An ontology describing congenital heart defects data

Charalampos Moschopoulos[1,2]✉, **Jeroen Breckpot**[3], **Yves Moreau**[1,2]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Katholieke Universiteit Leuven, Leuven, Belgium
[2]iMinds Future Health Department, KU Leuven, Belgium
[3]Center for Human Genetics, KU Leuven, Leuven

Received 9 September 2013; **Accepted** 9 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Congenital heart defects (CHDs) are a group of diseases characterized by a structural anomaly of the heart that is present at birth. They are considered as the commonest cause of childhood death in developed countries. The causes of congenital heart disease are still under investigation, but their strongly presumed to be genetic or a combination of genetic and environmental factors. In order to store the derived knowledge, a collaborative knowledge base called CHDWiki has been developed, which collects all the information about the genetic basis of CHDs. However, this dedicated web resource suffers the same problems with similar portals where heterogeneous information is hosted. Further steps should be taken in order to ensure the interoperability of the available data. Also, the hosted data should be offered in a machine-readable format in order to direct be used by other Bioinformatic tools. In order to solve these problems, the life science scientific community tends to use semantic web technologies, which have proved their efficiency through numerous examples including VariO ontology, Gene Ontology (GO), Orphanet Ontology of rare diseases (OntoOrpha) and many more. In this contribution, we present an ontology, which describes CHD data, developed around three main data categories: genotype, phenotype (CHDs) and clinical reports. Retrieving data from CHDWiki, this ontology describes the relationships between genes and human phenotypes, derived from published data or single clinical cases, providing a useful tool to geneticists, molecular biologists and clinicians. This ontology hosts information about syndromic genes, chromosomal aberrations that may cause CHDs and associations between genes and CHDs. Further information is included such as structural variations and single point mutations.

Motivation and Objectives

Congenital heart defects (CHDs) are a group of diseases characterised by a structural anomaly of the heart that is present at birth. They are considered as the commonest cause of childhood death in developed countries; see, e.g., the web site of the American Heart Association[1]. The causes of congenital heart disease are still under investigation, but they are strongly presumed to be genetic or a combination of genetic and environmental factors (Jenkins *et al.*, 2007). In order to store the derived knowledge, a collaborative knowledge base called CHDWiki (Barriot *et al.*, 2010) has been developed, which collects all the information about the genetic basis of CHDs.

However, this dedicated web resource suffers the same problems with similar portals where heterogeneous information is hosted. Further steps should be taken in order to ensure the interoperability of the available data. Also, the hosted data should be offered in a machine-readable format in order to directly be used by other bioinformatic tools. In order to solve these problems, the life science scientific community tends to use semantic web technologies, which have proved their efficiency through numerous examples including VariO ontology[2], Gene Ontology (GO) (Ashburner *et al.*, 2000), Orphanet Ontology of rare diseases (OntoOrpha) (Aime *et al.*, 2012), and many more.

In this contribution, we present an ontology, which describes CHD data, developed around three main data categories: genotype, phenotype (CHDs) and clinical reports. Retrieving data from CHDWiki, this ontology describes the relationships between genes and human phenotypes derived from published data or single clinical cases, providing a useful tool to geneticists, molecular biologists and clinicians. This ontology hosts information about syndromic genes, chromosomal aberrations that may cause CHDs and associations between genes and CHDs. Further information is included such as structural variations and single point mutations.

Methods

As mentioned before, the CHD ontology is composed by three main concepts (genes, CHDS and clinical reports) and the between them relationships. These three concepts are further classified using a `is _ a` relation as it can be seen in Figure 1. The CHD family of diseases is hierarchi-

[1] http://www.heart.org/HEARTORG

[2] http://variationontology.org/

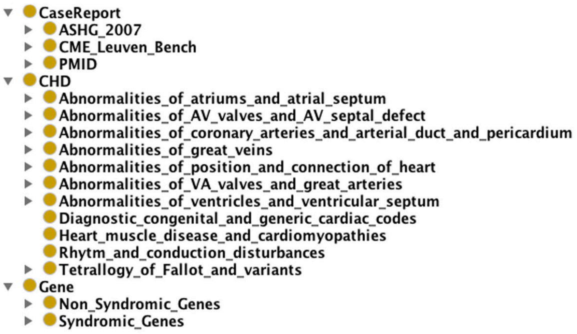
cally structured, starting from the CHD concept which is further refined in 11 subclasses (e.g., "Abnormalities of atriums and atrial septum", "Abnormalities of great veins", "Rhythm and conduction disturbances", etc). These concepts, in some cases, are further refined to more specific congenital heart defects. The other two main classes of the CHD ontology are also refined: the gene entities, which are subclasses of the concept Gene, are belonging either to the subclasses of the "Non Syndromic Genes" or "Syndromic Genes". The clinical case entities fall into one of the following categories: "ASHG" which stands for American Society of Human Genetics, "CME_Leuven_Bench" which refers to unreported patients with causal CNVs from Leuven Hospital, and "PMID" which stands for case reports in literature (PubMed ID).

Two further relationships were used in our ontology: `connected_with`, which describes an association between a gene and a CHD, and `involved_in`, which describes an association between a gene or a CHD and a case report. It has to be noted that CHDWiki users have manually curated each association between entities, providing high quality data.

We also created a vocabulary that describes the relationships between the ontology terms. As we consider the extensibility of ontologies a very important property, we used words into our vocabulary derived from two well-known ontologies: the Dublin Core Metadata Initiative[3] (DCMI) and the RDF Schema[4] (RDFS). Each class includes a set of properties such as structure variation (where deviation, chromosome, start and end region are referred) and single point mutation (where DNA mutation, peptide mutation and relevant reference are referred). Moreover, whenever was possible, we used external URIs for each class property. For instance, the property "label" of a gene refers to the corresponding gene page at Ensemble web portal. Also a variety of external URLs are provided under the "seeAlso" property of each Gene and CHD subclass.

To annotate the basic CHD ontology classes using unique URIs, we used well-known IDs such as the shortlist of the Association for European Pediatric Cardiology[5] (AEPC) for the CHDs, the Hugo IDs (Gray *et al.*, 2013) for the associated genes and the CHDWiki case report IDs for the recorded clinical cases. Specifically, the AEPC URI annotation for the CHDs is considered more accurate than using OMIM[6] annotation as this disease family is not so well characterised by it. Besides that, it has to be noted that the created URIs of the CHDs and the clinical cases refer to CHDWiki pages, while the associated gene URIs refer to the HUGO Gene Nomenclature Committee[7] (HGNC) web pages.

Table1. The class hierarchy of the CHD ontology (snapshot from Protégé 4.2).



- CaseReport
  - ASHG_2007
  - CME_Leuven_Bench
  - PMID
- CHD
  - Abnormalities_of_atriums_and_atrial_septum
  - Abnormalities_of_AV_valves_and_AV_septal_defect
  - Abnormalities_of_coronary_arteries_and_arterial_duct_and_pericardium
  - Abnormalities_of_great_veins
  - Abnormalities_of_position_and_connection_of_heart
  - Abnormalities_of_VA_valves_and_great_arteries
  - Abnormalities_of_ventricles_and_ventricular_septum
  - Diagnostic_congenital_and_generic_cardiac_codes
  - Heart_muscle_disease_and_cardiomyopathies
  - Rhytm_and_conduction_disturbances
  - Tetrallogy_of_Fallot_and_variants
- Gene
  - Non_Syndromic_Genes
  - Syndromic_Genes

Results and Discussion

In our project, we used Protégé 4.2[8] and the ontology was built in OWL format. We chose not to use OBO format as OWL is a more expressive language and we could also describe the relations between our ontology classes, which were not always an `is_a` instance. As Protégé reasoner presents some limitations concerning the class properties querying, we also created our ontology in simple RDF turtle format. Our ontology is also hosted at the Bioportal (Whetzel *et al.*, 2011), which is the most popular open repository of biomedical ontologies.

As a future work, we will continue the enrichment of the created ontology and will test its biological usefulness by applying queries on very specific research questions. Also, we plan to connect the CHDWiki data ontology with other well-known bio-ontologies such as the Human Phenotype Ontology (HPO) (Robinson *et al.*, 2008). This way, queries with increased information value for doctors and geneticists could be applied. Finally, we plan to create a SPARQL[9] endpoint that, accompanied with a friendly in-

[3] http://dublincore.org/
[4] http://www.w3.org/TR/rdf-schema/
[5] http://www.aepc.org/
[6] http://omim.org/
[7] http://www.genenames.org/
[8] http://protege.stanford.edu/
[9] http://www.w3.org/TR/rdf-sparql-query/

terface, will be hosted on CHDWiki portal. The goal is to encourage untrained users to query the available data without having any previous knowledge regarding semantic technologies such as RDF and SPARQL. While the scientific interest on rare diseases is continuously increasing, ontologies could play a vital role into the integration and further analysis of the generated data

Acknowledgements

We wish to thank Dr. Peter Robinson (Charité - Universitätsmedizin Berlin) for the helpful discussions.

Funding: Research supported by: Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, IOF: HB/12/022 Endometriosis. Flemish Government: FWO: PhD/postdoc grants, projects: Hercules Stichting: Hercules III PacBio RS, iMinds: SBO 2013. Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate). Action BM1006: NGS Data analysis network.

References

Aime X, Charlet J, *et al.* (2012) Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM. *Studies in health technology and informatics* **180**, 88-92. doi:10.3233/978-1-61499-101-4-88

Ashburner M, Ball CA, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29. doi:10.1038/75556

Barriot R, Breckpot J, *et al.* (2010) Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med* **2**, 16. doi:10.1186/gm137

Gray KA, Daugherty LC, *et al.* (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* **41**, D545-552. doi:10.1093/nar/gks1066

Jenkins KJ, Correa A, *et al.* (2007) Noninherited risk factors and congenital cardiovascular defects: current knowledge: a scientific statement from the American Heart Association Council on Cardiovascular Disease in the Young: endorsed by the American Academy of Pediatrics. *Circulation* **115**, 2995-3014. doi:10.1161/CIRCULATIONAHA.106.183216

Robinson PN, Kohler S, *et al.* (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610-615. doi:10.1016/j.ajhg.2008.09.017

Whetzel PL, Noy NF, *et al.* (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**(Web Server Issue), W541-545. doi:10.1093/nar/gkr469