Using SPARQL to explore human protein data in neXtProt and beyond

Lydie Lane,
OpenRiskNet final workshop, Amsterdam, Oct 24 2019
www.nextprot.org is the SIB database on human proteins

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.
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
**Release statistics**

Data release: 2019-08-22  
Application release: v2.23.2

**Data statistics**

<table>
<thead>
<tr>
<th>ENTRIES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein entries</td>
<td>20399</td>
</tr>
<tr>
<td>Isoforms (produced by splicing)</td>
<td>42410</td>
</tr>
<tr>
<td>Identifiers</td>
<td>878422</td>
</tr>
<tr>
<td>Binary interactions</td>
<td>240010</td>
</tr>
<tr>
<td>Post-translational modifications (PTMs)</td>
<td>190921</td>
</tr>
<tr>
<td>Natural peptides (seen by MS)</td>
<td>1772990</td>
</tr>
<tr>
<td>Variants (including disease mutations)</td>
<td>6019871</td>
</tr>
<tr>
<td>Controlled vocabularies and ontology terms</td>
<td>252243</td>
</tr>
<tr>
<td>Total number of publications</td>
<td>495928</td>
</tr>
<tr>
<td>Links between protein entries and terms</td>
<td>3977548</td>
</tr>
</tbody>
</table>
Each protein entry has multiple views
Proteomics view

- Which MS/MS peptides from my protein have been found?
- Which PTM?
Where is this protein expressed, at mRNA and protein levels?

Expression view

Expression information is integrated from Bgee (MicroArray and EST) and from HPA (IHC protein and RNA-seq)
✓ Is this particular variant externally accessible or internal to the structure?
Searching for information in neXtProt
neXtProt SPARQL for beginners
www.nextprot.org RDF data model

A wide variety of experimental and predicted annotations

Extensive use of CV

Isoform-centric

Full traceability
Advanced search (SPARQL) interface 1/2

Retrieves lists of entries with associated information
Advanced search (SPARQL) interface 2/2

>140 pre-made queries available
SNORQL interface https://snorql.nextprot.org

Retrieves any data as a table

>170 pre-made queries available

What are the 25 most frequent families with member count

<table>
<thead>
<tr>
<th>familylabel</th>
<th>membercnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;G-protein coupled receptor 1 family&quot;</td>
<td>723</td>
</tr>
<tr>
<td>&quot;Krueppel C2H2-type zinc-finger protein family&quot;</td>
<td>546</td>
</tr>
<tr>
<td>&quot;Protein kinase superfamily&quot;</td>
<td>492</td>
</tr>
<tr>
<td>&quot;Small GTPase superfamily&quot;</td>
<td>162</td>
</tr>
<tr>
<td>&quot;Immunoglobulin superfamily&quot;</td>
<td>130</td>
</tr>
<tr>
<td>&quot;Peptidase S1 family&quot;</td>
<td>121</td>
</tr>
<tr>
<td>&quot;Major facilitator superfamily&quot;</td>
<td>99</td>
</tr>
</tbody>
</table>
A detailed help on neXtProt RDF entities is available
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

<table>
<thead>
<tr>
<th>Tags</th>
<th>federated</th>
</tr>
</thead>
<tbody>
<tr>
<td>NXQ_00094 - Proteins which are targets of antipsychotic drugs and highly expressed in brain</td>
<td>drug, expression, federated query, tutorial</td>
</tr>
<tr>
<td>NXQ_00096 - Proteins which are targets of drugs for cardiac therapy</td>
<td>drug, federated query, tutorial</td>
</tr>
<tr>
<td>NXQ_00139 - Protein kinases which are drug targets according to CHEMBL</td>
<td>CHEMBL, drug, federated query, tutorial</td>
</tr>
<tr>
<td>NXQ_00140 - Proteins that interact with viral proteins</td>
<td>federated query, interaction, PPI, tutorial, UniProt</td>
</tr>
<tr>
<td>NXQ_00141 - Human proteins highly expressed in brain and observed in a PDB structure involving a virus protein</td>
<td>expression, federated query, interaction, PDB, snore-only, tutorial, UniProt</td>
</tr>
<tr>
<td>NXQ_00246 - Proteins which are enzymes catalyzing a reaction involving lipids</td>
<td>enzyme, federated query, tutorial</td>
</tr>
<tr>
<td>NXQ_00253 - Human pathways in which at least one protein is mitochondrial GOLD</td>
<td>federated query, pathway, snore-only, subcellular location, tutorial</td>
</tr>
<tr>
<td>NXQ_00254 - Proteins with associated pathways in WikiPathways</td>
<td>federated query, pathway, snore-only, tutorial</td>
</tr>
</tbody>
</table>

And soon to come...
Federating neXtProt and WikiPathways
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)

<table>
<thead>
<tr>
<th>entry</th>
<th>gen</th>
<th>pathwayname</th>
<th>locs</th>
</tr>
</thead>
<tbody>
<tr>
<td>entry:NX_Q07812 (neXtProt link)</td>
<td>&quot;BAX&quot;^xsd:string</td>
<td>&quot;Nanomaterial induced apoptosis&quot;@en</td>
<td>&quot;BAX complex,Bcl-2 family protein complex,Cytoplasm,Mitochondrial membrane,cell periphery,cytoplasm,cytosol,endoplasmic reticulum membrane,mitochondrial outer membrane,mitochondrial transition pore complex,mitochondrion,nuclear envelope,nucleus,plasmalemma,plasmalemma vesicles,plasmalemma vesicle complex&quot;</td>
</tr>
<tr>
<td>entry:NX_Q14249 (neXtProt link)</td>
<td>&quot;ENDOG&quot;^xsd:string</td>
<td>&quot;Nanomaterial induced apoptosis&quot;@en</td>
<td>&quot;Mitochondrion,mitochondrial inner membrane,mitochondrion,nuclear envelope,plasmalemma,plasmalemma vesicles,plasmalemma vesicle complex&quot;</td>
</tr>
<tr>
<td>entry:NX_Q16611 (neXtProt link)</td>
<td>&quot;BAK1&quot;^xsd:string</td>
<td>&quot;Nanomaterial induced apoptosis&quot;@en</td>
<td>&quot;BAK complex,Mitochondrion outer membrane,Integral component of membrane,mitochondrial outer membrane,mitochondrial outer membrane,mitochondrion,pentamer,pentameric ring,pentameric ring complex,pentamer complex&quot;</td>
</tr>
<tr>
<td>entry:NX_P10415 (neXtProt link)</td>
<td>&quot;BCL2&quot;^xsd:string</td>
<td>&quot;Nanomaterial induced apoptosis&quot;@en</td>
<td>&quot;Endoplasmic reticulum membrane,Mitochondrion outer membrane,Nucleoplasm,Nucleus membrane,Integrin membrane,Integrin membrane complex&quot;</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>entry</th>
<th>callret-1</th>
<th>cpathways</th>
<th>drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>entry:NX_O15530 -- (neXtProt link)</td>
<td>&quot;PDPK1&quot;</td>
<td>&quot;Endometrial cancer&quot;</td>
<td>&quot;Celecoxib&quot;</td>
</tr>
<tr>
<td>entry:NX_P00352 -- (neXtProt link)</td>
<td>&quot;ALDH1A1&quot;</td>
<td>&quot;Folate-Alcohol and Cancer Pathway Hypotheses&quot;</td>
<td>&quot;Tretinoin&quot;</td>
</tr>
<tr>
<td>entry:NX_P00374 -- (neXtProt link)</td>
<td>&quot;DHFR&quot;</td>
<td>&quot;Retinoblastoma Gene in Cancer&quot;</td>
<td>&quot;Methotrexate,Pemetrexed&quot;</td>
</tr>
<tr>
<td>entry:NX_P00519 -- (neXtProt link)</td>
<td>&quot;ABL1&quot;</td>
<td>&quot;Apoptosis-related network due to altered Notch3 in ovarian cancer,Integrated Breast&quot;</td>
<td>&quot;Dasatinib,Imatinib&quot;</td>
</tr>
</tbody>
</table>
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

Web: https://www.nextprot.org/
Twitter: @neXtProt_news
ResearchGate: neXtProt project
A one-day tutorial on 9 independent SPARQL endpoints hosted by the SIB, including neXtProt, UniProt, GlyConnect, Rhea, OrthoDB, OMA and Bgee
Thank you!

E-mail: lydie.lane@sib.swiss