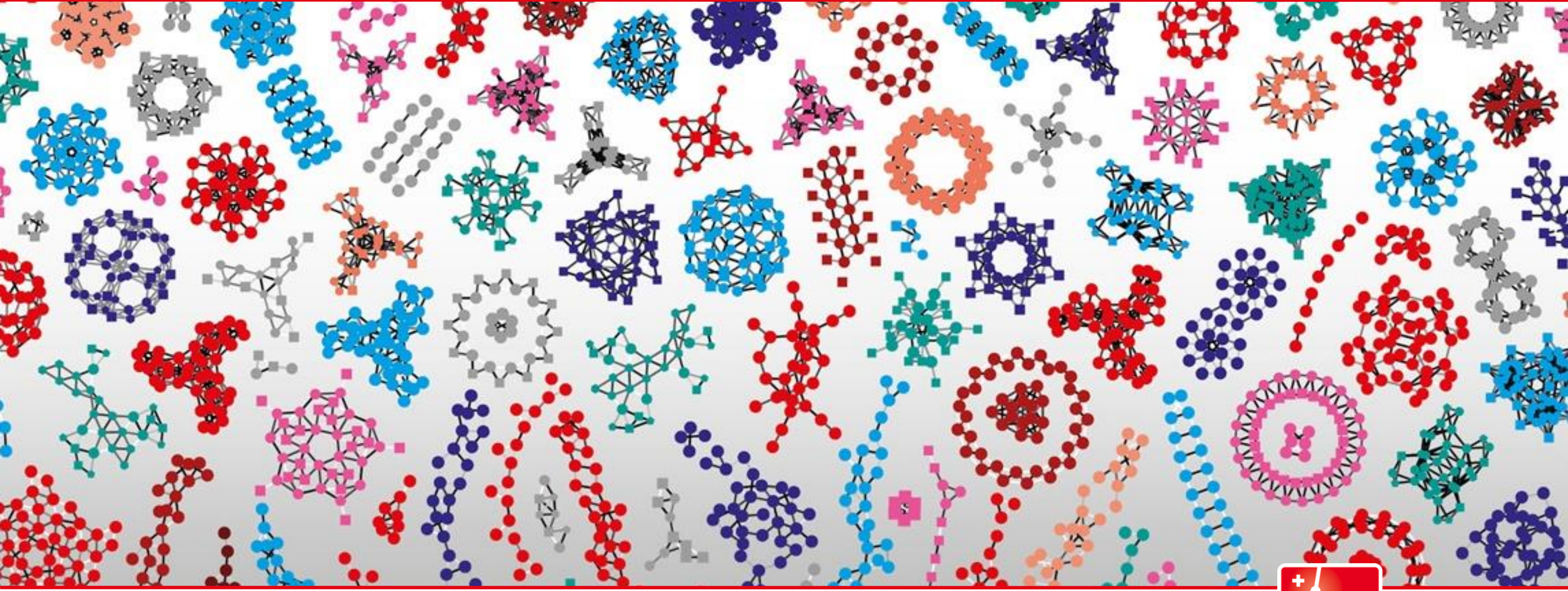
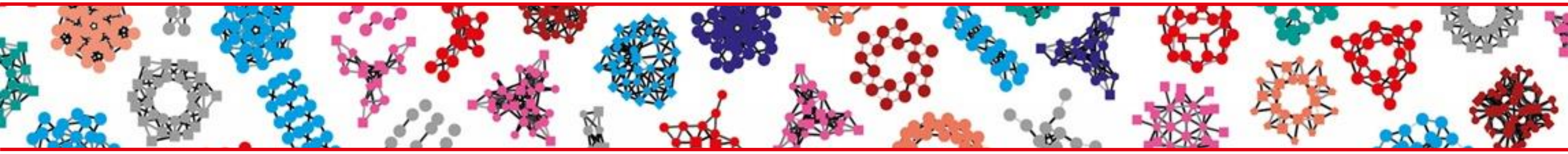


Swiss Institute of Bioinformatics

SPARQLing neXtProt data

SWAT4HCLS, Edinburgh, United Kingdom
Monique Zahn

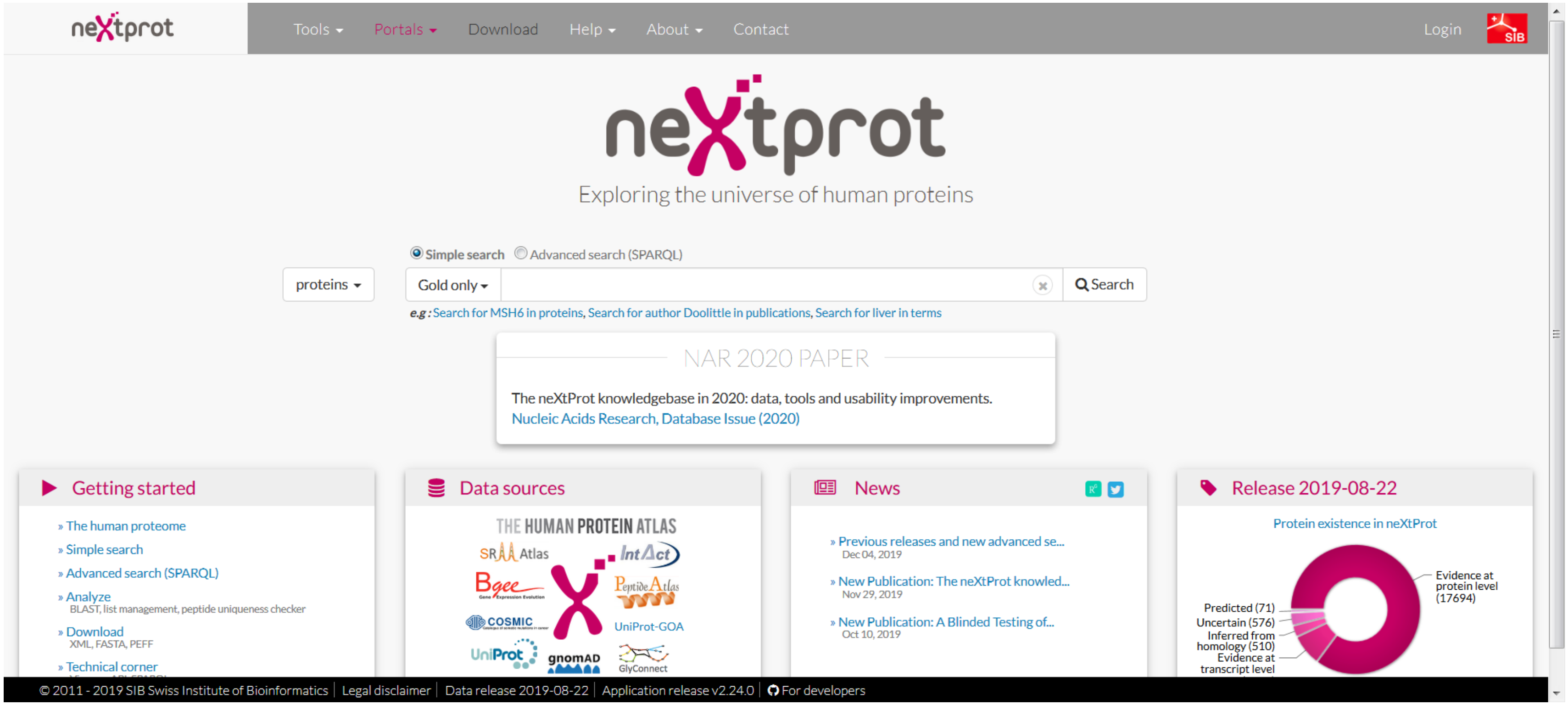
Overview



- 01 **Introduction to neXtProt**
- 02 Data model
- 03 SPARQLing in neXtProt
- 04 Summary

neXtProt – the SIB knowledgebase on human proteins

https://www.nextprot.org/



Exploring the universe of human proteins

e.g : Search for MSH6 in proteins, Search for author Doolittle in publications, Search for liver in terms

NAR 2020 PAPER

The neXtProt knowledgebase in 2020: data, tools and usability improvements.
Nucleic Acids Research, Database Issue (2020)

Getting started

- » The human proteome
- » Simple search
- » Advanced search (SPARQL)
- » Analyze
  BLAST, list management, peptide uniqueness checker
- » Download
  XML, FASTA, PEFF
- » Technical corner

Data sources

News

- » Previous releases and new advanced se...
  Dec 04, 2019
- » New Publication: The neXtProt knowled...
  Nov 29, 2019
- » New Publication: A Blinded Testing of...
  Oct 10, 2019

Release 2019-08-22

Protein existence in neXtProt

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.24.0 | For developers

Data sources

THE HUMAN PROTEIN ATLAS

SRM Atlas

IntAct

Bgee
Gene Expression Evolution

PeptideAtlas

COSMIC
Catalogue of somatic mutations in cancer

UniProt-GOA

UniProt

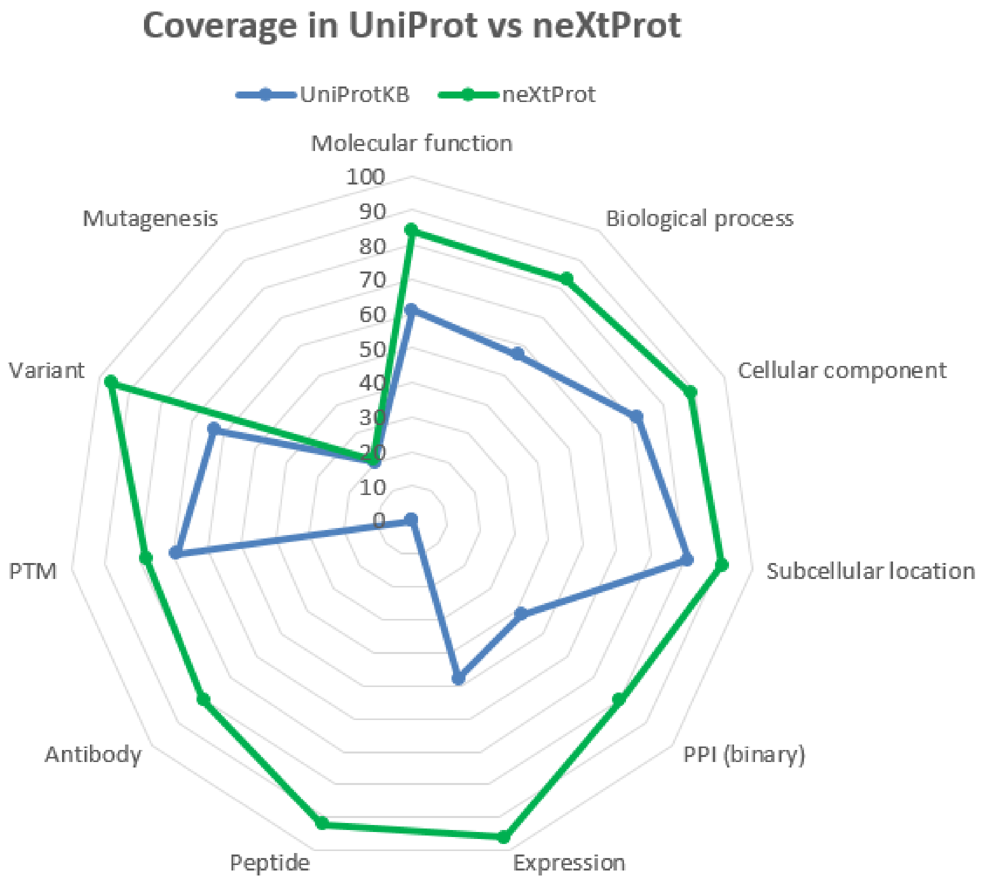
gnomAD

GlyConnect

COMING SOON : MassIVE
Mass Spectrometry
Interactive Virtual Environment

“We stand on the shoulders of giants.”

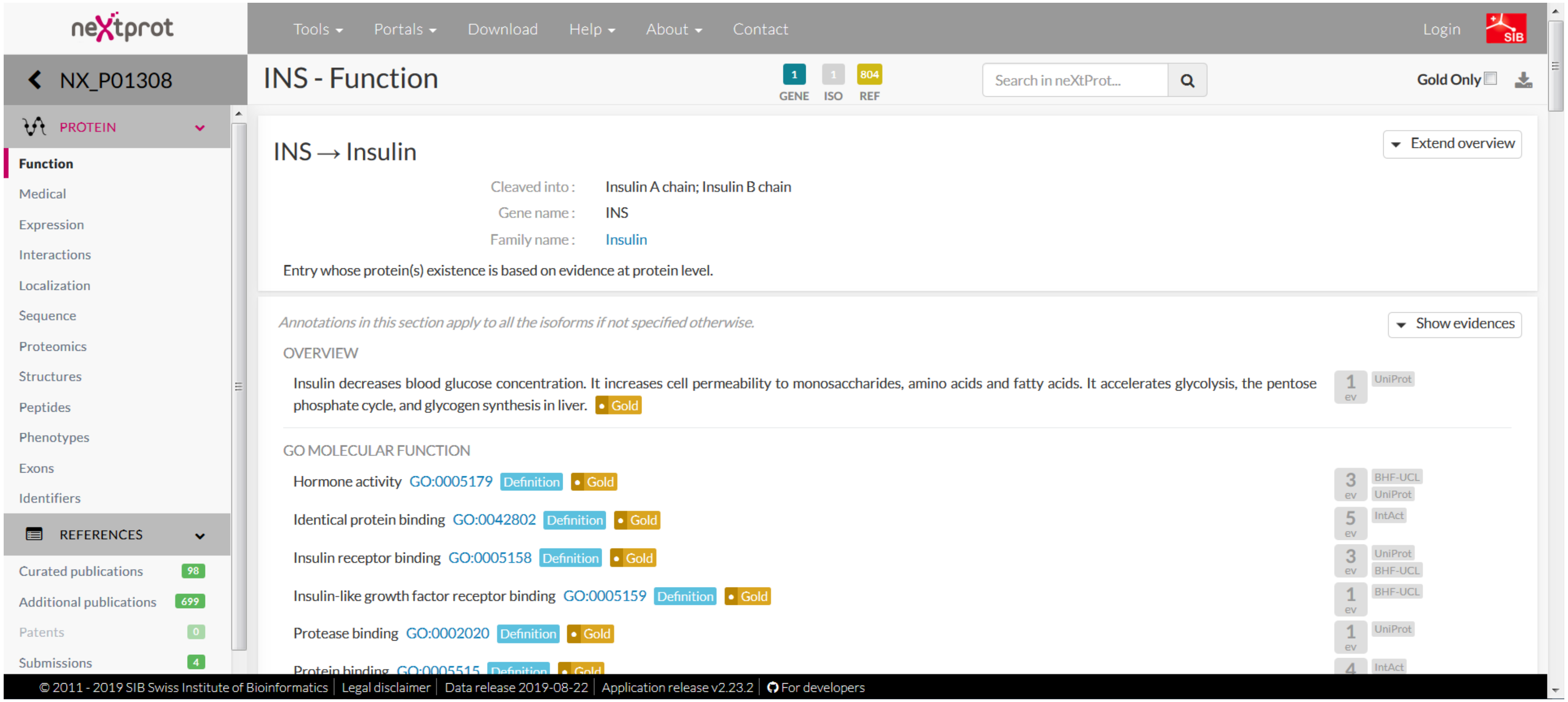
Improved coverage through data integration



Data on human proteins

https://www.nextprot.org/entry/NX_P01308/

Views



INS - Function

INS → Insulin

Cleaved into : Insulin A chain; Insulin B chain
Gene name : INS
Family name : Insulin

Entry whose protein(s) existence is based on evidence at protein level.

Annotations in this section apply to all the isoforms if not specified otherwise.

OVERVIEW

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver. • Gold — 1 ev — UniProt

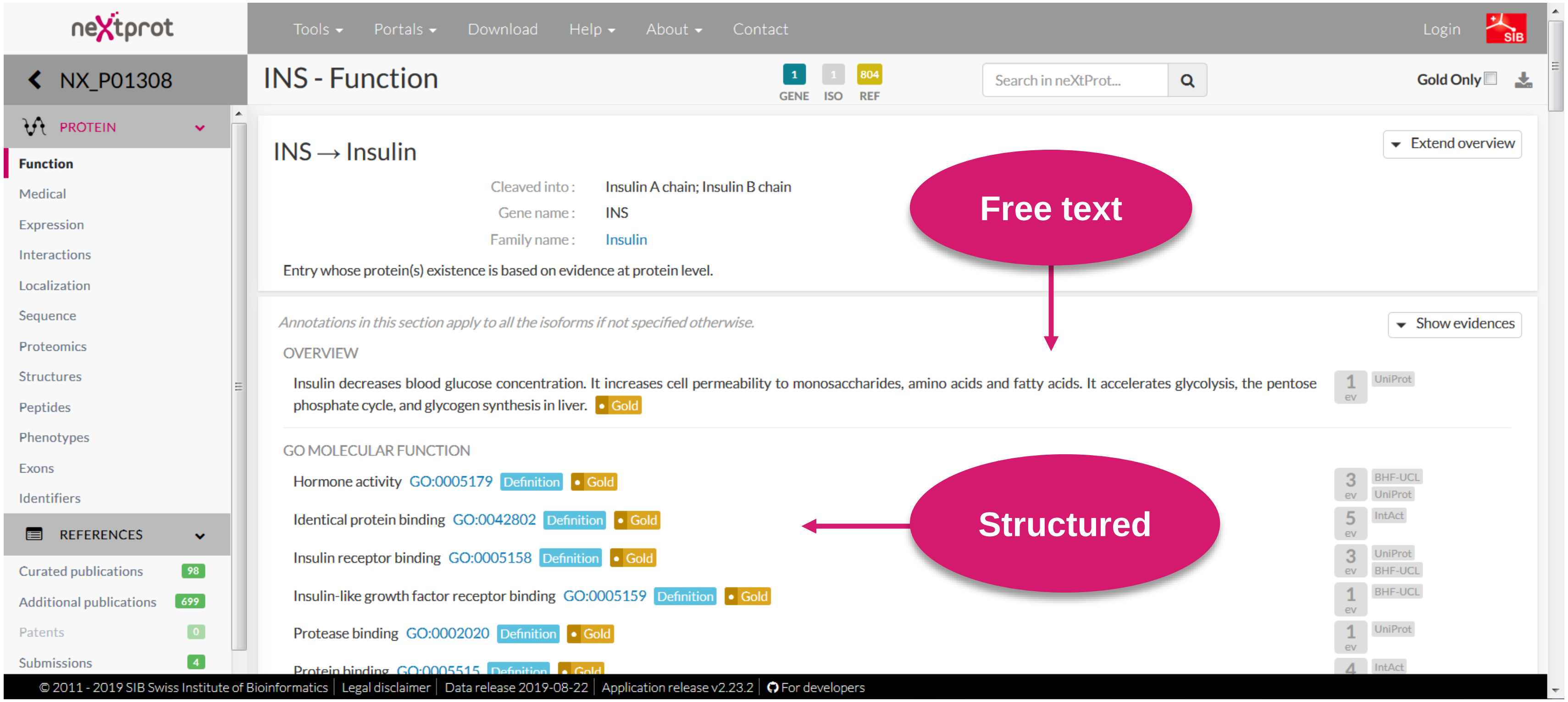
GO MOLECULAR FUNCTION

Hormone activity	GO:0005179	Definition	• Gold	3 ev	BHF-UCL UniProt
Identical protein binding	GO:0042802	Definition	• Gold	5 ev	IntAct
Insulin receptor binding	GO:0005158	Definition	• Gold	3 ev	UniProt BHF-UCL
Insulin-like growth factor receptor binding	GO:0005159	Definition	• Gold	1 ev	BHF-UCL
Protease binding	GO:0002020	Definition	• Gold	1 ev	UniProt
Protein binding	GO:0005515	Definition	• Gold	4	IntAct

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.23.2 | For developers

Types of data (1)

https://www.nextprot.org/entry/NX_P01308/



INS - Function

INS → Insulin

Cleaved into : Insulin A chain; Insulin B chain
Gene name : INS
Family name : Insulin

Entry whose protein(s) existence is based on evidence at protein level.

Annotations in this section apply to all the isoforms if not specified otherwise.

OVERVIEW

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver. • Gold — 1 ev UniProt

GO MOLECULAR FUNCTION

Function	GO term			Evidence	Source
Hormone activity	GO:0005179	Definition	• Gold	3 ev	BHF-UCL UniProt
Identical protein binding	GO:0042802	Definition	• Gold	5 ev	IntAct
Insulin receptor binding	GO:0005158	Definition	• Gold	3 ev	UniProt BHF-UCL
Insulin-like growth factor receptor binding	GO:0005159	Definition	• Gold	1 ev	BHF-UCL
Protease binding	GO:0002020	Definition	• Gold	1 ev	UniProt
Protein binding	GO:0005515	Definition	• Gold	4	IntAct

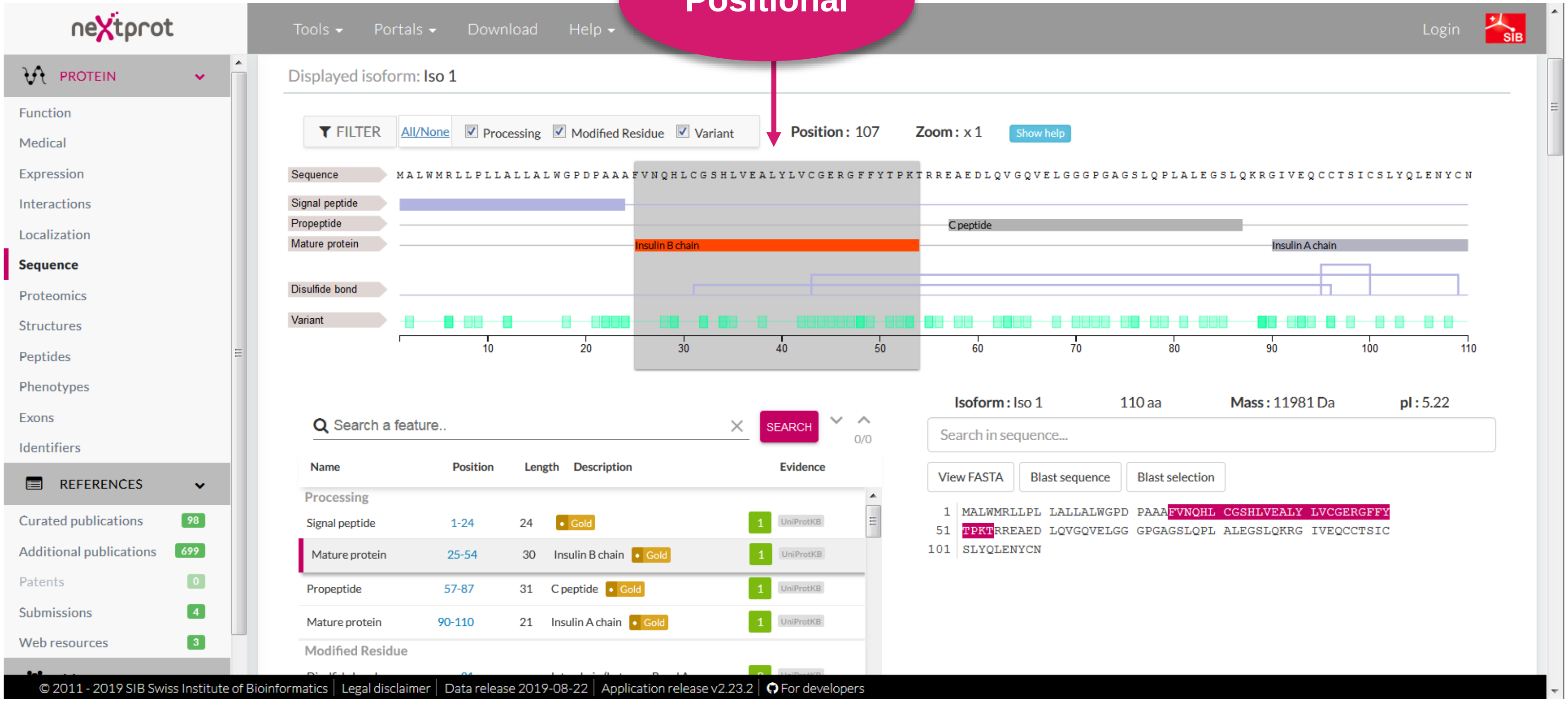
© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.23.2 | For developers

Free text

Structured

Types of data (2)

https://www.nextprot.org/entry/NX_P01308/sequence



Positional

- PROTEIN
  - Function
  - Medical
  - Expression
  - Interactions
  - Localization
  - **Sequence**
  - Proteomics
  - Structures
  - Peptides
  - Phenotypes
  - Exons
  - Identifiers
- REFERENCES
  - Curated publications 98
  - Additional publications 699
  - Patents 0
  - Submissions 4
  - Web resources 3

Displayed isoform: Iso 1

FILTER All/None ☑ Processing ☑ Modified Residue ☑ Variant

Position : 107 **Zoom :** x 1 Show help

Sequence: M A L W M R L L P L L A L L A L W G P D P A A A F V N Q H L C G S H L V E A L Y L V C G E R G F F Y T P K T R R E A E D L Q V G Q V E L G G G P G A G S L Q P L A L E G S L Q K R G I V E Q C C T S I C S L Y Q L E N Y C N

Search a feature.. SEARCH 0/0

Name	Position	Length	Description	Evidence
Processing				
Signal peptide	1-24	24	Gold	1 UniProtKB
Mature protein	25-54	30	Insulin B chain Gold	1 UniProtKB
Propeptide	57-87	31	C peptide Gold	1 UniProtKB
Mature protein	90-110	21	Insulin A chain Gold	1 UniProtKB
Modified Residue				

Isoform : Iso 1 110 aa **Mass :** 11981 Da **pI :** 5.22

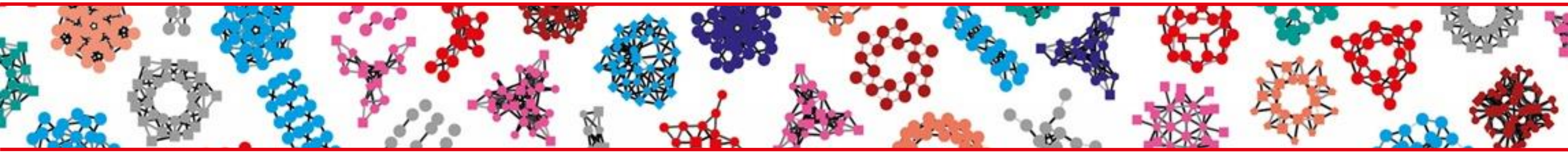
Search in sequence...

View FASTA | Blast sequence | Blast selection

```
  1 MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY
 51 TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC
101 SLYQLENYCN
```

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.23.2 | For developers

Overview



01 Introduction to neXtProt

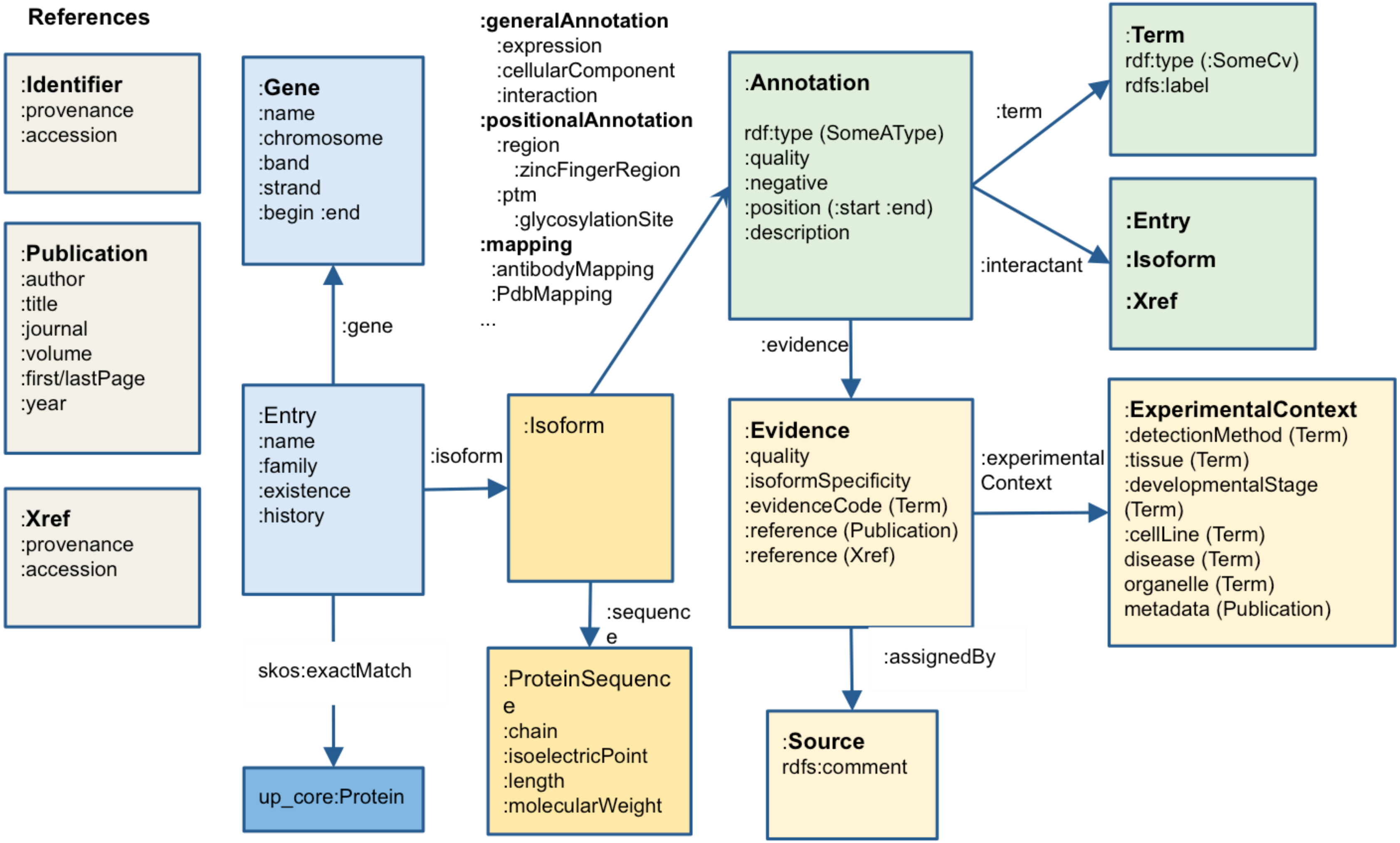
02 **Data model**

03 SPARQLing in neXtProt

04 Summary

Data model

https://www.nextprot.org/help/data-model



References

:Identifier
:provenance
:accession

:Publication
:author
:title
:journal
:volume
:first/lastPage
:year

:Xref
:provenance
:accession

:Gene
:name
:chromosome
:band
:strand
:begin :end

:gene

:Entry
:name
:family
:existence
:history

skos:exactMatch

up_core:Protein

:isoform

:Isoform

:sequence

:ProteinSequence
:chain
:isoelectricPoint
:length
:molecularWeight

:generalAnnotation
:expression
:cellularComponent
:interaction
:positionalAnnotation
:region
:zincFingerRegion
:ptm
:glycosylationSite
:mapping
:antibodyMapping
:PdbMapping
...

:Annotation

rdf:type (SomeAType)
:quality
:negative
:position (:start :end)
:description

:term

:interactant

:Term
rdf:type (:SomeCv)
rdfs:label

:Entry

:Isoform

:Xref

:evidence

:Evidence
:quality
:isoformSpecificity
:evidenceCode (Term)
:reference (Publication)
:reference (Xref)

:experimental Context

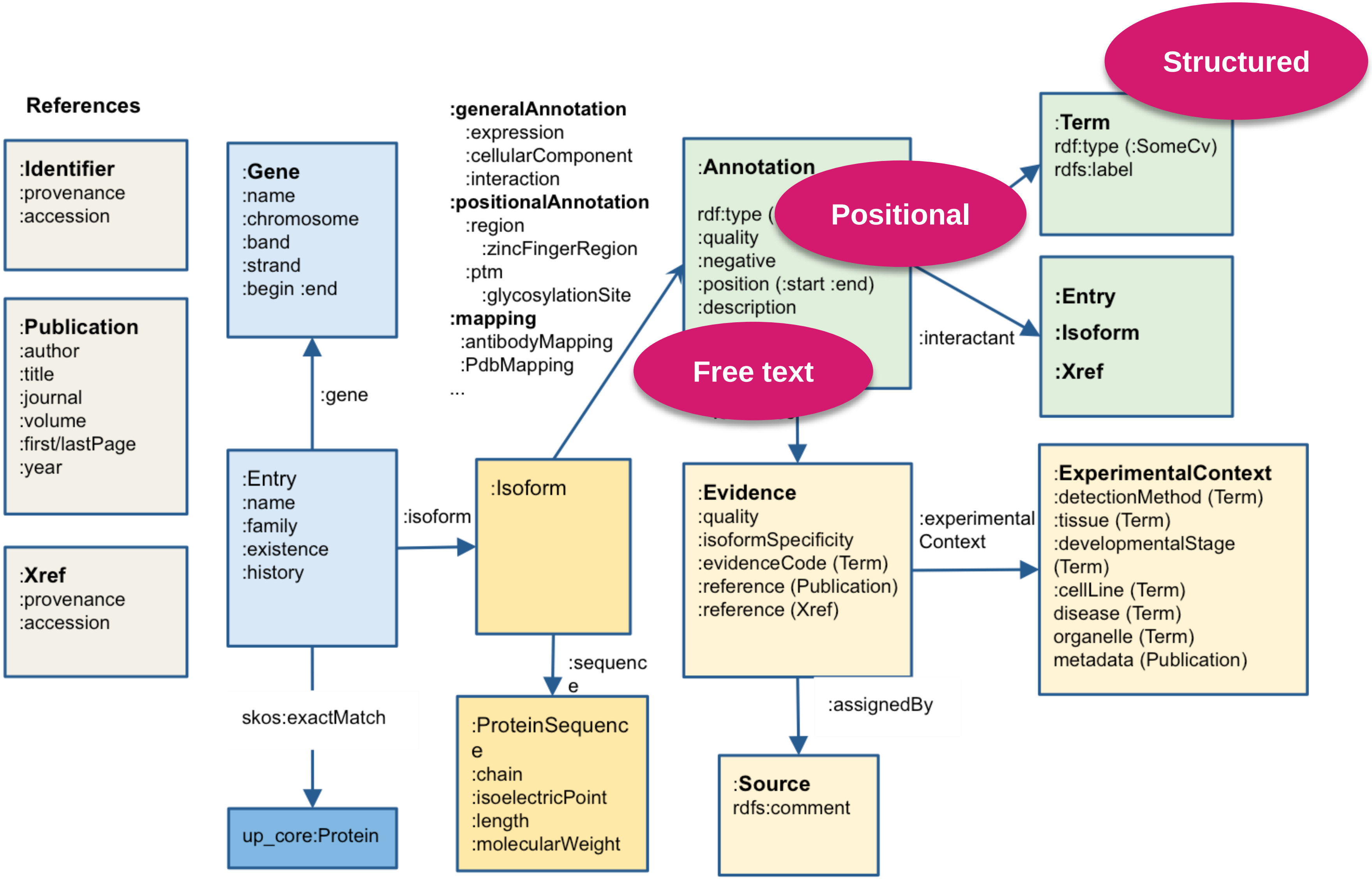
:ExperimentalContext
:detectionMethod (Term)
:tissue (Term)
:developmentalStage (Term)
:cellLine (Term)
disease (Term)
organelle (Term)
metadata (Publication)

:assignedBy

:Source
rdfs:comment

Single interoperable model (RDF)

Types of data



References

:**Identifier**
:provenance
:accession

:**Publication**
:author
:title
:journal
:volume
:first/lastPage
:year

:**Xref**
:provenance
:accession

:**Gene**
:name
:chromosome
:band
:strand
:begin :end

:gene

:Entry
:name
:family
:existence
:history

skos:exactMatch

up_core:Protein

:generalAnnotation
:expression
:cellularComponent
:interaction
:positionalAnnotation
:region
:zincFingerRegion
:ptm
:glycosylationSite
:mapping
:antibodyMapping
:PdbMapping
...

:isoform

:Isoform

:sequence

:ProteinSequence
:chain
:isoelectricPoint
:length
:molecularWeight

:**Annotation**
rdf:type (
:quality
:negative
:position (:start :end)
:description

Positional

Free text

Structured

:**Term**
rdf:type (:SomeCv)
rdfs:label

:interactant

:Entry
:Isoform
:Xref

:**Evidence**
:quality
:isoformSpecificity
:evidenceCode (Term)
:reference (Publication)
:reference (Xref)

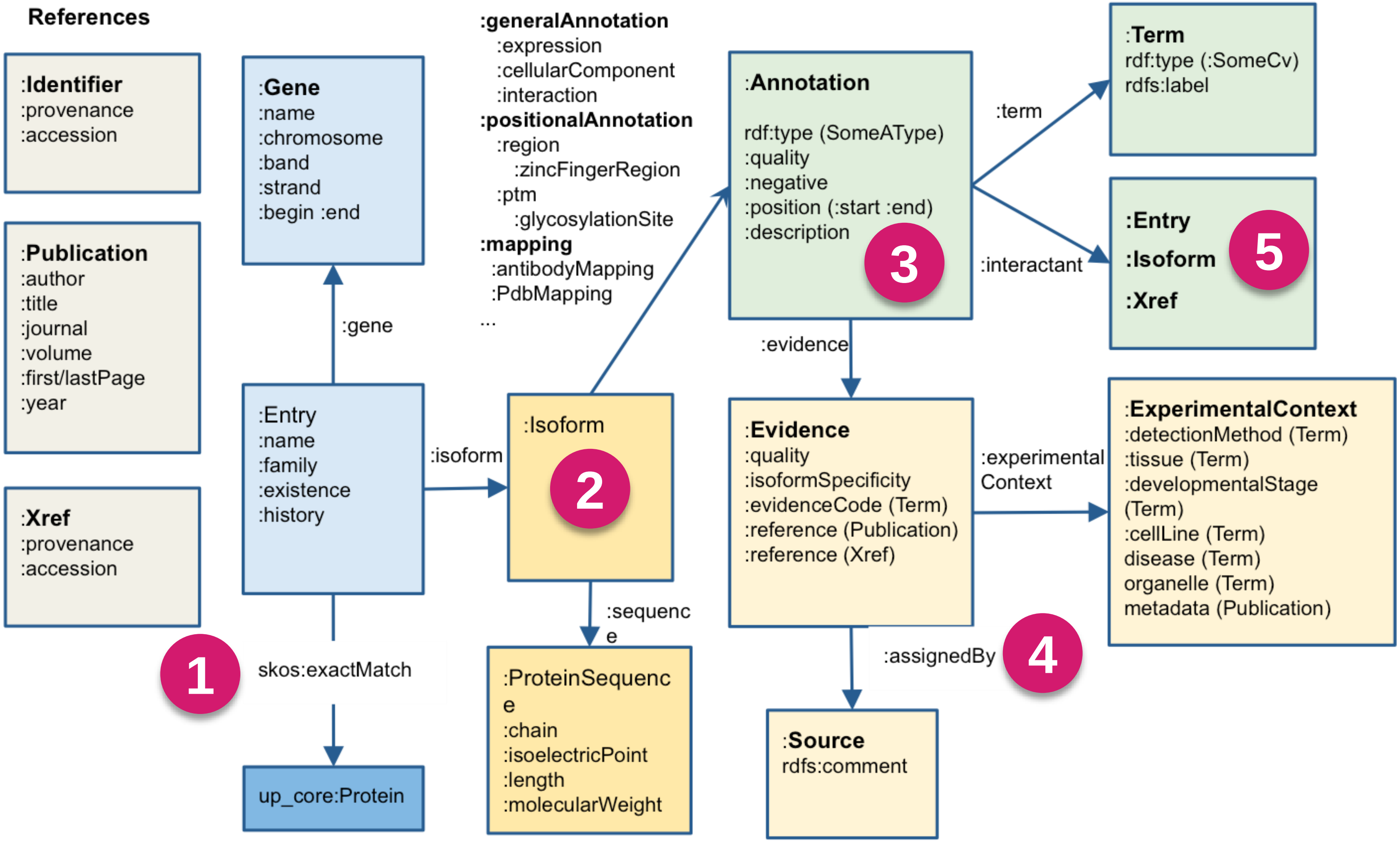
:experimental Context

:**ExperimentalContext**
:detectionMethod (Term)
:tissue (Term)
:developmentalStage (Term)
:cellLine (Term)
disease (Term)
organelle (Term)
metadata (Publication)

:assignedBy

:**Source**
rdfs:comment

Distinguishing features



References

:Identifier
:provenance
:accession

:Publication
:author
:title
:journal
:volume
:first/lastPage
:year

:Xref
:provenance
:accession

:Gene
:name
:chromosome
:band
:strand
:begin :end

:gene

:Entry
:name
:family
:existence
:history

skos:exactMatch

up_core:Protein

:isoform

:Isoform

:sequence

:ProteinSequence
:chain
:isoelectricPoint
:length
:molecularWeight

:generalAnnotation
:expression
:cellularComponent
:interaction
:positionalAnnotation
:region
:zincFingerRegion
:ptm
:glycosylationSite
:mapping
:antibodyMapping
:PdbMapping
...

:Annotation
rdf:type (SomeAType)
:quality
:negative
:position (:start :end)
:description

:term

:interactant

:Term
rdf:type (:SomeCv)
rdfs:label

:Entry
:Isoform
:Xref

:evidence

:Evidence
:quality
:isoformSpecificity
:evidenceCode (Term)
:reference (Publication)
:reference (Xref)

:experimentalContext

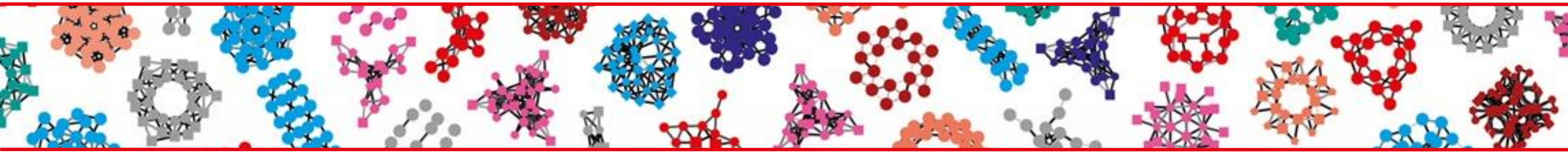
:ExperimentalContext
:detectionMethod (Term)
:tissue (Term)
:developmentalStage (Term)
:cellLine (Term)
disease (Term)
organelle (Term)
metadata (Publication)

:assignedBy

:Source
rdfs:comment

1
2
3
4
5

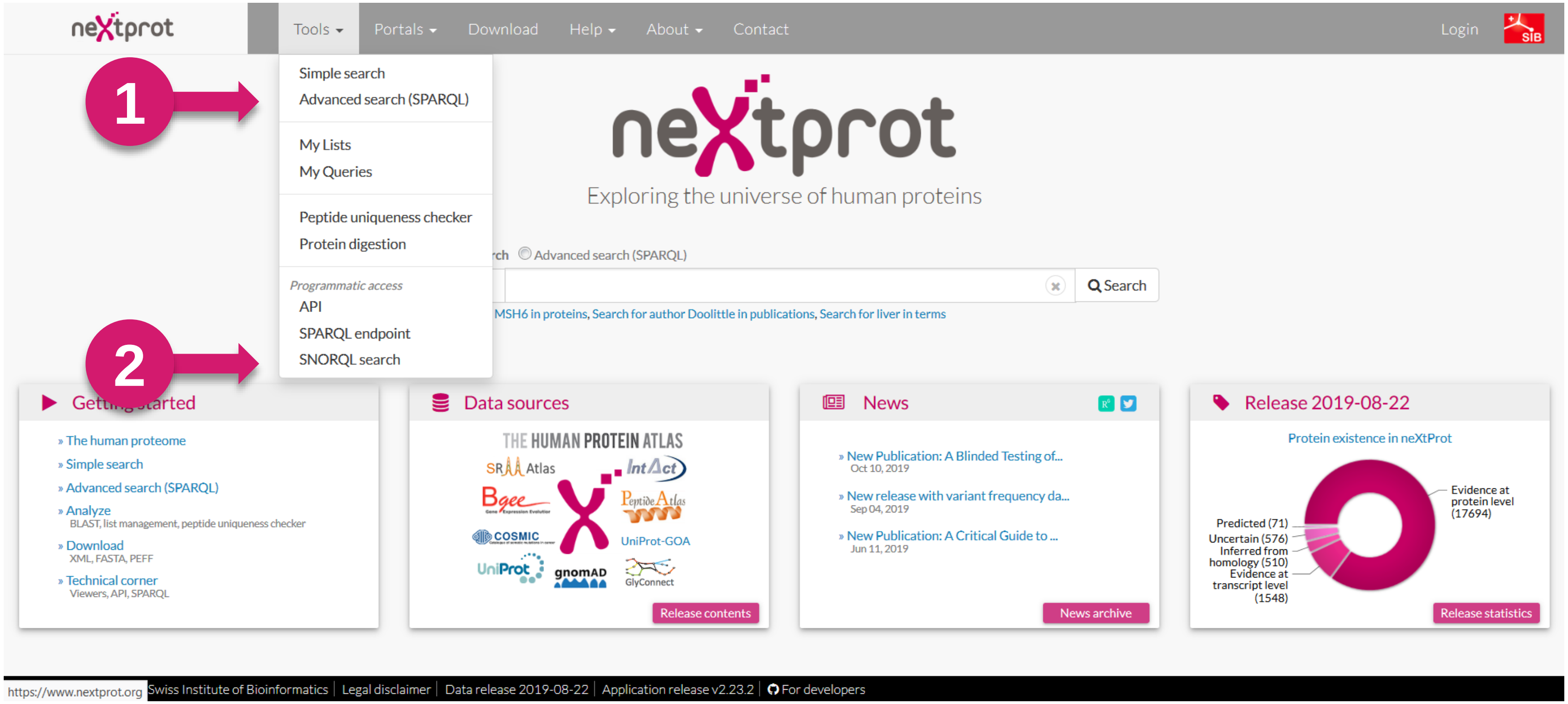
Overview



- **01** Introduction to neXtProt
- **02** Data model
- **03** SPARQLing in neXtProt
- **04** Summary

SPARQL user interfaces

https://www.nextprot.org/



3 SPARQL endpoint : https://api.nextprot.org/sparql
(SERVICE https://sparql.nextprot.org)

neXtProt Advanced Search

Find answers to complex questions:

1. Proteins whose genes are on chromosome 13 and are associated with a disease

2. Proteins with at least one variant of the types "A->R" or "R->A“

3. Proteins with alternative acetylation or Ubl conjugation (SUMO or Ubiquitin) at the same positions

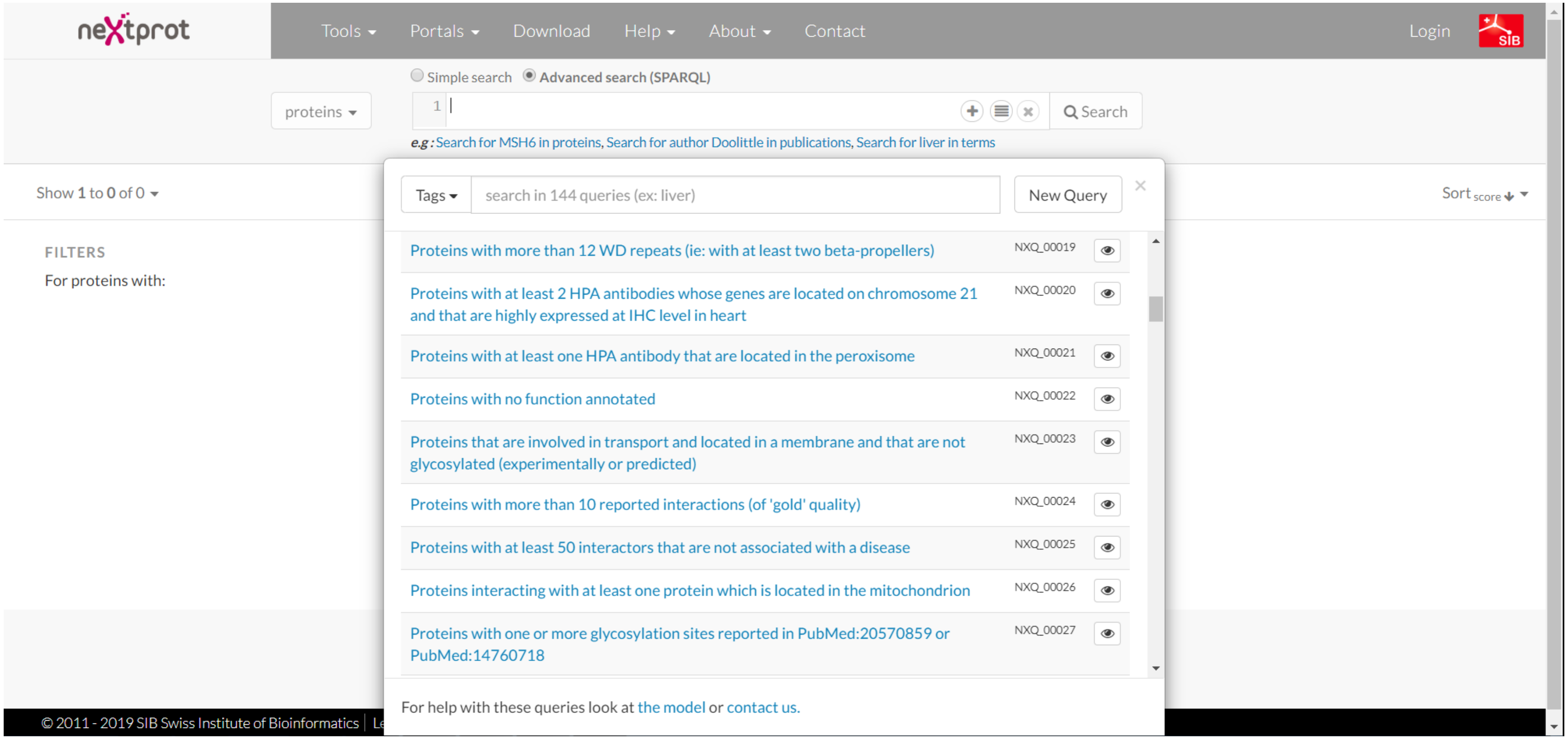
4. Proteins with at least two antibodies available from Human Protein Atlas that have associated tissue expression annotations from immunohistochemistry studies

These queries cannot be performed using the simple search.

neXtProt Advanced Search

https://www.nextprot.org/proteins/search?mode=advanced

Retrieve **entries** matching criteria



e.g : Search for MSH6 in proteins, Search for author Doolittle in publications, Search for liver in terms

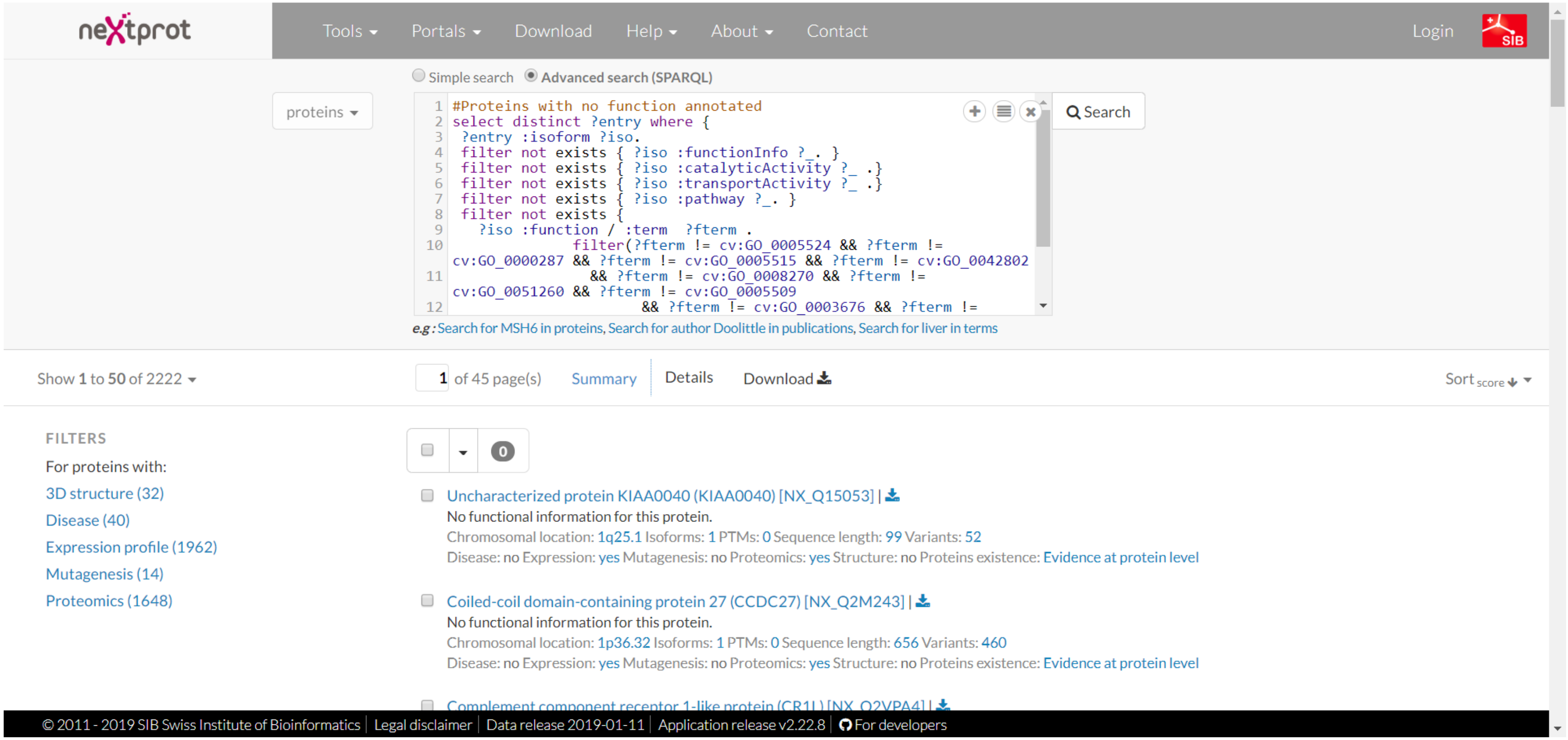
Proteins with more than 12 WD repeats (ie: with at least two beta-propellers)	NXQ_00019
Proteins with at least 2 HPA antibodies whose genes are located on chromosome 21 and that are highly expressed at IHC level in heart	NXQ_00020
Proteins with at least one HPA antibody that are located in the peroxisome	NXQ_00021
Proteins with no function annotated	NXQ_00022
Proteins that are involved in transport and located in a membrane and that are not glycosylated (experimentally or predicted)	NXQ_00023
Proteins with more than 10 reported interactions (of 'gold' quality)	NXQ_00024
Proteins with at least 50 interactors that are not associated with a disease	NXQ_00025
Proteins interacting with at least one protein which is located in the mitochondrion	NXQ_00026
Proteins with one or more glycosylation sites reported in PubMed:20570859 or PubMed:14760718	NXQ_00027

For help with these queries look at the model or contact us.

Advanced Search results

https://www.nextprot.org/proteins/search?mode=advanced&queryId=NXQ_00022

NXQ_00022 **Proteins** with no function annotated



○ Simple search ● Advanced search (SPARQL)

```
#Proteins with no function annotated
select distinct ?entry where {
 ?entry :isoform ?iso.
 filter not exists { ?iso :functionInfo ?_. }
 filter not exists { ?iso :catalyticActivity ?_ .}
 filter not exists { ?iso :transportActivity ?_ .}
 filter not exists { ?iso :pathway ?_. }
 filter not exists {
   ?iso :function / :term  ?fterm .
              filter(?fterm != cv:GO_0005524 && ?fterm !=
cv:GO_0000287 && ?fterm != cv:GO_0005515 && ?fterm != cv:GO_0042802
                && ?fterm != cv:GO_0008270 && ?fterm !=
cv:GO_0051260 && ?fterm != cv:GO_0005509
                      && ?fterm != cv:GO_0003676 && ?fterm !=
```

e.g : Search for MSH6 in proteins, Search for author Doolittle in publications, Search for liver in terms

Show **1** to **50** of 2222

1 of 45 page(s)

FILTERS

For proteins with:

- 3D structure (32)
- Disease (40)
- Expression profile (1962)
- Mutagenesis (14)
- Proteomics (1648)

☐ 0

☐ Uncharacterized protein KIAA0040 (KIAA0040) [NX_Q15053] |

No functional information for this protein.

Chromosomal location: 1q25.1 Isoforms: 1 PTMs: 0 Sequence length: 99 Variants: 52

Disease: no Expression: yes Mutagenesis: no Proteomics: yes Structure: no Proteins existence: Evidence at protein level

☐ Coiled-coil domain-containing protein 27 (CCDC27) [NX_Q2M243] |

No functional information for this protein.

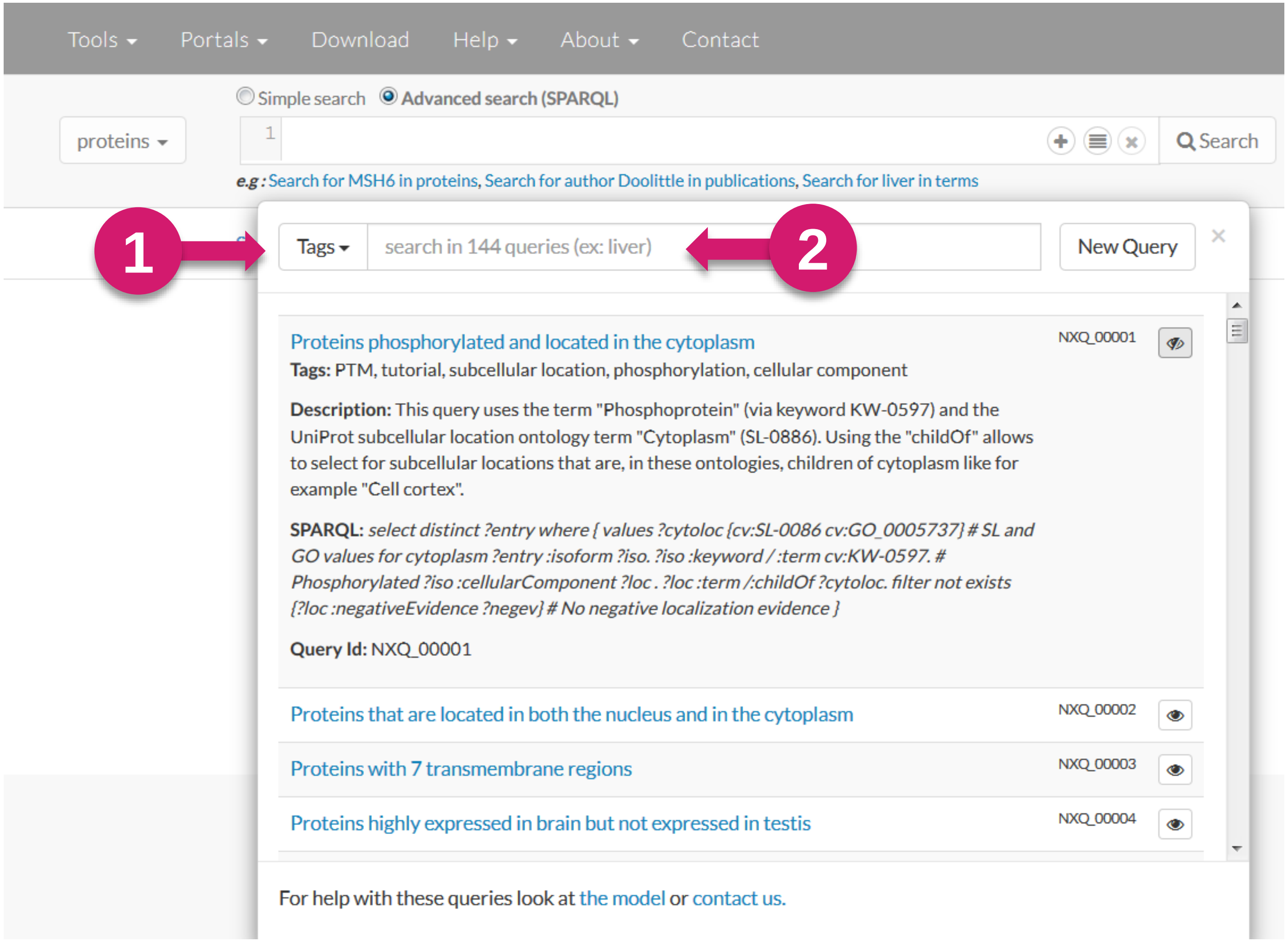
Chromosomal location: 1p36.32 Isoforms: 1 PTMs: 0 Sequence length: 656 Variants: 460

Disease: no Expression: yes Mutagenesis: no Proteomics: yes Structure: no Proteins existence: Evidence at protein level

☐ Complement component receptor 1-like protein (CR1L) [NX_Q2VPA4] |

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-01-11 | Application release v2.22.8 | For developers

SPARQL query examples



e.g : Search for MSH6 in proteins, Search for author Doolittle in publications, Search for liver in terms

Proteins phosphorylated and located in the cytoplasm

Tags: PTM, tutorial, subcellular location, phosphorylation, cellular component

Description: This query uses the term "Phosphoprotein" (via keyword KW-0597) and the UniProt subcellular location ontology term "Cytoplasm" (SL-0886). Using the "childOf" allows to select for subcellular locations that are, in these ontologies, children of cytoplasm like for example "Cell cortex".

SPARQL: *select distinct ?entry where { values ?cytoloc {cv:SL-0086 cv:GO_0005737} # SL and GO values for cytoplasm ?entry :isoform ?iso. ?iso :keyword / :term cv:KW-0597. # Phosphorylated ?iso :cellularComponent ?loc . ?loc :term /:childOf ?cytoloc. filter not exists {?loc :negativeEvidence ?negev} # No negative localization evidence }*

Query Id: NXQ_00001

Proteins that are located in both the nucleus and in the cytoplasm

Proteins with 7 transmembrane regions

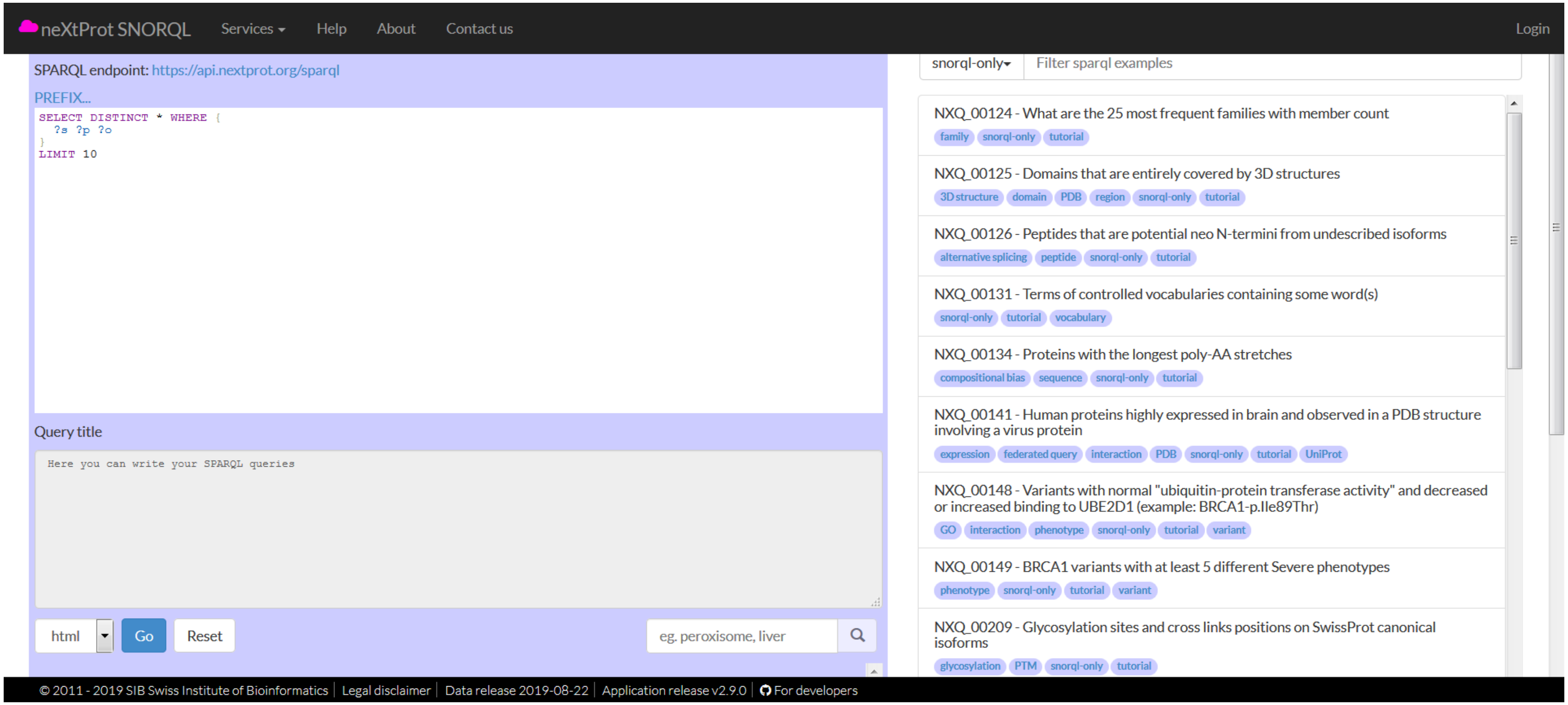
Proteins highly expressed in brain but not expressed in testis

For help with these queries look at the model or contact us.

neXtProt SNORQL Search

https://snorql.nextprot.org/

Queries to retrieve **any information** matching criteria



SPARQL endpoint: https://api.nextprot.org/sparql

PREFIX...

```
SELECT DISTINCT * WHERE {
  ?s ?p ?o
}
LIMIT 10
```

Query title

Here you can write your SPARQL queries

NXQ_00124 - What are the 25 most frequent families with member count
family · snorql-only · tutorial

NXQ_00125 - Domains that are entirely covered by 3D structures
3D structure · domain · PDB · region · snorql-only · tutorial

NXQ_00126 - Peptides that are potential neo N-termini from undescribed isoforms
alternative splicing · peptide · snorql-only · tutorial

NXQ_00131 - Terms of controlled vocabularies containing some word(s)
snorql-only · tutorial · vocabulary

NXQ_00134 - Proteins with the longest poly-AA stretches
compositional bias · sequence · snorql-only · tutorial

NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein
expression · federated query · interaction · PDB · snorql-only · tutorial · UniProt

NXQ_00148 - Variants with normal "ubiquitin-protein transferase activity" and decreased or increased binding to UBE2D1 (example: BRCA1-p.Ile89Thr)
GO · interaction · phenotype · snorql-only · tutorial · variant

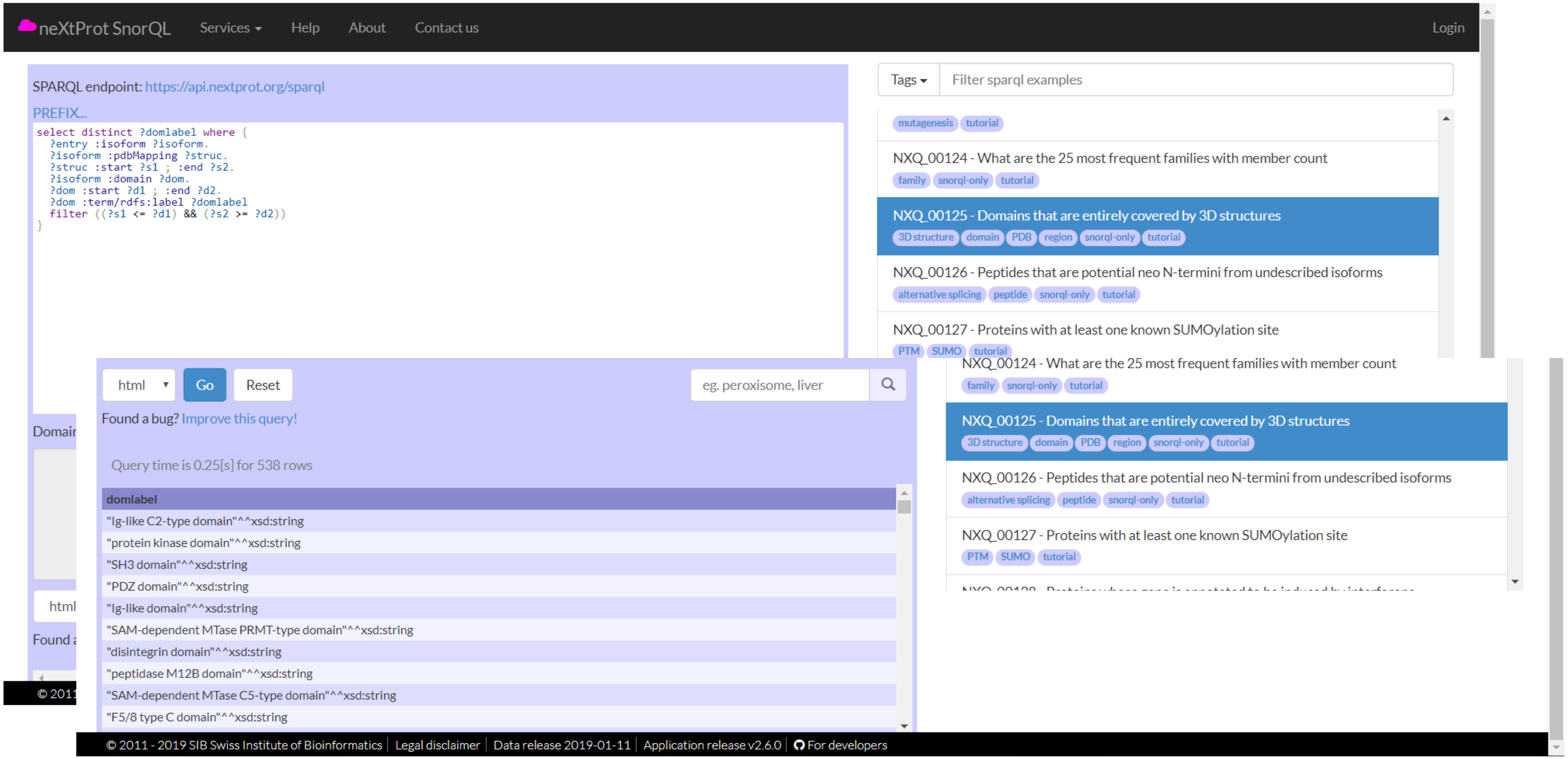
NXQ_00149 - BRCA1 variants with at least 5 different Severe phenotypes
phenotype · snorql-only · tutorial · variant

NXQ_00209 - Glycosylation sites and cross links positions on SwissProt canonical isoforms
glycosylation · PTM · snorql-only · tutorial

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.9.0 | For developers

neXtProt SNORQL Search

NXQ_00125 **Domains** that are entirely covered by 3D structures



SPARQL endpoint: https://api.nextprot.org/sparql

PREFIX...

```
select distinct ?domlabel where {
  ?entry :isoform ?isoform.
  ?isoform :pdbMapping ?struc.
  ?struc :start ?s1 ; :end ?s2.
  ?isoform :domain ?dom.
  ?dom :start ?d1 ; :end ?d2.
  ?dom :term/rdfs:label ?domlabel
  filter ((?s1 <= ?d1) && (?s2 >= ?d2))
}
```

Found a bug? Improve this query!

Query time is 0.25[s] for 538 rows

domlabel
"Ig-like C2-type domain"^^xsd:string
"protein kinase domain"^^xsd:string
"SH3 domain"^^xsd:string
"PDZ domain"^^xsd:string
"Ig-like domain"^^xsd:string
"SAM-dependent MTase PRMT-type domain"^^xsd:string
"disintegrin domain"^^xsd:string
"peptidase M12B domain"^^xsd:string
"SAM-dependent MTase C5-type domain"^^xsd:string
"F5/8 type C domain"^^xsd:string

mutagenesis tutorial

NXQ_00124 - What are the 25 most frequent families with member count
family snorql-only tutorial

NXQ_00125 - Domains that are entirely covered by 3D structures
3D structure domain PDB region snorql-only tutorial

NXQ_00126 - Peptides that are potential neo N-termini from undescribed isoforms
alternative splicing peptide snorql-only tutorial

NXQ_00127 - Proteins with at least one known SUMOylation site
PTM SUMO tutorial

NXQ_00124 - What are the 25 most frequent families with member count
family snorql-only tutorial

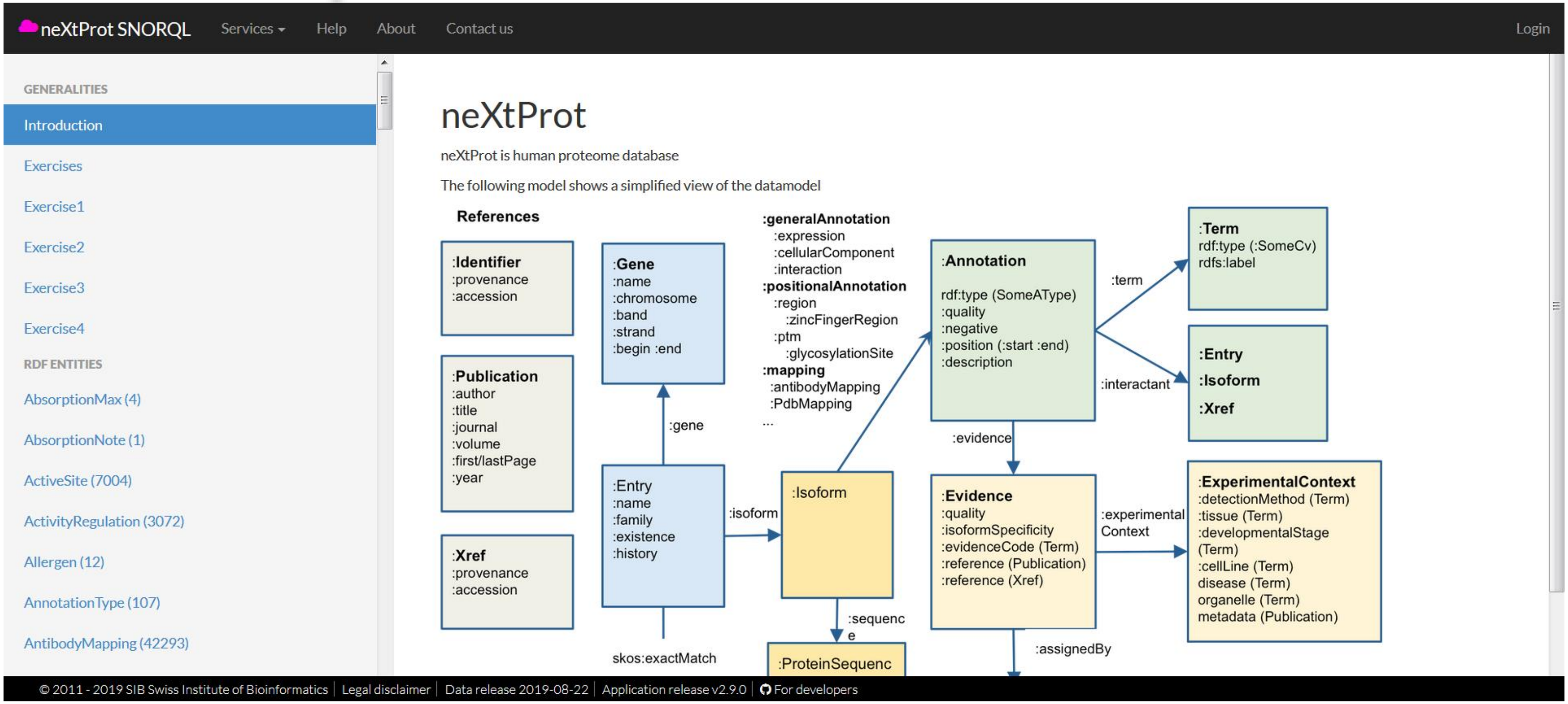
NXQ_00125 - Domains that are entirely covered by 3D structures
3D structure domain PDB region snorql-only tutorial

NXQ_00126 - Peptides that are potential neo N-termini from undescribed isoforms
alternative splicing peptide snorql-only tutorial

NXQ_00127 - Proteins with at least one known SUMOylation site
PTM SUMO tutorial

Documentation

https://snorql.nextprot.org/help/doc/introduction



neXtProt

neXtProt is human proteome database

The following model shows a simplified view of the datamodel

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.9.0 | For developers

Documentation for :disease

https://snorql.nextprot.org/help/entity/Disease

Disease Values▾ 37028

Involvement in disease: disease(s) associated with protein defect(s)

`:Isoform` `:disease`

`:Isoform` `:generalAnnotation`

`:Isoform` `:medical`

:Disease

`:entryAnnotationId`	`xsd:string`	37028	Example: AN_A0A0B4J1V
`:evidence`	`:Evidence`	70836	
`:quality`	`:QualityQualifier`	37028	
`:term`	`:UniprotDiseaseCv`	14856	
`rdfs:comment`	`xsd:string`	37028	Example: 3MC syndrome

Example

```
annotation:NX_P29597-1-AN_P29597_001466 :entryAnnotationId "AN_P29597_001466" .
annotation:NX_P29992-1-AN_P29992_XR_2718683 :evidence evidence:300002718683 .
annotation:NX_P29597-1-AN_P29597_001466 :quality :GOLD .
annotation:NX_P29597-1-AN_P29597_001466 :term cv:DI-02224 .
annotation:NX_P29597-1-AN_P29597_001466 rdfs:comment "Immunodeficiency 35 (IMD35) [MIM:611521]: A primary immunodeficiency characterized by recurrent skin abscesses, pneumonia, and highly elevated serum IgE. The disease is caused by mutations affecting the gene represented in this entry." .
```

Description

Subjects for :disease

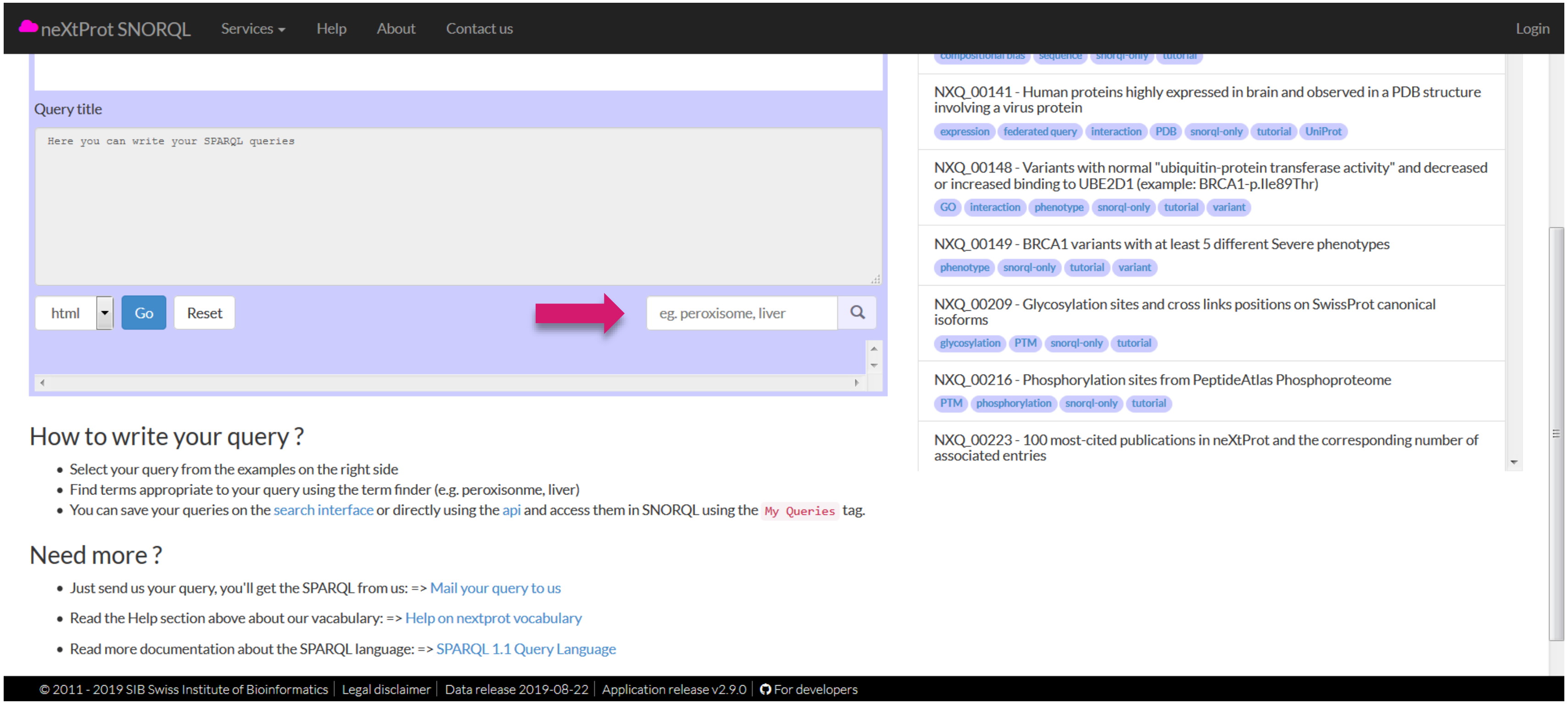
Objects for :disease

Values

Searching for terms

https://snorql.nextprot.org/

Find controlled vocabulary terms quickly



NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein

NXQ_00148 - Variants with normal "ubiquitin-protein transferase activity" and decreased or increased binding to UBE2D1 (example: BRCA1-p.Ile89Thr)

NXQ_00149 - BRCA1 variants with at least 5 different Severe phenotypes

NXQ_00209 - Glycosylation sites and cross links positions on SwissProt canonical isoforms

NXQ_00216 - Phosphorylation sites from PeptideAtlas Phosphoproteome

NXQ_00223 - 100 most-cited publications in neXtProt and the corresponding number of associated entries

How to write your query ?

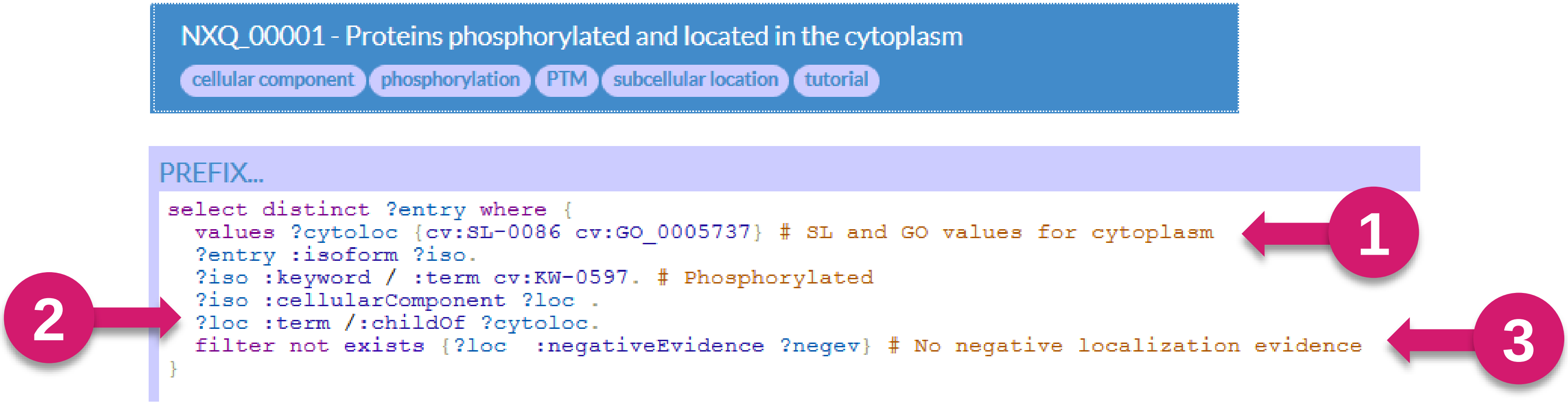
- Select your query from the examples on the right side
- Find terms appropriate to your query using the term finder (e.g. peroxisonme, liver)
- You can save your queries on the search interface or directly using the api and access them in SNORQL using the `My Queries` tag.

Need more ?

- Just send us your query, you'll get the SPARQL from us: => Mail your query to us
- Read the Help section above about our vacabulary: => Help on nextprot vocabulary
- Read more documentation about the SPARQL language: => SPARQL 1.1 Query Language

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.9.0 | For developers

Tips



NXQ_00001 - Proteins phosphorylated and located in the cytoplasm

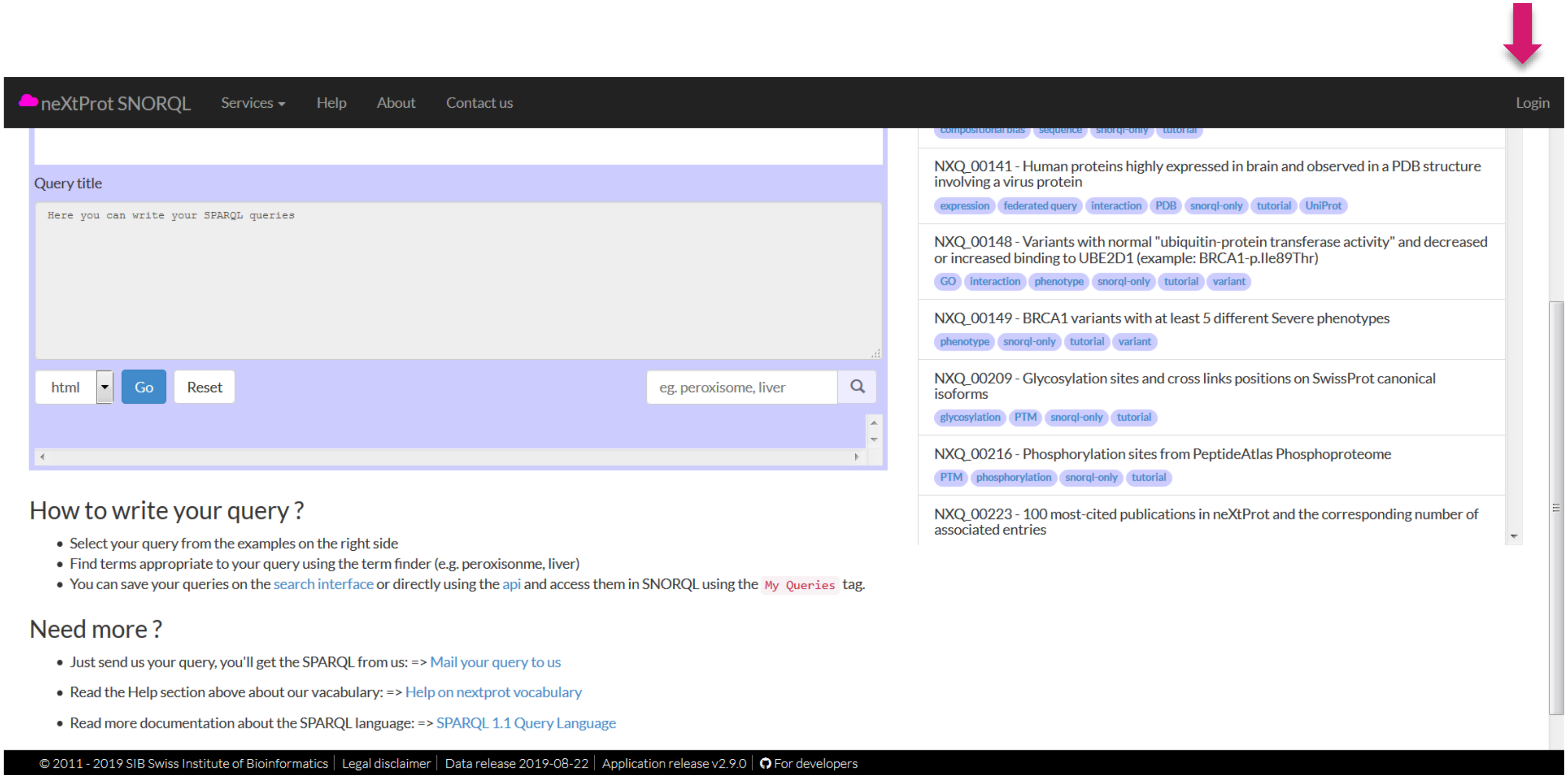
PREFIX...

```
select distinct ?entry where {
  values ?cytoloc {cv:SL-0086 cv:GO_0005737} # SL and GO values for cytoplasm
  ?entry :isoform ?iso.
  ?iso :keyword / :term cv:KW-0597. # Phosphorylated
  ?iso :cellularComponent ?loc .
  ?loc :term /:childOf ?cytoloc.
  filter not exists {?loc  :negativeEvidence ?negev} # No negative localization evidence
}
```

1. For cellular localization, use both controlled vocabularies (GO-CC and UniProtKB subcellular locations)
2. For hierarchical ontologies (i.e. GO, UniProtKB subcellular locations, neXtProt anatomy), use :childOf to include annotations to child terms
3. Remember to exclude negative annotations

Saving queries

https://snorql.nextprot.org/



Query title

Here you can write your SPARQL queries

NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein

expression · federated query · interaction · PDB · snorql-only · tutorial · UniProt

NXQ_00148 - Variants with normal "ubiquitin-protein transferase activity" and decreased or increased binding to UBE2D1 (example: BRCA1-p.Ile89Thr)

GO · interaction · phenotype · snorql-only · tutorial · variant

NXQ_00149 - BRCA1 variants with at least 5 different Severe phenotypes

phenotype · snorql-only · tutorial · variant

NXQ_00209 - Glycosylation sites and cross links positions on SwissProt canonical isoforms

glycosylation · PTM · snorql-only · tutorial

NXQ_00216 - Phosphorylation sites from PeptideAtlas Phosphoproteome

PTM · phosphorylation · snorql-only · tutorial

NXQ_00223 - 100 most-cited publications in neXtProt and the corresponding number of associated entries

How to write your query ?

- Select your query from the examples on the right side
- Find terms appropriate to your query using the term finder (e.g. peroxisonme, liver)
- You can save your queries on the search interface or directly using the api and access them in SNORQL using the `My Queries` tag.

Need more ?

- Just send us your query, you'll get the SPARQL from us: => Mail your query to us
- Read the Help section above about our vocabulary: => Help on nextprot vocabulary
- Read more documentation about the SPARQL language: => SPARQL 1.1 Query Language

© 2011 - 2019 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2019-08-22 | Application release v2.9.0 | For developers

Federated SPARQL queries

Used to explore the data in neXtProt and other resources

NXQ_00094 - Proteins which are targets of antipsychotic drugs and highly expressed in brain

drug · expression · federated query · tutorial

NXQ_00096 - Proteins which are targets of drugs for cardiac therapy

drug · federated query · tutorial

NXQ_00139 - Protein kinases which are drug targets according to CHEMBL

CHEMBL · drug · federated query · tutorial

NXQ_00140 - Proteins that interact with viral proteins

federated query · interaction · PPI · tutorial · UniProt

NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein

expression · federated query · interaction · PDB · snorql-only · tutorial · UniProt

NXQ_00246 - Proteins which are enzymes catalyzing a reaction involving lipids

enzyme · federated query · tutorial

NXQ_00253 - Human pathways in which at least one protein is mitochondrial GOLD

federated query · pathway · snorql-only · subcellular location · tutorial

NXQ_00254 - Proteins with associated pathways in WikiPathways

federated query · pathway · snorql-only · tutorial

DRUGBANK

ChEMBL

UniProt

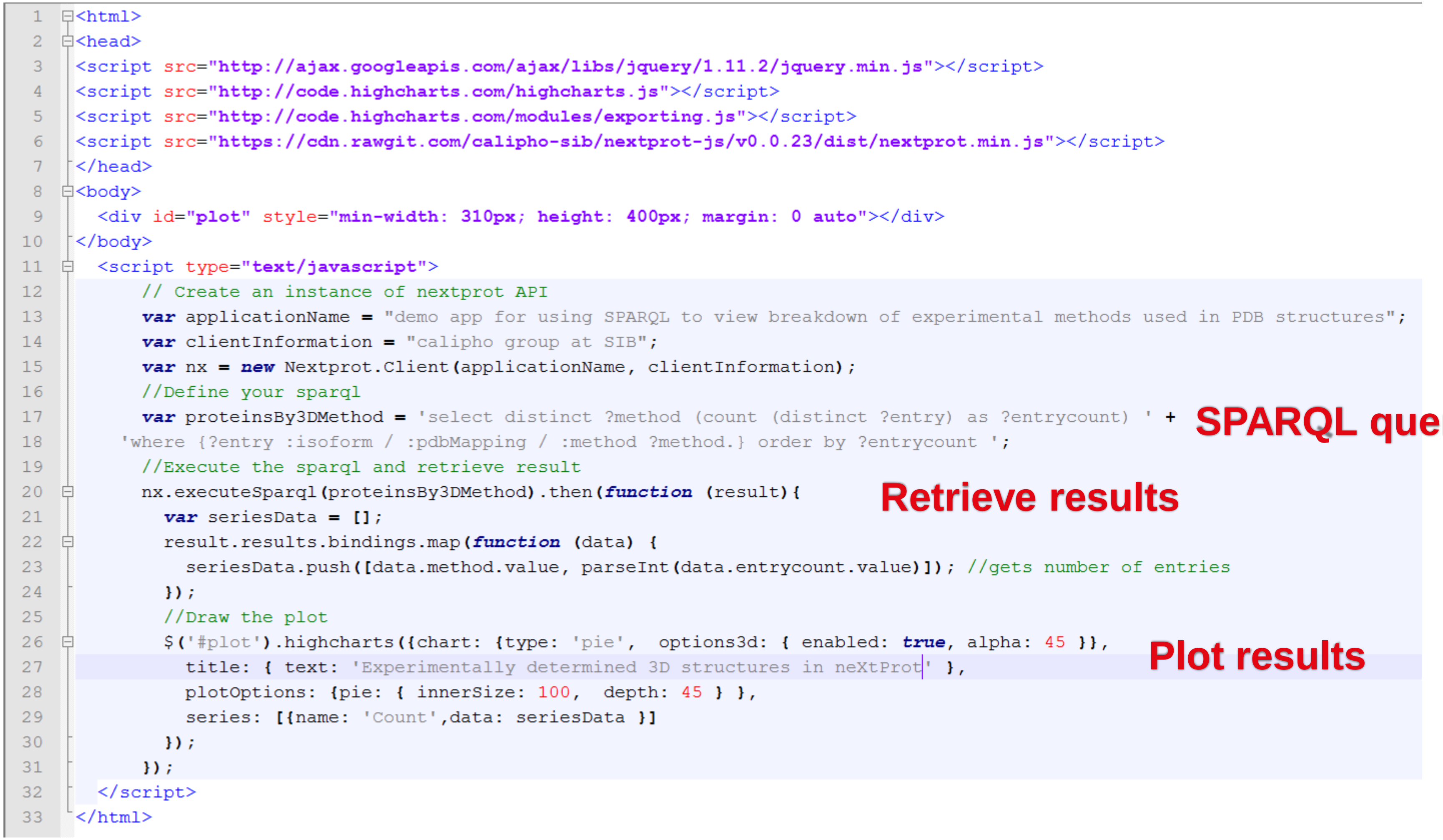
PDBj
Protein Data Bank Japan

Rhea

WikiPathways
Pathways for the People

Data visualization

Use javascript to see SPARQL query results using Highcharts.com



```
<html>
<head>
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.11.2/jquery.min.js"></script>
<script src="http://code.highcharts.com/highcharts.js"></script>
<script src="http://code.highcharts.com/modules/exporting.js"></script>
<script src="https://cdn.rawgit.com/calipho-sib/nextprot-js/v0.0.23/dist/nextprot.min.js"></script>
</head>
<body>
  <div id="plot" style="min-width: 310px; height: 400px; margin: 0 auto"></div>
</body>
  <script type="text/javascript">
      // Create an instance of nextprot API
      var applicationName = "demo app for using SPARQL to view breakdown of experimental methods used in PDB structures";
      var clientInformation = "calipho group at SIB";
      var nx = new Nextprot.Client(applicationName, clientInformation);
      //Define your sparql
      var proteinsBy3DMethod = 'select distinct ?method (count (distinct ?entry) as ?entrycount) ' +
    'where {?entry :isoform / :pdbMapping / :method ?method.} order by ?entrycount ';
      //Execute the sparql and retrieve result
      nx.executeSparql(proteinsBy3DMethod).then(function (result){
        var seriesData = [];
        result.results.bindings.map(function (data) {
          seriesData.push([data.method.value, parseInt(data.entrycount.value)]); //gets number of entries
        });
        //Draw the plot
        $('#plot').highcharts({chart: {type: 'pie',  options3d: { enabled: true, alpha: 45 }},
          title: { text: 'Experimentally determined 3D structures in neXtProt' },
          plotOptions: {pie: { innerSize: 100,  depth: 45 } },
          series: [{name: 'Count',data: seriesData }]
        });
      });
  </script>
</html>
```

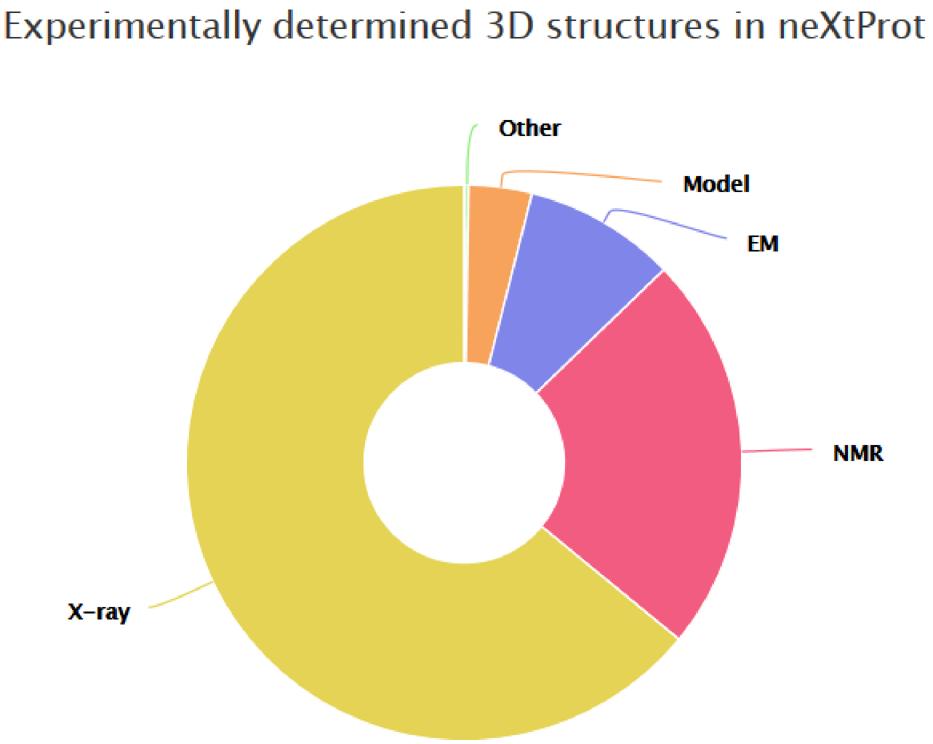
SPARQL query

Retrieve results

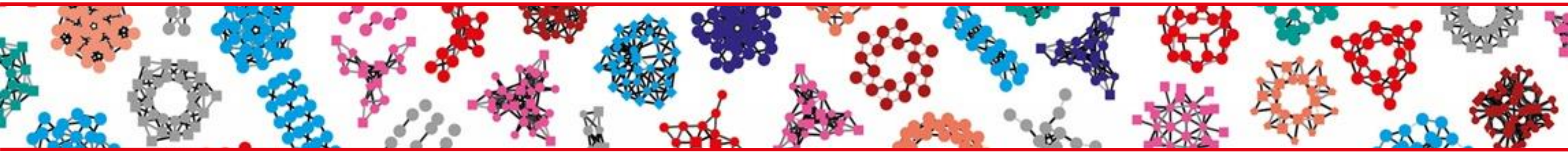
Plot results

Visualizing SPARQL query results

Automatically updated at each data release



Overview



- 01 Introduction to neXtProt
- 02 Data model
- 03 Uses of SPARQL
- **04 Summary**

Querying neXtProt using SPARQL

Access

SPARQL endpoint
https://api.nextprot.org/sparql

Advanced search
https://www.nextprot.org/proteins/search?mode=advanced

SNORQL
https://snorql.nextprot.org/

Data model
https://www.nextprot.org/help/data-model

Help > RDF entities
https://snorql.nextprot.org/help/doc/introduction

Sample queries
Found in the Advanced search (SPARQL) and SNORQL interfaces

User guide
https://doi.org/10.7490/f1000research.1116829.1

Uses of SPARQL

- **Answer complex biological questions**

  Execute complex queries, across resources

- **Retrieve data**

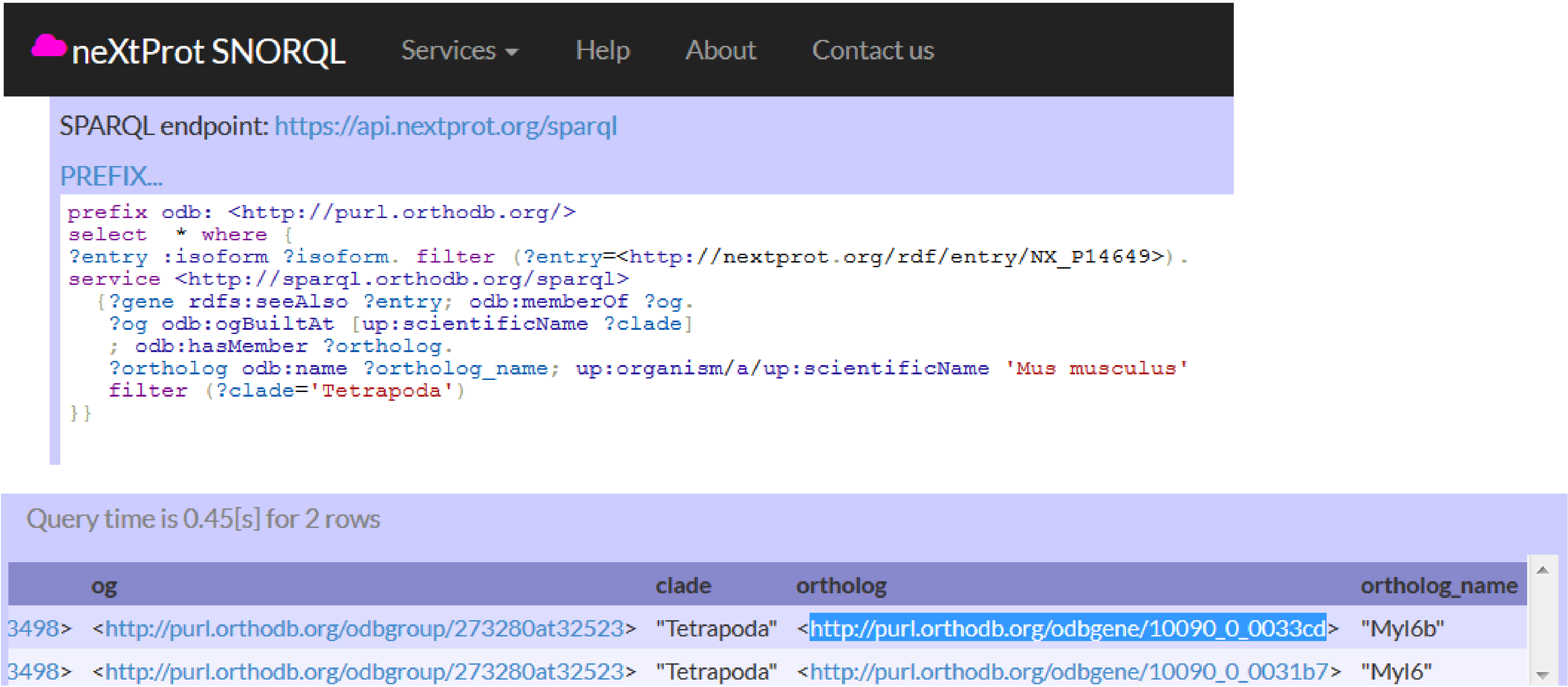
  For benchmarking, display in web sites, etc.

- **Biocuration**
- **Quality control**

  For spot checks and global checks of the data

Federated query with OrthoDB

Mouse ortholog(s) found at **Tetrapoda level** for human myosin light chain 6B (MYL6B) (neXtProt entry **NX_P14649**)

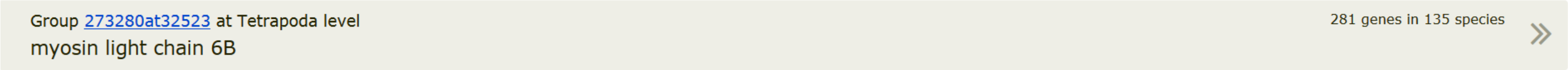


SPARQL endpoint: https://api.nextprot.org/sparql

```
prefix odb: <http://purl.orthodb.org/>
select  * where {
?entry :isoform ?isoform. filter (?entry=<http://nextprot.org/rdf/entry/NX_P14649>).
service <http://sparql.orthodb.org/sparql>
  {?gene rdfs:seeAlso ?entry; odb:memberOf ?og.
   ?og odb:ogBuiltAt [up:scientificName ?clade]
   ; odb:hasMember ?ortholog.
   ?ortholog odb:name ?ortholog_name; up:organism/a/up:scientificName 'Mus musculus'
   filter (?clade='Tetrapoda')
}}
```

Query time is 0.45[s] for 2 rows

	og	clade	ortholog	ortholog_name
3498>	<http://purl.orthodb.org/odbgroup/273280at32523>	"Tetrapoda"	<http://purl.orthodb.org/odbgene/10090_0_0033cd>	"Myl6b"
3498>	<http://purl.orthodb.org/odbgroup/273280at32523>	"Tetrapoda"	<http://purl.orthodb.org/odbgene/10090_0_0031b7>	"Myl6"



Group 273280at32523 at Tetrapoda level
myosin light chain 6B

281 genes in 135 species

Acknowledgements

Photograph taken by Chris James

neXtProt

Directors
Amos Bairoch, Lydie Lane

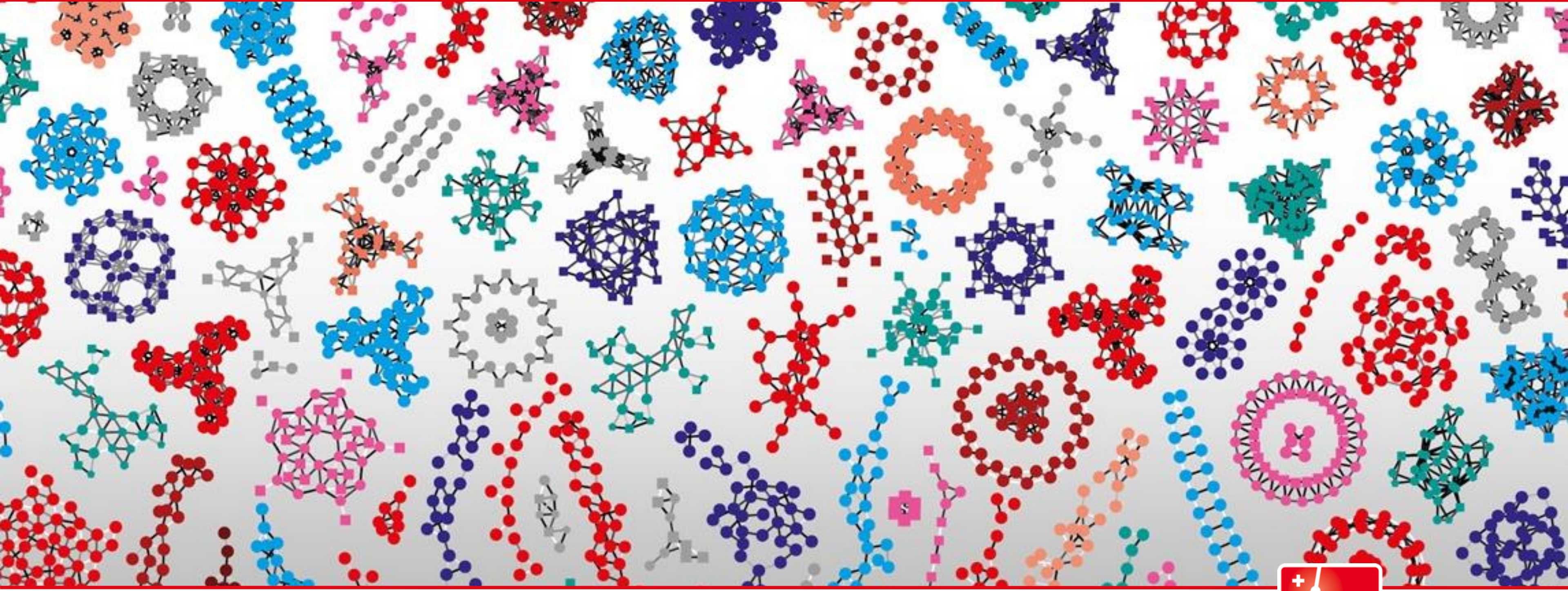
Biocurators
Paula Duek

Developers
Pierre-André Michel, Alain Gateau, Valentine Rech de Laval

Researcher
Mathieu Schaeffer, Kasun Samarasinghe, Vimel Rattina

Quality assurance
Monique Zahn

Web: https://www.nextprot.org/
Twitter: @neXtProt_news
ResearchGate: neXtProt project



Swiss Institute of
Bioinformatics

Thank you!

E-mail: monique.zahn@sib.swiss

www.sib.swiss