


Swiss Institute of  
Bioinformatics

# Using SPARQL to explore human protein data in neXtProt and beyond

Lydie Lane,  
OpenRiskNet final workshop, Amsterdam, Oct 24 2019

[www.nextprot.org](http://www.nextprot.org) is the SIB database on human proteins



Search

ABOUT US SERVICES PLATFORMS COMMUNITIES EVENTS NEWS INTRANET

Home » About us » Who we are » ELIXIR Nodes »

ABOUT US

What we do

▼ Who we are

▶ ELIXIR Hub

ELIXIR Nodes

How countries join

Why ELIXIR is needed

▶ How we are funded

▶ Governance

Implementation Studies


Collaborations

## ELIXIR Switzerland

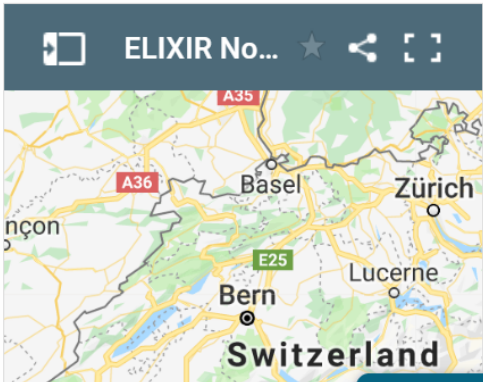
The [SIB Swiss Institute of Bioinformatics](#) is an academic not-for-profit organization whose mission is to lead and coordinate the field of bioinformatics in Switzerland. Its data science experts join forces to advance biological and medical research and enhance health.

SIB (i) provides the national and international life science community with a state-of-the-art bioinformatics infrastructure, including services, resources, expertise; and (ii) federates world-class researchers and delivers training in bioinformatics.

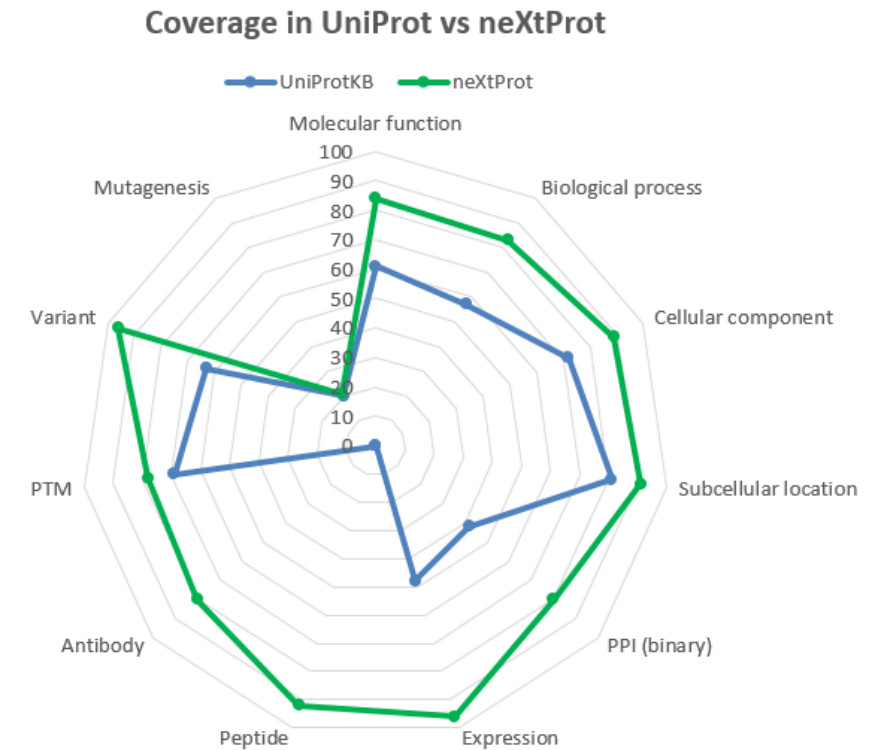
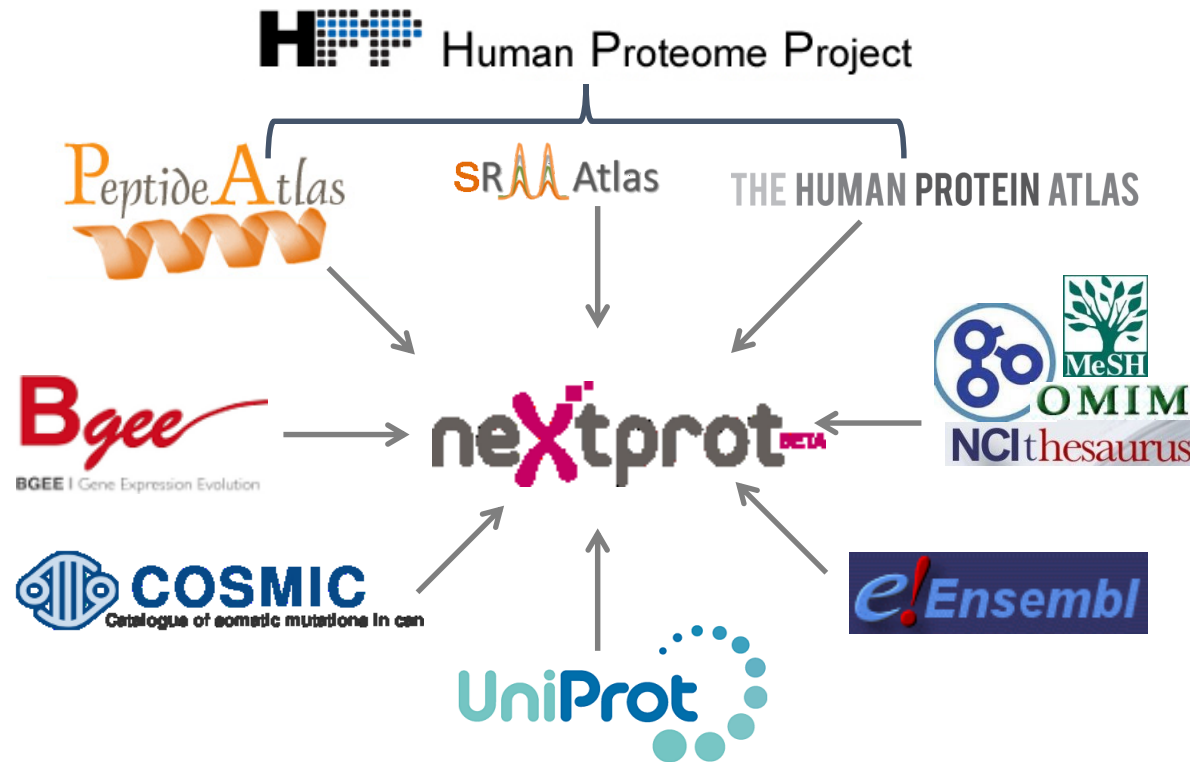
The institute includes some 70 world-class research and service groups including 800 scientists in the fields of genomics, proteomics, evolution and phylogeny, systems biology, structural biology, text mining and machine learning and personalized health.



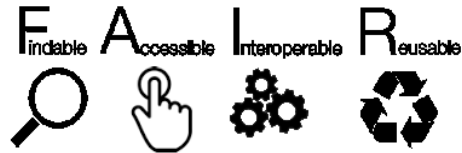
Swiss Institute of Bioinformatics



[www.nextprot.org](http://www.nextprot.org) collects and represents knowledge on **human proteins** at genomic, transcriptomic and proteomic levels.



# Access to neXtProt data and tools



## FINDABLE

- **ENTRIES** have a globally unique and persistent ID

## ACCESSIBLE

- **Web:** human readable
- **API:** machine parsable
- **FTP:** XML, PEFF, RDF/ttl

## INTEROPERABLE

- **Use controlled vocabularies and ontologies**
- **SPARQL end-point**

## REUSABLE

- **Detailed provenance**
- **Data license CC BY 4.0**
- **All data since 2011 on FTP**
- **Code on GitHub**



### Web site

<https://www.nextprot.org>



### File transfer protocol (FTP)

<ftp://ftp.nextprot.org/>



### Application programming interface (API)

<https://api.nextprot.org/>



### SPARQL endpoint

<https://api.nextprot.org/sparql>



# neXtProt in numbers

## Release statistics

Data release: 2019-08-22

Application release: v2.23.2

## Data statistics

### ENTRIES


Protein entries	20399
Isoforms (produced by splicing)	42410
Identifiers	878422
Binary interactions	240010
Post-translational modifications (PTMs)	190921
Natural peptides (seen by MS)	1772990
Variants (including disease mutations)	6019871
Controlled vocabularies and ontology terms	252243
Total number of publications	495928
Links between protein entries and terms	3977548

## Each protein entry has multiple views

The screenshot displays the neXtProt website interface for the protein entry NX\_P01350. The left sidebar shows a navigation menu with categories: PROTEIN (selected) and GENE. Under PROTEIN, the 'Function' view is active, with other options like Medical, Expression, Interactions, Localization, Sequence, Proteomics, Structures, Identifiers, Peptides, and Phenotypes. Under GENE, options for Exons and Identifiers are visible. The main content area is titled 'GAST - Function' and includes a search bar and a 'Gold Only' filter. The protein is identified as GAST → Gastrin. Key information includes: Cleaved into: Big gastrin; Gastrin; Gastrin-14; Gastrin-52; Gastrin-6; Gastrin-71; Gene name: GAST; Family name: Gastrin/cholecystokinin. A note states: 'Entry whose protein(s) existence is based on evidence at protein level.' There are buttons to 'Extend overview' and 'Show evidences'. The 'Keywords' section lists 'MOLECULAR FUNCTION' and 'Hormone' with a link to 'KW-0372 Definition'. The 'Further external links' section is currently empty. The footer contains copyright information: '© 2011 - 2017 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2017-08-01 | Application release v2.10.0 | For developers'.

neXtprot


Tools ▾ Portals ▾ Download ▾ Help ▾ About ▾ Contact ▾


Lydie Lane ▾ 


< NX\_P01350

GAST - Function

1 GENE 1 ISO 87 REF

Search in neXtProt... 

Gold Only 

 PROTEIN ▾

**Function**

Medical

Expression

Interactions

Localization

Sequence


Proteomics

Structures

Identifiers


Peptides

Phenotypes

 GENE ▾

Exons

Identifiers


GAST → Gastrin 


Cleaved into : Big gastrin; Gastrin; Gastrin-14; Gastrin-52; Gastrin-6; Gastrin-71

Gene name : [GAST](#)

Family name : [Gastrin/cholecystokinin](#)

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. 




Keywords

MOLECULAR FUNCTION

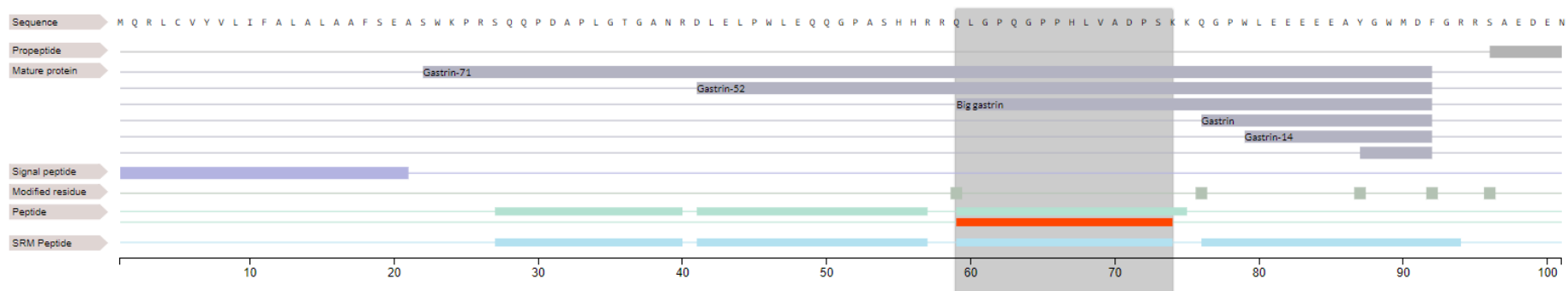
Hormone [KW-0372](#) [Definition](#)

Further external links

© 2011 - 2017 SIB Swiss Institute of Bioinformatics | Legal disclaimer | Data release 2017-08-01 | Application release v2.10.0 |  For developers

# Proteomics view

- Which MS/MS peptides from my protein have been found ?
- Which PTM?



Q Search a feature.. SEARCH 0/0

Name	Position	Length	Description	Evidence
Modified residue	87	1	Sulfonyltyrosine, partial • Gold	3 UniProtKB
Modified residue	92	1	Phenylalanine amide • Gold	2 UniProtKB
Modified residue	96	1	Phosphoserine • Gold	2 UniProtKB
Peptide				
Peptide	27-40	14	PAP00377936 unique • Gold	2 PeptideAtlas human set(s)
SRM Peptide	27-40	14	PAP00377936 unique • Gold	1 SRMAtlas
Peptide	41-57	17	PAP01528086 unique • Gold	1 PeptideAtlas human set(s)
SRM Peptide	41-57	17	PAP01528086 unique • Gold	1 SRMAtlas
Peptide	59-75	17	PAP00508568 unique • Gold	1 PeptideAtlas human set(s)
Peptide	59-74	16	PAP01525629 unique • Gold	1 PeptideAtlas human set(s)
SRM Peptide	59-74	16	PAP01525629 unique • Gold	1 SRMAtlas

Isoform: Iso 1 101 aa Mass: 11394 Da pI: 5.08





































Search in sequence...

View FASTA Blast sequence Blast selection

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

# Expression view

✓ *Where is this protein expressed, at mRNA and protein levels?*

Tissue / Cell type	▼ Expand all	mRNA expression			Protein expression	
		Microarray	RNA-seq	EST	IHC	
▼ Alimentary system <a href="#">TS-1293</a>		 	 		 	
▼ Gastrointestinal tract <a href="#">TS-0407</a>		 	 		 	
▶ Intestine <a href="#">TS-0490</a>		 				
▶ Esophagus <a href="#">TS-0700</a> • Gold		 2		 1		
▼ Stomach <a href="#">TS-0980</a> • Gold		 2			 	
▼ Stomach mucosa <a href="#">TS-0404</a>					 	
▼ Stomach epithelium <a href="#">TS-2068</a>					 	
Stomach glandular cell <a href="#">TS-1284</a> • Gold					  2	
▶ Mouth <a href="#">TS-1315</a> • Gold		 7				
▼ Pancreas <a href="#">TS-0736</a> • Gold		 1	 1			
▶ Exocrine pancreas <a href="#">TS-1241</a>						

Expression information is integrated from Bgee (MicroArray and EST) and from HPA (IHC protein and RNA-seq)



# Structure view

✓ *Is this particular variant externally accessible or internal to the structure?*

**Isoforms** NX\_P40424-1 NX\_P40424-2 NX\_P40424-3 [See isoforms](#)

**FILTER** [All/None](#) ☒ Region ☒ Site ☒ Variant ☒ Conflict ☒ Secondary structure **Position : 412** **Zoom : x 1** [Show help](#)

Sequence  
DNA binding  
Composition bias  
Site  
Variant  
Conflict  
Helix  
Turn

Homeobox-TALE-type

Variant  
Variant  
Variant  
Variant  
Variant  
Variant  
Variant  
Variant  
Variant  
Variant

**Bionext**

**2 3D Structures**

**OPTIONS**

Choose PDB ID:


Choose chain:  ☒ SHOW CONTEXT (ALL)

[Click on the 3D viewer then press \*\*i\*\* to display help.](#)

**3D STRUCTURES**

PDB	Method	Resolution (Å)	Chain(s)	Positions	Further links
<a href="#">1B72</a>	X-ray	2.35	B	233-319	<a href="#">PDBsum</a> <a href="#">Proteopedia</a>
<a href="#">1PUF</a>	X-ray	1.90	B	233-305	<a href="#">PDBsum</a> <a href="#">Proteopedia</a>

# Searching for information in neXtProt



Exploring the universe of human proteins

☒ Simple search ☐ Advanced search (SPARQL)

proteins ▼ Gold only ▼

*e.g.:* [Search for MSH6 in proteins](#), [Search for author Doolittle in publications](#), [Search for liver in terms](#)

# neXtProt SPARQL for beginners

**F1000Research**  
Open for Science

Search

SUBMIT YOUR RESEARCH

BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ▾ ABOUT ▾ BLOG MY RESEARCH ▾ SIGN IN


DOCUMENT

NOT PEER REVIEWED

Metrics | 27 Views | 3 Downloads

VIEW FULL SCREEN

Introduction to bioinformatics



The neXtProt Knowledgebase  
Querying using SPARQL


A CRITICAL GUIDE

DOWNLOAD  
2.89 MB

SHARE

CITE

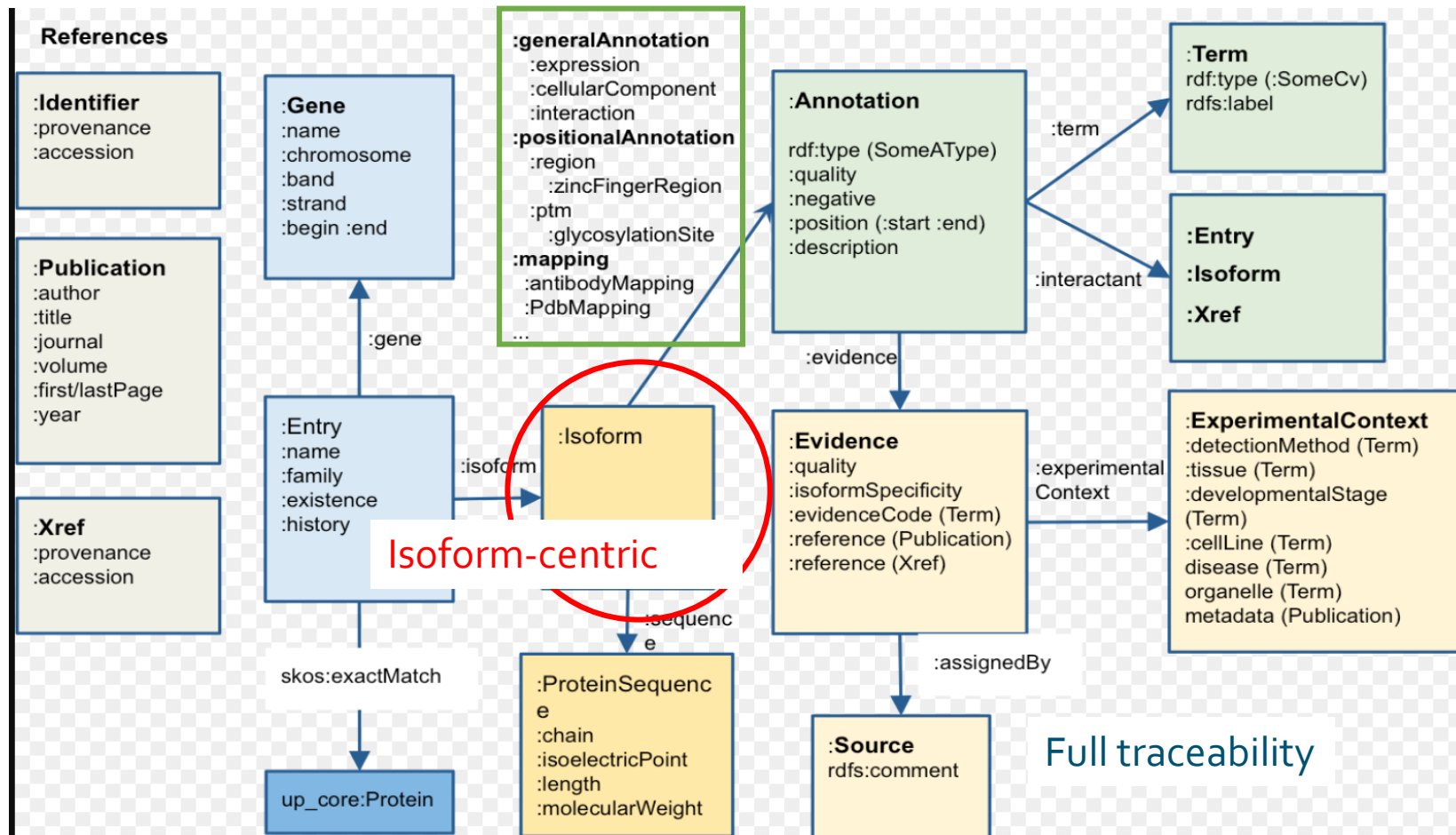
PART OF THE COLLECTION

**Bioinformatics  
Education and Training  
Collection**

# [www.nextprot.org](http://www.nextprot.org) RDF data model

A wide variety of experimental and predicted annotations

Extensive use of CV



# Advanced search (SPARQL) interface 1/2

Retrieves lists of **entries** with associated information

Simple search ☒ Advanced search

proteins ▼

```
1 #Proteins phosphorylated and located in
2 #the cytoplasm
3
4 select distinct ?entry where {
5   ?entry :isoform ?iso.
6   ?iso :keyword / :term cv:KW-0597.
7 }
8
```

Q Search

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

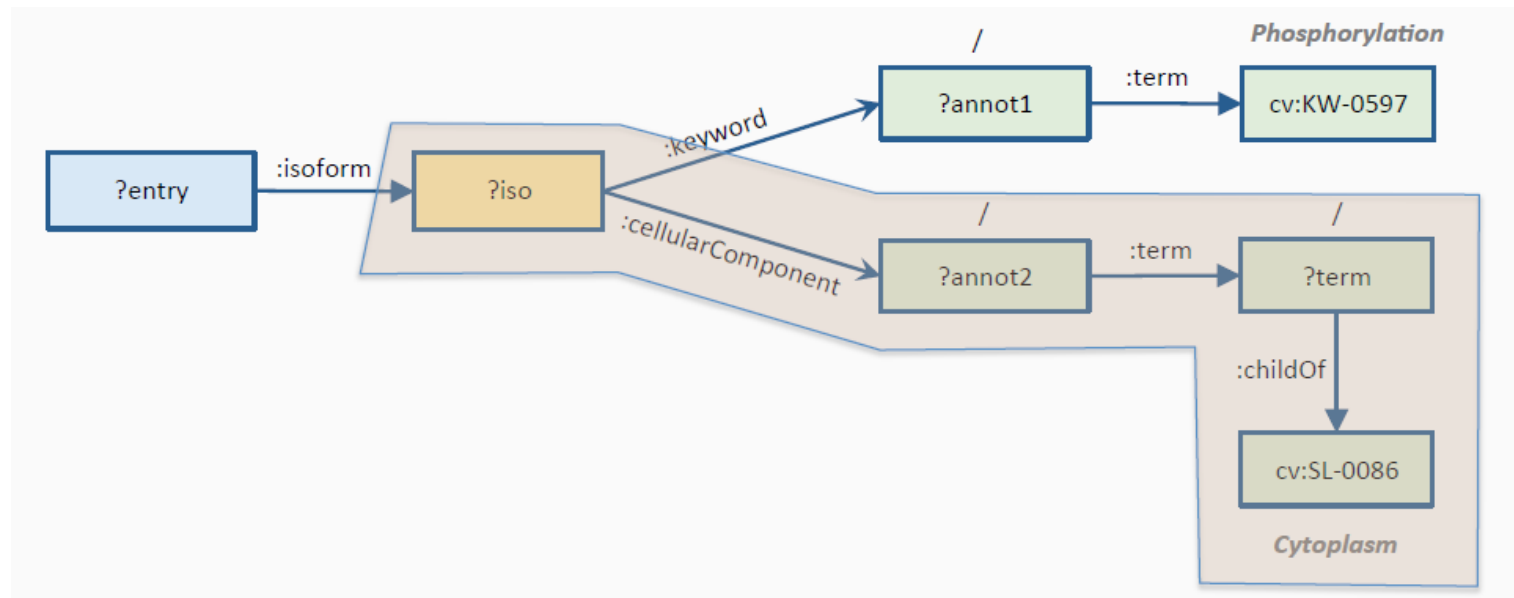
Show 1 to 50 of 5618 ▼

1 of 113 page(s)

Summary

Details

Download 📄



# Advanced search (SPARQL) interface 2/2

proteins ▾

Simple search ☐ Advanced search (SPARQL) ☒

1

+ ≡ × Search

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

Tags ▾ inter

New Query ×

Proteins that interact with protein RBM17 and that are involved in splicing	NXQ_00012	
Proteins with a PDZ domain that interact with at least one protein which is expressed in brain	NXQ_00015	
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 which are located on the genome next to a protein which is involved in spermatogenesis	NXQ_00058	
Proteins whose gene is annotated		
Proteins with a 3D structure in complex with another human protein which is not reported as binary interactant	NXQ_00132	
Proteins that potentially interact with Class I PDZ domains (whose C-terminal	NXQ_00137	

For help with these queries look at [the model](#) or [contact us](#).

>140 pre-made queries available



# SNORQL interface <https://snorql.nextprot.org>

Retrieves **any data** as a table

>170 pre-made queries available

NXQ\_00124 - What are the 25 most frequent families with member count

family snorql-only tutorial

NXQ\_00256 - Variants identified in exome datasets in a frequent homozygote state

evidence sequence variation snorql-only tutorial variant

What are the 25 most frequent families with member count

html Go Reset

eg. peroxisome, liver

Found a bug? [Improve this query!](#)

Query time is 0.291[s] for 25 rows

familylabel	membercnt
"G-protein coupled receptor 1 family"	723
"Krueppel C2H2-type zinc-finger protein family"	546
"Protein kinase superfamily"	492
"Small GTPase superfamily"	162
"Immunoglobulin superfamily"	130
"Peptidase S1 family"	121
"Major facilitator superfamily"	99

# A detailed help on neXtProt RDF entities is available

The screenshot shows the neXtProt SNORQL web interface. The browser address bar displays `snorql.nextprot.org/help/entity/Allergen`. The page header includes the neXtProt logo, navigation links (Services, Help, About, Contact us), and a user profile (Lydie Lane). A left sidebar lists 'GENERALITIES' (Introduction, Exercises, Exercise1-4) and 'RDF ENTITIES' (AbsorptionMax (4), AbsorptionNote (1), ActiveSite (7004), ActivityRegulation (3072), Allergen (12), AnnotationType (107), AntibodyMapping (42293), BetaStrand (147952), BgeeDevelopmentalStageCv (192)). The 'Allergen (12)' entity is selected and highlighted in blue.

The main content area for the 'Allergen' entity shows a 'Values' dropdown with '12' items. Below this, a description states: 'Allergenic properties: information relevant to allergenic proteins'.

Three property-value pairs are listed on the left:

- `:Isoform :allergen`
- `:Isoform :generalAnnotation`
- `:Isoform :medical`

On the right, a detailed view of the `:Allergen` entity is shown with the following properties and values:

- `:entryAnnotationId` `xsd:string` 12 `AN_O43290_0377`
- `:evidence` `:Evidence` 12
- `:quality` `:QualityQualifier` 12
- `rdfs:comment` `xsd:string` 12 `Causes an allergic react`

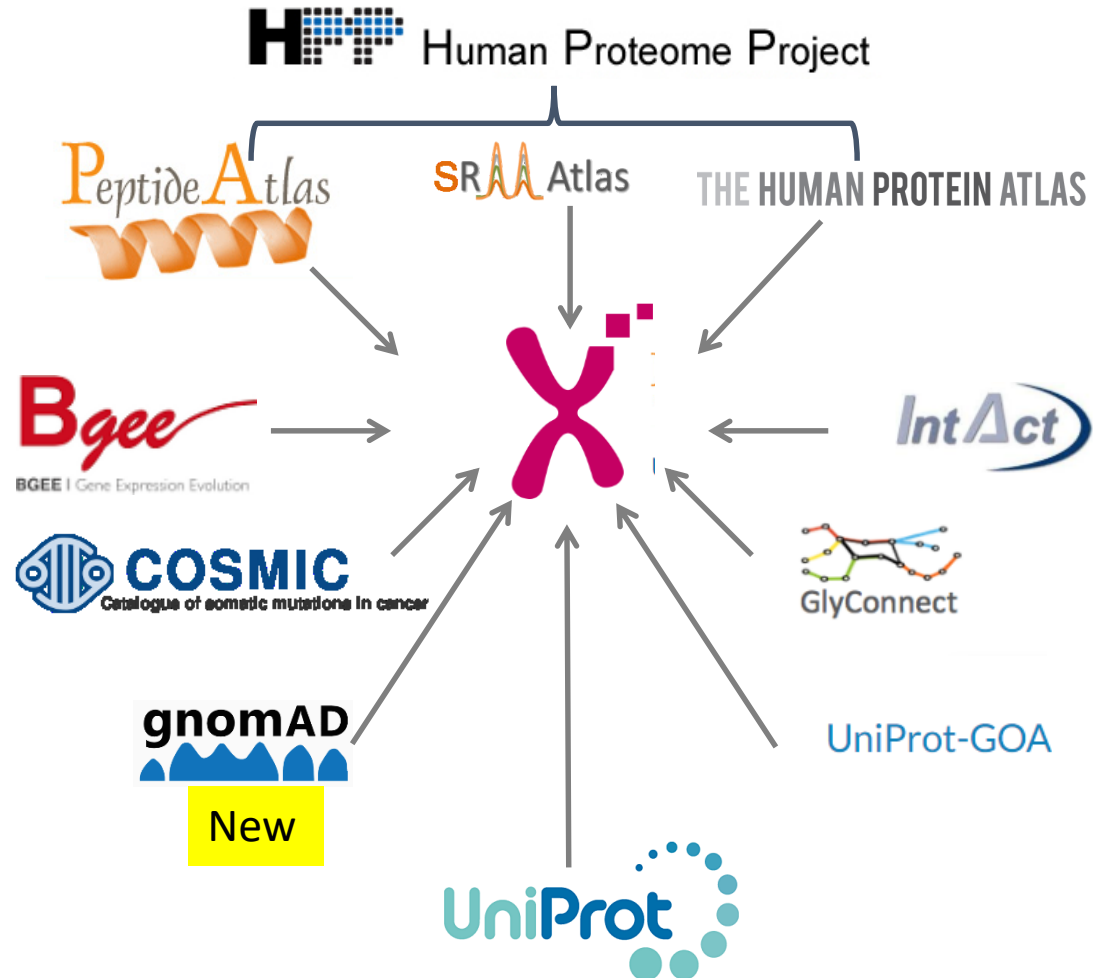
Below this, an 'Example' section provides a list of RDF annotations:

```
annotation:NX_P02538-1-AN_P02538_0222 :entryAnnotationId "AN_P02538_0222" .
annotation:NX_P02538-1-AN_P02538_0222 :evidence evidence:85311179 .
annotation:NX_Q9BPX6-2-AN_Q9BPX6_0220 :quality :GOLD .
annotation:NX_P02538-1-AN_P02538_0222 rdfs:comment "Causes an allergic reaction in human. Binds to IgE from atopic dermatitis (AD) patients. Identified as an IgE autoantigen in atopic dermatitis (AD) patients with severe skin manifestations." .
```

A second 'Example' section at the bottom shows query patterns:

```
?entry :isoform/:allergen ?statement
?entry :isoform/:binaryInteraction/:interactant/:allergen ?statement
?entry :isoform/:binaryInteraction/:interactant/:generalAnnotation ?statement
```

# [www.nextprot.org](http://www.nextprot.org) does not contain everything....



- ✓ MS-data
- ✓ Antibody-based data
- ✓ RNA-seq data
- ✓ PTMs
- ✓ Variants (genomic and somatic)
- ✓ PPIs
- ✓ Functional annotations
- Phylogenetic information
- Data on model organisms
- Pharmacology data
- Toxicology data
- Clinical proteogenomic data
- Structural data
- Protein interactions with pathogens
- ...

# Going further with federated queries

Tags ▾ federated

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](#)



And soon to come...



# Federating neXtProt and WikiPathways

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

PREFIX...

```
PREFIX wp: <http://vocabularies.wikipathways.org/wp#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dcterms: <http://purl.org/dc/terms/>

select distinct ?pathwayname group_concat(distinct ?gen , ',' ) as ?genes where {
service <http://sparql.wikipathways.org/sparql> {
SELECT DISTINCT ?pathwayname ?gen WHERE
{
  { ?geneProduct a wp:GeneProduct . }
  union
  { ?geneProduct a wp:Protein . }
  ?geneProduct rdfs:label ?gen .
  filter(!regex(?gen,"[ a-z-]")). # ensures official gene names for subsequent neXtprot matching
  ?geneProduct rdfs:label ?gen .
  ?geneProduct dcterms:isPartOf ?pathway .
  ?pathway a wp:Pathway .
  ?pathway wp:organism ?organism .
  FILTER(contains(str(?organism),"9606"))
  ?pathway dc:title ?pathwayname .
}
```

?entry a :Entry .  
?entry :gene / :name ?gen .  
?entry :isoform / :cellularComponent ?loc .  
values ?mitoloc {cv:SL-0173 cv:GO\_0005739 } # SL and GO values for mitochondrion  
?loc :term / :childOf ?mitoloc . # mitochondrion

Tags ▾ [pathw](#)

NXQ\_00022 - Proteins with no function annotated  
[function](#) [tutorial](#)

NXQ\_00044 - Proteins involved in the Reactome pathway "Respiratory electron transport"  
[pathway](#) [tutorial](#)

NXQ\_00135 - Proteins involved in both Wnt and Hippo signaling pathways  
[function](#) [pathway](#) [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](#)

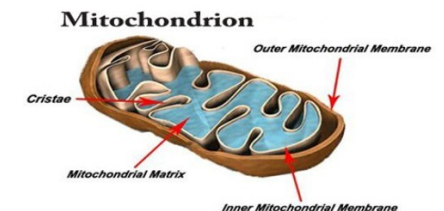
html ▾ Go Reset

html  
json  
**csv**  
xml

Improve this query!

is 0.502[s] for 781 rows

pathwayname	genes
"22q11.2 Deletion Syndrome"@en	"AIFM3,ALDH4A1,GLUD1,MRPL40,OAT,PRODH,RAF1,RTL10,SLC25A1,TXNRD2"
"4-hydroxytamoxifen, Dexamethasone, and Retinoic Acids Regulation of p27 Expression"@en	"MAP2K1,MAP2K2,MAPK1,MAPK3,MTOR,RAF1"
"ABC-family proteins mediated transport"@en	"ABCB10,ABCB6,ABCB8,ABCD1,ABCG1,PSMB3,RNF185,RNF5,UBB"
"ACE Inhibitor Pathway"@en	"CYP11B2"
"AGE/RAGE pathway"@en	"ATF2,CASP3,CYCS,FOXO1,LGALS3,MAP2K1,MAPK1,MAPK3,MAPK8,MMP2,MTOR,MYD88,PRKCA,RAF1,SHC1,SOD1,SRG"



# Retrieving proteins involved in pathways of interest, and their properties

Example: 106 entries associated with pathways named “nano...” (according to Wikipathways), and their associated subcellular location (GOLD according to neXtProt)

Query time is 0.663[s] for 106 rows			
entry	gen	pathwayname	locs
<a href="#">entry:NX_Q07812 -- (neXtProt link)</a>	"BAX"^^xsd:string	"Nanomaterial induced apoptosis"@en	"BAX complex,Bcl-2 family protein complex,Cytoplasm,Mitochondrion outer membrane,cell periphery,cytoplasm,cytosol,endoplasmic reticulum membrane,mitochondrial outer membrane,mitochondrion,transition pore complex,mitochondrion,nuclear envelope,nucleus,p
<a href="#">entry:NX_Q14249 -- (neXtProt link)</a>	"ENDO"^^xsd:string	"Nanomaterial induced apoptosis"@en	"Mitochondrion,mitochondrial inner membrane,mitochondrion,nucle
<a href="#">entry:NX_Q16611 -- (neXtProt link)</a>	"BAK1"^^xsd:string	"Nanomaterial induced apoptosis"@en	"BAK complex,Mitochondrion outer membrane,integral component of outer membrane,mitochondrial outer membrane,mitochondrion,p
<a href="#">entry:NX_P10415 -- (neXtProt link)</a>	"BCL2"^^xsd:string	"Nanomaterial induced apoptosis"@en	"Endoplasmic reticulum membrane,Mitochondrion outer membrane,Nucleoplasm,Nucleus membrane,cytoplasm,endoplasmic reticulum,endoplasmic reticulum membrane,membrane,mitochondrion



# Federating neXtProt, Wikipathways and DrugBank

169 proteins are part of “cancer pathways” and “druggable” according to DrugBank  
(35 are known targets of anti-neoplastic agents)



entry	callret-1	cpathways	drugs
<a href="#">entry:NX_O15530 -- (neXtProt link)</a>	"PDPK1"	"Endometrial cancer"	"Celecoxib"
<a href="#">entry:NX_P00352 -- (neXtProt link)</a>	"ALDH1A1"	"Folate-Alcohol and Cancer Pathway Hypotheses"	"Tretinoin"
<a href="#">entry:NX_P00374 -- (neXtProt link)</a>	"DHFR"	"Retinoblastoma Gene in Cancer"	"Methotrexate,Pemetrexed"
<a href="#">entry:NX_P00519 -- (neXtProt link)</a>	"ABL1"	"Apoptosis-related network due to altered Notch3 in ovarian cancer,Integrated Breast	"Dasatinib,Imatinib"

# Going further and filling gaps

---

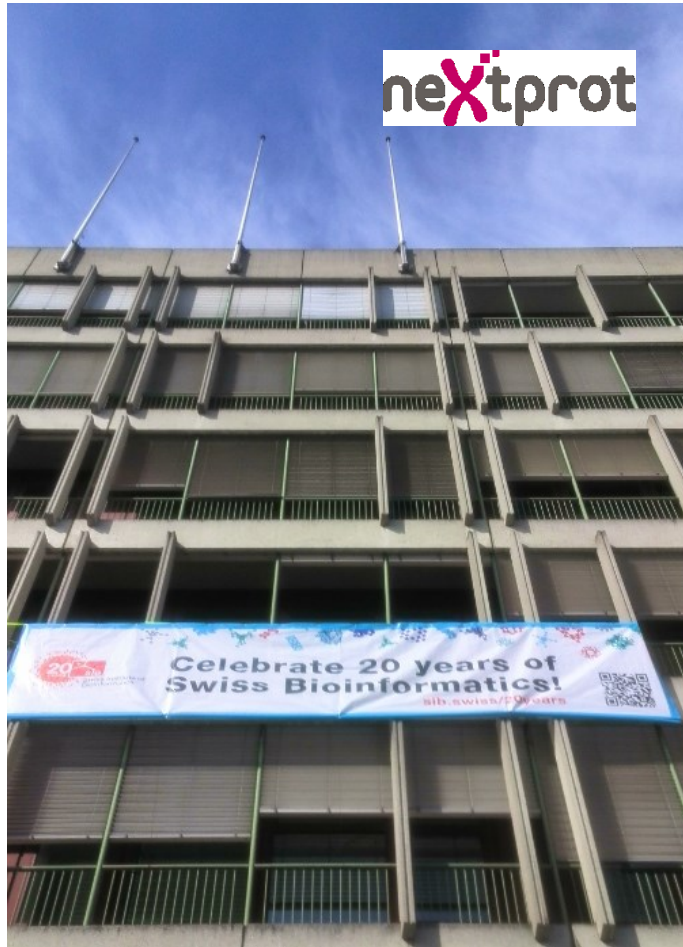
- **Annotating toxicology data?**

*Examples: proteins that bind nanoparticles? Proteins that are differentially expressed upon nanoparticle exposure?*

- **Federating with other SPARQL endpoints from OpenRisknet?**
- **Other ideas?**

***Happy to discuss use cases and possible applications with you!***

# The current neXtprot team @ SIB - Geneva



## **Directors**

Amos Bairoch, Lydie Lane

## **Biocurator**

Paula Duek

## **Developers**

Pierre-André Michel, Alain Gateau, Valentine Rech de Laval, Kasun Samarasinghe, Mathieu Schaeffer (PhD student)

## **Quality assurance**

Monique Zahn

**Contact:** [support@nextprot.org](mailto:support@nextprot.org)

**Web:** <https://www.nextprot.org/>

**Twitter:** @neXtProt\_news

**ResearchGate:** neXtProt project



# 12th International SWAT4HCLS Conference

Semantic Web Applications and Tools for Health Care and Life Sciences



## Edinburgh 2019

### LINKS

[About SWAT4HCLS](#)

### SWAT4LS 2019 IS SPONSORED BY:

[Read more about sponsorship opportunities.](#)

SWA  
venu  
of se  
Start  
expa  
and

A one-day tutorial on 9 independent SPARQL endpoints hosted by the SIB, including neXtProt, UniProt, GlyConnect, Rhea, OrthoDB, OMA and Bgee

RCH

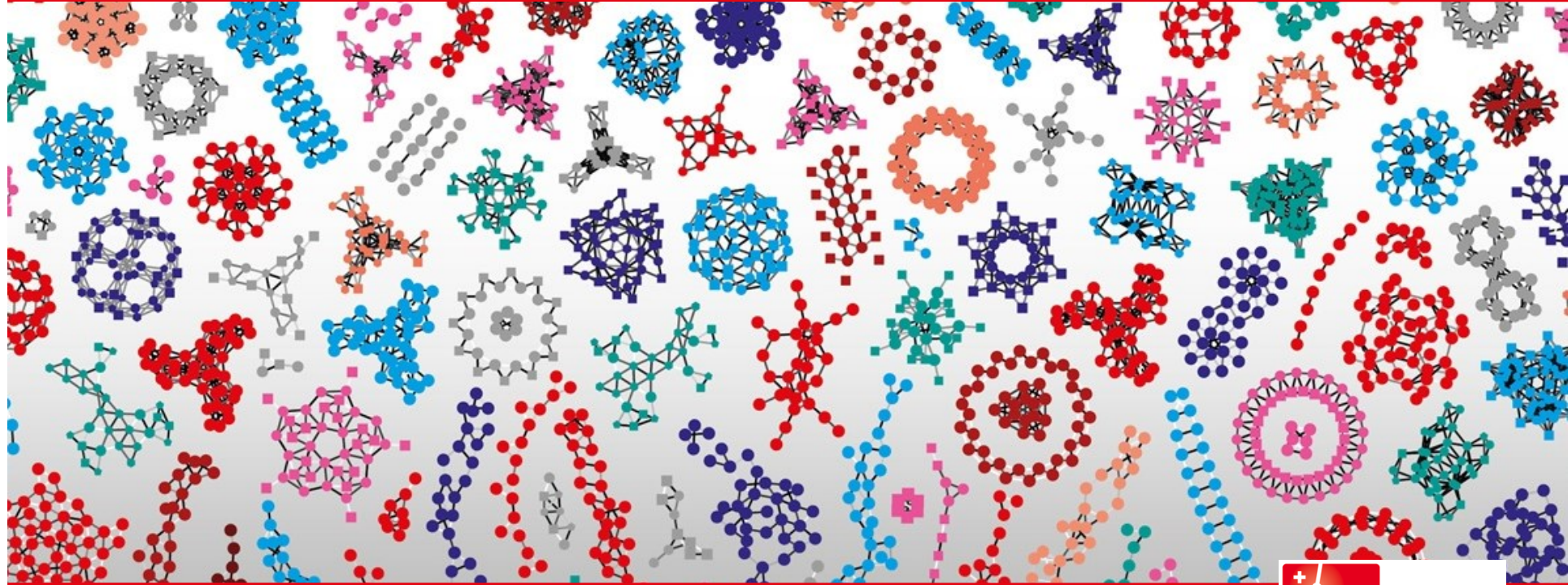
SWAT4LS intends to provide an open and vibrant discussion space cross cutting across academia, industry and institutional participants.

[Antwerp, 2018](#)  
[Rome, 2017](#)  
[Amsterdam, 2016](#)  
[Cambridge, 2015](#)

### LATEST NEWS

[Updates from SWAT4HCLS](#)





Swiss Institute of  
Bioinformatics

# Thank you!

E-mail: [lydie.lane@sib.swiss](mailto:lydie.lane@sib.swiss)



[www.sib.swiss](http://www.sib.swiss)