

The Proteasix Ontology

Mercedes Arguello Casteleiro¹, Julie Klein² and Robert Stevens¹

¹School of Computer Science, University of Manchester, Oxford road, Manchester, United Kingdom.

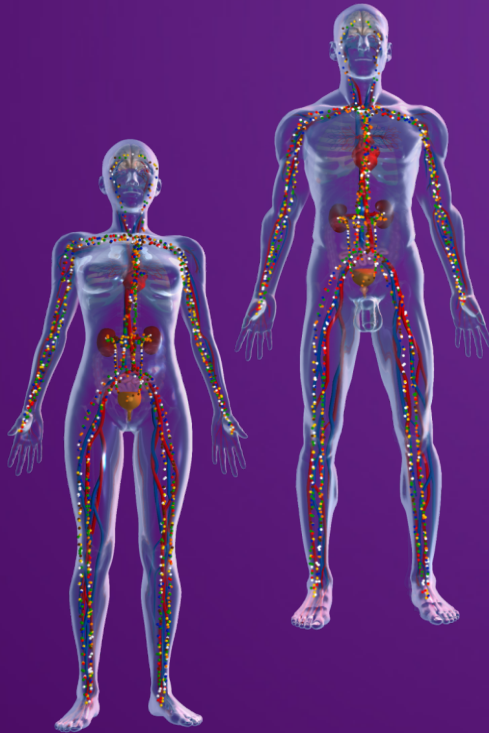
²Institut National de la Santé et de la Recherche Médicale (INSERM), U1048, Toulouse, France. And Université Toulouse III Paul-Sabatier, Toulouse, France.

Overview

- Biomedical background: Peptides as Biomarkers
- Current Tools & Applications
- DEMO 1 - Proteasix Web App
- Current limitations of Proteasix Web App
- Semantic Web technologies & Proteasix Ontology (PxO)
- Concluding remarks

why omics?

Omics: studying “all” molecules collectively



Complex diseases cannot be adequately described by **single features**:

- Interindividual differences
- Mechanism multiplicity

Exemplifying the challenges: CKD

Detect

Predict evolution

Renal biopsy
(kidney status)



But too invasive



Too late

Albuminuria
(glomerular filtration)



Not always present



Not sensitive enough

eGFR
(glomerular filtration)



Too late



Too late

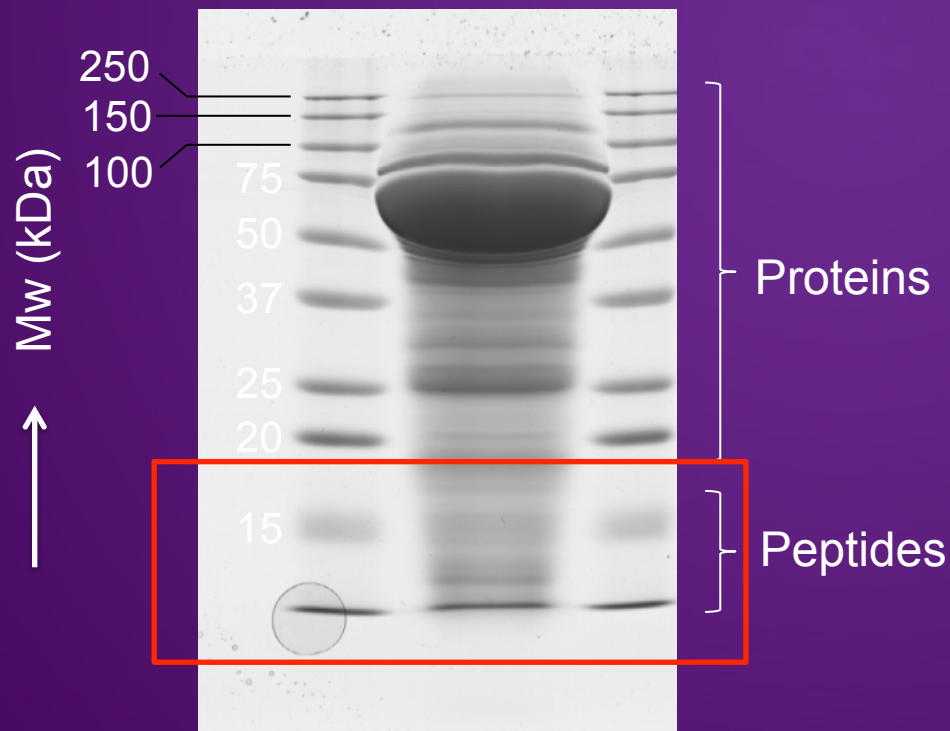
New biomarkers: non-invasive, early changes, kidney status

(To be used in addition/combination with gold-standards)

Peptides as Biomarkers

Urinary peptides

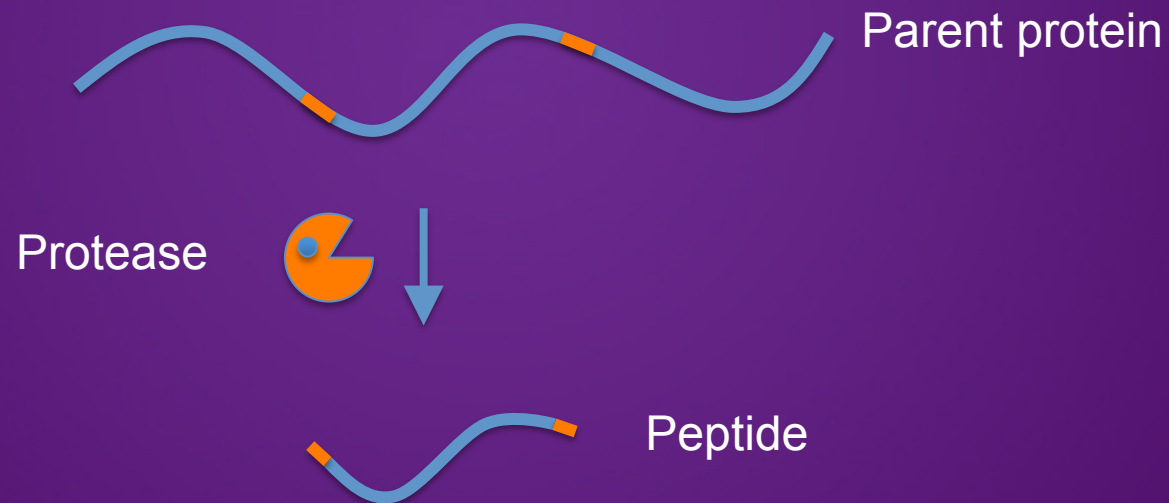
Peptide: proteolysis product (terminal post-translational modification)



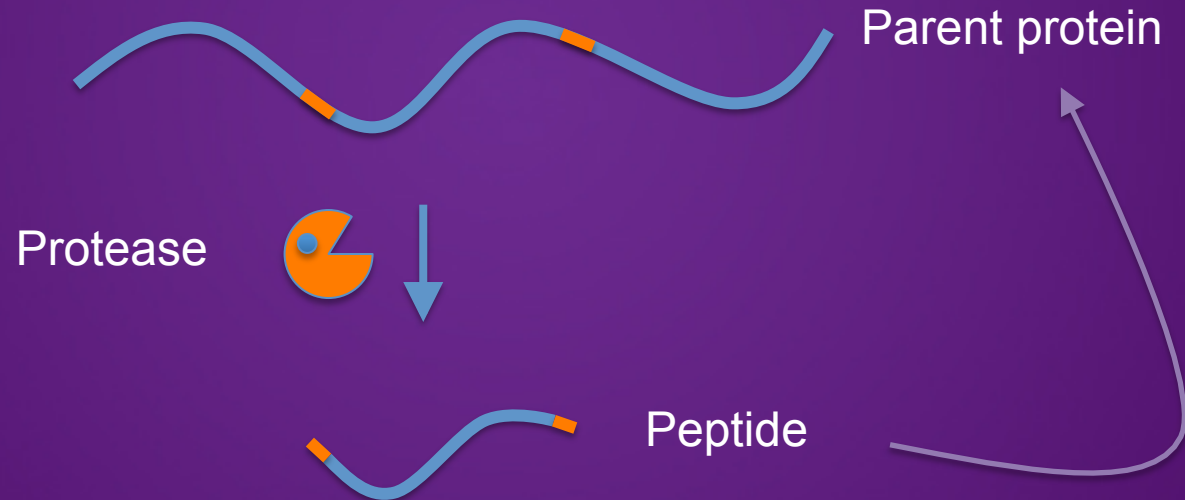
Courtesy: C. Lacroix, Toulouse

- Reduced pre-analytical handling before mass spectrometry analysis
no digestion
- Improved resistance to degradation
already degradation products
- Peptides are filtered under physiological conditions
detection of early events, before alteration of the filtration barrier

Links of urinary peptides to known processes in CKD

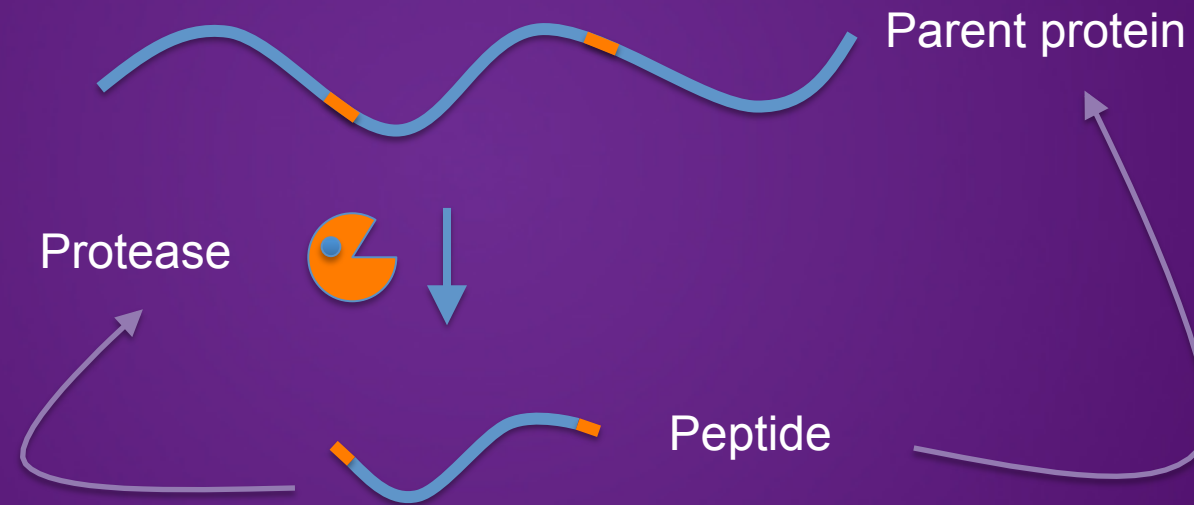


Links of urinary peptides to known processes in CKD



Parent protein is associated to pathophysiology?

Links of urinary peptides to known processes in CKD



Parent protein is associated to pathophysiology?

And/or

Protease activity is associated to pathophysiology?

Links of urinary peptides to known processes in CKD

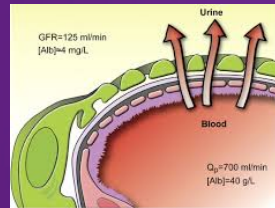
Parental proteins

Filtration plasma proteins

Serum albumin

Apo-A1

Beta-2-microglobulin



Inflammation

Osteopontin

Protein S100 A9

CD99



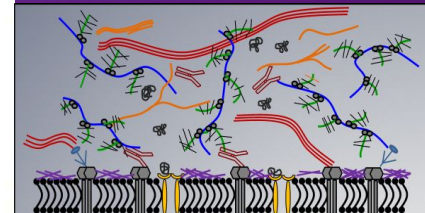
Tissue repair

Clusterin

Annexin A1



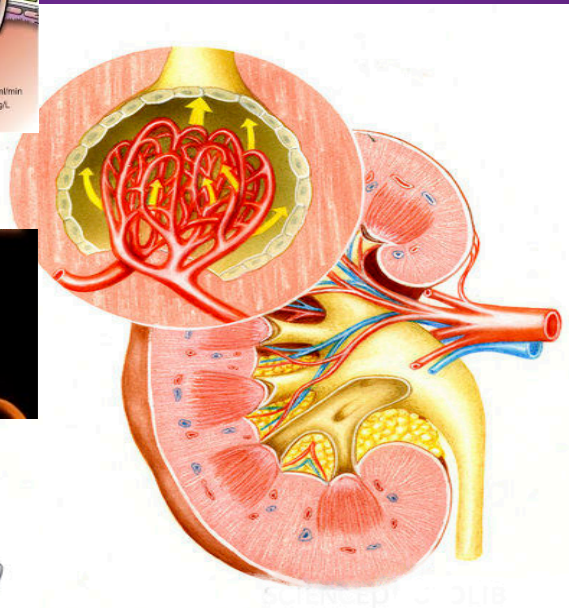
Aberrant expression of ECM (Fibrosis)



Collagen alpha I

Collagen alpha II

Collagen alpha III



Links of urinary peptides to known processes in CKD

Proteases

Protease identification: better understanding of disease mechanisms

Acute renal allograft rejection
Versus
Stable graft



Urine



Mass spectrometry



Peptide biomarkers



MMP8

IHC confirmation:
Neutrophils surrounding
capillaries

Clinical application

Mechanisms

Collagen degradation: inflammatory cell infiltration

AngII generation: macrophage recruitment

Sequence information of identified altered collagen $\alpha(I)$ and $\alpha(III)$ chain fragments in rejection samples suggested an involvement of matrix metalloproteinase-8 (MMP-8).

Links of urinary peptides to known processes in CKD

Proteases

Nomenclature:

- Cleavage site: 8 amino acid sequence (P4-P3-P2-P1 | P1'-P2'-P3'-P4')
- Scissile bond: peptide bond between P1 and P1'

Some proteases have very **broad specificity** (e.g. Trypsin X-X-X-R/K| X-X-X-X)

Some proteases have very **strong specificity** (e.g. Masp2 S-L-G-R| K-I-Q-I)

Some proteases have **exopeptidase activity**



Some proteases have **endopeptidase activity**



From peptides to enzymes? Cathepsin in Diabetic nephropathy

Smith et al. Hypertension 2012

- In the large conduit arteries, **elastin** is important in maintaining vascular compliance.
- In predialysis chronic kidney disease (CKD), **elastin degradation** is an important determinant of arterial stiffness and is associated with all-cause mortality.
- **Elastin degradation** is mediated by several **proteases**, including matrix metalloproteinase 2 and cathepsin S.
- Higher **matrix metalloproteinase 2** and **elastin-derived peptide levels** were also independently associated with preexisting cardiovascular disease.

From peptides to enzymes? Cathepsin in Diabetic nephropathy



Urinary peptides in DN



Proteasix



Some
math

(Holger Husi
Glasgow, UK)



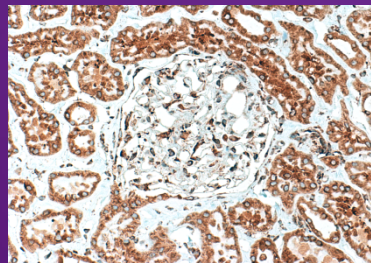
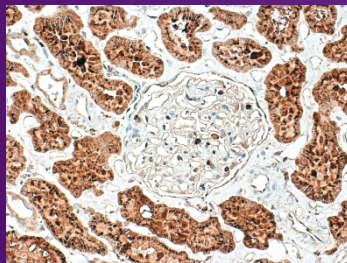
Top Predicted enzymes

↑Cathepsin S
↑ Cathepsin V (L2)

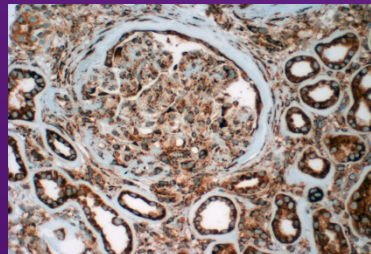
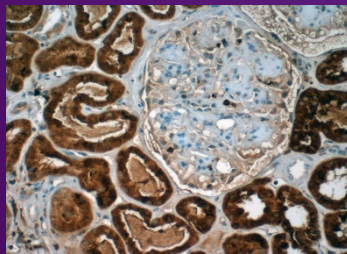
Cathepsin S

Cathepsin V (L2)

Zero
hour
biopsy



DN



Unpublished (Gert Mayer, Innsbruck, Austria)

Cathepsins

- Known roles in extracellular matrix (ECM) degradation.
- Cathepsin S increased in serum of CKD patients.

Smith et al. Hypertension 2012

- Cathepsin V (L2) ...??

Tools & Applications

Peptides represent a new source of very useful biomarkers in kidney and other disease.

Peptides can help better understanding the **pathophysiological mechanisms:**

Parental proteins

Proteases

Tools & Applications

Peptides represent a new source of very useful biomarkers in kidney and other disease.

Peptides can help better understanding the **pathophysiological mechanisms:**

Parental proteins

Proteases



Initially:
Very tedious work!

HELP ?

Tools & Applications

Peptides can help better understanding the
pathophysiological mechanisms

Initially:
Very tedious work!

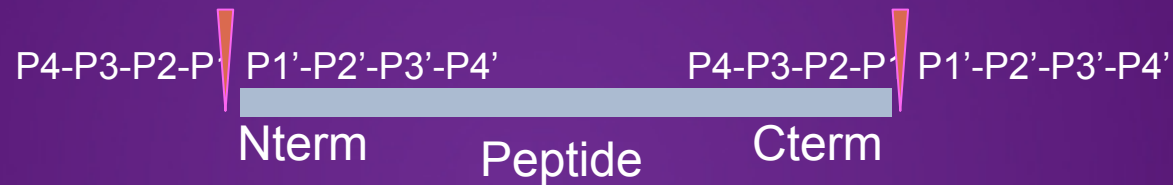


Let's get some HELP



Tools & Applications

What do we need?



- 2 cleavage sites/peptides (N and C Term)
- Cleavage sites are incomplete → reconstruct full cleavage sites
- Identify proteases that can cleave these cleavage sites

We need an **automatic tool** to reconstruct the cleavage sites

We need to collect information about **cleavage site/protease**

Tools & Applications

Existing **tools** for linking proteases, substrates, cleavage sites:

- CutDB (Proteolysis Map Project; PMAP)
- MEROPS
- TopFIND2
- Proteasix

CutDB

Database of protease/cleavage site

Not peptide-centric; No batch searches



PMAP-CutDB Proteolytic Event Database

Welcome to CutDB

The CutDB focuses on the annotation of individual proteolytic events, both actual and predicted.

Protease

Merops code

Organism:protease

Substrate

Organism:substrate

Disease

Search

ID	Protease	Organism [protease]	Substrate	Organism [substrate]	Cut-site	Structure	Update	Detail
<input type="checkbox"/> 18127	matrix metalloproteinase-2 M10.003	Gallus gallus	collagen, type I, alpha 1	Bos taurus	GPQG-IAGQ 952-953		May 25 2006 YI	Detail

MEROPS

Database of protease specificity

Not peptide-centric; No batch searches

Specificity matrix									
Amino acid	P4	P3	P2	P1	P1'	P2'	P3'	P4'	
Gly	252	121	430	482	60	125	590	290	
Pro	196	769	70	210	8	15	29	236	
Ala	253	368	642	506	110	235	538	332	
Val	199	447	129	45	280	444	274	272	
Leu	259	364	178	81	1389	390	112	297	
Ile	159	288	70	33	557	351	81	168	
Met	82	100	44	67	150	97	47	59	
Phe	128	81	48	58	177	143	100	126	
Tyr	97	51	48	54	131	91	61	53	
Trp	11	16	9	39	44	42	25	23	
Ser	178	116	483	333	79	229	486	235	
Thr	145	93	86	112	61	252	221	162	
Cys	37	42	51	39	70	24	11	40	
Asn	88	67	130	376	35	120	167	139	
Gln	166	79	226	152	129	220	128	149	
Asp	148	12	87	206	9	25	161	278	
Glu	169	42	161	266	33	123	147	241	
Lys	169	54	180	133	66	309	171	190	
Arg	138	26	192	105	13	99	26	38	
His	92	80	52	120	14	81	42	76	

WELCOME

TopFIND is the first public knowledgebase and analysis resource for protein termini and protease processing

More than 290,000 N- and C-termini and more than 33,000 cleavages listed

Covers H. sapiens, M. musculus, S. cerevisiae, A. thaliana and E. coli

Database

WHAT TopFIND PROVIDES

Integration of protein termini & function with proteolytic processing, alternative transcription & translation

Displays proteases and substrates within their protease web including detailed evidence information

TopFINDER automates analysis and functional annotation of proteomics-derived termini sets

PathFINDER identifies indirect protease-substrate connections for the evaluation of complex processes

Tool

MANCHESTER
1824

The University of Manchester

Proteasix Web-based version

Content:

Human proteases (230), mouse proteases (160), rat proteases (105)
>20000 protease/cleavage site association (CutDB, MEROPS)
318 protease specificity matrices for prediction (MEROPS)

Proteasix Web-based version

Content:

Human proteases (230), mouse proteases (160), rat proteases (105)
>20000 protease/cleavage site association (CutDB, MEROPS)
318 protease specificity matrices for prediction (MEROPS)

Functionality:

Visualisation of protease specificity
Cleavage site reconstruction (based on UniprotKB)
Prediction of proteases

MANCHESTER
1824

The University of Manchester



Proteasix Web-based version

<http://sysvasc.cs.man.ac.uk>

Current Limitations of Proteasix Web App

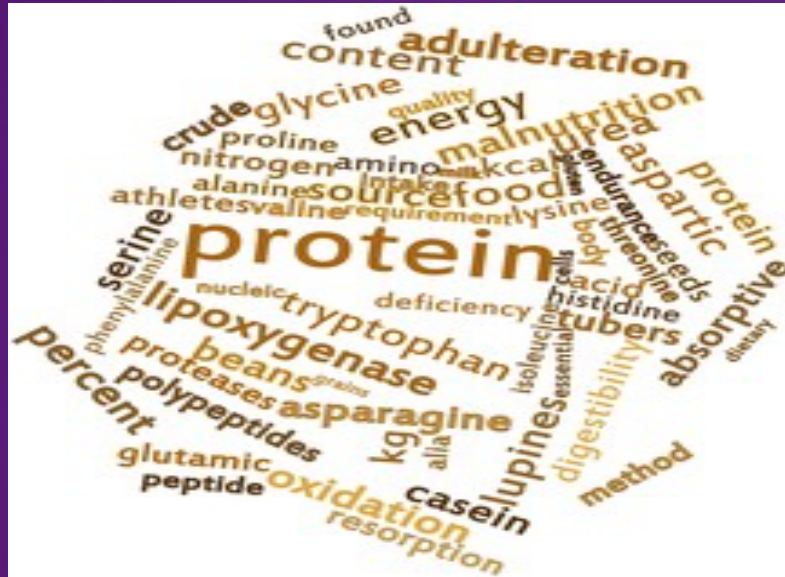
- To refine the predictions, we need to know:
 - the species in which our proteins exist
 - the function of the protease
 - where the proteins live, i.e. which cellular component
- At the moment there is NO distinction between
 - Exopeptidase activity**
 - Endopeptidase activity**

Current Limitations of Proteasix Web App

- To refine the predictions, we need to know:
 - the species in which our proteins exist
 - the function of the protease
 - where the proteins live, i.e. which organelles
- At the moment there is a gap between
 - Exopeptidases**
 - Endopeptidases**

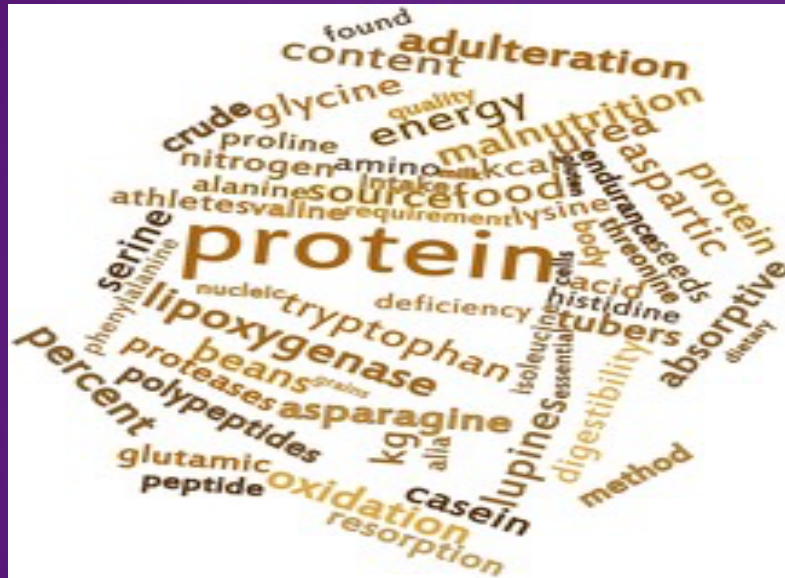
We need to incorporate
more knowledge

Ontology



In informatics and computer science, an **ontology** is a representation of the **shared background knowledge** for a community.

Ontology



In informatics and computer science, an **ontology** is a representation of the **shared background knowledge** for a community.

UniProtKB - P02768 (ALBU_HUMAN)

Protein | Serum albumin

Gene | ALB

Organism | *Homo sapiens* (Human)

MANCHESTER
1824

The University of Manchester

The **goal** with a computer science ontology
is to make **knowledge of a domain**
computationally useful

UniProtKB - P02768 (ALBU_HUMAN)

Protein | **Serum albumin**

Gene | **ALB**

Organism | *Homo sapiens (Human)*

MANCHESTER
1824

The University of Manchester

The **goal** with a computer science ontology is to make **knowledge of a domain computationally useful.**

UniProtKB - P02768 (ALBU_HUMAN)

Protein | Serum albumin

Gene | ALB

Organism | *Homo sapiens (Human)*

An **ontology** is a description of the **concepts** and **relationships** that can exist

Categories or classes:

Protein

Gene

Organism

The **goal** with a computer science ontology is to make **knowledge of a domain computationally useful.**

UniProtKB - P02768 (ALBU_HUMAN)

Protein | **Serum albumin**

Gene | **ALB**

Organism | *Homo sapiens (Human)*

An **ontology** is a description of the **concepts** and **relationships** that can exist

Categories or classes:

Protein

Gene

Organism

Relationships:

P02768 is-a protein

ALB is-a gene

P02768 only_in_taxon Human

Current Work: *Proteasix Ontology* (PxO)

Concepts

The screenshot shows a web browser window displaying the Proteasix Ontology (PxO) class hierarchy. The browser address bar shows the URL: <http://sysvasc.cs.man.ac.uk/formats/proteasixInOwl/proteasix.owl>. The browser tabs include 'Active Ontology', 'Entities', 'Classes', 'Object Properties', 'Data Properties', and 'Annotation Properties'. The main content area displays the 'Class hierarchy (inferred): polypeptide_region'.

- Thing
 - cellular_component
 - 'chemical entity'
 - group
 - 'molecular entity'
 - 'amino acid chain'
 - protein
 - 'proteolytic cleavage product'
 - 'Amino acid entity'
 - 'Amide amino acid'
 - 'Charged amino acid'
 - 'Neutrally charged amino acid'
 - 'Nucleophilic amino acid'
 - 'main group molecular entity'
 - 'Database entry'
 - 'BRENDA database entry'
 - 'MEROPS database entry'
 - 'UniProt Knowledge Base database entry'
 - molecular_function
 - 'catalytic activity'
 - organism
 - Archaea
 - Bacteria
 - Eukaryota
 - 'other sequences'
 - 'unclassified sequences'
 - Viruses
 - process
 - quality
 - sequence_feature
 - gene
 - region
 - biological_region
 - polypeptide_region
 - 'C-terminus region'
 - 'Cleavage site region'
 - 'N-terminus region'

Current Work: *Proteasix* Ontology (PxO)

pro (http://sysvasc.cs.man.ac.uk/formats/proteasixInOwl/proteasix.owl)

Active Ontology x Entities x Classes x Object Properties x Data Properties x Annotation Properties x Individuals by class x VOWL x

Class hierarchy Class hierarchy (inferred)

Class hierarchy (inferred): 'peptidase activity'

- Thing
 - cellular_component
 - 'chemical entity'
 - 'Database entry'
 - molecular_function
 - 'catalytic activity'
 - 'hydrolase activity'
 - 'peptidase activity'**
 - 'isopeptidase activity'
 - 'oligopeptidase activity'
 - 'peptidase activity, acting on D-amino acid peptides'
 - 'peptidase activity, acting on L-amino acid peptides'
 - organism
 - process
 - quality
 - sequence_feature

Class Annotations Class Usage

Annotations: 'peptidase activity'

Annotations +

- label [type: string]
peptidase activity
- id [type: string]
GO:0008233
- has_obo_namespace [type: string]
molecular_function
- definition [type: string]
Catalysis of the hydrolysis of a peptide bond. A peptide bond is formed from the carboxyl group of one amino acid shares electron with the second amino acid.
- has_exact_synonym [type: string]
hydrolase, acting on peptide bonds
- has_exact_synonym [type: string]
peptide hydrolase activity
- has_exact_synonym [type: string]
protease activity
- has_exact_synonym [type: string]

Description: 'peptidase activity'

Equivalent To +

SubClass Of +

- 'hydrolase activity'
- 'part of' some proteolysis

General class axioms +

SubClass Of (Anonymous Ancestor)

Current Work: *Proteasix* Ontology (PxO)

pro (http://sysvasc.cs.man.ac.uk/formats/proteasixInOwl/proteasix.owl)

Active Ontology x Entities x Classes x Object Properties x Data Properties x Annotation Properties x Individuals by class x VOWL x

Class hierarchy Class hierarchy (inferred)

Class hierarchy (inferred): 'peptidase activity'

Annotations: 'peptidase activity'

Annotations +

label [type: string]
peptidase activity

id [type: string]
GO:0008233

has_obo_namespace [type: string]
molecular_function

definition [type: string]
Catalysis of the hydrolysis of a peptide bond. A peptide bond is formed from the carboxyl group of one amino acid shares electron with the second amino acid.

has_exact_synonym [type: string]
hydrolase, acting on peptide bonds

has_exact_synonym [type: string]
peptide hydrolase activity

has_exact_synonym [type: string]
protease activity

has_exact_synonym [type: string]

Description: 'peptidase activity'

Equivalent To +

SubClass Of +

- 'hydrolase activity'
- 'part of' some proteolysis

General class axioms +

SubClass Of (Anonymous Ancestor)


Concepts

Relationships

Proteasix Ontology (PxO) reuse ontologies

www.obofoundry.org

The OBO Foundry


[About](#) ▾ [Principles](#) ▾ [Ontologies](#) ▾ [Participate](#) ▾ [FAQ](#) ▾ [Legacy](#) ▾

The OBO Foundry

Welcome to the new OBO website! See the [Announcement](#) for more info.

Download table as: [[YAML](#) | [JSON-LD](#) | [RDF/Turtle](#)]

chebi	Chemical Entities of Biological Interest	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds. Detail							
doid	Human Disease Ontology 	An ontology for describing the classification of human diseases organized by etiology. Detail							
go	Gene Ontology 	An ontology for describing the function of genes and gene products Detail							
obi	Ontology for Biomedical Investigations 	An integrated ontology for the description of life-science and clinical investigations Detail							

MANCHESTER
1824

The University of Manchester

Proteasix Ontology (PxO) reuse ontologies

chemical entity
molecular entity
Alanine

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾

Search Ontobee

Submit

Chemical Entities of Biological Interest

A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds.

 Follow @chebit 265 followers

OntoBee

AberOWL

OLS

CHEBI

A structured classification of chemical compounds of biological relevance.

Products

chebi.owl

chebi.obo

ID Space chebi

PURL <http://purl.obolibrary.org/obo/chebi.owl>

Proteasix Ontology (PxO) reuse ontologies

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾

NCBI organismal classification

An ontology representation of the NCBI organismal taxonomy

The NCBITaxon ontology is an automatic translation of the [NCBI taxonomy database](#) into obo/owl.

The translation treats each taxon as an obo/owl class whose instances (for most branches of the ontology) would be individual organisms. For example:

```
'Craig Venter' instance_of NCBITaxon_9606 (Homo sapiens)
```

The translation faithfully reproduces all of the content of the source database, even where this contravenes OBO guidelines. For example:

- The root class is called 'root', rather than something like 'organism'
- Plural names are used (both Linnaean and common names). E.g. "mammals"
- The organismhood of certain classes might be contested - either biologically ("viruses") or ontologically ("environmental samples")
- Synonyms may include quotation marks as part of the text

PURLs

The purls for this ontology are:

- <http://purl.obolibrary.org/obo/ncbitaxon.owl> (official purl for *ontology*)

Human

Rat

Mouse

Proteasix Ontology (PxO) reuse ontologies

molecular_function

cellular_component

biological_process

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾



Gene Ontology

An ontology for describing the function of genes and gene products

Follow @news4go { 759 followers }

OntoBee

AberOWL

OLS

AmiGO

The goal of the GeneOntology (GO) project is to provide a uniform way to describe the functions of gene products from organisms across all kingdoms of life and thereby enable analysis of genomic data

Products

go.owl

go/extensions/go-plus.owl

GO-Plus

The core ontology plus axioms connecting to select external ontologies

go/extensions/go-taxon-groupings.owl

GO Taxon Groupings

Classes added to ncbitaxon for groupings such as prokaryotes

ID Space

go

PURL

<http://purl.obolibrary.org/obo/go.owl>

License

CC-BY

Proteasix Ontology (PxO) reuse ontologies

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾

Search Ontobee

Submit

PRotein Ontology (PRO)

an ontological representation of protein-related entities

OntoBee

AberOWL

OLS

PRotein Ontology (PRO) has been designed to describe the relationships of proteins and protein evolutionary classes (ontology for ProEvo), to delineate the multiple protein forms of a gene locus (ontology for protein forms), and to interconnect existing ontologies

Products

pr.owl

ID Space pr

PURL <http://purl.obolibrary.org/obo/pr.owl>

Protein

amino acid chain

proteolytic cleavage
product

The term “**Semantic Web**” refers to W3C’s vision of the *Web of linked data*

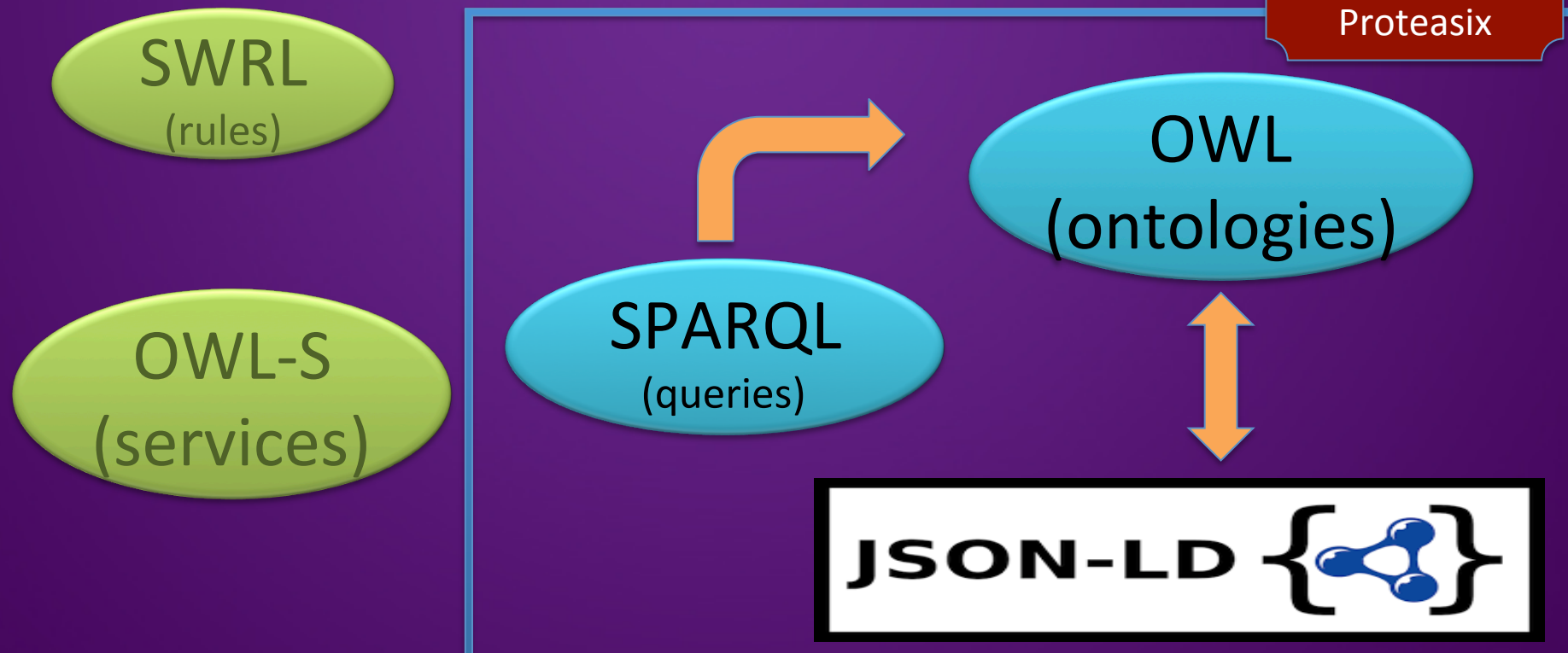


Ontologies are considered one of the pillars of the **Semantic Web**

Semantic Web
technologies
has reached to a
degree of maturity



Semantic Web technologies



Current Work: *Proteasix Ontology (PxO)*

- UniProt KB proteins (Swiss-prot and TrEMBL)
 - ❑ Organised by Taxons following PRO ontology
 - ❑ Annotated with GO biological_process; GO cellular_component; and GO molecular_function
- Model cleavage sites patterns for
 - ❑ **Exopeptidase activity**
 - ❑ **Endopeptidase activity**

For using Peptides as Biomarkers,
we need
more data.. and data linkage..

Semantic Web technologies

OWL
(ontologies)

Let's glue the
relevant knowledge
together



Class: polypeptide_region

SubClassOf:

biological_region

part_of some protein,

associated_with some 'proteolytic cleavage product',

only_in_taxon some organism

For using Peptides as Biomarkers,
we need
more data.. and data linkage..

In Swiss-Prot there are proteins annotated with GO:0004252,
which is serine-type endopeptidase activity

Semantic Web technologies

SPARQL
(queries)

Let's get them!

```
SELECT ?x FROM <file:./ontofiles/Swiss-Prot.owl>
{
  ?x rdf:type owl:Class;
    rdfs:subClassOf [ a owl:Restriction ;
                    owl:onProperty pr:has_function ;
                    owl:someValuesFrom obo:GO_0004252 ] .
}
```



For using Peptides as Biomarkers,
we need
more data.. and data linkage..

Can we retrieve from Swiss-Prot ALL the proteins annotated
with any DESCENDANT of Peptidase activity GO_0008233 ?

Semantic Web technologies

SPARQL
(queries)

Let's get Peptidase Activities from GO,
e.g. serine-type endopeptidase activity GO:0004252

```
GRAPH <file:./ontofiles/  
GO_TAXONOMY_molecular_function.owl>  
{?C rdfs:subClassOf+ obo:GO_0008233 .} .
```



For using Peptides as Biomarkers,
we need
more data.. and data linkage..

Can we retrieve from Swiss-Prot ALL the proteins annotated with any DESCENDANT of Peptidase activity GO_0008233 ?

Semantic Web technologies

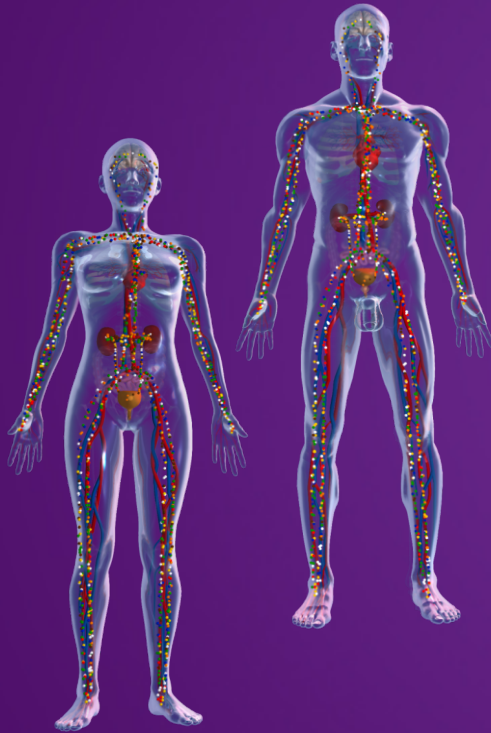
SPARQL
(queries)

YES! let's get proteins annotated with Peptidase Activities from GO

```
SELECT ?x FROM <file:./ontofiles/Swiss-Prot.owl>
FROM NAMED <file:./ontofiles/GO_TAXONOMY_molecular_function.owl>
{
  GRAPH <file:./ontofiles/GO_TAXONOMY_molecular_function.owl>
    {?C rdfs:subClassOf+ obo:GO_0008233 .} .
  {
    ?x rdf:type owl:Class;
      rdfs:subClassOf [ a owl:Restriction ;
        owl:onProperty pr:has_function ;
        owl:someValuesFrom ?C ] .
  }
}
```



Concluding remarks



TopFIND2 and Proteasix can help to automatically predict modification of protease activity

Current work: exploring the benefits of using Proteasix ontology (PxO) in the Semantic-Web version of Proteasix

Hard Limits:

- Prediction always need validation
- Data available (over-representation of some proteases, other are missing)

ACKNOWLEDGMENT: to Julie Klein for providing

- ❑ Biomedical background
- ❑ Comparison of Current Tools & Applications

