Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Text-mining for Swiss-Prot curation

a story of success and failure

Anne-Lise Veuthey, Swiss Institute of Bioinformatics

SIB

swissprot

already a long story…

- 8 years of text-mining activities at Swiss-Prot to help curators screen the scientific literature.
- participation in European projects:
  - **BioMinT** (FP5 " Quality of Life" project)
  - **FELICS** (FP6 " Research Infrastructure" project)
  - **SLING** (FP6 " Research Infrastructure" project)
- projects funded by the SNF:
  - **EAGL**
  - **UniMed**

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

…with three main lessons learnt

- Human knowledge and expertise is difficult to emulate
- Discrepancies exist between research objectives and curation requirements
- Integration in an annotation platform is mandatory

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

IR: which search engine?

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

- Many tools exist to retrieve and rank documents according to a query (Lucene, easyIR, txtmatch).
- **But**, curators have their own strategy to retrieve information:
  - They use contextual knowledge,
  - They directly screen full-text articles,
  - They use connections provided by citations,
  - They use multiple sources of information (databases, Google, etc.)
- In conclusion, other IR engines hardly compete with PubMed.

The precision/recall trade-off

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Two principal classes of methods for text-mining based on:

- Patterns and rules:
  - can be manually tuned for good precision
  - creation of patterns and rules is time-consuming
  - designed only for a specific application
- Machine learning:
  - able to generalize, usually good recall but often insufficient precision
  - needs large annotated corpora for training, which necessitate curation expertise.

Precision is more important than recall in text-mining tools for curators.

A successful niche

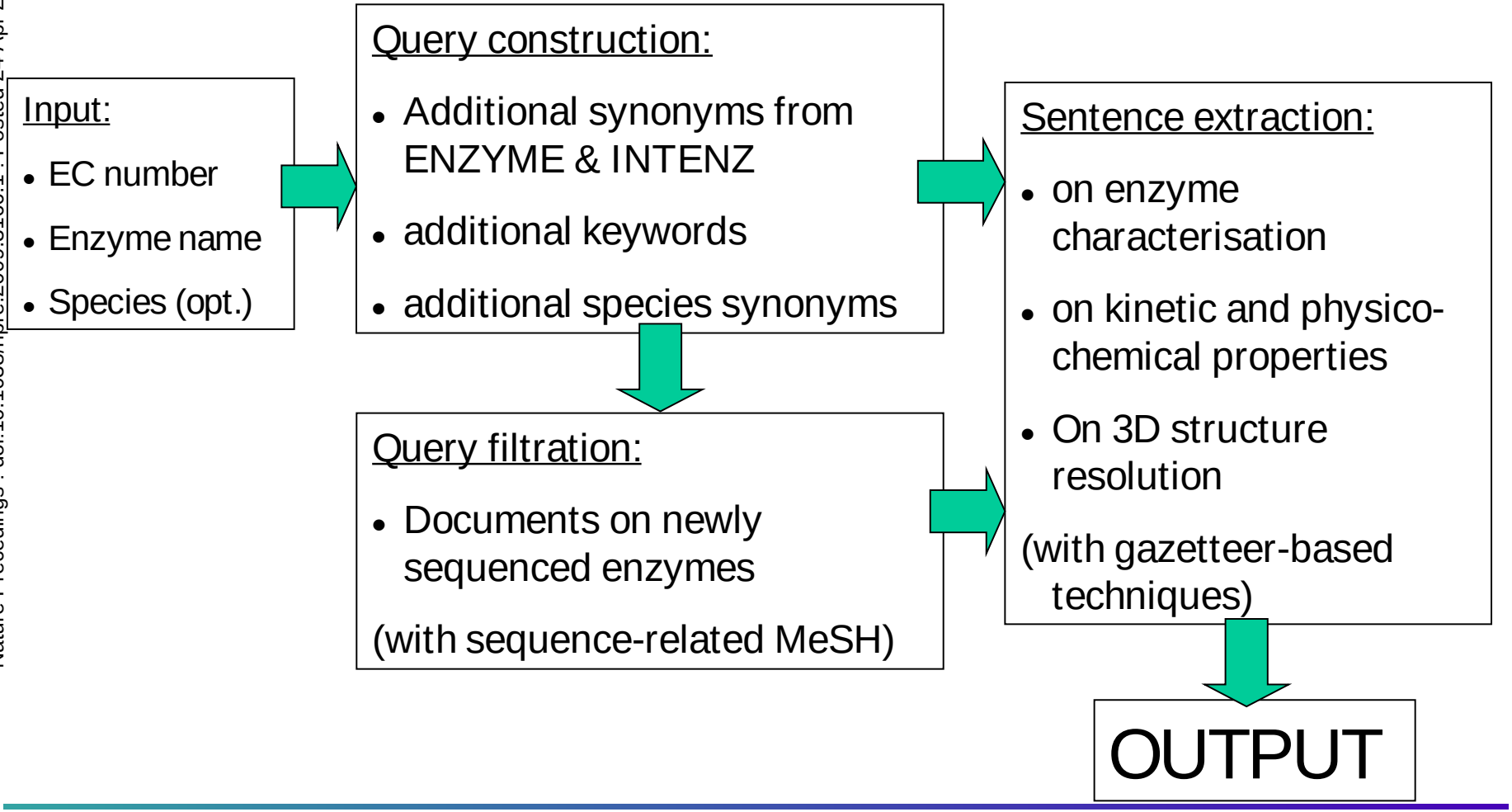
Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Failing to provide a global system of IR/IE for the UniProt curation, we have redirected our efforts to more modest applications based on patterns and rules, namely:

- The follow-up of the literature on enzymes
- The update of information on protein variants

A pipeline to follow up newly published information on enzymes

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009



Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 2 Apr 2009

Documents on enzyme structure: Results

bottom

9 / 195 document(s) contain(s) information on enzyme structure.

View query terms in abstracts show

Highlighted sentences contain information on enzyme structure according to the text-mining process.

Back New search

Documents sorted by: date or score score information

Index

18495156
16098517
15159562
14646138
14646103
12962631
11722571
11358521
10818354

9. PMID 8033912 --- collapse

Desulfoviridin, a multimeric-dissimilatory sulfite reductase from Desulfovibrio vulgaris (Hildenborough). Purification, characterization, kinetics and EPR studies.

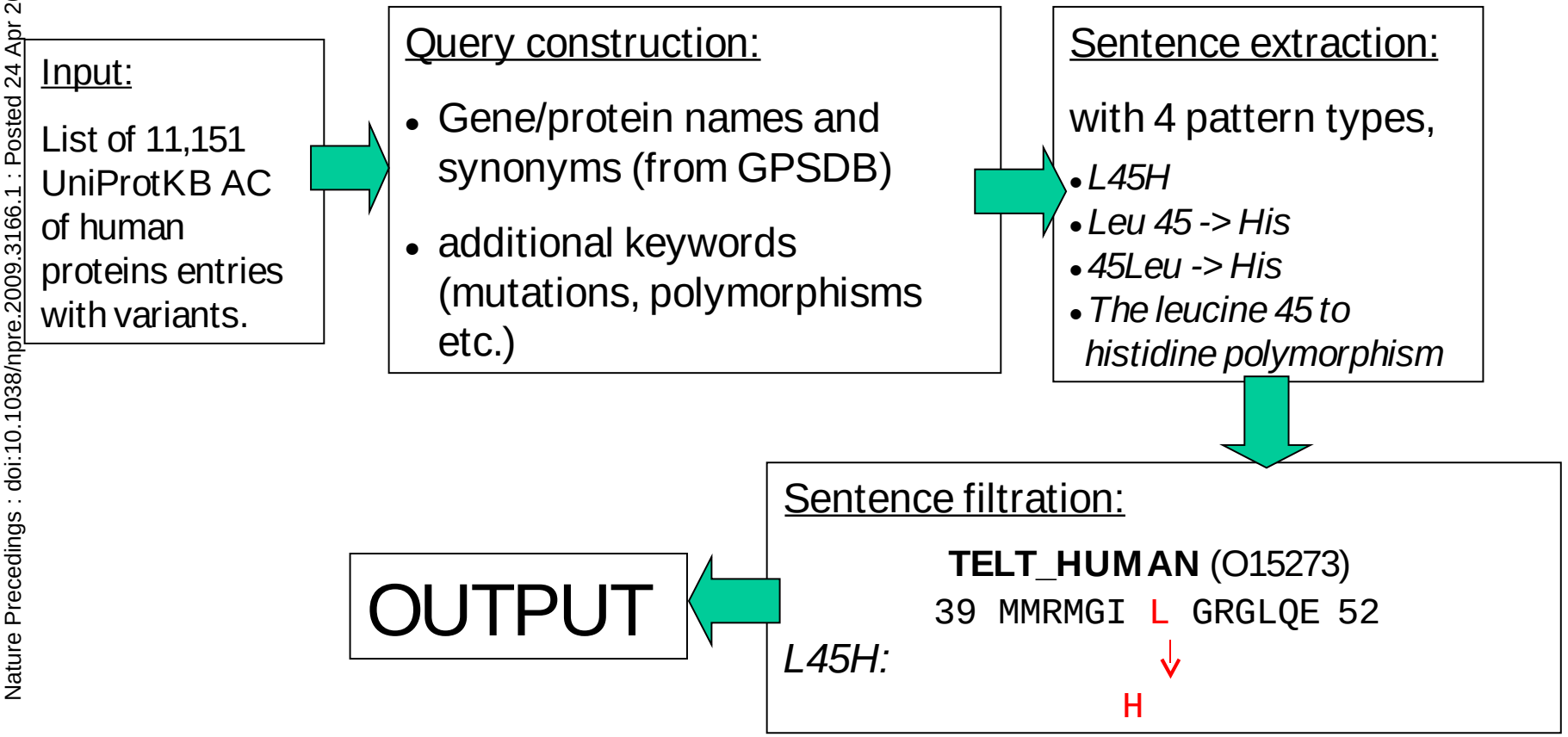
Wolfe BM, Lui SM, Cowan JA.

Eur J Biochem. 1994 Jul 1;223(1):79-89. EDAT: 1994/07/01

Conditions for the rigorous purification of desulfoviridin , the dissimilatory sulfite reductase from the sulfate-reducing bacterium Desulfovibrio vulgaris (Hildenborough) have been established . A final purification by fast protein liquid chromatography yields at least three distinct bands that each exhibit the characteristic absorption spectrum of desulfoviridin . Two of these have been extensively characterized by amino acid analysis , isoelectric focusing , polyacrylamide gel electrophoresis , and formulation of the prosthetic centers . Each contains two pairs of [Fe4S4] and siroheme units . These results stand in marked contrast to recent work claiming significant demetallation of siroheme , excess iron content , and the presence of Fe6S6 clusters . These proposals are critically assessed in light of our results and other published work . Steady-state kinetic parameters have been determined : kcat(SO3(2-) = 0.31 mol SO3(2 -) .s-1 .mol heme-1 , Km = 0.06 mM ; kcat(NO2-) = 0.038 mol NO2-.s-1 .mol heme-1 , Km = 0.028 mM ; kcat(NH2OH) = 29 mol NH2OH.s-1 .mol heme-1 , Km = 48 mM . A detailed comparison is made with the Escherichia coli and spinach assimilatory sulfite reductase enzymes and spinach nitrite reductase . Highly purified samples of dissimilatory sulfite reductase display an electron paramagnetic resonance spectrum characteristic of rhombic high spin ferric heme centers , while the fully reduced enzyme shows EPR features typical of [Fe4S4] clusters . The magnetic properties of the prosthetic centers are further characterized by variable temperature experiments and spin quantitation .

A pipeline for variant information update

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009



Yip Y.L., Lachenal N., Pillet V., Veuthey A-L.: *J. Bioinform. Comput. Biol.* 5:1215-1231(2007).

Swiss-Prot variant: VAR_015076 in UniProtKB/Swiss-Prot Q9Y6L6

General Information · Information on the variant · Sequence features · Structural features · References for the variant · Cross-references for the variant · Additional references for the variant (retrieved by text-mining)

Note: Most headings are clickable, even if they don't appear as links. They link to the UniProtKB user manual or variant pages documentation.

General information

Hide | Top

Swiss-Prot ID (AC)	**SO1B1_HUMAN (Q9Y6L6)**
Gene symbol(s)	Official: SLCO1B1 Synonym(s): LST1 or OATP2 or OATPC or SLC21A6
Chromosomal location	12p12.1
Protein name	Solute carrier organic anion transporter family member 1B1
Length of the protein	691

Information on the variant

Hide | Top

FTId	VAR_015076
Amino acid position of the variant	174
Residue change	From **Valine** (V) to **Alanine** (A), V174A, p.Val174Ala
Physico-chemical property	Change from medium size and hydrophobic (V) to small size and hydrophobic (A)
BLOSUM score	0
Status	Unclassified
Comment	Decreased transport activity

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Additional references for the variant (retrieved by text-mining)

Hide | Top

[1] **"The effect of SLCO1B1 polymorphism on repaglinide pharmacokinetics persists over a wide dose range. "**
Kalliokoski A., Neuvonen M., Neuvonen P.J., Niemi M.
Br J Clin Pharmacol. 2008 Sep 23.[Pubmed: 18823304][Abstract]
In abstract : "*Val174Ala) polymorphism on the pharmacokinetics of repaglinide is dose-dependent . ...*"

[2] **"Pharmacokinetic comparison of the potential over-the-counter statins simvastatin, lovastatin, fluvastatin and pravastatin. "**
Neuvonen P.J., Backman J.T., Niemi M.
Clin Pharmacokinet. 2008;47(7):463-74.[Pubmed: 18563955][Abstract]
In abstract : "*Val174Ala) genetic polymorphism of SLCO1B1 (encoding OATP1B1) considerably increases the plasma concentrations of simvastatin acid and moderately increases those of pravastatin but seems to have no significant effect on fluvastatin . ...*"

[3] **"SLCO1B1 521T-->C functional genetic polymorphism and lipid-lowering efficacy of multiple-dose pravastatin in Chinese coronary heart disease patients. "**
Zhang W., Chen B.L., Ozdemir V., He Y.J., Zhou G., Peng D.D., Deng S., Xie Q.Y., Xie W., Xu L.Y., Wang L.C., Fan L., Wang A., Zhou H.H.
Br J Clin Pharmacol. 2007 Sep;64(3):346-52. Epub 2007 Apr 18.[Pubmed: 17439540][Abstract]
In abstract : "*The aim of the present study was to evaluate the impact of SLCO1B1 521T -- > C (Val174Ala) functional genetic polymorphism on the lipid-lowering efficacy of multiple-dose pravastatin in Chinese patients with CHD . ...*"

[4] **"Role of organic anion transporter OATP1B1 (OATP-C) in hepatic uptake of irinotecan and its active metabolite, 7-ethyl-10-hydroxycamptothecin: in vitro evidence and effect of single nucleotide polymorphisms. "**
Nozawa T., Minami H., Sugiura S., Tsuji A., Tamai I.
Drug Metab Dispos. 2005 Mar;33(3):434-9. Epub 2004 Dec 17.[Pubmed: 15608127][Abstract]
In abstract : "*[...uptake of [(3)H]estrone-3-sulfate. Among the variants examined, OATP1B1*15 (N130D and V174A; reported allele frequency 10-15%) exhibited decreased transport activities for SN-38 as well as...]* "

[5] **"Functional analysis of single nucleotide polymorphisms of hepatic organic anion transporter OATP1B1 (OATP-C). "**
Iwai M., Suzuki H., Ieiri I., Otsubo K., Sugiyama Y.

Some challenging developments (1)

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Information extraction from full-text articles:

- gene/protein names, orf names
- PTMs, isoforms, mutations and polymorphisms
- experimental procedures

- Technical challenges:
  - HTML and PDF processing
  - analysis of tables, figures and supplementary material
  - information filtering through article' s section detection (Background, Methods, Results, Discussion)
  - combination of information, eg. results with related experimental procedure

Some challenging developments (2)

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Novelty detection:

- In UniProtKB/Swiss-Prot over 100’ 000 protein sequences have no functional annotation, it is important to follow the literature to find:
  - new functional characterization
  - changed or additional functions
  - new protein structures
- Technical challenges:
  - identification of the concerned protein (detection of orf names and species, usually in full-text)
  - detection of new facts out of redundant observations reported in the literature.

Metadata on publications

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

- Annotation of articles with metadata, eg. official gene names, GO terms, species, etc.
- Directly provided by the authors or suggested by text-mining tools and corrected by the authors.
- Some journals started to implement such data (*FEBS* initiative, *Plant Physiology* with *A. Thaliana* orf names).

…and don’ t forget

- Human knowledge and expertise is difficult to emulate
- Discrepancies exist between research objectives and curation requirements
- Integration in an annotation platform is mandatory

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

Acknowledgments

Nature Precedings : doi:10.1038/npre.2009.3166.1 : Posted 24 Apr 2009

- Violaine Pillet
- Nathalie Lachenal
- Marc Zehnder
- Pavel Dobrokhotov

and the
Swiss-Prot curation team
for their help and their feedback