Representation of Roles in Biomedical Ontologies: a Case Study in Functional Genomics

Anita Burgun[1], M.D., Ph.D., Olivier Bodenreider[2], M.D., Ph.D., Franck Le Duff[1], M.D., Fouzia Moussouni[3], Ph.D., Olivier Loréal[3], M.D., Ph.D.

[1]Laboratoire d'Informatique Médicale, Faculté de Médecine, 35033 Rennes, France
[2]U.S. National Library of Medicine, Bethesda, Maryland
[3]INSERM U522, C.H.R.U Pontchaillou, 35033 Rennes, France
Anita.Burgun@univ-rennes1.fr

ABSTRACT

Objective: Representing roles, i.e. functions of proteins, sequences and structures, is the cornerstone of knowledge representation in functional genomics. The objective of this study is to investigate representation of roles as functional categories or associative relations. We focus on GeneOntology (GO) and the UMLS and take examples from iron metabolism. Methods: The terms corresponding to the main proteins involved in iron metabolism were mapped to GO (including the annotations) and the UMLS. The representation of their biological roles was then analyzed. Results: Functional aspects are represented in both GO and the UMLS. However, the granularity may not be appropriate. Discussion: Advantages and limits of functional categories and associative relations are discussed.

INTRODUCTION

Biological knowledge is evolving from structural genomics towards functional genomics. The tremendous amount of DNA sequence information that is now available provides the foundation for studying how the genome of an organism is functioning, and high-throughput technologies provide detailed information on the mRNA, protein, and metabolite components of organisms. It makes it possible for researchers to discover new metabolic pathways, to model metabolic and regulatory networks in living organisms, and ultimately to understand the pathogenesis of diseases. Beyond their structure, the functions of the genes become essential information. In this context, it is fundamental that the knowledge representation systems supporting, for example, knowledge discovery provide an accurate representation of the roles and functions in the biomedical domain. Knowledge resources include GeneOntology™ (GO) [1], which focuses on genomics, and the Unified Medical Language System® (UMLS®) [2], which covers the whole biomedical domain.

Representing roles has been a central issue in conceptual modeling, e.g. [3]. While taxonomies of concepts (*is-a* hierarchies) organize things according to their essential features (what x is), and meronomies (*part-of* hierarchies) represent their constitutive features (what x is made of), two major options may be considered to represent functions: functional categories and associative relations. **Functional categories** are used in a system of hierarchy, where properties are inherited. For example, 'protein' can be combined with the function 'carrier' in order to generate the functional category 'carrier protein'. In the biomedical domain, many concepts are bound to a specific function. For example, 'endocrine cell' refers to the secretory function of the cell as well as its structure. In genomics, concepts such as 'exon', 'intron', and even 'gene' are defined relative to an activity that must be identified in order to properly understand them. Representing roles by functional categories requires rules (e.g., functional categories cannot subsume categories that are not functional) that allow ontology designers to incorporate functional categories into structures built upon *is-a* and *part-of* relations while preserving consistency. **Associative relations** are the other means for representing roles. For example, Yu & al. developed an ontology concerning genomic concepts that was based on the UMLS Semantic Network [4]. For their purpose, they proposed to extend the UMLS Semantic Network by adding sixteen semantic relations, mostly related to roles that structures can play, e.g., 'promotes'. The work we are reporting on in this paper is part of a wider-scope project that aims at integrating knowledge and data from heterogeneous sources in the context of functional genomics and transcriptomic analysis for iron metabolism and liver diseases. Although there is a general awareness that roles are an important modeling entity, roles are represented diversely in existing systems. This work is a preliminary study that analyzes and discusses representation of roles from examples related to iron metabolism in GO and UMLS.

BACKGROUND

Criteria for an explicit notion of roles have been proposed. The notion of essence, provided by Aristotle, is central to ontology. Strawson introduces the notion of sortal predicates, i.e. those that allow us to identify a thing as a particular kind and are temporally stable [5]. Not all the categories represented in ontologies are sortal, e.g., roles are not sortal predicates. Sowa, in [6] distinguishes between natural types that relate to the essence of entities, and roles that depend on an accidental relationship to some other entity. In Sowa's modeling of conceptual graphs, which relies on a type lattice, roles are subtypes of natural types. For example, Protein (essence) and Enzyme (role) would be subtypes of Substance, and Dehydrogenase would be a hybrid child of both Protein and Enzyme. However, further theoretical basis and pragmatic rules are needed for ontology design and modeling. A step forward, Guarino and Welty have promoted ontological distinctions that rely on the notions of identity, rigidity and dependence [7].

- Identity. The property of carrying an identity condition (IC), i.e. a condition that is both necessary and sufficient for identity (an instance can be recognized as a specific individual).
- Rigidity. A property P is rigid if, for each x, if P(x) is true in one possible world, then it is also true in all possible worlds. Protein is a rigid property, since one cannot lose the property without losing its identity. Carrier, on the other hand, is not a rigid property, since we can imagine something moving in and out the carrier property according to the context, while being the same substance.
- Dependence. A property P is dependent if, necessarily, whenever P(x) holds, then Q(y) holds, with x y. For example, carrier is dependent, since to be a carrier is related to the fact there is something to transport. By contrast, protein is not dependent.

A first distinction can be made between CONCEPTS[1] (we will use upper case in order to distinguish this notion from other occurrences of the word 'concept') and RELATIONS, according to the number of arguments. Among CONCEPTS, Guarino and Welty make distinctions between TYPES and ROLES according to their properties. A TYPE, e.g., 'protein', is rigid and carries an IC. TYPES may also be called sortal, natural or essential types. ROLES, e.g., 'carrier', are anti-rigid, and always dependent. Material roles like carrier do have an IC, while formal roles like part do not. However, the IC of material roles is only indirect, since they do not introduce any specific IC, but rather they inherit it from a subsuming TYPE.

	TYPE	MATERIAL ROLE	FORMAL ROLE
Identity	Yes	Yes	No
Rigidity	Yes	No	No
Dependence	Yes or No	Yes	Yes

Table 1 Some basic kinds of concepts

MATERIAL

GeneOntology: GeneOntology™ (GO) is organized under three top categories:

- Molecular Function: a task performed by gene products (e.g., transcription factor)
- Biological Process: a biological goal accomplished via one or more ordered assemblies of molecular functions (e.g., cAMP biosynthesis)
- Cellular Component: a subcellular structure or macromolecular complex (e.g., nucleus)

As of February 2002, ignoring concepts marked as obsolete in the database, GO contains 4542 process, 4894 molecular function and 929 component concepts[2] (called terms in GO). A gene product has one or more molecular functions is used in one or more biological processes; and may be associated with one or more cellular components. GO itself is not populated with gene products. GO concepts are to be used as attributes of gene products by collaborating external databases, which can make database cross-references between GO concepts and objects in their database (typically, gene products, or their surrogates, genes). Among the gene product databases, GO Annotation @EBI (GOA), Compugen Gene Ontology Gene Association Data, and Swiss-Prot contribute to assignments of gene products to the GO resource. For each term, they provide links towards molecular function (implicitly has-function) and biological process concepts (implicitly has-process) in GO.

The UMLS: The UMLS® comprises two major inter-related components: the Metathesaurus®, a large repository of concepts, and the Semantic Network, a limited network of 134 semantic types. The 2002 edition of the Metathesaurus includes 776,940 concepts and approximately 11,137,725 relationships. Several projects have mentioned the UMLS with application to genetics and molecular biology, e.g., [4, 8, 9, 10].

[1] Although Guarino and Welty use the term of Property instead of CONCEPT, we will use the latter, referring to the basic distinction between concepts and relations in many formalisms. Moreover, Category and Attribution which are other Properties are not represented here.

[2] http://www.geneontology.org/

A biological model for iron metabolism: Iron is central to the health of humans. Pathological conditions associated with altered iron metabolism are frequent and include hemochromatosis, which is characterized by iron overload, and anemia related to iron deficiency or inflammation. Many gene products are involved in iron metabolism [11, 12]. The processes can be complex. For example, L-ferritin synthesis is regulated by iron regulatory protein 1 (IRP1), via an iron-responsive element (IRE) on ferritin mRNA. IRP1 activity is related to iron levels. IRP1 is an iron-sensitive binding protein, i.e. the shape of IRP1 changes according to the iron level, which modifies the ability of interaction with the IRE, thus the synthesis of ferritin. In addition, a protein can play several roles simultaneously. For example, Ceruloplasmin, the major serum copper-containing protein, acts in iron metabolism due to its ferroxidase activity [13]. The main proteins involved in iron metabolism are listed in table 2. A few functions remain partially characterized. Furthermore, iron homeostasis is still being investigated, e.g., hepcidin is a putative iron-regulatory peptide [14].

Protein	Function	Localization
Dcytb	Enterocyte iron uptake Ferric reductase	membrane
DMT1	Enterocyte iron uptake Iron transport	membrane
Transferrin	Plasmatic iron transport	plasma
Transferrin receptor 1	Cellular iron transferrin uptake	membrane
HFE	Regulation of iron absorption ?	membrane
Transferrin receptor 2	Cellular iron transferrin uptake ? (hepatocyte)	membrane
Iron regulatory protein (IRP)	Iron metabolism regulation	cytosol
Ferritin	Cellular iron storage	cytosol
Frataxin	Iron transport	mitochondrion
Ferroportin	Cellular iron egress (enterocyte, macrophage)	membrane
Hephaestin	Enterocyte iron egress Ferroxidase activity	membrane
Ceruloplasmin	Enterocyte iron egress Ferroxidase activity	plasma

Table 2 - Main proteins of iron metabolism

METHODS

Mapping to GeneOntology: The terms corresponding to the twelve proteins of iron metabolism represented in table 2 were mapped to GOA and GO (Feb. 2002 public release) using approximate matching. The mapping was first restricted to human gene products. In case of failure, it was extended to the whole database. The resulting list was compared for validation to that obtained by selecting all the GOA items associated with 'Iron Homeostasis'. QuickGO was used to browse GO[3]. For each term, links towards molecular function and biological process in GO were explored.

Mapping to the UMLS Metathesaurus: The twelve proteins terms were mapped to the UMLS Metathesaurus 12th edition [14], using Knowledge Source Server functionalities (normalized string index) and the UMLS Semantic Navigator[4]. For each term, hierarchical and associative relationships in the Metathesaurus were analyzed as well as its semantic categorization according to the Semantic Network.

RESULTS

Among the 12 iron metabolism proteins that were studied, two gene products (Dcytb and hephaestin) are found neither in GOA nor in the UMLS. It may be noticed that the content of gene product databases is continuously updated. The names of gene products can change (e.g., DMT1 was previously named Nramp2 or DCT1) and all the synonyms may not be represented in a database (e.g. SLCLLA3 iron transporter was found in GOA in place of ferroportin). 10 proteins out of 12 are found in Annotations database. However, in GOA, DMT1 is present as a mouse, not human protein. Three proteins are present in the GO ontology strictly speaking, represented as molecular function concepts: Transferrin Receptor, Ferritin, and multicopper ferroxidase iron transport mediator for Ceruloplasmin. Eight proteins are represented in the UMLS. Every time a protein of the list is found in GO, GOA or UMLS, a function is assigned to it, either by a functional category or by an associative relationship. However, for DMT1 the only function is transporter, without precision.

Functional categories in GO, GOA and the UMLS: In GO, the class Molecular Function includes functional categories, e.g., Ligand binding protein or carrier. The UMLS Semantic Network allows for the categorization of chemicals concepts in the Metathesaurus with both an essence (Chemical viewed structurally and its subtypes) and a role (Chemical viewed functionally and its subtypes). For example, Ceruloplasmin is subsumed by Metalloprotein, which is a Chemical viewed structurally and Oxidoreductase, which is a Chemical viewed functionally.

Associative relations in GO, GOA and the UMLS: Implicit relationships between a gene product and a

[3] http://www.ebi.ac.uk/ego/index.html

[4] http://umlsks.nlm.nih.gov/ Resources Semantic Navigator

biological or molecular activity are designed by associating the gene product and the biological process in Gene Annotation files. For example, GOA files associate Ferritin with Iron Homeostasis and Iron Transport, which are Biological Processes in GO. In GO, the granularity may vary from very general terms, e.g., 'transport', to terms as precise as 'iron incorporation into iron-sulfur cluster via tris-L-cysteinyl-L-cysteine persulfido-bis-L-glutamato-L-histidino tetrairon'. In the UMLS, high-level associative relationships are represented among Semantic Types in the Semantic Network, resulting in predicates such as Biologically Active Substance affects (or complicates) Biologic Function. In addition, associative relationships are recorded among concepts in the Metathesaurus, representing factual knowledge. For example, Iron is related by an 'other' relationship (RO) to Ferritin. However, very few RO relationships are semantically defined in the Metathesaurus. Finally, information about the co-occurrence of MeSH descriptors in MEDLINE® citations is also recorded in the Metathesaurus. For example, Ferritin co-occurs in MEDLINE with Hemochromatosis, and the relation between their respective Semantic Types may be 'affects', 'causes', 'complicates' or 'produced by'. As in this example, however, the semantics of the relation between co-occurring concepts can often not be inferred unambiguously [15, 16]. In other cases, the relation, although unambiguous, remains poorly informative, e.g., Ceruloplasmin 'interacts with' Iron.

DISCUSSION

The representation of roles addresses conceptual conversion between relations and types, i.e. reification. For example, the construction "Transferrin transports Iron" represents a relation. By the cognitive operation of reification, it can be transformed into "Transferrin is an Iron Transporter".

There is a need for classifying biological concepts into functional classes. For example, one would need to list all the ferric iron transporters. While the reified representation of actions in functional categories permits a range of conceptual manipulations [17], major ontological constraints must be emphasized: (1) each entity must be assigned a TYPE in order to satisfy identity condition, (2) no mutual disjointedness is expected for ROLES, since an entity can have several roles, (3) design of hierarchies made of both TYPES and ROLES must follow strict fundamental ontological rules, e.g., a ROLE cannot subsume a TYPE, since the former is anti-rigid and the latter is rigid [7]. Explicit distinction between TYPES and ROLES is useful in presenting specific views, and provides a means to perform inferences. However, models whose underlying paradigm is that "classification by role does not depend on an entity's structure" [18:81], cannot apply to molecular biology. Structural patterns are, "by essence", associated with built-in functions. For example, DMT1 is highly homologous to yeast protein that transports manganese. It belongs to a conserved family of putative transmembrane transporters found in several organisms. Features of these proteins include multiple transmembrane domains, a glycosylated extracytoplasmic loop and a highly conserved intracellular motif [19].

As seen before, a fundamental property of ROLES is dependence. In other terms, ROLES depend on dyadic relation: if x is classified by a role, then x stands in a dyadic relation to some other entity y.

GENE PRODUCT	GO + GOA		UMLS
	GOA and links to GO (F: Molecular Function, B: Biological Process)	GO (F: Molecular Function, OTH: other)	Metathesaurus relationships (A: ascendant; OTH: other); Categorization (ST: Sem. Type)
Iron regulatory protein (IRP)	F: Hydro-lyase B: Metabolism	Not represented	A: Iron-Sulfur Proteins; RNA-binding Proteins ST: Amino Acid, Peptide, or Protein
Ferritin	F: Ligand binding protein or carrier; Ferric iron binding B: Iron homeostasis; Iron transport	OTH: intracellular iron storage	A: MetalloProteins, Iron compounds; OTH: Iron; ST: Amino Acid, Peptide, or Protein; Biologically Active Substance
Ferroportin	F: Iron transporter B: Iron transport; Iron homeostasis; Embryogenesis and Morphogenesis	Not represented	A: Carrier Proteins ST: Amino Acid, Peptide, or Protein
Ceruloplasmin	F: multicopper ferroxidase iron transport mediator; copper binding B: copper homeostasis; iron homeostasis	F: Oxidoreductase; Iron ion transporter OTH: copper homeostasis	A: Oxidoreductases; Alpha-Globulins; Carrier Proteins; Metalloproteins; Acute-Phase Proteins OTH: copper; copper oxidase; Menkes Kinsky Hair Syndrome ST: Amino Acid, Peptide, or Protein; Enzyme

Table 3 – Examples of functions as they are represented in GOA, GO, and UMLS

A protein x transports some entity y, e.g. iron. While functional categories focus on x and leave y implicit inside its definition, associative relations provide an explicit representation of the dyadic predicate, e.g., Transport(x,y). Moreover, functional categories are not adapted when functions are embedded one into another. For example, the GO term 'cation diffusion facilitator' represents in a single item a role (facilitator) whose target is itself a biological function (cation diffusion). By contrast, the representation of roles as relations tends to create a rather limited number of classes that can be combined, e.g., Prevent(Interact(x,y),z) which means that z prevents the interaction between x and z. Relation may involve more than two products, which must be represented as such. General biological mechanisms can be defined by means of associative relations. For example, definition of direct feed back of regulation protein activity (e.g., Iron interacts with IRP1) can be: Interact(x, regulator of the metabolism of x).

In addition, dynamic roles have to be represented. For example, the role of iron on the interaction between the IRE of L-ferritin and IRP1 varies according to its level. Such a role cannot be represented by means of functional categories. Modeling dynamic roles requires approaches such as UML that allow representation of complex entities and scenarios. Moreover, while relations in ontologies reflect discrete models of the world, continuous models are needed for representing gene activity, since intermediate gene activity levels or substance levels are important for some aspects of existing interactions, e.g. [20].

CONCLUSION

Functional categories provide a valuable means to classify biological entities according to their roles. However, they must be integrated into ontologies with respect of formal rules (e.g., Guarino's). Associative relations, on the other hand, provide explicit representation of dyadic predicates underlying roles and allow the representation of roles that apply to functions, and other complex predicates. Furthermore, models in functional genomics require representation of built-in functions and dynamic associative relations.

REFERENCES

1. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res 2001; 11(8): 1425-1433
2. Lindberg DAB, Humphreys BL, McCray AT, The Unified Medical Language System. Meth Inform Med, 1993, 32(4): 281-91
3. Pustejovsky J. The generative lexicon. MIT Press, MA, Cambridge, 1996
4. Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp 1999;:181-5
5. Strawson PF. Individuals: an essay in descriptive metaphysics, Routledge, London, 1959
6. Sowa J.F. Conceptual Structures: Information processing in mind and machine, Reading, MA: Addison Wesley, 1984
7. Guarino N, Welty C. A formal ontology of properties. In R. Dieng and O. Corby (eds.), Proc. EKAW2000. Springer Verlag: 97-112.
8. Sperzel WD et al. Biomedical database inter-connectivity: an experiment linking MIM, GENBANK, and META-1 via MEDLINE.Proc Annu Symp Comput Appl Med Care. 1991;:190-3.
9. Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. Proc AMIA Symp. 1999;127-31.
10. Ashburn M. On the representation of gene function in genetic databases, 1998, available at http://www.geneontology.org/
11. Lee PL, Gelbart T, West C, Halloran C, Felitti V, Beutler E. A Study of genes that may modulate the expression of hereditary hemochromatosis: Blood Cells Mol Dis 2001;27(5):783-802
12. Brissot P. Hemochromatosis at the intersection of classical medicine and molecular biology. C R Acad Sci III 2001, 324(9):795-804
13. Kaplan J, O'Halloran TV. Iron metabolism in eukaryotes: Mars and Venus at it again. Science 1996 15;271(5255):1510-1512
14. Pigeon C et al. A new mouse liver-specific gene, encoding a protein homologous to human antimicrobial peptide hepcidin, is overexpressed during iron overload. J Biol Chem 2001 Mar 16;276(11):7811-9
15. McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In *Semantics of Relationships*, Myeng SH and Green R Eds Kluwer; to appear
16. Burgun A, Bodenreider O. Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. Medinfo. 2001;10(Pt 1):171-5.
17. Talmy L. Toward a cognitive semantics. MIT Press, MA, Cambridge, 2000
18. Sowa JF. Knowledge Representation. Brooks Cole, 2000
19. Fleming MD et al. Microcytic anemia mice have a mutation in Nramp2, a candidate iron transporter gene. Nature Genetics, 1997, 16: 383-6
20. Murphy K, Mian S. Modeling gene expression data using dynamic bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA.