

UniMed: Mapping protein information to disease terminologies

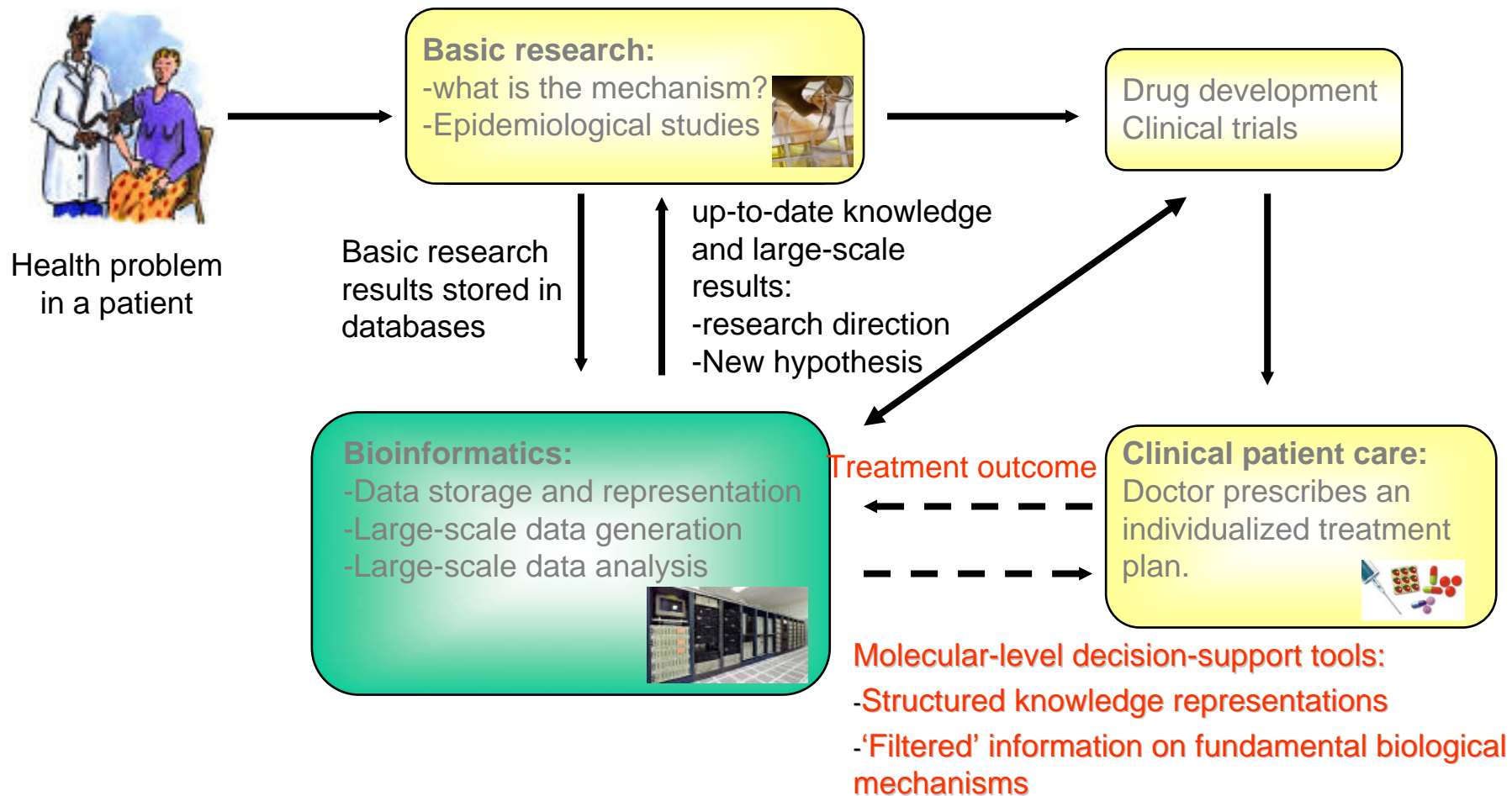
Anne-Lise Veuthey, Swiss Institute of Bioinformatics



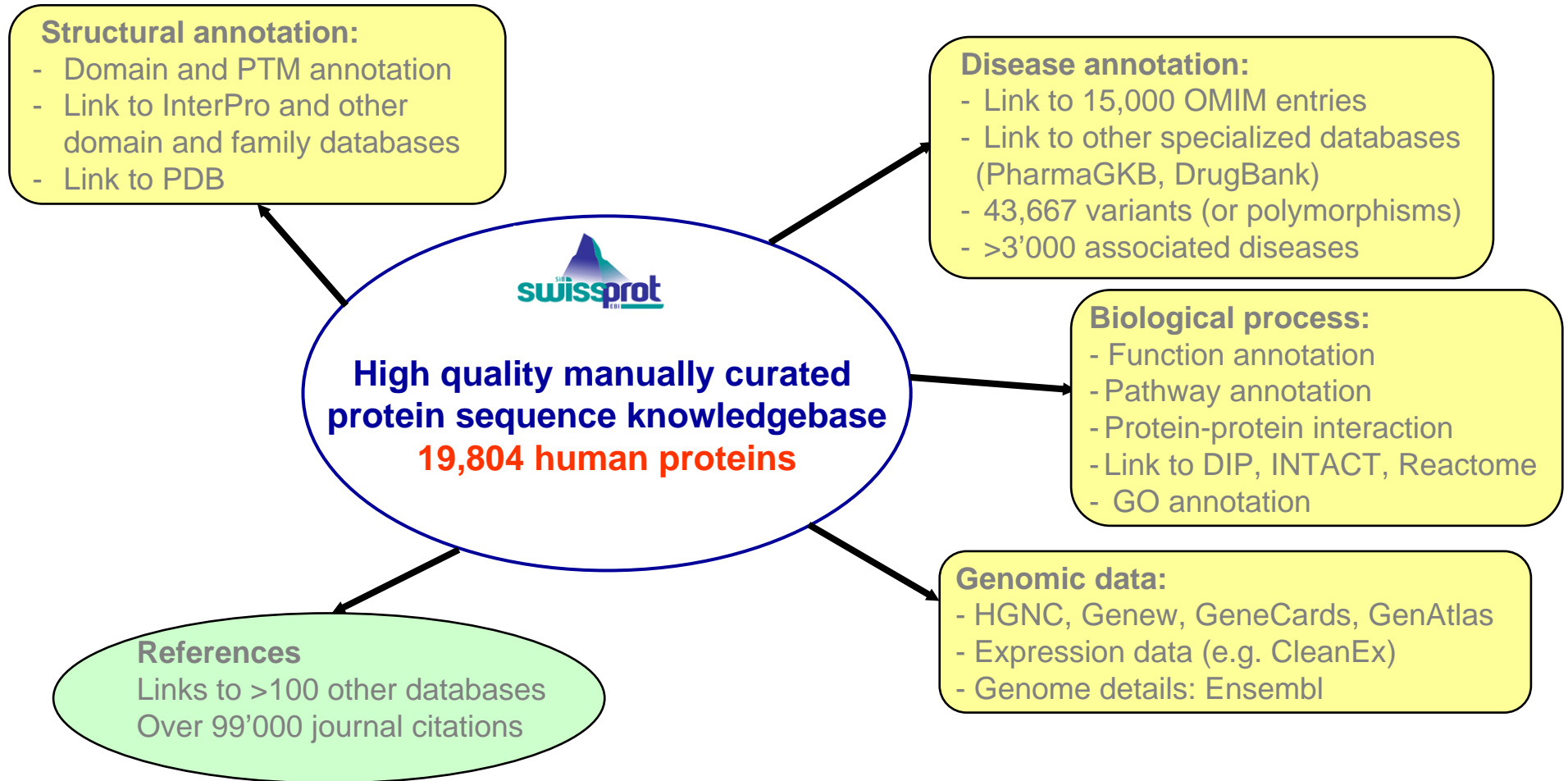
Outline

- Bioinformatics in translational research
- UniProtKB/Swiss-Prot
 - A hub to biological resources
 - Disease-related information content
- UniMed
 - Project objectives
 - Terminology mapping
 - Citation mapping
 - Discussion and work in progress

The role of bioinformatics in biomedical research and future clinical patient care



UniProtKB/Swiss-Prot



Disease information in UniProtKB/Swiss-Prot

<http://beta.uniprot.org/>

UniProtKB UniProtKB Downloads · Contact · Help

★ Reviewed, UniProtKB/Swiss-Prot **P00480** **#311250** GeneTests, Links
 Last modified May 20, 2008. Version 119. [History...](#)

Names and origin	
Protein names	Ornithine carbamoyltransferase <i>Also known as:</i> EC 2.1.3.3 OTCase Ornithine transcarbamoylase
Gene names	Name: OTC
Organism	Homo sapiens (Human)
General annotation (Comments)	
Catalytic activity	Carbamoyl phosphate + L-glutamine → L-glutamate + carbamoyl-L-glutamate
Pathway	Nitrogen metabolism 1/1.
Subunit structure	Homotrimer.
Subcellular location	Mitochondrion matrix.
Tissue specificity	Mainly in liver and intestinal mucosa
Involvement in disease	Defects in OTC are the cause of ornithine carbamoyltransferase deficiency (OTCD) [MIM: 311250]. OTCD is an X-linked disorder of the urea cycle which causes a form of hyperammonemia. Mutations with no residual enzyme activity are always expressed in hemizygote males by a very severe neonatal hyperammonemic coma that generally proves to be fatal. Heterozygous females are either asymptomatic or express orotic aciduria spontaneously or after protein intake. The disorder is treatable with supplemental dietary arginine and low protein diet. The arbitrary classification of patients into the "neonatal" group (clinical hyperammonemia in the first few days of life) and "late" onset (clinical presentation after the neonatal period) has been used to differentiate severe from mild forms.
Sequence similarities	Belongs to the ATCase/OTCase family .

Alternative titles; symbols

ORNITHINE CARBAMOYLTRANSFERASE DEFICIENCY
 OTC DEFICIENCY
 VALPROATE SENSITIVITY, INCLUDED

Gene map locus [Xp21.1](#)

TEXT

A number sign (#) is used with this entry because the disorder is caused by mutation in the gene encoding ornithine carbamoyltransferase (OTC; [300461](#)).

DESCRIPTION

Ornithine transcarbamylase deficiency is an X-linked inborn error of metabolism of the urea cycle which causes hyperammonemia. The disorder is treatable with supplemental dietary arginine and low protein diet.

Urea cycle disorders are characterized by the triad of hyperammonemia, encephalopathy, and respiratory alkalosis. Five disorders involving different defects in the biosynthesis of the enzymes of the urea cycle have been described: OTC deficiency, carbamyl phosphate synthetase deficiency ([237300](#)), argininosuccinate synthetase deficiency, or citrullinemia ([215700](#)), argininosuccinate lyase deficiency ([207900](#)), and arginase deficiency ([207800](#)).

Variant information

★ Reviewed, UniProtKB/Swiss-Prot **P00480** (OTC_HUMAN)
Last modified May 20, 2008. Version 119. [History...](#)

[Contribute](#)
[Send feedback](#)

Names and origin Hide | Top

Protein names	Ornithine carbamoyltransferase, mitochondrial [Precursor] <i>Also known as:</i> EC 2.1.3.3 OTCase Ornithine transcarbamylase
Gene names	Name: OTC
Organism	Homo sapiens (Human)

Sequence annotation (Features) Hide | Top

	Feature key	Position(s)	Length	Description	Graphical view
<input type="checkbox"/>	Natural variant	320	1	R → L in OTCD.	
<input type="checkbox"/>	Natural variant	326	1	E → K in OTCD.	
<input type="checkbox"/>	Natural variant	330	1	R → G in OTCD.	
<input type="checkbox"/>	Natural variant	333	1	T → A	
<input type="checkbox"/>	Natural variant	336	1	A → S in OTCD; late onset.	
<input type="checkbox"/>	Natural variant	337	1	V → L in OTCD; late onset.	
<input type="checkbox"/>	Natural variant	339	1	V → L in OTCD; neonatal.	
<input type="checkbox"/>	Natural variant	340	1	S → P in OTCD; late onset.	
<input type="checkbox"/>	Natural variant	341	1	L → P in OTCD.	
<input type="checkbox"/>	Natural variant	343	1	T → K in OTCD; late onset.	

Information access

- **UniProtKB/Swiss-Prot** contains information related to protein involvement in pathologies and provides access to many high quality resources with biological or medical relevance
- However this information is not easily accessible for medical-oriented queries.
- Two initiatives to resolve this issue:
 - Disease name standardisation
 - UniMed

UniMed

- Research project financed by the Swiss National Science Foundation
- Goal: increase the interoperability between molecular biology and clinical resources by:
 - **1st step:** indexing **UniProtKB/Swiss-Prot** with the medical terminologies **MeSH** and **ICD-10**
 - **2nd step:** building a search engine to improve the accessibility of variant pages of **UniProtKB/Swiss-Prot**
 - **3rd step:** providing a knowledge management system for biomedical data integration

Context

- Several systems were built in order to link **phenotypes to genotypes**:
 - GenesTrace (Cantor et al.),
 - BioMeKe (Marquet et al.)
 - PhenoGO (Lussier et al.)
 - MedGene DB (LaBaer)
 - GFINDER (Masseroli et al.)
- These systems use:
 - Text mining methods (NER, NLP, IR techniques)
 - Knowledge- and semantic based methods using ontological relationships of terminologies
- Our goal is to provide a framework which helps mine biological information to discover medical meaning

Project objective

Neurodegenerative Diseases

- PEX7 (O00628)
- TOR1A (O14656)
- TPP1 (O14773)
- CPLX1 (O14810)
- SPTBN2 (O15020)
- OPA1 (O60313)
- KIF1B (O60333)
- WFS1 (O76024)
- CDKL5 (O76039)
- ELP1 (O95163)
- EPM2A (O95278)
- VAPB (O95292)
- COX1 (P00395)
- HPTR1 (P00492)
- CRYAB (P02511)
- APOE (P02649)
- ...

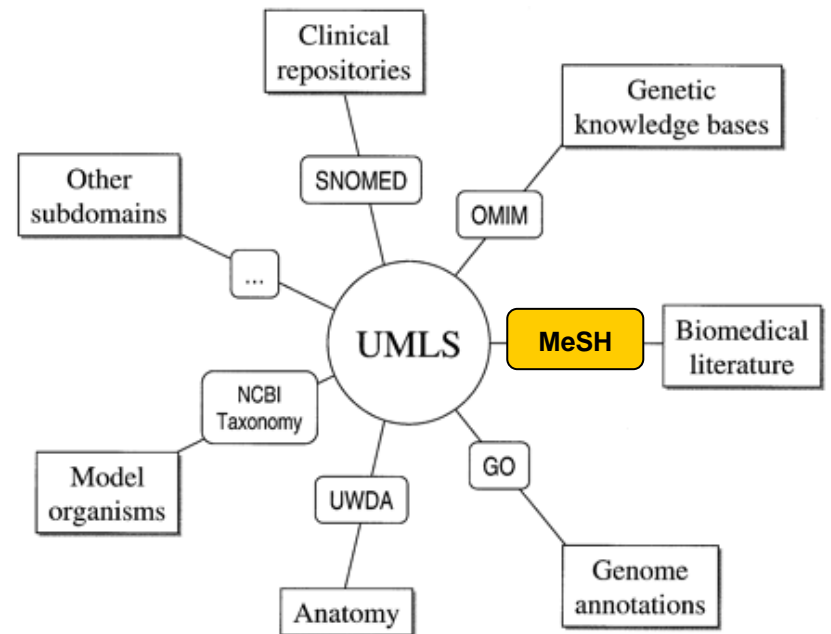
MeSH

refsum disease
dystonia musculorum deformans
neuronal ceroid-lipofuscinoses
parkinson disease
spinocerebellar ataxias
optic atrophy, autosomal dominant
charcot-marie-tooth disease
wolfram syndrome
rett syndrome
dysautonomia, familial
myoclonic epilepsies, progressive
amyotrophic lateral sclerosis
optic atrophy, hereditary, leber
lesch-nyhan syndrome
alexander disease
alzheimer disease

135 human proteins involved

Why MeSH?

- Controlled vocabulary thesaurus structured in a **hierarchy** of concepts
- Each concept includes a set of terms -**synonyms** and **lexical variants**
- MeSH is part of the **UMLS**, and, thus, linked to other medical terminologies
- MeSH is used to index the **biomedical literature**



The structure of MeSH

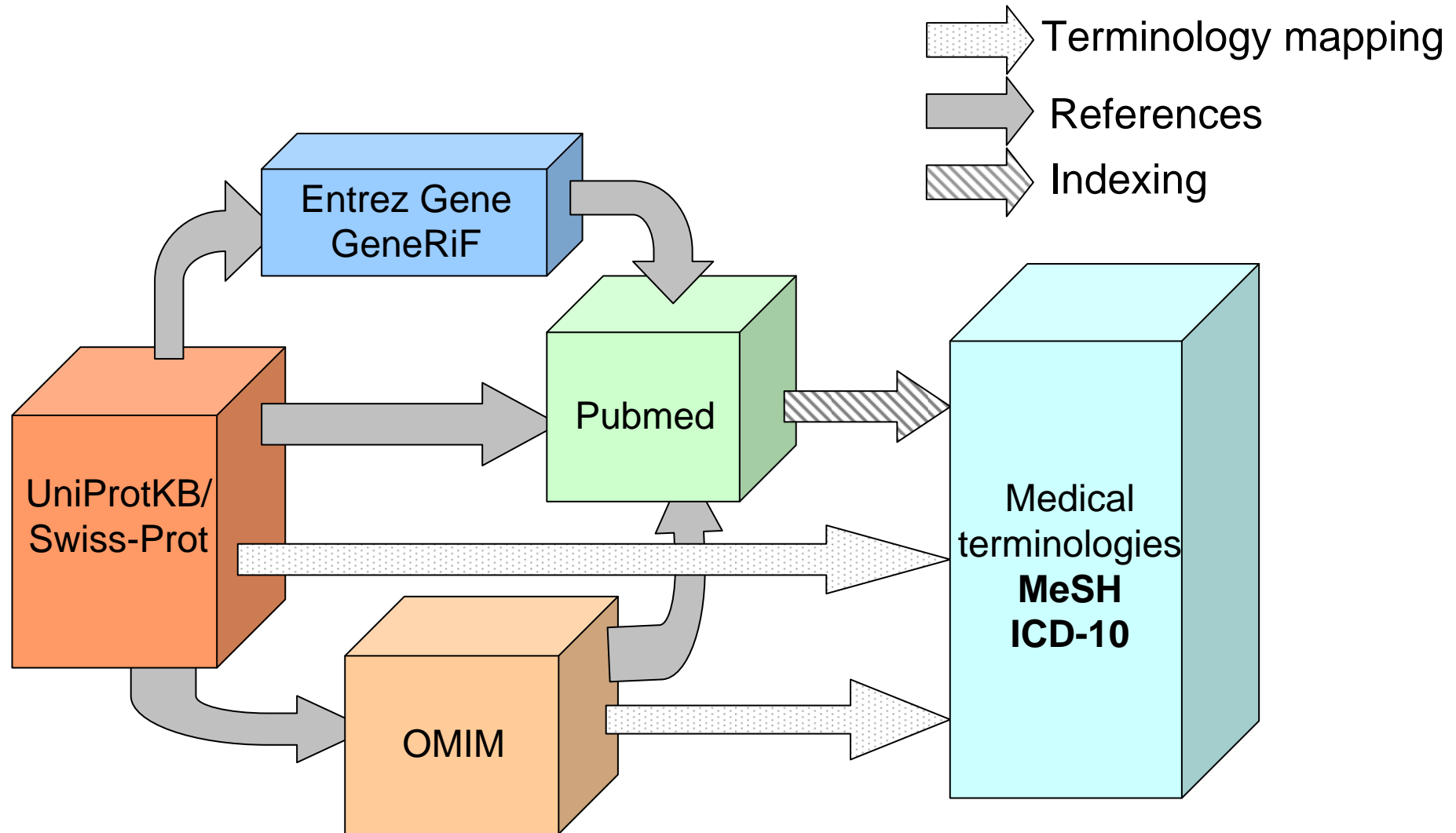
MeSH Heading	Parkinson Disease
Tree Number	C10.228.140.079.862.500
Tree Number	C10.228.662.600.400
Tree Number	C10.574.812
Annotation	drug ther: consider also ANTIPARKINSON AGENTS ; / chem ind = PARKINSON DISEASE , SECONDARY/chem ind
Concept 1 (Preferred)	Parkinson Disease
Scope Note	A progressive, degenerative neurologic disease characterized by a TREMOR that is maximal at rest, retropulsion (i.e. a tendency to fall backwards), rigidity, stooped posture, slowness of voluntary movements, and a masklike facial expression. Pathologic features include loss of melanin containing neurons in the substantia nigra and other pigmented nuclei of the brainstem. LEWY BODIES are present in the substantia nigra and locus coeruleus but may also be found in a related condition (LEWY BODY DISEASE, DIFFUSE) characterized by dementia in combination with varying degrees of parkinsonism. (Adams et al., Principles of Neurology, 6th ed, p1059, pp1067-75)
Term	Parkinson Disease
Term	Idiopathic Parkinson Disease
Term	Idiopathic Parkinson's Disease
Term	Lewy Body Parkinson Disease
Term	Lewy Body Parkinson's Disease
Term	Paralysis Agitans
Term	Parkinson Disease, Idiopathic
Term	Parkinson's Disease
Term	Parkinson's Disease, Idiopathic
Term	Parkinson's Disease, Lewy Body
Term	Primary Parkinsonism

[Nervous System Diseases \[C10\]](#)

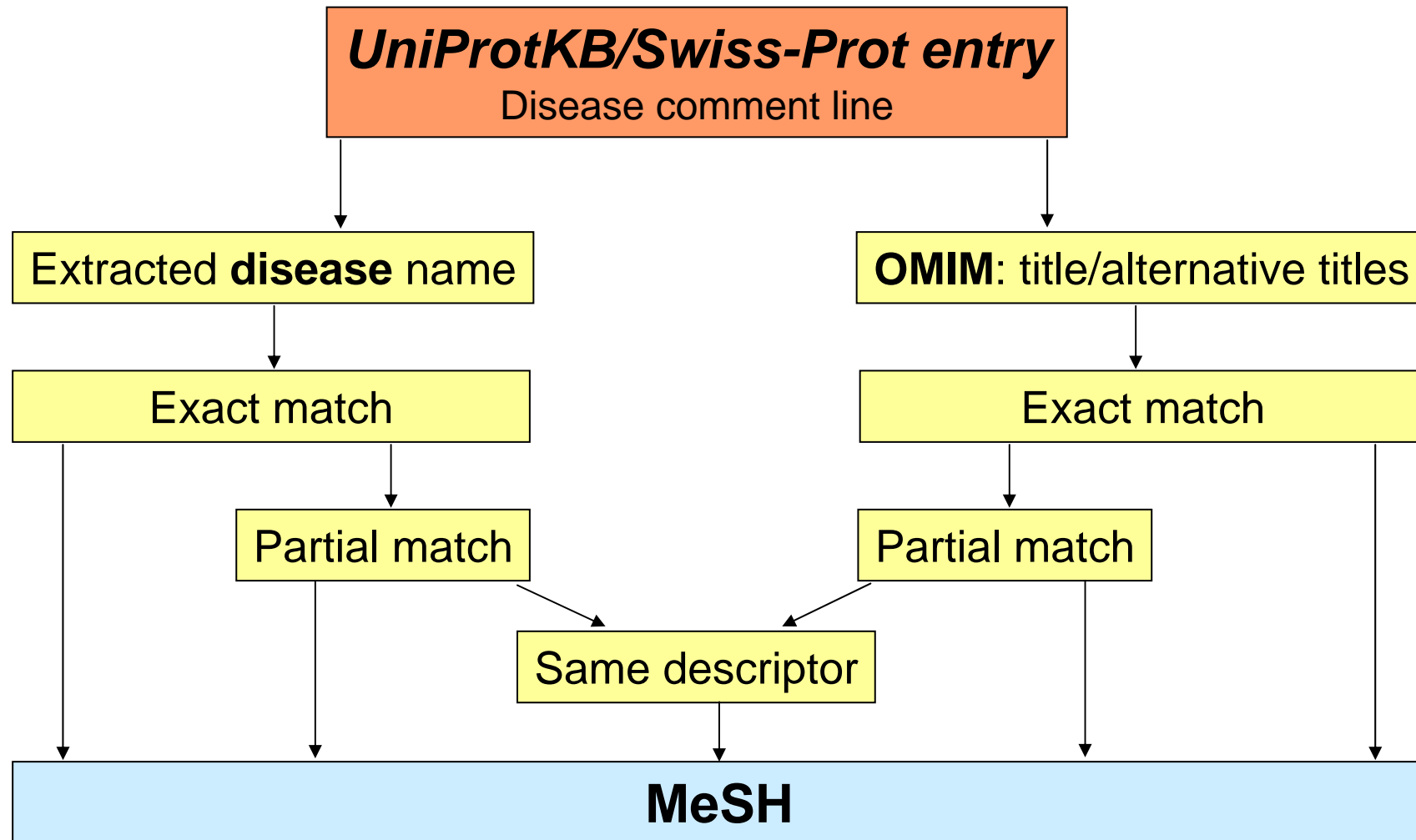
[Neurodegenerative Diseases \[C10.574\]](#)

[Heredodegenerative Disorders, Nervous System \[C10.574.500\]](#) +
[Lewy Body Disease \[C10.574.531\]](#)
[Motor Neuron Disease \[C10.574.562\]](#) +
[Multiple System Atrophy \[C10.574.625\]](#) +
[Olivopontocerebellar Atrophies \[C10.574.750\]](#)
[Paraneoplastic Syndromes, Nervous System \[C10.574.781\]](#) +
▶ [Parkinson Disease \[C10.574.812\]](#)
[Postpoliomyelitis Syndrome \[C10.574.827\]](#)
[Prion Diseases \[C10.574.843\]](#) +
[Shy-Drager Syndrome \[C10.574.875\]](#)
[Subacute Combined Degeneration \[C10.574.910\]](#)
[Tauopathies \[C10.574.945\]](#) +

Procedure



Terminology mapping



Disease extraction

Involvement in disease	Defects in NF2 are the cause of neurofibromatosis 2 (NF2) [MIM:101000]; also known as central neurofibromatosis. NF2 is a genetic disorder characterized by bilateral vestibular schwannomas (formerly called acoustic neuromas), schwannomas of other cranial and peripheral nerves, meningiomas, and ependymomas. It is inherited in an autosomal dominant fashion with full penetrance. Affected individuals generally develop symptoms of eighth-nerve dysfunction in early adulthood, including deafness and balance disorder. Although the tumors of NF2 are histologically benign, their anatomic location makes management difficult, and patients suffer great morbidity and mortality.
------------------------	--

Extraction using regular expressions
'are the cause of'
'involved in'
etc.

#101000

NEUROFIBROMATOSIS, TYPE II; NF2

Alternative titles; symbols

NEUROFIBROMATOSIS, CENTRAL TYPE
ACOUSTIC SCHWANNOMAS, BILATERAL
BILATERAL ACOUSTIC NEUROFIBROMATOSIS; BANF
ACOUSTIC NEURINOMA, BILATERAL; ACN

MeSH
'Neurofibromatosis 2'

Term matching procedure

- **Term pre-processing:** term normalisation and transformation in bag-of-words
- **Exact matches:** same length, case insensitive
- **Partial matches:** calculation of a similarity score between terms based of the TF/IDF schema used in information retrieval:

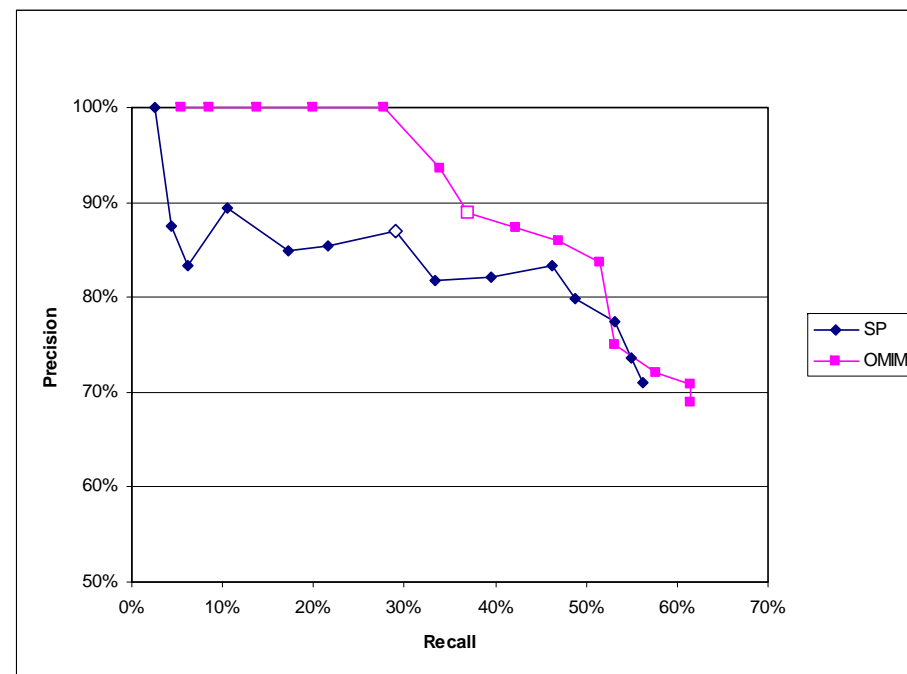
$$S = \frac{\sum_{cw} \log_2\left(\frac{1}{freq(cw)}\right) - \sum_{ncw} \log_2\left(\frac{1}{freq(ncw)}\right)}{size(disease)}$$

The term with the highest score was chosen.

Benchmark

200 disease names from 97 Swiss-Prot entries manually mapped to MeSH terms

- Used to evaluate the procedure in terms of **recall** and **precision**
- Used to set up a **score threshold**

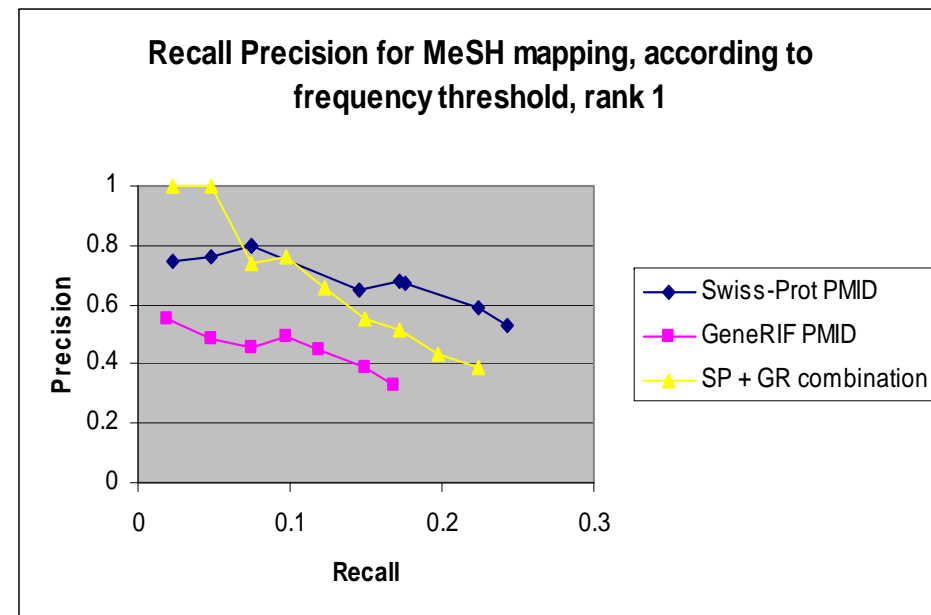


Result analysis

- On the benchmark:
 - 86% precision, 92% if restricted to partial matches to the same descriptors.
- On all Swiss-Prot disease lines
 - 68% coverage
- Problems with:
 - Lack of granularity of MeSH for genetic diseases
 - Differences in classification
 - Errors in disease extractions
- Mapping available at:
<http://research.isb-sib.ch/unimed>

Mapping through references

- Retrieval of **PubMed** documents cited in:
 - UniProtKB/Swiss-Prot entries
 - OMIM entries
 - GeneRiFs of the corresponding Entrez Gene entries
- Extraction of **indexing MeSH** terms from the disease category
- Ranking according to their relative frequencies



Result analysis

- On the benchmark:
 - Swiss-Prot references:
 - precision at rank 1: 67%
 - precision over all retrieved terms: 85%
 - OMIM references:
 - precision at rank 1: 48%
 - precision over all retrieved terms: 86%
 - GeneRiF references:
 - precision at rank 1: 51%
 - precision over all retrieved terms: 83%

Mapping combination

- Not easy because of the low precision of the “reference” mapping at rank 1
- Can be combined to improve the precision: if we discard “terminology” mapped terms not present in the “reference” mapped term list:
 - Precision: 93%
 - Recall: 63%
- The “reference” mapping could help users find the correct disease information

Mapping to ICD-10

The screenshot shows the ICD-10 website interface. At the top, it says "World Health Organization ICD Version 2007". On the left, there is a navigation menu with "List of Chapters", "Chapter Introduction", "List of Blocks", "Previous Block", and "Next Block". The main content area is titled "Chapter VI Diseases of the nervous system (G00-G99)" and "Extrapyramidal and movement disorders (G20-G26)". A search box on the left contains the word "parkinson" and has "Full search" selected. Below the search box, there are "OK" and "Help" buttons. A "Move to ICD code:" section has an empty input field and an "OK" button. The search results are as follows:

ICD-10 Code	Description
G20	Parkinson's disease Hemiparkinsonism Paralysis agitans Parkinsonism or Parkinson's disease: <ul style="list-style-type: none">· NOS· idiopathic· primary
G21	Secondary parkinsonism
G21.0	Malignant neuroleptic syndrome Use additional external cause code (Chapter XX), if desired, to identify drug.
G21.1	Other drug-induced secondary parkinsonism Use additional external cause code (Chapter XX), if desired, to identify drug.
G21.2	Secondary parkinsonism due to other external agents Use additional external cause code (Chapter XX), if desired, to identify external agent.
G21.3	Postencephalitic parkinsonism
G21.8	Other secondary parkinsonism
G21.9	Secondary parkinsonism, unspecified
G22*	Parkinsonism in diseases classified elsewhere Syphilitic parkinsonism (A52.1+)

- Lower performance on the benchmark:
 - Precision : 68%
 - Recall : 38%
- Problems:
 - term management
 - granularity of ICD-10 terms for genetic diseases
 - no morphologic variant terms

Discussion

- The mapping system was tuned for **high precision** to provide a fully automated procedure.
- But we need to **improve the recall** by:
 - Including NLP techniques in the disease extraction and matching procedures;
 - Trying to map to other terminologies such as SNOMed-CT, Disease Ontology to solve the granularity problem
- The mapping through references can be useful for database curation purposes and to provide additional disease information to users

Work in progress

- A search engine on variant pages with:
 - protein centric view
 - disease centric view
 - use of hierarchical indexing
- Implementaion of additional information in variant pages:
 - from the structure: mutation environment, interaction domain
 - from the sequence analysis: conservation score, sequence feature proximity
 - from the literature: via text mining

Aknowledgements



- Anaïs Mottaz
- Lina Yip
- Livia Famiglietti
- Arnaud Gos



- Patrick Ruch
- Robert Baud
- Julien Gobeil
- Abdel Kader Fall



SWISS NATIONAL SCIENCE FOUNDATION