Ontology Driven Modeling for the Knowledge of Genetic Susceptibility to Disease

Lin, Yu

(Degree)
博士（医学）
(Date of Degree)
2009-03-25
(Date of Publication)
2012-12-25
(Resource Type)
doctoral thesis
(Report Number)
甲4675
(URL)
https://hdl.handle.net/20.500.14094/D1004675

氏　　名	林　郁
博士の専攻分野の名称	博士（医学）
学 位 記 番 号	博い第 4675 号
学位授与の 要 件	学位規則第 5 条第 1 項該当
学位授与の 日 付	平成 21 年 3 月 25 日

【 学位論文題目 】

Ontology Driven Modeling for the Knowledge of Genetic Susceptibility to Disease(遺伝的疾患感受性に関する知識のためのオントロジー駆動型モデリング)

審 査 委 員

主 査	教 授	東　健
	教 授	清野 進
	教 授	西尾 久英

Problems: No formalization for the knowledge of genetic susceptibility to disease

Unlike Mendelian disease, the cause of a complex disease such as diabetes, hypertension and so on, is usually the interaction of the genetic factors and the environmental factors. The genetic susceptibility is realized when a genetic factor increases the probability of a person to develop one specific disease, for example Mary's Diabetes, and this genetic factor can be called as a genetic susceptibility factor. The general methodology to identify the genetic susceptibility to complex disease is a combination of linkage study and association study in biological experimental science. The evidence for proving the genetic susceptibility is built on the statistical measurement in the population-based association study. To determine a genetic susceptibility factor needs not only the value of odds ratio (OR) obtained from the association study, but also the sample size, the population and the replication results. This complexity makes an obstacle for managing and reusing the knowledge with the help of machine. Further more, without a formalization of the concept of the genetic susceptibility; the interoperability between systems will not be achieved.

Method: Ontology-driven modeling by using OWL-DL language

Originally *ontology* was used as a philosophical term for the study of the nature of existence. In both computer science and information science, an ontology is "a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group." Ontologies provide the basis for interoperability between systems, and they are considered to be an important technology for the Semantic Web. Ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field. As a method to formalize knowledge model, many ontologies have been developed and used in several areas, including bioinformatics and systems biology.

The use of biomedical ontologies has grown dramatically since the Gene Ontology (GO) Consortium was initiated in 1998 by three model organisms groups: FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD). Further milestones were the establishment in 2001 of OBO (Open Biomedical Ontologies) to serve as "an umbrella body for the developers of life-science ontologies" and to provide an OBO ontology repository, which in turn led to the creation in 2005 of the OBO Foundry, an experiment directed towards the creation of a suite of interoperable ontology modules designed to support life science research. Although they have the corresponding OWL version files, those successful ontologies are using the OBO format to write their ontology files.

OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. OWL builds on RDF and RDF Schema and uses RDF's XML-based syntax. W3C's Web Ontology Working Group defined OWL as three different sublanguages: OWL Full, OWL DL and OWL Lite. Compare to RDF and RDFS, the expressive power of these three languages is decreasing, and the reasoning supporting is increasing.

Compare to OWL, the OBO format lacks formal computational definitions for its constructs and tools, for example, the DL reasoners in OWL, to facilitate ontology development or maintenance. As for the semantics supporting, the OBO ontologies are less well defined than OWL ontologies. Thus, in our research, we prefer OWL as the modeling language, and decided to develop the OWL-DL ontologies to solve our problems.

The ontology engineering technologies we applied in developing the ontologies are the text mining, by which the core conceptions for ontology modeling were generated, and the Competency Questions technology, by which the ontologies were characterized. Except for that, we also applied the Ontology Design Patterns (ODPs) for modeling, such as the n-ary relation pattern.

The tools we were using in this thesis including Protégé-OWL 4.0 build 101, which is used for developing the ontologies, and the reasoner Pellete 1.5 embodied in Protégé-OWL 4.0, by which we achieved the automatical classification of the genetic susceptibility factors.

Results:

1. The n-ary relation pattern representing the genetic susceptibility to disease

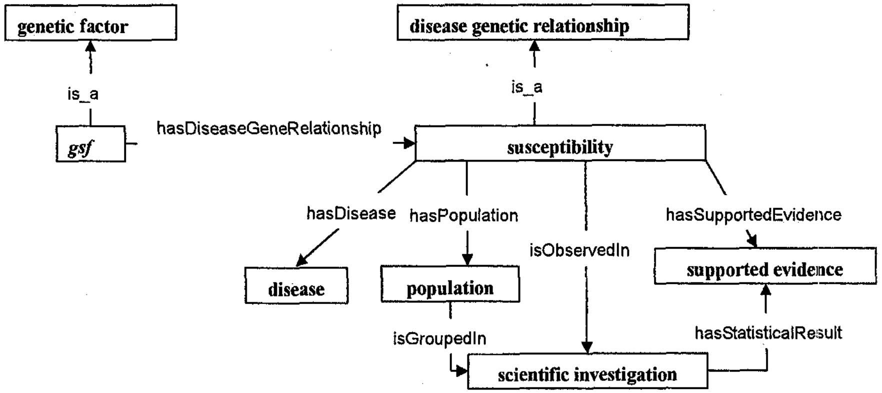

Figure 1. Full version of n-ary relation pattern representing the genetic susceptibility to disease

2. The hierarchy of the class: *DiseaseGeneticRelationship*

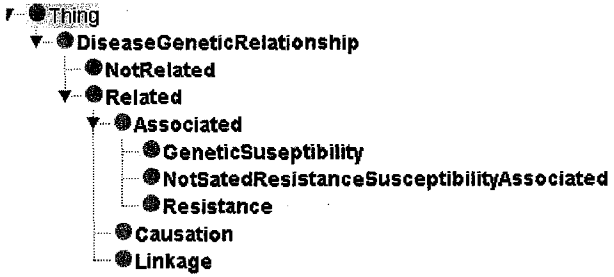
Thing
DiseaseGeneticRelationship
NotRelated
Related
Associated
GeneticSuseptibility
NotSatedResistanceSusceptibilityAssociated
Resistance
Causation
Linkage

Figure 2 a screen shot of Protégé-OWL 4.0 shows the Hierarchy of the class : *DiseaseGeneticRelationship*

3. Using the Manchester Syntax of OWL, we formalized the definition of genetic susceptibility factor as following:

```
Class: GeneticSusceptibilityFactor SubClassOf: GeneticFactor
   EquivalentTo: GeneticFactor that
                        hasGeneticSusceptibilityRelationship SOME GeneticSusceptibility AND
                        hasGeneticSusceptibilityRelationship min 1 GeneticSusceptibilityFactor

Class: Genetic_Susceptibility SubClassOf: AssociatedRelationship
   EquivalentTo: AssociatedRelationship that
                     isObserveRelaionshipOf ONLY GeneticSusceptibilityFactor

Class: AssociatedRelationship SubClassOf: ObservedRelationship

Class: ObservedRelationship SubClassOf: StatisticalObservation
   EquivalentTo: StatisticalObservation that
                        (NotRelated OR Related) AND
                        isObservedIn SOME DieaseGeneStudyinPaper AND
                        isObserveRelaionshipOf SOME GeneticFactor AND
```

```
isRelationshipWith SOME (HumanDisease OR
                         Measurement OR
                         PopulationCharacteristic) AND
hasPopulation ONLY StudyPopulation AND
hasSupportingEvidence ONLY SupportingEvidence
```

According to this formalization, we gave the following explanation for the above classes:
A genetic susceptibility factor is a genetic factor, which has at least one genetic susceptibility relationship.
A genetic susceptibility relationship is an associated relationship, which is and only is the observed relationship of a genetic susceptibility factor.
An associated relationship is a kind of Observed Relationship.
An Observed Relationship is a Statistical Observation, which is either related or not related observation; it is observed in some scientific paper, and is the relationship of at least one genetic factor, the relationship with at least one of the human disease or measurement of population characteristic, and has only the study population as well as the supporting evidence.

4. Automatical classification: By using reasoner Pellete 1.5 in protégé 4.0, the program classified the genetic factor "T allele of rs7903146" to be an instance of the class: GeneticSusceptibilityFactor.

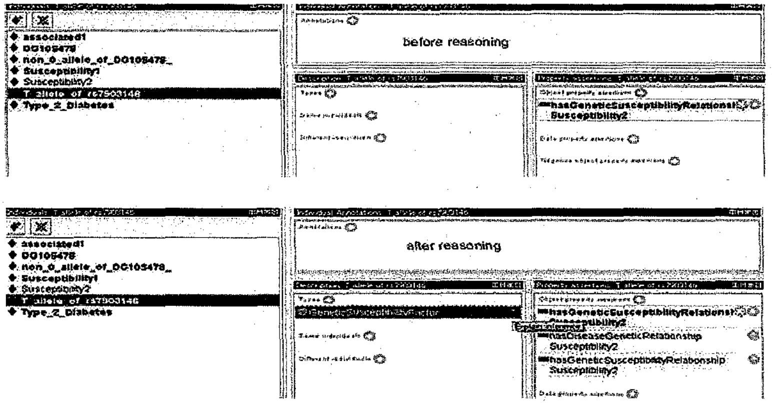
associated1
DO105478
non_0_allele_of_DO105478_
Susceptibility1
Susceptibility2
T_allele_of_rs7903146
Type_2_Diabetes
before reasoning
Types
Same individuals
Different individuals
Object property assertions
hasGeneticSusceptibilityRelationship Susceptibility2
Data property assertions
Negative object property assertions
after reasoning
GeneticSusceptibilityFactor
Explain inference
hasDiseaseGeneticRelationship Susceptibility2
hasGeneticSusceptibilityRelationship Susceptibility2

Figure 3. Before and after reasoning of the classification of "T_allele_of_rs7903146"

Discussion

In this paper, we have applied the ontological modeling method to represent the knowledge of genetic susceptibility to disease. There are obviously large areas of the world of biology that can be represented using OWL-DL with great success, one of the good practices is the protein classification by OWL reasoning. In our case, we applied reasoning for testing class membership, i.e., testing a given individual of Genetic Factor class is an instance of the Genetic Susceptibility Factor class, which has been formally declared in ontology. It is important to note that class membership of individuals can usually only automatically be recognized if the class description is complete.

Except for the practice on class membership, our study also applied the Ontology Design Patterns (ODPs) for modeling, such as the n-ary relation pattern. The n-ary relation pattern is a very important ODP for the biological world.

OWL uses Open World Assumption, and under this Open World Assumption, if a statement can not be proved to be true using current knowledge, we can not draw the conclusion that the statement is false. Compare to other language, such as prolog or SQL, OWL's open world assumption fits better in with the knowledge about biology, which is certainly not complete. The knowledge discovery is based on the biological experiments, more exceptions to the current knowledge certainly will appear in the future.

There are no criteria to define a genetic susceptibility factor, such as using the value of OR, for example "OR greater than 2", to define "susceptibility" relationship is not suitable. More susceptibility genetic factors were confirmed with a result of OR between 1.1 and 1.5. Some researchers believe that "It would be helpful if qualified number restrictions would be added to OWL-DL." However in our case, the number restrictions might help the system work, which depends on the future development of OWL-DL on this issue. All together, the semantic relations between genetic factors and disease are more reasonable in the current situation.

The semantic web seems a better idea for using the ontologies we have built, which are essential for the database integration and system interoperability. The framework of this modeling will be the base for linking the data sources come from the public databases (such as pubmed, OMIM), other ontologies (sequence ontology, human disease ontology, and so on), and the HTML or XML documents. Finally, we will use these ontologies to associate the possible genetic factors with disease by semantic web technology.

神戸大学大学院医学系研究科（博士課程）

論文審査の結果の要旨

受付番号	甲　第２０１４号	氏　名	林　郁
論文題目 Title of Dissertation	Ontology Driven Modeling for the Knowledge of Genetic Susceptibility to Disease 遺伝的疾患感受性に関する知識のためのオントロジー駆動型モデリング		
審査委員 Examiner	主　査 Chief Examiner　東　健 副　査 Vice-examiner　清野進 副　査 Vice-examiner　西尾久英		
審査終了日	平成２１年２月１８日		

（要旨は１，０００字～２，０００字程度）

In recent years, countless disease-associated genes have been published, however, to distinguish the true susceptibility factors from irrelevant DNA polymorphisms remains far less satisfaction. Considering the complex and unstructured information on this topic, a well-developed ontology is important to help to explicit the concepts and their relationships in the domain of the genetic susceptible factors to diseases. There is no formalization for the knowledge of genetic susceptibility to disease. Thus, the ontology modeling language OWL was used for formalization in this study. Ontology languages allow users to write explicit formal conceptualizations of domain models.

After introducing the Semantic Web and OWL language propagated by W3C, the candidate applied text mining technology combined with competency questions to specify the classes of the ontology. Then, an N-ary pattern was adopted to describe the relationships among these defined classes. Based on the former work of OGSF-DM (Ontology of Genetic Susceptibility Factors to Diabetes Mellitus), the candidate formalized the definition of "Genetic Susceptibility", "Genetic Susceptibility Factor" and other classes by using OWL-DL modeling language; and a reasoning automatically performed the classification of the class "Genetic Susceptibility Factor".

The ontology driven modeling is used for formalization the knowledge of genetic susceptibility to complex diseases. More importantly, when a class has been completely formalized in an ontology, the OWL reasoning can automatically compute the classification of the class, in this case, the class of "Genetic Susceptibility Factors". With more types of ontology always needs to be refined and many new classes must be taken into account to harmonize with the ontology. Using the ontology to develop the semantic web needs to be applied in the future. The semantic web seems a better idea for using the ontology the candidate has built, which are essential for the database integration and system interoperability.

The candidate, having completed studies on the ontology driven modeling for the knowledge of genetic susceptibility to disease, and having advanced the field of knowledge in the area of the ontology, is hereby recognized as having qualified for the degree of Ph.D. (Medicine).