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

Adrien Coulet, Malika Smaïl-Tabbone, Pascale Benlian, Amedeo Napoli and Marie-Dominique Devignes

Laboratoire Lorrain de Recherche en Informatique et ses Applications

(CNRS, INRIA, Nancy University),

Nancy, France

07/21/2006 – DILS'06

Outline

- Motivation
- The problem
- Ontology-based solution
- SNP-Converter
- Conclusion

Genomic variations (1/2)

- Human genome: 3.10^9 base pairs
  - 99,7 % constant
  - 0,3 % variant (10^6 positions that differ)
- Genomic variations
  - Polymorphism, variant, Mutation,
  - Single Nucleotide Polymorphism (or SNP)
    - Most genomic variations (90%) are SNPs
    - Frequency >1% in studied population

Patient 1: ...CACATC**G**ATCGA...
Patient 2: ...CACATC**A**ATCGA...

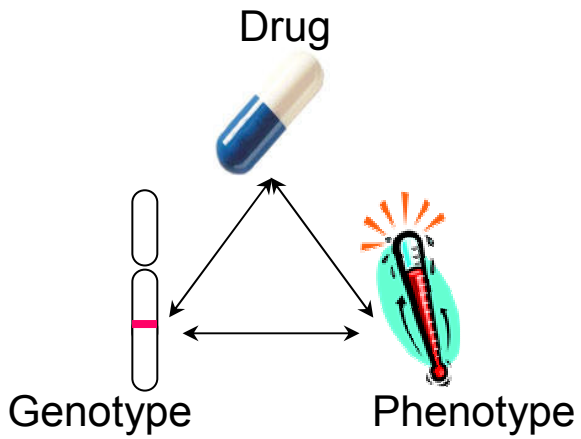
- Allow to study how genomic factors determine phenotype
  - with respect to environmental factors, *e.g. drug administration*

= Pharmacogenomics

Pharmacogenomics definition

- How genomic variations lead to variations in drug response
- Studies relationships between three domains

e.g. chemical properties, drug administration



e.g. individual genomic variations

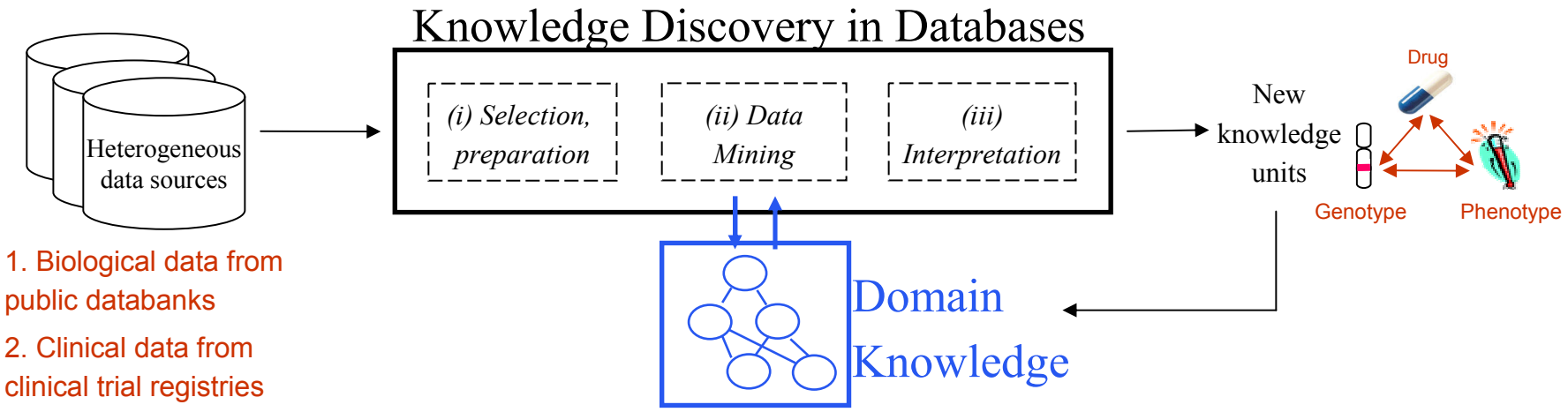
e.g. clinical data, molecular analysis

Examples in pharmacogenomics

- Codeine and CYP2D6
  - Codeine (opioid) must be metabolized into morphine to be active
  - Gene CYP2D6 code for the enzyme that performs this metabolism
  - 7% of Caucasians have an inactive variant of CYP2D6 enzyme
  - Codeine stays inactive in these patients

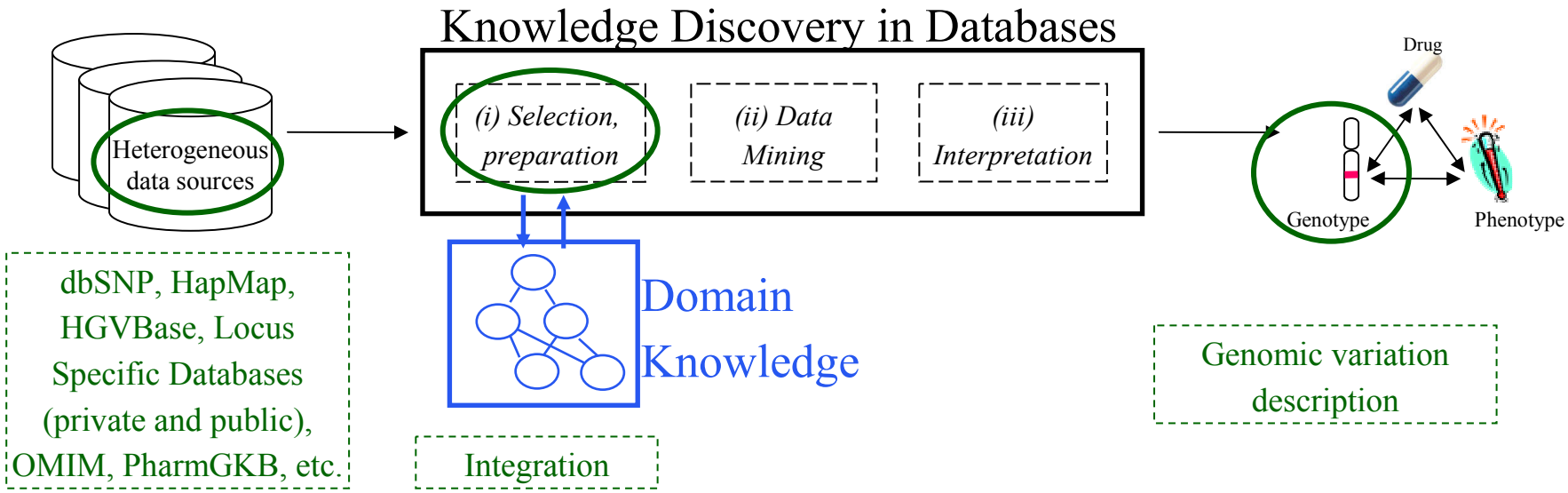
- Bidil
  - Drug for heart failure
  - Clinical trials => no benefits
  - Benefit in subgroup (African-descent)
  - Bidil approved by FDA for this subgroup

Pharmacogenomic Knowledge Discovery



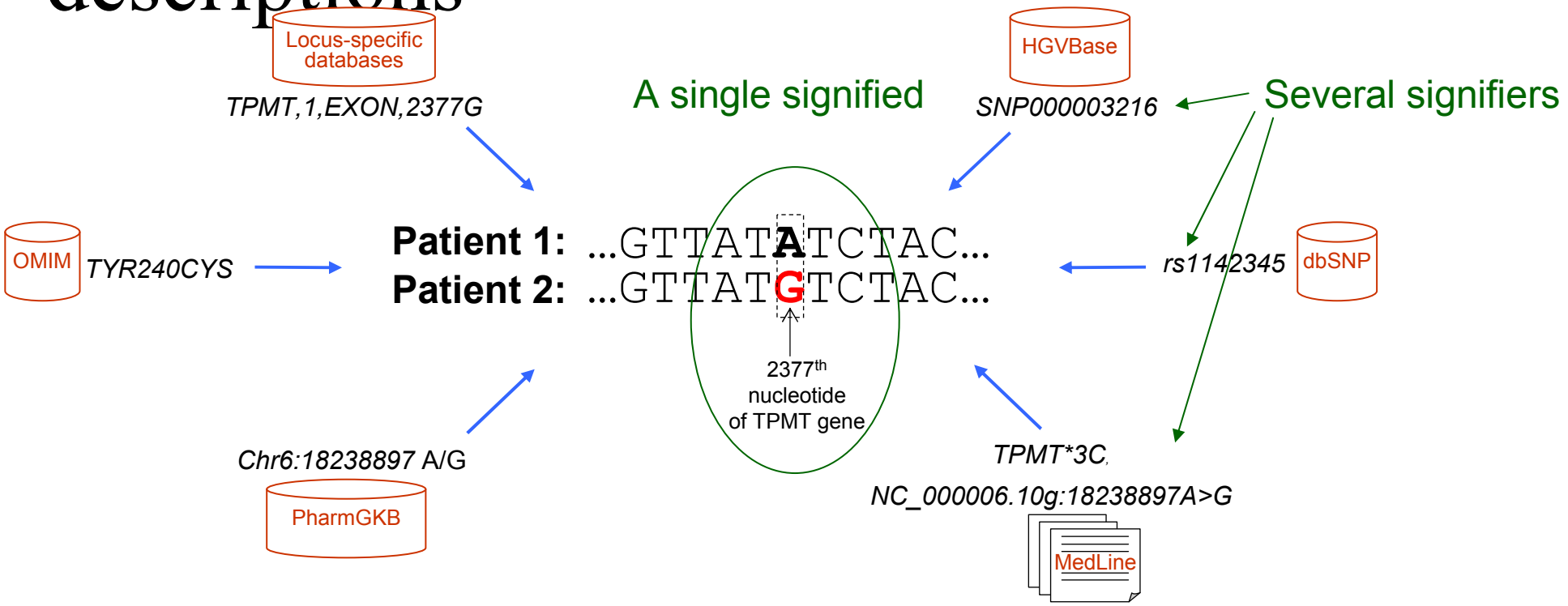
- Discovering new knowledge units thanks to
  - KDD process (Knowledge Discovery in Database) -- guided by --
  - Formal representation of the domain
- Applied to pharmacogenomics

Integration of genomic variation data



- We focus on the genotype vertex of pharmacogenomic triangle
- And on the first step of KDD process
- For a knowledge-based integration of genomic variation data

Heterogeneity in genomic variation descriptions



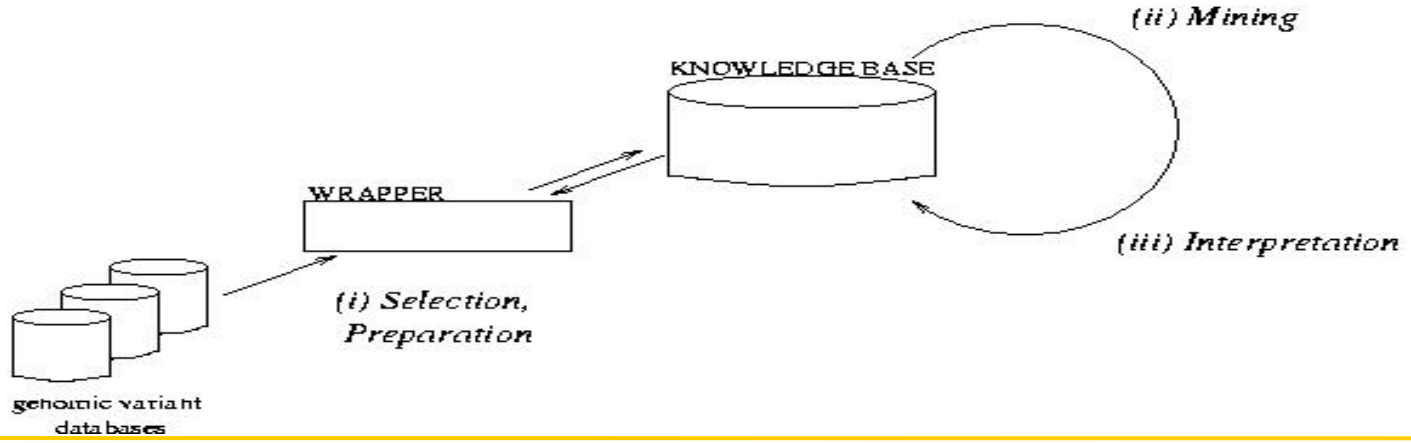
- Explicit *vs* implicit signifiers

Goals

- Our goal: To integrate genomic variation data
  - converting one genomic variation format into another,
  - establishing equivalence between variations displaying different representations

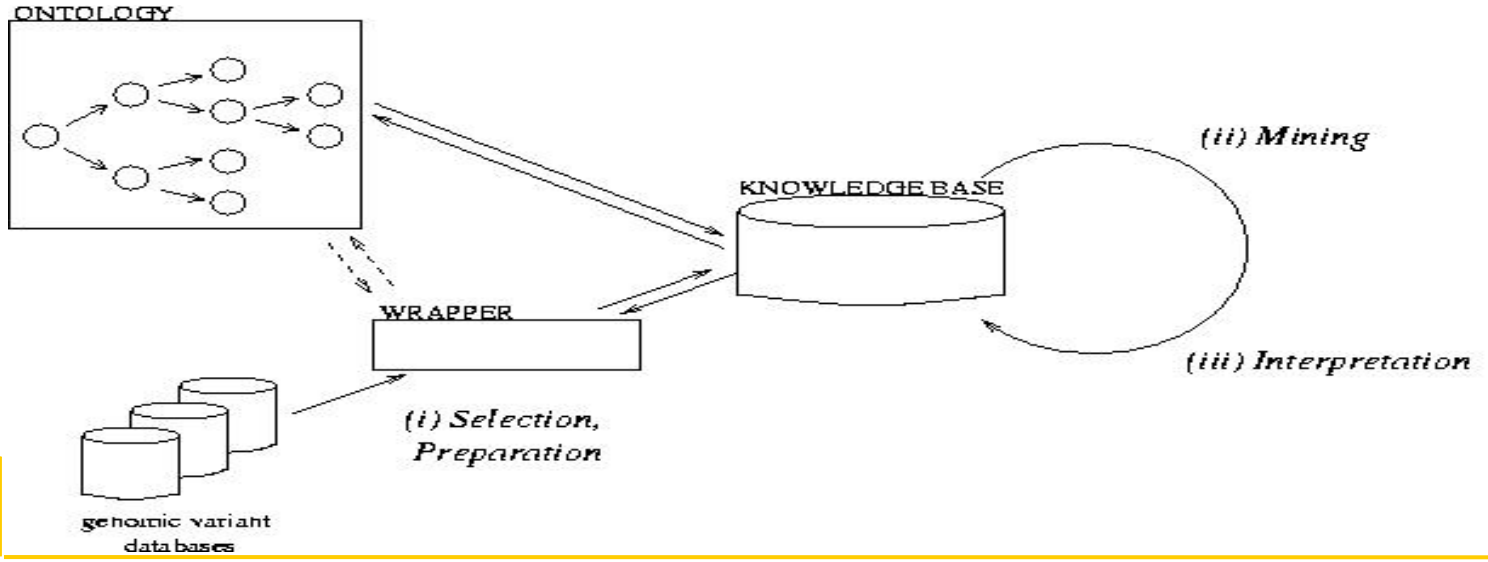
TPMT,1,EXON,2377G ≡ *NC_000006.10g:18238897A>G* ≡ *TPMT*3C* ≡ *TYR240CYS*

- For populating a knowledge base



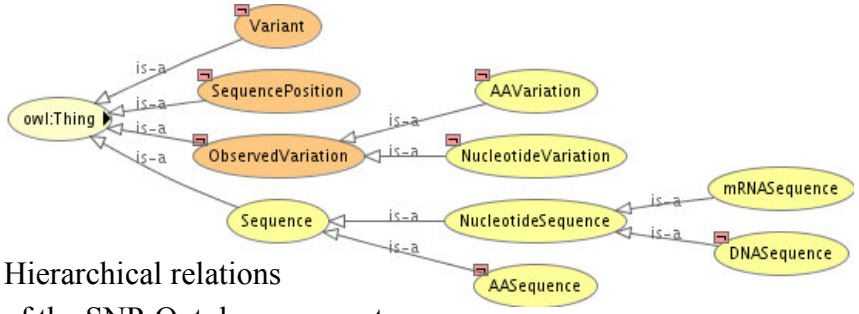
Ontology-based system

- Why using an ontology?
  - Limits of existing syntactic integration (*dbSNP*, *HGVBase*, etc.)
    - Static integration of some descriptions
    - Unable to automatically incorporate new data
  - Need of explicit domain knowledge
    - i.e. genome structure, transcript definition, genetic code, etc.



SNP-Ontology (1/3)

- What is an ontology?
  - An ontology is an explicit formalization of a conceptualization [Gruber *et al*, 1993]
  - An ontology is encoded within a knowledge representation language (associated to inference mechanisms)
- SNP-Ontology



Hierarchical relations of the SNP-Ontology concepts

$$\text{variant} \sqsubseteq (\forall \text{hasNucleotideVariation.nucleotide\_variation} \sqcap \forall \text{ hasPosition.}(\exists \text{ inRefSequence.nucleotide\_sequence}))$$

$$\text{nucleotide\_variation} \sqsubseteq \text{observed\_variation} \sqcap (\forall \text{ hasAlleleSequence.nucleotide\_sequence})$$

$$\text{nucleotide\_sequence} \sqsubseteq \text{sequence}$$

Formal definitions (in DL) of three concepts of the SNP-Ontology

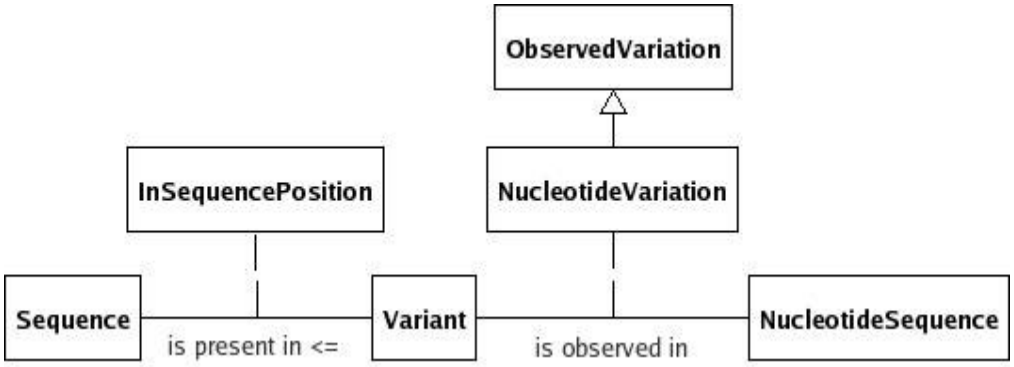
SNP-Ontology (2/3)

- Designed with an UML-based methodology

- Conceptualization:

  UML class diagrams of the genomic variation domain

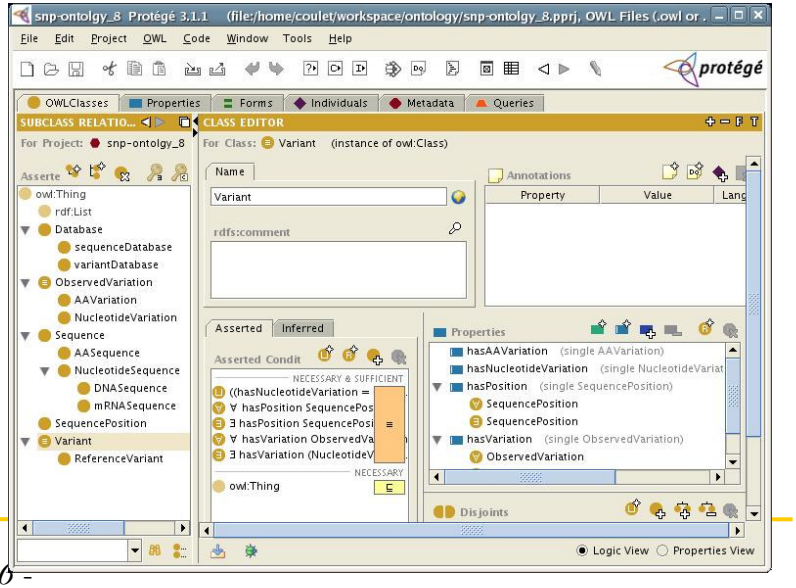
  - Classes, attributes, associations



- Formalization and implementation:

  The ontology is implemented in OWL-DL

  - Classes, datatype and object properties

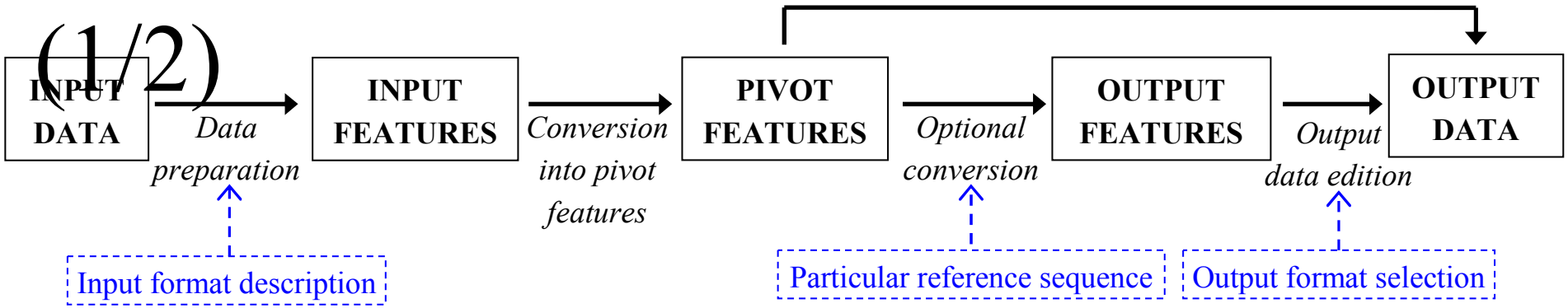


SNP-Ontology (3/3)

- A formal ontology designed with knowledge engineers and domain experts
- Implemented with the Protégé knowledge-base editor
- Available at www.loria.fr/~coulet/ontology/snpontology.owl
- To be submitted to the biological ontology community

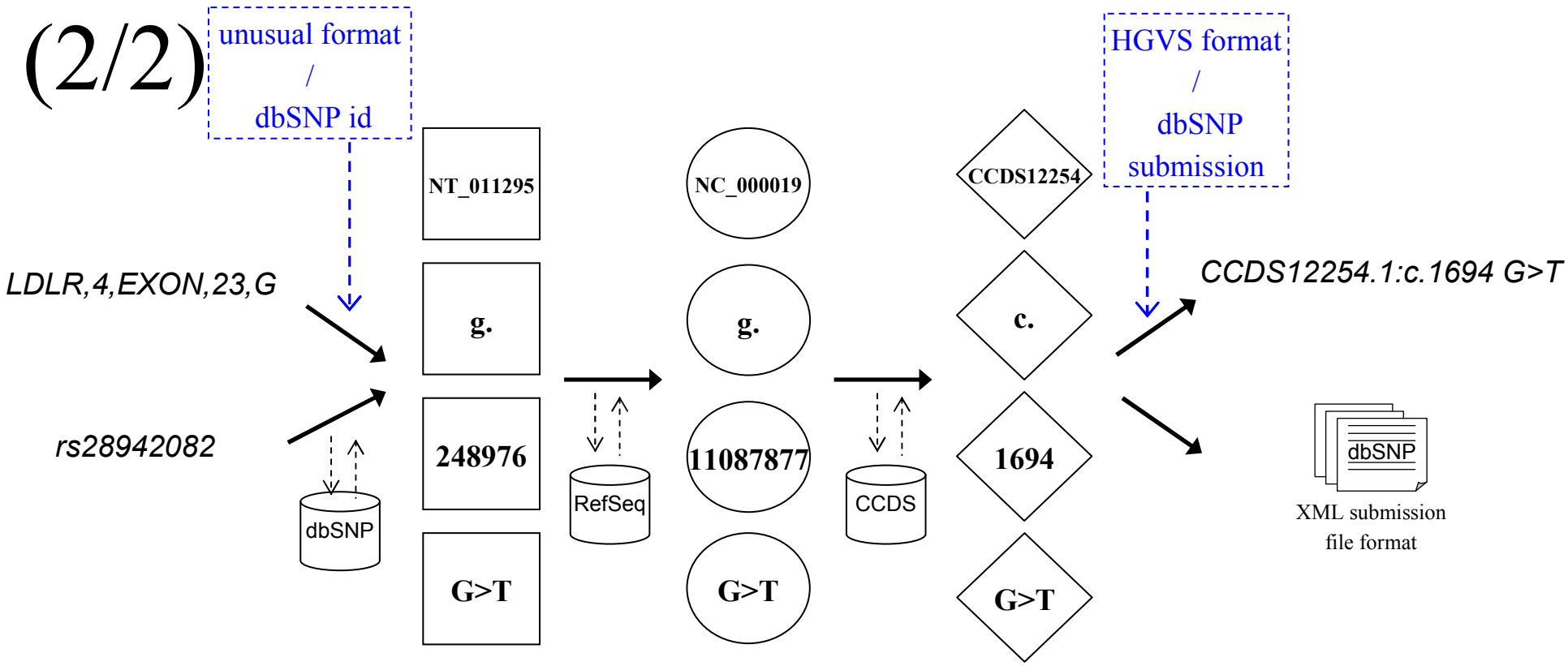
- But in itself unable to achieve data integration
  - i.e. to convert variation format or to establish equivalences

SNP-Converter: process description (1/2)



- Four features
  - 1. The identifier of a reference sequence — SNPO:sequence
  - 2. The type of the reference sequence — SNPO:aa_sequence ⊔ SNPO:nucleotide_sequence
  - 3. The position in the reference sequence — SNPO:sequence_position
  - 4. The observed variation — SNPO:observed_variation
- mapped on concepts of the ontology

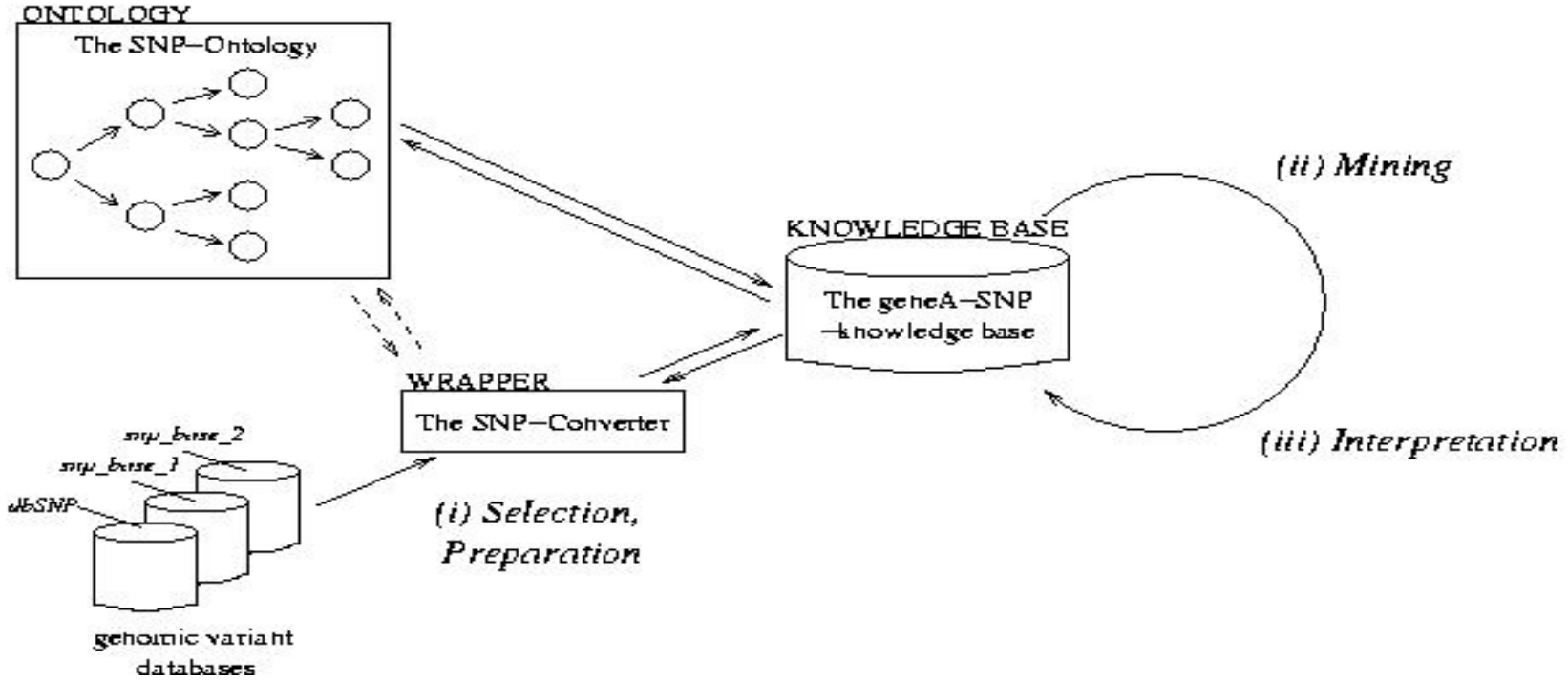
SNP-Converter: process description (2/2)



- A standalone application
  - Typical scenario: converting interesting SNPs from a private database into the dbSNP submission file format

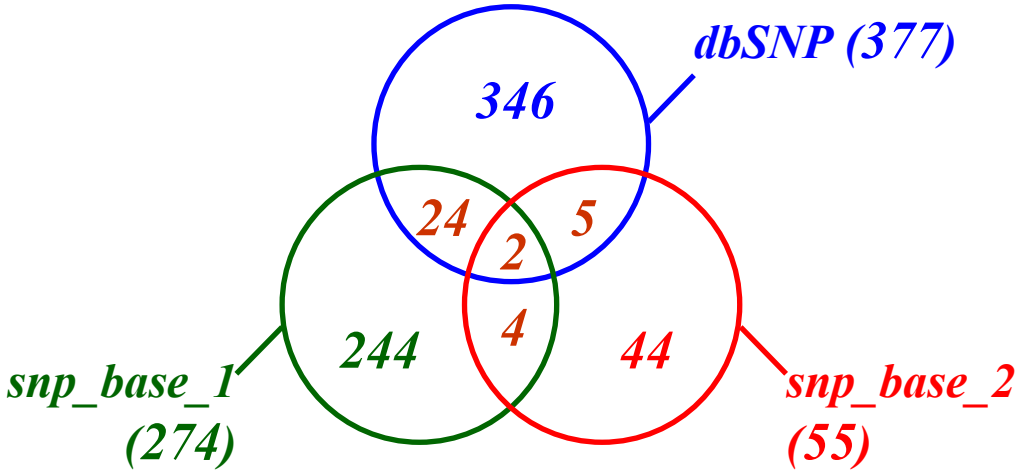
SNP-Converter: semantic integration use case

- A wrapper for semantic data integration
  - For extracting data from variation sources
  - For populating the knowledge base



SNP-Converter: example

- The **SNP-Knowledge Base** has been populated with variants from
  - ***dbSNP***: the genomic variation repository from the NCBI
  - ***snp_base_1*** and ***snp_base_2***: two private locus-specific databases
- The unified description allows to establish equivalence between variants



Conclusion

- Semantic data integration
  - SNP-Converter: A procedural extension of description-logics-based reasoners
  - A proof of concept limited to the genotype summit
- Limited to current version of genome assembly
- On going
  - Classifying genomic variations: e.g. insertion, deletion, etc.
  - Extension to the whole pharmacogenomics for achieving the KDD process
- **WANTED**: Clinical data with genotype
  - Collaboration

Thanks

- INSERM UMRS 538
  - Dr Pascale Benlian
- KIKA Medical
  - Romain Demoustier
- Phenosystems
  - David Atlan
- ORPAILLEUR team in LORIA
  - Amedeo Napoli,
  - Mathieux d'Aquin,
  - Marie-Dominique Devignes ,
  - Nicolas Jay,
  - Jean Lieber ,
  - Sandy Maumus,
  - Malika Smaïl-Tabbone

What ontology brings?

- 1. Expressing that to things are the same thing
  - *rs1142345* `same_as` *NC_000006.10g:18238897A>G*
- 2. Allowing, thanks to associated reasoning mechanisms to validate the knowledge representation
  - check consistency, satisfiability, etc.
- 3. Reasoning on concepts
  - *e.g. Because intron sequences are not translated,*
    Because aa sequences are resulting from translation,
    Then SNPs positioned on intron sequence have no impact on aa sequences
- 4. Reasoning on individuals
  - *e.g. The patient X has a G at the nucleotide 18238897th on the chromosome 6 and a*
    A at the nucleotide 12345689th on the chromosome 6
    *Then his genotype is TPMT*3C and then is a poor metabolizer of mercaptopurine*