Genomic CDS: an example of a complex ontology for pharmacogenetics and clinical decision support

Matthias Samwald[1]

[1] Medical University of Vienna, Vienna, Austria
matthias.samwald@meduniwien.ac.at

Abstract. Individual genetic data can be used to better predict the efficacy and safety of medications for individual patients. The Genomic Clinical Decision Support (Genomic CDS) ontology aims to utilize advanced Web Ontology Language 2 (OWL 2) reasoning for this task. The important, clear-cut medical use case, the complex axioms in the ontology and the heavy use of qualified cardinality restrictions make the ontology an interesting test object for new OWL 2 reasoners with improved performance.

Keywords: OWL, pharmacogenetics, clinical decision support

1 Motivation

Different patients can react drastically different to the same type of medication (**Fig. 1**). The goal of personalized medicine and pharmacogenetics is to predict an individual patient's response by analyzing genetic markers that influence how medications are metabolized or able to bind to their targets.
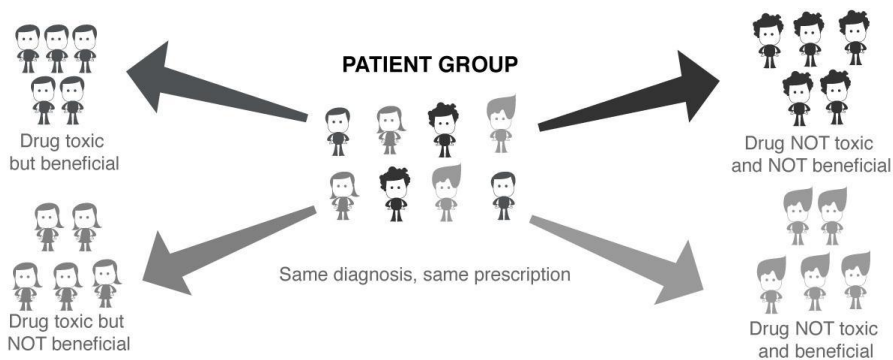

Fig. 1. The efficacy and safety of medications can drastically vary between patients. The goal of pharmacogenetics is to classify patients into subgroup based on genetic markers, to better predict which treatments could help and which could do harm.

To produce clinically valid and trustworthy predictions, no errors or ambiguities should arise in the process of inferring a patient's likely response from raw genetic

data. Current formalisms, data infrastructures and software applications leave many opportunities for introducing such errors and ambiguities. Ontologies formalized with the Web Ontology Language 2 (OWL 2) could be an excellent choice for tackling this problem, but the complexity and potentially large scale of ontologies in this domain also pose formidable challenges to currently available OWL 2 reasoners.

2   The Genomic CDS ontology

The Genomic Clinical Decision Support (Genomic CDS) ontology is an OWL 2 ontology aimed at representing pharmacogenetic knowledge and providing clinical decision support based on pharmacogenetic data. It is being developed by members of the Clinical Pharmacogenomics Task Force, which is part of the Health Care and Life Science Interest Group of the World Wide Web Consortium (W3C). The OWL files of the ontology, as well as 'demo' files containing example patient data can be downloaded from http://www.genomic-cds.org/ont/snapshot-june-2013
We also created a simplified version of the Genomic CDS ontology, called 'Genomic CDS light', which does not contain some of the axioms of the full ontology. Both versions of the ontology have ALCQ expressivity. They are characterized by extensive use of qualified cardinality restrictions.

The goals of developing the ontology are:
- Providing a simple and concise formalism for representing pharmacogenetic knowledge,
- Finding errors and lacking definitions in pharmacogenetic knowledge bases
- Automatically assigning alleles and phenotypes to patients
- Matching patients to clinically appropriate pharmacogenetic guidelines and clinical decision support messages

In the most common scenario, genetic patient data in OWL format is combined with the axioms of the Genomic CDS ontology, and an OWL reasoner is used to infer matching pharmacogenetic treatment recommendations. Several inference steps are needed to derive matching treatment recommendations from raw data about genetic markers (**Fig. 2**). The raw data consists of small variants in the genetic code, which in most cases are so-called *single nucleotide polymorphisms* (SNPs)*, such as an 'A' instead of a 'G'. *Alleles* are variants of a gene that are defined by containing sets of such small variants. *Phenotypes* are referring to the specific effects that certain small variants and alleles can have on the organism, e.g., how quickly a patient metabolizes a specific drug. Clinical guidelines can use small variants, alleles and/or phenotypes to match patients with treatment recommendations

The human genome usually contains two copies of each gene (one from the father, one from the mother), with each copy potentially bearing multiple genetic variants. Because of this, the ontologies rely heavily on qualified cardinality restrictions with

cardinalities of two, which seems to cause performance issues with most current OWL reasoners.
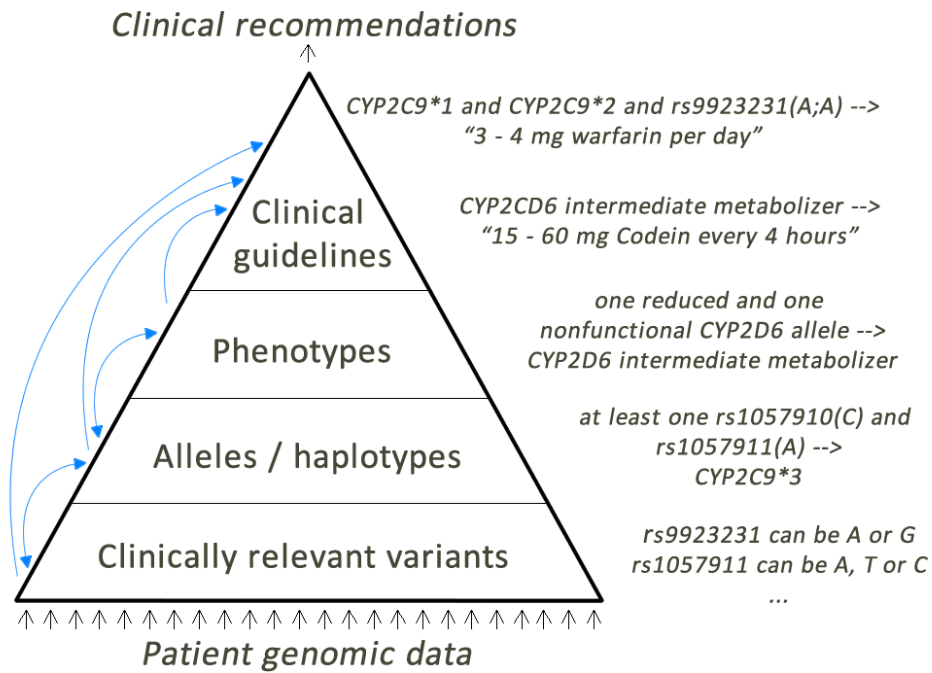

Fig. 2. : Through a series of inference steps, matching pharmacogenetic treatment guidelines are inferred from raw genetic patient data.

A simplified example of a rule for inferring an allele (CYP2C9*3) and its single nucleotide polymorphisms (SNPs) from a so-called 'tagging SNP' (a SNP that is necessary and sufficient for inferring the presence of the allele) looks like this in Manchester syntax:

```
Class: 'human with CYP2C9*3'
    EquivalentTo:
        has some rs1057910_C
    SubClassOf:
      has some 'CYP2C9 *3',
     (has some rs1057910_C)
      and (has some rs1057911_A)
      and (has some rs1799853_C)
      and (has some rs2256871_A)
      and (has some rs72558188_AGAAATGGAA)
```

An example of an axiom for inferring an adequate clinical decision support message for the anticoagulant drug warfarin (based on a combination of alleles and SNPs according to an official recommendation in the drug label):

```
Class: 'human triggering CDS rule 7'
  EquivalentTo:
   (has some 'CYP2C9*1') and (has some 'CYP2C9*3')
    and (has exactly 2 rs9923231_C)
   Annotations:
     label "human triggering CDS rule 7",
    CDS_message "3-4 mg warfarin per day should
       be considered as a starting dose range for
       a patient with this genotype according to
       the Warfarin drug label (Bristol-Myers
       Squibb)."
```

We used two OWL 2 reasoners with our ontology: TrOWL[1] [1] and HermiT[2] [2]. We also evaluated other OWL 2 reasoners (Fact++[3], Pellet[4]) in early stages of the project, but excluded them from further tests because they did not terminate or crashed even with small, preliminary versions of the ontology we developed. We compared the performance of the two reasoners on a virtual machine running on the Amazon Elastic Cloud Computing (EC2) cloud[5]. The machine was of the "High-Memory Extra Large Instance" type, running Microsoft Windows Server 2008, with 17.1 GB of memory, a 64-bit platform, and two virtual cores with 3.25 EC2 compute units each.

The reasoners were run as plugins in the 64 bit version of the Protégé 4.2 ontology editor. The initial heap size for Protégé was 10^{10} bytes (10 GB), and the maximum allowed heap size was 1.5×10^{10} bytes (15 GB). The TrOWL reasoner plugins with version 0.6 and 1.1 were each run three times for each ontology, and the mean of the time needed for classification was calculated. The HermiT 1.3.8 plugin was run once for each version of the ontology.

These preliminary tests showed TrOWL to be significantly more performant than HermiT for classifying the ontologies (
Table 1). However, HermiT was able to identify biologically meaningful inconsistencies present in genomic-cds-demo.owl (but not present in the light version of the ontology). TrOWL did not recognize these inconsistencies, most likely because it only partially covers the OWL 2 DL ruleset. These results show that only TrOWL is performant enough to be used in realistic settings (e.g. for clinical decision support), but that HermiT could serve to test and validate the results from TrOWL during develop-

---

[1]   http://trowl.eu
[2]   http://www.hermit-reasoner.com/
[3]   http://code.google.com/p/factplusplus/
[4]   http://clarkparsia.com/pellet/
[5]   http://aws.amazon.com/en/ec2/

ment (possibly comparing the results of the two reasoners for smaller ontology fragments).

Table 1. Reasoning performance: TrOWL is significantly more performant than HermiT in classifying our demo ontology (OWL 2 DL with ALCQ expressivity)

	HermiT 1.3.8	**TrOWL 1.1**	**TrOWL 0.6**
genomic-cds-light-demo.owl (2150 classes, 9500 axioms)	3 hours 48 minutes	1.5 seconds	18 seconds
genomic-cds-demo.owl (2300 classes, 11000 axioms)	detected inconsistencies	5.8 seconds	54 seconds

3   Conclusions and outlook

The Genomic CDS ontology is an example of an OWL 2 ontology for clinical genetics and decision support. Even though it is focused on a relatively small set of the most important pharmacogenetic markers, the ontology poses a significant challenge to currently available OWL 2 reasoners. There is great need for reasoners that are optimized for the kinds of OWL axioms encountered in ontologies dealing with clinical genomics.

4   Acknowledgements

References

1.   Thomas, E., Pan, J.Z., Ren, Y.: TrOWL: Tractable OWL 2 Reasoning Infrastructure. the Proc. of the Extended Semantic Web Conference (ESWC2010) (2010).
2.   Motik, B., Shearer, R., Horrocks, I.: Hypertableau Reasoning for Description Logics. J. Artif. Intell. Res. 36, 165–228 (2009).