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

Research Paper

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. 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].

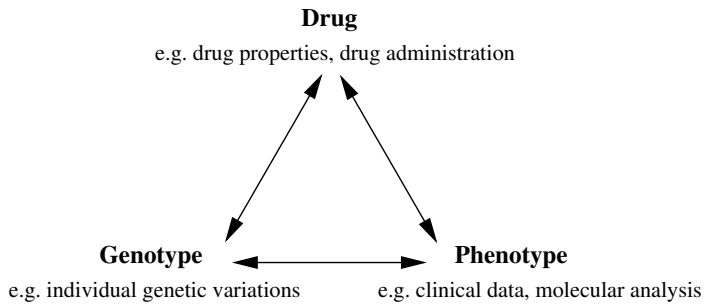

Fig. 1. Triangular schematization of the pharmacogenomics domain.

The present research work focuses on the genotype summit of the pharmacogenomics triangle since its complexity is often underestimated, and since major difficulties arise when locally observed genotype data 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 easy to achieve [8].

In [9], we have introduced the SNP-Ontology represented in the OWL language as a contribution towards building a semantic frame for pharmacogenomic studies. Our purpose is to use this ontology to formally represent the knowledge on genomic variants (i.e., SNP-knowledge base) as the first step of a KDD process as in [10, 11]. We thus developed the SNP-Converter application which acts as a wrapper for entering variant individuals in the SNP-knowledge base, starting from data extracted from various SNP-related databases as in [12, 13].

The section 2 introduces the various SNP representations and the existing attempts for integration. Section 3 presents the SNP-Converter application: rationale and functionalities. Usage of this application in the frame of the SNP-knowledge base is described in section 4. Section 5 discusses the issues of the solution presented here, and the perspectives of this work in terms of contribution to future pharmacogenomic studies are proposed.

2 Heterogeneity and Integration of Genomic Variations Data

2.1 Heterogeneous Representations of Genomic Variations

By definition, a genomic variation is originally associated to a position in a genomic (chromosome) sequence. However, when it affects a transcribed region, it is propagated to transcript sequence and, if the position is in a coding region, to protein

sequence. Variation databases indifferently represent variations in DNA, RNA or protein. Thus, they represent as well the original variation and its repercussions. For illustration, the substitution of a guanine by a thymine can be represented by G/T in the DNA sequence, GGC/GTC in the affected codon, g/u in the corresponding RNA, Gly/Val in the translated protein. In addition, the representation within the databases of the variant position differs depending on the reference sequence (and its version) used to locate it. Let us take an example : the G/T substitution is at position 11,087,877 in the chromosome 19 sequence, which has the accession number NC_000019 in the RefSeq database, at position 2,489,679 in the NT_011295 contig sequence, and at position 565 in the NP_000518 protein sequence (on the second nucleotide of the codon). The substitution can also be localized at position 26,747 in one of the associated gene sequences, or at position 108 in the eleventh exon of this gene. Various syntaxes can be used to represent these variants, which are also often referred by their accession numbers in given databases. For example, the variant described above would be cited in the PharmGKB database as the G/T variant at position chr19:11087877, and in the dbSNP database as the rs28942082 polymorphism. A generic syntax has been recommended by the Human Genome Variation Society (HGVS). According to this proposed standard, our variant should be described by the following expression: NC_000019.8:g.11087877G>T, where NC_000019.8 is the unique accession number (in the NCBI RefSeq database) of the sequence used to position the variant, the letter g means that the sequence is genomic, by opposition to p for instance which is used for a protein sequence, 11087877 corresponds to the position in the referred sequence, and G>T describes the substitution itself (http://www.hgvs.org/mutnomen/recs.html) [14]. However this nomenclature has not been universally adopted yet. Previous nomenclatures sometimes subsist for historical reasons. For example our variant is still found in OMIM as the “FH NAPLES” or “Gly544Val”, that is to say with denominations related to the historical context of its discovery. In addition, private and disease- or locus-specific databases continue using non-conventional representations that enlarge the set of possible nomenclatures. Figure 2 illustrates the numerous alternative manners of designating a unique genome variant in private and public databases. It is worth noting that some of the non-conventional notations (c) are ambiguous: the first one does not mention the reference nucleotide, the third and fourth ones refer to two different versions of the same protein.

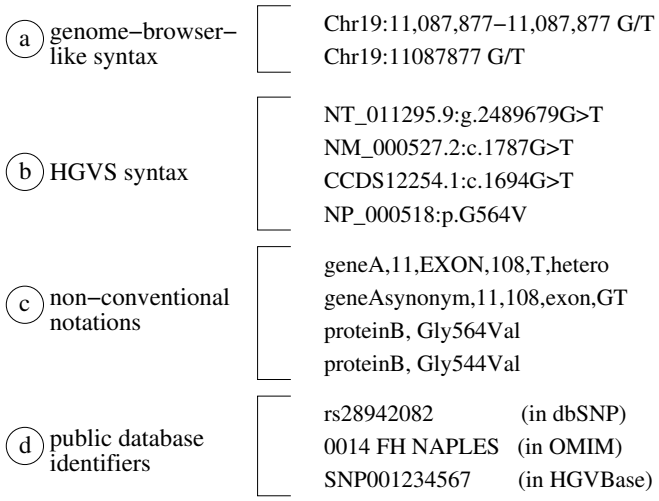

Fig. 2. Various notations or references for the same variant.

Finding intersection between several genomic variation databases is a critical issue for genetic diagnosis and "variome" exploration [15, 16]. However, as shown above, this task is not easy because of the amount of alternative and equivalent representations. Thus a system capable of establishing equivalence i.e. aligning between the different representations of a given variant is needed for investigating genome variations, and for being a basis for further pharmacogenomic studies.

2.2 Integrated Solutions

A first solution for solving the problem of heterogeneous representation of genomic variations is to build integrated databases providing a single access to variants pertaining from various sources. The NCBI dbSNP database lists over 9 million human polymorphisms, and constitutes the largest source of variants over the web [17]. Indeed, together with directly submitted SNP data, dbSNP integrates data from other large public databases of variants such as the NCI CGAP-GAI database, the TSC (The SNP Consortium, Ltd) variation initiative, HGVBase, HapMap, PharmGKB, Perlgen. Furthermore, dbSNP is fully integrated with NCBI databases (GenBank, PubMed, LocusLink, Human Genome Project Data) leading to a rich set of data.

HGVbase (Human Genome Variation Database, formerly HGBase) is the product of a collaboration between the Karolinska Institute (Sweden) and the European Bioinformatics Institute (UK). It has been constructed as a means for gathering polymorphisms from all possible public sources [18]. Thanks to both collection and submission, this relational database is cataloguing more than 8 million polymorphisms and proposes interesting text-based search facilities. HGVbase has been interfaced with SRS (Sequence Retrieval System). An originality of this work is that the authors pro-

pose the first controlled vocabulary, the Mutation Event Controlled Vocabulary[1], to facilitate polymorphism data integration. Each HGVbase record contains all the information necessary to re-construct the variant description in the HGVS standard syntax.

TAMAL (Technology And Money Are Limiting) is based on a materialized data warehouse that integrates five SNP sources (HapMap, Perlgen, Affymetrix, dbSNP and the UCSC genome browser), and that offers querying facilities through current versions of these resources (updated quarterly) in view of facilitating SNP selection for genetic study design [19]. To help selecting SNPs that are likely involved in the genetic determination of human complex traits, various properties of SNP have been integrated such as SNP localisation (in coding regions, in promoters) or haplotype tagging.

LS-SNP (Large-Scale annotation of coding non-synonymous SNPs) is an original work aimed at enriching dbSNP annotations of non-synonymous coding SNPs with information about protein sequences, functional pathways and comparative protein structure models in order to predict polymorphism impacts on produced proteins [20]. This resource can be a precious guide for SNP selection before a clinical study.

The pharmacogenomics knowledge base (PharmGKB) contains data sets linking genotype and phenotype information [7]. This integrated resource presents two major interests. First, original polymorphisms are directly submitted to PharmGKB as results of clinical trials, enabling to link them to individual clinical data. Second, PharmGKB allows extended navigation through cross-referenced sources such as NCBI databanks, UCSC Genome Browser and Gene Ontology. This makes PharmGKB a valuable resource for interactively enriching annotations on given variants. PharmGKB data are structured according to an XML schema that defines the relationships between the different handled objects. However, as far as we know, PharmGKB is not exploitable for automatic data extraction and mining.

This brief panorama of integrated databases in the domain of genomic variations shows that each project has to solve in some way the problem of integrating heterogeneous variant representations. Methods used are rarely explicited since they must fit the data model associated to the database, and cannot be reused for other purposes. More general propositions have been made to promote integration of variant representations. A controlled vocabulary (the Mutation Event Controlled Vocabulary quoted above) has been proposed by the HGVbase. The Polymorphism Markup Language offers the possibility of exchanging data on sequence variations [21]. The associated DTD (Document Type Definition) describes polymorphism variation, frequency, population, assay, submitter and publication. DDBJ and JBIC recommend the use of PML for interoperability of data on SNPs and other genomic variations. Under the supervision of the Object Management Group, the SNP object has been precisely specified [22]. This work takes into account a large view of the data linked to SNPs in existing data sources. The HGVS participates in this effort of knowledge representation as one of the rare propositions looking at the genetic variation concept, and not simply at the representations of variants in databases [14]. It should be noted that the unequivocal identification of genomic variants does not mean here unique identifier,

1 http://www.ebi.ac.uk/mutations/recommendations/mutclass.txt

since the generic syntax proposed by HGVS allows multiple references to various types of sequences (chromosomes, contigs, transcripts, proteins). Finally, the XML PharmGKB schema presented as an ontology by the authors includes the representation of domain concepts and their relationships in a structured formalism [23].

2.3 Semantic Integration

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, and genetic code. This is one reason leading us to the design of the SNP-Ontology [9]. Indeed a specific ontology in the field of genomic variations is useful, because it embodies the abstract knowledge required for data integration and analysis. Existing initiatives mentioned above such as the PharmGKB ontology and the OMG SNP specification contributed to the early stages of this work. Several additional concepts were defined to provide the SNP-Ontology with the capacity of hosting any variant, whatever their description, as individuals instantiating the ontology concepts and properties. The SNP-Ontology has been coded with the OWL (Web ontology language) formalism and edited with the Protégé knowledge base framework [24][25]. OWL is the standard representation language for the semantic web. Its foundations are both description logics and web standard languages (XML and RDF-S). It allows building a knowledge base equipped with reasoning mechanisms such as subsumption, classification, consistency checking and instantiation. These mechanisms once plugged in Protégé lead to new inferred knowledge that can enrich the knowledge base. However integrating variant descriptions requires handling of concrete data (e.g. string, integer) that is not yet fully allowed by the description logic framework [26]. Thus, we have developed the SNP-Converter application that can be used either as a standalone application for format conversion purposes, or in the frame of an ontology-driven knowledge base for integrating datasets of genomic variants.

3 The SNP-Converter

3.1 Inputs and Outputs

A variant is considered as an observed variation located at a specific position along a sequence. The observed variation can be a nucleotide or an amino acid variation depending whether the sequence serving as reference for localisation is nucleic acid or protein. This definition, that follows the HGVS nomenclature standard, leads to represent a variant by four features:

(i) the identifier of a reference sequence (i.e. its accession number in a public sequence database) ;

(ii) the type of concerned sequence (genomic, coding : cDNA, mRNA or protein coded respectively by g., c., r. or p. according to HGVS standards) ;

(iii) the position of the variant in the reference sequence ;

(iv) the observed variation (G/T, G >-, ->T, GT>AG, g>u, Gly>Val, etc.).

Conjunction of these four features yields an unequivocal representation of the variant. As mentioned above, a given variant can be represented by several sets of features depending on the selected reference sequence identifier. The core of the SNP-Converter application takes a set of four features as input, and converts them into an alternative set of four features representing the same variant. Because most representations do not explicitly provide the input features, a data preparation step is embedded in the SNP-Converter application. Present implementation of the preparation step allows the extraction of input features from dbSNP, HapMap, HGVBase records (in XML format) and from flat files or spreadsheets of two private databases that follow non conventional notations such as the first and second ones in figure 2.c. Reciprocally the converted output features may be processed to comply with the output format adapted to their envisaged usage. The SNP-Converter is currently able to produce several output formats: a simple text file using HGVS nomenclature, dbSNP XML and submission file formats. A typical scenario of SNP-Converter usage is the conversion of interesting SNPs from a private database into the dbSNP submission file format.

3.2 The Conversion Process

The SNP-Converter process, shown in Figure 3a, can be decomposed into 4 steps: (1) data preparation, (2) conversion of the four input features into pivot features, (3) an optional additional conversion into specific output features, and finally (4) the edition of output data. A simple instantiation of this process is illustrated in Figure 3b.

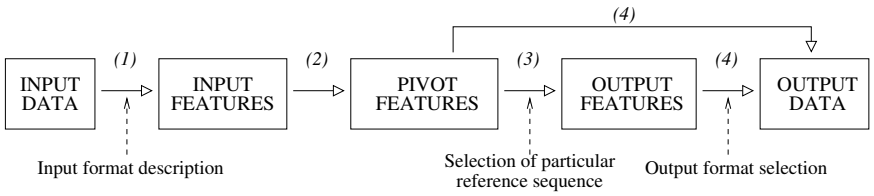

Fig. 3a. The SNP-Converter global process.

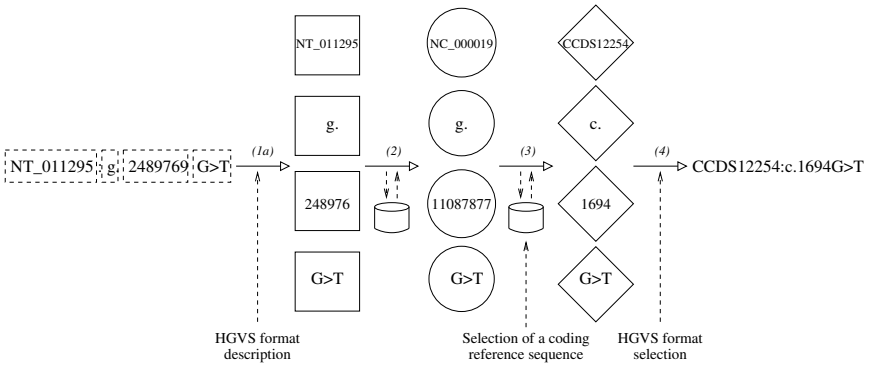

Fig. 3b. Illustration of the enactment of the SNP-Converter process on a given variant representation.

(1) The data preparation step consists in extracting the four input features from input data and depends on each specific source format. This preparation step also depends on whether the variant description is explicit (e.g. Genome-browser-like syntax or HGVS syntax) or implicit (e.g. database identifier). (1a) When the description is explicit, the four input features can be directly extracted by parsing the description according to a format-specific scheme. (1b) When the description is implicit, input data should first be completed in view of extracting input features. For example if the input data is a dbSNP identifier, it can be used to query the database and extract from the variant record the explicit data composing the input features.

(2) Pivot features consist in the particular set of features obtained for a given variant when the reference sequence is the complete chromosome sequence (RefSeq accession number, e.g. NC_000019.8) that includes the input reference sequence. Since the pivot sequence type is genomic, the variant position and the nature of observed variation must be computed. The input reference sequence is first localized on the complete chromosome sequence using alternative data sources. For instance the relative position of a gene can be found thanks to the gene symbol in the RefSeq complete chromosome entry ("FEATURES/gene" section). Exon genomic positions can also be retrieved in the "FEATURES/mRNA" section. When the variant position is expressed relative to translation start (ATG), genomic coordinates of coding sequence can be retrieved from the NCBI CCDS database. The appropriate coordinate conversion can then be computed to finally produce the position of the variant relative to the complete chromosome sequence. Finally the observed variation must be converted into a variation at the genomic level. If the input variation is described at the DNA level, this feature remains unchanged. Alternatively, if the observed variation is at the mRNA level, uracil must be converted into thymine. An observed variation described at the protein level should be converted according to the genetic code. Due to the genetic code degeneration, several codons can code for the same amino acid. Thus the conversion from amino acid to nucleic acid variation can lead to more than one set of features. The SNP-Converter outputs all these possibilities.

(3) The next step of reference features conversion is optional since it appears unnecessary when the desired output format fits to the pivot features. If not, the output reference sequence should be selected by the user, and can be DNA, cDNA, mRNA, or protein. The conversion process then follows the same rationale as the previous one to produce new relative position and observed variation in the new reference representation.

(4) Output data should finally be formatted depending on the purpose of the conversion. A first possibility is to edit the output features according to the HGVS syntax or any other syntax. A second possibility is to build a variant description in a specific format for database submission. Finally, another interesting possibility is to enter the output data in a knowledge-base-compliant formalism such as OWL to allow its assertion in a knowledge base (see below).

The SNP-Converter is implemented as a Java application, and has been tested on a set of variants composed of the dbSNP variants mapping on chromosome 19, and of variants extracted from a private database. For this purpose, dbSNP variants were extracted from downloaded dbSNP XML file and other variants from private text files. The goal was to find which variants from the private database were missing in the dbSNP database. The SNP-Converter application allowed us to compare the pivot features of the private variants with those of dbSNP variants. This experience allowed us to determine the overlapping coverage of both databases, and to identify several variants that were not yet submitted to dbSNP.

4. The SNP-Converter as a Wrapper for Semantic Integration

The SNP-Ontology (see section 2.3) plays the role of a coherent domain-specific global schema for a knowledge-base. We have made a mapping between the four features handled by the SNP-Converter and the SNP-Ontology concepts allowing the SNP-Converter to assert variants as individuals in the knowledge base. Since these four features are extracted from input data, the whole process leads to an indirect mapping of source schemas on the ontology. In practice we found relevant for any new variant, to insert in the SNP-knowledge base, not only its original set of features (for sake of traceability), but also the pivot features computed by the SNP-Converter. Thanks to these pivot features, the SNP-Converter is capable of qualifying as equivalent variants initially represented by distinct descriptions (see Fig. 4). The equivalence checking performed by the SNP-Converter is used here as a procedural extension of description-logics-based reasoners, aimed at enriching the knowledge base.

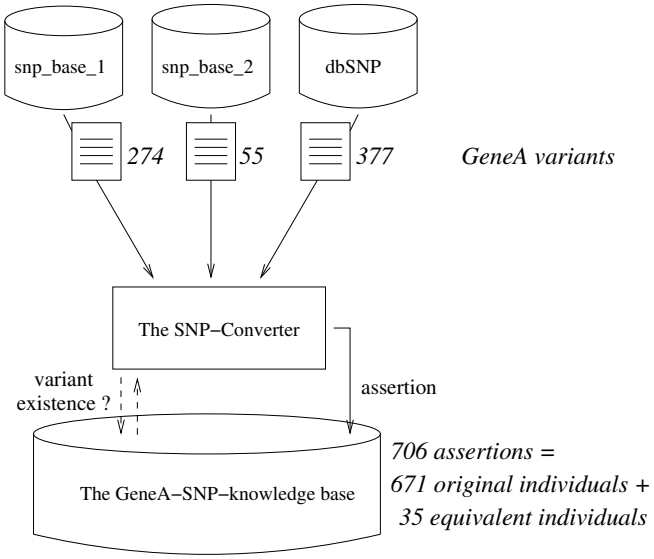

Fig. 4. Schematization of the use of the SNP-Converter application as a wrapper coupled to a knowledge base.

Figure 4 also shows the result of an experience carried on variants of a specific gene (named here geneA). Three sets of data were processed by the SNP-Converter application : 274 and 55 variants from private databases snp_base_1 and snp_base_2 respectively, and 377 variants from dbSNP. Among the 706 assertions created by the wrapper, 671 could be qualified as original individuals, and 35 were found equivalent to existing individuals.

5. Conclusion and Discussion

The SNP-Converter application has been developed to face the heterogeneity problem in genomic variation representation. The SNP-Converter can be used standalone to pass from one variation description to another. As such it constitutes a valuable tool for several use-cases: confronting private and public variant data, preparing submission of new variants to public databases, facilitating variant annotation retrieval from heterogeneous databases, guiding the choice of relevant variants to include in clinical trials, etc. The core of the SNP-Converter was designed to be generic thanks to the mapping with the SNP-Ontology. However, the handling of new sources requires some ad hoc adaptations for driving the extraction of the input features. This task will be facilitated by an administration interface. It should be noted that the SNP-Converter works with constant RefSeq versions and therefore is faced to the common problem of managing sequences pertaining to different assemblies.

With respect to the KDD process, our objective is to settle a semantic frame facilitating semantic data integration, data mining and incremental knowledge acquisition. In particular we consider semantic data integration as the design of an ontology-based knowledge base. This work demonstrates the importance and necessity of adequate wrappers preceding the semantic data integration stage as a consequence of the limits of existing knowledge management tools.

Our methodology differs from already described integrated solutions (see Sect. 2) and more general ones such as BioMart [27] or YeastHub [28] since most of these approaches are limited to 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 constitutes a proof of concept limited to one of the pharmacogenomics triangle summits (see Fig. 1), and to the first step of the KDD process. Nevertheless it allows us to proceed in the data mining process. Complete demonstration will necessitate extension of the ontology to include the two other summits (drug and phenotype) and the testing of our methodology for these two domains.

Acknowledgement

This work has been partly funded by the EUREKA-labeled GenNet research and development contract between KIKA medical, PhenoSystems and Loria-CNRS. AC benefits from a CIFRE fellowship. Special thanks to Romain Demoustier from KIKA medical and to David Atlan from Phenosystems for stimulating discussions.

References

1. Kruglyak,L., Nickerson,D. Variation is the spice of life. Nat Genet. 27, 3 (2001) 234-6.
2. Frawley,W., Piatetsky-Shapiro,G., Matheus,C. Knowledge Discovery in databases: An Overview. Knowledge Discovery in Databases, AAAI/MIT Press.(1991) 1-30.
3. Janetzko,D., Cherfi,H., Kennke,R., Napoli,A., Toussaint,Y. Knowledge-based Selection of Association Rules for Text Mining. 16h European Conference on Artificial Intelligence, ECAI'04, Valencia (2004).
4. Vetere,G., Lenzerini,M. Models for Semantic. Interoperability in Service Oriented Architectures, IBM Systems. Journal, 44 (2005).
5. Gruber,T.R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition. 5 (1993) 199-220.
6. Evans,W., Relling,M. Pharmacogenomics: moving toward individualized medicine, Nature. 29 (2004) 464-468.
7. Klein,T., Chang,J., Cho,M., Easton,K., Fergerson,R., Hewett,M., Lin,Z., Liu,Y., Liu,S., Oliver,D. et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenom. J. 1 (2001) 167–170.
8. Marsh,S., Kwok,P., McLeod,H. SNP databases and pharmacogenetics: great start, but a long way to go. Hum Mutat. 20, 3 (2002) 174-9.

9. Coulet,A., Smaïl-Tabbone,M., Napoli,A., Benlian,P., Devignes M.D. SNPOntology for semantic integration of genomic variation data. ISMB 2006, Fortaleza. [OnLine]. https://hal.inria.fr/inria-00067863
10. Anand,S., Bell,D., Hughes, J. The role of domain knowledge in data mining, Conference on Information and Knowledge Management CIKM'95, Baltimore, USA (1995).
11. Euler,T., Scholz,M. Using Ontologies in a KDD workbench, ECML/PKDD'04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa (2004).
12. Catarci,T., Lenzerini,M. Representing and using inter-schema knowledge in cooperative information systems. Journal of Intelligent and Cooperative Information Systems. 2 (1993) 375-398.
13. Levy,A. Logic-Based Techniques in Data Integration Logic Based Artificial Intelligence. Jack Minker. Kluwer Publishers (2000).
14. den Dunnen,J., Antonarakis,S. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 15 (2000) 7–12.
15. den Dunnen,J., Paalman,M. Standardizing mutation nomenclature: why bother? Hum Mutat. 22 (2003) 181–182.
16. Cotton,R.G.H., Kazazian,H.H. Toward a human variome project. Hum Mutat. 26,6 (2005) 499.
17. Sherry,S., Ward,M., Sirotkin,K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. 9 (1999) 677–679.
18. Fredman,D., Munns,G., Rios,D., Sjoholm,F., Siegfried,M., Lenhard,B., Lehvaslaiho,H., Brookes,A. HGVbase : a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res. 32 (2004) D516-9.
19. Hemminger,B., Saelim,B., Sullivan,P. TAMAL: an integrated approach to chosing SNPs for genetics studies of human compex traits. Bioinformatics. 22 (2006) 626-627.
20. Karchin,R., Diekhaux,M., Kelly L., Thomas D., Pieper,U., Eswar,N., Haussler,D., Sali,A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 21 (2005) 2814-2820.
21. Sugawara,H., Mizushima,H., Kano,T., Shigemoto,Y., Hashimoto,Y., Tomabechi,I., Sakagami,N. et al Polymorphism Markup Language (PML) for the interoperability of data on SNPs and other sequence variations, 19th International CODATA Conference (2004).
22. OMG Single Nucleotide Polymorphisms specification (2005) [Online] http://www.omg.org/cgi-bin/doc/dtc/05-02-06.pdf.
23. Oliver,D., Rubin,D., Stuart,J, Hewett,M., Klein,T., Altman,R. Ontology development for a pharmacogenetics knowledge base. Pac Symp Biocomput. (2002) 65-76.
24. Horrocks,P., Patel-Schneider,F., van Harmelen,F. From SHIQ and RDF to OWL: The making of a web ontology language, Journal of Web Semantics, 1, 1 (2003) 7-26.
25. Noy,N., Sintek,M., Decker,S., et al. Creating Semantic Web contents with Protege-2000. IEEE Intelligent Systems 16. (2001) 60-71.
26. W3C Web Ontology Working Group (WOWG), (2004) Owl web ontology language semantics and abstract syntax. W3C recommendation [Online]. http://www.w3.org/TR/owl-ref/.
27. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. Birney,E. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. Genome Res. 14 (2004) 160-169.
28. Cheung,K.H., Kevin Y. Yip,K.Y., Smith,A., deKnikker,R., Masiar,A., Gerstein,M. YeastHub: a semantic web use case for integrating data in the life sciences domain**,** Bioinformatics.21 (2005) i85-i96. 28.