SNP-Converter: An Ontology-Based Solution to Reconcile Heterogeneous SNP Descriptions for Pharmacogenomic Studies

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
{coulet, malika, napoli, devignes}@loria.fr
[3] Université Pierre et Marie Curie - Paris6, INSERM UMRS 538, Biochimie – Biologie Moléculaire, Paris, France
pascale.benlian@sat.ap-hop-paris.fr

Abstract. Pharmacogenomics explores the impact of individual genomic variations in health problems such as adverse drug reactions. Records of millions of genomic variations, mostly known as Single Nucleotide Polymorphisms (SNP), are available today in various overlapping and heterogeneous databases. Selecting and extracting from these databases or from private sources a proper set of polymorphisms are the first steps of a KDD (Knowledge Discovery in Databases) process in pharmacogenomics. It is however a tedious task hampered by the heterogeneity of SNP nomenclatures and annotations. Standards for representing genomic variants have been proposed by the Human Genome Variation Society (HGVS). The SNP-Converter application is aimed at converting any SNP description into an HGVS-compliant pivot description and vice versa. Used in the frame of a knowledge system, the SNP-Converter application contributes as a wrapper to semantic data integration and enrichment.

1 Introduction

One of the great challenges in the post-genomic area consists in exploring the involvement of individual genomic variations in biological processes. Technical advances in high-throughput genotyping enable rapid sampling of thousands of genotypes. 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, very few are known to have an obvious pathological effect. These are named *mutations*. More general terms, such as *polymorphism* or *variant*, are preferred to characterize the general concept of variation [1]. Around 90% of the genome variations are limited to one-nucleotide substitutions (for example a guanine replaces a thymine at a given position in the genome) designated as single nucleotide polymorphism or SNP.

The challenge mentioned above, i.e. to explore the involvement of individual genomic variations in biological process, can be considered as a data mining problem.

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 82 – 93, 2006.

Knowledge discovery in databases (KDD) is a process aimed at extracting from large databases information units that can be interpreted as knowledge units [2]. This process comprises three major steps: (i) the selection and preparation of data, (ii) the data mining operation, and finally (iii) the interpretation of the extracted units. Various integration problems may arise along the process. The first step often requires to integrate data from public and private databases in order to guide the selection step or to enrich the selected set of data. The last step also necessitates to assess the extracted information units with respect to existing knowledge [3]. In both cases, integration tasks will consist in establishing equivalence, consistency or discrepancy between data or concepts, as well as classifying new data or concepts among existing ones. This type of integration should therefore rely on a semantic conceptual frame in which reasoning mechanisms are available. Indeed, ontologies contribute to build such an environment [4].

An ontology is a formalization of a conceptualisation [5], that is to say the definition and the representation for a given domain of concepts and their relationships allowing human and machine agents to share knowledge about this domain, and to reason with respect to this knowledge. By providing a semantic conceptual frame to a data mining process, an ontology should play a valuable role to facilitate data integration as well as knowledge acquisition.

Pharmacogenomics is a multi-dimensional domain where genome variations, phenotypic data and drug properties can be mined together in order to find out possible associations of variations with individual good or adverse drug responses [6]. More and more pharmaceutical firms are willing to include the exploration of particular genomic variants in their drug clinical trials in order to detect relationships between the following three summits of the pharmacogenomics triangle (Figure 1): (1) drug (properties and administration), (2) phenotype (biological and clinical data), and (3) genotype (genome variations).

Integration of the genotype dimension in clinical trials is not straightforward partially because of the large number of variants present all along the genome. Indeed many genes contain more SNPs than can be conceivably genotyped in current studies. Thus the choice of a relevant subset of SNPs to be included in studies should be somehow guided. A knowledge base called PharmGKB participates in this effort by offering a repository for storing experimental data sets related to pharmacogenomic studies [7].

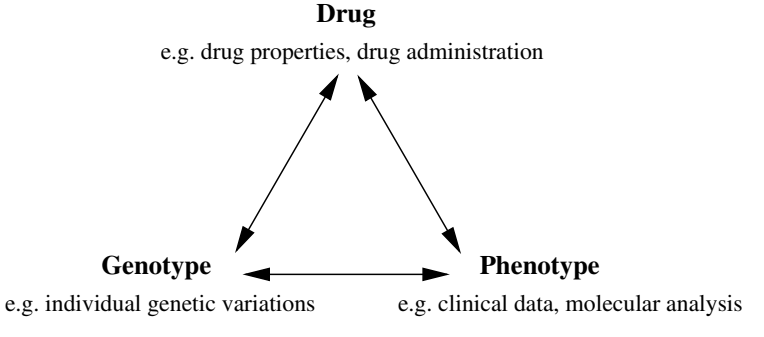

Fig. 1. Triangular schematization of the pharmacogenomics domain