# SNP-Ontology for semantic integration of genomic variation data

**Adrien Coulet[1,2], Malika Smaïl-Tabbone[2], Pascale Benlian[3], Amedeo Napoli[2] and Marie-Dominique Devignes[2]**

[1]KIKA Medical, 35 rue de Rambouillet 75012 Paris, France. [2]LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP), Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France. [3]Université Pierre et Marie Curie - Paris6, INSERM UMRS 538, Biochimie - Biologie Moléculaire, Paris, France

One of the great challenges in the post-genomic area consists in exploring the involvement of individual genomic variations in biological processes. Among the large amount of individual variations (more than 10 millions displaying a frequency higher than 1% in studied populations) dispersed all along the genome, most are Single Nucleotide Polymorphism (SNPs). SNP are specially studied in the field of pharmacogenomics for evaluating their impact on individual drug responses.

Major difficulties arise when locally observed genotype data on genomic variations have to be confronted to existing data in public databases. Particularly the nomenclatures used to describe the SNPs are heterogeneous within the public databases themselves (dbSNP, UCSC genome browser, HapMap, PharmGKB), and when compared to private data sources, so that variant identification and correspondence between two heterogeneous sources is not obvious. We propose here to formalize knowledge on genomic variations in view of integrating the data and its associated knowledge in a uniform semantic environment defined by an ontology.

Variation databases indifferently represent variations in DNA, RNA or protein. Thus, they represent as well the original variation and its repercussions. In addition, the representation of a variant position differs depending on the reference sequence (and its version) used to locate it. A generic syntax has been recommended by the Human Genome Variation Society (http://www.hgvs.org/mutnomen/recs.html). However this nomenclature has not been universally adopted yet. As a result finding intersection between several genomic variation databases is not easy because of the amount of synonymous representations.

We present here an approach for ontology-driven semantic integration of genomic variation data. It is composed of the design of an ontology on the considered domain, and the populating of a corresponding knowledge base in accordance to the ontology semantics.

In the first step of the ontology design, UML class diagrams are used as a support for conceptually modeling the domain. By analogy with database schema and with global schema used for data integration, we express with UML an *ontology schema* that describes concepts and relationships between them [1]. Once conceptualized, the domain knowledge is formalized in OWL (Web Ontology Language), and implemented within a knowledge base capable of housing individuals. The ontology editor currently used is Protégé. This editor is plugged with inference mechanisms as an OWL classifier for completing and validating the knowledge base.

As an application of the exposed approach, the SNP-Ontology has been designed and coded with the OWL formalism (see Figure 1). Converting one SNP format into another one, and establishing equivalence between variants displaying different representations, calls for explicit domain knowledge about gene structure, transcript definition, genetic code. These concepts and other specific ones have been defined for taking into account any variant representation. Existing

initiatives such as the PharmGKB ontology and the OMG SNP specification have inspired the modeling task [2][3].

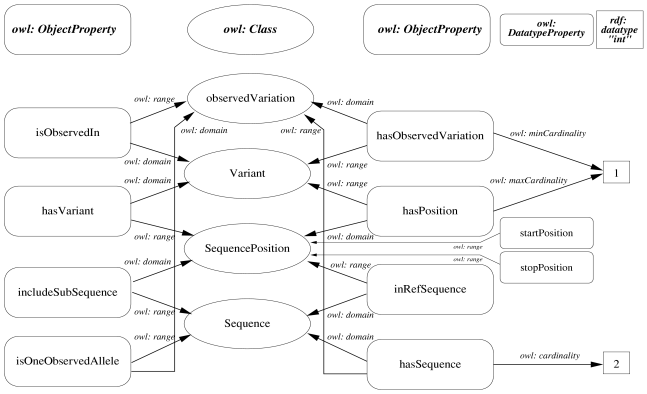


**Figure 1.** Illustration of some OWL classes and properties of the SNP-Ontology.

In parallel, we develop a wrapper for extracting variants from databases, for transforming them in accordance to the SNP-Ontology semantics, and for entering variant individuals in a SNP-specialized knowledge base. Thanks to the wrapper, we populated the SNP-Ontology knowledge base with a dataset of N genomic variations (N=706) pertaining from dbSNP and two private databanks *snp_base_1* and *snp_base_2*. Since these variants concern a gene of interest arbitrarily named GeneA, we called this first version of the knowledge base the GeneA-SNP-knowledge base. This integration test shows that the overlapping between the three databases is really low. Indeed less than 5% of variants are found in more than two databases. This result justify the relevance of data integration between public and private genomic variation resources. The variant integration procedure is validated by using the reasoner coupled to the knowledge base to check each assertion. The resulting fully consistent GeneA-SNP-knowledge base offers facilities to query its variant population according to their properties such as position on a given sequence. This functionality contributes to the data selection step in the early stages of a pharmacogenomic KDD process.

Our approach differs from integrated solutions for variant and more general ones such as BioMart or YeastHub since most of these approaches aim at facilitating integrated access to heterogeneous data, whereas our goal is to facilitate data mining and integration of data-mining results in a knowledge base. The work reported here concerns the first step of a pharmacogenomic KDD process. Complete implementation will necessitate extension of the ontology to include drug and phenotype modeling, according to the approach described here.

[1] OMG OWL Full and UML 2.0 Compared (2005) [Online] http://www.omg.org/cgi-bin/docs/ontology/04-03-01.txt.

[2] Klein T, Chang J, Cho M, Easton K, Fergerson R, et al. (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenom. J. 1:167–170.

[3] OMG Single Nucleotide Polymorphisms specification (2005) [Online] http://www.omg.org/cgi-bin/doc/dtc/05-02-06.pdf.