



UniMed: From UniProt to Medicine

# Automatic Annotation of Pathological Functions in UniProt

Patrick Ruch

[patrick.ruch@sim.hcuge.ch](mailto:patrick.ruch@sim.hcuge.ch)



UNIVERSITÉ DE GENÈVE

June 25-27, 2007 - PRBB, Barcelona



1. Data overload is massive:  
genetic sequences,  
**textual data...**

Need automatic  
approaches

2. Genomics and proteomics  
database curation is  
partial: 20% in DB, vs.  
80% in text

Speed up database  
curation/annotation  
[BioCreative]

3. Medical outcomes are  
limited: clinical research,  
medicinal chemistry...

**Filling the gap between  
medicine and molecular  
biology...**



## By adding medical annotation in UniProt

- Make explicit and enrich medical contents, focusing on two objectives:
  - Disease-related information
  - Drug-related information
- **Priorities**
  - P1: Proteins with known pathological functions [OMIM]  
~ make explicit well known and established facts
  - P2: Other proteins  
~ discover new phenotype/genotype associations

# Swiss-Prot curation of DPYD: Comments and OMIM Links

## Comments

- **FUNCTION:** Involved in pyrimidine base degradation. Catalyzes the reduction of uracil and thymine. Also involved in the degradation of the chemotherapeutic drug 5-fluorouracil.
- **CATALYTIC ACTIVITY:** 5,6-dihydrouracil + NADP<sup>+</sup> = uracil + NADPH.
- **COFACTOR:** Binds 2 FAD.
- **COFACTOR:** Binds 2 FMN.
- **COFACTOR:** Binds 2 4Fe-4S clusters. Contains approximately 33 iron atoms per molecule.
- **PATHWAY:** Amino-acid biosynthesis; beta-alanine biosynthesis.
- **SUBUNIT:** Homodimer.
- **SUBCELLULAR LOCATION:** Cytoplasm.
- **TISSUE SPECIFICITY:** Found in most tissues with greatest activity found in liver and peripheral blood mononuclear cells.
- **DISEASE:** Defects in DPYD are the cause of dihydropyrimidine dehydrogenase deficiency (DPYD deficiency) [MIM:274270]; also known as hereditary thymine-uraciluria or familial pyrimidinemia. DPYD deficiency is a disease characterized by persistent urinary excretion of excessive amounts of uracil, thymine and 5-hydroxymethyluracil. Patients suffering from this disease show a severe reaction to the anticancer drug 5-fluorouracil. This reaction includes stomatitis, Leukopenia, thrombocytopenia, hair loss, diarrhea, fever, marked weight loss, cerebellar ataxia, and neurologic symptoms, progressing to semicoma.
- **WEB RESOURCE:** NAME=GeneReviews; URL="http://www.genetests.org/query?gene=DPYD".

## Copyright

Copyrighted by the UniProt Consortium, see <http://www.uniprot.org/terms>. Distributed under the Creative Commons Attribution-NoDerivs License.

## Cross-references

Sequence databases	
EMBL	U09178; AAA57474.1; U20938; AAB51366.1; AB003063; BAA8973.1; X95670; CAA64973.1; U57655; AAB07049.1
PIR	A54718; A54718.
UniGene	Hs.335034
3D structure databases	
HSSP	Q28943; 1GTE.
SMR	Q12882; 2-1017.
Protein family/group databases	
GermOnline	ENSG00000188641
Enzyme and pathway databases	
Reactome	REACT_1698.1; Nuc
Organism-specific gene databases	
HGNC	HGNC:3012; DPYD.
GeneCards	DPYD.
GeneLynx	DPYD; Homo sapien
GenAtlas	DPYD.
MIM	274270; gene+phen
Gene expression databases	
CleanEx	HGNC:3012; DPYD.
ArrayExpress	Q12882; -.
Ontologies	
	GO:0005737; Cellular component: cytoplasm ( <i>non-traceable author statement from UniProtKB</i> ).
	GO:0017113; Molecular function: dihydrovrimidine dehvdroeasenase (NADP+) activitv ( <i>non-traceable author statement from UniProtKB</i> ).

Defects in DPYD are the cause of dihydropyrimidine dehydrogenase deficiency [MIM:274270]; also (...) severe reaction to the anticancer drug 5-fluorouracil. This reaction includes stomatitis, leukopenia, hair loss, cerebellar ataxia, (...)

→ Disease

→ Patient safety/Adverse effects not listed in TERIAC !

# Useful links for DPYD: MEDLINE

## UniProtKB/Swiss-Prot entry Q12882

### Entry information

Entry name **DPYD\_HUMAN**  
Primary accession number **Q12882**  
Secondary accession numbers Q16694 Q16761 Q96TH1  
Integrated into Swiss-Prot on November 1, 1997  
Sequence was last modified on November 1, 1997 (Sequence version 1)  
Annotations were last modified on December 12, 2006 (Entry version 69)

### Name and origin of the protein

Protein name **Dihydropyrimidine dehydrogenase [NADP+] [Precursor]**  
Synonyms **EC 1.3.1.2  
DPD  
DHPDHase  
Dihydrouracil dehydrogenase  
Dihydrothymine dehydrogenase**  
Gene name **Name: DPYD**  
From Homo sapiens (Human) [TaxID: 9606]  
Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

### References

- [1] NUCLEOTIDE SEQUENCE [MRNA].  
**TISSUE=Liver;**  
PubMed=8083224  
Yokota H., Fernandez-Salguero P., Furuya H., Lin K., McBride O.W., Podschun B., Schnackerz K.D., Gonzalez F.J.;  
"cDNA cloning and chromosome mapping of human dihydropyrimidine dehydrogenase, an enzyme associated with 5-fluorouracil toxicity and congenital thymine uraciluria.";  
J. Biol. Chem. 269:23192-23196(1994).
- [2] NUCLEOTIDE SEQUENCE.  
PubMed=9135003  
Johnson M.R., Wang K., Tillmanns S., Albin N., Diasio R.B.;  
"Structural organization of the human dihydropyrimidine dehydrogenase gene.";  
Cancer Res. 57:1660-1663(1997).
- [3] NUCLEOTIDE SEQUENCE.  
DOI=10.1016/S0304-3835(97)00377-7; PubMed=9464498  
Ogura K., Nishiyama T., Takubo H., Kato A., Okuda H., Arakawa K., Fukushima M., Nagayama S., Kawaguchi Y., Watabe T.;  
"Suicidal inactivation of human dihydropyrimidine dehydrogenase by (E)-5-(2-bromovinyl)uracil derived from the antiviral, sorivudine.";  
Cancer Lett. 122:107-113(1998).
- [4] NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 581-635.  
**TISSUE=Liver;**  
PubMed=8892022  
Vreken P., van Kuilenburg A.B.P., Meinsma R., Smit G.P.A., Bakker H.D., de Abreu R.A., van Gennip A.H.;  
"A point mutation in an invariant splice donor site leads to exon skipping in two unrelated Dutch patients with dihydropyrimidine dehydrogenase deficiency.";  
J. Inher. Metab. Dis. 19:645-654(1996).
- [5] NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 581-635.  
PubMed=9170156  
Fernandez-Salguero P.M., Sapone A., Wei X., Holt J.R., Jones S., Idle J.R., Gonzalez F.J.;

# MEDLINE contents for PMID:9439663 [DPYD-linked]

1: [Hum Genet.](#) 1997 Dec;101(3):333-8.

[Related Articles, Links](#)



## Dihydropyrimidine dehydrogenase (DPD) deficiency: identification and expression of missense mutations C29R, R886H and R235W.

[Vreken P](#), [Van Kuilenburg AB](#), [Meinsma R](#), [van Gennip AH](#).

Academic Medical Center, University of Amsterdam, The Netherlands. [p.vreken@amc.uva.nl](mailto:p.vreken@amc.uva.nl)

Dihydropyrimidine dehydrogenase (DPD) deficiency (McKusick 274270) is an autosomal recessive disease characterized by thymine-uraciluria in homozygous-deficient patients and associated with a variable clinical phenotype. **Cancer patients with this defect should not be treated with the usual dose of 5-fluorouracil** because of the expected lethal toxicity. In addition, heterozygosity for mutations in the *DPD* gene increases the risk of toxicity in cancer patients treated with this drug. Sequence analysis in a patient with complete DPD deficiency, previously shown to be heterozygous for the delta C1897 frame-shift mutation, revealed the presence of a novel missense mutation, R235W. Expression of this novel mutation and previously identified missense mutations C29R and R886H in *Escherichia coli* showed that both C29R and R235W lead to a mutant DPD protein without significant residual enzymatic activity. The R886H mutation, however, resulted in about 25% residual enzymatic activity and is unlikely to be responsible for the DPD-deficient phenotype. We show that the *E. coli* expression system is a valuable tool for examining DPD enzymatic variants. In addition, two new patients who were both heterozygous for the C29R mutation and the common splice donor site mutation were identified. Only one of these patients showed convulsive disorders during childhood, whereas the other showed no clinical phenotype, further illustrating the lack of correlation between genotype and phenotype in DPD deficiency.

### MeSH Terms:

- [Adult](#)
- [Antimetabolites, Antineoplastic/adverse effects](#)
- [Cloning, Molecular](#)
- [DNA, Complementary/genetics](#)
- [Dihydrouracil Dehydrogenase \(NADP\)](#)
- [Escherichia coli/genetics](#)
- [Female](#)
- [Fluorouracil/adverse effects](#)
- [Heterozygote](#)
- [Humans](#)
- [Male](#)
- [Mutation\\*](#)
- [Oxidoreductases/deficiency\\*](#)
- [Oxidoreductases/genetics](#)
- [Polymerase Chain Reaction](#)
- [Purine-Pyrimidine Metabolism, Inborn Errors/complications](#)
- [Purine-Pyrimidine Metabolism, Inborn Errors/genetics\\*](#)
- [Recombinant Proteins/metabolism](#)
- [Seizures/complications](#)
- [Sequence Analysis, DNA](#)

### Substances:

- [Antimetabolites, Antineoplastic](#)
- [DNA, Complementary](#)
- [Recombinant Proteins](#)
- [Fluorouracil](#)
- [Oxidoreductases](#)
- [Dihydrouracil Dehydrogenase \(NADP\)](#)

Cancer patients with this defect should not be treated with the usual dose of 5-fluorouracil (...)

→ Pharmacovigilance/Signaling

MeSH Terms:  
Antimetabolites, Antineoplastic

Substances:  
Fluorouracil

PMID: 9439663 [PubMed - indexed for MEDLINE]

Entrez PubMed  
Overview  
Help | FAQ  
Tutorials  
New/Noteworthy  
E-Utilities

PubMed  
Services  
Journals Database  
MeSH Database  
Single Citation  
Matcher  
Batch Citation  
Matcher  
Clinical Queries  
Special Queries  
LinkOut  
My NCBI

Related  
Resources  
Order Documents  
NLM Mobile  
NLM Catalog  
NLM Gateway  
TOXNET  
Consumer Health  
Clinical Alerts  
ClinicalTrials.gov  
PubMed Central

# Swiss-Prot curation of PAX3: Variants + Phenotype

Key	From	To	Length	Description	FTId
CHAIN	1	479	479	Paired box protein Pax-3.	PRO_0000050178
DOMAIN	34	161	128	Paired.	
DNA_BIND	219	278	60	Homeobox.	
SITE	319	320	2	Breakpoint for translocation to form PAX3-NCOA1 oncogene.	
VAR_SEQ	196	215		ASAPQSDEGSDIDSEPDLP -> GKRWRLGRRTCVWTWRASAS (in isoform Pax3A).	VSP_002355
VAR_SEQ	196	206		ASAPQSDEGSD -> GKALVSGVSSH (in isoform Pax3B).	VSP_002357
VAR_SEQ	207	479		Missing (in isoform Pax3B).	VSP_002358
VAR_SEQ	216	479		Missing (in isoform Pax3A).	VSP_002356
VARIANT	45	45	1	F -> L (in WS1).	VAR_003790
VARIANT	47	47	1	N -> H (in WS3).	VAR_003791
VARIANT	47	47	1	N -> K (in CDHS).	VAR_003792
VARIANT	48	48	1	G -> R (in WS1).	VAR_017533
VARIANT	50	50	1	P -> L (in WS1; important hearing loss).	VAR_003793
VARIANT	56	56	1	R -> L (in WS1; associated with meningomyelocele).	VAR_003794
VARIANT	59	59	1	I -> F (in WS1).	VAR_003795
VARIANT	59	59	1	I -> N (in WS1).	VAR_003796
VARIANT	60	60	1	V -> M (in WS1).	VAR_003797
VARIANT	62	62	1	M -> V (in WS1).	VAR_003798
VARIANT	63	67	5	Missing (in WS1).	VAR_003799
VARIANT	73	73	1	S -> L (in WS1).	VAR_013640
VARIANT	78	78	1	V -> M (in WS1).	VAR_017534
VARIANT	81	81	1	G -> A (in WS1; original)	
VARIANT	84	84	1	S -> F (in WS3).	
VARIANT	85	85	1	K -> E (in WS1).	
VARIANT	90	90	1	Y -> H (in WS3).	
VARIANT	99	99	1	G -> D (in WS1).	
VARIANT	238	238	1	F -> S (in WS1).	
VARIANT	265	265	1	V -> F (in WS1).	
VARIANT	266	266	1	W -> C (in WS1).	
VARIANT	270	270	1	R -> C (in WS1 and WS3).	
VARIANT	271	271	1	R -> C (in WS1).	VAR_017537
VARIANT	271	271	1	R -> G (in WS1).	VAR_003806
VARIANT	271	271	1	R -> H (in WS1; associated with Lys-273 in one family).	VAR_017538
VARIANT	273	273	1	R -> K (associated with His-271 in one Waardenburg syndrome type I family).	VAR_017539
VARIANT	315	315	1	T -> K (in dbSNP:rs2234675) [NCBI/Ensembl].	VAR_003807
VARIANT	391	391	1	Q -> H (in WS1).	VAR_013641
CONFLICT	108	108		Missing (in Ref. 5).	

Meningomyeloceles  
Hearing loss  
→ Abnormality/Symptoms

## Sequence information

Length: 479 AA [This is the length of the unprocessed precursor]

# Inference... for retrieving/accessing data

<b>MeSH Heading</b>	Meningomyelocele
<b>Tree Number</b>	<a href="#">C10.500.680.610</a>
<b>Tree Number</b>	<a href="#">C16.131.666.680.610</a>
<b>Scope Note</b>	Congenital, or rarely acquired, herniation of meningeal and spinal cord tissue through a bony defect in the vertebral column. The majority of these defects occur in the lumbosacral region. Clinical features include <a href="#">PARAPLEGIA</a> , loss of sensation in the lower body, and incontinence. This condition may be associated with the <a href="#">ARNOLD-CHIARI MALFORMATION</a> and <a href="#">HYDROCEPHALUS</a> . (From Joynt, Clinical Neurology, 1992, Ch55, pp35-6)
<b>Entry Term</b>	Myelocele
<b>Entry Term</b>	Myelomeningocele
<b>Entry Term</b>	Acquired Meningomyelocele
<b>Entry Term</b>	Myelomeningocele, Acquired
<b>Allowable Qualifiers</b>	<a href="#">BL</a> <a href="#">CF</a> <a href="#">CI</a> <a href="#">CL</a> <a href="#">CO</a> <a href="#">DH</a> <a href="#">DI</a> <a href="#">DT</a> <a href="#">EC</a> <a href="#">EH</a> <a href="#">EM</a> <a href="#">EN</a> <a href="#">EP</a> <a href="#">ET</a> <a href="#">GE</a> <a href="#">HI</a> <a href="#">IM</a> <a href="#">ME</a> <a href="#">MI</a> <a href="#">MO</a> <a href="#">NU</a>
<b>Previous Indexing</b>	<a href="#">Spina Bifida</a> (1966-1974)
<b>Online Note</b>	search SPINA BIFIDA 1966-74; use SPINA BIFIDA to search MYELOCELE
<b>History Note</b>	78(75); was see under SPINA BIFIDA 1963-77; MYELOCELE was see under
<b>Unique ID</b>	D008591

## MeSH Tree

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

#### [Nervous System Malformations \[C10.500\]](#)

##### [Neural Tube Defects \[C10.500.680\]](#)

- [Anencephaly \[C10.500.680.196\]](#)
- [Arnold-Chiari Malformation \[C10.500.680.291\]](#)
- [Encephalocele \[C10.500.680.488\]](#)
- [Meningocele \[C10.500.680.598\]](#)
- ▶ [Meningomyelocele \[C10.500.680.610\]](#)
- [Spinal Dysraphism \[C10.500.680.800\] +](#)

### [Congenital, Hereditary, and Neonatal Diseases and Abnormalities \[C16\]](#)

#### [Abnormalities \[C16.131\]](#)

##### [Nervous System Malformations \[C16.131.666\]](#)

##### [Neural Tube Defects \[C16.131.666.680\]](#)

- [Anencephaly \[C16.131.666.680.196\]](#)
- [Arnold-Chiari Malformation \[C16.131.666.680.291\]](#)
- [Encephalocele \[C16.131.666.680.488\]](#)
- [Meningocele \[C16.131.666.680.598\]](#)
- ▶ [Meningomyelocele \[C16.131.666.680.610\]](#)
- [Spinal Dysraphism \[C16.131.666.680.800\] +](#)

## Synonyms

Spina Bifida (historical)

Myelocele

Myelomeningocele

Meningomyelocele

## Ontological inference

Neural Tube Defects

Nervous System Diseases

Nervous System Malformations

Meningomyelocele

Abnormalities

Broader semantics → UMLS



## Output of medical annotation for DPYD

### Statistical Ranking

+

dihydropyrimidine dehydrogenase deficiency (Disease or Syndrome T047)  
 dihydropyrimidine dehydrogenase (Amino Acid...T116/Enzyme T126 )  
 thrombocytopenias (Disease or Syndrome T047)  
 cerebellar ataxia (Sign or Symptom T184)

5-fluorouracil (Nucleic Acid T114/Pharmacologic Substance T121)

humans (Population Group T098)

leukopenias (Laboratory or Test Result T034/Pathologic Function T046)

stomatitis (Disease or Syndrome T047)

thymine (Nucleic Acid T114/ Biologically Active Substance T123)

-

*cancers* (Neoplastic Process T191)

#### Disease-related

dihydropyrimidine dehydrogenase deficiency

*+ Traceable statement*

thrombocytopenias

cerebellar ataxia

leukopenias

stomatitis

*neoplams*

#### Populations

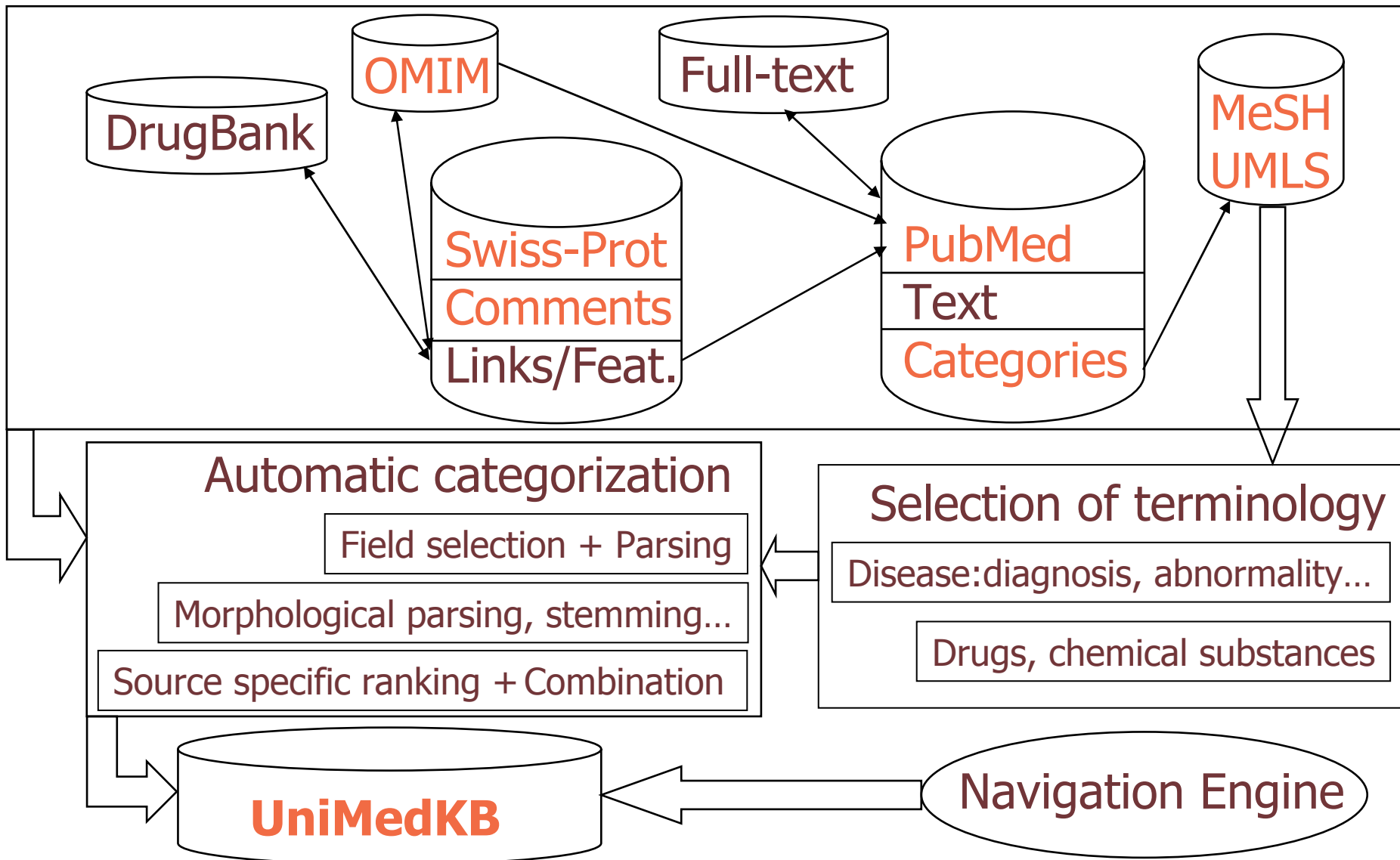
Humans

#### Drug-related

dihydropyrimidine dehydrogenase

5 fluorouracil

# Project Architecture: Current Experiments



SPROT	Name	MeSH/UMLS	Score	ID (OMIM, MEDL...)	Traceable Statement
Q12882	dpyd	Cancer	0.5	PMID-9439663	Cancer patients with this defect should not be treated with the usual dose of 5-fluorouracil.
Q12882	dpyd	5-fluorouracil	0.889	SP-Q12882-DISEASE	



## Current status of the project

- **Benchmark**

- 60 random UniProt entries

- 92 disease comments ( $\sim 2.14$ )

- 82 refs to OMIM

V1: Diseases, i.e.  $\sim 2.5$  categories per protein

V3: All medical entities: anatomy, drugs, populations...



## Method 1: High precision

- Focus on diagnosis  
Regular expressions

*cause of, involve in, contribute(s) to, cause(s) +  
Disease + [MIM#]*

*OMIM phenotypes and gene+pheno (#+)*

Titles + alternative symbols

→ Mapping to MeSH



## Method 2: Categorization by Ranking

- Applied to the whole CC DISEASE [Ruch and al 2006]
- FSA Pattern matcher to detect adjacent features  
 $\text{word}_1 \dots \text{word}_n \rightarrow \text{word}_1 \dots \text{word}_n \quad N = \text{Len}(w)$   
 $\text{word}_1 \dots \text{word}_n \rightarrow \text{word}_1 \dots \_^{[* , 1]} \dots \text{word}_n$   
→ Boolean scoring
- Vector Space: Porter stems + TF\*IDF weighting [VS]  
→ Data-driven scoring
- Indexing units:
  - Stems
  - Linguistically-motivated phrases
  - Thesaurus

## Vector space parameters

- TF:  $\text{weight}_{\text{term}} = f(\text{term frequency})$
- IDF:  $\text{weight}_{\text{term}} = f(\text{document frequency}^{-1})$
- Normalization: to balance long and short documents

Term Frequency	
First Letter	$f(tf)$
n (natural)	$tf$
l (logarithmic)	$1 + \log(tf)$
a (augmented)	$\alpha + \beta \times (\frac{tf}{\max(tf)})$ , where $\alpha = 1 - \beta$ and $0 < \alpha < 1$
Inverse Document Frequency	
Second Letter	$f(\frac{1}{df})$
n(no)	1
t(full)	$\log(\frac{N}{df})$
Normalization	
Third Letter	$f(\text{length})$
n(no)	1
c(cosine)	$\frac{1}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{j=1}^t w_{j,q}^2}}$

$$\text{dtu: } w_{ij} = \frac{(\text{Ln}(\text{Ln}(tf_{ij})+1)+1) \cdot idf_j}{(1-\text{slope}) \cdot \text{pivot} + \text{slope} \cdot nt_i}$$

$$\text{dtn: } w_{ij} = idf_j \cdot (\text{Ln}(\text{Ln}(tf_{if}) + 1) + 1)$$

## Mean average precision: MAP

	Precision at 3	Top Precision	MAP
RegEx + VS + Thesaurus	0.80	0.92	0.20

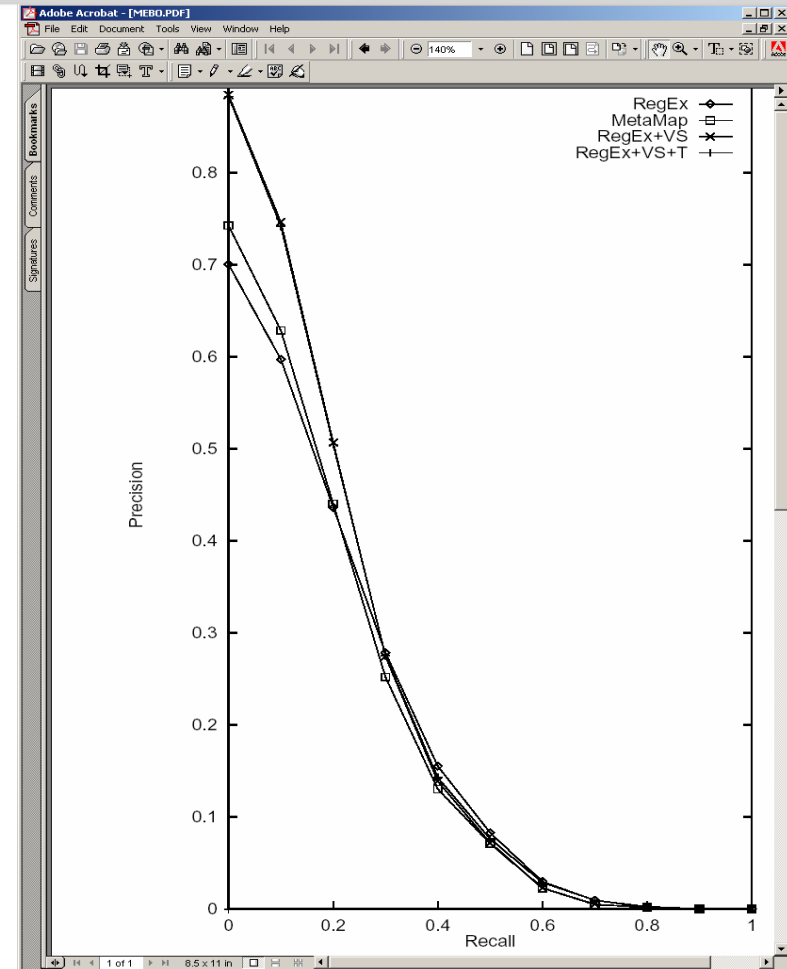
Ruch 2006: MeSH ~ 92%  
GO ~ 17%

Now: MeSH ~ 95%  
GO ~ 30% (+76%)



GOCat: A Gene Ontology Annotation Tool

<http://www.geneontology.org/GO.tools.annotation.shtml#gocat>





## Method 3: Combination with Bibliometrics

- Follow all bibliographical references to MEDLINE
  - In Swiss-Prot
  - In OMIM, via CC-DISEASE
- Rank Medical Subject Headings by frequency
  - MeSH-wide
- Combined with previous statistical estimates
- Filter by UMLS Semantic Types



## Results: Method 1 [Pattern matching]

- Recall out of 92 diseases [Exact match]

16 from Swiss-Prot	17%	
21 from OMIM	23%	
27 from both	29%	(P=100%)
- + Partial match:  
**recall = 49%**  
**precision = 80%**

	1	2* + 3*	1 + 2* + 3*
Precision at recall = 0	0.80	0.83	<b>0.84 (+ 5%)</b>
Recall after 15 categories	0.49	0.57	<b>0.57 (+ 16%)</b>

\* Applied on all the content of DISEASE fields

- The improvement is significant for recall
- Marginal for precision...

# Error analysis: Case with MAP = 0 [Method 2 + 3]

Queryid (Num): 1468  
Total number of documents over all queries  
Retrieved: 7  
Relevant: 2  
Rel\_ret: 0  
Interpolated Recall - Precision Averages:

043548

CC -!- DISEASE: Defects in TGM5 are a cause of peeling skin syndrome CC acral type (APSS) [MIM:609796, 270300]. Peeling skin syndrome CC (PSS) is an autosomal recessive genodermatosis characterized by CC the continuous shedding of the outer layers of the epidermis from CC birth and throughout life. In some cases of PSS, skin peeling is CC accompanied by erythema, vesicular lesions, or, in rare cases, CC other ectodermal features, like fragile hair and nail CC abnormalities. Two main subtypes, noninflammatory type A and CC inflammatory type B, have been suggested. However, it is clear CC from the dermatology literature that there are additional CC subtypes. In some families, an acral form of PSS (APSS) has been CC reported, in which skin peeling is strictly limited to the dorsa CC of the hands and feet, and, again, ultrastructural and CC histological analysis shows a level of blistering high in the CC epidermis at the stratum granulosum-stratum comeum junction.

043548 Skin Disease, Genetic | T037953 | M  
043548 Skin Abnormalities | T037935 | M0019

? id	sprot_alias	mesh_id	mesh_description
1622	043548	D000013	abnormalities
1623	043548	D001768	blisters
1624	043548	D004890	erythemas
1625	043548	D005533	Foot Dermatoses
1626	043548	D006229	Hand Dermatoses
1627	043548	D009264	nail abnormalities
1628	043548	D012871	Skin Diseases
1629	043548	D012872	Skin Diseases, Vesiculobullous
1630	043548	D013577	syndromes

- OMIM provides better synonyms than MeSH
- The benchmark contain only diagnosis, while other categories are relevant: signs, symptoms...  
→ Benchmark must be enriched
- Use domain-specific regularities  
Skin diseases → Skin diseases, **genetic**
- Enrich existing terminologies  
Protein X → Protein X **deficiency**



## Conclusion

- Automatic medical annotation is feasible
- Using existing curated KB is effective
- Nosography<sub>OMIM</sub> > MeSH for genetic diseases
  - obtain a better vocabulary for diseases...
  - importance of additional resources such as SNOMED or ICD

## Acknowledgements

Anaïs Mottaz, Lina Yip, Anne-Lise Veuthey,  
Julien Gobeill, Imad Tbahriti, Robert Baud

