BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 22 2011, pages 3193–3199
doi:10.1093/bioinformatics/btr531

Databases and ontologies

Advance Access publication September 22, 2011

PIDO: the primary immunodeficiency disease ontology

Nico Adams[1,2,*,†] Robert Hoehndorf[1], Georgios V. Gkoutos[1], Gesine Hansen[3] and Christian Hennig[3]

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

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Primary immunodeficiency diseases (PIDs) are Mendelian conditions of high phenotypic complexity and low incidence. They usually manifest in toddlers and infants, although they can also occur much later in life. Information about PIDs is often widely scattered throughout the clinical as well as the research literature and hard to find for both generalists as well as experienced clinicians. Semantic Web technologies coupled to clinical information systems can go some way toward addressing this problem. Ontologies are a central component of such a system, containing and centralizing knowledge about primary immunodeficiencies in both a human- and computer-comprehensible form. The development of an ontology of PIDs is therefore a central step toward developing informatics tools, which can support the clinician in the diagnosis and treatment of these diseases.

Results: We present PIDO, the primary immunodeficiency disease ontology. PIDO characterizes PIDs in terms of the phenotypes commonly observed by clinicians during a diagnosis process. Phenotype terms in PIDO are formally defined using complex definitions based on qualities, functions, processes and structures. We provide mappings to biomedical reference ontologies to ensure interoperability with ontologies in other domains. Based on PIDO, we developed the PIDFinder, an ontology-driven software prototype that can facilitate clinical decision support. PIDO connects immunological knowledge across resources within a common framework and thereby enables translational research and the development of medical applications for the domain of immunology and primary immunodeficiency diseases.

Availability: The Primary Immunodeficiency Disease Ontology is available under a Creative Commons Attribution 3.0 (CC-BY 3.0) licence at http://code.google.com/p/pido/. The most recent public release of the ontology can always be found at http://purl.org/scimantica/pido/owl/pid.owl. An instance of the PIDFinder software can be found at http://pidfinder.appspot.com

Contact: nico.adams@csiro.au

Received on May 13, 2011; revised on September 13, 2011; accepted on September 18, 2011

1 INTRODUCTION

1.1 Immunological and clinical motivation

Primary immunodeficiency diseases (PIDs) are Mendelian diseases of low incidence, caused by defects in genes involved in the development, maintenance and regulation of the immune system. PIDs most often affect toddlers and infants, but can also manifest much later in life and into adulthood (Riminton and Limaye, 2004). As a disease group, PIDs are extremely heterogeneous and according to the most recent classification of the International Union of Immunological Societies Primary Immunodeficiency Disease Classification Committee, >150 distinct forms of PID have been identified (Geha *et al.*, 2007), although >200 PID genes are known at the time of writing (Keerthikumar *et al.*, 2009).

These facts make primary immunodeficiencies a challenging group of diseases for both the practicing clinician and the biomedical researcher alike. The first stumbling block is the comparatively low incidence of these diseases: a recent study suggests that the average prevalence of a PID in US households is ∼1 in 1200 persons (Boyle and Buckley, 2007), although other sources suggest incidences of up to 1 in 2 000 000. Many general practitioners as well as clinicians have little or no familiarity with PIDs and consequently the time that elapses from the first manifestation of symptoms to a confirmed diagnosis is often long: a recent study investigating children with *Common Variable Immunodeficiency Disorders* (CVID) found that the mean time between the manifestation of symptoms and the induction of immunoglobulin substitution therapy is 5.8 years (Urschel *et al.*, 2009). A second stumbling block is the fact that the phenotypic variation associated with PIDs is usually very high: in the case of patients with defects in the Wiskott–Aldrich Syndrome Protein (WASP) gene, for example, the exact nature of the gene defect (e.g. deletion versus missense or nonsense mutation, precise location of splice-site abnormalities) will determine, whether patients exhibit fully developed Wiskott–Aldrich Syndrome (WAS) (X-linked thrombocytopenia, B-cell lymphoma, frequent bacterial and fungal infections, eczema, small platelet sizes, etc.) or a milder form (X-linked thrombocytopenia or neutropenia), which, in turn, gives rise to a less-complex phenotype. A final stumbling block for clinicians is associated with information retrieval: there are very few information resources containing structured and both human and machine-comprehensible information related to PIDs and their phenotypes. Examples of dedicated domain databases include the ImmunoDeficiency Resource, (Väliaho *et al.*, 2005), Info4PI (Samarghitean and Vihinen, 2009) and the ESID Registry

*To whom correspondence should be addressed.

†Present address: CSIRO Materials Science and Engineering, Bayview Avenue, Clayton, VIC 3168, Australia.

(Guzman *et al.*, 2007). Valuable information concerning primary immunodeficiencies is also contained in general bioinformatics databases such as the Online Mendelian Inheritance in Man database (Hamosh *et al.*, 2005), which contains descriptions of phenotypes associated with Mendelian diseases in free text form, ArrayExpress (Parkinson *et al.*, 2009) or the Gene Expression Omnibus (Barrett *et al.*, 2009) for functional genomics data, UniProt (Consortium, 2010) (information about proteins), IntAct (Aranda *et al.*, 2010) (protein/protein interactions) or KEGG (Kanehisa *et al.*, 2010) (protein interactions and pathways). In practice, however, these resources are difficult to use for the average clinician and only present fragmented information, rather than an integrated picture, which is often needed by the clincal practitioner.

Biomedical researchers, wishing to engage in research in the field, are faced with similar problems. On the basis of literature searches, it is currently extremely difficult to even identify which primary immunodeficiencies exist, what their phenotypes are and to find and integrate all the relevant information across bioinformatics resources. For example, patients with suspected PIDs of unclear origin are often given the diagnosis CVID resulting in a large and diverse group of diseases bearing a common label. As such, the problems faced by clinicians and biomedical researchers significantly overlap. Computational support for both the clinician involved in the care of PID patients as well as the biomedical researcher is therefore highly desirable and necessary. Specifically, such support should lead to 'knowledge centralisation' in two distinct ways. First, knowledge centralization in the sense of data integration integrates PID-related data across the various resources mentioned above and presents a unified view to the end user. Knowledge centralization in the sense of the development of both a human-comprehensible as well as a machine-computable representation of information about PID can, for example, lead to the development of expert systems that utilize the represented information. Ontologies are both human-comprehensible and machine computable specifications of the knowledge in a domain of discourse and, as such, are well suited to play a 'centralising' role. Here, we present the Ontology of Primary Immunodeficiency Diseases (PIDO), describe how it interacts with other relevant ontologies and demonstrate its application in a clinical decision support system.

2 SYSTEMS AND METHODS

2.1 Ontological motivation

In order to bridge the genotype–phenotype gap and to develop successful computable representations of primary immunodeficiencies, a number of domain as well as granularity boundaries must be traversed and integrated. PIDs arise from complex interactions between gene products, pathways, tissues, organs and the interaction with the environment. The description and representation of these interactions should be computable and contain an adequate theory of biological, chemical and immunological functions and functionings. Consequently, an ontology of disease and phenotype should provide the means to define phenotypes based on the processes, objects and functions which give rise to a particular phenotype. For example, it should be possible to infer from a phenotype such as agammaglobulinemia that a patient does not have gamma-globulin *as part* and therefore that the biological function of gamma-globulin cannot be realized in this patient. To obtain these inferences and achieve interoperability with other relevant ontologies such as anatomy, process, phenotype and disease ontologies, we have adopted a method for formally defining phenotypes (Hoehndorf *et al.*, 2010b) for the development and axiomatization of PIDO.

In particular, our method of defining phenotypes enables basic interoperability with ontologies of anatomy and physiology, and we have included links to the FMA and the GO for this purpose. PIDO does not develop or use a classification of PIDs such as the one proposed by the International Union of Immunological Societies' Primary Immunodeficiency Disease Classification Committee (Geha *et al.*, 2007). However, combining PIDO's phenotype and disease definitions with anatomy ontologies can be used for the generation of a novel PID classification based on the axioms in an anatomy or physiology ontology. For example, we can create a class called 'Agammaglobulinemias' based on a combination of 'Having Primary Immunodeficiency Disease' and 'Having Agammaglobulinemia', and infer which particular PIDs satisfy the definition of 'Agammaglobulinemias'. The generation of PID classification based on complex class description enables a flexible and expressive access to PIDs and their associated phenotypes.

2.2 Biomarkers

PIDO characterizes PIDs based on both their phenotypes and associated biomarkers. We define a biomarker as a role (Loebe, 2005) that a phenotype plays within a clinical diagnosis process. The biomarker role is a role in an observation process:

```
Phenotype and (plays_role some BiomarkerRole)
BiomarkerRole subclassOf
  (role_of some Clinical_Diagnosis_Process)
```

By specifying the type of clinical diagnosis process, the type of biomarker may be further specified. An imaging biomarker, for example, is a biomarker that 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 provides an extensive hierarchy of biomarkers, which is useful for the further classification of phenotypes and any one phenotype will be able to assume multiple biomarker roles. In the first instance, the classification in terms of biomarkers will mirror the way in which most clinicians classify phenotypes. The formal integration of biomarkers and diagnostic processes in the PID ontology is the subject of future work.

Typical biomarkers are, for example, genomic biomarkers, cell functional biomarkers, laboratory findings, etc. Collectively, the phenotypes playing the role of biomarkers form the phenotype of a PID. All biomarkers were manually extracted from the recent primary clinical and research literature by domain experts. Incorporated biomarkers have been annotated with the PubMed identifier of the manuscript from which they were taken.

2.3 Relation to other ontologies

The ontology of PIDS draws on terms from many different domains such as genetics, anatomy, chemistry and proteins. The PID ontology only uses those terms which are necessary to achieve the desired expressivity and coverage and provides mappings to multiple established domain ontologies. The mappings are constructed in such a way as to map class names and definitions, but without importing axiomatizations.

The PID ontology's classes have been mapped to key resources in the biomedical domain to facilitate interoperability with other ontologies. In particular, classes describing anatomical parts have been mapped to the corresponding classes in the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) and the NCI Thesaurus (Sioutos *et al.*, 2007). Genes have been annotated with official gene symbols, alternative gene symbols, gene names and corresponding associated Mendelian Inheritance in Man (MIM) (Hamosh *et al.*, 2005) phenotypes, all of which were retrieved from Entrez Gene (Maglott *et al.*, 2007) as well as NCI Thesaurus terms. Phenotypes have been annotated with corresponding terms from the Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010). The manual annotation with corresponding phenotype terms derived from

Table 1. Ontologies and domains currently cross-referenced within PIDO

Ontology	Domain
FMA	Anatomy
NCI Thesaurus	General Medical Vocabulary
HPO	Human Phenotypes
Chemical Entities of Biological Interest	Chemical Entities
GO	Genes and Processes
PO	Proteins

the Mammalian Phenotype Ontology (Smith and Eppig, 2010) is currently ongoing, and annotations will be released in future versions of the ontology.

Furthermore, the mapping of terms referring to cellular components and biological processes, to proteins, cell types and qualities is still in progress, and mappings to the Gene Ontology (GO) (Ashburner *et al.*, 2000), the Protein Ontology (PO) (Natale *et al.*, 2011), the Cell Type Ontology (CL) (Bard *et al.*, 2005) and the Phenotypic Attribute and Trait Ontology (PATO) (Gkoutos *et al.*, 2005) will be created (Table 1). The mappings are constructed either manually or semi-automatically by using fuzzy string matches between ontologies. Fuzzy matches are determined using the Levenshtein (Navarro, 2001) or Needleman–Wunsch (Needleman and Wunsch, 1970) algorithms and validated by a human curator.

To demonstrate interoperability with other biomedical ontologies, in particular phenotype and anatomy ontologies, we use automated reasoning over phenotype and anatomy ontologies to derive representations of PID phenotypes in the phenotype ontologies for other species. Such mappings are of considerable use even for organisms, such as worms, for example, which are very far apart from humans: simple model organisms that can be used to develop comprehensive experimental analyses of the genetic and molecular makeup of complex phenotypic traits have been the staple of biomedical research for some time. For example, much work has been done in the past to study the immune system of nematodes as a model system for innate immunity (Schulenburg *et al.*, 2004). In particular, we used the PhenomeBLAST software (http://phenomeblast.googlecode.com) to generate mappings to the mouse, fly and worm phenotype ontologies, and we make these mappings available on our website.

PIDO uses the General Formal Ontology (GFO) (Herre *et al.*, 2006) as an upper ontology, because the GFO facilitates the integration of objects and processes, contains an expressive theory of relational and processual roles (Loebe, 2005) and provides expressive axioms in its OWL version. The taxonomic structure of the GFO is presented in Figure 1.

This strategy facilitates then reuse of existing resources and ontologies while at the same time avoiding reasoning problems that arise due to the large size of many reference ontologies. Furthermore, a single ontology that combines the necessary fragments of relevant biomedical domain ontologies is often easier to maintain and use (Bard and Rhee, 2004).

3 IMPLEMENTATION

The PID ontology has been formalized in the Web Ontology Language (OWL). The ontology has been edited using either the Protege 4 (Knublauch *et al.*, 2004) or TopBraid Composer (http://www.topbraidcomposer.com) OWL editors as well as the Manchester OWL API (Horridge and Bechhofer, 2009).

The PIDFinder Web Application was developed in the Java programming language using the Google Web Toolkit (GWT, http://code.google.com/webtoolkit/) and Google App Engine (GAE, http://code.google.com/appengine/) frameworks.

4 DISCUSSION

4.1 Canonical disease phenotypes

In the context of PIDO, we focus on the representation of the *canonical* phenotype of PIDs: a *canonical* disease phenotype consists of every observed phenotypic manifestation of the disease. We emphasize, however, that patients do not commonly exhibit all manifestations that are associated with a canonical disease phenotype. As a consequence, we intend to explore the relation between the observed phenotype in a patient and the canonical phenotype of a disease using a measure of semantic similarity that is able to account for incomplete and noisy information. Including all possible observations associated with a disease in the description of canonical disease phenotypes will enable us to increase the similarity, if patients and disease phenotype overlap in any of these phenotypes, and to decrease the similarity if they do not. We would then count patient and disease as phenotypically most similar when they fully overlap in their phenotypes. This method has already been applied successfully for the identification of novel gene–disease association in the PhenomeNET system (Hoehndorf *et al.*, 2011).

4.2 Use-case: WAS-related PIDs

WAS-related primary immunodeficiencies may arise as a consequence of defects in the *WAS* gene. Several recent studies have established good correlations between the type of gene defect and the corresponding human phenotype (Imai *et al.*, 2004; Jin *et al.*, 2004). Patients with insertions into the *WAS* gene or complete deletions of the gene have been found to develop full Wiskott-Aldrich Syndrome, whereas those with splice abnormalities develop either the full form or a milder version, depending on the exact nature of the splice abnormality. In PIDO, the following (clinical) phenotypes are associated with Wiskott Aldrich Syndrome presence of anti-DNA antibody, inflammation of the joints and colon, autoimmune hemolytic anemia, autoimmune thrombocytopenia, B-cell and hematopoietic cell neoplasms, bacterial, viral and fungal infections, bloody diarrhea, CD8$^+$ T-cell dependent cytotoxicity defect, eczema, immune-complex glomerulonephritis, myelodysplastic syndrome, petechia and small platelet size.

We use the Wiskott Aldrich Syndrome as an example to show how this complex phenotype can be modeled in terms of more basic phenotypes.

One commonly occurring phenotype in WAS patients is an elevated level of anti-DNA antibodies present in the serum. We formalize this as a phenotype of patients that have anti-DNA antibodies as part which are present in an increased concentration:

```
Phenotype and phenotype_of some
  (has_part some
    (Anti-DNA_Antibody and (has_property
      some Increased_Concentration)))
```

This enables the inference that entities with this phenotype have *Anti-DNA Antibodies* as part and therefore enables interoperability with anatomy ontologies that also use the *has_part* and *part_of* relations (Hoehndorf *et al.*, 2010b). Based on this assertion, further inferences across ontologies become possible. For example, anti-DNA antibodies realize certain functions under a given set of conditions. If the inference of presence or absence of a part is subsequently combined with an ontology specifying, for example,

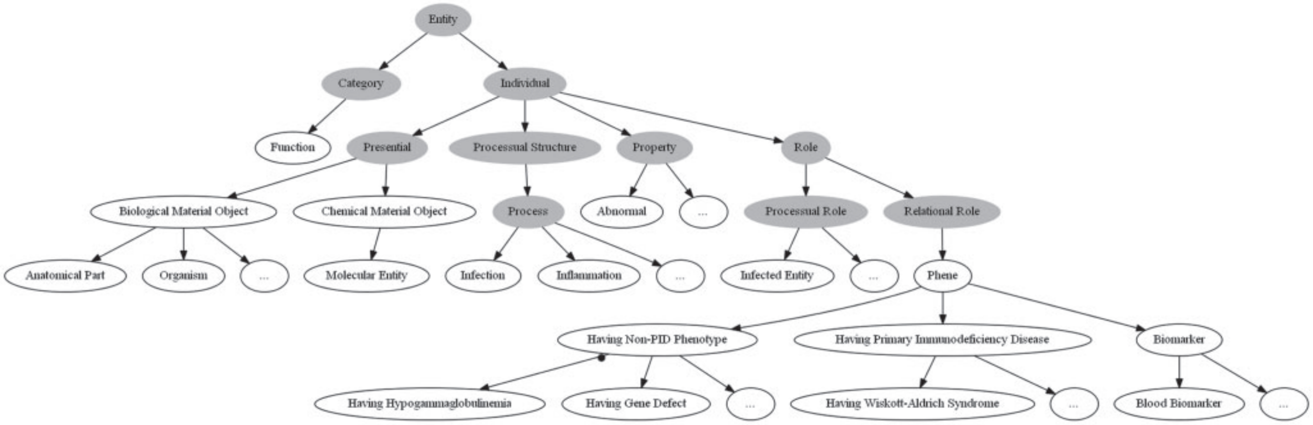


Fig. 1. Simplified schematic of the GFO and top-level classes of PIDO. The GFO is an ontology both of categories and individuals. Individuals are divided into abstract individuals (such as numbers), concrete individuals (such as processes and material objects) and spatio-temporal individuals (regions of space or time). Processes are subclasses of concrete individuals and phenotypes are considered to be relational roles. GFO classes are shown in gray.

the function of the part, then inference over presence or absence of function—although not currently implemented in PIDO—also now becomes a possibility (Hoehndorf *et al.*, 2010a).

Another phenotype often associated with Wiskott–Aldrich patients are frequently recurring bacterial, viral and fungal infections. Infections are processes in the GFO. Within GFO, processes can be characterized by the modes in which entities participate in a process, and these modes of participation are considered to be the processual roles of a process (Herre *et al.*, 2006; Loebe, 2005). For the process of infection, at least two such processual roles may be identified: the role of *Infectious Agent* and the role of *Infected Entity*. Formally, we formulate the following axioms:

```
Infection SubClassOf: gfo:Process and
  (has_role some Infectious_Agent) and
  (has_role some Infected_Entity)

Infectious_Agent SubClassOf:
  gfo:Processual_role
    and (role_of some gfo:Process)
```

Processual roles can be *played* by entities. Therefore, we can define a *Bacterial Infection* as an infection in which the role of the *Infectious agent* is played by a *Bacterium*:

```
Bacterial_Infection EquivalentTo:
  (Infection and (has_role some
    (Infectious_Agent and
      (played_by some Bacterium))))
```

Analogously, we are able to specify an infection by site rather than by causative agent. For example, we can define a *Gastrointestinal Tract Infection*:

```
Gastrointestinal_Tract_Infection EquivalentTo:
  (Infection and (has_role some
    (Infected_Entity and played_by
      some Gastrointestinal_Tract)))
```

This definition states that a *Gastrointestinal Tract Infection* is an *Infection* in which the role of the *Infected Entity* is played by the *Gastrointestinal Tract*. Based on these defined classes, we can then define *Bacterial Gastrointestinal Tract Infection* as follows:

```
Gastrointestinal_Tract_Infection EquivalentTo:
  (Infection and (has_role some
    (Infected_Entity and played_by
      some Gastrointestinal_Tract)) and
      (has_role some (Infectious_Agent and
        played_by some Bacterium)))
```

After defining the process *Gastrointestinal Tract Infection*, we are able to define the phenotype *Having Bacterial Infection*:

```
Having_Bacterial_Infection EquivalentTo:
  phenotype_of some (plays-role some
    (Infected_Entity and
       role_of some Bacterial_Infection))
```

This definition states that the phenotype *Having Bacterial Infection* is a phenotype of things that play the role of the *Infected entity* within a *Bacterial Infection* process.

All process-based phenotype definitions in PIDO follow the same pattern. As a future extension to these phenotype definition patterns, we could further define phenotypes such as ‘Having Frequent Recurring Infections’ by explicitly referring to an increase in the *rate* of occurrence of infections. However, while definitions following such a pattern are used in other phenotype ontologies and their inclusion will enable a basic form of interoperability with these ontologies, a full definition of the intended meaning of ‘frequently recurring’ is a substantial challenge for ontology representation languages that may be addressed in the future.

We can combine basic phenotypes to describe complex phenomena such as syndromes by asserting the class representing the complex phenomena as equivalent to or a subclass of all the phenotypes that characterize it. For example, the Omenn Syndrome is commonly associated with primary immunodeficiencies arising due to defects of the *RAG1* or *RAG2* genes (Santagata *et al.*, 2000). Patients presenting with the syndrome either have very large or very small lymph nodes, hepatosplenomegaly, lymphocytosis and alopecia. In PIDO, this can be formalized as:

```
Having_Omenn_Syndrome SubClassOf:
  Having_Alopecia and
  Having_Hepatosplenomegaly and
  Having_Lymphocytosis and
  (Having_Small_LymphNode or
    Having_Large_LymphNode)
```

Canonical representations of primary immunodeficiency diseases can be developed in an analogous manner by describing them as intersections of simpler phenotypes.

4.3 Description Logic-based querying

The phenotype formalism developed here allows the use of queries based on Description Logic. For example, when confronted with a patient suspected of suffering from a primary immunodeficiency and presenting with thrombocytopenia, a clinician might be interested in PIDs that are associated with this phenotype. PIDO can facilitate such queries. Currently, a reasoner will determine the subclasses of the class

```
Having_Primary_Immunodeficiency_Disease and
Having_Thrombocytopenia
```

to be *Common Variable Immunodeficiency* caused by *CD81 Gene Defect*, *Goods Syndrome*, *Hyper-IgM Syndrome Type 2*, *ORAI1 Defect*, *RAG1/RAG2 SCID Phenotype with Expansion of Gamma-Delta T-Cells*, *STIM1-Defect* and *WAS*. If the patient subsequently goes on to develop a B-cell lymphoma, the query can be expanded to

```
Having_Primary_Immunodeficiency_Disease and
Having_Thrombocytopenia and
Having_B_Cell_Lymphocytic_Neoplasm
```

In this case, the results narrow to the Wiskott-Aldrich Syndrome alone. Queries of this type can contribute to a diagnosis based on the phenotypes observed in a patient. Using the same query, a researcher might also determine, which genes are commonly associated with *Thrombocytopenia* in PIDO. Because of the particular axiomatization of phenotypes that includes possible gene defects, the genetic causes underlying a PID can be retrieved using querying in Description Logics, thereby leading to an integration across levels of granularity.

4.4 The PID Finder

A prime motivation for PIDO's development is to assist clinicians in gaining a rapid overview over existing relevant PID knowledge as well as to contribute to the diagnosis of PIDs. To demonstate PIDO's utility in clinical decision support, we developed the PIDFinder (Fig. 2). The PIDFinder is a prototype web application, which presents the information contained in the ontology in an easily accessible manner and allows, apart from access to PIDO's content, the phenotypic comparison of PIDs and the generation of diagnosis hypotheses. One central feature of PIDFinder is to allow a physician to specify a set of phenotypes that are observed in a patient with a suspected primary immunodeficiency disorder. The phenotypes are displayed in a faceted manner using the biomarker classification as facets. The faceted classification of phenotypes is automatically inferred based on axioms in PIDO (Fig. 2c).

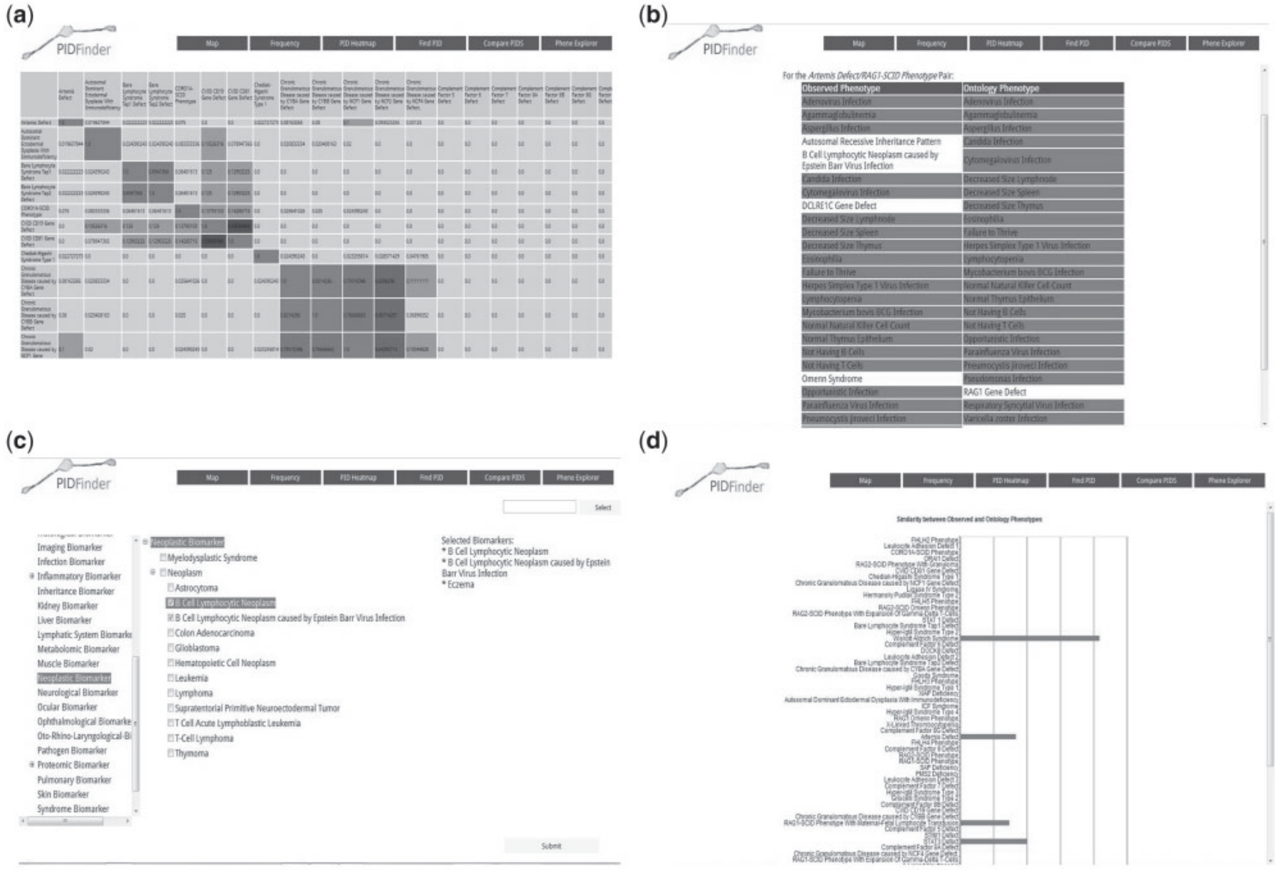
Due to the axioms in PIDO, phenotypes can appear in several facets: *Hepatomegaly*, for example, is inferred to be both an *Imaging Biomarker* as it can be observed in an imaging process, as well as a *Liver Biomarker* as it describes biological variation of the liver. The interface also offers a simple text-based suggest-box mechanism as an alternative search mode. Once phenotypes have been selected, they are collected into a set and form the *observed phenotype* of the patient. The patient phenotype set is subsequently compared with the *canonical* set of phenotypes characterizing each PID using the Tanimoto distance metric (Fig. 2d). The greater the calculated Tanimoto coefficient for an observed patient phenotype set compared with a PID phenotype set, the greater the phenotypic similarity between the observed phenotype of the patient and the canonical disease phenotype. The result of the comparison is a ranking of possible PIDs that are phenotypically similar to the patient phenotype. The results of the comparison are subsequently visualized and returned to the user either in a graph-based form or a side-by-side comparison of overlapping and non-overlapping phenotypes.

The Tanimoto distance is further used to compare phenotypic similarity between PIDs themselves. The result is visualized in the form of a heatmap (Fig. 2a) as well as for using a side-by-side comparison of overlap and non-overlap between PID phenotypes (Fig. 2b).

It should be noted that the PID Finder is only one of several use-cases for PIDO and as such does not leverage all the possibilities that the use of expressive and well-axiomatized ontologies offers. For example, we have not implemented the ability to run DL queries in the current version of the software and are also not currently using formal reasoning to arrive at diagnosis suggestions, but have rather opted for a semantic similarity approach using Tanimoto distances. However, this does not mean that the development of a well-axiomatized ontology is a wasted effort: the PIDFinder, for example, makes use of reasoning to generate the multifaceted presentation of phenotypes. The use of classifiers and formal descriptions also helps to deal with ambiguity in the specification of observed phenotypes: if, for example, a clinician only selects the 'Arthritis' phenotype, a classifier will infer that there are more specialized subforms of arthritis (e.g. polyarthritis, oligoarthritis, etc.) in the ontology and expand the selection made by the clinician in the PidFinder user interface to include these phenotypes in the selection unless a more defined subtype is subsequently chosen. Finally, implementing functionality allowing OWL DL Queries of the type discussed in the previous section in the PID Finder also remains a possibility.

4.5 Limitations and future research

Future research related to PIDO will pursue several different directions. First, the continued enrichment of the ontology with content and cross-references to other ontologies and terminologies will be of the highest priority. As such, we will continue to extract and critically evaluate phenotypic information from the primary research literature and incorporate it into the ontology. A second area of future extension is enriching the ontology with data from other biomedical databases and ontologies. In particular, we intended to fully incorporate the FMA and GO ontologies and utilize them for the classification of PIDs and their phenotypes. Third, we intend to work with the developers and maintainers of PID registries



For the *Artemis Defect/RAG1-SCID Phenotype* Pair:

Observed Phenotype	Ontology Phenotype
Adenovirus Infection	Adenovirus Infection
Agammaglobulinemia	Agammaglobulinemia
Aspergillus Infection	Aspergillus Infection
Autosomal Recessive Inheritance Pattern	Candida Infection
B Cell Lymphocytic Neoplasm caused by Epstein Barr Virus Infection	Cytomegalovirus Infection
Candida Infection	Decreased Size Lymphnode
Cytomegalovirus Infection	Decreased Size Spleen
DCLRE1C Gene Defect	Decreased Size Thymus
Decreased Size Lymphnode	Eosinophilia
Decreased Size Spleen	Failure to Thrive
Decreased Size Thymus	Herpes Simplex Type 1 Virus Infection
Eosinophilia	Lymphocytopenia
Failure to Thrive	Mycobacterium bovis BCG Infection
Herpes Simplex Type 1 Virus Infection	Normal Natural Killer Cell Count
Lymphocytopenia	Normal Thymus Epithelium
Mycobacterium bovis BCG Infection	Not Having B Cells
Normal Natural Killer Cell Count	Not Having T Cells
Normal Thymus Epithelium	Opportunistic Infection
Not Having B Cells	Parainfluenza Virus Infection
Not Having T Cells	Pneumocystis jiroveci Infection
Omenn Syndrome	Pseudomonas Infection
Opportunistic Infection	RAG1 Gene Defect
Parainfluenza Virus Infection	Respiratory Syncytial Virus Infection
Pneumocystis jiroveci Infection	Varicella zoster Infection

Imaging Biomarker
Infection Biomarker
Inflammatory Biomarker
Inheritance Biomarker
Kidney Biomarker
Liver Biomarker
Lymphatic System Biomarke
Metabolomic Biomarker
Muscle Biomarker
Neoplastic Biomarker
Neurological Biomarker
Ocular Biomarker
Ophthalmological Biomarke
Oto-Rhino-Laryngological-Bi
Pathogen Biomarker
Proteomic Biomarker
Pulmonary Biomarker
Skin Biomarker
Syndrome Biomarker

Neoplastic Biomarker
□ Myelodysplastic Syndrome
□ Neoplasm
□ Astrocytoma
☑ B Cell Lymphocytic Neoplasm
☑ B Cell Lymphocytic Neoplasm caused by Epstein Barr Virus Infection
□ Colon Adenocarcinoma
□ Glioblastoma
□ Hematopoietic Cell Neoplasm
□ Leukemia
□ Lymphoma
□ Supratentorial Primitive Neuroectodermal Tumor
□ T Cell Acute Lymphoblastic Leukemia
□ T-Cell Lymphoma
□ Thymoma

Selected Biomarkers:
* B Cell Lymphocytic Neoplasm
* B Cell Lymphocytic Neoplasm caused by Epstein Barr Virus Infection
* Eczema

Select

Submit

Fig. 2. Screenshots of the PIDFinder Web Application: (**a**) heatmap characterizing the phenotypic similarity of PIDs; (**b**) phenotype-by-phenotype comparison of PIDs; (**c**) selection of phenotypes observed in a patient and (**d**) PID diagnosis suggestion based on a set of observed phenotypes.

to both apply PIDO for the classification and analysis of PID-related information, and to extend PIDO with classes that are relevant within the established PID resources. Finally, we intend to further address the challenges involved in describing canonicity and non-canonicity in biomedical ontologies (Hoehndorf *et al.*, 2007, 2010b).

The appropriate definition of phenotypes such as *Having Small Platelets Size* (a phenotype for Wiskott-Aldrich Syndrome) or *Thrombocytopenia* remains an open issue. The former refers to the fact that the average of the platelet size distribution in a patient is shifted to lower values with respect to that which is considered normal, and thrombocytopenia denotes the situation in which the number of platelets in the blood of a patient is reduced with respect to the number that is considered normal.

4.6 Conclusions

We have developed PIDO, which characterizes PIDs by defining a semantically rich representation of the basic observable characteristics in organisms. The characteristics are based on descriptions of quality, function, structure or process and are interoperable with other biomedical domain ontologies. Based on PIDO, we have developed PIDFinder, a prototypical web-based clinical decision support system that enables access to the knowledge contained in the ontology and is capable of generating diagnosis hypotheses. PIDO connects immunological knowledge across resources within a common framework and thereby enables translational research and the development of medical applications for the domain of immunology and PIDs.

ACKNOWLEDGEMENT

We thank Dr Ulrich Baumann (Hannover Medical School) for the helpful discussions and support.

Funding: BBSRC Grant (BB/G022747/1 to N.A.); European Commission's 7th Framework Programme, RICORDO project, Grant (248502 to R.H.); BBSRC Grant (BB/G0043581 to G.G.); Hannover Medical School C4-Proferssorship to G.H. And, Deutsche Forschungsgemeinschaft (DFG) Grant (SFB587, TP14 to C.H.).

Conflict of Interest: none declared.

REFERENCES

Aranda,B. *et al.* (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bard,J.B.L. and Rhee,S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.*, **5**, 213–222.

Bard,J. *et al.* (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Boyle,J.M. and Buckley,R.H. (2007) Population prevalence of diagnosed primary immunodeficiency diseases in the United States. *J. Clin. Immunol.*, **27**, 497–502.

Consortium,U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

Geha,R. *et al.* (2007) Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *J. Allergy Clin. Immunol.*, **120**, 776–794.

Gkoutos,G.V. *et al.* (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.

Guzman,D. *et al.* (2007) The ESID Online Database network. *Bioinformatics* , **23**, 654–655.

Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Herre,H. *et al.* (2006) General Formal Ontology (GFO) a foundational ontology integrating objects and processes. *Technical Onto-Med Report No. 8*, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig.

Hoehndorf,R. *et al.* (2007) Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics*, **8**, 377.

Hoehndorf,R. *et al.* (2010a) Applying the functional abnormality ontology pattern to anatomical functions. *J. Biomed. Semant.*, **1**, 4.

Hoehndorf,R. *et al.* (2010b) Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, **26**, 3112–3118.

Hoehndorf,R. *et al.* (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.*, **39**, e119.

Horridge,M. and Bechhofer,S. (2009) The OWL API: a Java API for working with OWL 2 ontologies. In *OWLED 2009, 6th OWL Experienced and Directions Workshop*. Chantilly, Virginia.

Imai,K. *et al.* (2004) Clinical course of patients with WASP gene mutations. *Blood*, **103**, 456–464.

Jin,Y. *et al.* (2004) Mutations of the Wiskott-Aldrich Syndrome Protein (WASP): hotspots, effect on transcription, and translation and phenotype/genotype correlation. *Blood*, **104**, 4010–4019.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Keerthikumar,S. *et al.* (2009) Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA Res.*, **16**, 345–351.

Knublauch,H. *et al.* (2004) The protege owl plugin : an open development environment for semantic web applications. *Design*, **3298**, 229–243.

Loebe,F. (2005) Abstract vs social roles: a refined top-level ontological analysis. In Boella,G. *et al.* (eds) *Proceedings of the 2005 AAAI Fall Symposim 'Roles, an Interdisciplinary Perspective: Ontologies, Languages and Multiagent Systems'.* AAAI Press, Menlo Park, California, USA.

Maglott,D. *et al.* (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.

Natale,D.A. *et al.* (2011) The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.*, **39**, D539–D545.

Navarro,G. (2001) A guided tour to approximate string matching. *ACM Comput. Surv.*, **33**, 31–88.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Parkinson,H. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

Riminton,D.S. and Limaye,S. (2004) Primary immunodeficiency diseases in adulthood. *Intern. Med. J.*, **34**, 348–354.

Robinson,P.N. and Mundlos,S. (2010) The human phenotype ontology. *Clin. Genet.*, **77**, 525–534.

Rosse,C. and Mejino,J.L.V. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J. Biomed. Informat.*, **36**, 478–500.

Samarghitean,C. and Vihinen,M. (2009) Bioinformatics services related to diagnosis of primary immunodeficiencies. *Curr. Opin. Allergy Clin. Immunol.*, **9**, 531–536.

Santagata,S. *et al.* (2000) The genetic and biochemical basis of Omenn syndrome. *Immunol. Rev.*, **178**, 64–74.

Schulenburg,H. *et al.* (2004) Evolution of the innate immune system: the worm perspective. *Immunol. Rev.*, **198**, 36–58.

Sioutos,N. *et al.* (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Informat.*, **40**, 30–43.

Smith,C.L. and Eppig,J.T. (2010) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisc. Rev. Syst. Biol. Med.*, **1**, 390–399.

Urschel,S. *et al.* (2009) Common variable immunodeficiency disorders in children: delayed diagnosis despite typical clinical presentation. *J. Pediatr.*, **154**, 888–894.

Väliaho,J. *et al.* (2005) BMC Medical Informatics and Distribution of immunodeficiency fact files with XML âŁ" from Web to WAP. *BMC Med. Informat. Decis. Mak.*, **11**, 1–11.