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

Kobe J. Med. Sci., Vol. 55, No. 3, pp. E53-E66, 2009

Ontology Driven Modeling for the Knowledge of Genetic Susceptibility to Disease

YU LIN*, and NORIHIRO SAKAMOTO

Department of Sociomedical Informatics, Kobe University Graduate School of Medicine

Received 18 December 2008/ Accepted 14 January 2009

Key Words: Ontology, Modeling, Knowledge Representation, OWL, Genetic Susceptibility

For the machine helped exploring the relationships between genetic factors and complex diseases, a well-structured conceptual framework of the background knowledge is needed. However, because of the complexity of determining a genetic susceptibility factor, there is no formalization for the knowledge of genetic susceptibility to disease, which makes the interoperability between systems impossible. Thus, the ontology modeling language OWL was used for formalization in this paper. After introducing the Semantic Web and OWL language propagated by W3C, we 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), we formalized the definition of "Genetic Susceptibility", "Genetic Susceptibility Factor" and other classes by using OWL-DL modeling language; and a reasoner automatically performed the classification of the class "Genetic Susceptibility Factor". Conclusion: 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 our case, the class of "Genetic Susceptibility Factors". With more types of genetic susceptibility factors obtained from the laboratory research, our ontologies always needs to be refined, and many new classes must be taken into account to harmonize with the ontologies. Using the ontologies to develop the semantic web needs to be applied in the future.

Semantic Web and Ontologies

The Semantic Web as an extension for the current Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. (1) The term *Semantic Web* comprises techniques that promise to dramatically improve the current WWW and its use. These technologies including: explicit metadata, ontologies, logic, and agents. Semantic Web agents will make use of all the technologies above:

- Metadata will be used to identify and extract information from Web sources.
- Ontologies will be used to assist in Web searches, to interpret retrieved information, and to communicate with other agents.
- Logic will be used for processing retrieved information and for drawing conclusions. (2)

Phone: +81-78-382-5777 E-mail: linikujp@gmail.com

The key idea of the Semantic Web is the use of machine-processable web information. The Semantic Web is already in place and is characterized by a widespread production of formalized knowledge models (the ontologies and metadata), from a variety of different groups and individuals.

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.”(3) Ontologies provide the basis for interoperability between systems, and they are considered to be an important technology for the Semantic Web. As a method to formalize knowledge models, many ontologies have been developed and used in several areas, including bioinformatics and systems biology. (4, 5)

Basic Features and Notations of OWL

Ontology languages allow users to write explicit formal conceptualizations of domain models. OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. It is intended to provide a language that can be used to:

- conceptualize domains by defining classes and properties of those classes,
- construct individuals and assert properties about them,
- reason about these classes and individuals.

OWL is developed as a vocabulary extension of the Resource Description Framework (RDF) (6). The underlying structure of any expression in RDF is a collection of triples; each triple consists of a subject, a predicate (also called a property) and an object. Such triples can also be represented as the “RDF graph”, in which nodes correspond to the “subject” and “object”, and the directed arc corresponds to the “predicate” as shown in Figure 1.

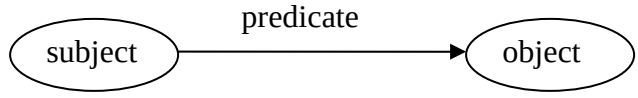


Figure1 The graphical representation for a generic RDF

However, RDF uses only binary properties. This restriction seems quite restrict because we would often like to use predicates with more than two arguments. Luckily, such predictions can be simulated by a number of binary predicates. XML-based syntax of RDF is well suited for machine processing but is not particularly human-friendly. RDF Schema (RDFS) makes assumptions about a particular application domain, and defines the semantics of any domain through classes and properties. RDF and RDFS allow the representation of some ontological knowledge.

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. The expressive power of these three languages is decreasing, and the reasoning supporting is increasing, however all are more expressive than RDF or RDFS.We introduce some basic primitives as follows:

<owl:Class>: This primitive is used to create ontology concepts as classes. Each class will be named with a class identifier. For example, “<owl:Class rdf:ID=”Disease”>” defines a *owl:Class* instance ”Disease”.

• <owl:DatatypeProperty>: This primitive is used to express the relations between instances of classes and RDF literals (7) and XML Schema datatypes (8).

• <owl:ObjectProperty>: This primitive is used to describe the relations between instances of two classes.

• <owl:Restriction>: This primitive is used to specify the constraints on ontology concepts or classes.

• <owl:onProperty>: This primitive indicates which property is restricted.

The detailed information about features and primitives of the language is to be founded in (9), (10), (11). In the next subsection we will show how to use OWL-DL to model an ontology for applications in biomedicine field.

Ontologies in Biomedical Field

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. (12) Ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field.

However, those successful ontologies used the OBO format to write their ontology files. The Open Biomedical Ontologies (OBO) format created by the GO consortium is a very successful format for biomedical ontologies, including the Gene Ontology. As an ontology representation language, the OBO flat file format is popular in biomedical field. Many early developed biomedical ontologies, such as SO (Sequence Ontology), FMA (Foundational Model of Anatomy), GO (Gene Ontology) and so on, used the OBO format. (13)

By using tag-Value pairs, the OBO flat file format can describe one GO term by id, namespace, definition and five types’ relationships with other GO terms: is_a, part_of, regulates, positively_regulates and negatively_regulates. Figure 2 shows a part of the OBO files of GO.

```
[Term]
id: GO:0000005
name: ribosomal chaperone activity
namespace: molecular_function
def: "OBSOLETE. Assists in the correct assembly of ribosomes or ribosomal subunits in vivo, but is not a component of the assembled riboso
comment: This term was made obsolete because it refers to a class of gene products and a biological process rather than a molecular functi
is_obsolete: true
consider: GO:0042254
consider: GO:0051082

[Term]
id: GO:0000006
name: high affinity zinc uptake transmembrane transporter activity
namespace: molecular_function
def: "Catalysis of the transfer of a solute or solutes from one side of a membrane to the other according to the reaction: Zn2+(out) = Zn2
is_a: GO:0005385 ! zinc ion transmembrane transporter activity
```

Figure2 The OBO file (a part of Gene Ontology)

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. The community of the OBO ontologies' users has realized this problem, and now most OBO format ontologies have the corresponding OWL files for sharing and reusing their ontologies.

Since we are concerning more with the semantics supporting ability of our ontologies, in our research, we prefer OWL as the modeling language, and decided to develop an OWL-DL ontology to solve our problems.

MATERIALS AND METHODS

The Knowledge of Genetic Susceptibility in Biology

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. Genetic susceptibility is realized when a genetic factor increases the probability of a person developing a specific disease, for example Mary's Diabetes, and this genetic factor can be called as a genetic susceptibility factor. However, the genetic susceptibility factors contributing susceptibility to common disease may not be obvious mutations; and it is more likely a combination of subtle changes on several genes, which may be quite common in the healthy population. Moreover, the main determinants of susceptibility may be different in different populations. (14) The current status of determining susceptibility genetic factors remains far less satisfactory.

The general methodology to identify the genetic susceptibility to complex disease is a combination of linkage study and association study in biological experimental science. At first, by using the family-based samples, researchers conduct the linkage analysis, through which researchers obtain a number of broad linked regions that represent several mega bases of DNA. To narrow down such a region to a susceptible gene (or genes), a population-based approach is required. Case-control study of the unrelated individuals is a wildly used approach in this step. The working hypothesis is that variants in linkage disequilibrium (LD) with the susceptibility locus will define the genomic region responsible for the original linkage signal. However, the extent of LD in various regions of the genomic DNA has been shown to be highly variable. (15)

The evidence for proving the genetic susceptibility is built on the statistical measurement in the population-based association study. By using odds ratio (OR) measurement in a case-control study, a genetic variant in a case-group is considered as an event compare to a control group; if the odds ratio (OR) is greater than 1, the event is more relative with the case group than the control group. Thus, an association relationship has been observed, and the observed genetic factors can be considered as the disease related genes. However, to be related to a disease doesn't mean that the genes contribute susceptibility to the related disease. According to Wang et al.'s review in (16): "Most irrefutable disease-susceptibility variants that have been identified so far — mainly from functional candidate association studies — have allelic odds ratios that are in the order of 1.1–1.5." Conversely, for an odds ratio of 2, even for an allele with a Minor Allele Frequencies of 0.005 there is 76% power. However, such high odds ratios must be rare in common diseases. (16) Moreover, the value of OR is not the only criteria, the sample size, the population and the replication results of the observed genetic factors must be taken into account in this stage as well.

Modeling the knowledge of genetic susceptibility to disease

According to the above analysis, the definition of a related gene is distinct from a susceptibility gene. Thus, we choose the original papers from PubMed but not the other existing databases of disease related genes as our original information pool to start our work. We use Text mining technology to generate the core conceptions for ontology modeling, and then conduct the Competency Questions technology to characterize the ontology.

Text Mining

Although it is the most complex and unstructured data source to search, literature is the most powerful resource to support the knowledge we want to model. Taking Diabetes Mellitus as an example disease, we retrieved original research papers from PubMed database, by using the following query:

((gene[TW] OR genetic*[TW] OR genome*[TW] OR "susceptibility gene"[TIAB] OR "susceptibility genes"[TIAB]) AND (Diabetes[TIAB] OR "Diabetes Mellitus"[TIAB] OR Diabetic[TIAB] OR hyperglycemia[TW]) NOT "diabetes insipidus"[TW]) AND hasabstract[text] NOT (Review[PT] OR Editorial[PT] OR meta-Analysis[PT] OR Comment[PT])*

A total number of 26220 abstracts have been obtained by 12:00am of 26th, Aug. 2008. After manually excluding the irrelevant abstracts, we obtained a corpus of 5873 abstracts, which composes the basic literature source of the modeling construction.

By analyzing the titles of those abstracts in corpus, we finally obtained 5 types basic structure of the core conceptions in this domain:

1. (Genetic Variants of X Gene) and (Y Disease)
2. the association of (Polymorphism of X Gene) with (Y Disease)
3. (Genetic Variants of X Gene) associates with (Z Population)
4. (Genetic Variants of X Gene) associates with (P Phenotype) in (Z Population) with (Y Disease)
5. lack of association of (Genetic Variants of X Gene) with (Y Disease) in (Z Population)

The primary blocks for representing this knowledge have been decided in this step. These primary blocks became the modular ontologies for the whole ontology in the knowledge modeling procedure. The most essential part of this modeling includes the blocks of “disease or phenotype”, “disease-gene relationship” and “gene or genetic factor”, which has been shown in Figure 3.

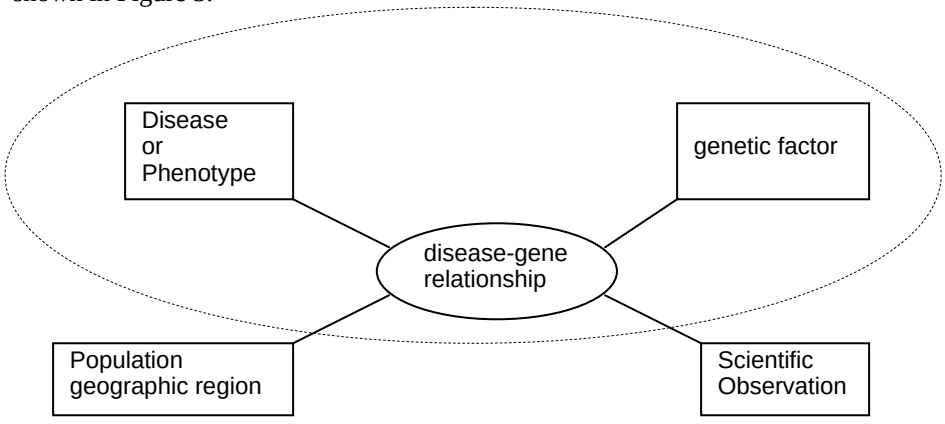


Figure3 The primary blocks for genetic susceptibility to complex disease

Competency Questions

Professor Asunción Gómez-Pérez describes competency questions as: "..a set of natural language questions, called competency questions, are used to determine the scope of the ontology. These questions and their answers are both used to extract the main concepts and their properties, relations, and formal axioms of the ontology...Given the set of informal scenarios, a set of informal competency questions are identified. Informal competency questions are those written in natural language to be answered by the ontology once the ontology is expressed in a formal language. The competency questions play the role of a type of requirement specification against which the ontology can be evaluated." (17)

Here, we use the TCF7L2 gene and its susceptibility to Type 2 Diabetes as an example scenario. The following Informal Competency Questions were designed by the domain experts. To answering those questions, we searched the literatures of OMIM® (Online Mendelian Inheritance in Man®) and a patent WO/2006/137085 for TCF7L2, and summarized the alleles of markers and their susceptibility or resistance to Type 2 Diabetes in Table 1.

Marker	**Allele**	**Susceptibility or Resistance to T2D**
DG10S478	0 allele	resistance
	non-0 allele	susceptibility
rs7903146	C allele	resistance
	T allele	susceptibility
rs12255372	G allele	resistance
	T allele	susceptibility
rs7895340	G allele	resistance
	A allele	susceptibility
rs11196205	G allele	resistance
	C allele	susceptibility
rs7901695	T allele	resistance
	C allele	susceptibility
rs12243326	C allele	susceptibility
rs4506565	T allele	susceptibility

Table1 The susceptibility or resistance alleles of markers to T2D

The following are the Competence Questions and their answers:

1. CQ: What LD (linkage disequilibrium) region is associated with Type 2 Diabetes?
   Answer1: A LD between a SNP, rs12255372, and a microsatellite marker, DG10S478, associated with type 2 diabetes.
   Answer2: A LD between a SNP, rs7903146, and a microsatellite marker, DG10S478, associated with type 2 diabetes
2. CQ: What is the location relationship of rs12255372 and TCF7L2 gene?
   Answer: rs12255372 is located in the intron 4 of the TCF7L2 gene.
3. CQ: What polymorphisms on TCF7L2 are associated with Type 2 Diabetes?
   Answer: a microsatellite DG10S478;

SNP: rs12255372, rs7903146;
LD block: exon 4 LD block of TCF7L2

4. CQ: What allele of DG10S478 is susceptible to T2D?
   Answer: non-0-allele
5. CQ: What allele of rs12255372 is susceptibility to T2D?
   Answer: T allele
6. CQ: What allele of rs12255372 is resistance to T2D?
   Answer: A allele
7. CQ: In what population DG10S478 is susceptible to T2D?
   Answer: Icelandic individuals, Mexican
8. CQ: List out all the scientific investigations which have done research on the susceptibility of DG10S478 to T2D. (which is a requirement for the system)
   Answer:
   Pubmed ID: 17470138
   Pubmed ID: 17340123
   Pubmed ID: 17317761
   Pubmed ID: 16415884
   Pubmed ID: 16936218

Tools

We used the Protégé-OWL 4.0 build 101 to develop the ontology. The N-ary relations patterns were adopted to model the relations between gene and disease, which will be described in detail in the following section.

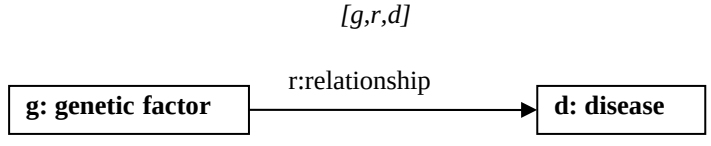
RESULT

The N-ary relations between the basic blocks

In Semantic Web languages, such as RDF and OWL, a property is a *binary* relation: it is used to link two individuals or an individual and a value. However, in some cases, the natural and convenient way to represent certain concepts is to use relations to link an individual to more than just one individual or value. These relations are called *n-ary relations*. (18) The genetic susceptibility to disease is not the relation which can be completely described only by the binary relation between a genetic factors and a disease; more elements are needed for representing this term, such as the populations, the experiments, and the supporting evidence. Thus we adopted the N-ary pattern 1 suggested by W3C in (18) in our research.

As we have mentioned before, the knowledge of genetic susceptibility to disease is about the relationships between the genetic factor and a peculiar disease.

The relation between the genetic factor and disease can be described as such a triple *[g,r,d]*, which can be represented as a node connected graph as being showed in Figure 4.



g designates the genetic factors; d designates the disease; r designates the relationships.

Figure4 Node connected graph of the triple *[g,r,d]*

Since the RDF and OWL's primitive modeling is based on the binary relationship, above graph can be alternatively represented as being showed in Figure 5.

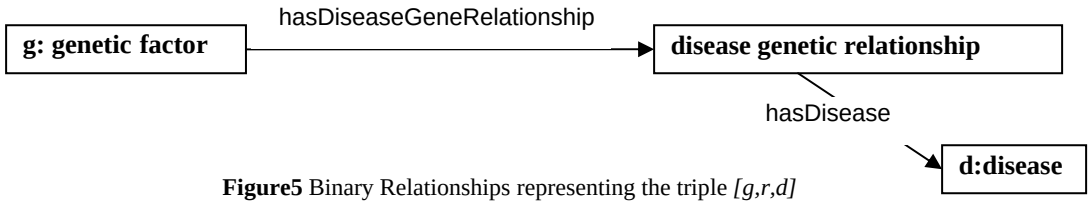
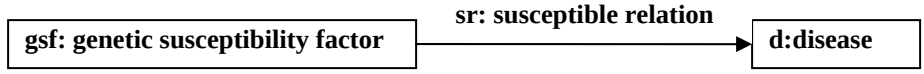


Figure5 Binary Relationships representing the triple *[g,r,d]*

When the relationship between a genetic factor and a disease is specified as a susceptibility relationship, the gene factor can be called genetic susceptibility factor. We represented it as the triple: *[gsf, sr, d]*. Figure 6 shows the node connected graph of the triple *[gsf, sr, d]* and Figue 7 shows using the N-ary relations pattern to model the triple *[gsf, sr, d]*.

[gsf, sr, d]



gsf designates the genetic susceptibility factor; sr designates the susceptibility relationship; d designates the disease

Figure6 Node connected graph of the triple *[gsf,sr,d]*

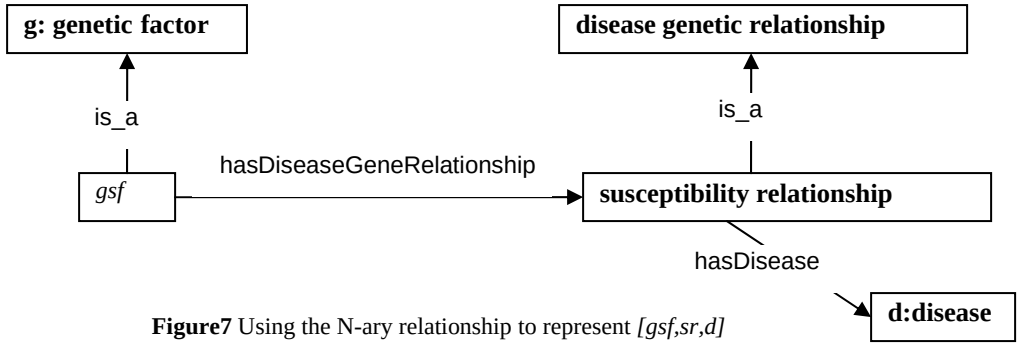


Figure7 Using the N-ary relationship to represent *[gsf,sr,d]*

Here, the *is_a* relations are formal *is_a* relations.

So, if we say *genetic susceptibility factor is_a genetic factor*, which means all the instances of *genetic susceptibility factor* are the instances of *genetic factor*. The relation "*is_a*" is equal to "*subclass_of* " in all models of ontology.

Since many other constrains such as populations, scientific investigations are necessary for defining a *gsf*, the whole N-ary relations between all the primary blocks are shown in Figure 8.

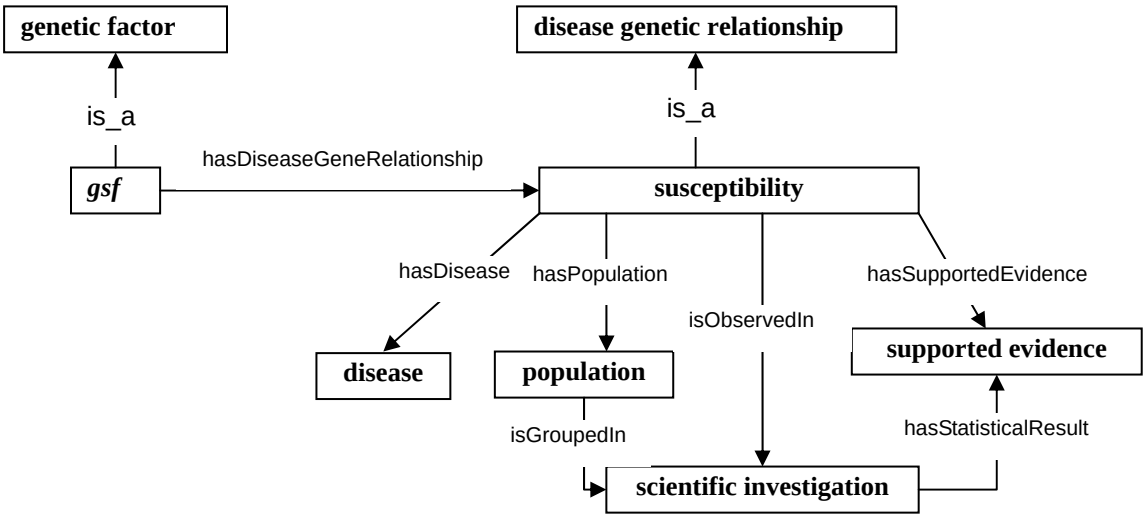


Figure8 Full version of n-ary relation pattern representing the genetic susceptibility to disease

Previous Work: OGSF-DM

We have established an OWL-DL ontology OGSF-DM (Ontology of Genetic Susceptibility Factors to Diabetes Mellitus), which has been created to describe the genetic susceptibility factors to Diabetes Mellitus(19) on the study of relevant abstracts retrieved from PubMed. Being built under the framework of upper ontology BFO, OGSF-DM includes three ontologies: the Ontology of Genetic Susceptibility Factors (OGSF), which serves as main ontology embodied with two imported sub-ontologies: the Ontology of Glucose Metabolism Disorders (OGMD) and the Ontology of Geographical Regions (OGR). All those ontologies are available on the BioPortal site, which is the National Center for Biomedical Ontology's ontology repositories. The URL of BioPortal is http://bioportal.bioontology.org/

OGSF-DM includes representations of entities drawn from the following five interrelated domains:

i) Human disease;
ii) Phenotypes and observed quantity parameters at the cell, organ, or (human) organism level;
iii) Genetic entities;
iv) Geographical regions;
v) Entities relevant to disease genetics introduced in the original papers.

The sub-ontology OGMD covers scopes i) and ii); OGR covers scope iv); the main ontology OGSF covers scopes iii) and v). (Figure 9)

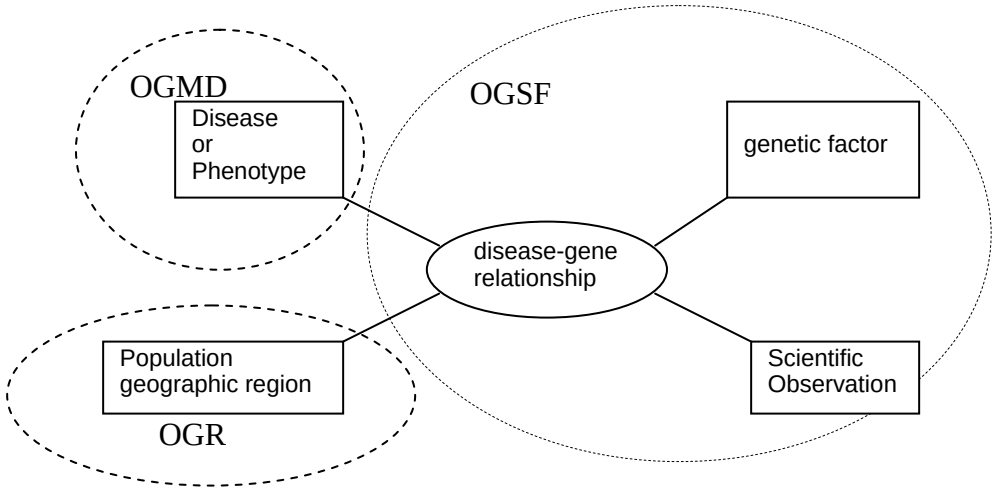


Figure9 The three ontologies cover five domains in OGSF-DM

Because the coverage domains necessary to OGSF-DM make overlapping with OBO Foundry ontologies unavoidable, this means that OGSF-DM cannot itself be a reference ontology within the OBO Foundry. Rather, it is an application ontology built on the Foundry as its basis.

We defined a small ontology to show the hierarchy of *DiseaseGeneticRelationship* by using Protégé-OWL 4.0 build 101.(Figure 10)

Figure10 Hierarchy of class : *DiseaseGeneticRelationship*

Genetic Susceptibility in OWL

Researches on genetic susceptibility are currently focused on statistically correlating specific genetic markers with specific diseases or abnormal phenotypes in the particular populations. The investigated genetic markers are used to detect DNA sequences with salient variation characteristics, such as Microsatellites or SNPs. Alleles or allelic variants, which are the alternative forms of such polymorphic sequences, contribute either susceptibility or resistance to the development of a disease.

As we mentioned before, OWL-DL supports those users who want the maximum expressiveness without losing computational completeness (all entailments are guaranteed to be computed) and decidability (all computations will finish in finite time) of reasoning systems.(20) OWL-DL is underpinned by Description Logics, which is a field of research

that has studied a particular decidable fragment of first order logic. This means that an OWL-DL ontology is expressed in a formalism with well-defined semantics and over which the automated reasoning can take place.

Using the Manchester Syntax of OWL, we defined the 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.

Classification by reasoning

One of the key features of OWL-DL ontologies is that they can be processed by a reasoner. One of the main services offered by a reasoner is to test whether or not one class is a subclass of another class. By performing such tests on the classes in an ontology, it is possible for a reasoner to compute the inferred ontology class hierarchy, as well as the classification of instances.

In the following Figure 9, 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*.

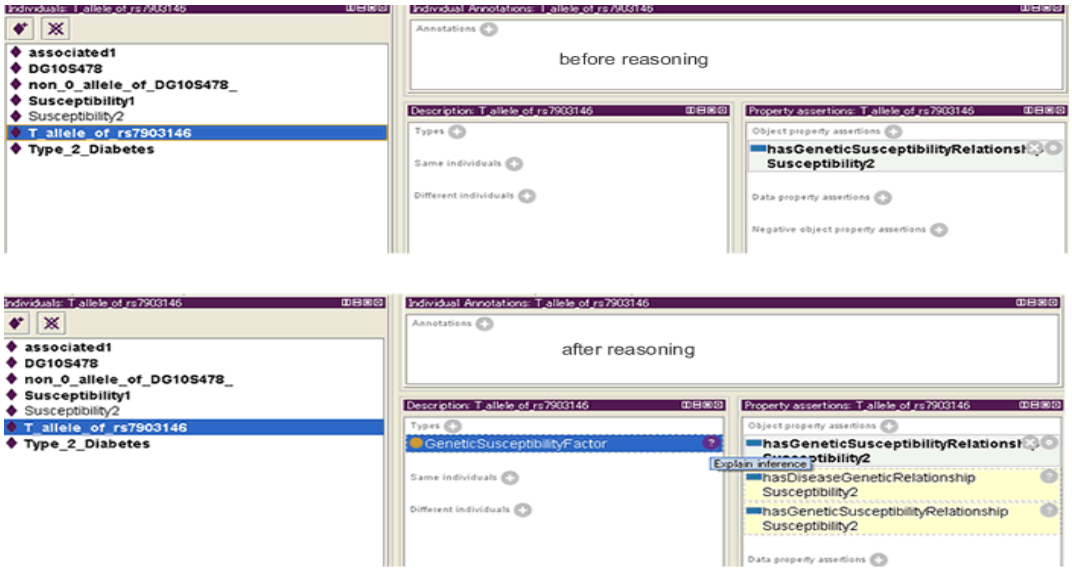


Figure9 Before and after reasoning of the classification of "T_allele_of_rs7903146"

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, such as protein classification by OWL reasoning. (21) In our case, we applied reasoning for testing class membership, i.e. , testing a given individual of the 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. In our practice, we defined the Genetic Susceptibility Factor has at least one genetic susceptibility relationship. After asserting the individual A has a susceptibility relationship B, the reasoner will infer that this A should be the instance of class "GeneticSusceptibilityFactors".

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. Whilst much can be modeled with binary relationships, it is often the case that we need to say more about the relationships between things other than that it is simply in existence. Evidence of observations; probabilities; sources; evidence; etc. are just a few of the cases in which n-ary relationships will be desirable.(22)

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 is no criteria to define a genetic susceptibility factor, as we have mentioned before, most susceptibility genetic factors were confirmed with a result of OR that are in the order of 1.1–1.5.(16) Some researchers believe that "It would be helpful if qualified number restrictions would be added to OWL-DL." (22) In our case, number restrictions might help improve the performance of our system, however this is dependent on the future development of OWL-DL. We feel semantical relations between genetic factors and disease are more reasonable in the current situation. For instance, the rs7903146 T allele is significantly susceptible to T2D (OR greater than 4)in a recent study in Korean people (23); whereas in another Brazilian team, they reported that the same allele is associated with a 1.57 increased risk for type 2 diabetes in a Brazilian cohort of patients with known coronary heart disease. However, it is not significant enough for judging the risk in the general population in Brazil. (24)

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 to link 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, in which the relations between the genetic factors and disease will be in a hierarchy as we showed in Figure 8.

REFERENCES

1. **BERNERS-LEE, T., HENDLER, J. and LASSILA, O.** 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American
2. **Antoniou G. and van Harmelen F.** 2004. A Semantic Web Primer p.1-15. The MIT Press, Cambridge, Massachusetts, London, England.
3. **Studer, R. Benjamins VR., Fensel D.** 1998. Knowledge Engineering: Principles and Methods. IEEE Transactions on Data and Knowledge Engineering **25(1-2):** 161-197
4. OBO, Open Biomedical Ontologies; http://obo.sourceforge.net/
5. **Lambrix, P**. 2004. Ontologies in bioinformatics and systems biology. In Artificial Intelligence Methods and Tools for Systems Biology, ed. Dubitzky and Azuaje, 129--146. Springer.
6. Resource description framework (rdf): Concepts and abstract syntax. W3C. [Online]. http://www.w3.org/TR/rdf-concepts
7. Rdf/xml syntax specification (revised). W3C. [Online]. Available: http://www.w3.org/TR/rdf-syntax-grammar/
8. Xml schema part 2: Datatypes. W3C. [Online]. Available: http://www.w3.org/TR/xmlschema-2
9. **Bechhofer,S., van Harmelen, F., Hendler, J., Horrocks, I., Deborah, L., McGuinness, P., Patel-Schneider, F., Stein, L.A.** 2004. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref
10. Owl web ontology language guide. W3C. [Online]. Available: http://www.w3.org/TR/owl-guide/
11. Owl web ontology language semantics and abstract syntax. W3C. [Online]. Available: http://www.w3.org/TR/owl-semantics/

12. **Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R. H., Shah, N., Whetzel, P. L., Lewis, S.** 2007.The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25:**1251–1255.
13. **Day-Richter,J.** 2004. The OBO Flat File Format Specification, version 1.0 [Online]. Available: http://www.geneontology.org/GO.format.obo-1_0.shtml
14. **Strachan, T., and Read A.P.** 2004. Human Molecular Genetics 3, pp. 447-448. Garland Science, New York, USA.
15. **Love-Gregory, L. D., Wasson J., Ma J., Jin C. H., Glaser B., Suarez, B. K. and Permutt, M. A.** 2004. A Common Polymorphism in the Upstream Promoter Region of the Hepatocyte Nuclear Factor-4 Gene on Chromosome 20q Is Associated With Type 2 Diabetes and Appears to Contribute to the Evidence for Linkage in an Ashkenazi Jewish Population. Diabetes Apr, p.1134-1140.
16. **Wang, W. Y. S., Barratt, B. J., Clayton, D. G. and Todd, J.A.** 2005. Genome-wide Association Studies: theoretical and practical concerns. Nat Rev Genet. **6(2):**109-18.
17. **Gómez-Pérez, A., Fernández-López, M., Corcho, O.** 2004. Ontological Engineering , p.119-120. Springer-Verlag London Limited, London, UK.
18. Defining N-ary Relations on the Semantic Web. W3C. [Online]. Available: http://www.w3.org/TR/swbp-n-aryRelations/
19. **Lin, Y., Sakamoto, N.** 2008. Ontology of Genetic Susceptibility Factors to Diabetes Mellitus (OGSF-DM), pp.99-104. Interdisciplinary Ontology Proceedings of the First Interdisciplinary Ontology Meeting, Tokyo, Japan.
20. OWL Web Ontology Language Overview. W3C. [Online]. Available : http://www.w3.org/TR/owl-features/
21. **Wolstencroft, K., Stevens, R. and Haarslev, V.** 2007. Applying OWL reasoning to genomic data, p.225-248. Semantic Web Revolutionzing Knowledge Discovery in the Life Sciences, Springer Science+Business Media, LLC, New York, USA.
22. **Stevens,R., Aranguren, M. E., Wolstencroft, K., Sattler, U., Drummond, N., Horridge, M., Rector, A.** 2007. Using OWL to model biological knowledge. Int. J. Human-Computer Studies **65:** 583–594
23. **Yu J, Steck, A.K., Babu, S., Yu, L., Miao, D., McFann, K., Hutton, J., Eisenbarth, G.S., Klingensmith, G.** 2008 Single Nucleotide Transcription factor 7-like 2 (TCF7L2) Gene Polymorphisms in Anti-Islet Autoantibody Negative Patients at Onset of Diabetes. J Clin Endocrinol Metab. doi: 10.1210/jc.2007-2694
24. **Marquezine, G.F., Pereira, A.C., Sousa, A.G., Mill, J.G., Hueb, W.A., Krieger, J.E.** 2008. TCF7L2 variant genotypes and type 2 diabetes risk in Brazil: significant association, but not a significant tool for risk stratification in the general population. BMC Med Genet. **9(1):**106.