BIOINFORMATICKÉ NÁSTROJE

VYUŽITÍ V PRAXI
VYPRACOVÁNÍ PROJEKTU

Zkouška:

Test + přednáška/prezentace + nová data (výběr termínu)

Prezentace - Analýza proteinu

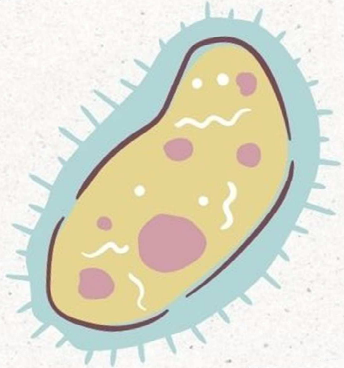
Obsah této přednášky

Konkrétní nová data – článek (< 5 let) o komplexu (nebo proteinu)

Ujasnit si souvislosti, rozšířit si znalosti, aplikovat poznatky z přednášek

Analýza proteinu

- Popis a funkce
- Konzervovanost (Alignment)
- Evoluce (fylogenetický strom)
- Zjištění přítomnosti domén daného proteinu
- Zjištění interakčních partnerů
- Struktura proteinu – Alphafold/Colabfold
- Úprava proteinového modelu v PyMol
    - Zvýraznění interakčních aminokyselin
    - Zvýraznění domén



JJ0

UniProt a jeho možnosti

Komplexní, vysoce kvalitní a volně přístupný zdroj sekvenčních a funkčních informací o proteinech

Hledání proteinu
Dle názvu
Dle sekvence

Souvislosti
Funkce
Interaktom
Mutace

Vizualizace
Alphafold
Lokalizace proteinu

Stahování dat

Analýzy
BLAST
Alignment

https://www.uniprot.org/

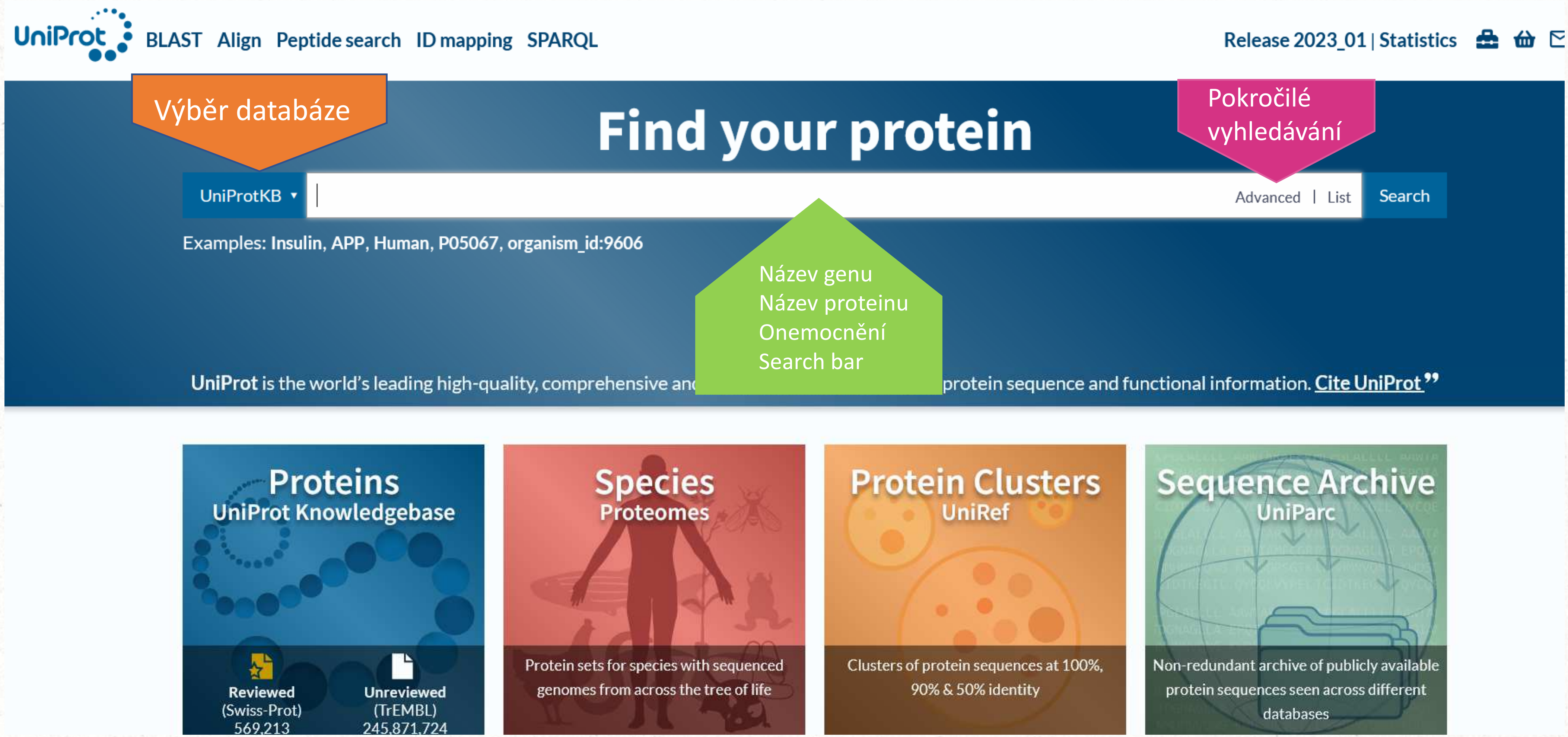
Snímek 4

JJ0 Uniprot - odrazový můstek, vytažení podstatných informací, SNP, hrubý nástřel domén - provedení po webu

Jitka Jemelková; 2023-04-15T18:20:41.865

UniProt

https://www.uniprot.org/



UniProt BLAST Align Peptide search ID mapping SPARQL Release 2023_01 | Statistics

Výběr databáze

Pokročilé vyhledávání

Find your protein

UniProtKB ▾ | Advanced | List | Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

Název genu
Název proteinu
Onemocnění
Search bar

UniProt is the world's leading high-quality, comprehensive and … protein sequence and functional information. Cite UniProt

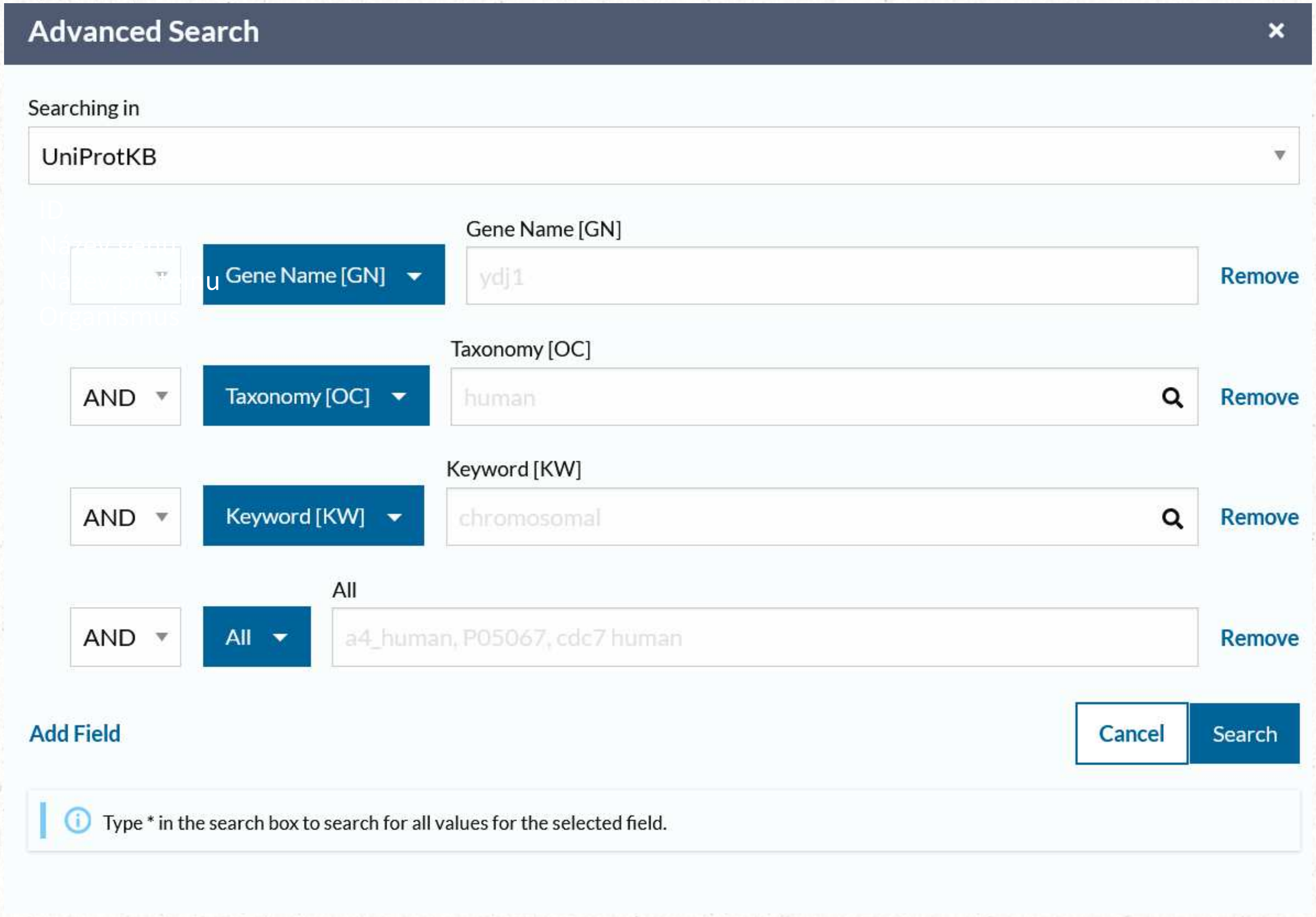
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 569,213
Unreviewed (TrEMBL) 245,871,724

Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity

Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

UniProt – Pokročilé vyhledávání



Searching in: UniProtKB

ID
Název genu
Název proteinu
Organismus

Gene Name [GN]: ydj1

AND Taxonomy [OC]: human

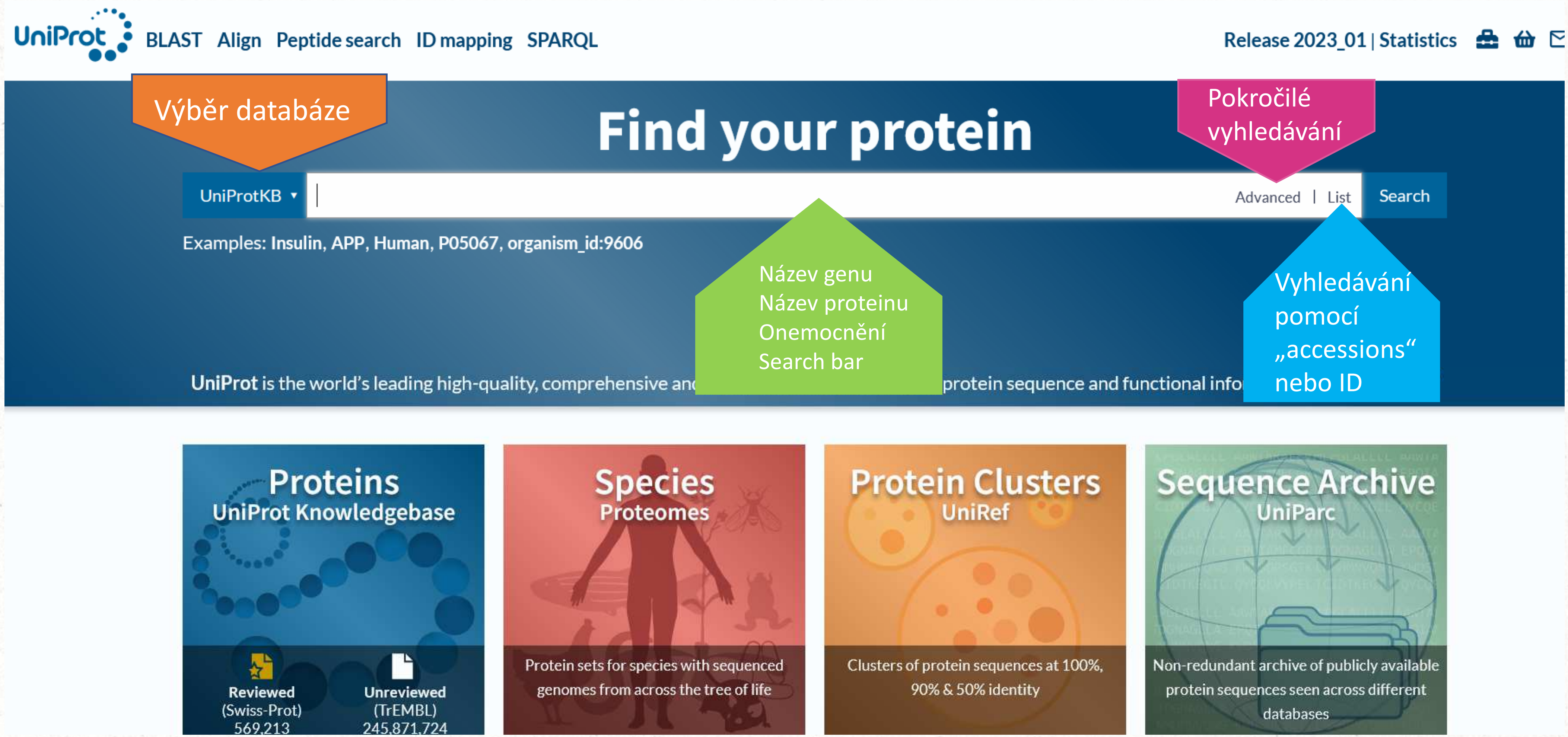
AND Keyword [KW]: chromosomal

AND All: a4_human, P05067, cdc7 human

Type * in the search box to search for all values for the selected field.

UniProt

https://www.uniprot.org/



Find your protein

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and … protein sequence and functional info…

Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 569,213
Unreviewed (TrEMBL) 245,871,724

Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity

Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

Výběr databáze

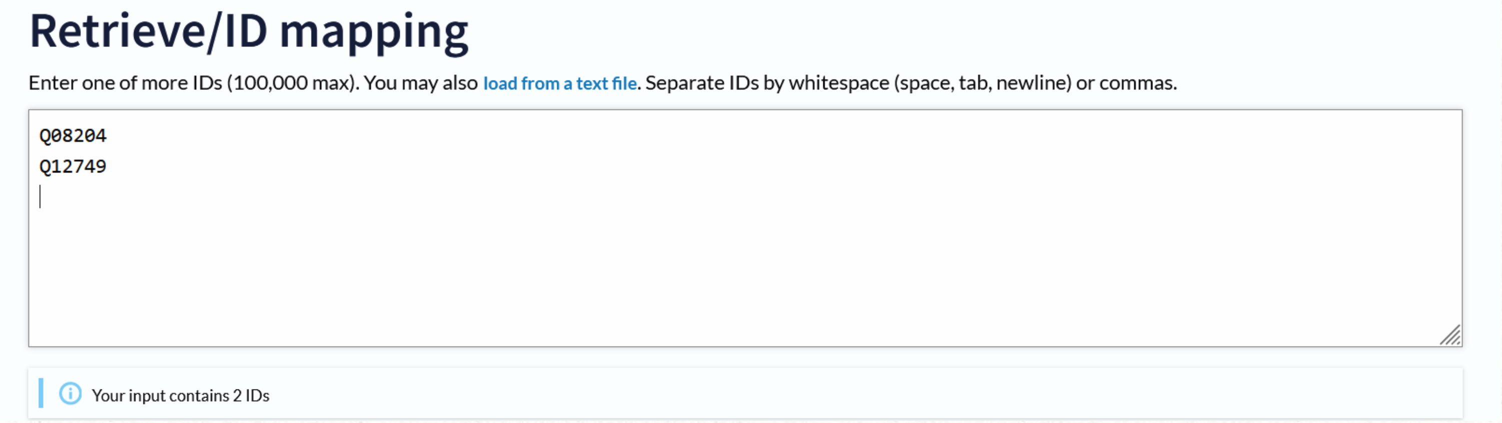
Pokročilé vyhledávání

Název genu
Název proteinu
Onemocnění
Search bar

Vyhledávání pomocí „accessions“ nebo ID

UniProt – ID mapping

Vyhledávání více proteinů najednou

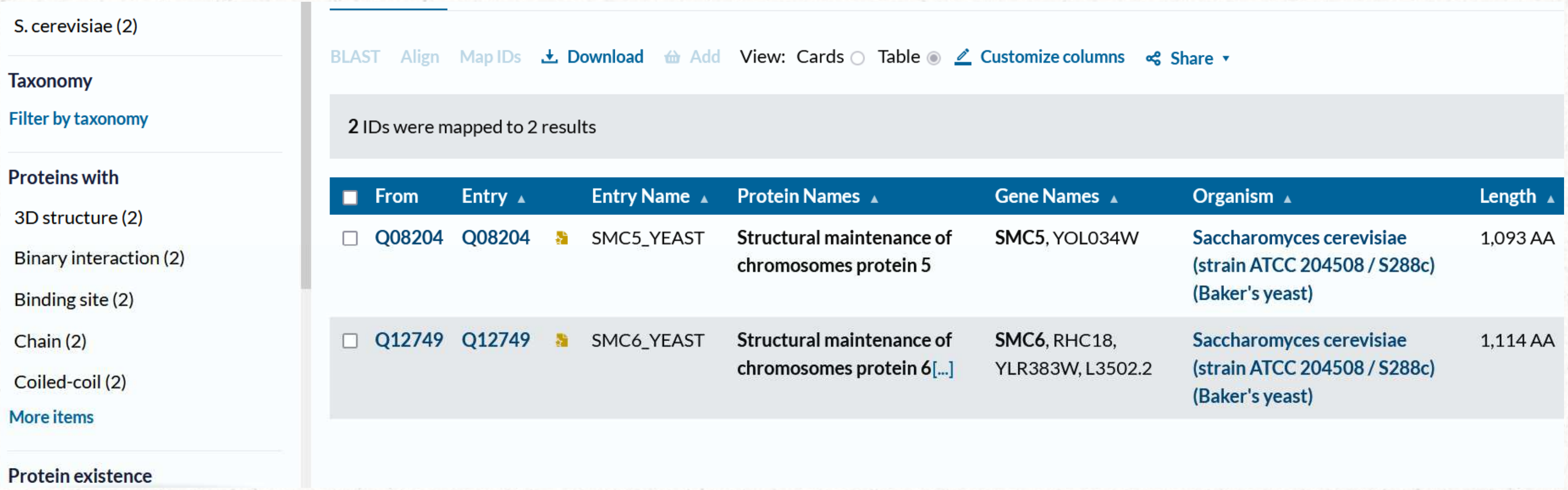


Retrieve/ID mapping

Enter one of more IDs (100,000 max). You may also load from a text file. Separate IDs by whitespace (space, tab, newline) or commas.

```
Q08204
Q12749
```

Your input contains 2 IDs



2 IDs were mapped to 2 results

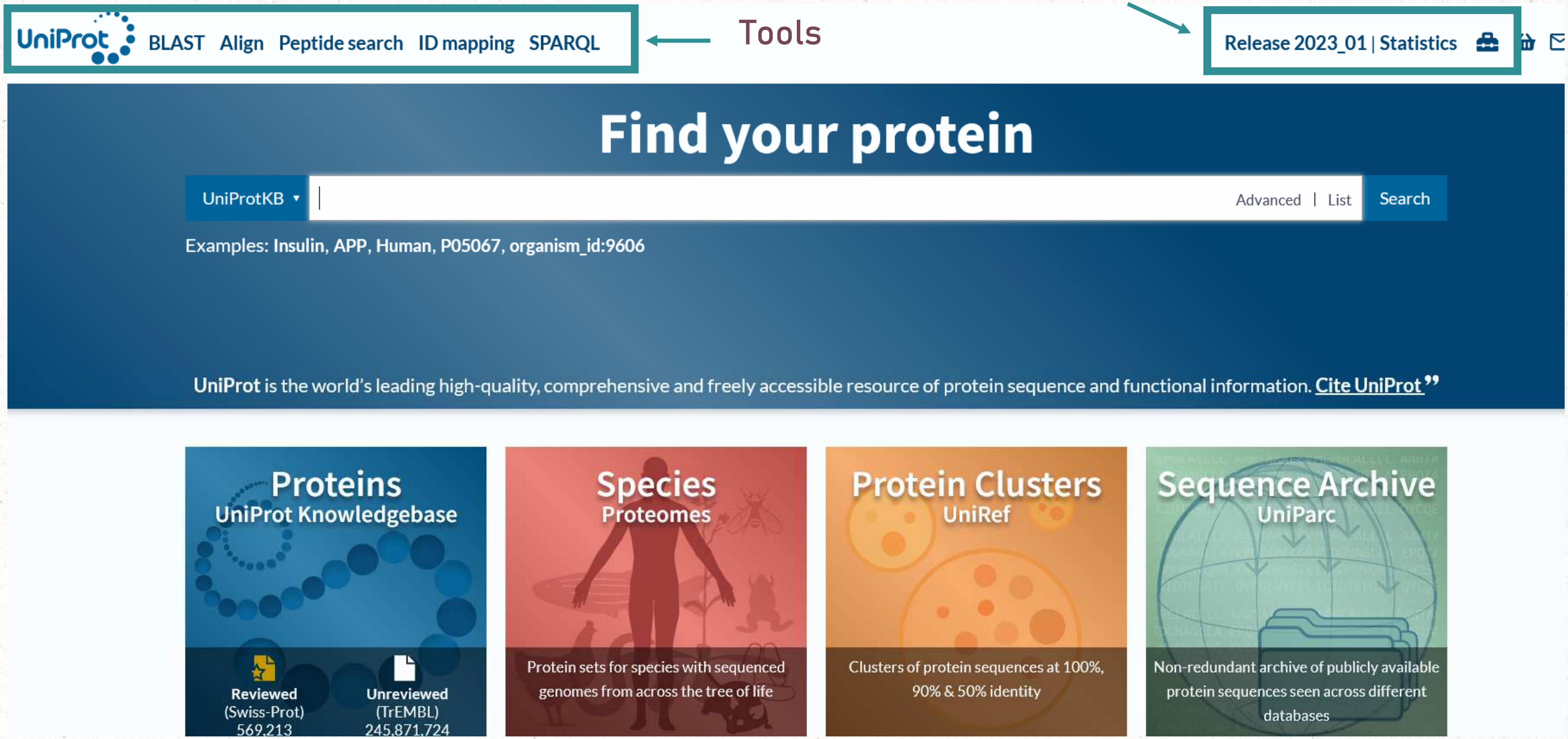
☐	From	Entry	Entry Name	Protein Names	Gene Names	Organism	Length
☐	Q08204	Q08204	SMC5_YEAST	Structural maintenance of chromosomes protein 5	**SMC5**, YOL034W	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1,093 AA
☐	Q12749	Q12749	SMC6_YEAST	Structural maintenance of chromosomes protein 6[...]	**SMC6**, RHC18, YLR383W, L3502.2	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1,114 AA

UniProt

https://www.uniprot.org/

Uložená data/analýzy

Tools



Find your protein

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. Cite UniProt

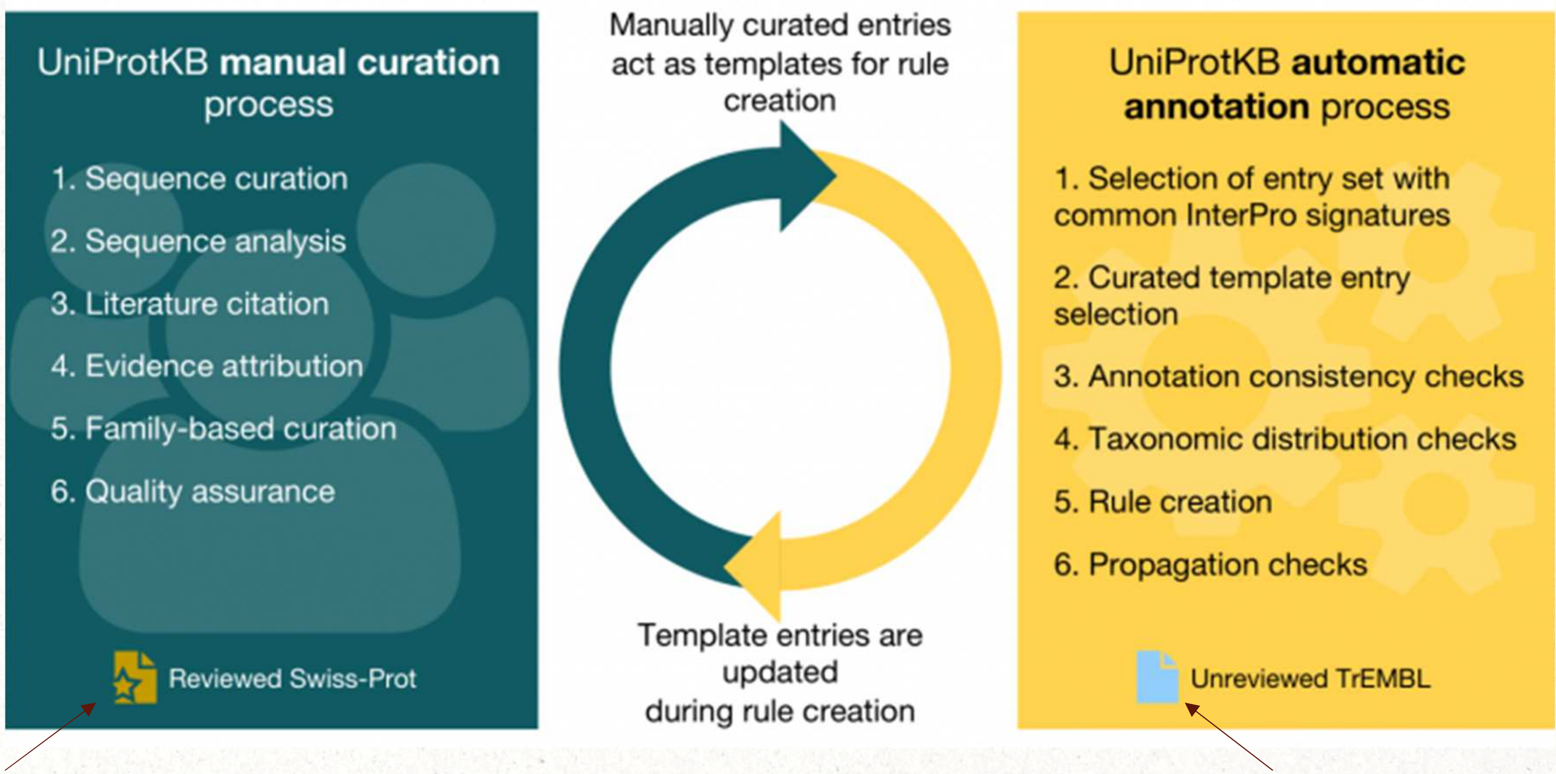
Proteins
UniProt Knowledgebase
Reviewed (Swiss-Prot) 569,213
Unreviewed (TrEMBL) 245,871,724

Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life

Protein Clusters
UniRef
Clusters of protein sequences at 100%, 90% & 50% identity

Sequence Archive
UniParc
Non-redundant archive of publicly available protein sequences seen across different databases

Na co si dát pozor



UniProtKB manual curation process

1. Sequence curation
2. Sequence analysis
3. Literature citation
4. Evidence attribution
5. Family-based curation
6. Quality assurance

Reviewed Swiss-Prot

Manually curated entries act as templates for rule creation

Template entries are updated during rule creation

UniProtKB **automatic annotation** process

1. Selection of entry set with common InterPro signatures
2. Curated template entry selection
3. Annotation consistency checks
4. Taxonomic distribution checks
5. Rule creation
6. Propagation checks

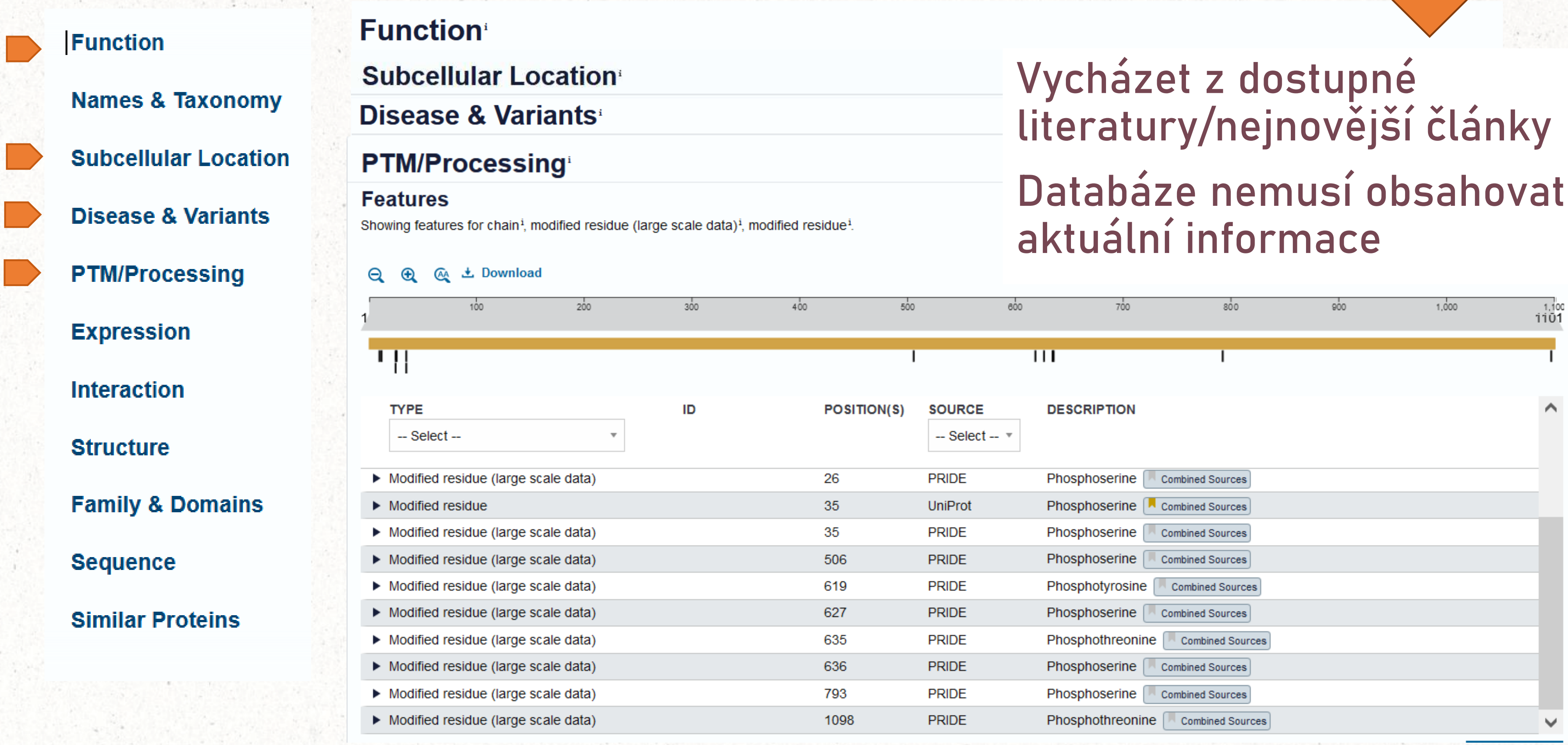
Unreviewed TrEMBL

Data, která jsou ručně zkontrolována
Označení žlutým listem s hvězdičkou

Automaticky anotované

UniProt – Vyhledávání proteinu

HsSMC5 protein



Features

Showing features for chain, modified residue (large scale data), modified residue.

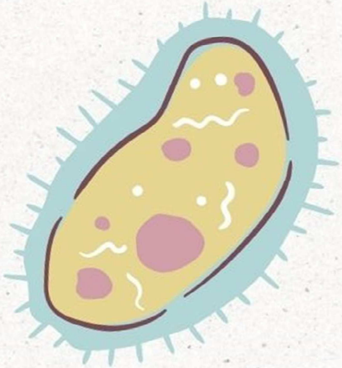
TYPE	ID	POSITION(S)	SOURCE	DESCRIPTION
-- Select --			-- Select --	
Modified residue (large scale data)		26	PRIDE	Phosphoserine Combined Sources
Modified residue		35	UniProt	Phosphoserine Combined Sources
Modified residue (large scale data)		35	PRIDE	Phosphoserine Combined Sources
Modified residue (large scale data)		506	PRIDE	Phosphoserine Combined Sources
Modified residue (large scale data)		619	PRIDE	Phosphotyrosine Combined Sources
Modified residue (large scale data)		627	PRIDE	Phosphoserine Combined Sources
Modified residue (large scale data)		635	PRIDE	Phosphothreonine Combined Sources
Modified residue (large scale data)		636	PRIDE	Phosphoserine Combined Sources
Modified residue (large scale data)		793	PRIDE	Phosphoserine Combined Sources
Modified residue (large scale data)		1098	PRIDE	Phosphothreonine Combined Sources

Vycházet z dostupné literatury/nejnovější články

Databáze nemusí obsahovat aktuální informace

Přednáška: Analýza proteinu

- Popis a funkce ☑
- Konzervovanost (Alignment)
- Evoluce (fylogenetický strom)
- Zjištění přítomnosti domén daného proteinu



Alignment

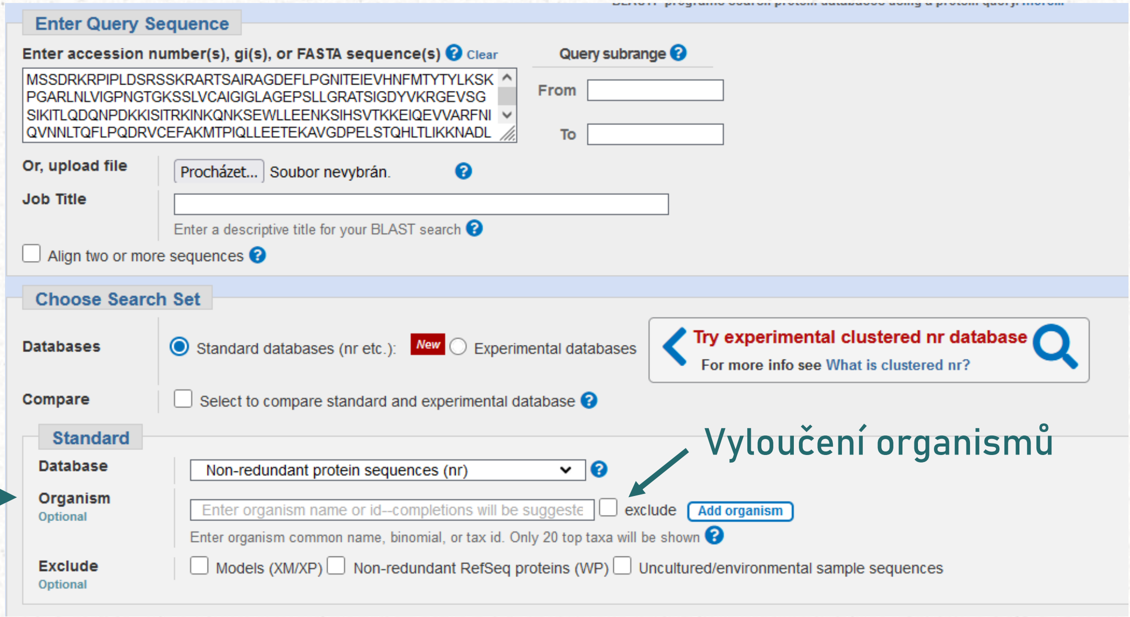
- Porovnání příbuzných proteinů (Multiple sequence alignment)
  - Sekvenční podobnost
  - Strukturní podobnost (Alphafold)

- Souvisí s přítomností konzervovaných domén/aminokyselin
  - Funkční, strukturní souvislosti
  - Evoluční, fylogenetické vztahy

- Strukturní podobnost
  - šroubovice konzervované
  - Hledání podobného „patternu“ – hydrofobní, nabité aminokyseliny

Základní alignment

- BLAST – hledání příbuzných sekvencí
    - Lze vybrat i organismy, u kterých bude vyhledávat podobné proteiny
    - Podívat se i na fylogeneticky vzdálenější organismy
    - Grafické rozhraní – napoví o konzervovanosti



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

```
MSSDRKRPIPLDSRSSKRARTSAIRAGDEFLPGNITEIEVHNFMTYTYLKSK
PGARLNLVIGPNGTGKSSLVCAIGIGLAGEPSLLGRATSIGDYVKRGEVSG
SIKITLQDQNPDKKISITRKINKQNKSEWLLEENKSIHSVTKKEIQEVVARFNI
QVNNLTQFLPQDRVCEFAKMTPIQLLEETEKAVGDPELSTQHLTLIKKNADL
```

Query subrange ?

From

To

Or, upload file Procházet... Soubor nevybrán. ?

Job Title

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Databases ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare ☐ Select to compare standard and experimental database ?

Standard

Database Non-redundant protein sequences (nr) ?

Organism Optional — Enter organism name or id--completions will be suggeste ☐ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude Optional — ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Vyloučení organismů

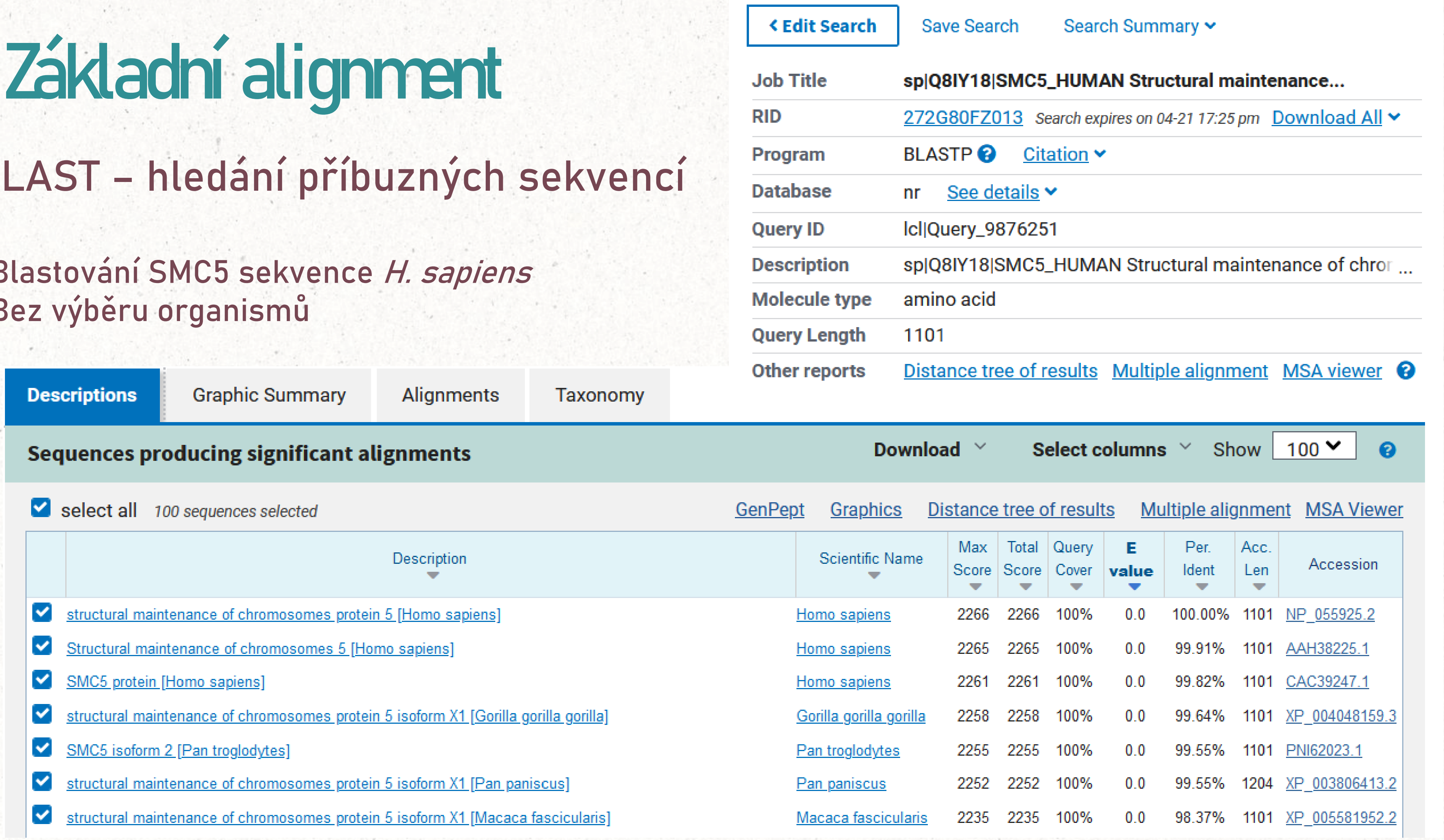
Výběr organismů →

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Základní alignment

- BLAST – hledání příbuzných sekvencí

Blastování SMC5 sekvence *H. sapiens*
Bez výběru organismů

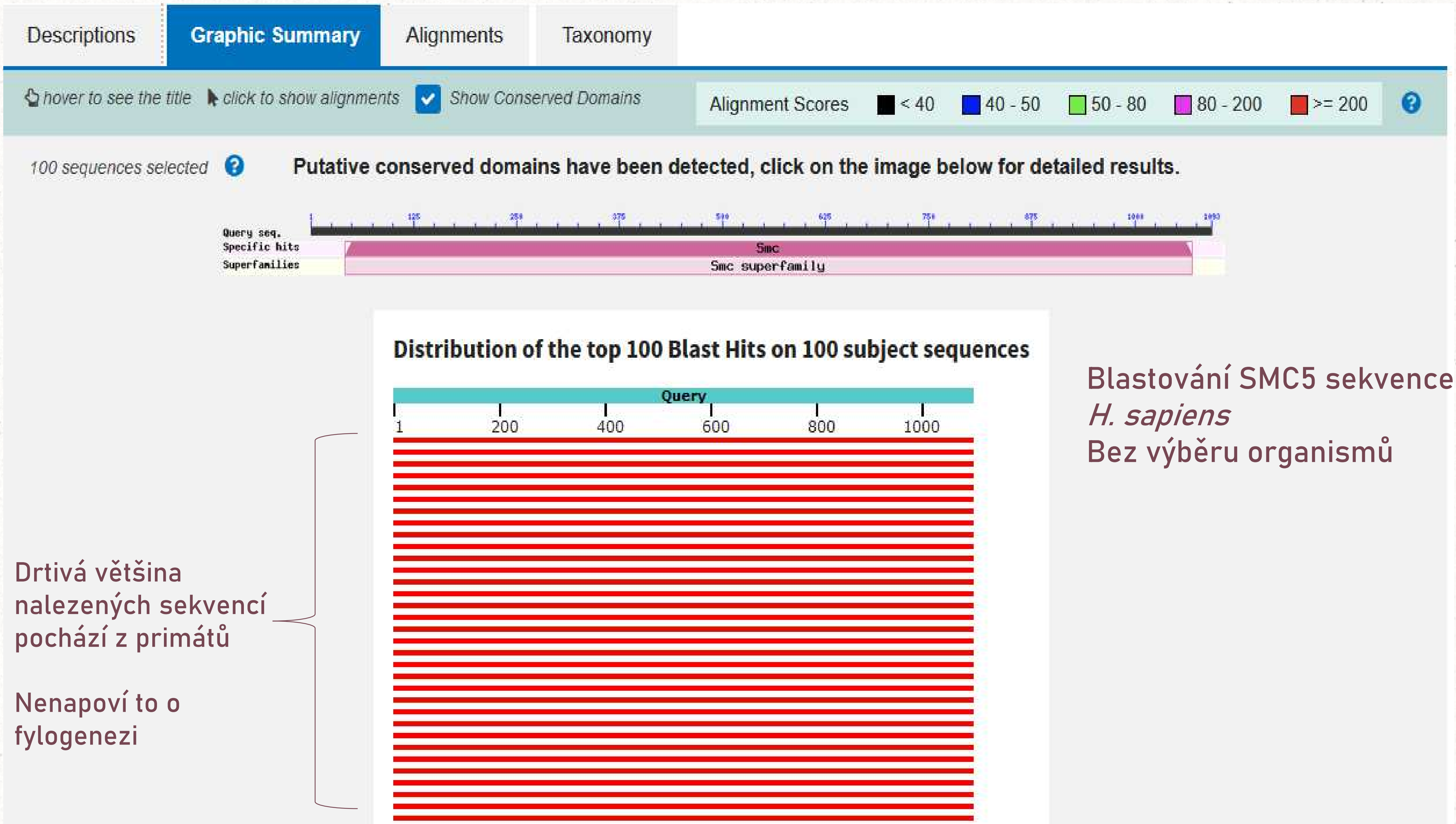


Job Title	**sp|Q8IY18|SMC5_HUMAN Structural maintenance...**
RID	272G80FZ013 *Search expires on 04-21 17:25 pm* Download All ⌄
Program	BLASTP Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_9876251
Description	sp|Q8IY18|SMC5_HUMAN Structural maintenance of chror ...
Molecule type	amino acid
Query Length	1101
Other reports	Distance tree of results Multiple alignment MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	structural maintenance of chromosomes protein 5 [Homo sapiens]	Homo sapiens	2266	2266	100%	0.0	100.00%	1101	NP_055925.2
☑	Structural maintenance of chromosomes 5 [Homo sapiens]	Homo sapiens	2265	2265	100%	0.0	99.91%	1101	AAH38225.1
☑	SMC5 protein [Homo sapiens]	Homo sapiens	2261	2261	100%	0.0	99.82%	1101	CAC39247.1
☑	structural maintenance of chromosomes protein 5 isoform X1 [Gorilla gorilla gorilla]	Gorilla gorilla gorilla	2258	2258	100%	0.0	99.64%	1101	XP_004048159.3
☑	SMC5 isoform 2 [Pan troglodytes]	Pan troglodytes	2255	2255	100%	0.0	99.55%	1101	PNI62023.1
☑	structural maintenance of chromosomes protein 5 isoform X1 [Pan paniscus]	Pan paniscus	2252	2252	100%	0.0	99.55%	1204	XP_003806413.2
☑	structural maintenance of chromosomes protein 5 isoform X1 [Macaca fascicularis]	Macaca fascicularis	2235	2235	100%	0.0	98.37%	1101	XP_005581952.2



100 sequences selected

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

Distribution of the top 100 Blast Hits on 100 subject sequences

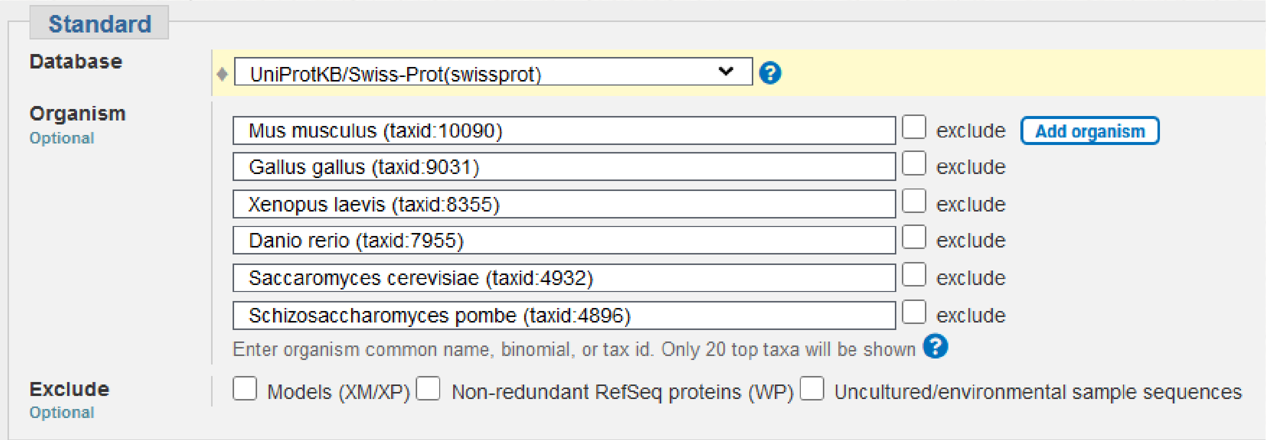
Blastování SMC5 sekvence
H. sapiens
Bez výběru organismů

Drtivá většina nalezených sekvencí pochází z primátů

Nenapoví to o fylogenezi

Základní alignment

Blastování SMC5 sekvence *H. sapiens s* výběrem organismů



Standard

Database: UniProtKB/Swiss-Prot(swissprot)

Organism Optional

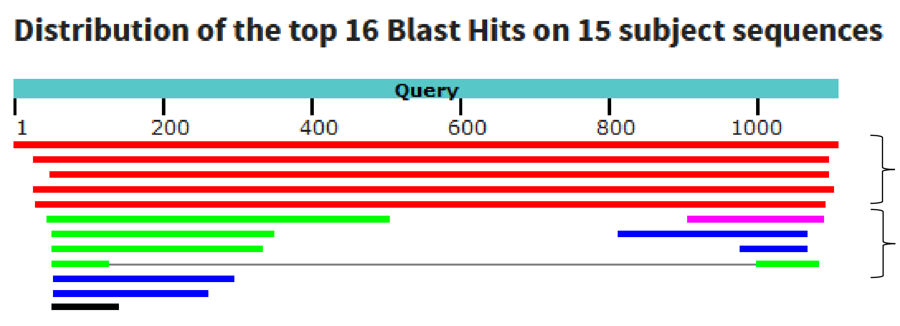
- Mus musculus (taxid:10090) ☐ exclude
- Gallus gallus (taxid:9031) ☐ exclude
- Xenopus laevis (taxid:8355) ☐ exclude
- Danio rerio (taxid:7955) ☐ exclude
- Saccaromyces cerevisiae (taxid:4932) ☐ exclude
- Schizosaccharomyces pombe (taxid:4896) ☐ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Výběr organismů
Homo sapiens
Mus musculus
Gallus gallus
Xenopus laevis
Danio rerio
S. cerevisiae
S. pombe

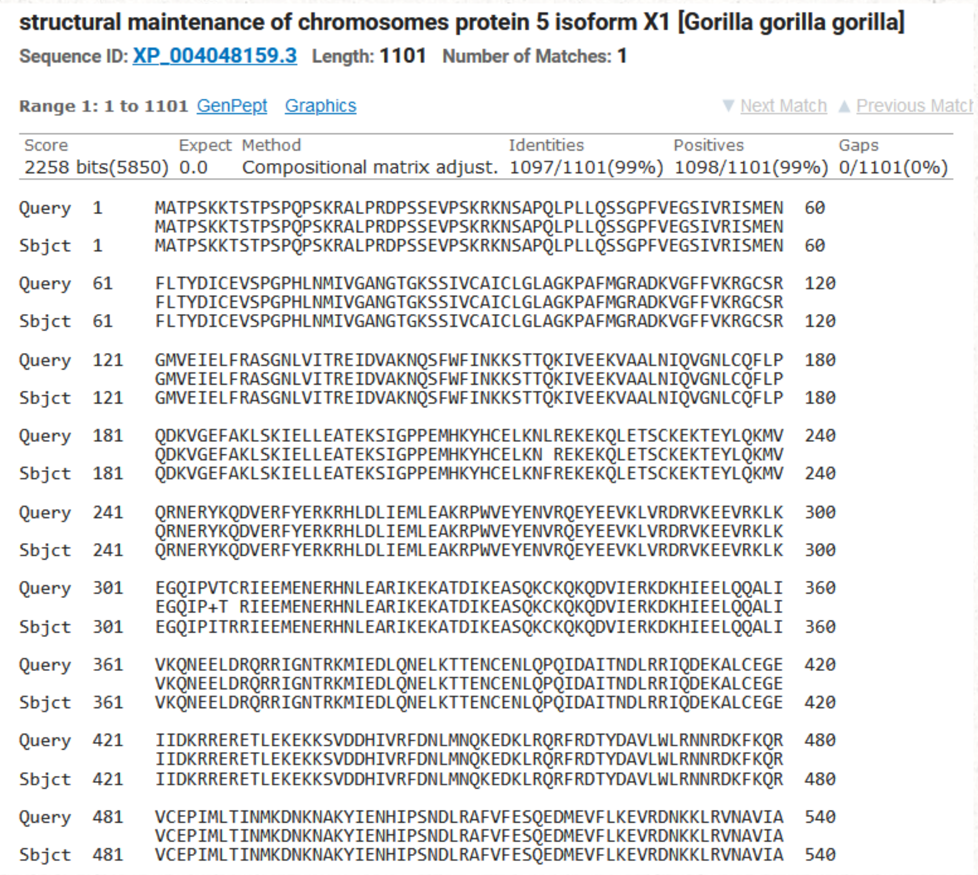
Vybrat i vzdálenější organismy. Čím více sekvencí, tím lépe



SMC5 proteiny vzdálených organismů

SMC6 proteiny

Porovnání sekvence proteinu SMC5
Homo sapiens a Gorilla gorilla



structural maintenance of chromosomes protein 5 isoform X1 [Gorilla gorilla gorilla]

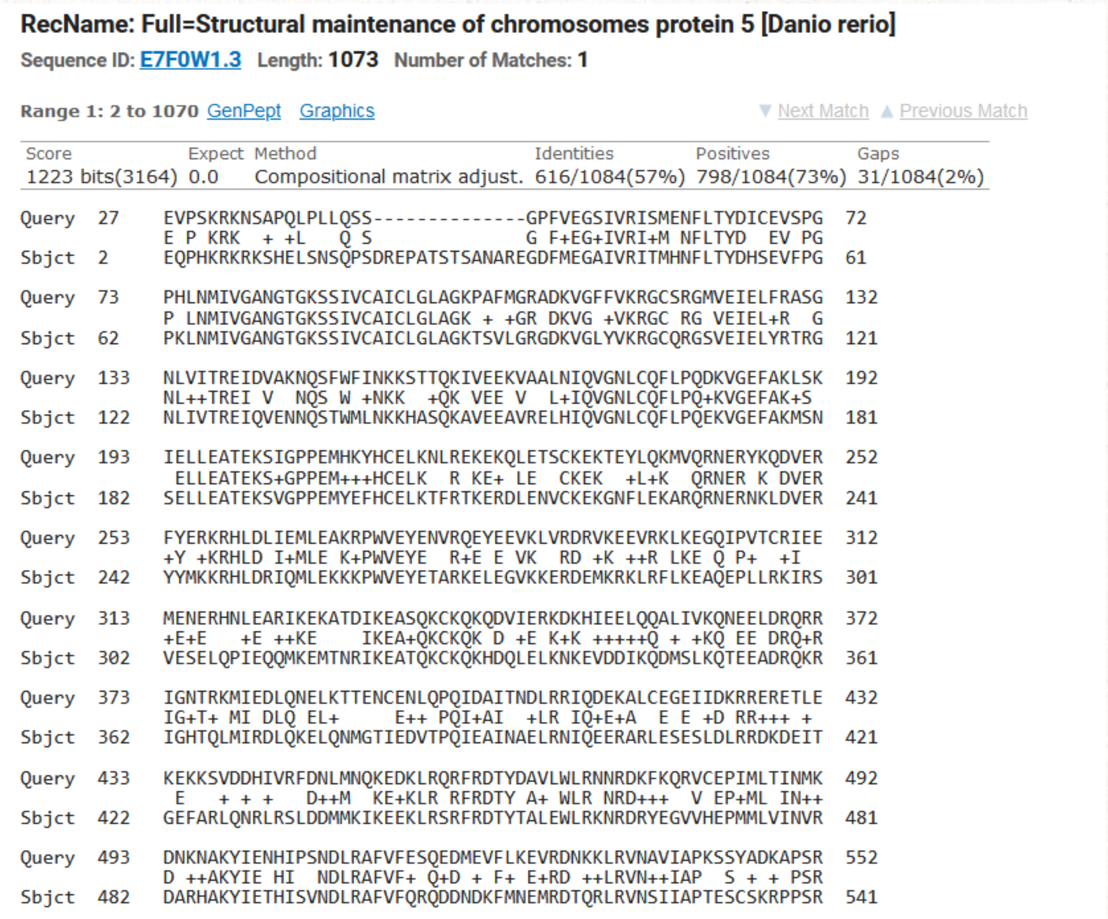
Sequence ID: XP_004048159.3 **Length:** 1101 **Number of Matches:** 1

Range 1: 1 to 1101 GenPept Graphics ▼ Next Match ▲ Previous Matcl

Score	Expect	Method	Identities	Positives	Gaps
2258 bits(5850)	0.0	Compositional matrix adjust.	1097/1101(99%)	1098/1101(99%)	0/1101(0%)

```
Query  1    MATPSKKTSTPSPQPSKRALPRDPSSEVPSKRKNSAPQLPLLQSSGPFVEGSIVRISMEN  60
            MATPSKKTSTPSPQPSKRALPRDPSSEVPSKRKNSAPQLPLLQSSGPFVEGSIVRISMEN
Sbjct  1    MATPSKKTSTPSPQPSKRALPRDPSSEVPSKRKNSAPQLPLLQSSGPFVEGSIVRISMEN  60

Query  61   FLTYDICEVSPGPHLNMIVGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSR  120
            FLTYDICEVSPGPHLNMIVGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSR
Sbjct  61   FLTYDICEVSPGPHLNMIVGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSR  120

Query  121  GMVEIELFRASGNLVITREIDVAKNQSFWFINKKSTTQKIVEEKVAALNIQVGNLCQFLP  180
            GMVEIELFRASGNLVITREIDVAKNQSFWFINKKSTTQKIVEEKVAALNIQVGNLCQFLP
Sbjct  121  GMVEIELFRASGNLVITREIDVAKNQSFWFINKKSTTQKIVEEKVAALNIQVGNLCQFLP  180

Query  181  QDKVGEFAKLSKIELLEATEKSIGPPEMHKYHCELKNLREKEKQLETSCKEKTEYLQKMV  240
            QDKVGEFAKLSKIELLEATEKSIGPPEMHKYHCELKN REKEKQLETSCKEKTEYLQKMV
Sbjct  181  QDKVGEFAKLSKIELLEATEKSIGPPEMHKYHCELKNFREKEKQLETSCKEKTEYLQKMV  240

Query  241  QRNERYKQDVERFYERKRHLDLIEMLEAKRPWVEYENVRQEYEEVKLVRDRVKEEVRKLK  300
            QRNERYKQDVERFYERKRHLDLIEMLEAKRPWVEYENVRQEYEEVKLVRDRVKEEVRKLK
Sbjct  241  QRNERYKQDVERFYERKRHLDLIEMLEAKRPWVEYENVRQEYEEVKLVRDRVKEEVRKLK  300

Query  301  EGQIPVTCRIEEMENERHNLEARIKEKATDIKEASQKCKQKQDVIERKDKHIEELQQALI  360
            EGQIP+T RIEEMENERHNLEARIKEKATDIKEASQKCKQKQDVIERKDKHIEELQQALI
Sbjct  301  EGQIPITRRIEEMENERHNLEARIKEKATDIKEASQKCKQKQDVIERKDKHIEELQQALI  360

Query  361  VKQNEELDRQRRIGNTRKMIEDLQNELKTTENCENLQPQIDAITNDLRRIQDEKALCEGE  420
            VKQNEELDRQRRIGNTRKMIEDLQNELKTTENCENLQPQIDAITNDLRRIQDEKALCEGE
Sbjct  361  VKQNEELDRQRRIGNTRKMIEDLQNELKTTENCENLQPQIDAITNDLRRIQDEKALCEGE  420

Query  421  IIDKRRERETLEKEKKSVDDHIVRFDNLMNQKEDKLRQRFRDTYDAVLWLRNNRDKFKQR  480
            IIDKRRERETLEKEKKSVDDHIVRFDNLMNQKEDKLRQRFRDTYDAVLWLRNNRDKFKQR
Sbjct  421  IIDKRRERETLEKEKKSVDDHIVRFDNLMNQKEDKLRQRFRDTYDAVLWLRNNRDKFKQR  480

Query  481  VCEPIMLTINMKDNKNAKYIENHIPSNDLRAFVFESQEDMEVFLKEVRDNKKLRVNAVIA  540
            VCEPIMLTINMKDNKNAKYIENHIPSNDLRAFVFESQEDMEVFLKEVRDNKKLRVNAVIA
Sbjct  481  VCEPIMLTINMKDNKNAKYIENHIPSNDLRAFVFESQEDMEVFLKEVRDNKKLRVNAVIA  540
```

Porovnání sekvence proteinu SMC5
Homo sapiens sapiens a *Danio rerio*



RecName: Full=Structural maintenance of chromosomes protein 5 [Danio rerio]

Sequence ID: E7F0W1.3 **Length:** 1073 **Number of Matches:** 1

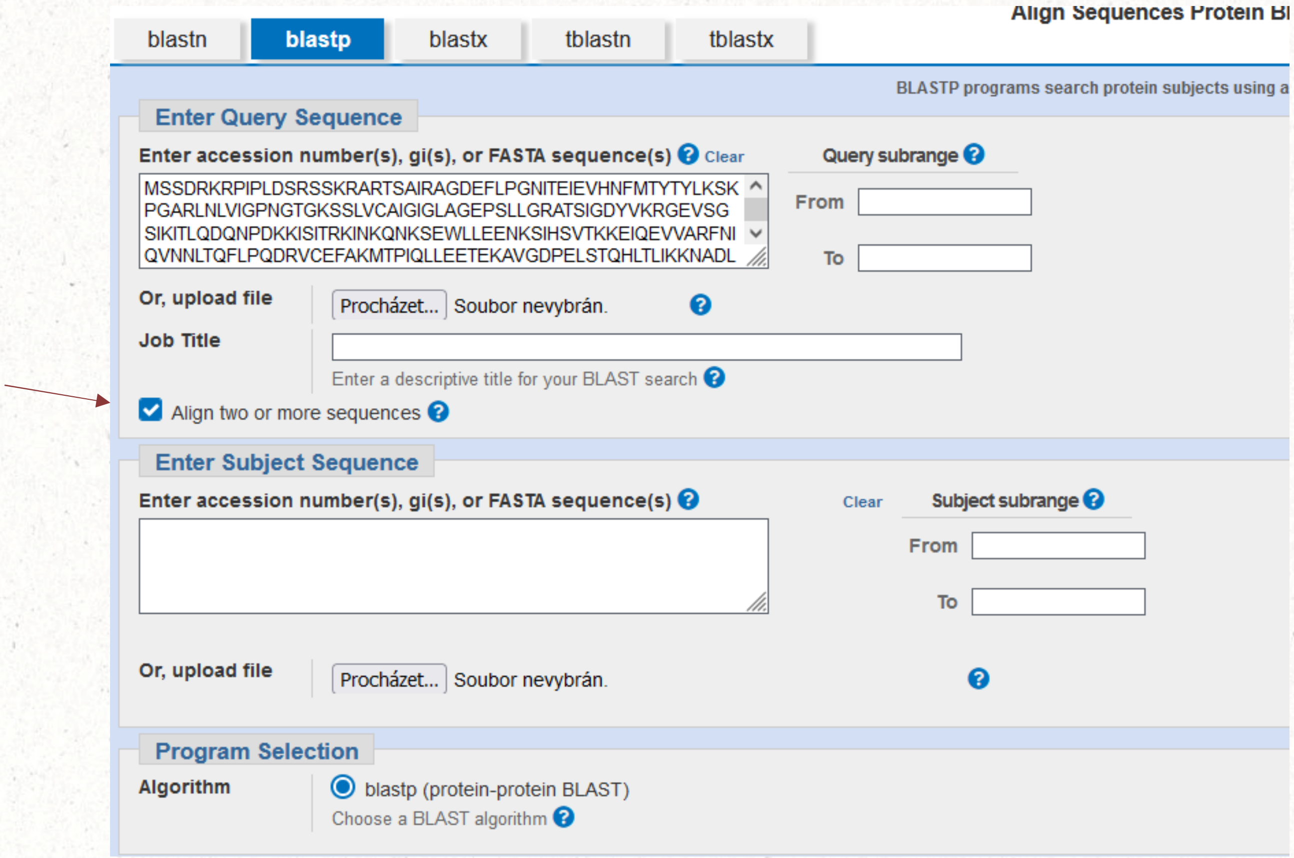
Range 1: 2 to 1070 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1223 bits(3164)	0.0	Compositional matrix adjust.	616/1084(57%)	798/1084(73%)	31/1084(2%)

```
Query  27   EVPSKRKNSAPQLPLLQSS--------------GPFVEGSIVRISMENFLTYDICEVSPG  72
            E P KRK  + +L   Q S              G F+EG+IVRI+M NFLTYD  EV PG
Sbjct  2    EQPHKRKRKSHELSNSQPSDREPATSTSANAREGDFMEGAIVRITMHNFLTYDHSEVFPG  61

Query  73   PHLNMIVGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSRGMVEIELFRASG  132
            P LNMIVGANGTGKSSIVCAICLGLAGK + +GR DKVG +VKRGC RG VEIEL+R  G
Sbjct  62   PKLNMIVGANGTGKSSIVCAICLGLAGKTSVLGRGDKVGLYVKRGCQRGSVEIELYRTRG  121

Query  133  NLVITREIDVAKNQSFWFINKKSTTQKIVEEKVAALNIQVGNLCQFLPQDKVGEFAKLSK  192
            NL++TREI V  NQS W +NKK  +QK VEE V  L+IQVGNLCQFLPQ+KVGEFAK+S
Sbjct  122  NLIVTREIQVENNQSTWMLNKKHASQKAVEEAVRELHIQVGNLCQFLPQEKVGEFAKMSN  181

Query  193  IELLEATEKSIGPPEMHKYHCELKNLREKEKQLETSCKEKTEYLQKMVQRNERYKQDVER  252
             ELLEATEKS+GPPEM+++HCELK  R KE+ LE  CKEK  +L+K  QRNER K DVER
Sbjct  182  SELLEATEKSVGPPEMYEFHCELKTFRTKERDLENVCKEKGNFLEKARQRNERNKLDVER  241

Query  253  FYERKRHLDLIEMLEAKRPWVEYENVRQEYEEVKLVRDRVKEEVRKLKEGQIPVTCRIEE  312
            +Y +KRHLD I+MLE K+PWVEYE  R+E E VK  RD +K ++R LKE Q P+  +I
Sbjct  242  YYMKKRHLDRIQMLEKKKPWVEYETARKELEGVKKERDEMKRKLRFLKEAQEPLLRKIRS  301

Query  313  MENERHNLEARIKEKATDIKEASQKCKQKQDVIERKDKHIEELQQALIVKQNEELDRQRR  372
            +E+E   +E ++KE    IKEA+QKCKQK D +E K+K +++++Q + +KQ EE DRQ+R
Sbjct  302  VESELQPIEQQMKEMTNRIKEATQKCKQKHDQLELKNKEVDDIKQDMSLKQTEEADRQKR  361

Query  373  IGNTRKMIEDLQNELKTTENCENLQPQIDAITNDLRRIQDEKALCEGEIIDKRRERETLE  432
            IG+T+ MI DLQ EL+     E++ PQI+AI  +LR IQ+E+A  E E +D RR+++ +
Sbjct  362  IGHTQLMIRDLQKELQNMGTIEDVTPQIEAINAELRNIQEERARLESESLDLRRDKDEIT  421

Query  433  KEKKSVDDHIVRFDNLMNQKEDKLRQRFRDTYDAVLWLRNNRDKFKQRVCEPIMLTINMK  492
             E   + + +   D++M  KE+KLR RFRDTY A+ WLR NRD+++  V EP+ML IN++
Sbjct  422  GEFARLQNRLRSLDDMMKIKEEKLRSRFRDTYTALEWLRKNRDRYEGVVHEPMMLVINVR  481

Query  493  DNKNAKYIENHIPSNDLRAFVFESQEDMEVFLKEVRDNKKLRVNAVIAPKSSYADKAPSR  552
            D ++AKYIE HI  NDLRAFVF+ Q+D + F+ E+RD ++LRVN++IAP  S + + PSR
Sbjct  482  DARHAKYIETHISVNDLRAFVFQRQDDNDKFMNEMRDTQRLRVNSIIAPTESCSKRPPSR  541
```

Základní alignment

- Pair-Alignment příbuzných sekvencí v BLAST



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
MSSDRKRPIPLDSRSSKRARTSAIRAGDEFLPGNITEIEVHNFMTYTYLKSK
PGARLNLVIGPNGTGKSSLVCAIGIGLAGEPSLLGRATSIGDYVKRGEVSG
SIKITLQDQNPDKKISITRKINKQNKSEWLLEENKSIHSVTKKEIQEVVARFNI
QVNNLTQFLPQDRVCEFAKMTPIQLLEETEKAVGDPELSTQHLTLIKKNADL
```

Query subrange
From
To

Or, upload file: Procházet... Soubor nevybrán.

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Subject subrange
From
To

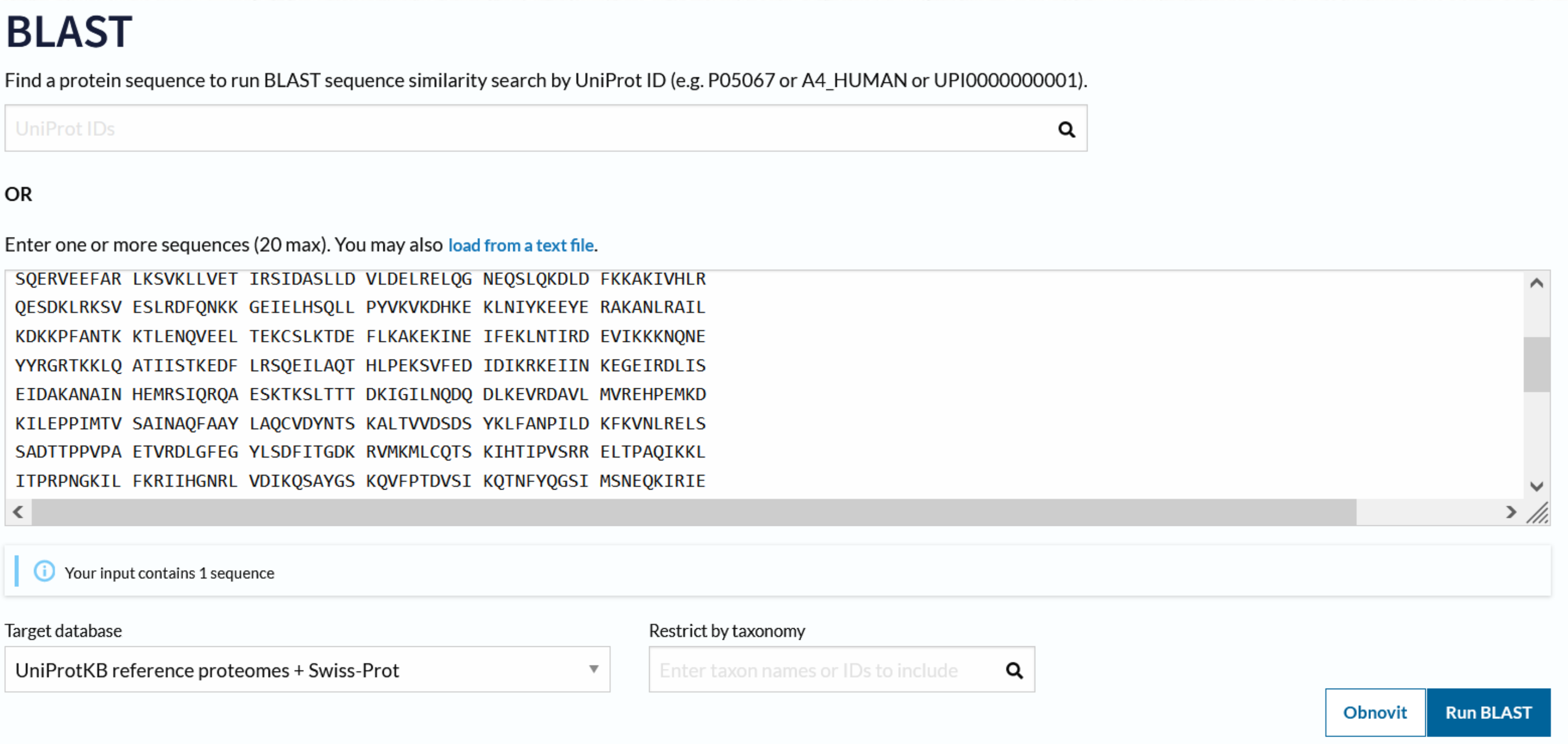
Or, upload file: Procházet... Soubor nevybrán.

Program Selection

Algorithm: ● blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST na UniProt



BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

OR

Enter one or more sequences (20 max). You may also load from a text file.

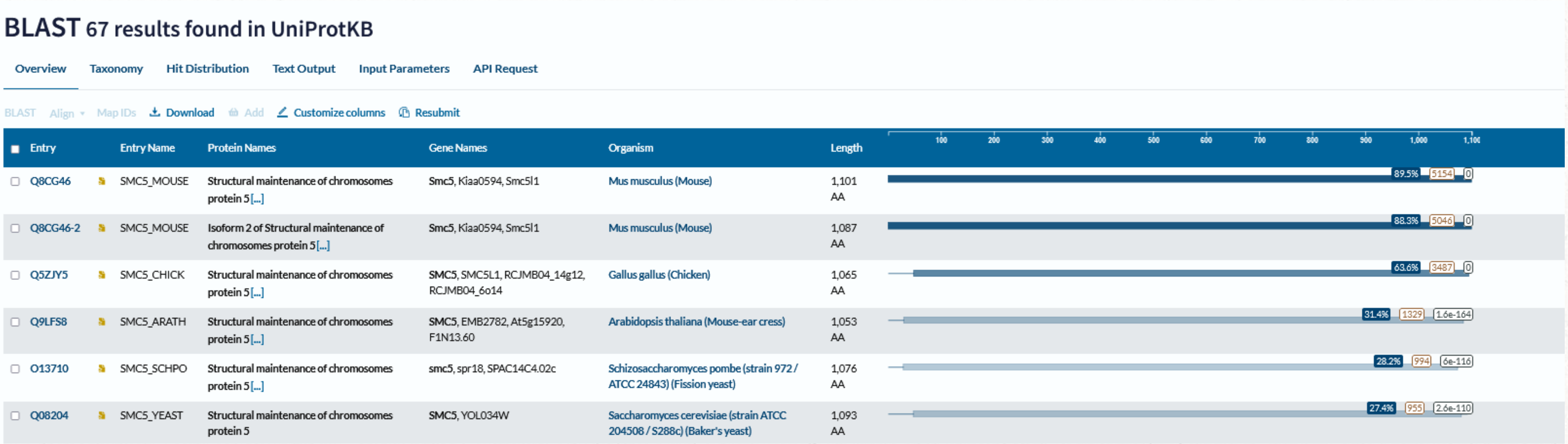
```
SQERVEEFAR LKSVKLLVET IRSIDASLLD VLDELRELQG NEQSLQKDLD FKKAKIVHLR
QESDKLRKSV ESLRDFQNKK GEIELHSQLL PYVKVKDHKE KLNIYKEEYE RAKANLRAIL
KDKKPFANTK KTLENQVEEL TEKCSLKTDE FLKAKEKINE IFEKLNTIRD EVIKKKNQNE
YYRGRTKKLQ ATIISTKEDF LRSQEILAQT HLPEKSVFED IDIKRKEIIN KEGEIRDLIS
EIDAKANAIN HEMRSIQRQA ESKTKSLTTT DKIGILNQDQ DLKEVRDAVL MVREHPEMKD
KILEPPIMTV SAINAQFAAY LAQCVDYNTS KALTVVDSDS YKLFANPILD KFKVNLRELS
SADTTPPVPA ETVRDLGFEG YLSDFITGDK RVMKMLCQTS KIHTIPVSRR ELTPAQIKKL
ITPRPNGKIL FKRIIHGNRL VDIKQSAYGS KQVFPTDVSI KQTNFYQGSI MSNEQKIRIE
```

Your input contains 1 sequence

Target database: UniProtKB reference proteomes + Swiss-Prot

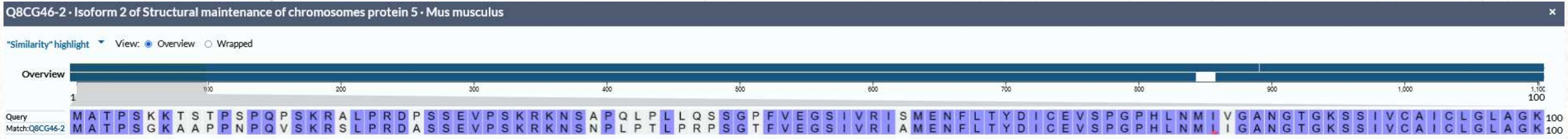
https://www.uniprot.org/blast

BLAST na UniProt



BLAST 67 results found in UniProtKB

☐	Entry	Entry Name	Protein Names	Gene Names	Organism	Length	100 200 300 400 500 600 700 800 900 1,000 1,100
☐	Q8CG46	SMC5_MOUSE	Structural maintenance of chromosomes protein 5 [...]	Smc5, Kiaa0594, Smc5l1	Mus musculus (Mouse)	1,101 AA	89.5% 5154 0
☐	Q8CG46-2	SMC5_MOUSE	Isoform 2 of Structural maintenance of chromosomes protein 5 [...]	Smc5, Kiaa0594, Smc5l1	Mus musculus (Mouse)	1,087 AA	88.3% 5046 0
☐	Q5ZJY5	SMC5_CHICK	Structural maintenance of chromosomes protein 5 [...]	SMC5, SMC5L1, RCJMB04_14g12, RCJMB04_6o14	Gallus gallus (Chicken)	1,065 AA	63.6% 3487 0
☐	Q9LFS8	SMC5_ARATH	Structural maintenance of chromosomes protein 5 [...]	SMC5, EMB2782, At5g15920, F1N13.60	Arabidopsis thaliana (Mouse-ear cress)	1,053 AA	31.4% 1329 1.6e-164
☐	O13710	SMC5_SCHPO	Structural maintenance of chromosomes protein 5 [...]	smc5, spr18, SPAC14C4.02c	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	1,076 AA	28.2% 994 6e-116
☐	Q08204	SMC5_YEAST	Structural maintenance of chromosomes protein 5	SMC5, YOL034W	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1,093 AA	27.4% 955 2.6e-110

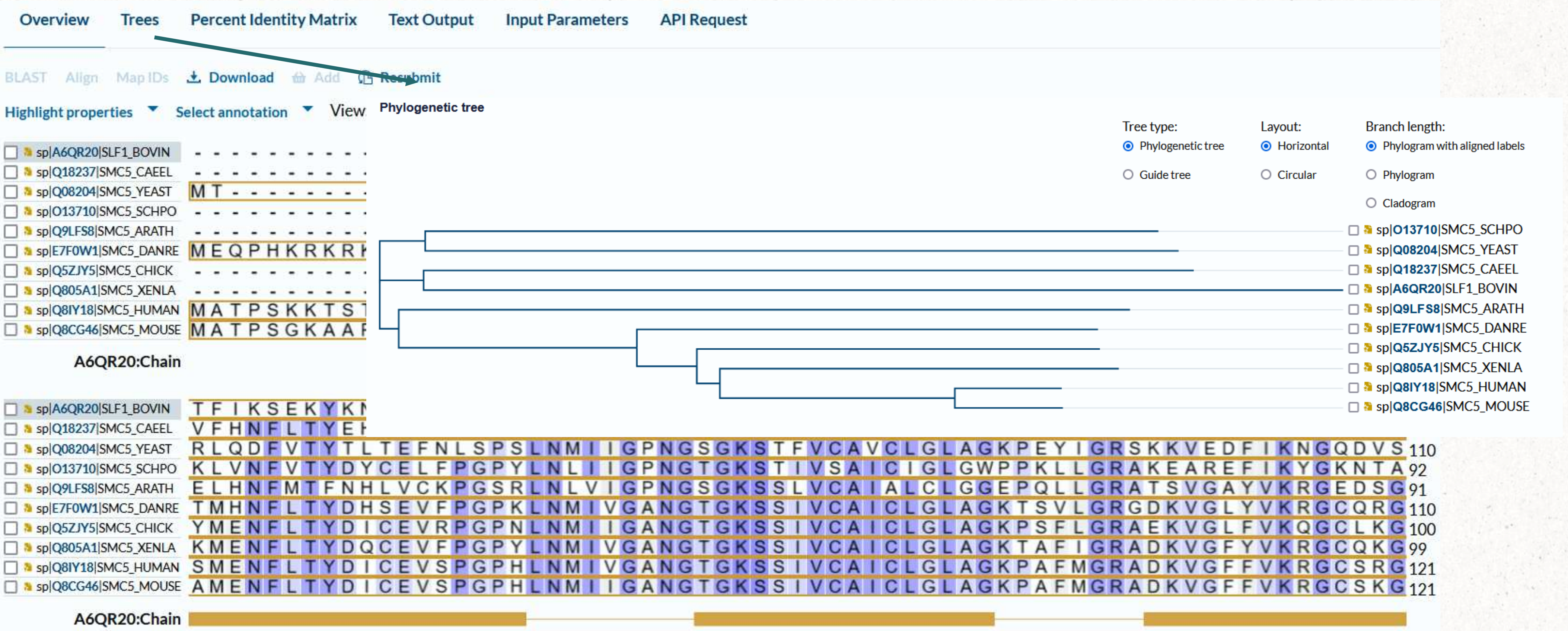


Q8CG46-2 · Isoform 2 of Structural maintenance of chromosomes protein 5 · Mus musculus

```
Query            M A T P S K K T S T P S P Q P S K R A L P R D P S S E V P S K R K N S A P Q L P L L Q S S G P F V E G S I V R I S M E N F L T Y D I C E V S P G P H L N M I V G A N G T G K S S I V C A I C L G L A G K 100
Match:Q8CG46-2   M A T P S G K A A P P N P Q V S K R S L P R D A S S E V P S K R K N S N P L P T L P R P S G T F V E G S I V R I A M E N F L T Y D I C E V S P G P H L N M L I G A N G T G K S S I V C A I C L G L A G K 100
```

https://www.uniprot.org/blast

Sekvenční podobnost/alignment na UniProt



Tree type:
● Phylogenetic tree
○ Guide tree

Layout:
● Horizontal
○ Circular

Branch length:
● Phylogram with aligned labels
○ Phylogram
○ Cladogram

```
☐ sp|A6QR20|SLF1_BOVIN   - - - - - - - - - -
☐ sp|Q18237|SMC5_CAEEL   - - - - - - - - - -
☐ sp|Q08204|SMC5_YEAST   M T - - - - - - - -
☐ sp|O13710|SMC5_SCHPO   - - - - - - - - - -
☐ sp|Q9LFS8|SMC5_ARATH   - - - - - - - - - -
☐ sp|E7F0W1|SMC5_DANRE   M E Q P H K R K R
☐ sp|Q5ZJY5|SMC5_CHICK   - - - - - - - - - -
☐ sp|Q805A1|SMC5_XENLA   - - - - - - - - - -
☐ sp|Q8IY18|SMC5_HUMAN   M A T P S K K T S
☐ sp|Q8CG46|SMC5_MOUSE   M A T P S G K A A
```

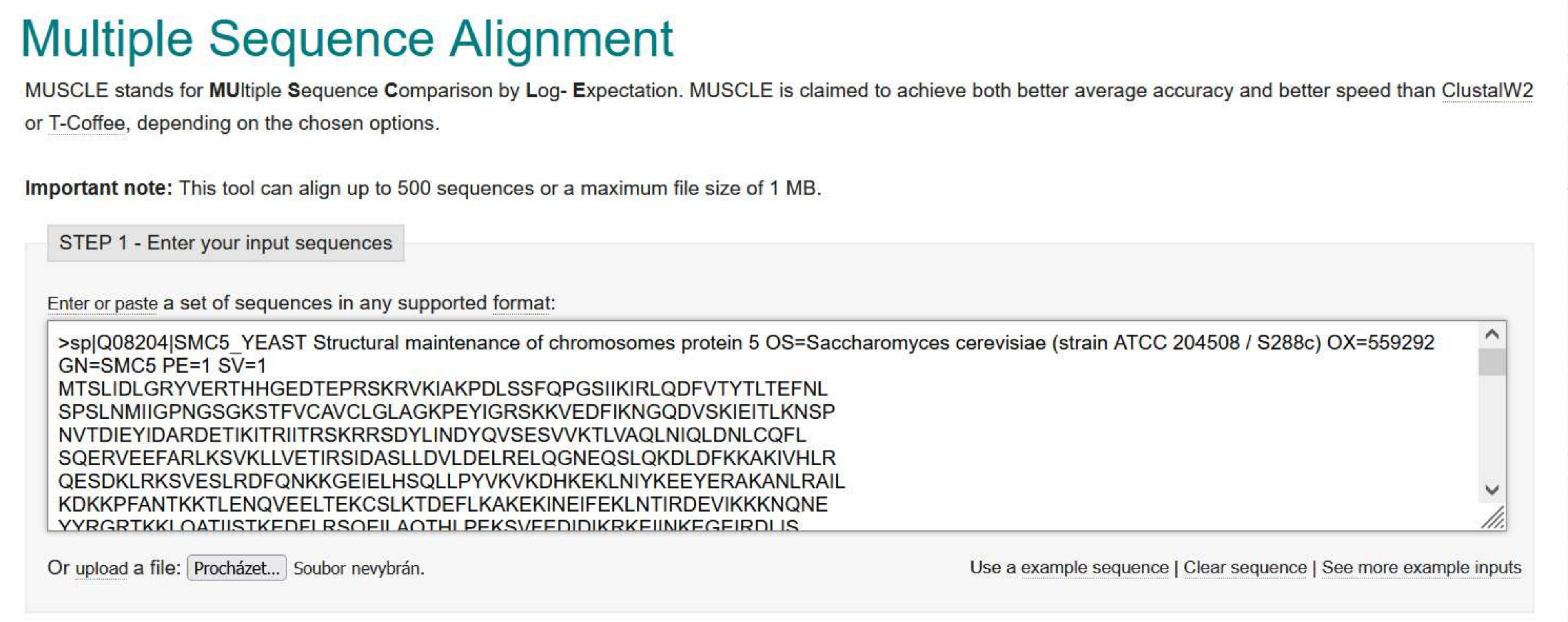
```
☐ sp|O13710|SMC5_SCHPO
☐ sp|Q08204|SMC5_YEAST
☐ sp|Q18237|SMC5_CAEEL
☐ sp|A6QR20|SLF1_BOVIN
☐ sp|Q9LFS8|SMC5_ARATH
☐ sp|E7F0W1|SMC5_DANRE
☐ sp|Q5ZJY5|SMC5_CHICK
☐ sp|Q805A1|SMC5_XENLA
☐ sp|Q8IY18|SMC5_HUMAN
☐ sp|Q8CG46|SMC5_MOUSE
```

```
☐ sp|A6QR20|SLF1_BOVIN   T F I K S E K Y K
☐ sp|Q18237|SMC5_CAEEL   V F H N F L T Y E
☐ sp|Q08204|SMC5_YEAST   R L Q D F V T Y T L T E F N L S P S L N M I I G P N G S G K S T F V C A V C L G L A G K P E Y I G R S K K V E D F I K N G Q D V S 110
☐ sp|O13710|SMC5_SCHPO   K L V N F V T Y D Y C E L F P G P Y L N L I I G P N G T G K S T I V S A I C I G L G W P P K L L G R A K E A R E F I K Y G K N T A 92
☐ sp|Q9LFS8|SMC5_ARATH   E L H N F M T F N H L V C K P G S R L N L V I G P N G S G K S S L V C A I A L C L G G E P Q L L G R A T S V G A Y V K R G E D S G 91
☐ sp|E7F0W1|SMC5_DANRE   T M H N F L T Y D H S E V F P G P K L N M I V G A N G T G K S S I V C A I C L G L A G K T S V L G R G D K V G L Y V K R G C Q R G 110
☐ sp|Q5ZJY5|SMC5_CHICK   Y M E N F L T Y D I C E V R P G P N L N M I I G A N G T G K S S I V C A I C L G L A G K P S F L G R A E K V G L F V K Q G C L K G 100
☐ sp|Q805A1|SMC5_XENLA   K M E N F L T Y D Q C E V F P G P Y L N M I V G A N G T G K S S I V C A I C L G L A G K T A F I G R A D K V G F Y V K R G C Q K G 99
☐ sp|Q8IY18|SMC5_HUMAN   S M E N F L T Y D I C E V S P G P H L N M I V G A N G T G K S S I V C A I C L G L A G K P A F M G R A D K V G F F V K R G C S R G 121
☐ sp|Q8CG46|SMC5_MOUSE   A M E N F L T Y D I C E V S P G P H L N M I I G A N G T G K S S I V C A I C L G L A G K P A F M G R A D K V G F F V K R G C S K G 121
```

https://www.uniprot.org/align

Multiple Sequence Alignment

https://www.ebi.ac.uk/Tools/msa/muscle/



Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

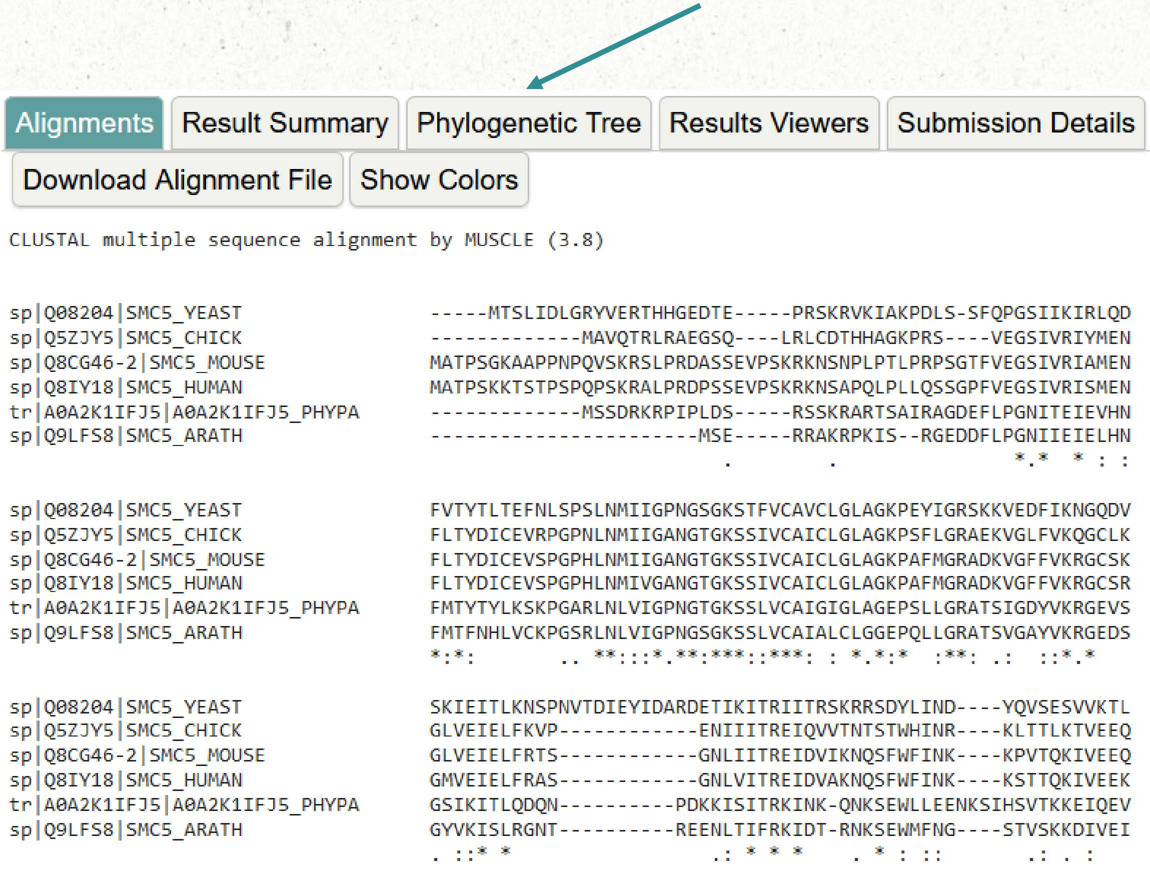
Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

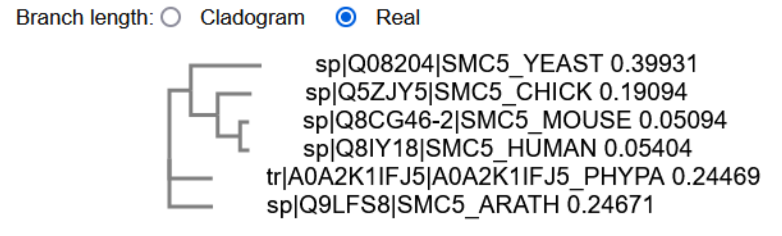
Enter or paste a set of sequences in any supported format:

```
>sp|Q08204|SMC5_YEAST Structural maintenance of chromosomes protein 5 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292
GN=SMC5 PE=1 SV=1
MTSLIDLGRYVERTHHGEDTEPRSKRVKIAKPDLSSFQPGSIIKIRLQDFVTYTLTEFNL
SPSLNMIIGPNGSGKSTFVCAVCLGLAGKPEYIGRSKKVEDFIKNGQDVSKIEITLKNSP
NVTDIEYIDARDETIKITRIITRSKRRSDYLINDYQVSESVVKTLVAQLNIQLDNLCQFL
SQERVEEFARLKSVKLLVETIRSIDASLLDVLDELRELQGNEQSLQKDLDFKKAKIVHLR
QESDKLRKSVESLRDFQNKKGEIELHSQLLPYVKVKDHKEKLNIYKEEYERAKANLRAIL
KDKKPFANTKKTLENQVEELTEKCSLKTDEFLKAKEKINEIFEKLNTIRDEVIKKKNQNE
YYRGRTKKLQATIISTKEDFLRSQEILAQTHLPEKSVFEDIDIKRKEIINKEGEIRDLIS
```

Multiple Sequence Alignment



```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


sp|Q08204|SMC5_YEAST                -----MTSLIDLGRYVERTHHGEDTE-----PRSKRVKIAKPDLS-SFQPGSIIKIRLQD
sp|Q5ZJY5|SMC5_CHICK                -------------MAVQTRLRAEGSQ----LRLCDTHHAGKPRS----VEGSIVRIYMEN
sp|Q8CG46-2|SMC5_MOUSE              MATPSGKAAPPNPQVSKRSLPRDASSEVPSKRKNSNPLPTLPRPSGTFVEGSIVRIAMEN
sp|Q8IY18|SMC5_HUMAN                MATPSKKTSTPSPQPSKRALPRDPSSEVPSKRKNSAPQLPLLQSSGPFVEGSIVRISMEN
tr|A0A2K1IFJ5|A0A2K1IFJ5_PHYPA      -------------MSSDRKRPIPLDS-----RSSKRARTSAIRAGDEFLPGNITEIEVHN
sp|Q9LFS8|SMC5_ARATH                -----------------------MSE-----RRAKRPKIS--RGEDDFLPGNIIEIELHN
                                                             .        .               *.*  * : :

sp|Q08204|SMC5_YEAST                FVTYTLTEFNLSPSLNMIIGPNGSGKSTFVCAVCLGLAGKPEYIGRSKKVEDFIKNGQDV
sp|Q5ZJY5|SMC5_CHICK                FLTYDICEVRPGPNLNMIIGANGTGKSSIVCAICLGLAGKPSFLGRAEKVGLFVKQGCLK
sp|Q8CG46-2|SMC5_MOUSE              FLTYDICEVSPGPHLNMIIGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSK
sp|Q8IY18|SMC5_HUMAN                FLTYDICEVSPGPHLNMIVGANGTGKSSIVCAICLGLAGKPAFMGRADKVGFFVKRGCSR
tr|A0A2K1IFJ5|A0A2K1IFJ5_PHYPA      FMTYTYLKSKPGARLNLVIGPNGTGKSSLVCAIGIGLAGEPSLLGRATSIGDYVKRGEVS
sp|Q9LFS8|SMC5_ARATH                FMTFNHLVCKPGSRLNLVIGPNGSGKSSLVCAIALCLGGEPQLLGRATSVGAYVKRGEDS
                                    *:*:       .. **:::*.**:***::***: : *.*:*  :**: .:  ::*.*

sp|Q08204|SMC5_YEAST                SKIEITLKNSPNVTDIEYIDARDETIKITRIITRSKRRSDYLIND----YQVSESVVKTL
sp|Q5ZJY5|SMC5_CHICK                GLVEIELFKVP------------ENIIITREIQVVTNTSTWHINR----KLTTLKTVEEQ
sp|Q8CG46-2|SMC5_MOUSE              GLVEIELFRTS------------GNLIITREIDVIKNQSFWFINK----KPVTQKIVEEQ
sp|Q8IY18|SMC5_HUMAN                GMVEIELFRAS------------GNLVITREIDVAKNQSFWFINK----KSTTQKIVEEK
tr|A0A2K1IFJ5|A0A2K1IFJ5_PHYPA      GSIKITLQDQN----------PDKKISITRKINK-QNKSEWLLEENKSIHSVTKKEIQEV
sp|Q9LFS8|SMC5_ARATH                GYVKISLRGNT----------REENLTIFRKIDT-RNKSEWMFNG----STVSKKDIVEI
                                    . ::* *                 .: * * *    . * : ::       .: . :
```



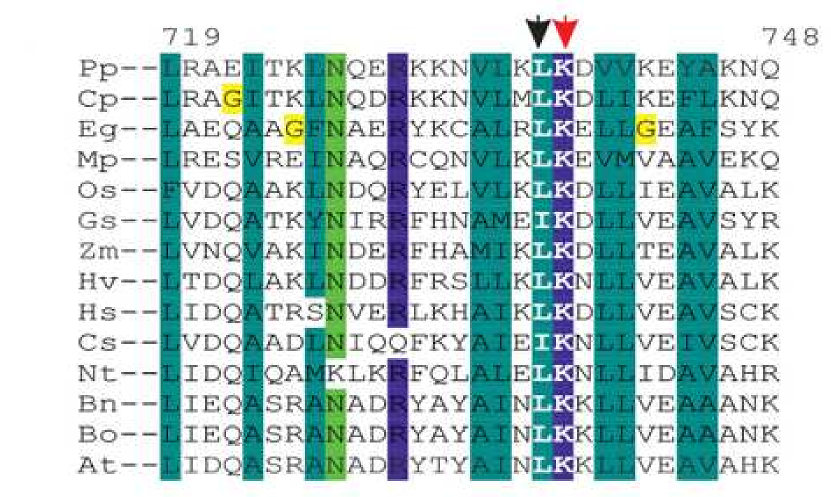
muscle-I20230503-064457-0144-4283286-p1m.clw

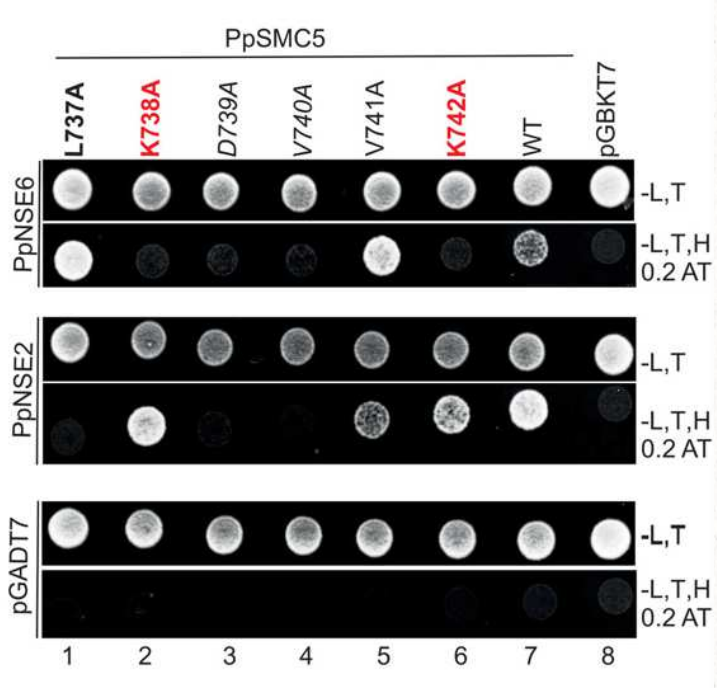
Prezentace výsledků o konzervovanosti (inspirace)

- Přenést základní sekvenci do wordu
- Postupně přidávat další sekvence
  - nejlépe i fylogeneticky vzdálenější – napoví o konzervovanosti
- Zaznačit aminokyseliny či pattern, který je podobný
  - Hydrofobní aminokyseliny: L,V,A,M,F,I,W
  - Záporně nabité aminokyseliny: D,E
  - Kladně nabité aminokyseliny: R,K
  - Nenabité aminokyseliny: T,S,C,N,Q
  - Malé aminokyseliny: P,G (otočka)

- Lze proložit i sekundární strukturou

Alignment podjednotky SMC5

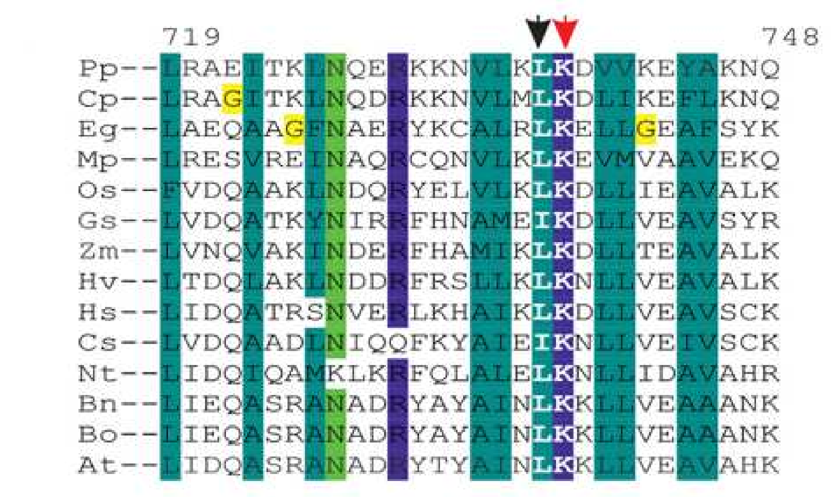




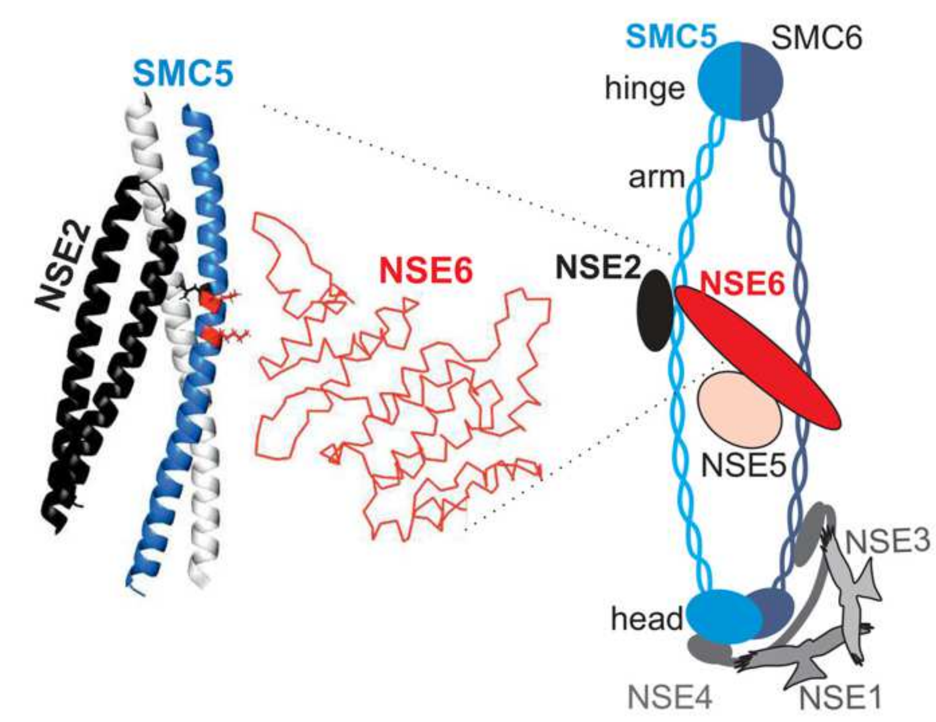
Hydrofobní aminokyseliny: L,V,A,M,F,I,W
Záporně nabité aminokyseliny: D,E
Kladně nabité aminokyseliny: R,K
Nenabité (neutrální) aminokyseliny: T,S,C,N,Q
Malé aminokyseliny: P,G (otočka)

doc. Jan Paleček
(Lelkes et al. 2023)

Alignment podjednotky SMC5



```
    719                                  748
Pp--LRAEITKLNQERKKNVLKLKDVVKEYAKNQ
Cp--LRAGITKLNQDRKKNVLMLKDLIKEFLKNQ
Eg--LAEQAAGFNAERYKCALRLKELLGEAFSYK
Mp--LRESVREINAQRCQNVLKLKEVMVAAVEKQ
Os--FVDQAAKLNDQRYELVLKLKDLLIEAVALK
Gs--LVDQATKYNIRRFHNAMEIKDLLVEAVSYR
Zm--LVNQVAKINDERFHAMIKLKDLLTEAVALK
Hv--LTDQLAKLNDDRFRSLLKLKNLLVEAVALK
Hs--LIDQATRSNVERLKHAIKLKDLLVEAVSCK
Cs--LVDQAADLNIQQFKYAIEIKNLLVEIVSCK
Nt--LIDQIQAMKLKRFQLALELKNLLIDAVAHR
Bn--LIEQASRANADRYAYAINLKKLLVEAAANK
Bo--LIEQASRANADRYAYAINLKKLLVEAAANK
At--LIDQASRANADRYTYAINLKKLLVEAVAHK
```



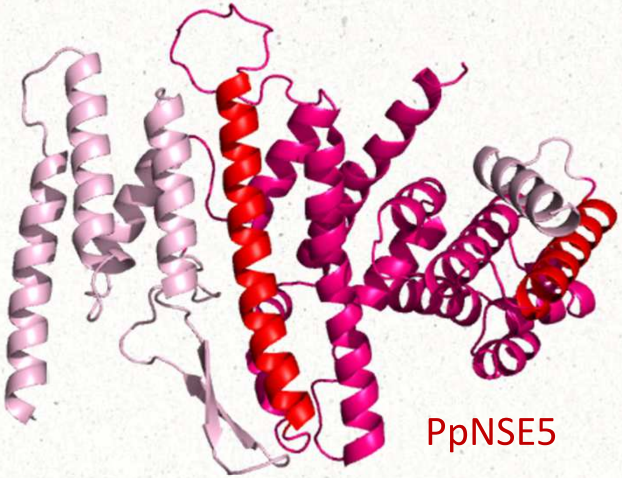
Hydrofobní aminokyseliny: L,V,A,M,F,I,W
Záporně nabité aminokyseliny: D,E
Kladně nabité aminokyseliny: R,K
Nenabité (neutrální) aminokyseliny: T,S,C,N,Q
Malé aminokyseliny: P,G (otočka)

doc. Jan Paleček
(Lelkes et al. 2023)

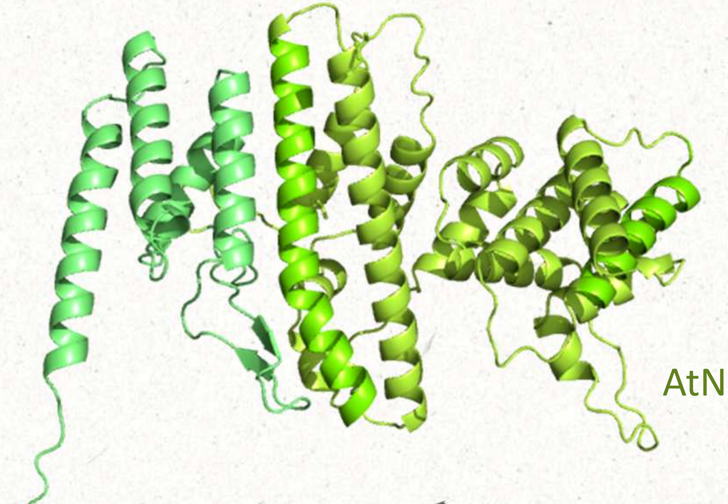
Pokročilý alignment

- BLAST nenajde příbuzné sekvence s vaším proteinem
- Nalezení proteinu, který by měl být homologem/ortologem
- Alignment příbuzných sekvencí v BLAST
- Přenesení výsledků do dokumentu (word)
- Postupné přidávání dalších sekvencí
- Využití informace
  - sekundární struktura (PSIPRED)
  - terciální struktura (Alphafold)
- Značení patternu

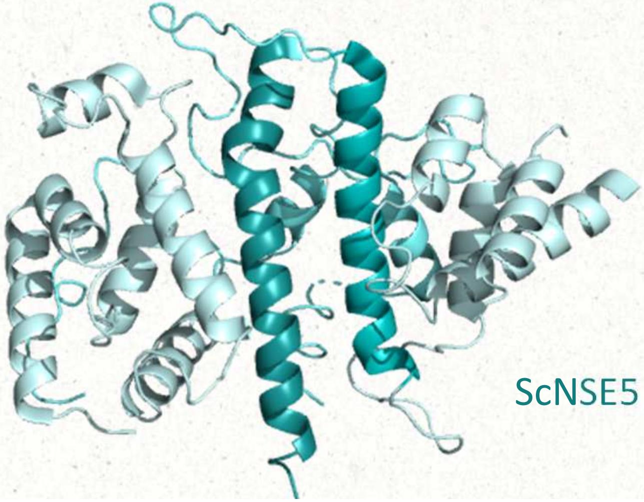
Strukturní podobnost podjednotky NSE5



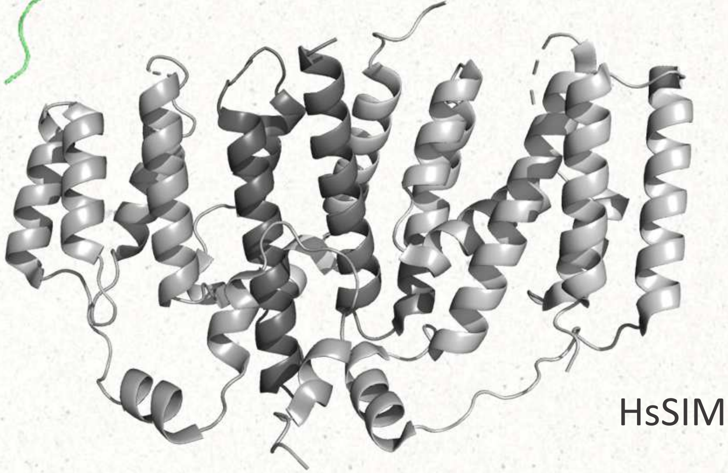
PpNSE5



AtNSE5



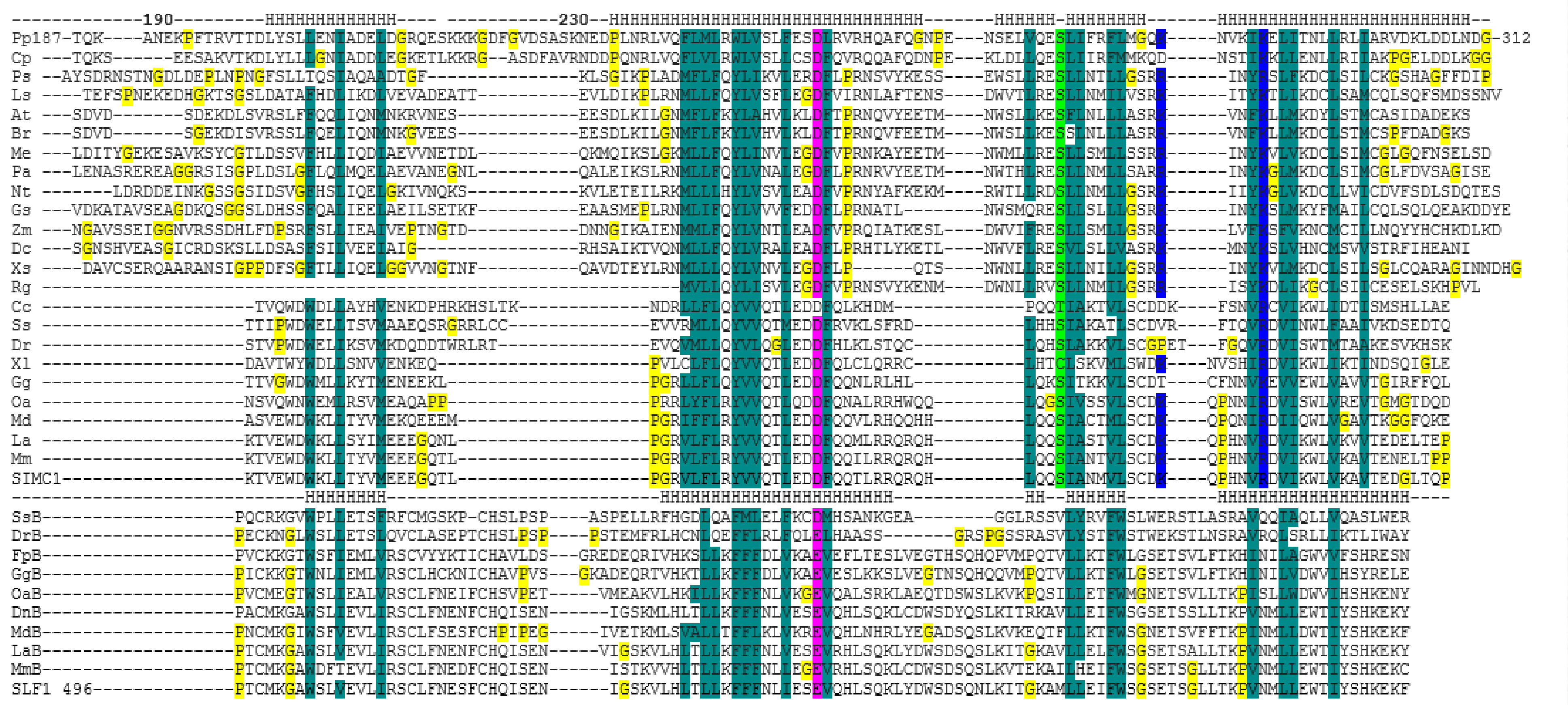
ScNSE5



HsSIMC1

Alphafold
doc. Jan Paleček
(unpublished)

Alignment podjednotky NSE5



```
-------------190---------HHHHHHHHHHHHH---- -----------230--HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------HHHHHH-HHHHHHHH-------HHHHHHHHHHHHHHHHHHHHHHHHHHH--
Pp187-TQK----ANEKPFTRVTTDLYSLLENIADELDGRQESKKKGDFGVDSASKNEDPLNRLVQFLMLRWLVSLFESDLRVRHQAFQGNPE---NSELVQESLIFRFLMGQK-----NVKIKELITNLLRLIARVDKLDDLNDG-312
Cp ---TQKS------EESAKVTKDLYLLLGNIADDLEGKETLKKRG--ASDFAVRNDDPQNRLVQFLVLRWLVSLLCSDFQVRQQAFQDNPE---KLDLLQESLIIRFMMKQD-----NSTIKKLLENLLRIIAKPGELDDLKGG
Ps --AYSDRNSTNGDLDEPLNPNGFSLLTQSIAQAADTGF---------------KLSGIKPLADMFLFQYLIKVLERDFLPRNSVYKESS----EWSLLRESLLNTLLGSRR------INYRSLFKDCLSILCKGSHAGFFDIP
Ls ----TEFSPNEKEDHGKTSGSLDATAFHDLIKDLVEVADEATT----------EVLDIKPLRNMLLFQYLVSFLEGDFVIRNLAFTENS----DWVTLRESLLNMIIVSRK------ITYKTLIKDCLSAMCQLSQFSMDSSNV
At ---SDVD-------SDEKDLSVRSLFFQQLIQNMNKRVNES------------EESDLKILGNMFLFKYLAHVLKLDFTPRNQVYEETM----NWSLLKESFLNLLLASRK------VNFKLLMKDYLSTMCASIDADEKS
Br ---SDVD-------SGEKDISVRSSLFQELIQNMNKGVEES------------EESDLKILGNMFLFKYLVHVLKLDFTPRNQVFEETM----NWSLLKESSLNLLLASRR------VNFKLLMKDCLSTMCSPFDADGKS
Me ---LDITYGEKESAVKSYCGTLDSSVFHLLIQDIAEVVNETDL----------QKMQIKSLGKMLLFQYLINVLEGDFVPRNKAYEETM----NWMLLRESLLSMLLSSRR------INYKVLVKDCLSIMCGLGQFNSELSD
Pa ---LENASREREAGGRSISGPLDSLGFLQLMQELAEVANEGNL----------QALEIKSLRNMLLFQYLVNALEGDFLPRNRVYEETM----NWTHLRESLLNMLLSARR------INYKGLMKDCLSIMCGLFDVSAGISE
Nt -------LDRDDEINKGSSGSIDSVGFHSLIQELGKIVNQKS-----------KVLETEILRKMLLLHYLVSVLEADFVPRNYAFKEKM----RWTLLRDSLLNMLLGSRK------IIYKGLVKDCLLVICDVFSDLSDQTES
Gs ---VDKATAVSEAGDKQSGGSLDHSSFQALIEELAEILSETKF----------EAASMEPLRNMLIFQYLVVVFEDDFLPRNATL--------NWSMQRESLISLLLGSRK------INYKSLMKYFMAILCQLSQLQEAKDDYE
Zm ---NGAVSSEIGGNVRSSDHLFDPSRFSLLIEAIVEPTNGTD-----------DNNGIKAIENMMLFQYLVNTLEADFVPRQIATKESL----DWVIFRESLLSMLLGSRK------LVFKSFVKNCMCILLNQYYHCHKDLKD
Dc ---SGNSHVEASGICRDSKSLLDSASFSILVEEIAIG----------------RHSAIKTVQNMLLFQYLVRALEADFLPRHTLYKETL----NWVFLRESVLSLLVASRK------MNYKSLVHNCMSVVSTRFIHEANI
Xs ----DAVCSERQAARANSIGPPDFSGFTLLIQELGGVVNGTNF----------QAVDTEYLRNMLLLQYLVNVLEGDFLP------QTS----NWNLLRESLLNILLGSRR------INYKVLMKDCLSILSGLCQARAGINNDHG
Rg ---------------------------------------------------------------MVLLQYLISVLEGDFVPRNSVYKENM----DWNLLRVSLLNMILGSRR------ISYKDLIKGCLSIICESELSKHPVL
Cc ---------------------TVQWDWDLLAYHVENKDPHRKHSLTK-------------NDRLLFLQYVVQTLEDDFQLKHDM-------------PQQTIAKTVLSCDDK----FSNVRCVIKWLIDTISMSHLLAE
Ss --------------------TTIPWDWELLTSVMAAEQSRGRRLCC-------------EVVRMLLQYVVQTMEDDFRVKLSFRD-----------LHHSIAKATLSCDVR----FTQVRDVINWLFAAIVKDSEDTQ
Dr --------------------STVPWDWELIKSVMKDQDDTWRLRT---------------EVQVMLLQYVLQGLEDDFHLKLSTQC-----------LQHSLAKKVLSCGPET---FGQVRDVISWTMTAAKESVKHSK
Xl --------------------DAVTWYWDLLSNVVENKEQ--------------------PVLCLFLQYVVQTLEDDFQLCLQRRC-----------LHTCLSKVMLSWDK----NVSHIRDVIKWLIKTINDSQIGLE
Gg --------------------TTVGWDWMLLKYTMENEEKL--------------------PGRLLFLQYVVQTLEDDFQQNLRLHL-----------LQKSITKKVLSCDT----CFNNVKEVVEWLVAVVTGIRFFQL
Oa --------------------NSVQWNWEMLRSVMEAQAPP--------------------PRRLYFLRYVVQTLQDDFQNALRRHWQQ---------LQGSIVSSVLSCDK----QPNNIRDVISWLVREVTGMGTDQD
Md --------------------ASVEWDWKLLTYVMEKQEEEM-------------------PGRIFFLRYVVQTLEDDFQQVLRHQHH----------LQQSIACTMLSCDK----QPQNIRDIIQWLVGAVTKGGFQKE
La --------------------KTVEWDWKLLSYIMEEEGQNL-------------------PGRVLFLRYVVQTLEDDFQQMLRRQRQH---------LQQSIASTVLSCDK----QPHNVRDVIKWLVKVVTEDELTEP
Mm --------------------KTVEWDWKLLTYVMEEEGQTL-------------------PGRVLFLRYVVQTLEDDFQQILRRQRQH---------LQQSIANTVLSCDK----QPHNVRDVIKWLVKAVTENELTPP
SIMC1------------------KTVEWDWKLLTYVMEEEGQTL-------------------PGRVLFLRYVVQTLEDDFQQTLRRQRQH---------LQQSIANMVLSCDK----QPHNVRDVIKWLVKAVTEDGLTQP
-----------------------------HHHHHHHH---------------------------HHHHHHHHHHHHHHHHHHHHHHHHH-------------HH--HHHHHH---------HHHHHHHHHHHHHHHHHHHH----
SsB-------------------PQCRKGVWPLLETSFRFCMGSKP-CHSLPSP----ASPELLRFHGDLQAFMLELFKCDMHSANKGEA--------GGLRSSVLYRVFWSLWERSTLASRAVQQIAQLLVQASLWER
DrB-------------------PECKNGLWSLLETSLQVCLASEPTCHSLPSP----PSTEMFRLHCNLQEFFLRLFQLELHAASS-------GRSPGSSRASVLYSTFWSTWEKSTLNSRAVRQLSRLLIKTLIWAY
FpB-------------------PVCKKGTWSFIEMLVRSCVYYKTICHAVLDS---GREDEQRIVHKSLLKFFFDLVKAEVEFLTESLVEGTHSQHQPVMPQTVLLKTFWLGSETSVLFTKHINILAGWVVFSHRESN
GgB-------------------PICKKGTWNLIEMLVRSCLHCKNICHAVPVS---GKADEQRTVHKTLLKFFFDLVKAEVESLKKSLVEGTNSQHQQVMPQTVLLKTFWLGSETSVLFTKHINILVDWVIHSYRELE
OaB-------------------PVCMEGTWSLIEALVRSCLFNEIFCHSVPET-----VMEAKVLHKILLKFFFNLVKGEVQALSRKLAEQTDSWSLKVKPQSILLETFWMGNETSVLLTKPISLLWDWVIHSHKENY
DnB-------------------PACMKGAWSLIEVLIRSCLFNENFCHQISEN------IGSKMLHLTLLKFFFNLVESEVQHLSQKLCDWSDYQSLKITRKAVLLEIFWSGSETSSLLTKPVNMLLEWTIYSHKEKY
MdB-------------------PNCMKGIWSFVEVLIRSCLFSESFCHPIPEG-----IVETKMLSVALLTFFLKLVKREVQHLNHRLYEGADSQSLKVKEQTFLLKTFWSGNETSVFFTKPINMLLDWTIYSHKEKF
LaB-------------------PTCMKGAWSLVEVLIRSCLFNENFCHQISEN-----VIGSKVLHLTLLKFFFNLVESEVRHLSQKLYDWSDSQSLKITGKAVLLELFWSGSETSALLTKPVNMLLEWTIYSHKEKY
MmB-------------------PTCMKGAWDFTEVLIRSCLFNEDFCHQISEN------ISTKVVHLTLLKFFFNLLEGEVRHLSQKLCDWSDSQSLKVTEKAILHEIFWSGSETSGLLTKPVNMLLEWTIYSHKEKC
SLF1 496--------------PTCMKGAWSLVEVLIRSCLFNESFCHQISEN------IGSKVLHLTLLKFFFNLIESEVQHLSQKLYDWSDSQNLKITGKAMLLEIFWSGSETSGLLTKPVNMLLEWTIYSHKEKF
```

doc. Jan Paleček
(unpublished)

Zjištění konzervovaných domén proteinu

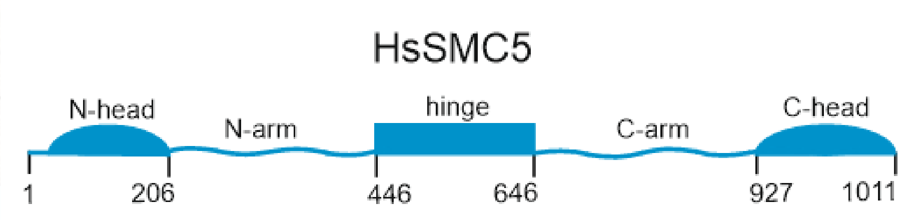
- NCBI Conserved domain search https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
- UniProt https://www.uniprot.org/

- Protein může být popsaný a charakterizovaný v publikacích
  - Příbuzné organismy - porovnání

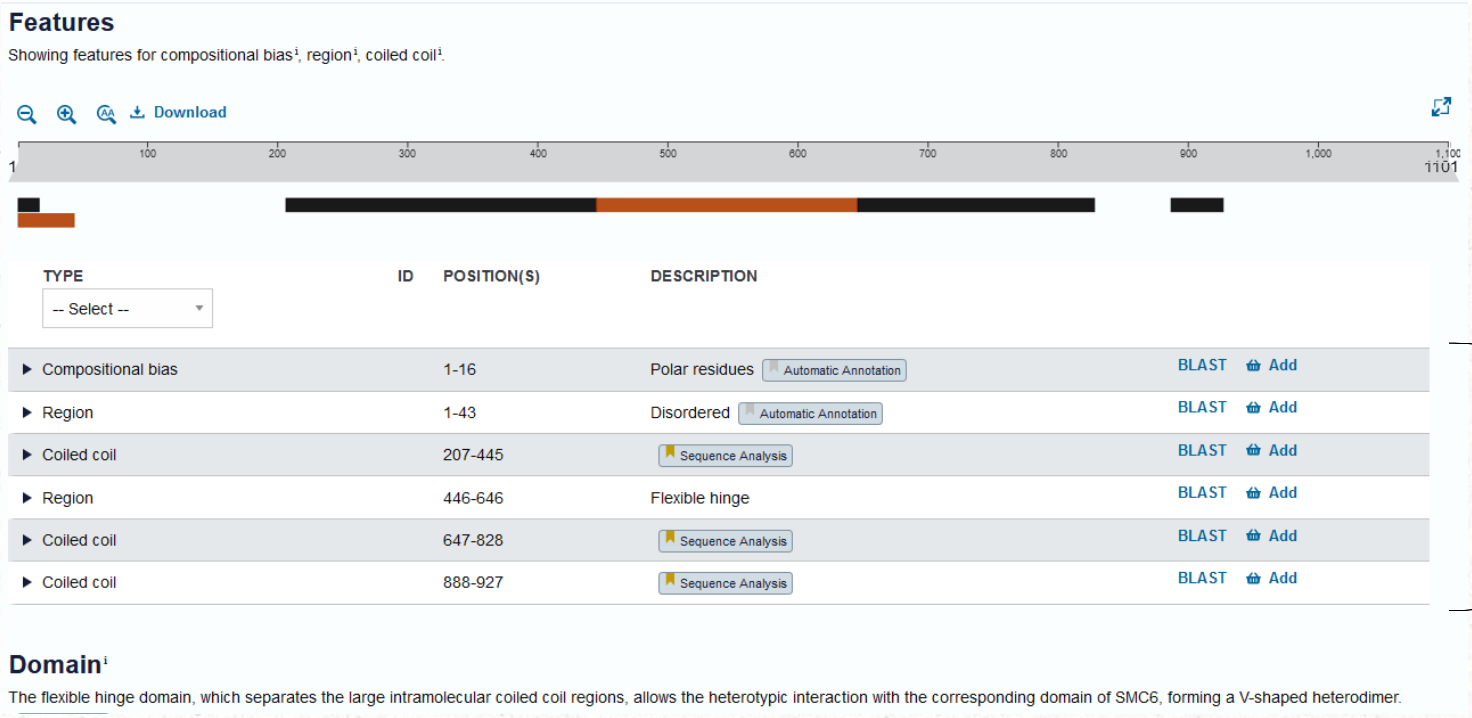
- Interakční motivy – kratší než domény (jednotky aminokyselin)
  - Online tools/programy na jejich vyhledávání

Zjištění konzervovaných domén proteinu

UniProt



Family & Domains



Features

Showing features for compositional bias[1], region[1], coiled coil[1].

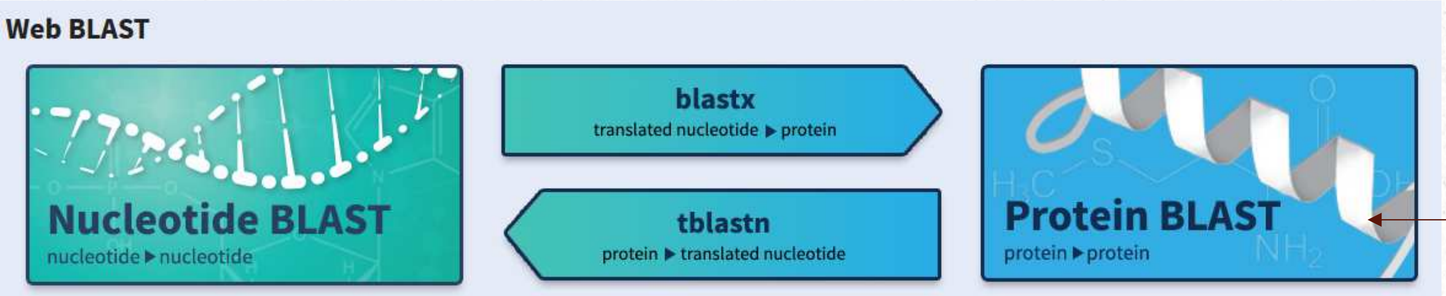
TYPE	ID	POSITION(S)	DESCRIPTION		
-- Select --					
Compositional bias		1-16	Polar residues Automatic Annotation	BLAST	Add
Region		1-43	Disordered Automatic Annotation	BLAST	Add
Coiled coil		207-445	Sequence Analysis	BLAST	Add
Region		446-646	Flexible hinge	BLAST	Add
Coiled coil		647-828	Sequence Analysis	BLAST	Add
Coiled coil		888-927	Sequence Analysis	BLAST	Add

Domain[1]

The flexible hinge domain, which separates the large intramolecular coiled coil regions, allows the heterotypic interaction with the corresponding domain of SMC6, forming a V-shaped heterodimer.

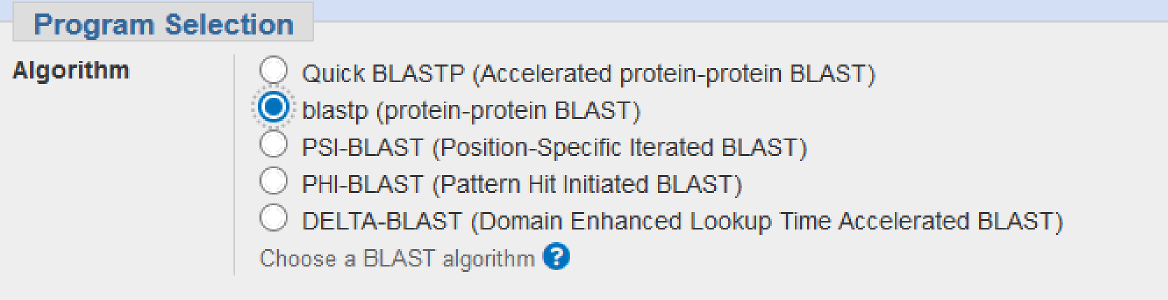
Vymezení strukturních/
funkčních oblastí proteinu

Zjištění konzervovaných domén proteinu



Web BLAST

Zadat sekvenci proteinu

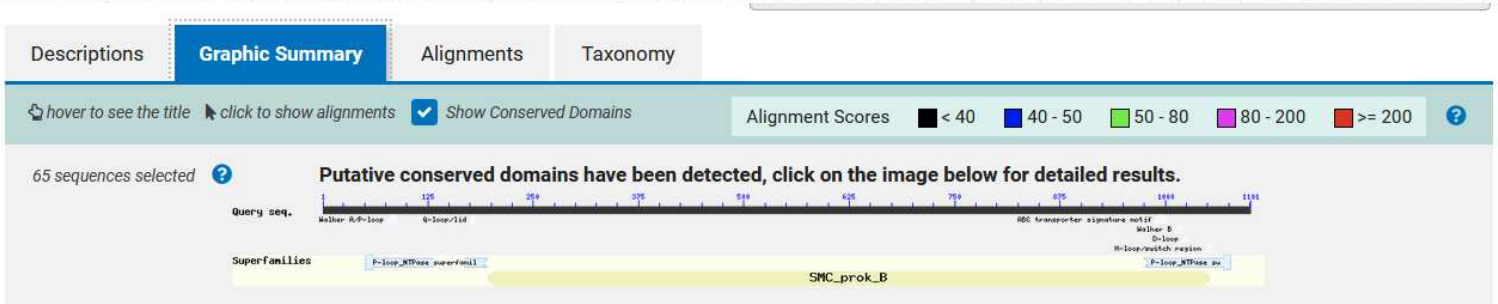


Program Selection

Algorithm

- ○ Quick BLASTP (Accelerated protein-protein BLAST)
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)
- ○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm



☑ Show Conserved Domains

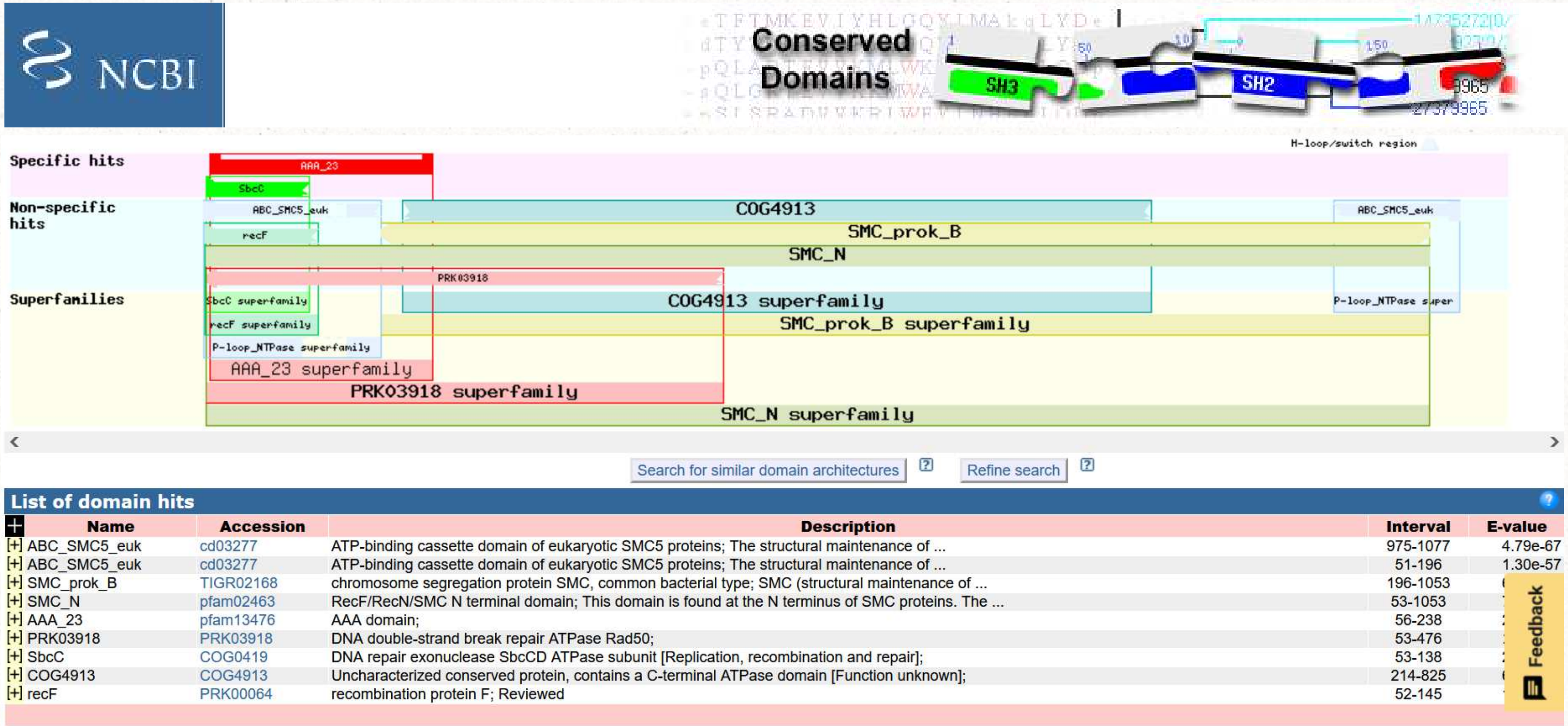
Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200

65 sequences selected

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

Propojení na NCBI Conserved domains

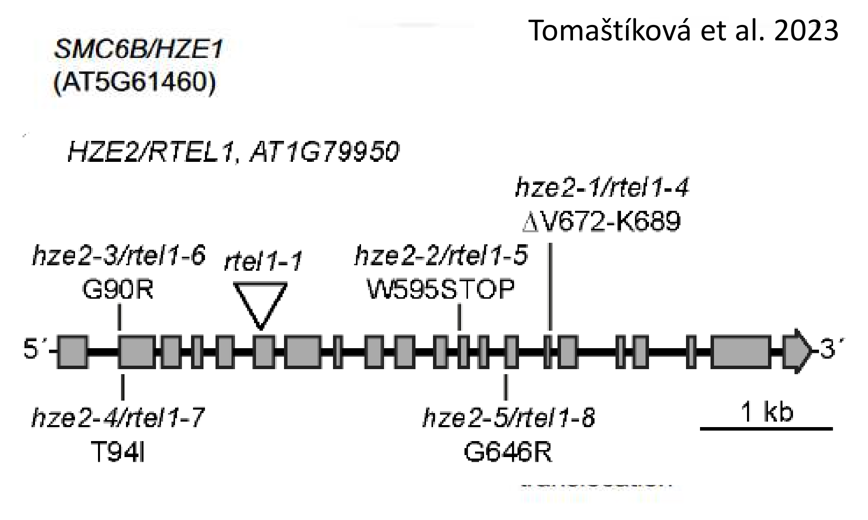
Zjištění konzervovaných domén proteinu



Name	Accession	Description	Interval	E-value
[+] ABC_SMC5_euk	cd03277	ATP-binding cassette domain of eukaryotic SMC5 proteins; The structural maintenance of ...	975-1077	4.79e-67
[+] ABC_SMC5_euk	cd03277	ATP-binding cassette domain of eukaryotic SMC5 proteins; The structural maintenance of ...	51-196	1.30e-57
[+] SMC_prok_B	TIGR02168	chromosome segregation protein SMC, common bacterial type; SMC (structural maintenance of ...	196-1053	[illegible]
[+] SMC_N	pfam02463	RecF/RecN/SMC N terminal domain; This domain is found at the N terminus of SMC proteins. The ...	53-1053	[illegible]
[+] AAA_23	pfam13476	AAA domain;	56-238	[illegible]
[+] PRK03918	PRK03918	DNA double-strand break repair ATPase Rad50;	53-476	[illegible]
[+] SbcC	COG0419	DNA repair exonuclease SbcCD ATPase subunit [Replication, recombination and repair];	53-138	[illegible]
[+] COG4913	COG4913	Uncharacterized conserved protein, contains a C-terminal ATPase domain [Function unknown];	214-825	[illegible]
[+] recF	PRK00064	recombination protein F; Reviewed	52-145	[illegible]

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Schémata genů/proteinů



Tomaštíková et al. 2023

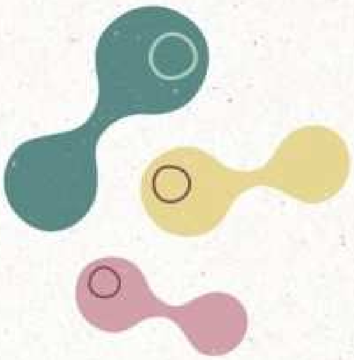
- Powerpoint
- BioRender
- CorelDRAW a jiné…

► Schémata genů
  - Zaznačení exonů, intronů, UTR sekvencí
  - Místa mutací – původní nukleotid/pozice/ změněn na jaký nukleotid (A325G)

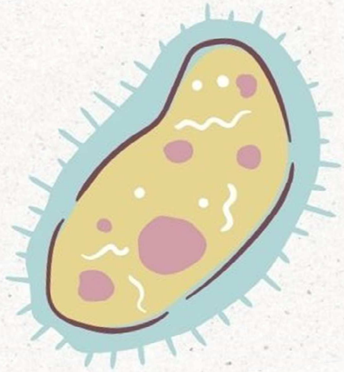
► Schémata proteinů
  - Domény
  - Jaké mají mutace efekt na proteinové úrovni?

Při publikacích pozor, ať obrázky neztrácí kvalitu při zoomu.

Přednáška: Analýza proteinu

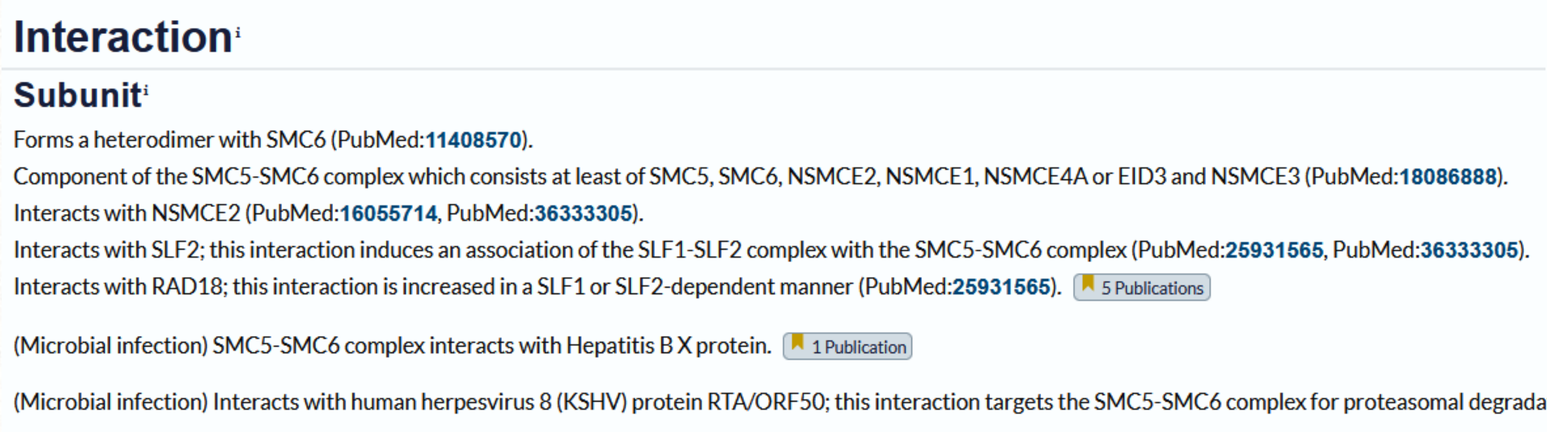


- Popis a funkce ☑
- Konzervovanost (Alignment)
- Evoluce (fylogenetický strom)
- Zjištění přítomnosti domén daného proteinu ☑
- Zjištění interakčních partnerů



Databáze protein-proteinových interakcí

Informace o interakcích i v Uniprot + odkazy na databáze



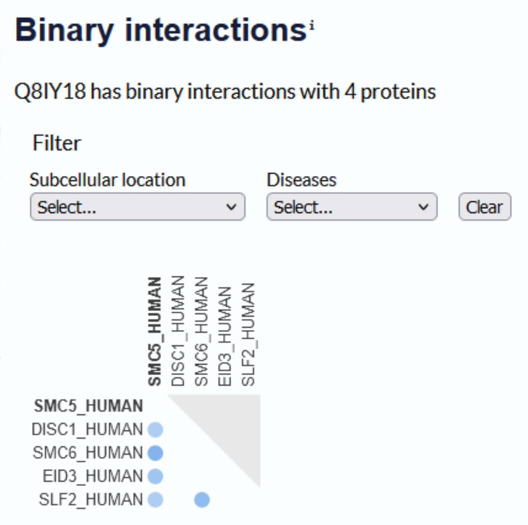
Interaction

Subunit

Forms a heterodimer with SMC6 (PubMed:**11408570**).
Component of the SMC5-SMC6 complex which consists at least of SMC5, SMC6, NSMCE2, NSMCE1, NSMCE4A or EID3 and NSMCE3 (PubMed:**18086888**).
Interacts with NSMCE2 (PubMed:**16055714**, PubMed:**36333305**).
Interacts with SLF2; this interaction induces an association of the SLF1-SLF2 complex with the SMC5-SMC6 complex (PubMed:**25931565**, PubMed:**36333305**).
Interacts with RAD18; this interaction is increased in a SLF1 or SLF2-dependent manner (PubMed:**25931565**). 5 Publications

(Microbial infection) SMC5-SMC6 complex interacts with Hepatitis B X protein. 1 Publication

(Microbial infection) Interacts with human herpesvirus 8 (KSHV) protein RTA/ORF50; this interaction targets the SMC5-SMC6 complex for proteasomal degradation. 1 Publication



Binary interactions

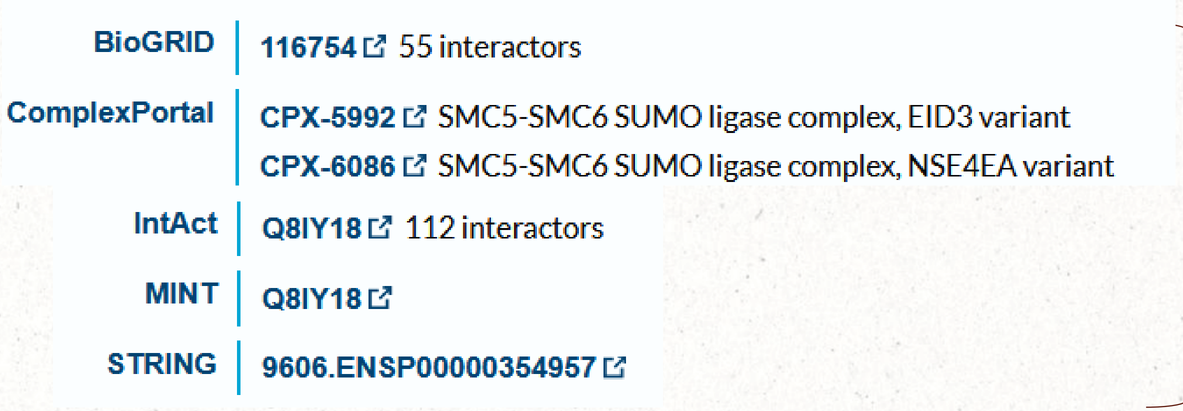
Q8IY18 has binary interactions with 4 proteins

Filter

Subcellular location: Select...

Diseases: Select...

	SMC5_HUMAN	DISC1_HUMAN	SMC6_HUMAN	EID3_HUMAN	SLF2_HUMAN
SMC5_HUMAN					
DISC1_HUMAN	●				
SMC6_HUMAN	●				
EID3_HUMAN	●				
SLF2_HUMAN	●		●		

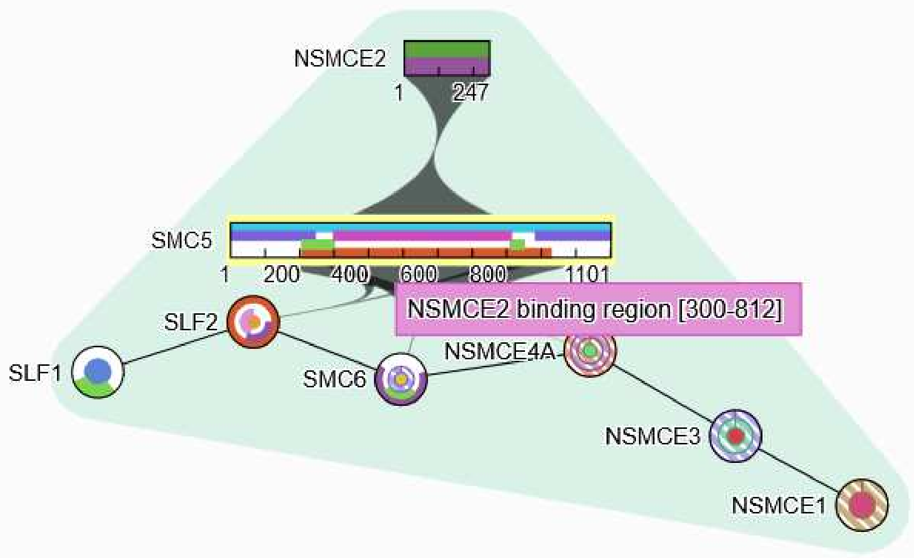


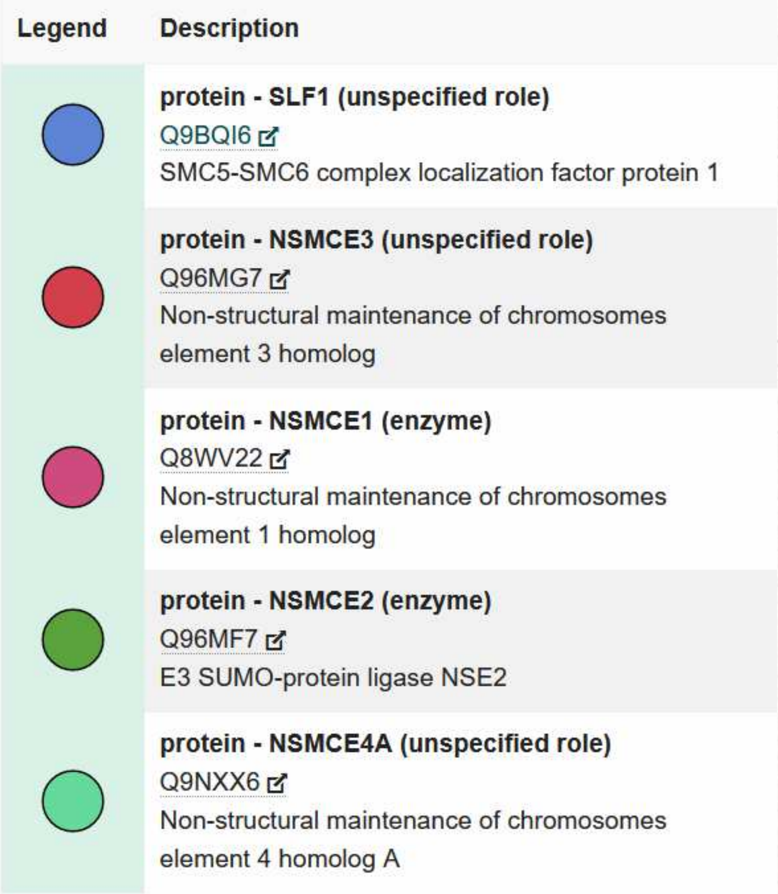
BioGRID	**116754** 55 interactors
ComplexPortal	**CPX-5992** SMC5-SMC6 SUMO ligase complex, EID3 variant **CPX-6086** SMC5-SMC6 SUMO ligase complex, NSE4EA variant
IntAct	**Q8IY18** 112 interactors
MINT	**Q8IY18**
STRING	**9606.ENSP00000354957**

Obsahují podobné informace
Nutné vyfiltrovat převážně fyzické interakce (protein-protein), binární

ComplexPortal

Protein je součástí komplexu



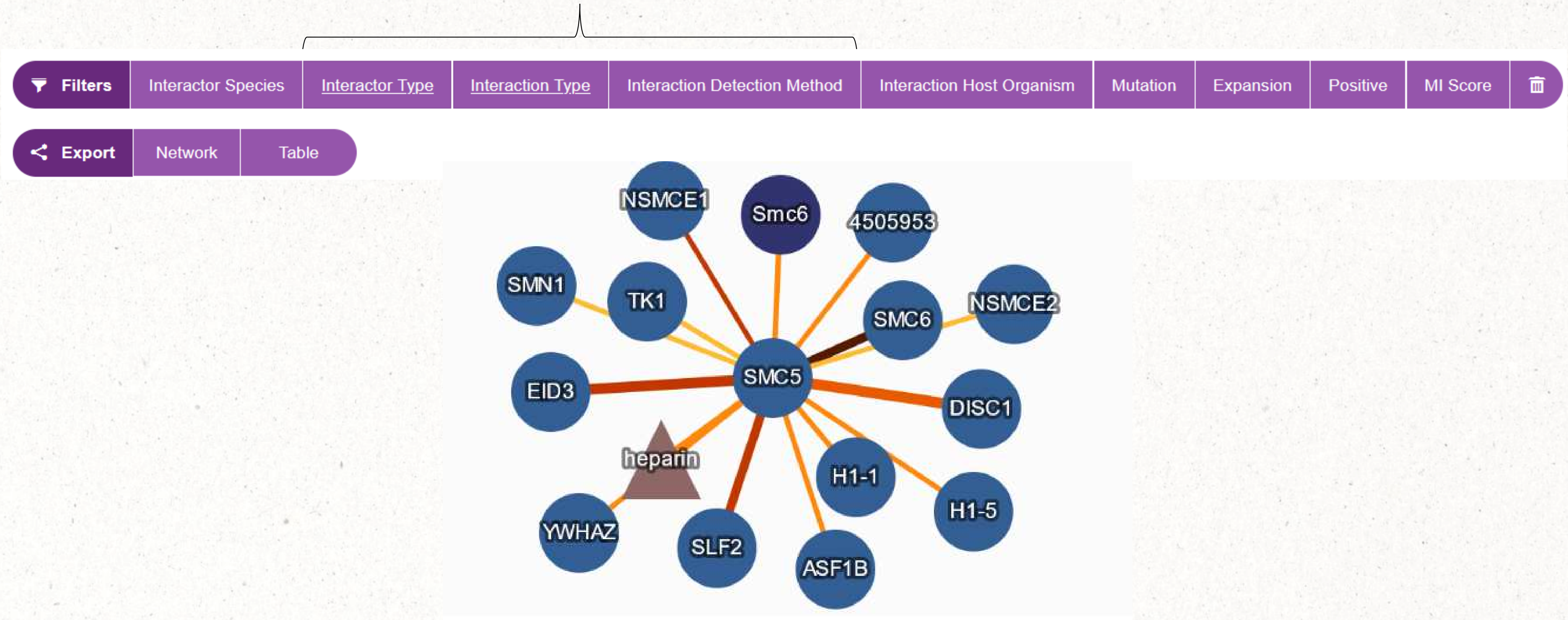


Legend	Description
🔵	**protein - SLF1 (unspecified role)** Q9BQI6 SMC5-SMC6 complex localization factor protein 1
🔴	**protein - NSMCE3 (unspecified role)** Q96MG7 Non-structural maintenance of chromosomes element 3 homolog
●	**protein - NSMCE1 (enzyme)** Q8WV22 Non-structural maintenance of chromosomes element 1 homolog
🟢	**protein - NSMCE2 (enzyme)** Q96MF7 E3 SUMO-protein ligase NSE2
●	**protein - NSMCE4A (unspecified role)** Q9NXX6 Non-structural maintenance of chromosomes element 4 homolog A

https://www.ebi.ac.uk/complexportal/home

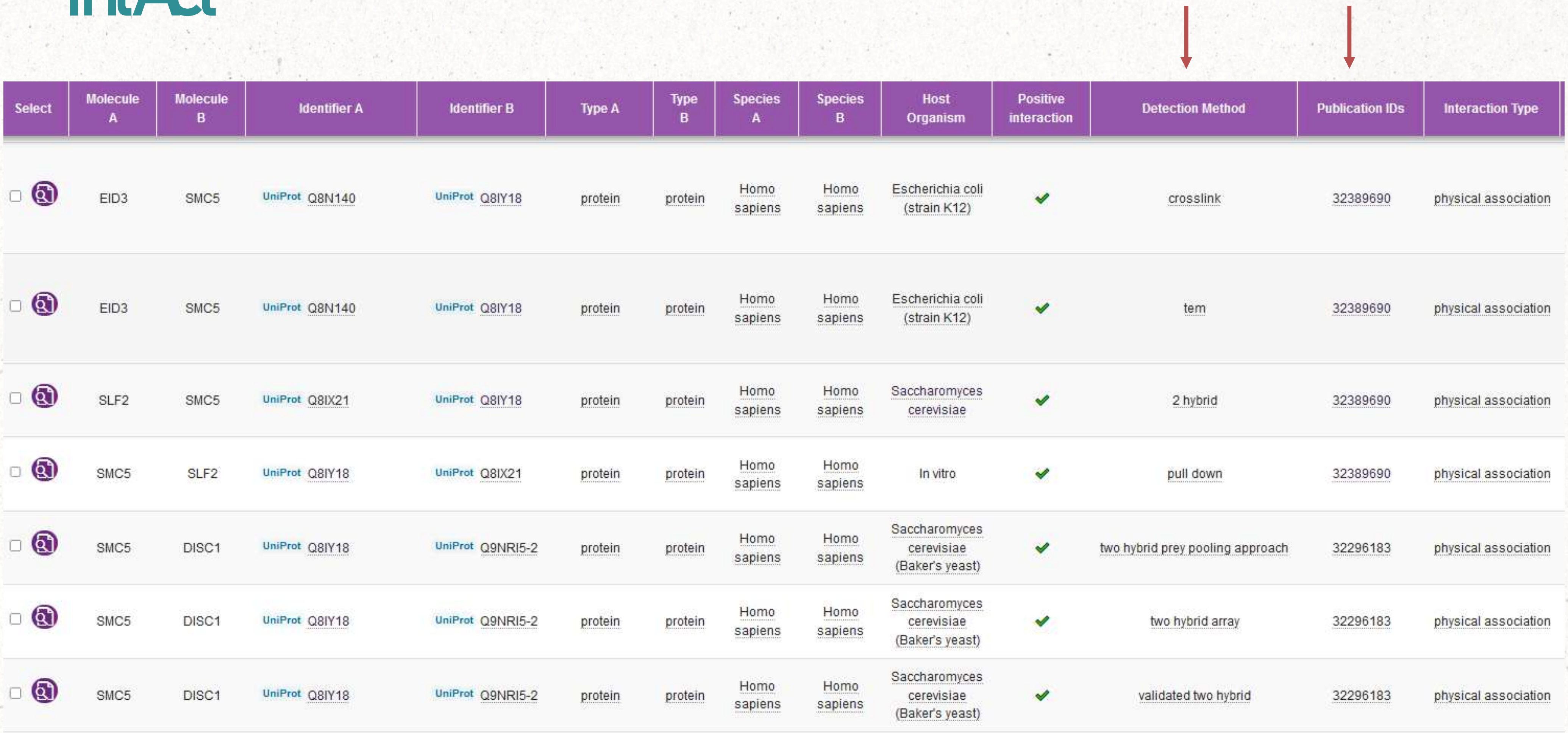
IntAct

Filtry – typ interakce
Detekční metoda interakce – binární, koimunoprecipitace?



https://www.ebi.ac.uk/intact/home

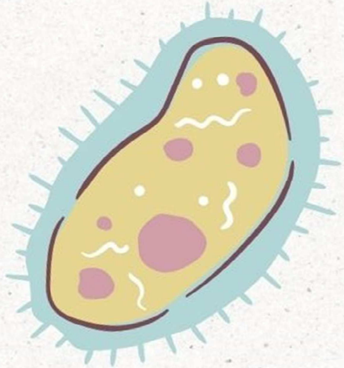
IntAct



Select	Molecule A	Molecule B	Identifier A	Identifier B	Type A	Type B	Species A	Species B	Host Organism	Positive interaction	Detection Method	Publication IDs	Interaction Type
☐	EID3	SMC5	UniProt Q8N140	UniProt Q8IY18	protein	protein	Homo sapiens	Homo sapiens	Escherichia coli (strain K12)	✓	crosslink	32389690	physical association
☐	EID3	SMC5	UniProt Q8N140	UniProt Q8IY18	protein	protein	Homo sapiens	Homo sapiens	Escherichia coli (strain K12)	✓	tem	32389690	physical association
☐	SLF2	SMC5	UniProt Q8IX21	UniProt Q8IY18	protein	protein	Homo sapiens	Homo sapiens	Saccharomyces cerevisiae	✓	2 hybrid	32389690	physical association
☐	SMC5	SLF2	UniProt Q8IY18	UniProt Q8IX21	protein	protein	Homo sapiens	Homo sapiens	In vitro	✓	pull down	32389690	physical association
☐	SMC5	DISC1	UniProt Q8IY18	UniProt Q9NRI5-2	protein	protein	Homo sapiens	Homo sapiens	Saccharomyces cerevisiae (Baker's yeast)	✓	two hybrid prey pooling approach	32296183	physical association
☐	SMC5	DISC1	UniProt Q8IY18	UniProt Q9NRI5-2	protein	protein	Homo sapiens	Homo sapiens	Saccharomyces cerevisiae (Baker's yeast)	✓	two hybrid array	32296183	physical association
☐	SMC5	DISC1	UniProt Q8IY18	UniProt Q9NRI5-2	protein	protein	Homo sapiens	Homo sapiens	Saccharomyces cerevisiae (Baker's yeast)	✓	validated two hybrid	32296183	physical association

Přednáška: Analýza proteinu

- Popis a funkce ☑
- Konzervovanost (Alignment)
- Evoluce (fylogenetický strom)
- Zjištění přítomnosti domén daného proteinu ☑
- Zjištění interakčních partnerů ☑
- Struktura proteinu – Alphafold/Colabfold
- Úprava proteinového modelu v PyMol
  - Zvýraznění interakčních aminokyselin
  - Zvýraznění domén



Predikce sekundární struktury

- PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred

Submission details

Protein Sequence

Protein Sequence

Help...

If you wish to test these services follow this link to retrieve a test fasta sequence.

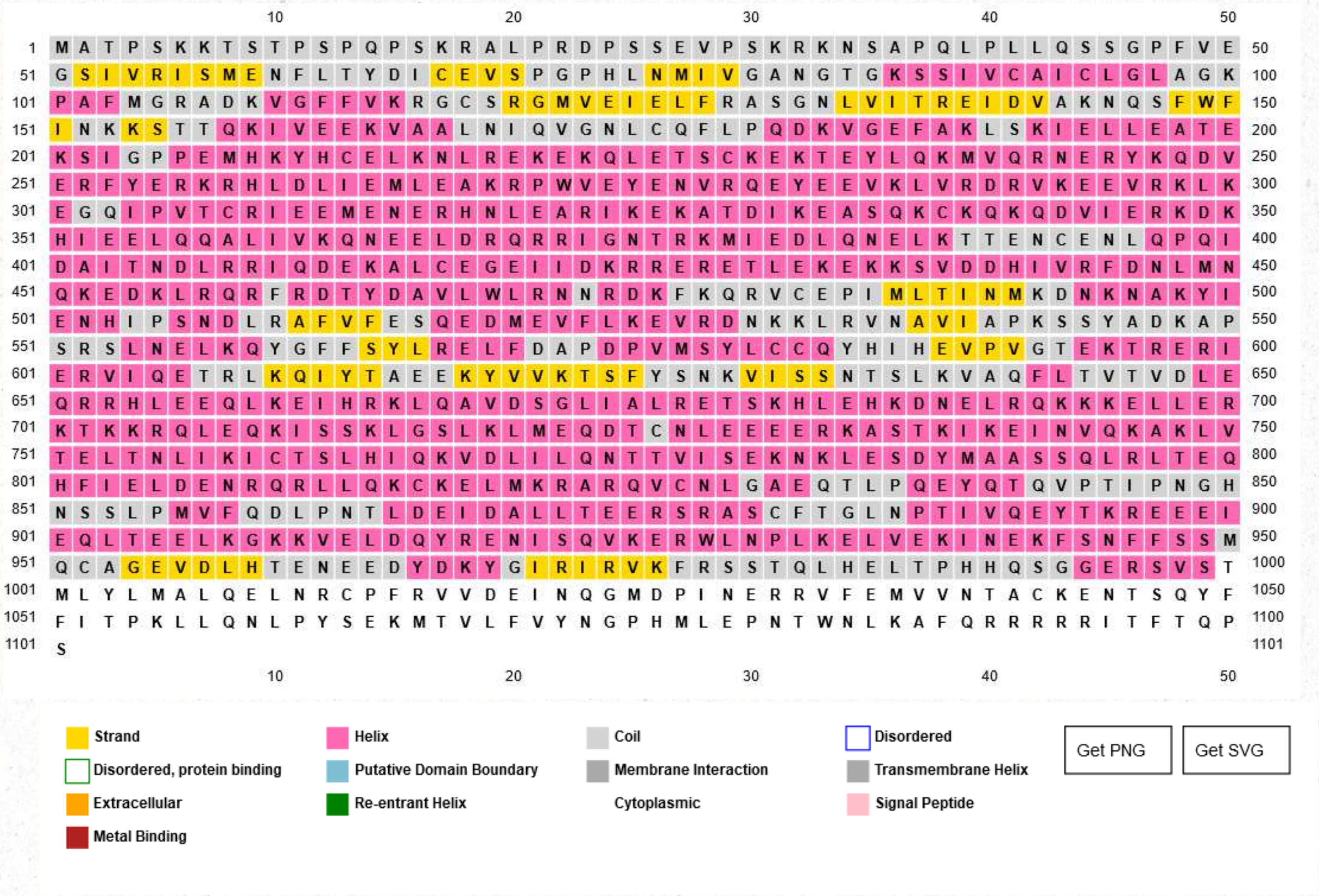
Job name

Job name

Email (optional)

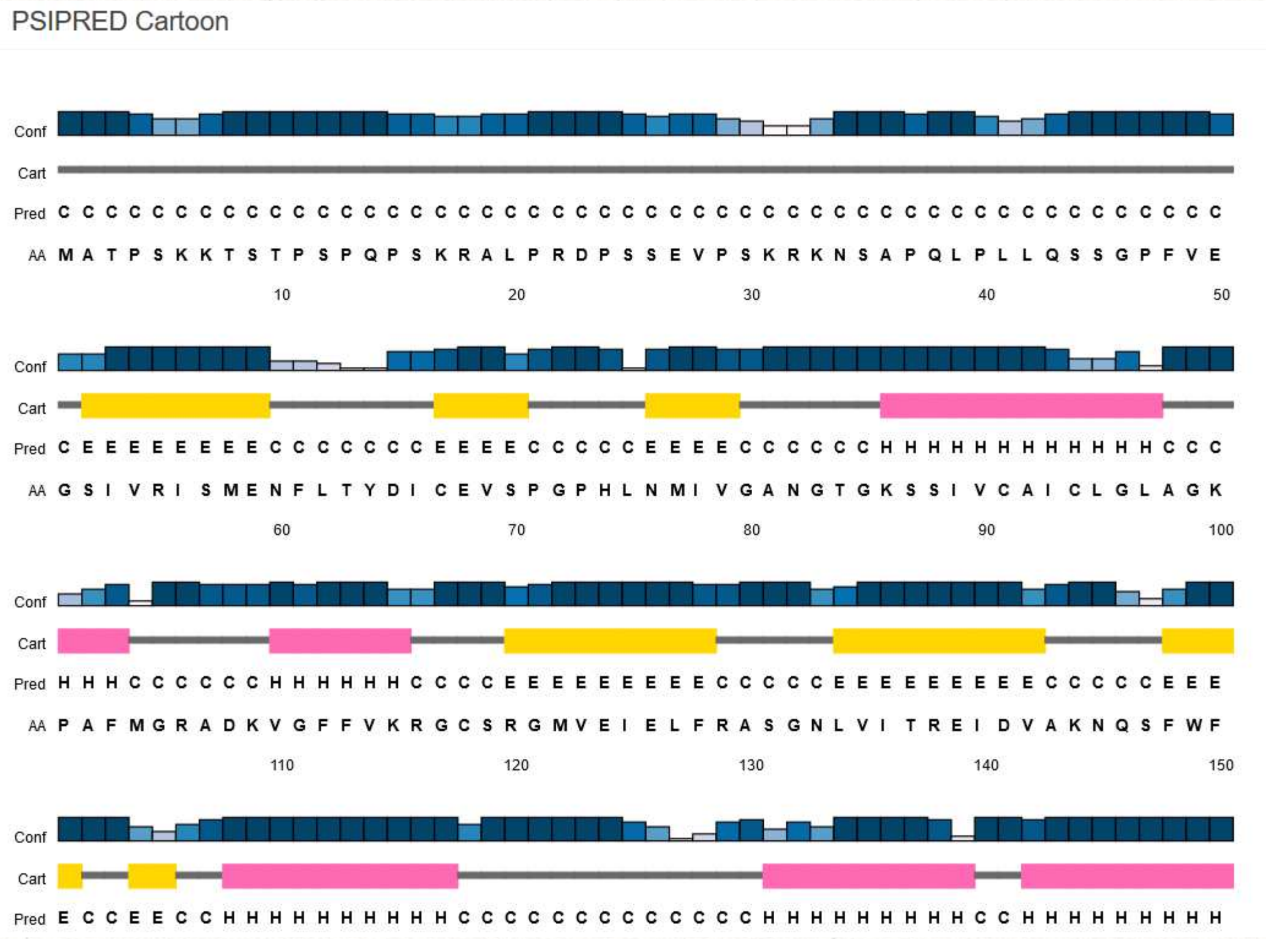
Email (optional)

PSIPRED – predikce sekundární struktury



```
        10        20        30        40        50
1    MATPSKKTSTPSPQPSKRALPRDPSSEVPSKRKNSAPQLPLLQSSGPFVE 50
51   GSIVRISMENFLTYDICEVSPGPHLNMIVGANGTGKSSIVCAICLGLAGK 100
101  PAFMGRADKVGFFVKRGCSRGMVEIELFRASGNLVITREIDVAKNQSFWF 150
151  INKKSTTQKIVEEKVAALNIQVGNLCQFLPQDKVGEFAKLSKIELLEATE 200
201  KSIGPPEMHKYHCELKNLREKEKQLETSCKEKTEYLQKMVQRNERYKQDV 250
251  ERFYERKRHLDLIEMLEAKRPWVEYENVRQEYEEVKLVRDRVKEEVRKLK 300
301  EGQIPVTCRIEEMENERHNLEARIKEKATDIKEASQKCKQKQDVIERKDK 350
351  HIEELQQALIVKQNEELDRQRRIGNTRKMIEDLQNELKTTENCENLQPQI 400
401  DAITNDLRRIQDEKALCEGEIIDKRRERETLEKEKKSVDDHIVRFDNLMN 450
451  QKEDKLRQRFRDTYDAVLWLRNNRDKFKQRVCEPIMLTINMKDNKNAKYI 500
501  ENHIPSNDLRAFVFESQEDMEVFLKEVRDNKKLRVNAVIAPKSSYADKAP 550
551  SRSLNELKQYGFFSYLRELFDAPDPVMSYLCCQYHIHEVPVGTEKTRERI 600
601  ERVIQETRLKQIYTAEEKYVVKTSFYSNKVISSNTSLKVAQFLTVTVDLE 650
651  QRRHLEEQLKEIHRKLQAVDSGLIALRETSKHLEHKDNELRQKKKELLER 700
701  KTKKRQLEQKISSKLGSLKLMEQDTCNLEEEERKASTKIKEINVQKAKLV 750
751  TELTNLIKICTSLHIQKVDLILQNTTVISEKNKLESDYMAASSQLRLTEQ 800
801  HFIELDENRQRLLQKCKELMKRARQVCNLGAEQTLPQEYQTQVPTIPNGH 850
851  NSSLPMVFQDLPNTLDEIDALLTEERSRASCFTGLNPTIVQEYTKREEEI 900
901  EQLTEELKGKKVELDQYRENISQVKERWLNPLKELVEKINEKFSNFFSSM 950
951  QCAGEVDLHTENEEDYDKYGIRIRVKFRSSTQLHELTPHHQSGGERSVST 1000
1001 MLYLMALQELNRCPFRVVDEINQGMDPINERRVFEMVVNTACKENTSQYF 1050
1051 FITPKLLQNLPYSEKMTVLFVYNGPHMLEPNTWNLKAFQRRRRRITFTQP 1100
1101 S                                                  1101
        10        20        30        40        50
```

PSIPRED – predikce sekundární struktury



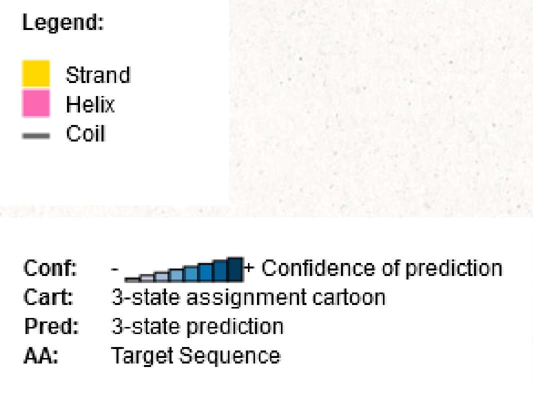
PSIPRED Cartoon

Conf
Cart
Pred C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C C
AA M A T P S K K T S T P S P Q P S K R A L P R D P S S E V P S K R K N S A P Q L P L L Q S S G P F V E
10 20 30 40 50

Conf
Cart
Pred C E E E E E E E E C C C C C C C E E E E C C C C C E E E E C C C C C C H H H H H H H H H H H H C C C
AA G S I V R I S M E N F L T Y D I C E V S P G P H L N M I V G A N G T G K S S I V C A I C L G L A G K
60 70 80 90 100

Conf
Cart
Pred H H H C C C C C C H H H H H H C C C C E E E E E E E E E C C C C C E E E E E E E E E C C C C C E E E
AA P A F M G R A D K V G F F V K R G C S R G M V E I E L F R A S G N L V I T R E I D V A K N Q S F W F
110 120 130 140 150

Conf
Cart
Pred E C C E E C C H H H H H H H H H H C C C C C C C C C C C C C H H H H H H H H H C C H H H H H H H H H



Legend:

🟨 Strand
🟪 Helix
— Coil

Conf: - + Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence

AlphaFold – predikce 3D struktury proteinu

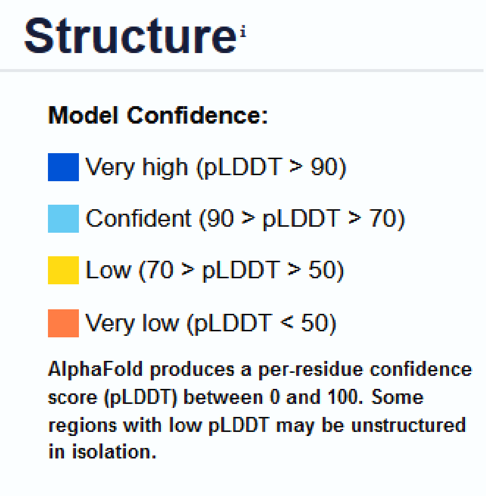
1) Využití databáze PDB (stažení PDB modelu) https://www.rcsb.org/


RCSB PDB PROTEIN DATA BANK

203,607 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

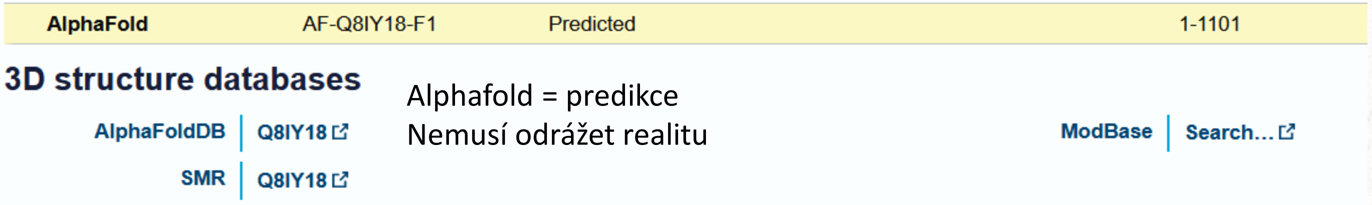
2) 3D predikce proteinu na UniProt – odkaz na Alphafold databázi


Structure

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

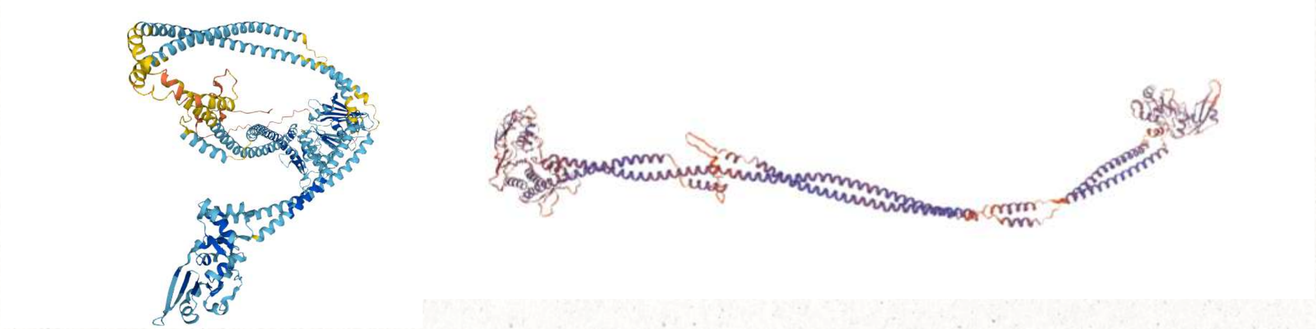
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.


AlphaFold	AF-Q8IY18-F1	Predicted	1-1101

3D structure databases

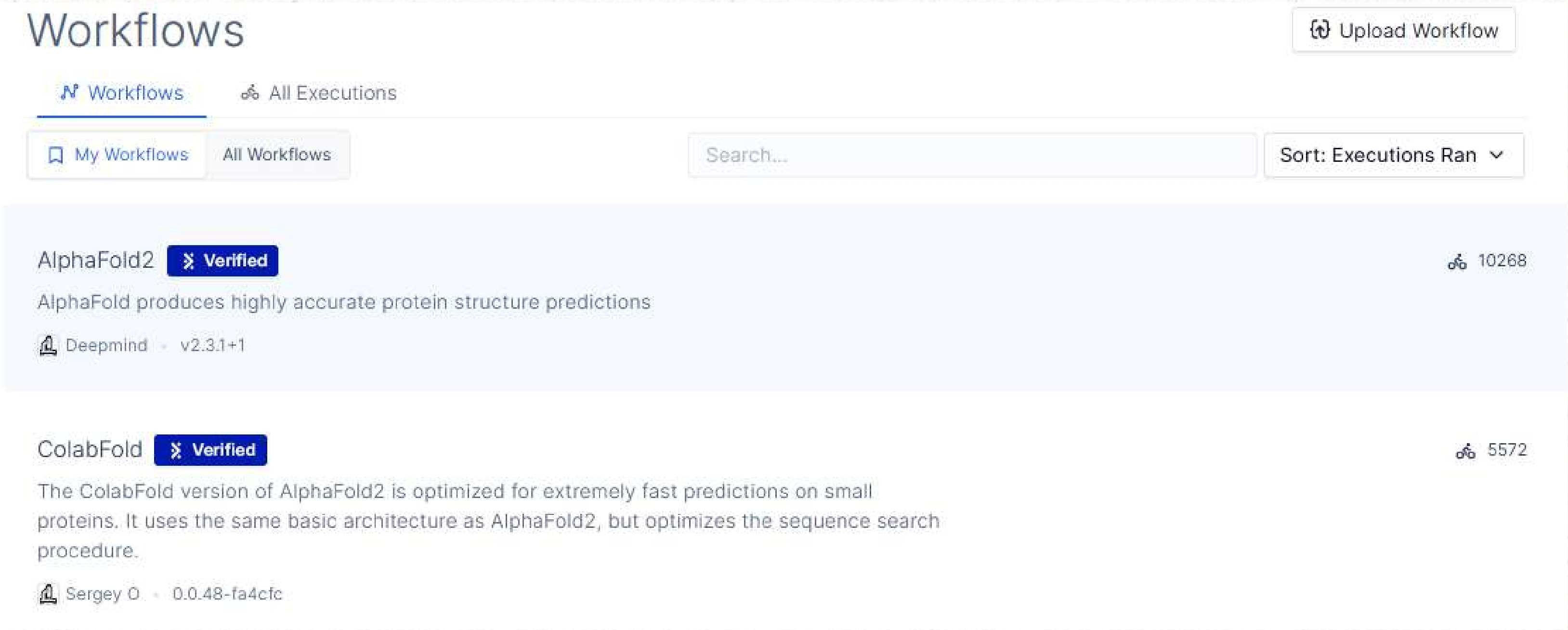
AlphaFoldDB	Q8IY18
SMR	Q8IY18
ModBase	Search...

Alphafold = predikce
Nemusí odrážet realitu



AlphaFold – predikce 3D struktury proteinu

- Využití online predikce Alphafold
  - LatchBio: https://console.latch.bio/workflows



AlphaFold2 Verified 10268

AlphaFold produces highly accurate protein structure predictions

Deepmind · v2.3.1+1

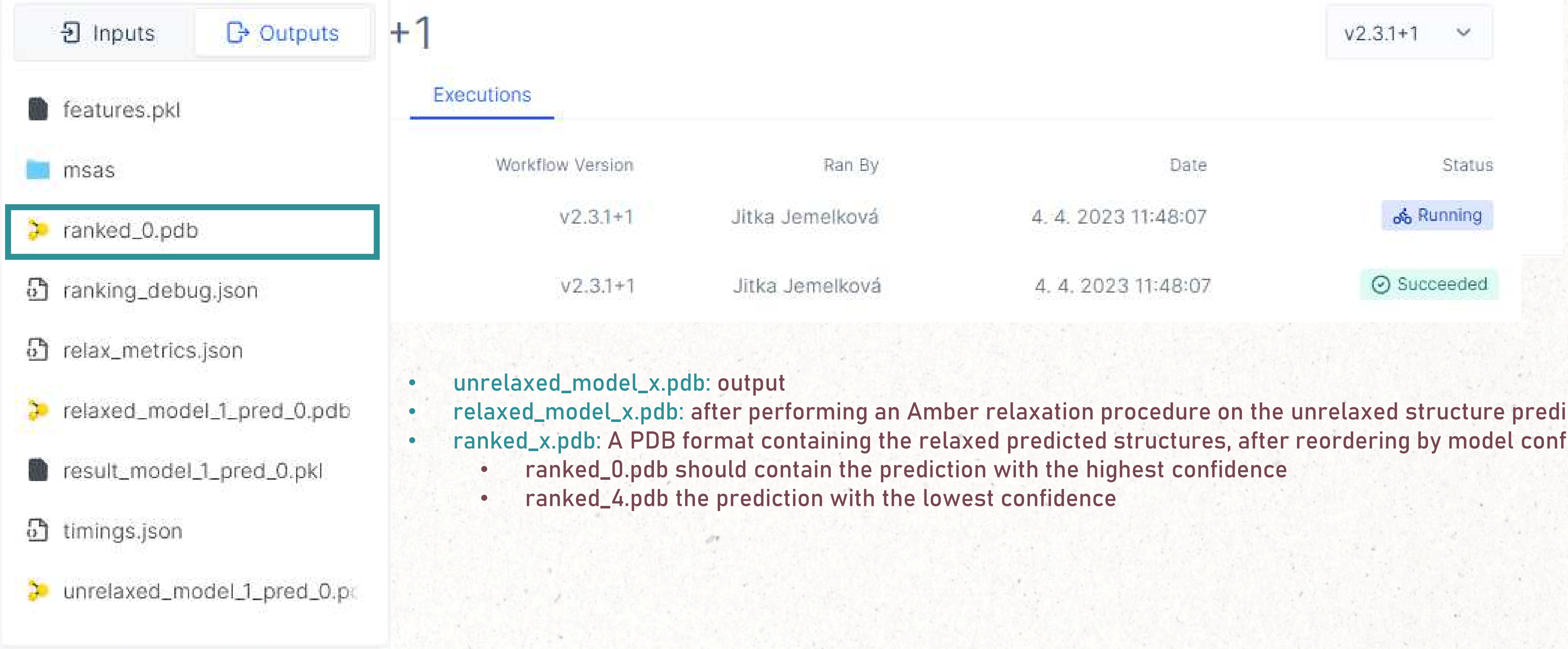
ColabFold Verified 5572

The ColabFold version of AlphaFold2 is optimized for extremely fast predictions on small proteins. It uses the same basic architecture as AlphaFold2, but optimizes the sequence search procedure.

Sergey O · 0.0.48-fa4cfc

- Umožňuje i predikci 3D struktury proteinového komplexu

LatchBio AlphaFold



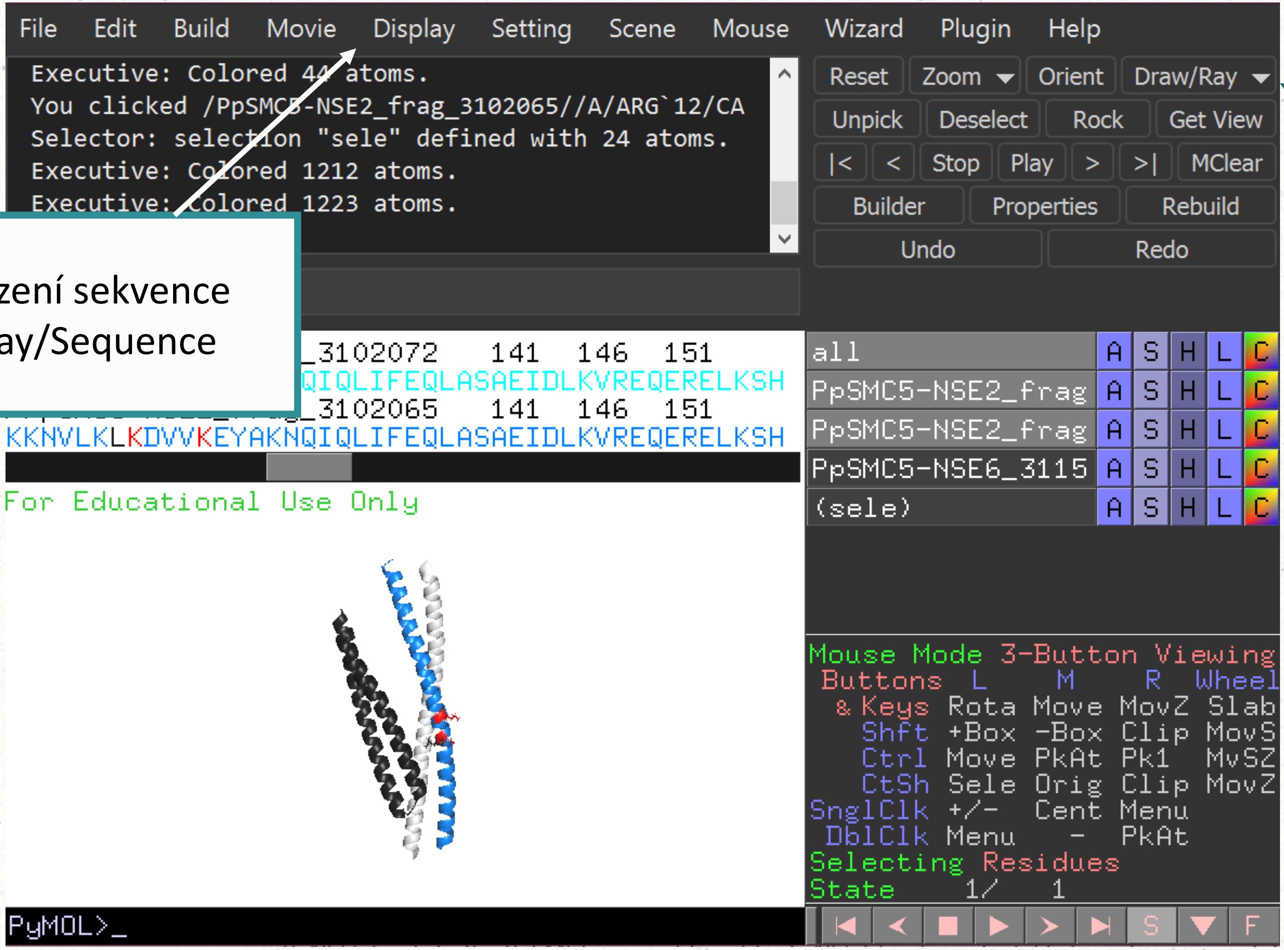
features.pkl
msas
ranked_0.pdb
ranking_debug.json
relax_metrics.json
relaxed_model_1_pred_0.pdb
result_model_1_pred_0.pkl
timings.json
unrelaxed_model_1_pred_0.p

Workflow Version	Ran By	Date	Status
v2.3.1+1	Jitka Jemelková	4. 4. 2023 11:48:07	Running
v2.3.1+1	Jitka Jemelková	4. 4. 2023 11:48:07	Succeeded

- unrelaxed_model_x.pdb: output
- relaxed_model_x.pdb: after performing an Amber relaxation procedure on the unrelaxed structure prediction
- ranked_x.pdb: A PDB format containing the relaxed predicted structures, after reordering by model confidence.
    - ranked_0.pdb should contain the prediction with the highest confidence
    - ranked_4.pdb the prediction with the lowest confidence

Zpracování 3D predikce (PyMOL)

ranked_0(2)



```
Executive: Colored 44 atoms.
You clicked /PpSMC5-NSE2_frag_3102065//A/ARG`12/CA
Selector: selection "sele" defined with 24 atoms.
Executive: Colored 1212 atoms.
Executive: Colored 1223 atoms.
```

```
_3102072     141   146   151
QIQLIFEQLASAEIDLKVREQERELKSH
_3102065     141   146   151
KKNVLKLKDVVKEYAKNQIQLIFEQLASAEIDLKVREQERELKSH
For Educational Use Only
```

```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Wheel
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl Move PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
 SnglClk +/-  Cent Menu
  DblClk Menu  -   PkAt
Selecting Residues
State     1/    1
PyMOL>_
```

Stahování obrázků

Zobrazení sekvence
Display/Sequence

A: Action
S: Show
H: Hide
L: Label
C: Color

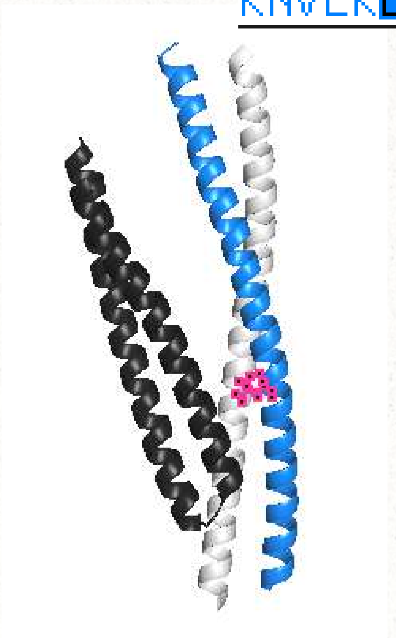
Úkony v PyMol

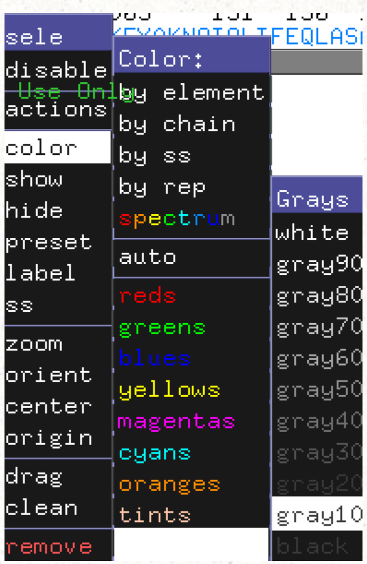
- Změna barvy celého proteinu
- Zvýraznění domén
- Zvýraznění sekundárních struktur
- Zvýraznění aminokyselin
    - Postranní řetězce

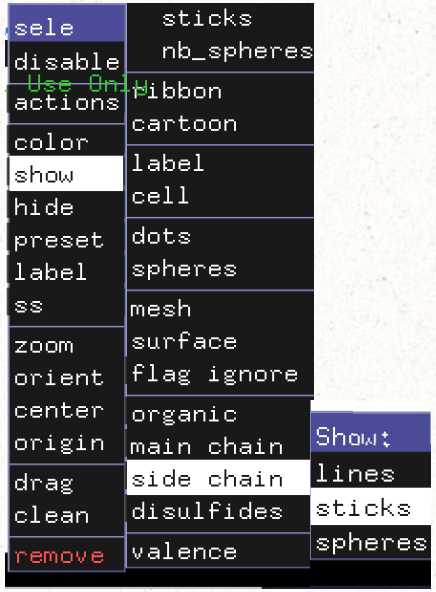
- Stahování obrázků
- Tvorba videa

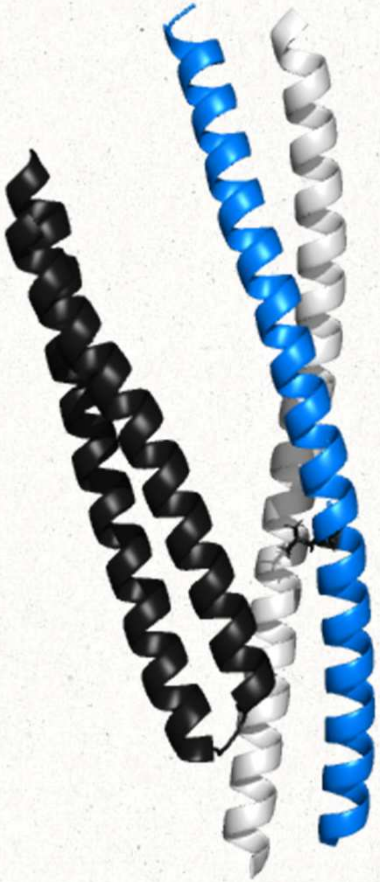
Zvýraznění aminokyseliny

- Označení aminokyseliny (levé tlačítko myši)
- Pravé tlačítko myši – Color
- Pravé tlačítko myši – Show → Side chain → Stick



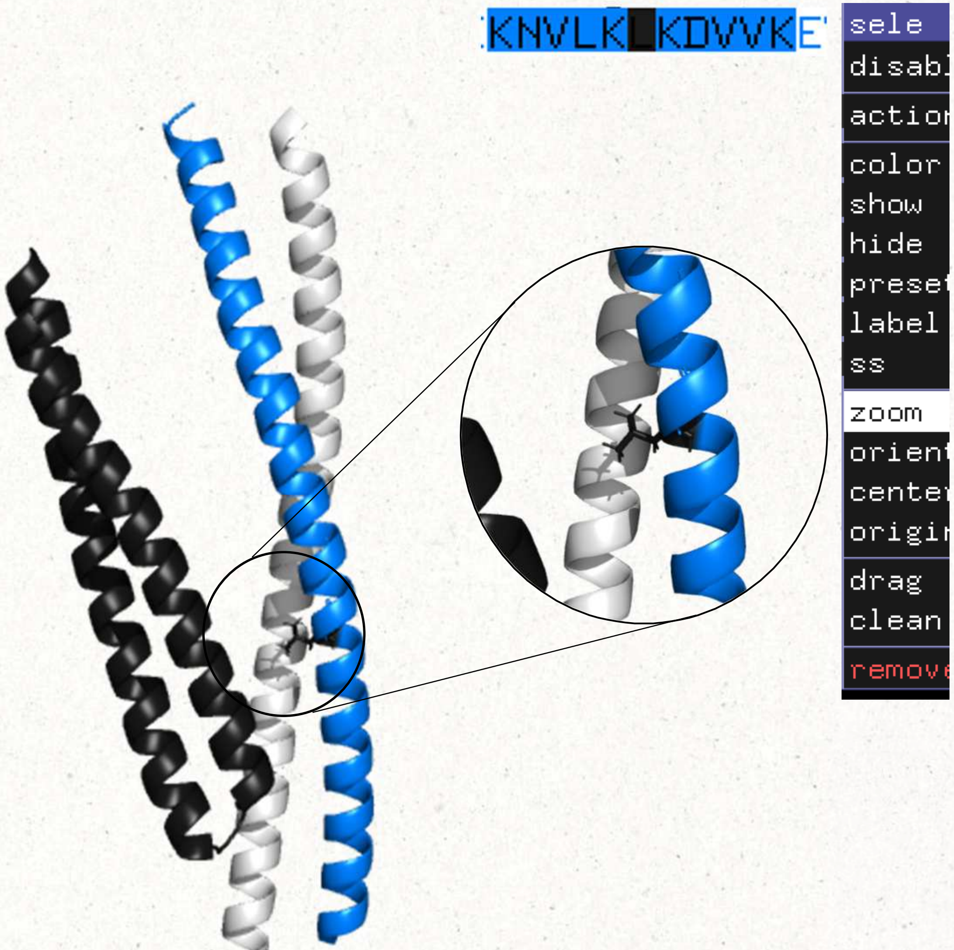






Zvýraznění aminokyseliny

- Označení aminokyseliny (levé tlačítko myši)
- Pravé tlačítko myši – Color
- Pravé tlačítko myši – Show → Side chain → Stick

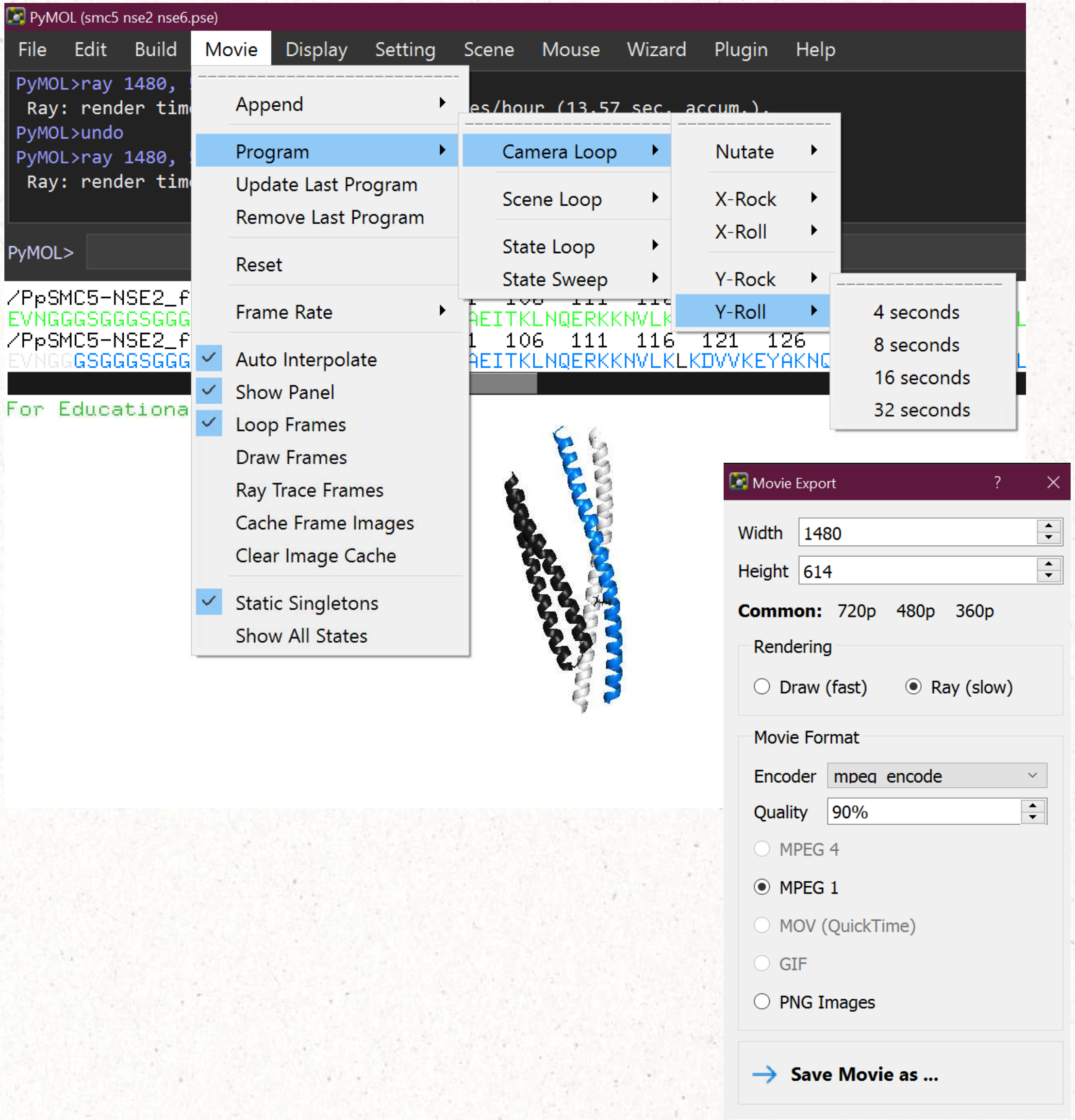


Tvorba videa v PyMol

Pymol:

Movie→ Program → Camera Loop → Y roll → 16 second

- Trial verze PyMol nepodporuje stažení videí, ale studentská verze ano



PyMOL (smc5 nse2 nse6.pse)

```
PyMOL>ray 1480,
 Ray: render tim                                    es/hour (13.57 sec. accum.).
PyMOL>undo
PyMOL>ray 1480,
 Ray: render tim
PyMOL>
/PpSMC5-NSE2_f
EVNGGGSGGGSGGG                                     AEITKLNQERKKNVLK
/PpSMC5-NSE2_f                                     1   106   111   116   121   126
EVNGGGSGGGSGGG                                     AEITKLNQERKKNVLKLKDVVKEYAKNQ
For Educationa
```

Movie Export

Width 1480

Height 614

Common: 720p 480p 360p

Rendering

○ Draw (fast) ● Ray (slow)

Movie Format

Encoder mpeg_encode

Quality 90%

○ MPEG 4

● MPEG 1

○ MOV (QuickTime)

○ GIF

○ PNG Images

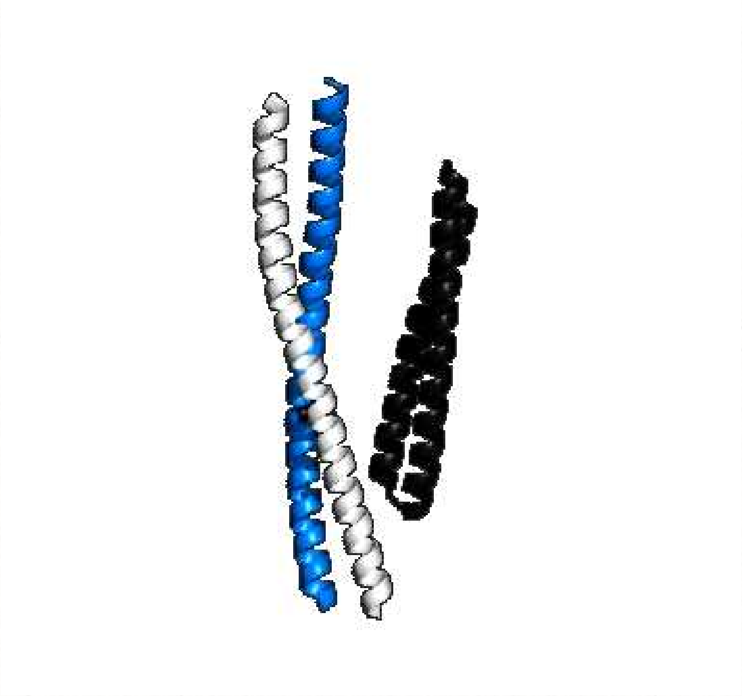
Tvorba videa v PyMol

Pymol:

Movie→ Program → Camera Loop → Y roll → 16 second

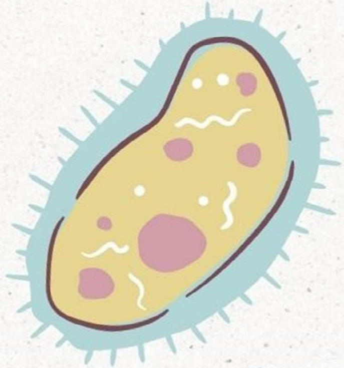
- Trial verze PyMol nepodporuje stažení videí, ale studentská verze ano

- Využití aplikace BandiCam https://www.bandicam.com/cz/
  Snímání obrazovky monitoru



Přednáška: Analýza proteinu

- Popis a funkce ☑
- Konzervovanost (Alignment)
- Evoluce (fylogenetický strom)
- Zjištění přítomnosti domén daného proteinu ☑
- Zjištění interakčních partnerů ☑
- Struktura proteinu – Alphafold/Colabfold
- Úprava proteinového modelu v PyMol ☑
  - Zvýraznění interakčních aminokyselin
  - Zvýraznění domén



Zkouška:

test 11.5. 2023 9:00 C2-211
+ přednáška/prezentace + nová data (výběr termínu)

Prezentace - Analýza proteinu ☑

Konkrétní nová data – článek (< 5 let) o komplexu (nebo proteinu)

Ujasnit si souvislosti, rozšířit si znalosti, aplikovat poznatky z přednášek ...