The Ontology of Primary Immunodeficiency Diseases (PIDs) – Using PIDs to Rethink the Ontology of Phenotypes

Nico Adams[1,3], Christian Hennig[2], Robert Hoehndorf[1,3], Anika Oellrich[1], Dietrich Rebholz-Schuhmann[1], Gesine Hansen[2]

[1]European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom,
[2]Department of Paediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, D-30625 Hannover, Germany,
[3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, United Kingdom

ABSTRACT

Primary immunodeficiency diseases (PIDs) are the consequence of genetic disorders and usually manifest themselves in very young patients. Because of their rarity, they are notoriously difficult to diagnose both for general practitioners and clinicians. In this paper, we present the foundations of an ontology of PIDs, which will be at the heart of an expert system designed to assist the clinician in the diagnosis of these diseases. To achieve this, the PIDOntology characterises Primary Immunodefieciencies in terms of Phenotypes. While there are a number of different ontologies already available that allow the description of phenotypes and phenotypic qualities, these have a number of associated ontological problems, which we will also address as part of this paper. We use the subtype of Hyper-IgE Syndrome caused by a STAT3 defects as an example of a primary immunodeficiency and show how the clinical phenotype of the disease can be modeled in terms of other phenotypes by introducing the notion of the "phene". Furthermore, we develop patterns for different types of phenes and show, that these patterns can be mapped onto more traditional entity-quality statements, which are the current state of the art in phenotypic modeling.

1 INTRODUCTION

Primary Immunodeficiency Diseases (PIDs) are a group of disorders caused by the absence of or by defects in genes involved in regulating and coordinating the body's immune system. Their incidence varies from relatively common (1:1,200) to extremely rare (1:2,000,000) [Lim and SJ, 2004]. PIDs can manifest themselves in newborns and toddlers, but also much later in life, which makes their identification and diagnosis difficult for both the general practitioner and the experienced clinician alike. The fact that primary immunodeficiencies represent a large and diverse group of disorders complicates the picture even further: currently, the International Union of Immunological Societies' classification recognises 140 different PIDs [Geha et al., 2007], although more than 200 primary immunodefiencies have now been identified.

Informatics resources that contain condensed and structured information, which can be used for the diagnosis of PIDs remain relatively scarce, and include the ImmunoDeficiency Resource [Vliaho et al., 2002] and INFO4PI [Foundation, 2010]. A recent review provides a comprehensive overview over the existing bioninformatics services relating to PIDs [Samargithean and Vihinen, 2009]. Furthermore, few of the available services for PIDs leverage the power of semantic web technologies, and allow the semantic codification of knowledge.

1.1 Why an ontology of Primary Immunodeficiency Diseases?

One important component of the semantic web are ontologies, which are formal and – importantly – computable specifications of a shared conceptualisation. The overall aim of our research is the development of an ontology-driven expert system for clinicians, which can assist in the diagnosis of primary immunodeficiency diseases.

Apart from diagnostic tools, we also expect that the ontology will be useful in a number of other areas in the future: it is reasonable to assume, that it will form one of the components for semantically-rich publishing of academic work relating to research in primary immunodeficiency disorders and that it will also be useful for the integration with knowledge contained in other ontologies as well as the integration of existing and new data resulting from the development of relevant new research and laboratory techniques such as iterative chip-based cytometry [Hennig et al., 2009]. In this contribution, we report on the development of the underlying ontological principles of the PID Ontology. In the future, we will address the delivery of an ontology-driven and web-delivered expert system as well as formal methods for the comparison of observed and formally codified phenotypes for primary immunodeficiency diseases.

2 METHODS

2.1 Constructing the PID Ontology

The scope of the PID ontology is to serve as both a reference and an application ontology, which defines a phenotype of a primary immunodeficiency disorder in terms of a set of biomarkers. There is little restriction on the type of biomarker the ontology references. Typical biomarkers are, for example, gene defects, disorders, clinical or laboratory findings and genetic inheritance patterns. Collectively, these biomarkers form the "canonical" phenotype of a primary immunodeficiency disorder.

The PID ontology has been formalized in the OWL 2 ontology language, which has a mechanism for annotating assertions in the ontology. This is of critical importance for the development of the ontology: every assertion relating a primary immunodeficiency to a symptom is annotated with the literature source, from which the assertion was derived by a human curator. An example implementation of the ontology can be downloaded from `http://bitbucket.org/na303/pidontologyexample/`.

In the development of the ontology we aim to be term-orthogonal with respect to other established ontologies in the biomedical domain – in particularly those developed under the umbrella of the OBO Foundry. Furthermore, we have decided to use the General Formal Ontology (GFO) [Herre et al., 2006] as an upper ontology due to its integration of objects and processes, its rich classification of roles [Loebe, 2005] and the expressive axioms in its OWL version.

2.2 Phenotypic characters and phenes

The dominant view in the biomedical ontology community is that phenotypic characters are qualities of entities. This is exemplified in the Entity-Quality (EQ) formalism [Mungall et al., 2010] and the databases that use EQ for their annotations, as well as in the PATO ontology [Gkoutos et al., 2004] which provides the qualities that **inhere in** an entity.

We note, however, that there are at least two fundamentally different types of qualities in the PATO: qualities of objects that can be *discovered* by an observer, and qualities of objects that are *established* by an observer. Qualities of the first kind are qualities like *red*. Qualities that are *established* by an observer are all those qualities that are defined with respect to an explicit or implicit *norm*: *increased* size or *lacking parts*.

In particular, there is a kind of quality in the PATO that should *not* be considered to be an ontological quality. PATO contains the category *count* with sub-categories *present* and *absent*. Yet *absent* cannot be a quality *of* some entity in the same sense as *weight* or *shape* can be qualities *of* an entity: qualities are existentially dependent on their

bearer, they cannot exist without the entity *of which* they are a quality. But in the case of *absent*, such a bearer does not exist [Hoehndorf et al., 2010]. *Absent* was introduced in PATO in order to facilitate the curators' need to annotate absent parts. We provide a formal treatment of absence using a relation to allow the inference of *has-part* and *lacks-part* relations from these qualities. This addresses the difficulty of using EQ formalisms to establish an information flow between an assertion about an entity's qualities to statements involving propositions about the parts, qualities or dispositions the entity has.

Additionally, having or lacking parts is not *primarily* a quality: first, there is the absence of parts, *not having an X as part*, a structural feature expressed using negation and the **part-of** relation, a feature of an organism that is distinct from an ontological quality.

Therefore, although qualities in the sense of PATO certainly seem to play an important role in phenotypes, and reference to qualities is often necessary to describe phenotypes, phenotypes may not always be qualities, at least not in the sense of PATO, as existentially dependent entities inhering in a bearer. To formalize PIDs, we use the EQ formalism to provide compatibility with the rich resources already available in this formalism, and also use a new method for the semantic representation of phenotypes that interoperates with OWL and Semantic Web technology and permits making the semantics of phenotypic descriptions explicit in these formalisms. To set these semantic phenotype descriptions apart from the EQ phenotypes, we call them *phene*s [Allan, 2008] in the context of this paper.

A *phene* is a basic observable characteristic possessed by an organism. Phenes are attributive individuals in the sense that they are existentially dependent on a bearer, and they are related to their bearer by the **pheneOf** relation. The **pheneOf** relation has an inverse which we call **hasPhene**.

We use a general definition pattern for phenes. This pattern is based on the observation that having a phene means exhibiting certain features and properties. We define a category of phenes *X* as the category of phenes "of entities with the property *Y*". The property *Y* is expressed as class-membership in description logic or unary predicates in first-order logics. The analysis of the ontological status *Y* is outside the scope of this manuscript.

```
X EquivalentClass Phene and pheneOf some Y
```

2.3 Biomarkers and Biomarker Roles

The notion of an entity acting as a biomarker is central to the PID ontology. Biomarkers are phenes that characterise variation in cellular or biological components, pathways, etc., and where the variation is objectively measured and observed. Therefore, biomarkers are phenes participating in clinical diagnosis processes in the biomarker role. The form of participation is dependent on the observer, and the PID ontology contains a classification of biomarkers based on the kind of role they play in the observation process. Generally, we may define a "biomarker" as

```
Phene and (plays-role some Biomarker_Role).
```

By specifying the type of observation process, we may further define the type of biomarker. An "imaging biomarker", for example, is a biomarker, which is observed in a radiological observation process (e.g. projection radiography or Computed Tomography Scanning). By analogy, a cellular biomarker is a biomarker observed during a cytometric experiment. The PID ontology will provide an extensive hierarchy of biomarkers, which is useful for the further classification of phenes and any one phene will be able to assume multiple biomarker roles. In the first instance, the classification of phenes in terms of biomarkers will mirror the way in which most clinicians classify phenes. However, due to the definition of biomarkers via the role mechanism, the axiomatic construction of inferred polytaxonomies is eminently feasible. The formal integration of PID ontology with diagnotic processes is the subject of future work.

2.4 Diseases and syndromes

Although the PID ontology uses terms referring to syndromes and diseases, it does not actually classify diseases or syndromes themselves. Instead, the PID ontology is an ontology of the *disease phenotype* of primary immune deficiency diseases.

In other ontologies, diseases are sometimes classified as processes, dispositions or qualities [Scheuermann et al., 2009]. We hold that a phenotypic description of diseases is more general than each approach. Using phenes permits the representation of dispositions, qualities, processes and other attributes of organisms in a single coherent framework.

2.5 Relation to other ontologies

Whenever possible, we reuse terms from ontologies in the OBO and OBO Foundry. In particular, we use the Foundational Model of Anatomy [Rosse and Mejino, 2003], PATO [Mungall et al., 2010], the Human Phenotype Ontology [Robinson et al., 2008], Mouse Pathology Ontology and OBI [Courtot et al., 2008].

In addition to defining PIDs in terms of phenes using OWL, we provide EQ definitions of the biomarkers in the PID ontology. These serve to facilitate compatibility with those resources that have

been annotated using the EQ formalism. Additionally, we document transitional patterns for formally defining the EQ statements using the method introduced here.

3 DISCUSSION

3.1 Use-case: Hyper-IgE syndrome caused by STAT3 Defects

The Hyper-IgE syndrome (HIES), sometimes also known as "Job's syndrome" is a collective term for a set of complex immunodeficiencies. First reported by Davis et al. [1966], it is characterised by increased serum IgE levels, chronic dermatitis and serious recurrent infections such as pneumonia and recurrent staphylococcal skin abscesses. Staphylococcal infections can also affect lungs, joints and other sites. Other signs and symptoms which have been observed are atypical eczema, pneumatoceles, and osteopenia. Furthermore, patients often have fair skin and red hair as well as "lion-like" facial features, caused by a high palate. The inheritance pattern is autosomal dominant or recessive and patients affected by the autosomal dominant form often fail to shed their primary teeth. Additionally, some patients also suffer from scoliosis [Grimbacher et al., 1999]. The presentation of Hyper-IgE syndrome varies from patient to patient and is also age-dependent. However, no patient will present with the whole gamut of clinical indicators for the disease, that have been observed over the totality of all patients.

We use the Hyper-IgE syndrome caused by STAT3 defects as an exemplar to show how this complex phenotype can be modeled in terms of phenes and how phenes themselves can be considered to be biomarkers in the context of a diagnosis process. We focus on the representation of the *canonical* phenotype in the ontology. Patients, however, can be non-canonical with respect to the canonical phenotype of the Hyper-IgE syndrome.

One phenotypic trait characterizing the syndrome is the absence of Th17 cells. Using the EQ method, this is formalized as

```
[Term]
id: absence_of_Th17_cell
intersection_of: PATO:0001557 ! lacking physical part
intersection_of: towards CL:0000899 ! Th17 cell
```

Using our method, this can be extended and formalized using the phene *Absence of Th17 cells* as

```
Phene and pheneOf some (not (has-part some CL:0000899))
```

which enables the inference that entities with this phene have no Th17 cells as part.

Another phene of the Hyper-IgE syndrome is pneumonia. In the HPO, a *pneumonia* is an inflammation of the lung:

```
[Term]
id: HP:0002090 ! pneumonia
intersection_of: PATO:0001561
                 ! having extra processual parts
intersection_of: inheres_in FMA:7195 ! lung
intersection_of: towards MPATH:212 ! inflammation
```

Inflammation is a process and there are different forms of participation in processes. In the General Formal Ontology, these are modeled using processual roles. In an inflammation process, we may distinguish between an inflammatory agent and an inflamed structure. In pneumonia, the lung plays the role of the inflamed structure and a definition of the inflammatory agent can be used to further differentiate the type of pneumonia (e.g. staphylococcal pneumonia, where a strain of Staphylococcus plays the role of the inflammatory agent). Using our notion of phenes, we may therefore rewrite the EQ definition in the following form:

```
Phene and pheneOf some (has-proper-part
    some (FMA:7195 and plays-role
    some Inflamed_Entity))
    and (plays-role some BiomarkerRole)
```

Inflamed Entity, in turn, can be defined as

```
ProcessualRole and role-of some MPATH:212
```

A processual role is a **role-of** a process, and playing the role implies participation in this process. Furthermore, the use of "has-proper-part" as opposed to "has-part" allows the distinction between a local inflammation (pneumonia is localised in the lung) and global inflammation (e.g. hemophagocytotic syndrome) and, more generally local and global parts.

We can use phenes to model additional parts as well. A related PID, the Wiscott-Aldrich-Syndrome, is characterized by B-cell lymphocytic neoplasms, which would be formalized using EQ and PATO and MPATH as:

```
[Term]
id: B_Cell_Lymphocytic_Neoplasm
intersection_of: PATO:0002002
                 ! having extra physical parts
intersection_of: inheres_in FMA:FMA:20394
                 ! human body
intersection_of: towards MPATH:516 ! B-cell neoplasms
```

The use of the PATO quality *having extra physical parts* hides the semantics of the phene, i.e., that the phene's bearer has a B-cell neoplasm as **part**. Therefore, we define a phene *Having B-Cell lymphocytic neoplasm* as

```
Phene and pheneOf some
                 (Human and has-part some MPATH:516)
```

This differs from the EQ statement in that it explicitly establishs a relation between having the neoplasm and the parts of the organism.

3.2 Comparison

Beyond the advantages of phenes and the problems associated with the EQ formalism discussed above, our use of the phene formalism makes much of the implicit semantics contained in entity-quality statements explicit: taking the B-cell lymphocytic neoplasm as an example, the use of the PATO quality "having extra physical parts" hides the semantics of the particular phenotype, namely that the phenotype's bearer has a B-cell lymphocytic neoplasm as **part**. The use of phenes as shown above explicitly establishes the relationship between having the phene and the parts of the organism that bear it. Furthermore, the phene formalism allows the correct use of qualities: absence (as in absence of Th17 cells) is not a quality as currently modeled by PATO and phenes allow us to model absence as not having a *part*. In spite of the differences in approach, the phene-formalism is "backwards compatible" with the

more traditional EQ approach and should allow the mapping and interconversion of ontologies using either of these two frameworks.

3.3 Future research

Future research will pursue several different strands. Firstly, the continued enrichment of the ontology with content will be the highest priorty. As discussed above, the use of relations such as **part-of** in the definition of phenes can lead to inconsistencies when combining these with canonical ontologies. In future research we will address this, by investigating how canonical ontologies formalized in OWL can be restructured using *Canonical* and *Non-canonical* classes.

The appropriate definition of phenes such as "small platelets size" (a phene for Wiskott-Aldrich Syndrome) or "thrombocytopenia" remains an unsolved issue. The former refers to the fact that the average of the platelet size distribution in a "non-canonical patient" is shifted to lower values with respect to that in the "canonical patient" and thrombocytopenia denotes the situation in which the number of platelets in the blood of a "non-canonical" patient is reduced with respect to the number found in a "canonical patient". The current state of the art is to use the EQ framework, for which, however, no explicit semantics is currently available to formalise concepts such as "small platelet size". Furthermore, some symptoms of a disease only manifest themselves in a fraction of all patients that share a common diagnosis.

Finally, we will address the development of formal methods for the comparison of canonical phenotypes (i.e. phenotypes encoded in an ontology) and observed phenotypes (the phenotype presented by a patient) and how these comparisons can be used to assist the clinician in the diagnosis of primary immunodeficiency diseases.

4 CONCLUSION

We have developed the ontological basis for the description of primary immunodeficiency syndromes by defining a semantically rich representation of basic observable characteristics in organisms. Building on this, we have developed a view of a primary immunodeficiency disease as a set of complex phenes, described by other simpler phenes. We show that this new formalism and the modeling of phenotypic characters using entity-quality statements are compatible.

The method we use to characterize the canonical disease phenotypes of primary immunodeficiency syndromes can be applied to all forms of disease. Thereby, the PID ontology serves as an example for the development of disease and phenotype ontologies in general. Application of the method used in constructing PID leads to an integration with canonical ontologies, allows for an explicit representation of abnormality and facilitates knowledge-based inferences over observed phenomena.

REFERENCES

Charlotte L. Allan. Schizophrenia: From genes to phenes to disease. *Current Psychiatry Reports*, 10(4), 2008.

Mélanie Courtot et al. The owl of biomedical investigations. In Catherine Dolbear, Alan Ruttenberg, and Ulrike Sattler, editors, *OWLED*, volume 432 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

SD Davis, J Schaller, and Wedgwood RJ. Job's syndrome : Recurrent, "cold", staphylococcal abscesses. *Lancet*, 287(7445), 1966.

Jeffrey Modell Foundation. Info4pi, 2010. URL `http://www.info4pi.org`.

R Geha et al. Primary immunodeficiency diseases: An update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *Journal of Allergy and Clinical Immunology*, 120(4):776–794, 2007.

G. V. Gkoutos, E.C.J. Green, A-M Mallon, J.M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biology*, 6(1):R8, 2004.

B Grimbacher, SM Holland, JL Gallin, F Greenberg, SC Hill, HL Malech, JA Miller, AC O'Connell, and Puck JM. Hyper-IgE syndrome with recurrent infections–an autosomal dominant multisystem disorder. *New England Journal of Medicine*, 340(9): 692–702, 1999.

Christian Hennig, Nico Adams, and Gesine Hansen. A versatile platform for comprehensive chip-based explorative cytometry. *Cytometry, Part A*, 75A(4):362–370, 2009.

Heinrich Herre, Barbara Heller, Patryk Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 2006.

Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo, and Janet Kelso. Applying the functional abnormality ontology pattern to anatomical functions. *Journal for Biomedical Semantics*, 2010. in press.

SM Lim and Elenitoba-Johnson SJ. The molecular pathology of primary immunodeficiencies. *Journal of Molecular Diagnostics*, 6(2):59–83, 2004.

Frank Loebe. Abstract vs. social roles: A refined top-level ontological analysis. In G. Boella, J. Odell, L. van der Torre, and H. Verhagen, editors, *Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems'*. AAAI Press, 2005.

Christopher Mungall, Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.

Peter N. Robinson et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, 83(5):610–615, 2008.

Cornelius Rosse and Jose L. V. Mejino. A reference ontology for bioinformatics: The foundational model of anatomy. *Journal of Biomedical Informatics*, (36):478–500, 2003.

C. Samargithean and M. Vihinen. Bioinformatics services related to diagnosis of primary immunodeficiencies. *Current Opinion in Allergy and Clinical Immunology*, 9(6):531–536, 2009.

Richard H. Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. In *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, pages 116–120, 2009.

J Vliaho, M Pusa, T Ylinen, and M Vihinen. IDR: the immunodeficiency resource. *Nucleic Acids Research*, 30(1): 232–234, 2002.