

# Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships

Adrien Coulet<sup>1,2\*</sup>, Malika Smaïl-Tabbone<sup>2</sup>, Pascale Benlian<sup>3</sup>, Amedeo Napoli<sup>2</sup> and Marie-Dominique Devignes<sup>2</sup>

<sup>1</sup>KIKA Medical, 35 rue de Rambouillet, Paris, France

<sup>2</sup>LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP), Campus Scientifique, Vandoeuvre-lès-Nancy, France

<sup>3</sup>Université Pierre et Marie Curie - Paris6, INSERM UMRS 538, Biochimie-Biologie Moléculaire, Paris, France

\*Corresponding author: adrien.coulet@loria.fr

## Abstract

Complexity of post-genomic data and multiplicity of mining strategies are two limits to Knowledge Discovery in Databases (KDD) in life sciences. Because they provide a semantic frame to data and because they benefit from the progress of semantic web technologies, bio-ontologies should be considered for playing a key role in the KDD process. In the frame of a case study relative to the search of genotype-phenotype relationships, we demonstrate the capability of bio-ontologies to guide data selection during the preparation step of the KDD process. We propose three scenarios to illustrate how domain knowledge can be taken into account in order to select or aggregate data to mine, and consequently how it can facilitate result interpretation at the end of the process.

## 1 Background

One of the promising interests of bio-ontologies is their use for guiding the process of Knowledge Discovery in Databases (KDD). The KDD process has been successfully used in various domains such as finance or biomedicine [8]. However application cases are limited by the fact that it still necessitates a close interaction between the system and domain experts. Because the data manipulated in life sciences are complex, and because data mining algorithms generate large amounts of raw results, the interpretation step of KDD in biology is particularly tricky and discouragingly time-consuming.

Existing bio-ontologies help in giving structure to the amounts of complex data in life sciences. The National Center for Biomedical Ontology (NCBO) has recently developed the Bioportal [3] that provides a unified panorama of bio-ontologies from a unique web site [12]. In computer science, ontologies formalise a shared understanding of knowledge about a particular domain [9].

All along the KDD process, domain knowledge, embedded within ontologies, can be used for guiding the various steps [2]:

- (1) During the data preparation step, it facilitates integration of heterogeneous data, and guides the selection of relevant data to mine,
- (2) During the mining step, domain knowledge allows specifying constraints for mining algorithms,
- (3) In the interpretation step, it helps experts for validating the extracted knowledge units before integrating them.

We distinguish here the data mining, which is limited to the execution of a mining algorithm, from the whole KDD process that includes data preparation and result interpretation.

In a previous work we stressed on the role of ontologies in data integration [5]. In this paper we focus on the selection of data that deserve mining. This step is usually performed by experts who use their own knowledge for selecting the most relevant data. We explore here how the availability of an ontology and its associated Knowledge Base (KB) can assist the expert in his task.

Filling the gap between genotype and phenotype is of principal interest in biology research. Large scale clinical studies enable the recording of many genomic and post-genomic data

thanks to new biotechnology tools (microarray, mass spectrometry, etc.). Recent studies [4,11] show that data mining methods are promising approaches for the exploration of correlations between genotype and phenotype data inside the large amounts of data that result from this kind of studies. However it also illustrates that such analyses are particularly delicate to manage because of the two difficulties mentioned previously: domain complexity and amount of data.

Section 2 presents a case study, with a dataset, an ontology and a KB that will be used for the data selection. Section 3 proposes three different scenarios that make use of the ontology in order to guide data selection. Section 4 discusses the obtained results and concludes on this work.

## 2 A Case Study

### ***Searching for Variant-Phenotype Trait Relationships***

We illustrate the benefit of using an ontology for the data selection problem in KDD with a real biological problem and a real dataset relative to Familial Hypercholesterolemia (FH). KDD process is used here to reveal relationships between genomic variants and phenotype traits. Such relationships could then be used to identify modulator variants, i.e. any variant (or any group of variants) related to a modulation in the disease or in a disease symptom. For instance, depending on allele versions of two genomic variants of the APOE gene (rs7418 and rs429358) various levels of severity are observed in the FH. Modulator variants are particularly interesting in pharmacogenomics since they are known to modulate the activity of drug pathways, and consequently to modulate the drug effect.

### ***The FH Dataset***

Our dataset concerns:

- ( $\alpha$ ) patients affected by the genetic hypercholesterolemia (FH),
- ( $\beta$ ) patients affected by a non-genetic hypercholesterolemia, and
- ( $\gamma$ ) patients without any hypercholesterolemia.

Genotype attributes describe observed alleles for genomic variants of the LDLR gene. An example of genotype attribute is the observed allele for the variant located at position Chr19:11085058 (e.g. AA). Phenotype attributes describe traits usually observed when studying the metabolism of lipids. Two examples of phenotype attributes are the LDL blood concentration (e.g.  $[\text{LDL}]_b = 3 \text{ g.l}^{-1}$ ) and the presence of xanthoma. Table 1 describes quantitatively the FH dataset.

**Table 1: Characteristics of the FH dataset**

	<i>Nature</i>	<i>Number</i>	<i>Total number</i>
Objects	Patients	125	125
Attributes	Genotype	292	304
	Phenotype	12	

### ***Preliminary Data Mining***

A preliminary exploration consisted in submitting the whole FH dataset to two different unsupervised data mining algorithms. The first one is association rule search with the Apriori algorithm [1], the second one is a clustering with the COBWEB algorithm [7]. Apriori and COBWEB implementations used for all the reported mining tasks are those of the Weka toolbox [15].

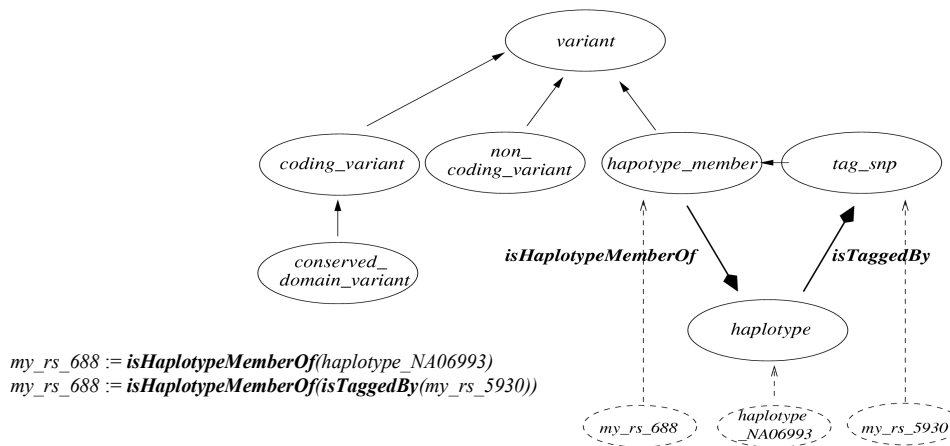
With Apriori (support=99%, confidence=0.9), the total number of itemsets for the generation of rules is 223, with 169 itemsets strictly larger than 2.

With COBWEB (acuity=1, cutoff=0.5), the total number of clusters is 187 (Table 2, 1st column).

These preliminary results revealed complex to interpret due to the large amount of variants involved and the lack of contextual data. For instance, variants (which are identified by a simple label) located in coding sequence are difficult to distinguish from those located in non coding sequence, similarly normal values of LDL blood concentration are difficult to distinguish from pathologic ones. In addition the existence of haplotypes generates noise in the form of trivial associations.

### SNP-Ontology and SNP-Knowledge Base about the Genomic Variations of the LDLR gene

The SNP-Ontology [13] embeds a formal description in OWL-DL of genomic variations and their related concepts: sequence in which they are observed, haplotype they belong to, proteins they modify, database in which they are stored, etc.. We populate for our case-study a SNP-Knowledge Base that follows the semantic structure of the SNP-Ontology and integrates knowledge about genomic variations of the LDLR gene. The method used to populate the SNP-KB is precisely described in [5]. We illustrate with the SNP-Ontology and the SNP-KB three different kinds of semantic operators that we use for guiding data selection (Figure 1): Subsumption, object property, and constraints on properties.



**Figure 1**  
**The SNP-Ontology and the SNP-KB.** Some classes of the SNP-Ontology and instances of the SNP-KB. Plain arrows represent subsumptions, dotted arrows represent instantiations, bold arrows represent object properties.

## 3 Guiding Data Selection with Domain Knowledge

### Attribute Selection thanks to Subsumption

The hierarchical structure of the ontology enables browsing top-down or bottom-up for a progressive selection of attribute. For instance the FH dataset can be progressively focused on more and more specific classes of variants. Table 2 shows the reduction of results on a data mining dealing first with all variants (*variant*), then focusing on variants from coding regions (*coding\_variant*), and finally on variants that induce a modification in conserved domain of proteins (*conserved\_domain\_variant*). Conversely, the way back permits to enlarge the

number of mined variants and to generalize associations that may have been revealed on smaller attribute sets.

**Table 2: Quantitative characterization of the complexity of data mining results depending on attribute selection**

	<i>variant</i>	<i>coding_variant</i>	<i>conserved_domain_variant</i>	<i>tag_snp</i>
Variants to Mine	289	231	150	176
Resulting Itemsets (L>2)	169	169	49	32
Resulting Clusters	187	178	82	46

### **Attribute Aggregation thanks to Object Properties**

Because relationships between instances in KB bring information about dependencies between corresponding attributes, they could serve to aggregate these attributes. Haplotypes illustrate appropriately this aggregation process. In simple words, a group of variants that segregate uniformly (the haplotype) could be replaced by a single variant (the tag-SNP). Since the SNP-KB does integrate haplotype information from the HapMap project [10], it enables to identify tag-SNPs and haplotype members (Figure 1). Thus attribute selection can be limited to tag-SNPs (*tag\_snp*) which considerably reduces the number of attributes (Table 2). The amount of results to interpret is therefore reduced not only because of decreasing attribute number but also because of reducing dependencies between selected attributes.

More generally, aggregation can be envisaged between attributes as soon as functional dependency exists between them (e.g. *date of birth* determines *age*). In the previous example, the haplotype definition is interpreted as a functional dependency between occurrences of *tag-SNP* and *haplotype members*. However, this interpretation is subordinated to the precision of haplotype construction and thus deserves attention.

### **Object Selection thanks to Class Description**

In contrast with the two previous scenarios dedicated to attribute (e.g. *variant*) selection, this subsection illustrates the object (*patient*) selection that reduces amount of data as well. Furthermore, this third scenario illustrates more particularly the selection of instances that correspond to a defined class description (including constraints in addition to subsumption and object properties).

In our FH case study we define groups of patients which are suspected to present specific genotype-phenotype profiles. For this purpose, we use the SO-Pharm ontology [14] which embeds knowledge about clinical studies in pharmacogenomics [6]. A SO-Pharm-KB has been populated with private data from the FH dataset according to the method mentioned in the last subsection of the section 2.

Classes and properties of SO-Pharm enable us to describe four classes of patients: one that already exists in SO-Pharm, and three others that have been specially defined.

```

patient (defined in SO-Pharm)
patient_α := patient  $\Pi$  presentsGenotypeItem (hasValue (LDLR_mutation))
patient_β := patient  $\Pi$  presentsGenotypeItem (hasValue (no_LDLR_mutation))
     $\Pi$  presentsPhenotypeItem (hasValue (high_LDL_in_blood))
patient_γ := patient  $\Pi$  presentsGenotypeItem (hasValue (no_LDLR_mutation))
     $\Pi$  presentsPhenotypeItem (hasValue(normal_LDL_in_blood))

```

Selecting instances from only one class enables to reduce the amount of data, and to mine the resulting subset in order to characterize it. This mining task may benefit from prior attribute selection and aggregation as detailed in previous scenarios.

## 4 Discussion and Conclusion

This paper demonstrates in the frame of a case study how domain knowledge captured in bio-ontologies and KB facilitates the KDD process. Proposed scenarios can be combined so as to define a KDD strategy in accordance with biomedical objectives. Additional scenarios can also be envisaged such as object aggregation. In our case study, object aggregation could consist in grouping together patients from the same family thereby retaining a unique representing member for each family.

The aggregation process is based on relationships between instances and is consequently dependent on data quality. If an instance is missing or is wrong, the aggregation will be impossible or wrong. However, knowledge about haplotypes could also serve for completing missing values about observed alleles of each haplotype member.

The next step of our work will deal with implementing a framework for guiding the data preparation in accordance with defined scenarios. This requires a knowledge base system that enables ontology edition, management of large OWL dataset, and instance retrieval, in order to ensure seamless junction between domain knowledge and data.

Challenging future works may focus on how to automatically formalize results of the KDD process in order to enrich both the ontology and the KB. Such a capability would enable to iteratively run the KDD process, using an enriched domain knowledge at each iteration.

## References

1. Agrawal R, Imielinski T, Swami AN: **Mining Association Rules between Sets of Items in Large Databases**. *SIGMOD* 1993. **22**(2):207.
2. Anand S, Bell D, Hughes J: **The Role of Domain Knowledge in Data Mining**. In *Proc of CIKM*, 1995. Baltimore, USA.
3. **Bioportal** [<http://www.bioontology.org/tools/portal/bioportal.html>]
4. Capriotti E, Fariselli P, Calabrese R, Casadio R: **Predicting Protein Stability Changes from Sequences Using Support Vector Machines**. *Bioinformatics* 2005. **21**: ii54-ii58.
5. Coulet A, Smaïl-Tabbone M, Benlian P, Napoli A, Devignes MD: **SNP-Converter: An Ontology-Based Solution to Reconcile Heterogeneous SNP Descriptions**. In *Proc of DILS* 2006. LNBI 4075: 82-93.
6. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes MD: **Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing**. In *Proc of KSinBIT* 2006. LNCS 4277: 648-57.
7. Fisher DH: **Knowledge Acquisition via Incremental Conceptual Clustering**. *Machine Learning* 1987. **2**:139-172.
8. Frawley W, Piatetsky-Shapiro G, Matheus C: **Knowledge Discovery in Databases: An Overview**. *Knowledge Discovery in Databases*, AAAI/MIT Press. 1-30; 1991.
9. Gruber TR: **A Translation Approach to Portable Ontology Specifications**. *Knowledge Acquisition* 1993. **5**, 199-220.
10. **HapMap** [<http://www.hapmap.org/>]
11. Li J, Zhou Y, Elston RC: **Haplotype-based Quantitative Trait Mapping Using a Clustering Algorithm**. *BMC Bioinformatics* 2006. **7**:258.
12. Rubin DL, Lewis SE, Mungall CJ *et al.*: **N.C.B.O.: Advancing Biomedicine through Structured Organization of Scientific Knowledge**. *OMICS* 2006. **10**:2, 185-198.
13. **SNP-Ontology** [[http://www.bioontology.org/files/6723/snponontology\\_full.owl](http://www.bioontology.org/files/6723/snponontology_full.owl)]
14. **SO-Pharm** [[http://www.loria.fr/~coulet/sopharm1.3\\_description.html](http://www.loria.fr/~coulet/sopharm1.3_description.html)]
15. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques**. 2nd Edition, Morgan Kaufmann, S-F; 2005.