Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2025/2026

„Protein bioinformatics I“

Retrieving protein sequences from databases (Uniprot: FASTA formate)

Computing amino-acids compositions, molecular weight, isoelectric point, and other parameters (SMS)

Prediction of proteases cutting (PeptideCutter)

Predicting elements of protein secondary structure, domains

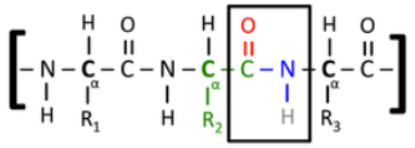
Predicting 3-D structure and the domain organization of proteins

Finding all proteins that share a similar sequence and Classifying proteins into families

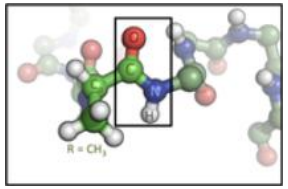
Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…



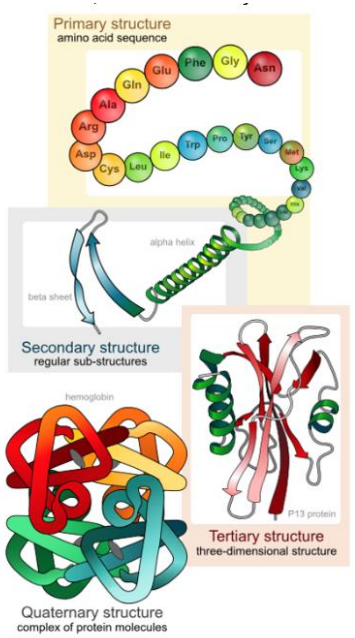
Proteins



20 Aminoacids – primary structure:

(Frederick Sanger-1958 Nobel prize for insulin sequencing)

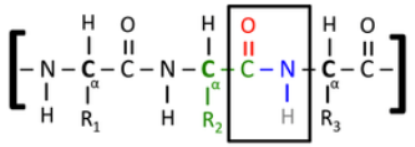
Secondary structure
Tertiary structure
Quaternary structure



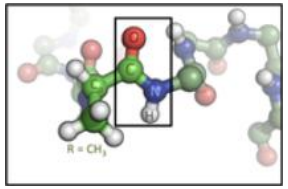
1-letter code	3-letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	Stop codon	TAA, TAG, TGA
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG

SEQUENCE ⇨ STRUCTURE ⇨ FUNCTION

Protein sequences are the fundamental determinants of biological structure and function.

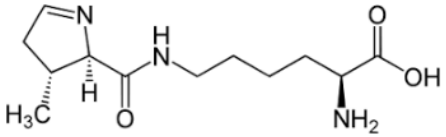


Proteins

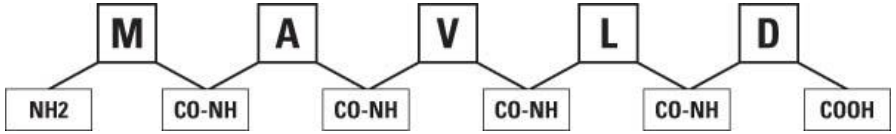


20 Aminoacids – primary structure:

J	Xle	Isoleucine/Leucine
O	Pyl	Pyrrolysine
U	Sec	Selenocysteine
X	Xaa	Any residue



N-terminus → C-terminus



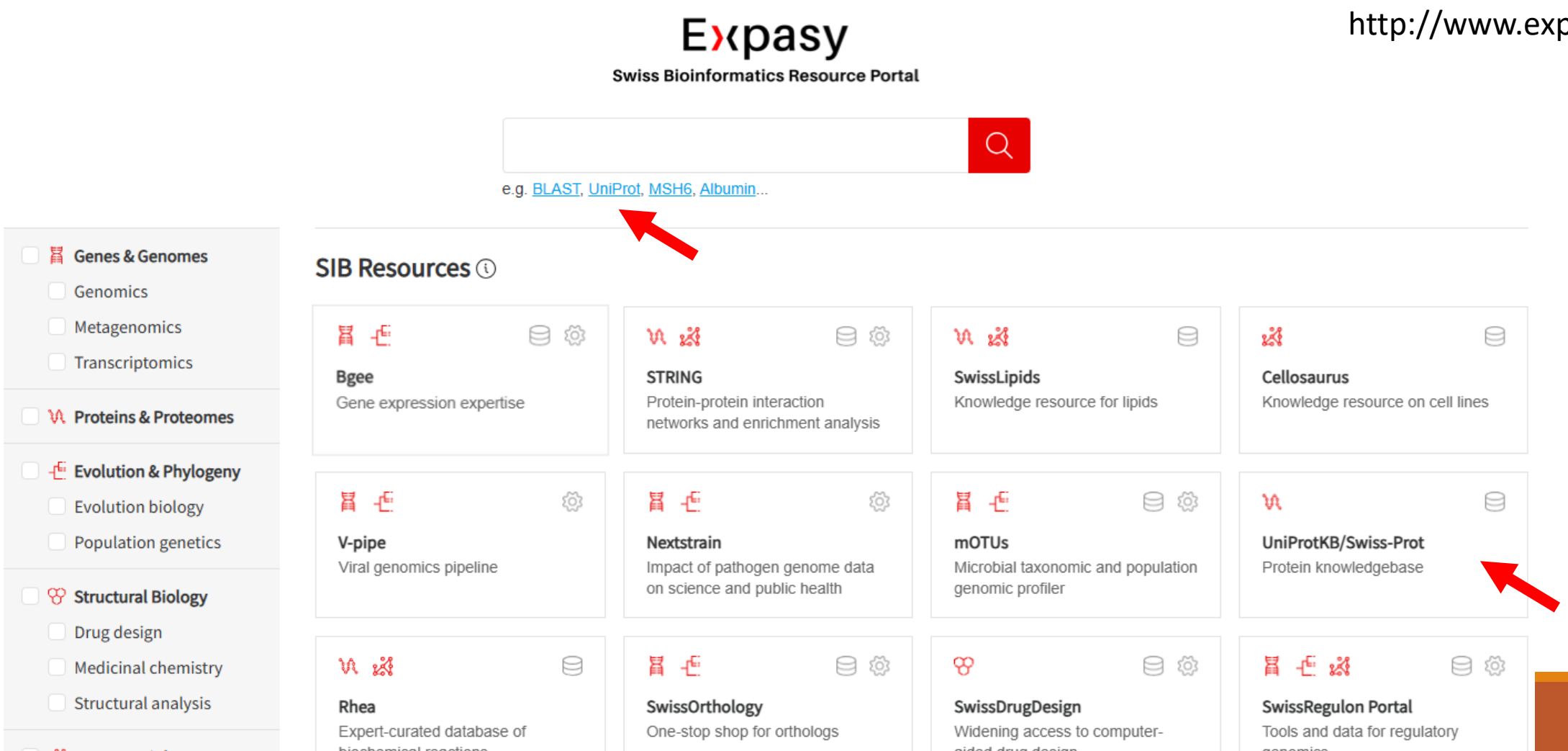
1-letter code	3-letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	Stop codon	TAA, TAG, TGA
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG

Databases

	Primary database	Secondary database
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	• ENA, GenBank and DDBJ (nucleotide sequence) • ArrayExpress Archive and GEO (functional genomics data) • Protein Data Bank (PDB; coordinates of three-dimensional macromolecular structures)	• InterPro (protein families, motifs and domains) • UniProt Knowledgebase (sequence and functional information on proteins) • Ensembl (variation, function, regulation and more layered onto whole genome sequences)

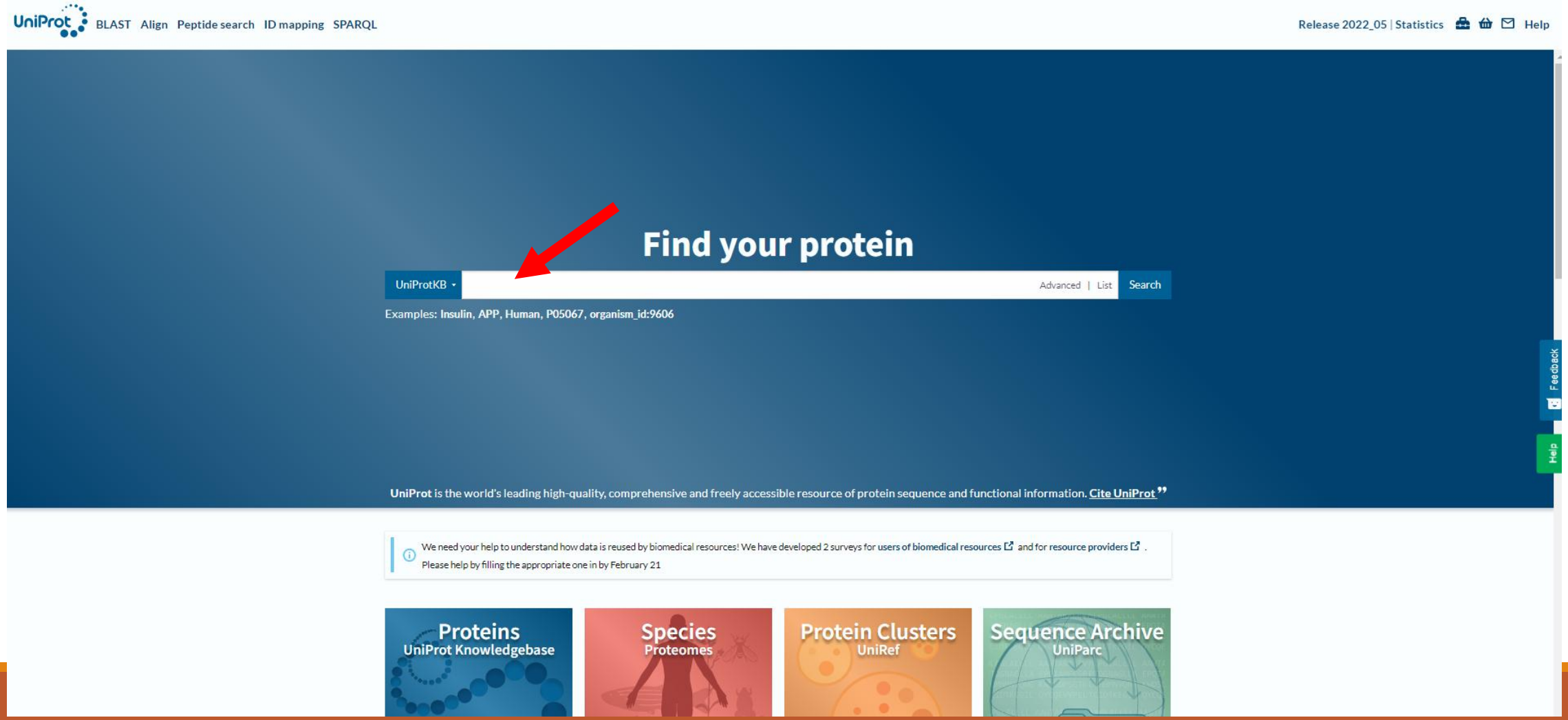
Protein database: Expasy/UniProt

http://www.expasy.org/



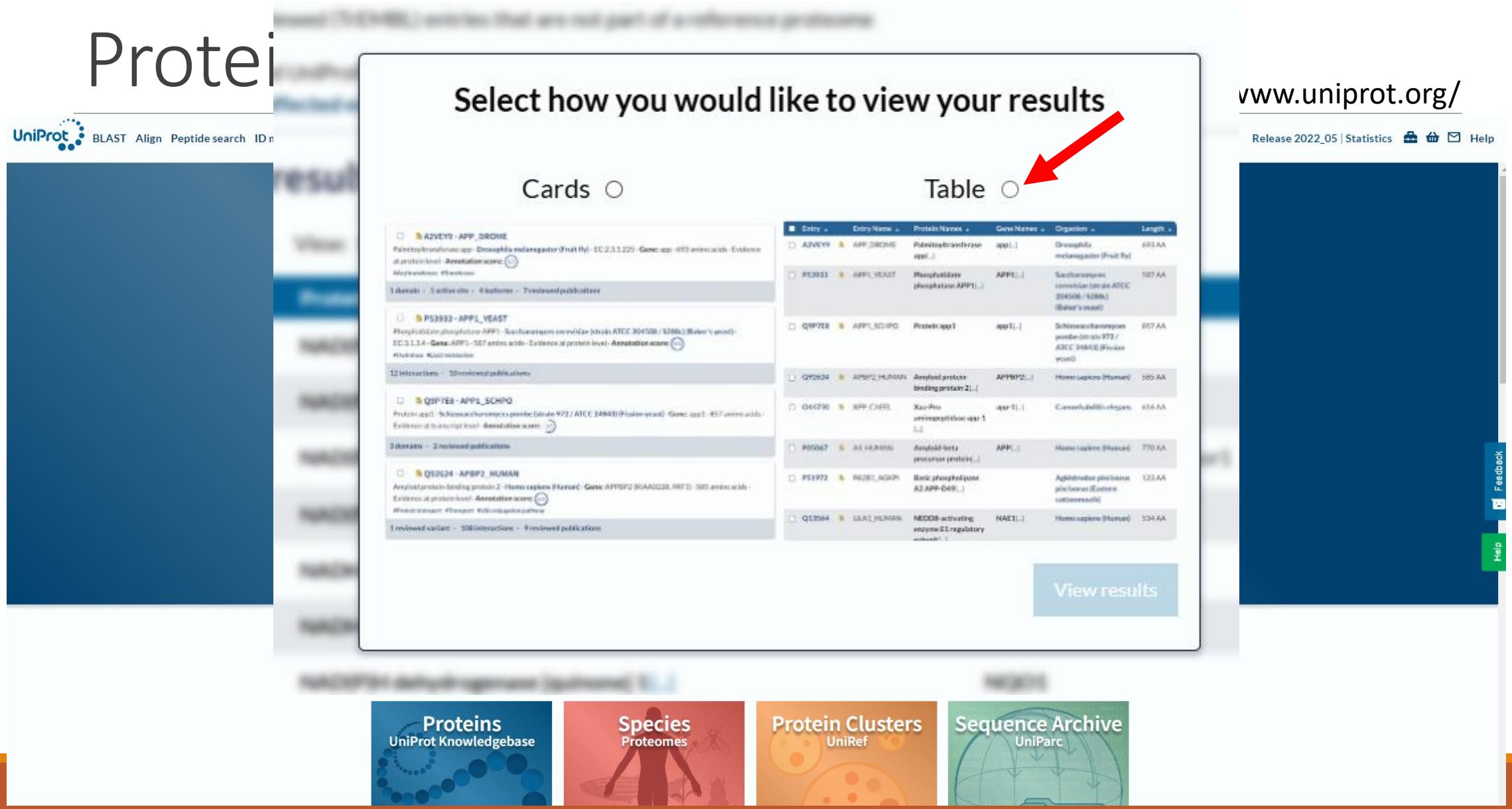
Protein database: UniProt

http://www.uniprot.org/



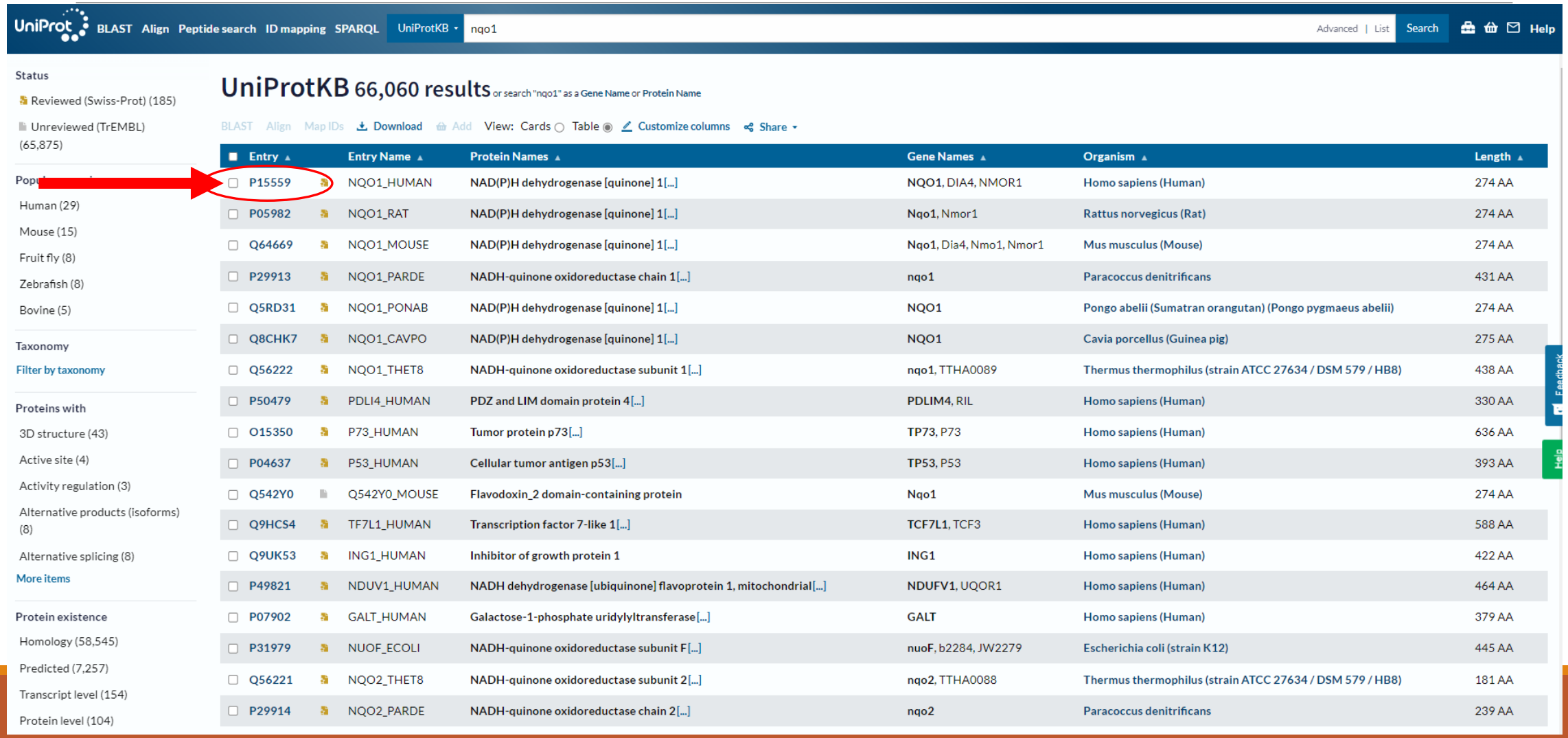
Protei

www.uniprot.org/

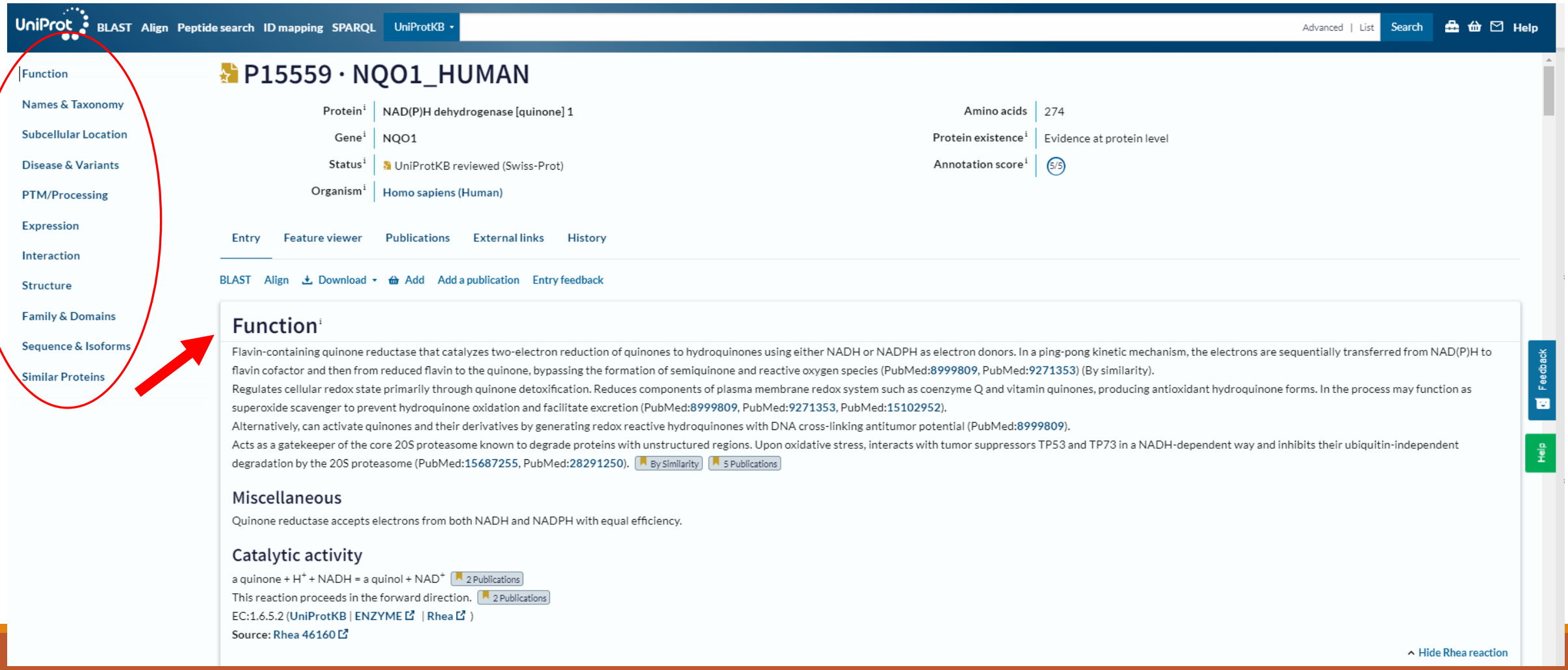


Protein database: UniProt

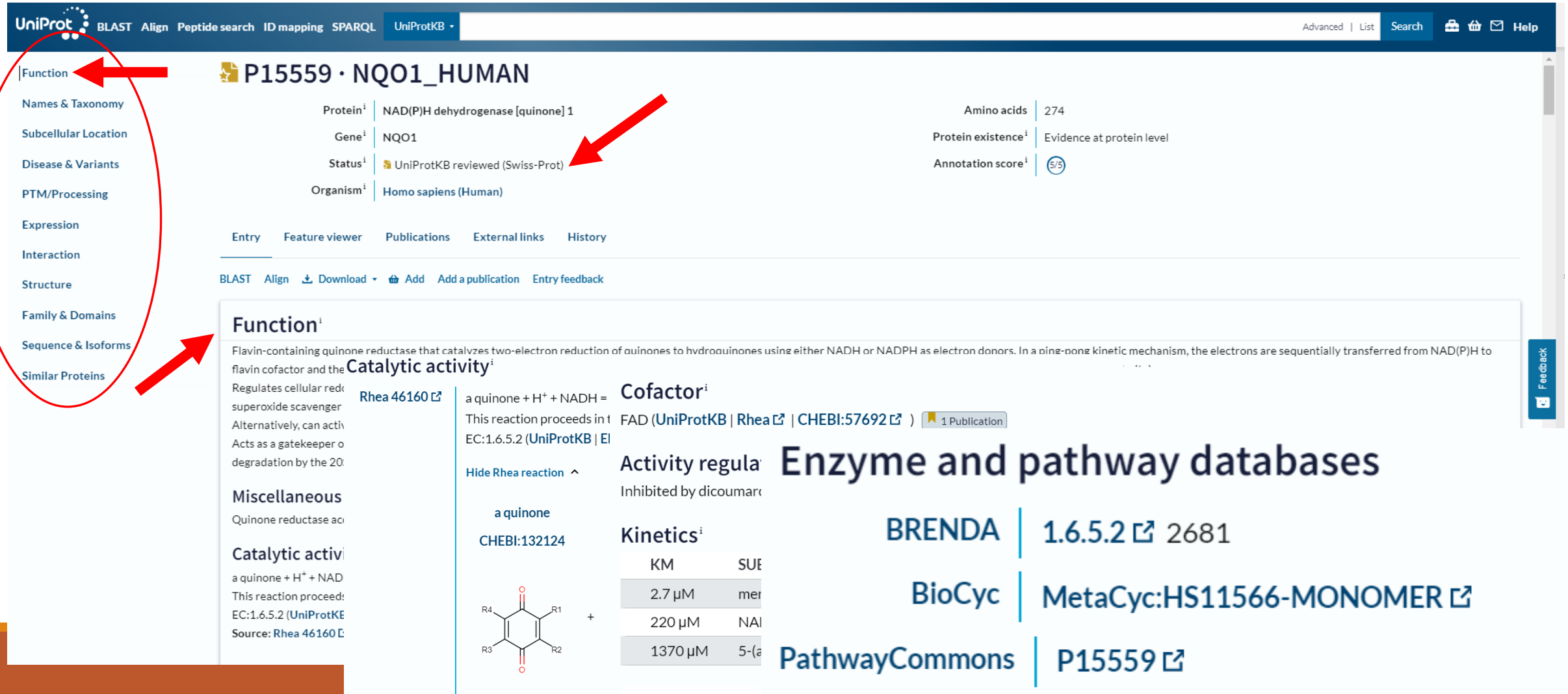
http://www.uniprot.org/



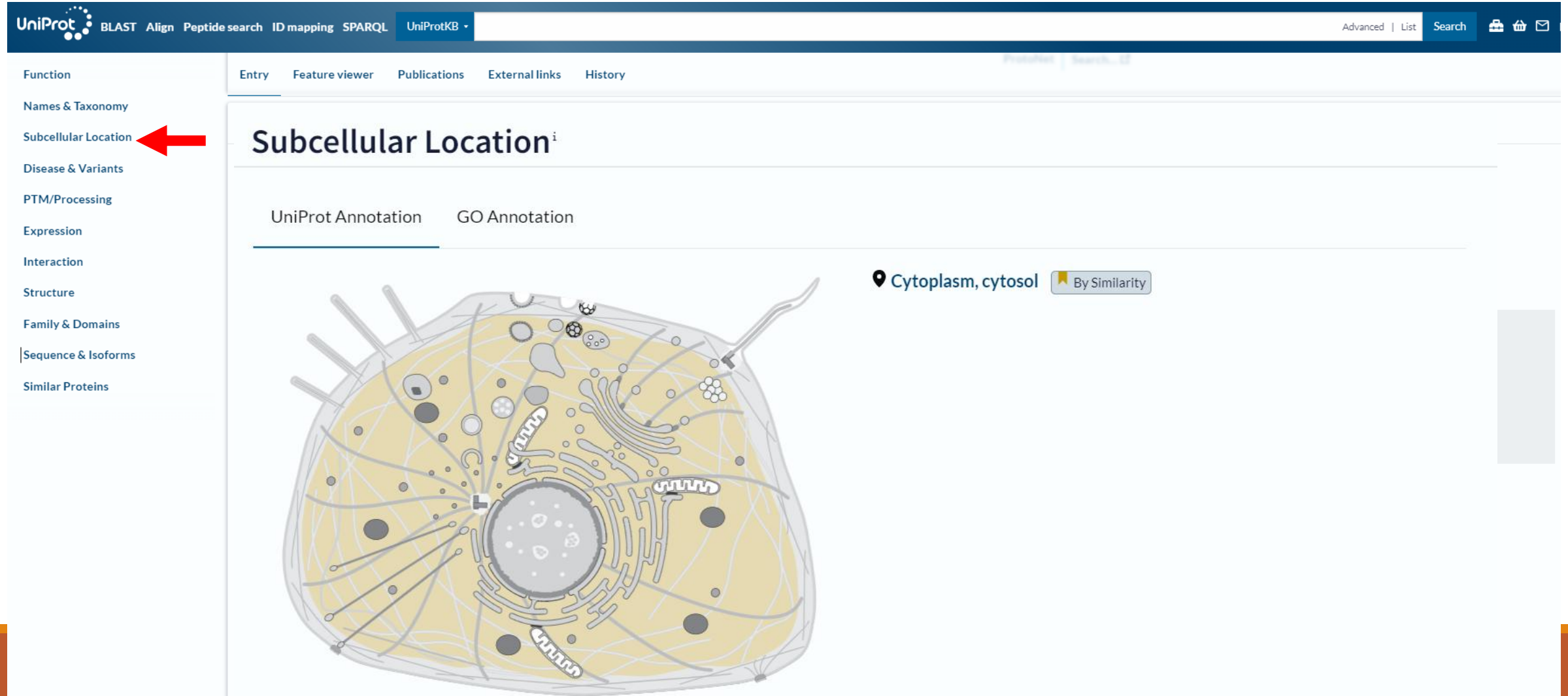
Protein database: UniProt



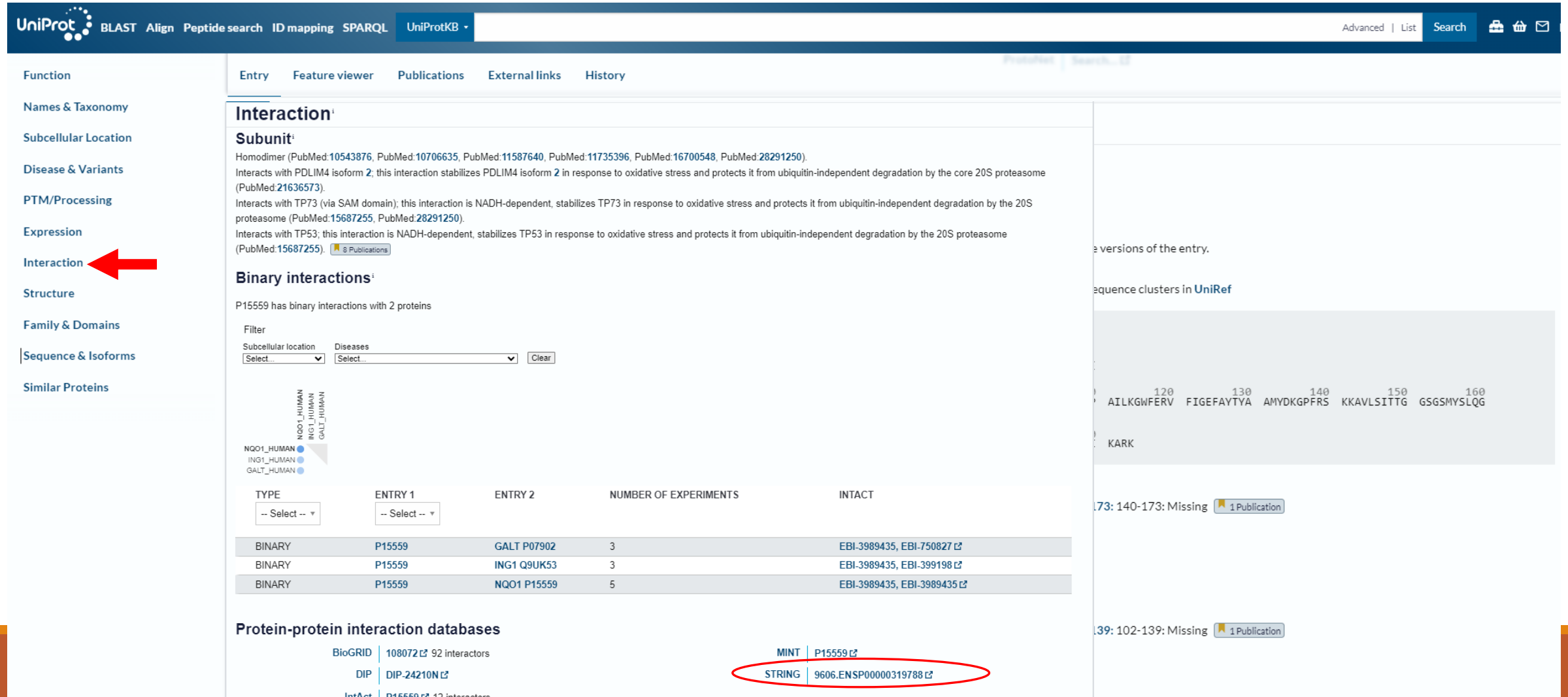
Protein database: UniProt



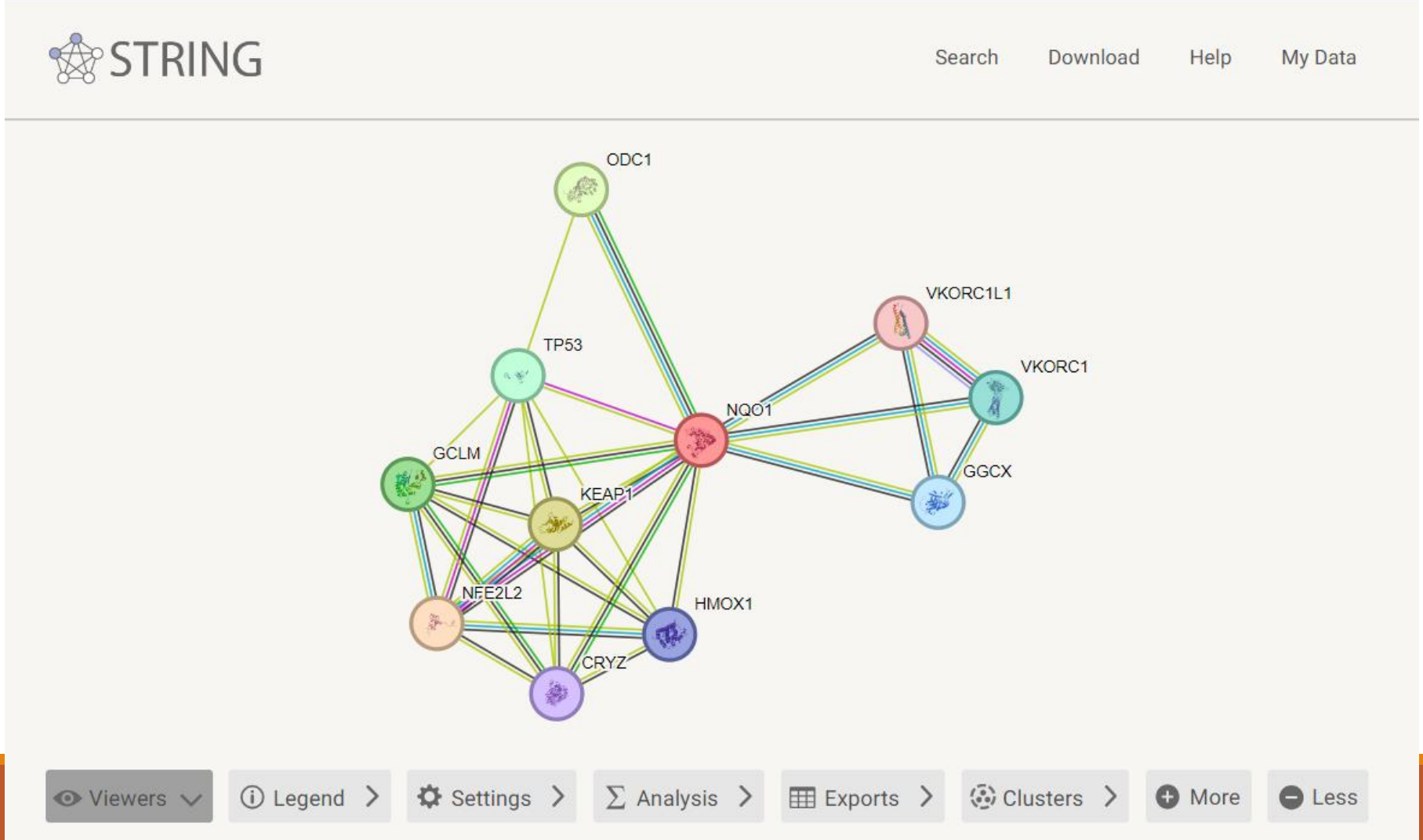
Protein database: UniProt



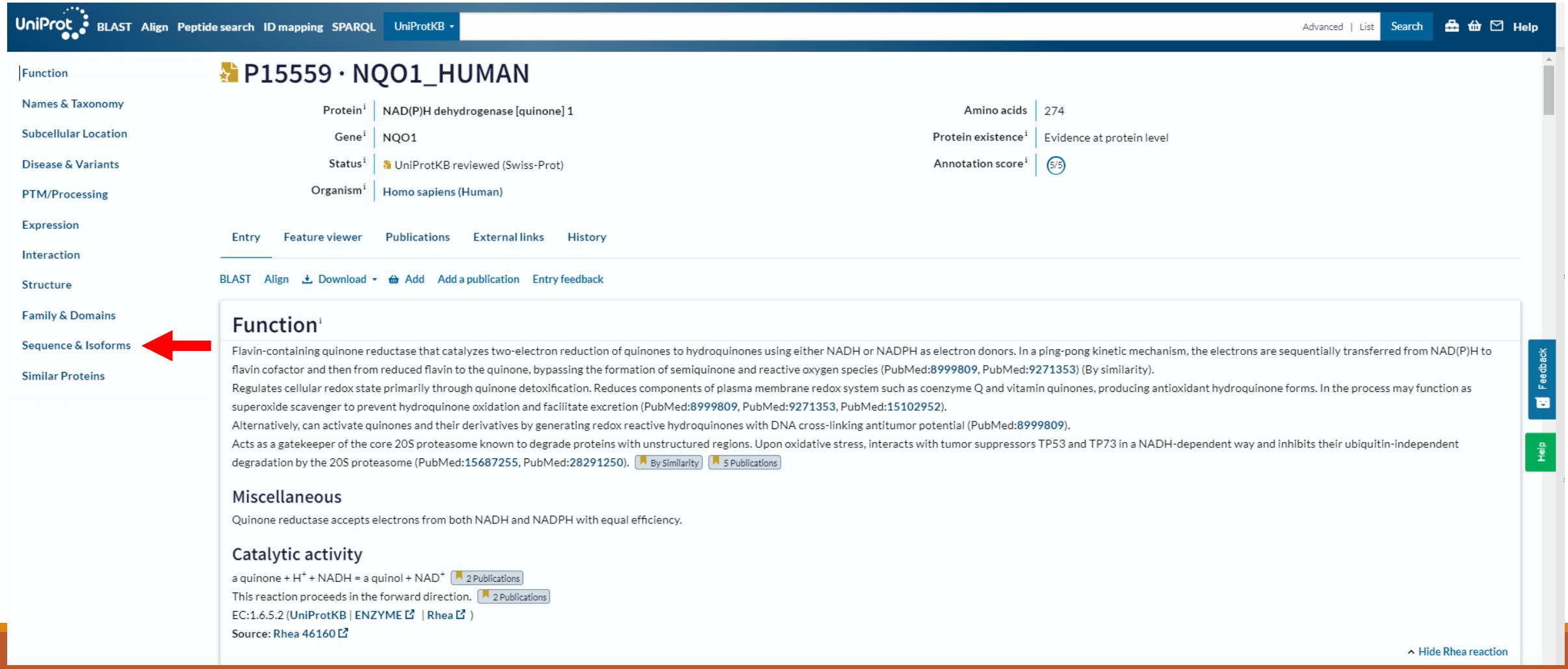
Protein database: UniProt



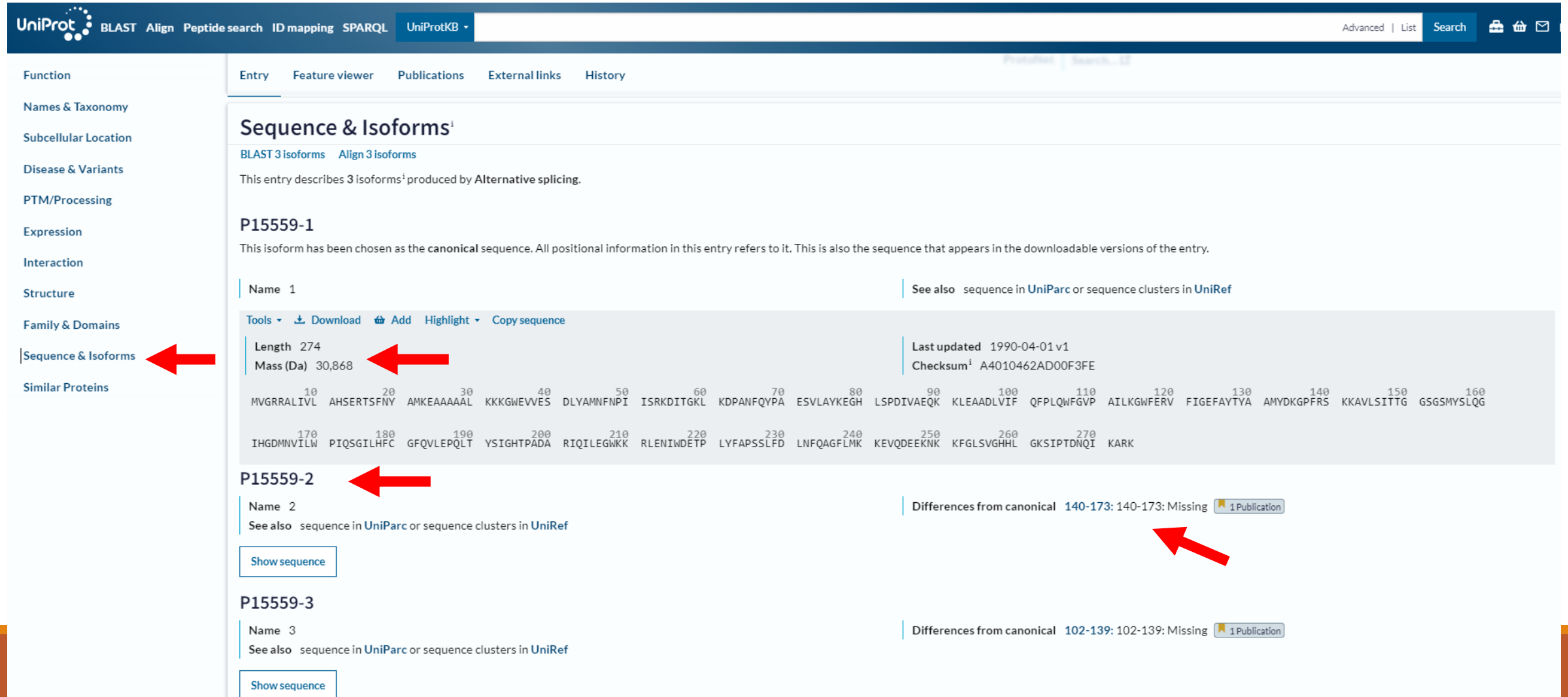
STRING: protein-protein interaction



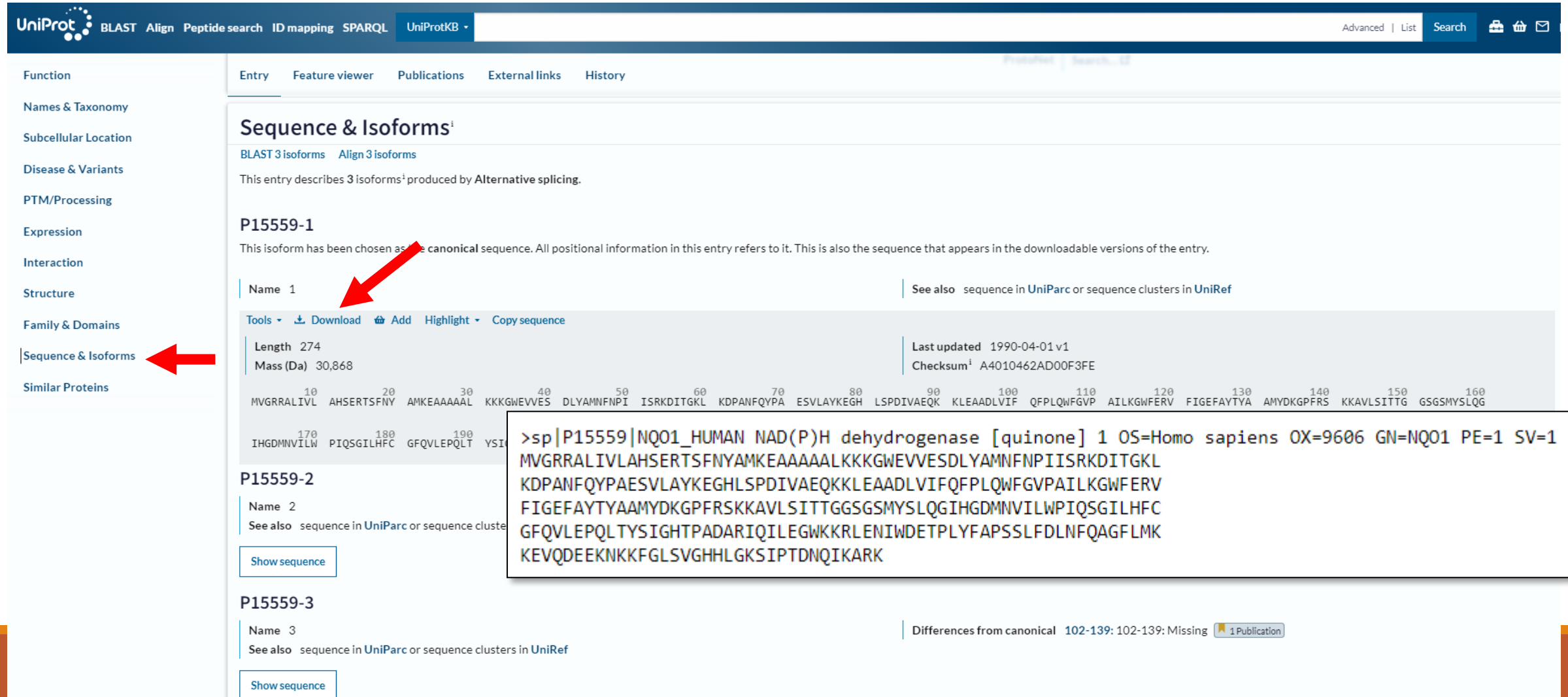
Protein database: UniProt



Protein database: UniProt



Protein database: UniProt



FASTA (and RAW) format

FASTA = popular tool for sequence comparison and database searching

W.R. Pearson a D.J. Lipman 1988

fasta format:

>NQO1_homo

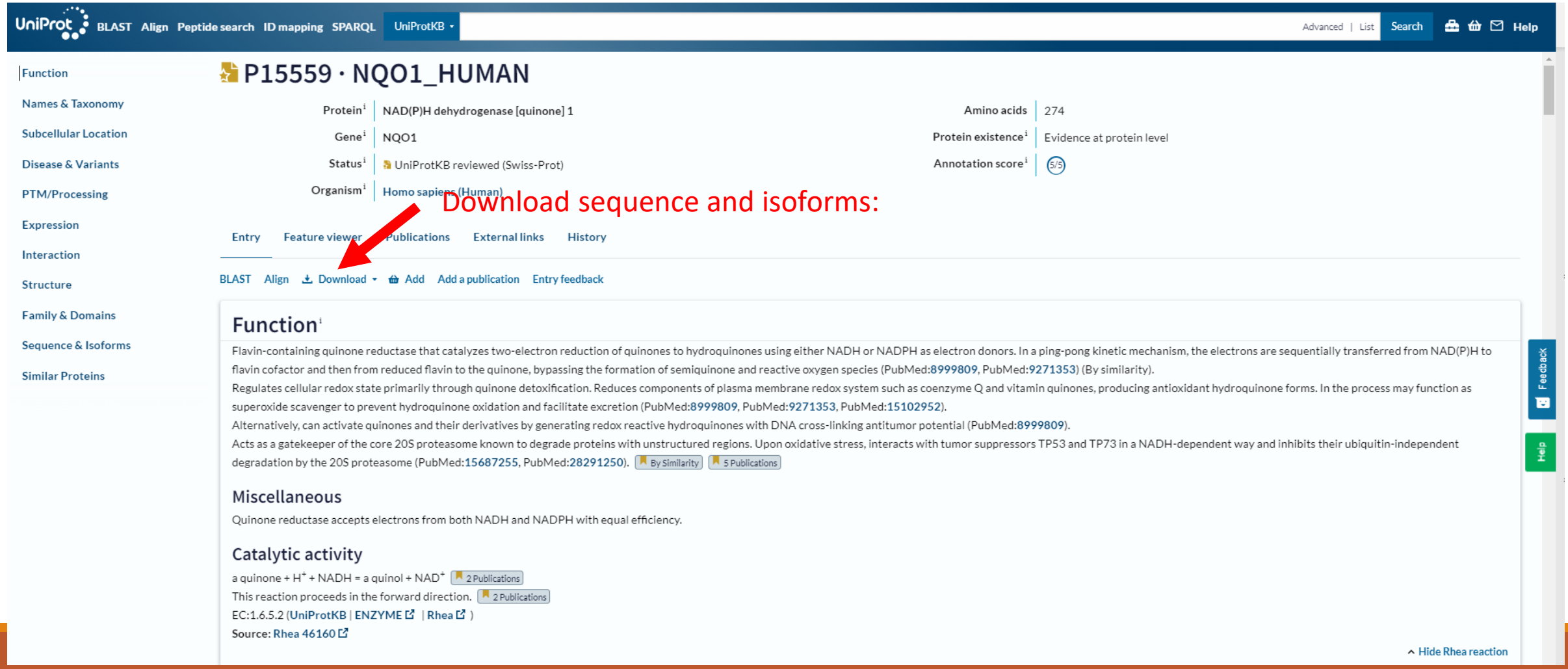
```
MVGRRALIVLAHSERTSFNYAMKEAAAAA
LKKKGWEVVESDLYAMNFNPIISRKDITG
KLKDPANFQYPA
```

```
MVGRRALIVLAHSERTSFNYAMKEAAAAA
LKKKGWEVVESDLYAMNFNPIISRKDITG
KLKDPANFQYPA
```

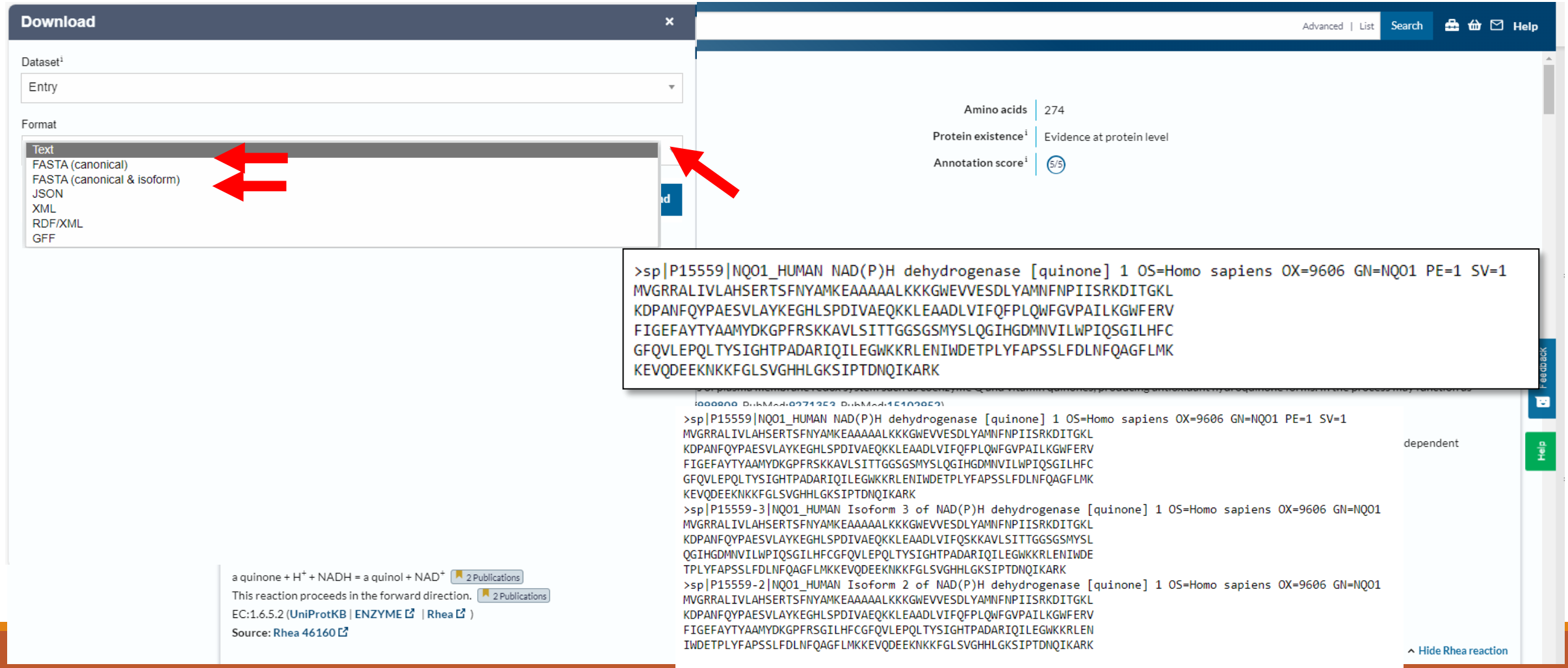
1. „definition“ line starts with **>** and unique identification follows.

2. line-a sequence (DNA/protein- single letter code).

Protein database: UniProt



Protein database: UniProt



Practical part in UniProt

Find the accession number of human cytochrome P450 3A4 (CYP3A4)

What function has and where is locallized protein Q9C0B1?

What function has enzyme O95251 and how many isoforms it has?

Practical part in UniProt

Find the accession number of „reviewed“ human cytochrome P450 3A4 (CYP3A4)

P08684 · CP3A4_HUMAN

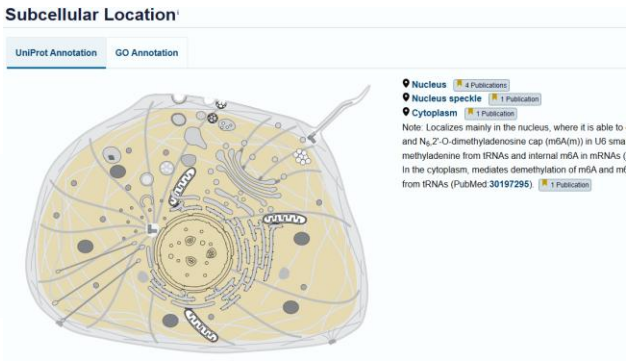
Protein	Cytochrome P450 3A4	Amino acids	503 (go to sequence)
Gene	CYP3A4	Protein existence	Evidence at protein level
Status	UniProtKB reviewed (Swiss-Prot)	Annotation score	5/5
Organism	Homo sapiens (Human)		

What function has and where is locallized protein Q9C0

- RNA demethylase

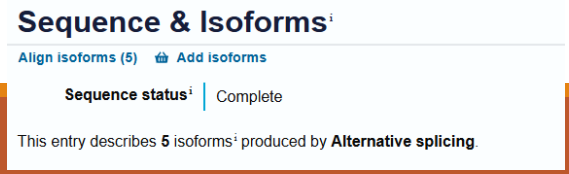
Q9C0B1 · FTO_HUMAN

Protein	Alpha-ketoglutarate-dependent dioxygenase FTO

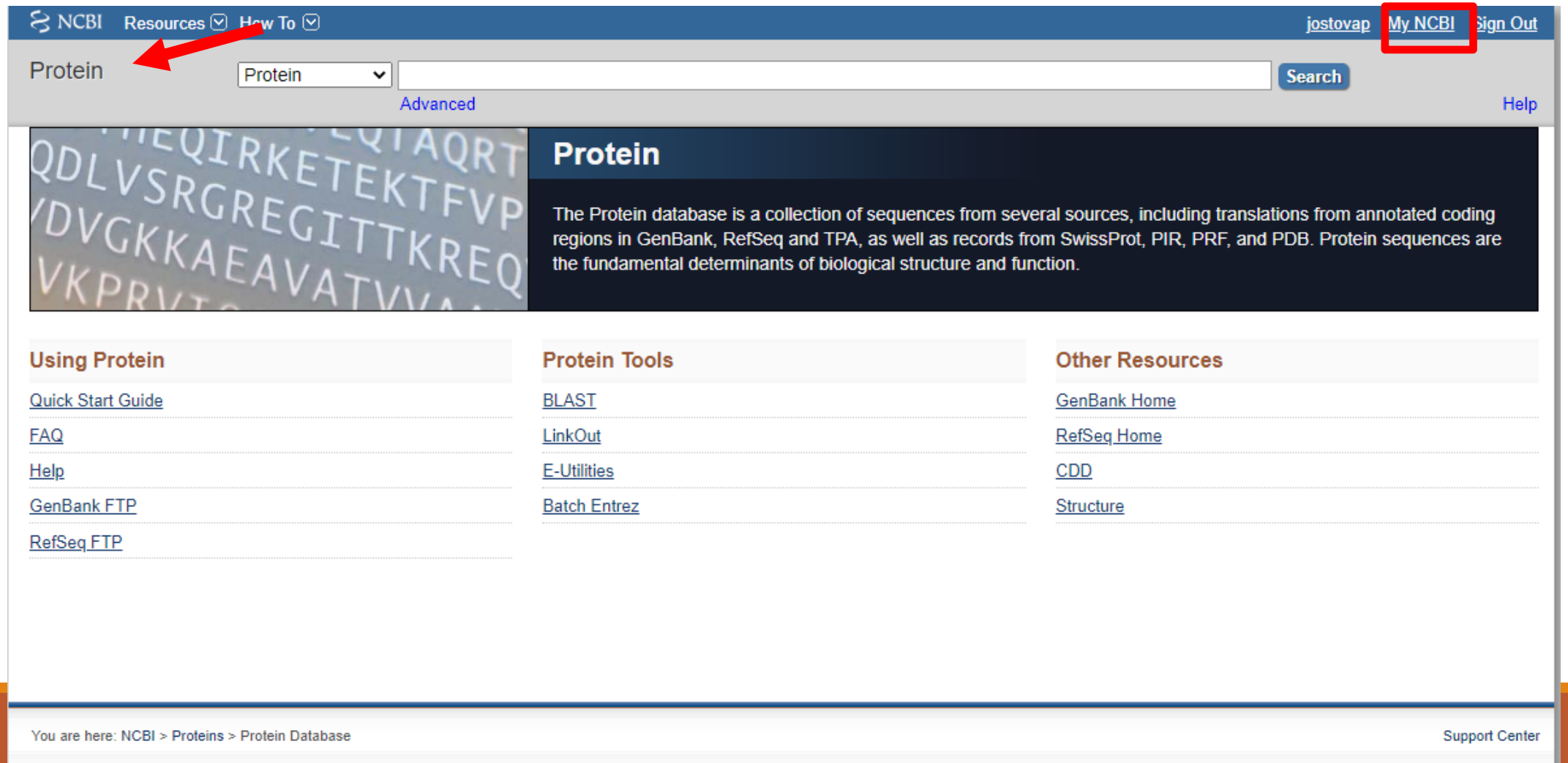


What function has enzyme O95251 and how many isoforms it has?

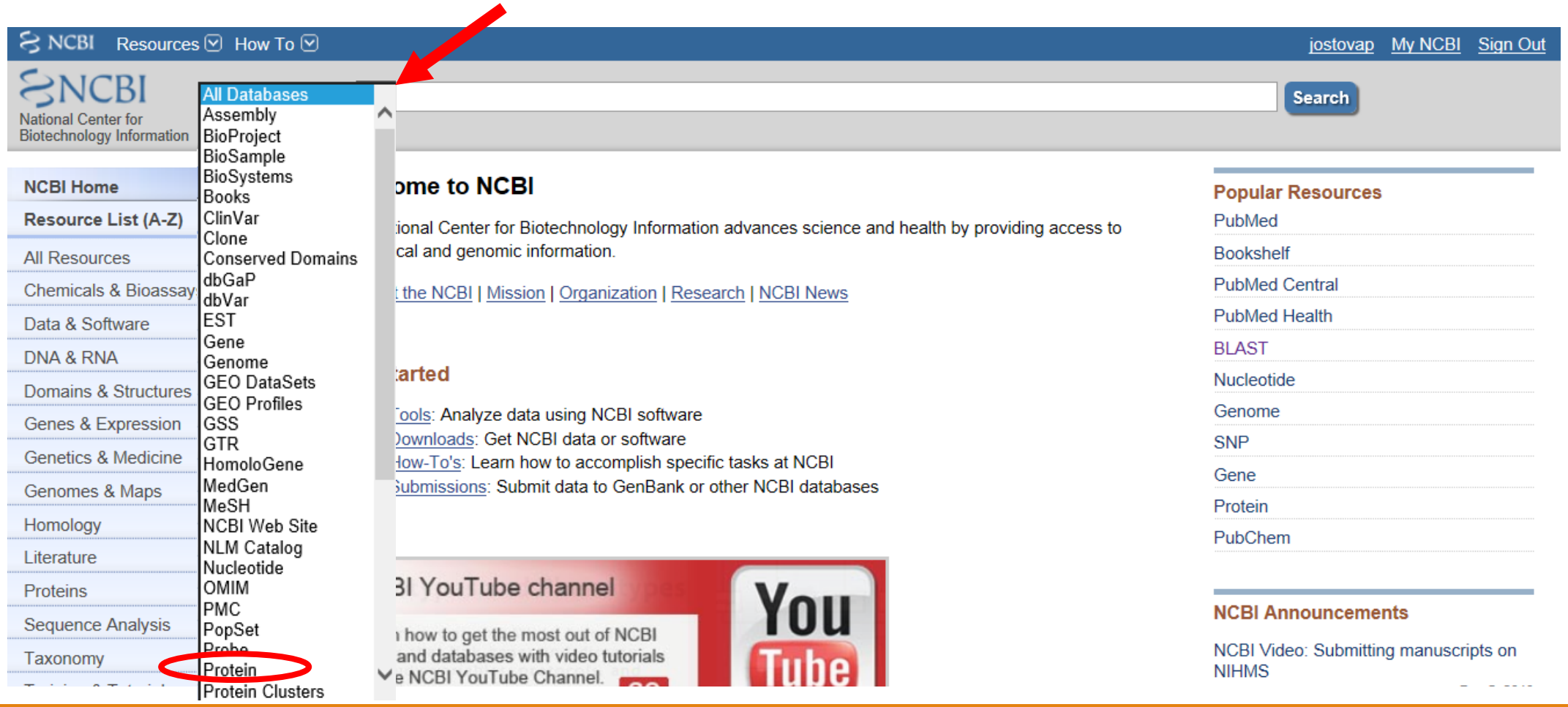
- **Histone acetyltransferase KAT7**



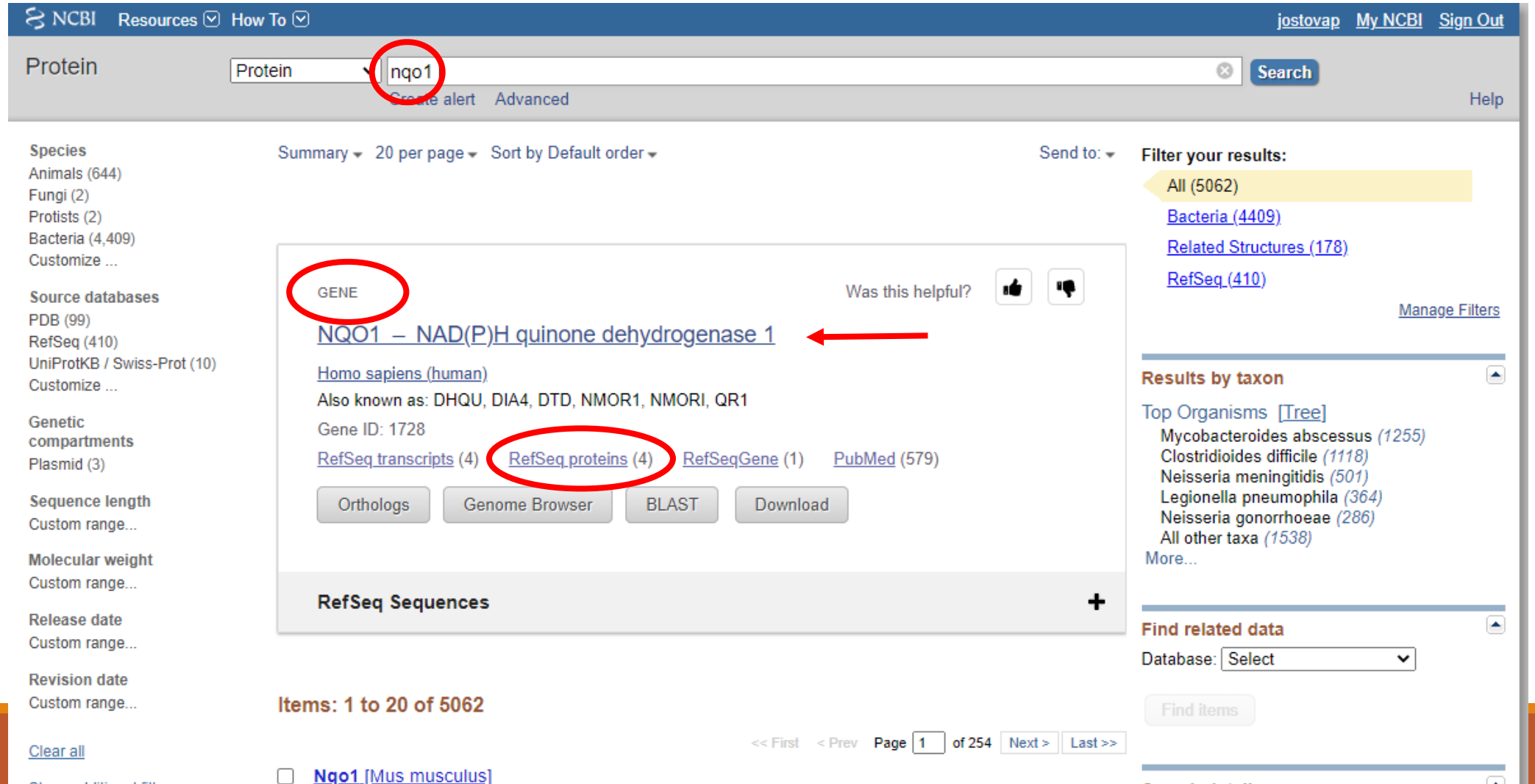
Protein database: NCBI



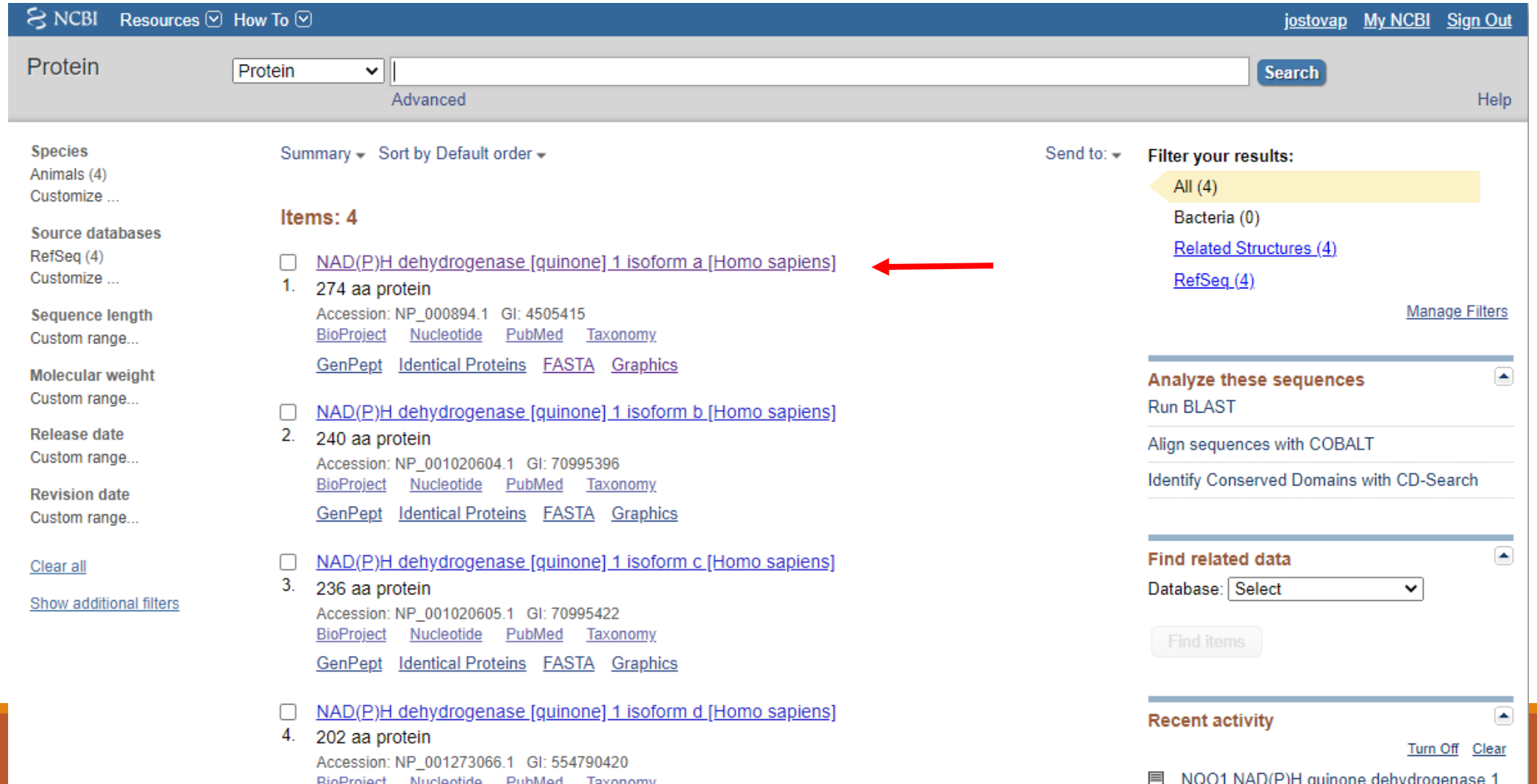
Protein database: NCBI



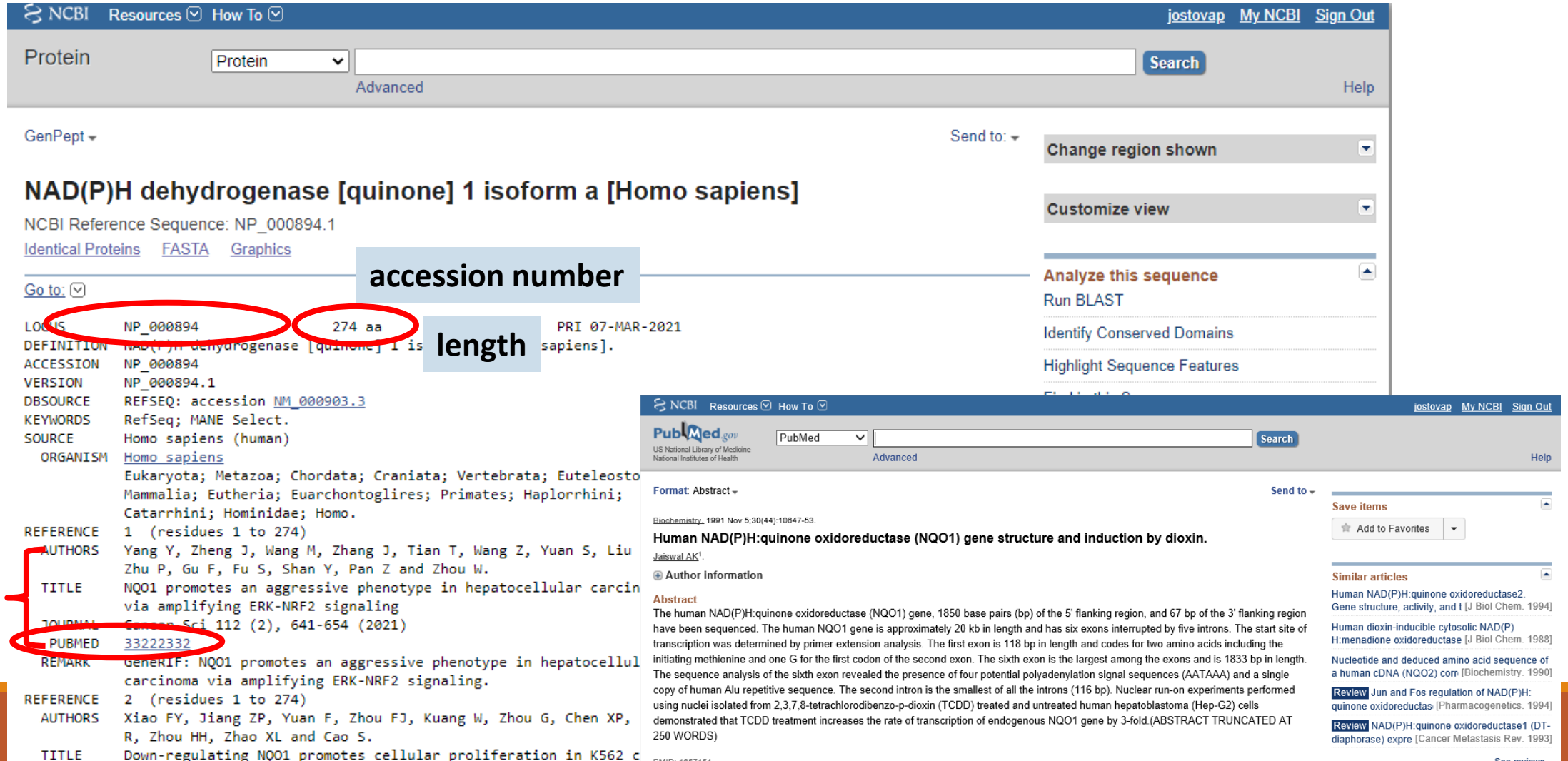
Protein database: NCBI



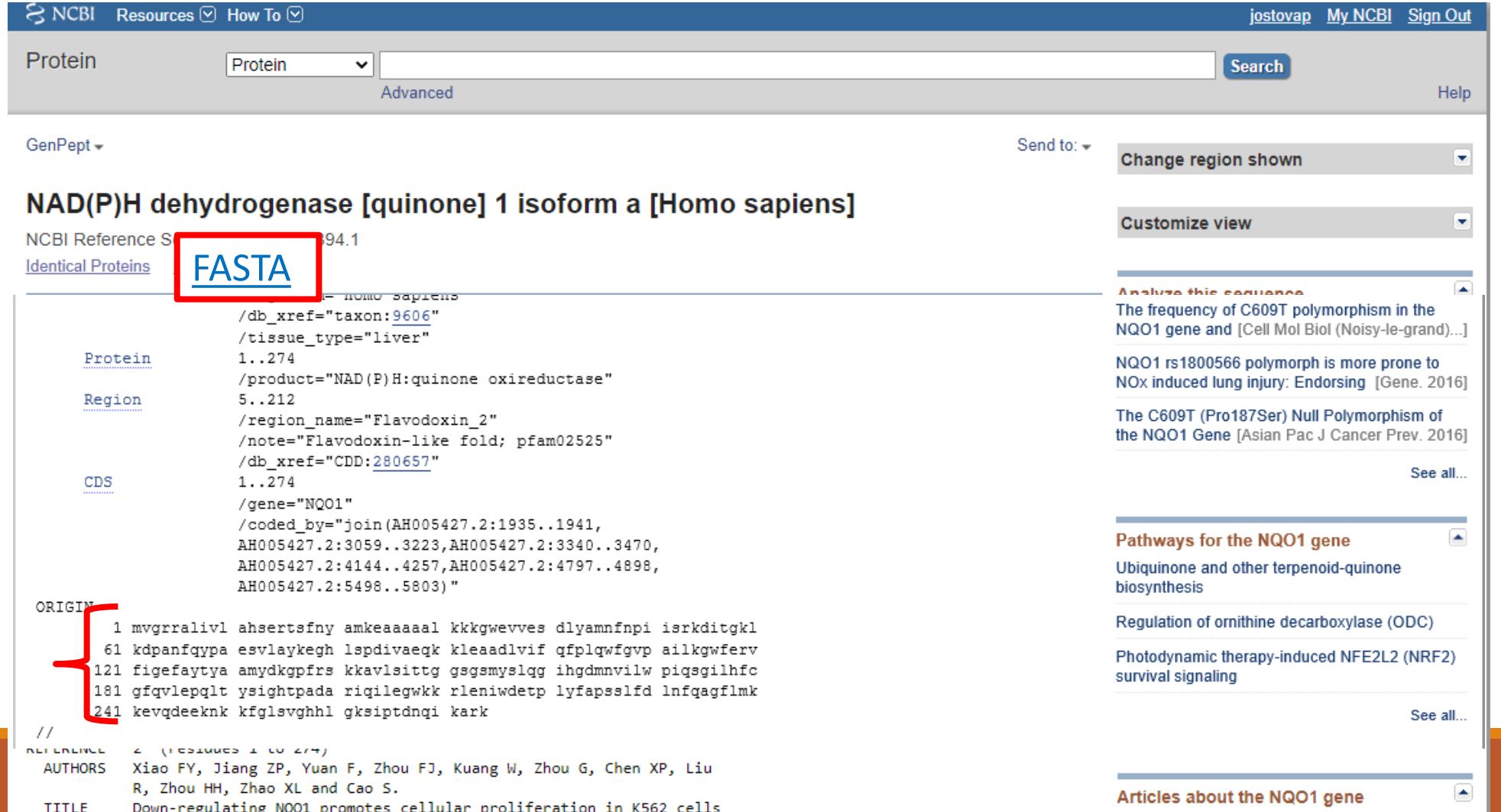
Protein database: NCBI



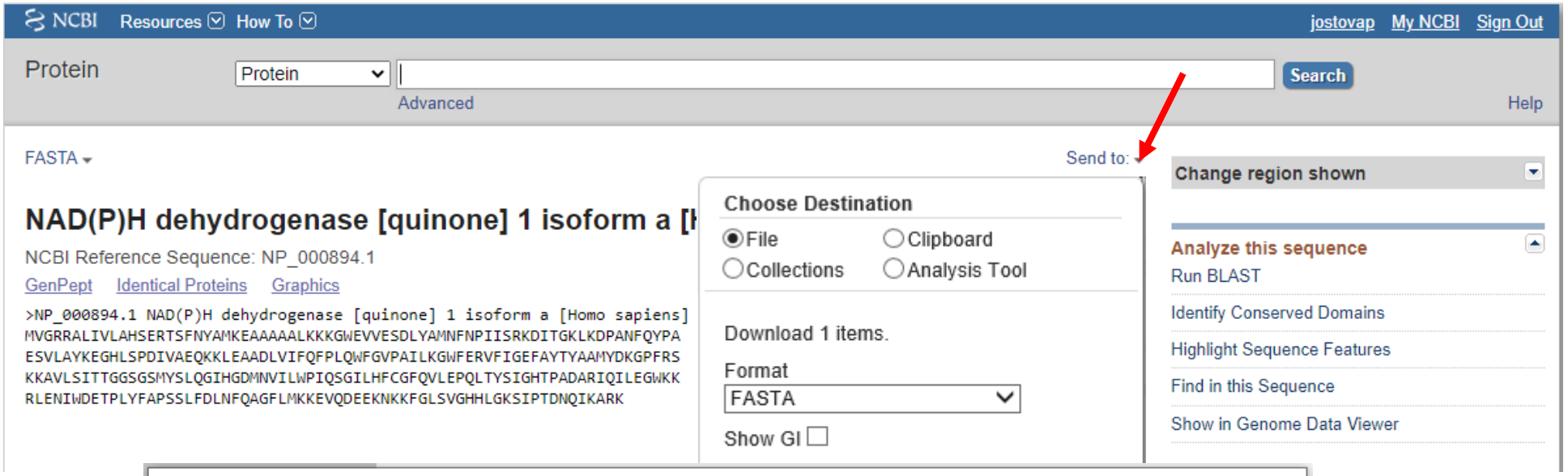
Protein database: NCBI

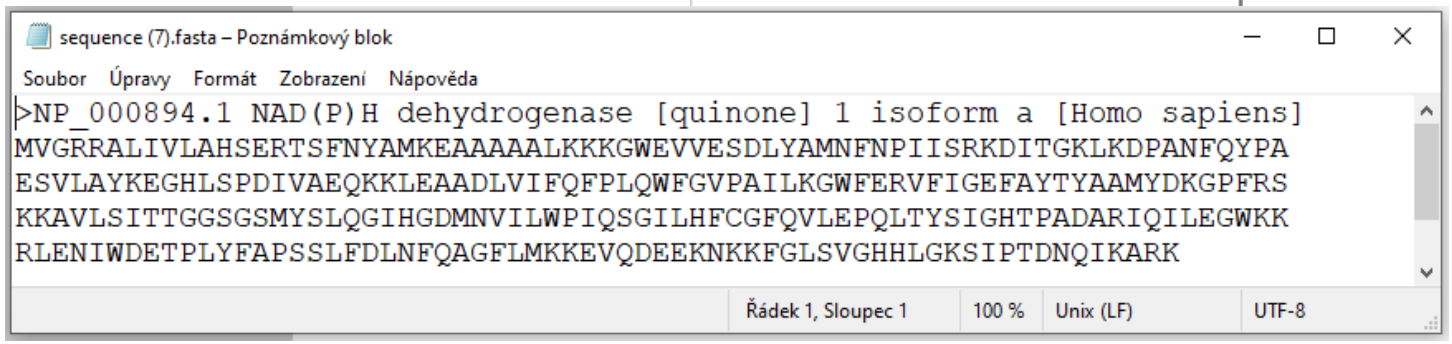


Protein database: NCBI

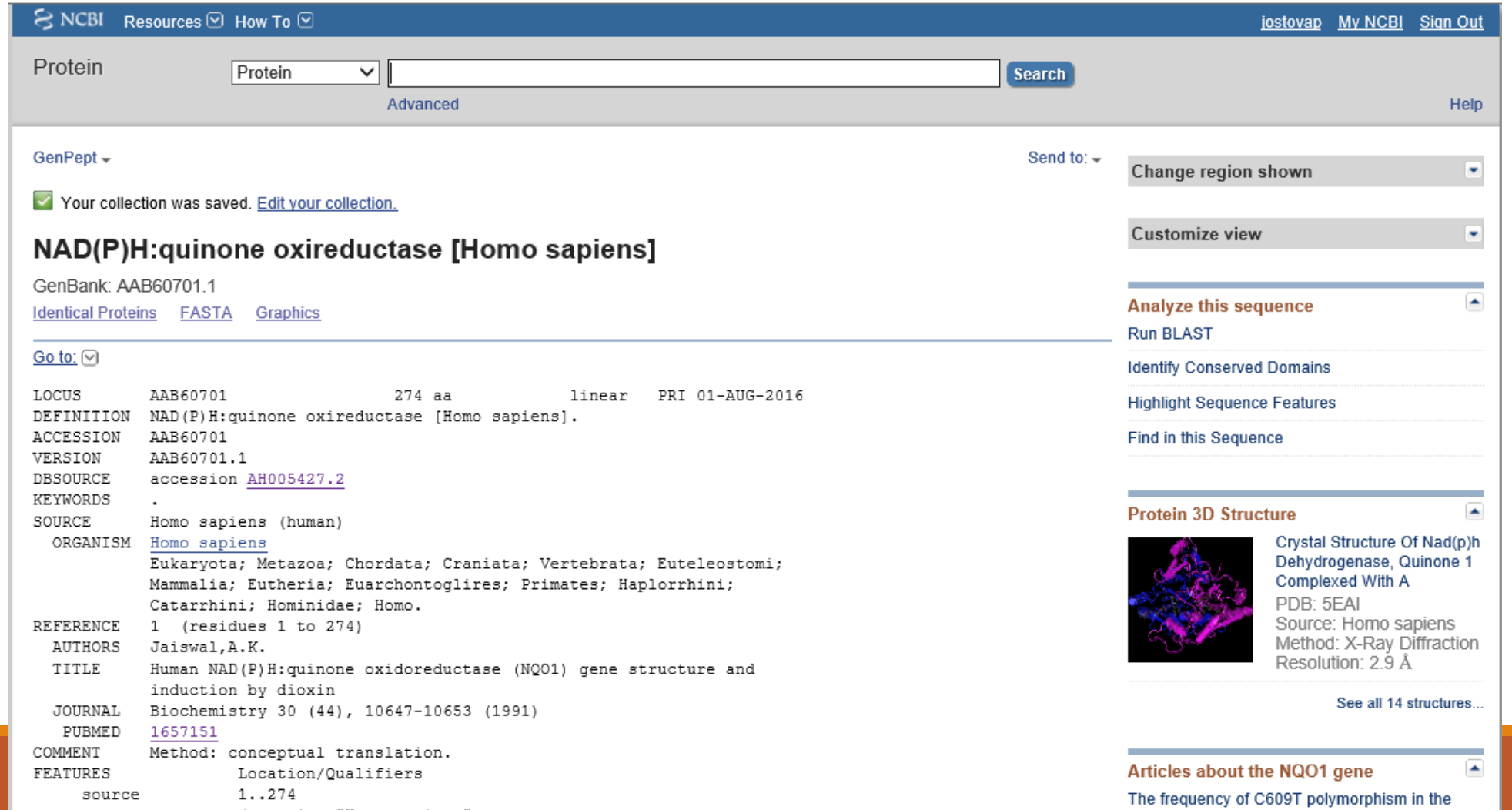


Protein database: NCBI

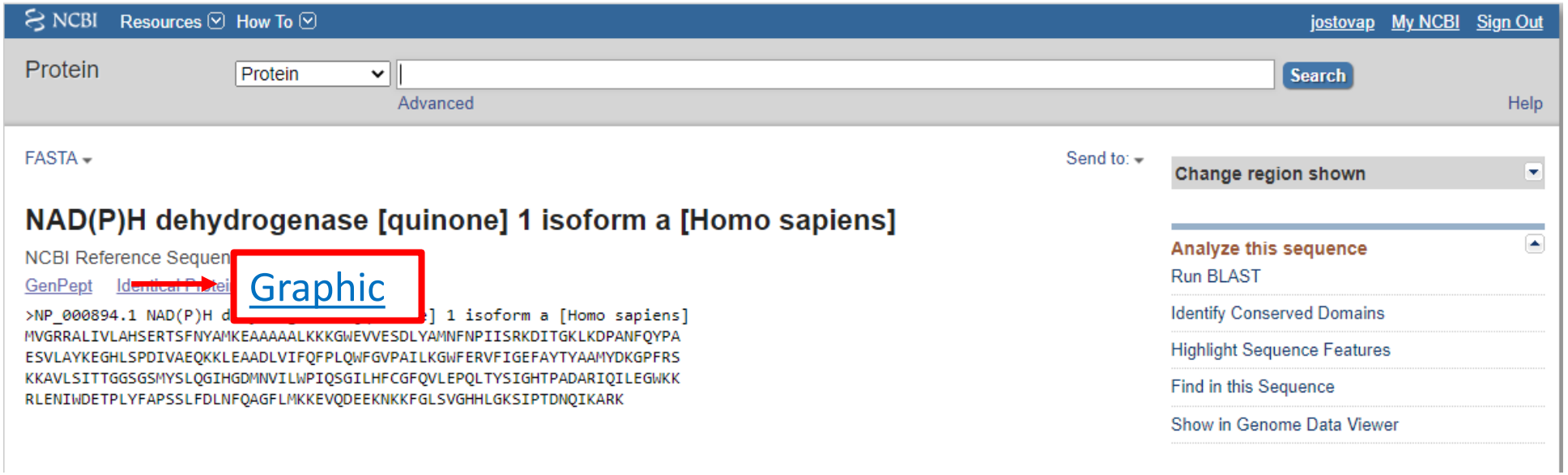




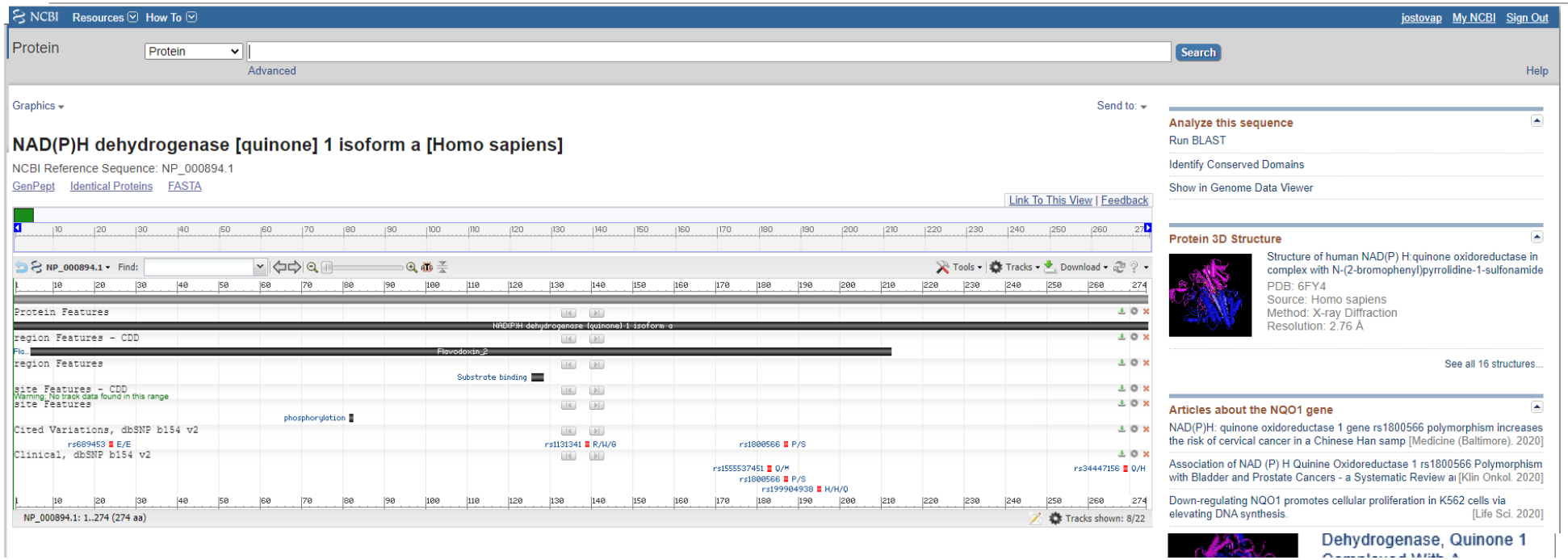
Protein database: NCBI → MyNCBI



Protein database: NCBI



Protein database: NCBI



Practical part with NQO1

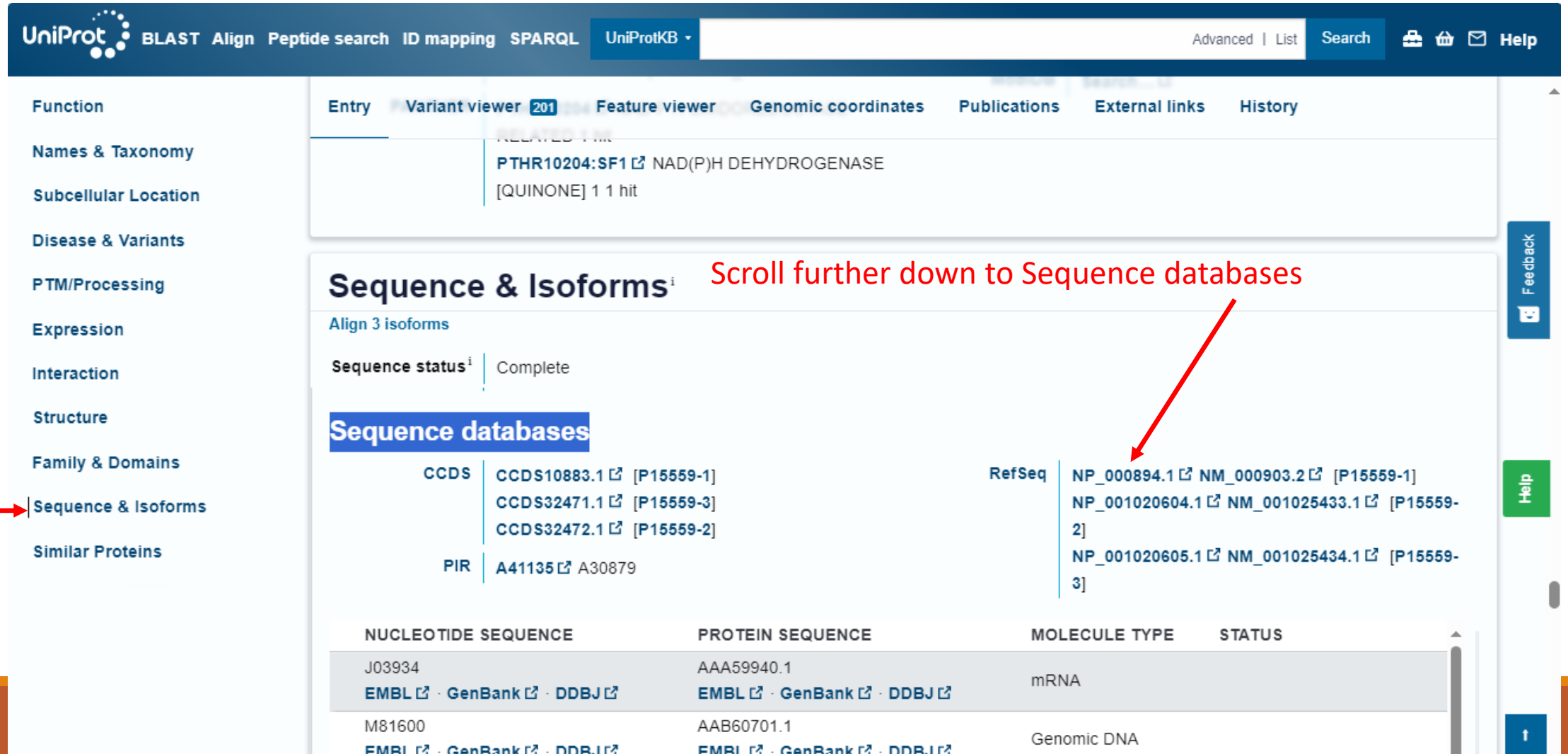
Try Uniprot and protein NCBI:

Find your sequnce

-what is the function of your protein?
-find subcellular location of your protein
-is it involved in any pathology?
-how many amino acids has your sequence?

-try direct search in NCBI
-**try link through Sequence databases → Refseq from Uniprot**

link through Sequence databases → Refseq from Uniprot



„Protein bioinformatics I“

Retrieving protein sequences from databases

Computing amino-acids compositions, molecular weight, isoelectric point, and other parameters

Prediction of proteases cutting

Predicting elements of protein secondary structure, domains

Predicting 3-D structure and the domain organization of proteins

Finding all proteins that share a similar sequence and Classifying proteins into families

Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…

Protein Sequence Analysis

SMS

Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein
Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats
-Restriction Digest
-Restriction Summary
-Reverse Translate
-Translate

Sequence Manipulation Suite:

About

The Sequence Manipulation Suite is written in JavaScript 1.5, which is a lightweight, cross-platform, object-oriented scripting language. JavaScript is now standardized by the ECMA (European Computer Manufacturers Association). The first version of the ECMA standard is documented in the ECMA-262 specification. The ECMA-262 standard is also approved by the ISO (International Organization for Standards) as ISO-16262. JavaScript 1.5 is fully compatible with ECMA-262, Edition 3.

Sequences submitted to the Sequence Manipulation Suite do not leave your computer and are instead manipulated by your web browser, which executes the JavaScript. The Sequence Manipulation Suite was written by Paul Stothard (University of Alberta, Canada). Send questions and comments to stothard@ualberta.ca.

Here are short descriptions of the programs that comprise the Sequence Manipulation Suite:

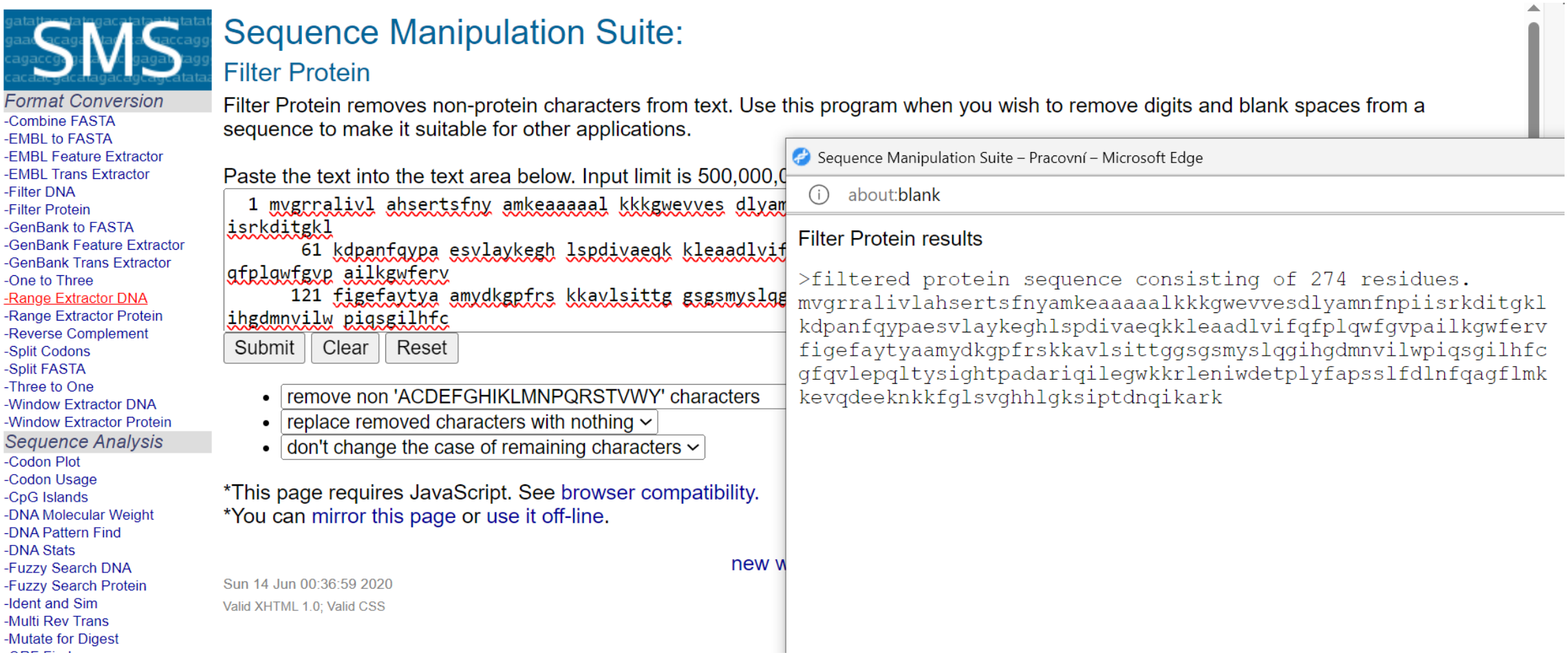
Format Conversion:

- **Combine FASTA** - converts multiple FASTA sequence records into a single sequence. Use Combine FASTA, for example, when you wish to determine the codon usage for a collection of sequences using a program that accepts a single sequence as input.
- **EMBL to FASTA** - accepts one or more EMBL files as input and returns the DNA sequence from each in FASTA format. Use this program when you wish to quickly remove all of the non-DNA sequence information from an EMBL file.
- **EMBL Feature Extractor** - accepts one or more EMBL files as input and reads the sequence feature information described in the feature tables. The program extracts or highlights the relevant sequence segments and returns each sequence feature in FASTA format. EMBL Feature Extractor is particularly helpful when you wish to derive the sequence of a cDNA from a genomic sequence that contains many introns.
- **EMBL Trans Extractor** - accepts one or more EMBL files as input and returns each of the protein translations described in the files in FASTA format. EMBL Trans Extractor can be used when you are more interested in the predicted protein translations of a DNA sequence than the DNA sequence itself.
- **Filter DNA** - removes non-DNA characters from text. Use this program when you wish to remove digits and blank spaces from a sequence to make it suitable for other applications.
- **Filter Protein** - removes non-protein characters from text. Use this program when you wish to remove digits and blank spaces from a sequence to make it suitable for other applications.
- **GenBank to FASTA** - accepts one or more GenBank files as input and returns the entire DNA sequence from each in FASTA format. Use this program when you wish to quickly remove all of the non-DNA sequence information from a GenBank file.
- **GenBank Feature Extractor** - accepts one or more GenBank files as input and reads the sequence feature information described in the feature tables, according to the rules outlined in the GenBank release notes. The program extracts or highlights the relevant sequence segments and returns each sequence feature in FASTA format. GenBank Feature Extractor is particularly helpful when you wish to derive the sequence of a cDNA from a genomic sequence that contains many introns.
- **GenBank Trans Extractor** - accepts one or more GenBank files as input and returns each of the protein translations described in the files in FASTA format. GenBank Trans Extractor should be used when you are more interested in the predicted protein translations of a DNA sequence than the DNA sequence itself.
- **One to Three** - converts single letter translations to three letter translations.
- **Range Extractor DNA** - accepts one or more DNA sequences along with a set of positions or ranges. The bases corresponding to the positions or ranges are returned, either as a single new sequence, a set of FASTA records

Protein Sequence Analysis

https://sites.ualberta.ca/~stothard/javascript/index.html

Protein Sequence Analysis



Protein Sequence Analysis

SMS

Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats
-Restriction Digest
-Restriction Summary
-Reverse Translate
-Translate

Sequence Figures
-Color Align Conservation
-Color Align Properties
-Group DNA
-Group Protein
-Primer Map
-Restriction Map
-Translation Map

Random Sequences
-Mutate DNA
-Mutate Protein
-Random Coding DNA
-Random DNA Sequence
-Random DNA Regions
-Random Protein Sequence
-Random Protein Regions
-Sample DNA
-Sample Protein
-Shuffle DNA
-Shuffle Protein

Miscellaneous

Sequence Manipulation Suite:

Protein Stats

Protein Stats returns the number of occurrences of each residue in the sequence you enter. Percentage totals are also given for each residue, and for certain groups of residues, allowing you to quickly compare the results obtained for different sequences.

Paste the raw sequence or one or more FASTA sequences into the text area below. Input limit is 500000 characters.

```
ESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAA
MYDKGPFRS
KKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADAR
IQILEGWKK
RLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIK
ARK
```

Please check the browser compatibility page before using this program.

Submit Clear Reset

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

Fri Jun 17 16:17:06 2016
Valid XHTML 1.0; Valid CSS

Sequence Manipulation Suite - Internet Expl...

about:blank

Soubor Úpravy Zobrazit Oblíbené položky Nástroje Nápověda

Protein Stats results

Results for 274 residue sequence "AAB60701.1 NAD(P)H:quinone oxireductase [Homo sapiens]" starting "MVGRRALIVL"

Pattern:	Times found:	Percentage:
A	25	9.12
B	0	0.00
C	1	0.36
D	12	4.38
E	17	6.20
F	17	6.20
G	21	7.66
H	7	2.55
I	19	6.93
K	24	8.76
L	25	9.12
M	7	2.55
N	9	3.28
P	13	4.74
Q	12	4.38
R	9	3.28
S	17	6.20
T	9	3.28

96%

Protein Sequence Analysis

SMS

Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats
-Restriction Digest
-Restriction Summary
-Reverse Translate
-Translate

Sequence Figures
-Color Align Conservation
-Color Align Properties
-Group DNA
-Group Protein
-Primer Map
-Restriction Map
-Translation Map

Random Sequences
-Mutate DNA
-Mutate Protein
-Random Coding DNA
-Random DNA Sequence
-Random DNA Regions

Sequence Manipulation Suite:

Protein Molecular Weight

Protein Molecular Weight accepts one or more protein sequences and calculates molecular weight. You can append copies of commonly used epitopes and fusion proteins using the supplied list. Use Protein Molecular Weight when you wish to predict the location of a protein of interest on a gel in relation to a set of protein standards.

Paste the raw sequence or one or more FASTA sequences into the text area below. Input limit is 200000 characters.

```
ESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAA
MYDKGPFRS
KKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADAR
IQILEGWKK
RLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIK
ARK
```

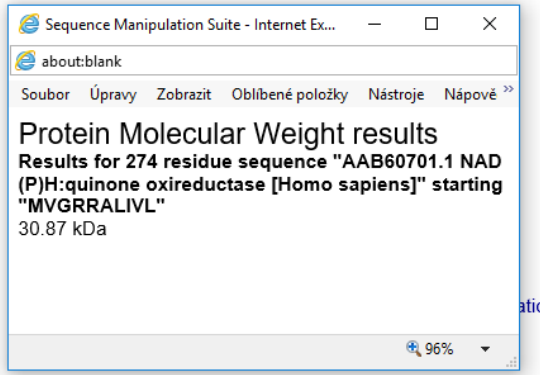
Please check the browser compatibility page before using this program.

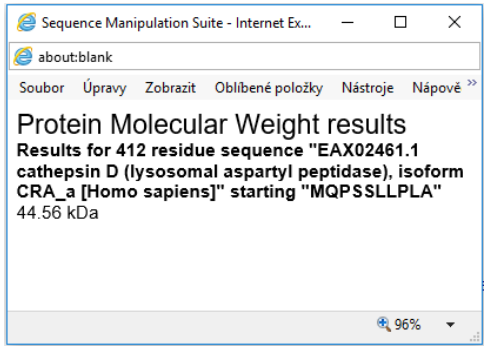
Submit Clear Reset

- Add 1 copies of Nothing to the above sequence.

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

Fri Jun 17 16:17:06 2016
Valid XHTML 1.0; Valid CSS





Protein Sequence Analysis

SMS

Format Conversion
-Combine FASTA
-EMBL to FASTA
-EMBL Feature Extractor
-EMBL Trans Extractor
-Filter DNA
-Filter Protein
-GenBank to FASTA
-GenBank Feature Extractor
-GenBank Trans Extractor
-One to Three
-Range Extractor DNA
-Range Extractor Protein
-Reverse Complement
-Split Codons
-Split FASTA
-Three to One
-Window Extractor DNA
-Window Extractor Protein

Sequence Analysis
-Codon Plot
-Codon Usage
-CpG Islands
-DNA Molecular Weight
-DNA Pattern Find
-DNA Stats
-Fuzzy Search DNA
-Fuzzy Search Protein
-Ident and Sim
-Multi Rev Trans
-Mutate for Digest
-ORF Finder
-Pairwise Align Codons
-Pairwise Align DNA
-Pairwise Align Protein
-PCR Primer Stats
-PCR Products
-Protein GRAVY
-Protein Isoelectric Point
-Protein Molecular Weight
-Protein Pattern Find
-Protein Stats
-Restriction Digest
-Restriction Summary
-Reverse Translate
-Translate

Sequence Figures

Sequence Manipulation Suite:

Range Extractor Protein

Range Extractor Protein accepts a protein sequence along with a set of positions or ranges. The residues corresponding to the positions or ranges are returned, either as a single new sequence, a set of FASTA records, as uppercase text, or as lowercase text. Use Range Extractor Protein to obtain subsequences using position information.

Paste a raw sequence or one or more FASTA sequences into the text area below. Input limit is 500,000,000 characters.

```
>sample sequence
MQKSPLEKASFISKLFFSWTTPILRKGYRHHLELSDIYQAPSADSADHLSEKLEREWDREQ
ASKKNPQLIHALRRCFFWRFLFYGILLYLGEVTKAVQPVLLGRIIASYDPENKVERSIAIY
LGIGLCLLFIVRTLLLHPAIFGLHRIGMQMRTAMFSLIYKKTLKLSSRVLDKISIGQLVSL
LSNNLNKFDEGLALAHFIWIAPLQVTLLMGLLWDLLQFSAFCGLGLLIILVIFQAILGKMM
VKYRDQRAAKINERLVITSEIIDNIYSVKAYCWESAMEKMIENLREVELKMTRKAAYMRFF
```

Enter the residue positions or ranges to be extracted. Use ".." to represent a range, and use a comma to separate entries. The words 'start', 'end', 'center', and 'length' can be used in place of digits, to represent the beginning, end, middle, and length of the sequence. Arithmetic expressions can be included in the ranges. For example, to obtain the last three residues of a sequence, the range '(end - 2)..end' can be used. To obtain the 30 bases on either side of the center residue along with the center residue, the ranges '(center - 30)..(center - 1), center, (center + 1)..(center + 30)' can be used.

1, 5, 10..12

Submit | Clear | Reset

- Sequence segments should be returned as a new sequence

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

new window | home | citation

Mon Nov 6 02:56:29 2017
Valid XHTML 1.0; Valid CSS

Try SMS suite

Analyze your sequence
- how many cysteins are in your sequence?

HW2

3) Extract peptide from 10. - 50. amino acid.
4) Count the molecular weight of this (short) peptide.
5) How many cysteins are in your (whole) sequence?

„Protein bioinformatics II“

Retrieving protein sequences from databases

Computing amino-acids compositions, molecular weight, isoelectric point, and other parameters

Prediction of proteases cutting

Predicting elements of protein secondary structure, domains

Predicting 3-D structure and the domain organization of proteins

Finding all proteins that share a similar sequence and Classifying proteins into families

Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…

Homework 2

1) find both (Uniprot/NCBI) accession numbers for reference sequence (isoform 1)

2) download your sequence in FASTA format

3) Extract peptide form positions 10 to 50

4) Predict molecular weight of this peptide

5) How many cysteins are in your sequence?

E.g use „výstřižky“ „snipping tool“

- Compile in „one note“ (or word, or pdf)

Homework 2: examples

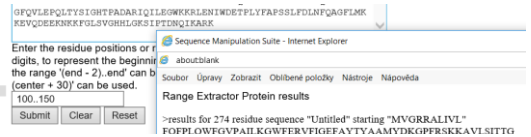
DÚ2

NAD(P)H:quinone oxireductase [Homo sapiens]

GenBank: AAB60701.1

UniProtKB - P15559 (NQO1_HUMAN)

>AAB60701.1 NAD(P)H:quinone oxireductase [Homo sapiens]
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPA
ESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRS
KKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKK
RLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK



Protein Molecular Weight results
Results for 51 residue sequence "Untitled" starting "FQFPLQWFGV"
5.88 kDa

Protein Stats results
Results for 274 residue sequence "Untitled" starting "MVGRRALIVL"

Pattern:	Times found:	Percentage:
A	25	9.12
B	0	0.00
C	1	0.36

NQO1:	AAB60701
	P15559
	>AAB60701.1 NAD(P)H:quinone oxireductase [Homo sapiens] MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPA ESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRS KKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKK RLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
	FQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTG 5.88 kDa
	Cystein:1
	Your collection was saved. Edit your collection.

NAD(P)H:quinone oxireductase [Homo sapiens]

GenBank: AAB60701.1

Identical Proteins FASTA Graphics