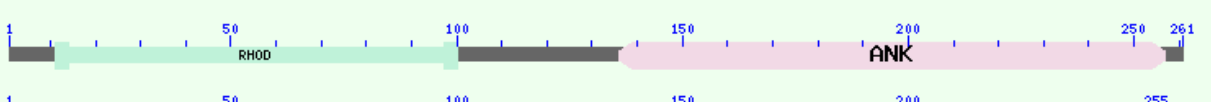


Total architectures: 736

Rhodanese-like protein with Ankyrin
 taxonomy span: Proteobacteria
 Similarity score: 2
 Total nr sequences: 23


[Lookup sequences in Entrez](#)

gij226946871|YP_002801944
 Rhodanese-like protein with



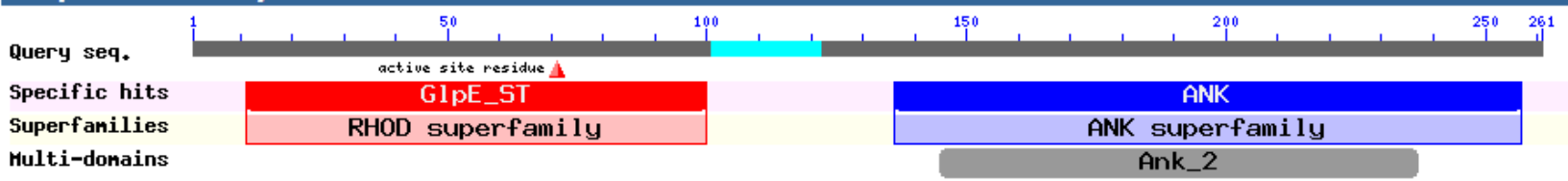
domain details >


Conserved domains on [gij226946871|ref|YP_002801944]

[View full result](#) 

Rhodanese-like protein with Ankyrin repeat [Azotobacter vinelandii DJ]

Graphical summary [show options >](#)





Query seq. 

Specific hits: **GlpE_ST** (red), **ANK** (blue)

Superfamilies: **RHOD superfamily** (red), **ANK superfamily** (blue)

Multi-domains: **Ank_2** (grey)

[Search for similar domain architectures](#)  [Refine search](#) 

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]GlpE_ST[cd01444], GlpE sulfurtransferase (ST) and homologs are members of the Rhodanese Homology Domain ...	29075	no	5.72e-30
[+]ANK[cd00204], ankyrin repeats; ankyrin repeats mediate protein-protein interactions in very diverse ...	29261	yes	4.03e-24
[+]Ank_2[pfam12796], Ankyrin repeats (3 copies);	205076	yes	5.63e-19

- Databáze proteinových rodin, vytvořená na základě Multiple sequence alignmentů (MSA) a Skrytých Markovových modelů (HMM)



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	<input type="text" value="enter any accession or ID"/> <input type="button" value="Go"/> <input type="button" value="Example"/>
	Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.
	Or view the help pages for more information

Recent Pfam [blog](#) posts

Hide this

[Dfam 1.1 released](#) (posted 15 November 2012)

We are pleased to announce that we've released Dfam 1.1. This version represents a few important changes from 1.0, including updated hit results, a new tab for each entry page showing relationships to other entries, and improved handling of redundant profile hits. New Hit Results The underlying database and set of entries have not changed from Dfam 1.0, but [...]

Úloha

- Pokuste se určit funkci následujícího proteinu pomocí databáze **Pfam**.

Sekvence:

MRYIRLCIISLLATLPLAVHASPQPLEQIKQSESQLSGRVGMIEMDLASGRTLTAWRADERFPMMSTFKVVLCGAMLA
RVDAGDKQLERKIHQRQQDLVDYSPVSEKHLADGMTVGELCAAITMSDNSAANLLLATVGGPAGLTAFLRQIGDNV
TRLDRWETELNEALPGDARDTTTPASMAATLRKLLTSQRLSARSQRQLLQWMVDDRVAGPLIRSVLPAGWFIADKTG
ASKRGARGIVALLGPNKAERIVVIYLRDTPASMAERNQQIAGIGAAL IEHWQR

- Odhadněte, které z vyznačených aminokyselin mají vliv na správnou funkci či strukturu proteinu.

PFAM <http://pfam.sanger.ac.uk>



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)

Sequence search results

[Show](#) the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment	
				Start	End	Start	End
Beta-lactamase2	Beta-lactamase enzyme family	Family	CL0013	47	257	48	256
#HMM	dtgeei.ginadeefpaaStiKvpil...eavlegelslderitvtkedivggsgilgkldgktslsvrdllelmiavSDNtAtnlLidrlg.lldavnawlrelglrdtrlrrklpdl.e.aldk						
#MATCH	+g+++ + +ade+fp++St+Kv++ +v++g +l+ +i++++d+v+ s++ +k+ + ++v +l++++i+ SDN A+nL++++g + ++a+lr++g + trl+r++ +l+ al+ c						
#PP	577888789*****976656699*****999*****99*****9988868999						
#SEQ	ASGRILTAWRADERFPMMSTFKVVLGgamLARVDAGDKQLERKIHVRQDDLVDYSPVSEKHLADGMTVGE LCAAAITMSDNSAANLLLATVGGPAGLTAFLRQIGDNVIRLDRWETELNeALPG						

Comments or questions on the site? Send a mail to pfam-help@sanger.ac.uk

The Wellcome Trust

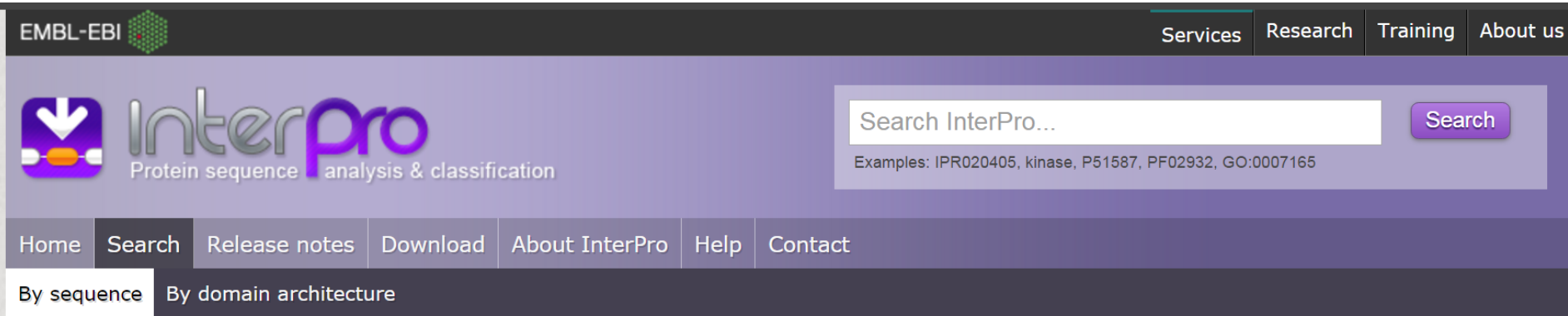
V kombinaci je síla...



InterProScan <http://www.ebi.ac.uk/Tools/pfa/iprscan/>

Kombinovaný nástroj pro analýzu proteinové sekvence pomocí různých databází

- 14 aplikací v jednom běhu



The screenshot shows the top navigation bar of the InterProScan website. On the left, it features the EMBL-EBI logo. On the right, there are links for 'Services', 'Research', 'Training', and 'About us'. Below the navigation bar is a large purple banner containing the InterPro logo (a stylized 'I' with a white arrow pointing down) and the text 'Interpro Protein sequence analysis & classification'. To the right of the logo is a search box with the placeholder text 'Search InterPro...' and a 'Search' button. Below the search box, there are example search terms: 'Examples: IPR020405, kinase, P51587, PF02932, GO:0007165'. Below the banner is a secondary navigation bar with links for 'Home', 'Search', 'Release notes', 'Download', 'About InterPro', 'Help', and 'Contact'. Below this is a dark grey bar with two tabs: 'By sequence' (which is selected) and 'By domain architecture'.

InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.

Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long.

Please note that you can only scan one sequence at a time.

Analyse your protein sequence

InterProScan



InterProScan is a sequence analysis application (nucleotide and protein sequences) that provides

InterProScan

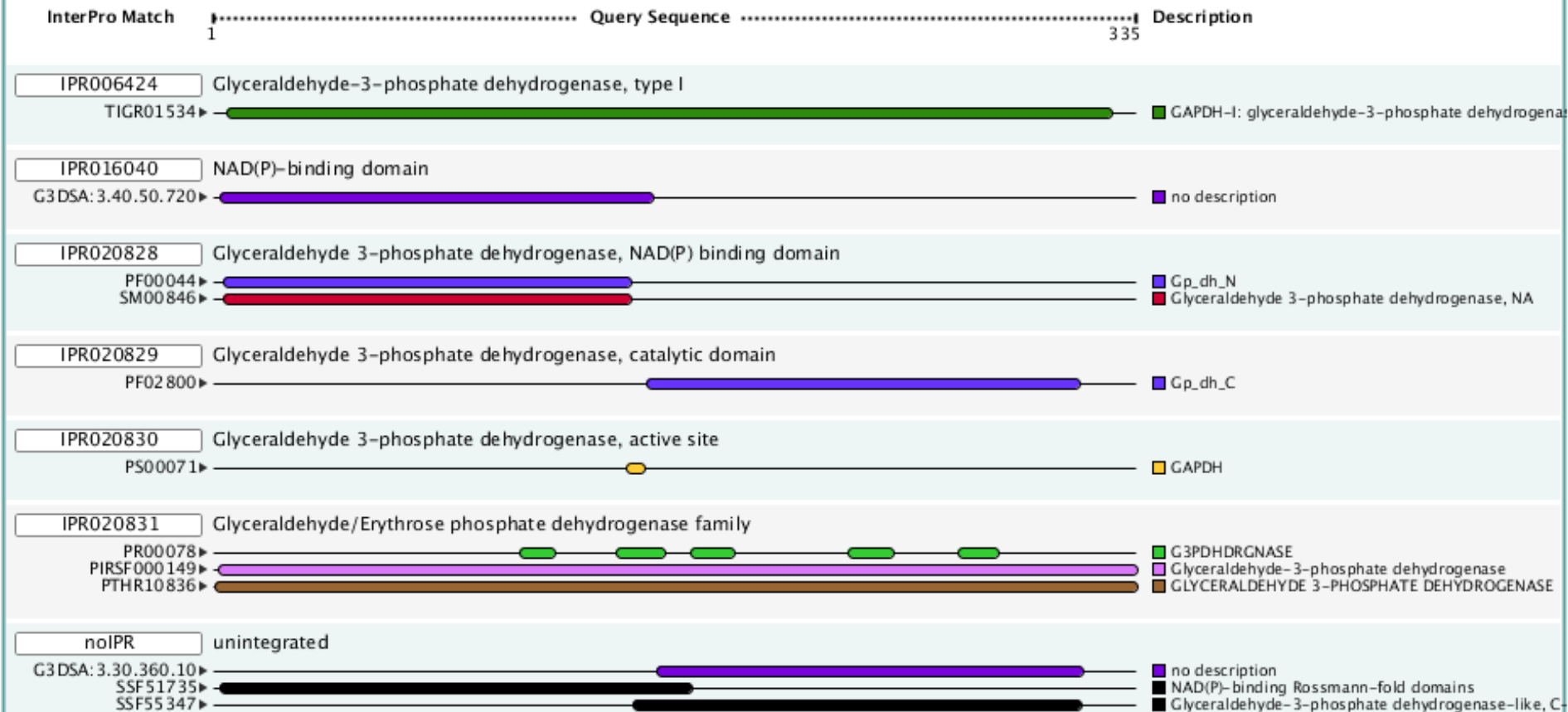
Launched Mon, Jan 28, 2013 at 14:46:10
Finished Mon, Jan 28, 2013 at 14:47:06

InterProScan (version: 4.8)

Sequence: Sequence_1

Length: 335

CRC64: C9C135E8AE3E8744



■ PRODOM ■ PRINTS ■ PIR ■ PFAM ■ SMART ■ TIGRFAMs ■ PROFILE
■ HAMAP ■ PROSITE ■ SUPERFAMILY ■ SIGNALP ■ TMHMM ■ PANTHER ■ GENE3D

Úloha

- Pokuste se určit funkci následujícího proteinu pomocí serveru **InterProScan**.

Sekvence:

MTELKNDRYLRALLRQPVDVTPVWMMRQAGRYLPEYKATRAQAGDFMSLCKNAELACEV
TLQPLRRYPLDAAILFSDILTIPDAMGLGLYFEAGEGPRFTAPVTCKADVDKLPIDPEDELGYV
MNAVRTIRRELKGEVPLIGFSGSPWTLATYMVEGGSSKAFTVIKKMMYADPQALHLLLDKLA
KSVTLYLNAQIKAGAQSVMIFDTWGGVLTGRDYQQFSLYYMHKIVDGLLRENDGRRVPVTLF
TKGGGQWLEAMAETGCDALGLDWTTDIADARRRVGHKVALQGNMDPSMLYAPPARIEDE
VATILAGFGQGEGHVFNLGHGIHQDVPPEHAGAFVEAVHRLSA QYHN

Filter view on

Entry type

- F** Family
- D** Domains
- R** Repeats
- S** Site

Status

- ?** Unintegrated

Colour by [help](#)

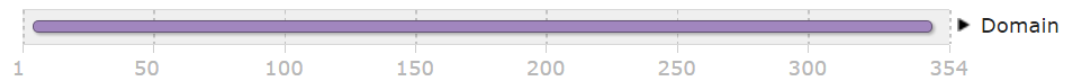
- domain relationship
- source database

Length 354 amino acids

Protein family membership

F Uroporphyrinogen decarboxylase HemE (IPR006361)

Domains and repeats

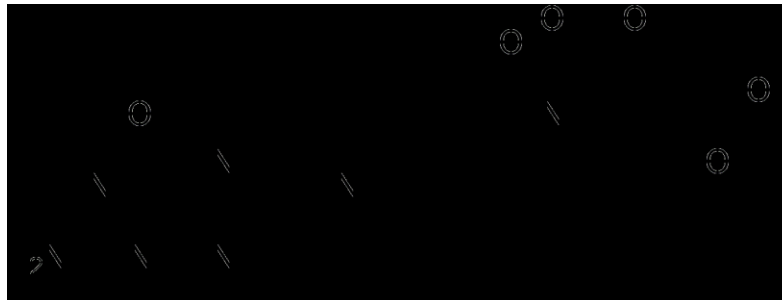
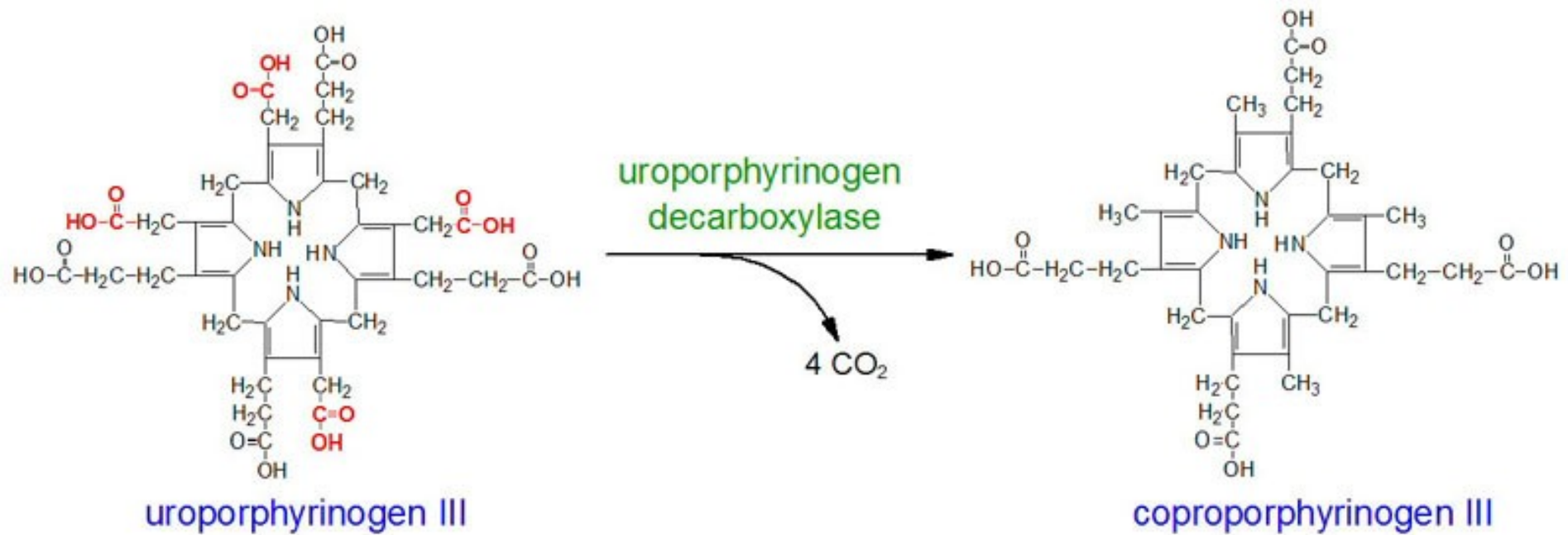


Detailed signature matches

F IPR006361	Uroporphyrinogen decarboxylase HemE	
		▶ MF_00218 (URO_D) ▶ TIGR01464
D IPR000257	Uroporphyrinogen decarboxylase (URO-D)	
		▶ PS00907 (UROD_2) ▶ PF01208 (URO-D) ▶ PS00906 (UROD_1)
? no IPR	Unintegrated signatures	
		▶ G3DSA:3.20.20... ▶ PTHR21091 (METHYLTE...) ▶ PTHR21091:SF2 (URO...) ▶ SSF51726 (UROD/MetE...)

GO term prediction

Úloha



rofolát

PROPSEARCH – Když selže alignment

<http://abcis.cbs.cnrs.fr/propsearch/>

The logo for PROPSEARCH is a rectangular button with a double border. The word "PROPSEARCH" is written in blue, uppercase letters in the center of the button.

PROPSEARCH

- Neprovádí alignment
- Porovnává složení (zastoupení) aminokyselin, molekulovou hmotnost, izoelektrický bod, atd.
> celkem 144 parametrů
- Snaha zařadit protein do funkční rodiny
- **Další studium sekvence nutné !**

Úloha

- Následující sekvenci hypotetického proteinu analyzujte pomocí serveru **Propsearch**.

Sekvence:

MASPSILKKYGKYFEYCPLEERMIELAKKGEIADAMLLFEKEKPSEFVYKGDAIEKRLRNIYLSTR
LGVKAKINFNDYVIPRDLRWMLDIYESYLNMGENKVFLILGGELRYLIDFFESYLQFKGFYLLVV
KEAKDLLRFRNTCHYDAIIFSDSSILEYQNVDELKNLFNSLETTLKVHNRKNSVKVLLSPALPKAI
MSSKPYKVLEQFFKEKGIEMEGILPYQLNADDKLLPPHFHNSEMEKSKEYRELESKTKVYIQEF
LKKANMNDENEGNDNQKNTN

Please cite: Uwe Hobohm and Chris Sander: "A sequence property approach to searching protein databases", J.Mol.Biol. 251 (1995) 390-399

For a successful application, please have a look in: Uwe Hobohm and Chris Sander: "Does the HIV Nef protein mimic the MHC ?", FEBS-letters 333(1993)211-213.

Query sequence

Paste your sequence into the text area. The sequence may contain **ONLY amino acid residue characters (one letter code)** and carriage returns but no blanks AND NO position numbers (All characters apart from amino acid letters might now be truncated by PropSearch).

```
MPPGVLDLNLNRNGIQCIASVTLLQRFPHYHRNHHTIPTMDQNRPTVGGVTAEVLKAVVSEALL
GKFTAVFP
KLFKQCQNKTYEGQKKFYANFSTDCDTSNSNNMSRKFCVGGNWKMGDQKSI AEIAKTLSSAA
LDPNTEVV
IGCPAIYLMYARNLLPCELGLAGQNAYKVAKGAFTGEISPAMLKDIGADWVILGHSERRAIF
GESDALIA
EKAHALAEGLKVIACIGETLEEREAGKTNEVVARQMCAYAQKIKDWNVVVAYEPVWAIGT
GQTATPDQ
AQEVHAFRLRQWLSNDSISKEVSASLRIQYGGSVTAANAKELAKKPDIDGFLVGGASLKPFEVD
IINARQ
```

Clear Form

Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	ycx8_euggr	0.00	281	1	281	7.51	Hypothetical 33.1 kDa protein in RBCL-ATPE intergenic region
2	pyrb_aquae	6.69	291	1	291	7.72	Aspartate carbamoyltransferase (EC 2.1.3.2) (Aspartate
3	y800_pyrab	7.35	326	1	326	6.09	Hypothetical ATP-binding protein PAB0800.
4	ynz4_caeel	7.37	297	1	297	9.38	Hypothetical 35.0 kDa protein T09A5.4 in chromosome III.
5	y983_aquae	7.44	292	1	292	8.89	Hypothetical protein AQ_983.
6	ye16_yeast	7.69	306	1	306	7.25	Hypothetical 35.9 kDa protein in ISC10 3'region.
7	cher_camje	7.76	262	1	262	9.38	Chemotaxis protein methyltransferase (EC 2.1.1.80).
8	bra2_chith	7.77	245	1	245	7.26	Balbani RING A 28 kDa protein precursor.
9	mag_yeast	7.81	296	1	296	8.16	DNA-3-methyladenine glycosylase (EC 3.2.2.21) (3-methyladenine DNA
10	y151_yeast	7.82	340	1	340	6.59	Hypothetical 39.8 kDa protein in MPT4-ACS2 intergenic region.
11	yj17_aquae	7.91	345	1	345	8.08	Hypothetical protein AQ_1917.
12	yz38_aquae	7.98	320	1	320	9.28	Hypothetical protein AA38.
13	rfbj_salmu	8.02	293	1	293	8.50	CDP-abequose synthase (EC 4.2.1.-).
14	ars2_aquae	8.02	299	1	299	6.73	Putative arsenical pump-driving ATPase 2 (EC 3.6.3.16) (Arsenite-
15	met8_yeast	8.02	274	1	274	6.27	Siroheme biosynthesis protein MET8 [Includes: Precorrin-2 oxidase
16	kkih_lacla	8.03	262	1	262	6.08	Probable aminoglycoside 3'-phosphotransferase (EC 2.7.1.95) (Kanamycin
17	yp76_borbu	8.07	265	1	265	8.61	Hypothetical protein BBA76.
18	rnh_bpt4	8.07	305	1	305	9.00	Ribonuclease H (EC 3.1.26.4) (RNase H).
19	mtm3_metja	8.08	289	1	289	9.46	Modification methylase MjaIII (EC 2.1.1.72) (Adenine-specific
20	ym2_yeast	8.12	240	1	240	6.60	Hypothetical 27.7 kDa protein in GBT1_SFG00 intergenic region

SMART <http://smart.embl-heidelberg.de/>



Analýza zastoupení **proteinových domén**
Prohledávání dle sekvencí, domén, taxonů

STRING <http://string-db.org>



Funkční vazby různých proteinů

Na základě výskytu v genomu, zapojení do
metabolických drah, textového hledání,...

Úloha

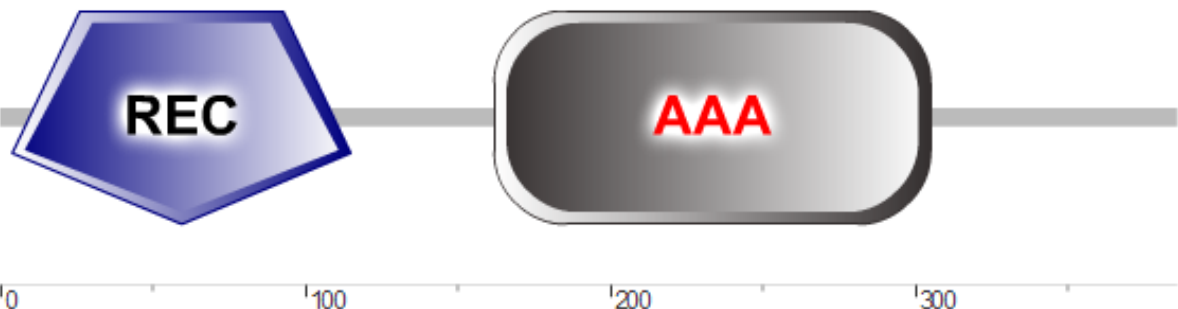
- Pomocí serveru **SMART** analyzujte domény v následující sekvenci a následně pomocí serveru **STRING** prozkoumejte pravděpodobné zapojení v metabolismu.

Sekvence:

MSIEHILIIDDDPHILALLSEILGARNFSVSSAPGVKQAIKQISNCPFDLIISDMNMPDGSGLDII
QYTKQHRPQTPILVITAFGTIQNAVEAMRFGAFNYLTKPFSPDALFTLIAKAEELQALQQDNLF
LQSQGSSISHPLIAESPSMKQLLDKARRAANSSANIFVHGESGCGKENLSFFIHKHSPRSTKPYI
KVNCAAIPDTLLESEFFGHEKGAFTGATTKKVGRFELAHQGTLLEDEITEIPIHLQAKLLRAIQE
QEFEHIGGIKTLPVNIRFLATSNRDLEEAIETKVLQRDLYYRLSVISLHIPPLRDRKEDILPLAHYYL
EKFCKMNNKPPKTLSELAQRNLLDYSWPGNVR ELSNVLERTVILENDPAITPSMLALL

Domains within *Chlamydia trachomatis* 434/Bu protein **B0B841_CHLT2 (B0B841)**

Two component system response regulator



Information

Architecture

Interactions

Orthology

Length 386 aa

Source database [UniProt](#)

Identifiers B0B841_CHLT2, B0B841

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the same piece of sequence; the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured**. Domains shown in the above diagram are marked as 'overlap' in the right side table below.

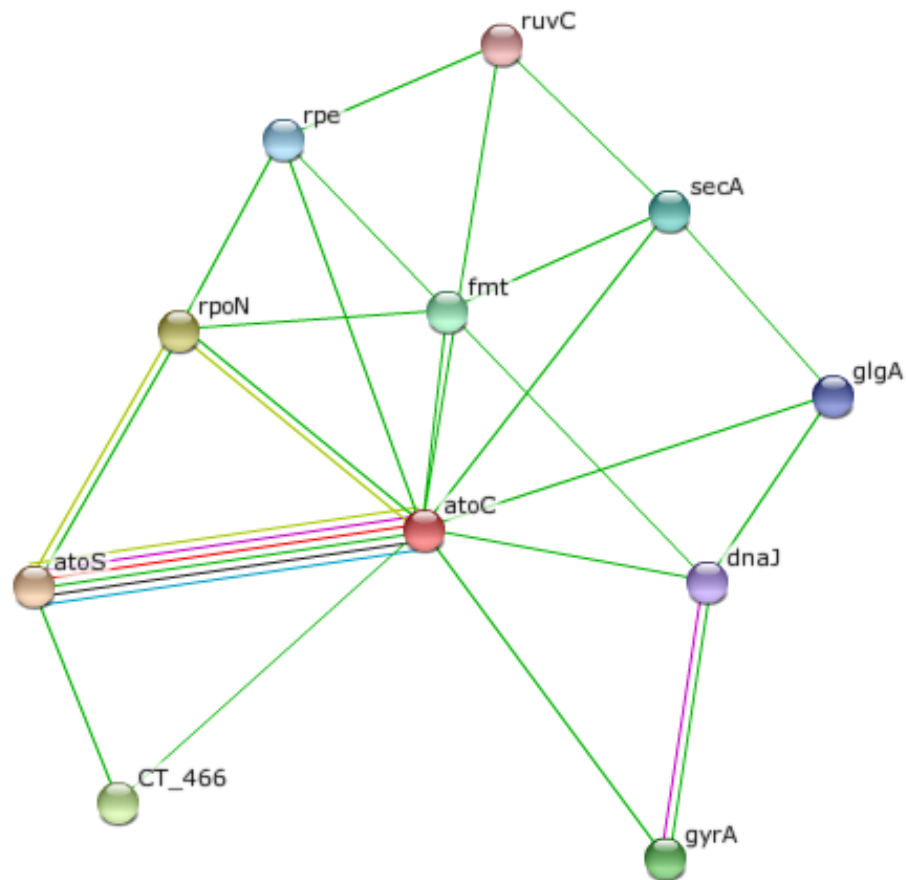
Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
REC	4	115	4.25e-32
AAA	162	305	6.98e-7

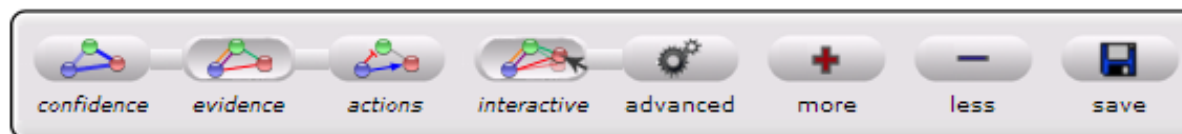
Features NO

Name	Start ▲	E-value
low complexity	6	1

Click on a row to highlight the feature in the diagram



This is the **evidence view**. Different line colors represent the types of evidence for the association.



(requires Flash player 10 or better)

Porovnání predikce a experimentu

- **Predikce:**

- + Rychlá (sekundy-hodiny), levná/dostupná (Freeware)
- + Spolehlivá pro známé (!) proteiny a pro proteiny s vysokou homologií
- Pouze kvalitativní
- Málo spolehlivá pro neznámé proteiny
- Nepoužitelná pro unikátní případy

- **Experiment:**

- + Teoreticky použitelný pro libovolný protein
- Finančně (i miliony Kč) a časově náročný (minuty-hodiny + příprava vzorku = týdny až roky)

Rady do života

- ★ **O daném proteinu získej maximum informací** ★
- ★ **Kombinuj různé predikční programy a přístupy** ★
- ★ **Kriticky kontroluj SW výstupy** ★