A key challenge in bioinformatics today is ensuring that biological data can be unequivocally communicated between experimentalists and bioinformaticians. Enabling such communication is not trivial, as every scientific field develops its own jargon with implicit understandings that can easily escape an outsider. We describe here our approach to enforce an explicit and exact data representation for the Immune Epitope Database (IEDB Peters et al. 2005) through the use of a formal ontology.

Being the first database of its scale in the immune epitope domain, it was necessary for the IEDB to devise an adequate data structure at the outset of the project with the goal that it should be capable of capturing the context of immune recognition. Early on, it became readily apparent that an unambiguous description of the information being captured is imperative for consistent curation across journal articles and among curators. Accordingly, an initial ontology was developed (Sathiamurthy et al. 2005) based upon consultations with domain experts and guidance from expert ontologists. The structure devised from this ontology proved capable of dealing with a great deal of immunological data over time.

However, after several years of curation, it became necessary to adjust the ontology and data structure to accommodate more and more exceptions. Moreover, the database had, from the beginning, excluded particular types of experiments as they were not easily accommodated in its structure. These experiments, known as adoptive transfer or passive immunization, are complex in nature, involving one organism in which the immune response is generated (donor organism) and another to which that immune response is transferred and studied (recipient organism). In order to accommodate adoptive transfer experiments, as well as improve how all experimental approaches were represented, the IEDB has recently developed a new version of its ontology (ONTology of Immune Epitopes, ONTIE 2.0) and simultaneously restructured its database scheme.

This book chapter gives an introduction into the biology of immune epitope recognition followed by a brief primer on ontology development in general, and
introduces the higher level ontologies we are building upon. We then present results from our specific approach to develop ONTIE 2.0, and the global reorganization of the IEDB data structure.

**Background: Immune Epitopes and the IEDB**

The vertebrate immune system has the capacity to detect nonself molecules through adaptive immune receptors. These receptors include the B cell receptors (BCRs), present on the surface of B cells and secreted as antibodies, and the T cell receptors (TCRs). The molecular entities recognized by BCRs and TCRs are referred to as immune epitopes. Epitopes found in proteins consist of either short linear stretches of amino acids (i.e., peptides) or are conformational epitopes, formed by the spatial arrangement of several amino acid residues within the three-dimensional protein structure. Epitopes can also be derived from other molecules, including carbohydrates and lipids.

BCRs and antibodies bind directly to their targets, tagging them for further action by immune effector cells, and sometimes interfering directly with the antigen’s function. TCRs recognize their target epitopes as part of a complex with a major histocompatibility complex (MHC) molecule. MHC molecules are found on the surface of antigen presenting cells and display a sample of peptides derived from digested proteins present inside the cell. After infection with a foreign pathogen, some MHC molecules will present peptides derived from nonself proteins. These peptide-MHC complexes are scanned by the TCR, allowing the T cell to probe the contents of cells and take action when nonself proteins are encountered.

The IEDB was created with the aim to gather all published experimental data characterizing immune epitopes. Details of the experimental context in which the epitopes are defined are captured along with the epitope structure itself. Capturing the experimental context is imperative, since being an epitope is not an intrinsic property of a molecular entity. In other words, to accurately describe an epitope requires not just detailing the molecular structure alone, but it must include a description of the context in which that molecular structure is recognized as an epitope by the immune receptors.

The data required to capture the experimental context are quite complex. The experiments performed to define epitopes are varied and comprise multiple parts and steps. For example, the first step of the experiment shown in Fig. 3 will require capturing:

1. The host organism in which the immune response was studied (C57BL/6 mouse);
2. The immunization processes that primed the immune response (subcutaneous injection); and
3. The immunogen used (SARS coronavirus nucleoprotein).

By providing the details of the experimental context, comparisons across different studies can be made and discrepancies between experiments can be better understood.
Additionally, refined retrieval of specific subsets of data (e.g., epitopes defined in humans, or those recognized by CD4+ T cells) is also made possible.

To capture the experimental context in which epitopes are defined, the IEDB utilizes over 300 database fields to describe an individual experiment. The database is primarily populated from experiments described in articles published in peer-reviewed, PubMed-indexed journals. The translation of the published data into the database fields is performed by a team of nine Ph.D.-level curators following a formalized curation process, with several quality control steps and mechanisms to enforce and adapt curation rules (Salimi and Vita 2006; Vita et al. 2006). Currently, more than 5,000 articles have been curated for a total of greater than 115,000 assays involving approximately 38,500 epitopes. The IEDB ontology provides the formal framework to enforce consistency of curation for this vast amount of data.

**Background: Why Ontology Development?**

Ontology is a fundamental branch of philosophy that goes back to the early Greek philosophers Aristotle and Plato and deals with the question of “what things exist?” This question is answered by listing individual entities, types of entities, and relationships among them.

The use of ontologies to annotate data in biological databases has become ubiquitous in recent years (Ashburner et al. 2000; Rosse and Mejino 2003; Whetzel et al. 2006a, b). This is a result of their increased application in the information sciences in general. Information science has revolutionized the ability to retrieve specific sets of records or to derive summary statistics on data by providing the capacity to store large sets of information in databases. To be useful, the meaning of the stored information has to be clear to all users. Agreement between the users of the database and the data providers regarding how each entry in each field is interpreted can be achieved by an unambiguous mapping to a formal ontology.

A formal ontology represents the most complex of three approaches commonly used in information science to enforce data consistency:

- **Controlled vocabulary** – a list of terms and their definitions
- **Taxonomy** – the terms of a controlled vocabulary are organized into a hierarchical structure, typically an is a hierarchy (parent – child)
- **Formal ontology** – the is a hierarchy of a taxonomy is expanded to include multiple types of relationships between terms, such as part of and source of

Fig. 1 illustrates how several terms from an immunization protocol (Fig. 3) would be represented in each of these approaches. A controlled vocabulary (Fig. 1a) would arrange the terms as a list with each term having a specific definition, but cannot accommodate explicit relationships. Under a taxonomy (Fig. 1b), the terms are also explicitly defined and have exactly one is a relationship to their parent term. The fact that nucleoprotein is a protein and a protein is a object allows the transitive conclusion to be drawn that nucleoprotein is a object. Although this relationship
may be obvious, the ability to express it in formal terms is a step towards automating the discovery of new relationships.

A formal ontology (Fig. 1c) expands upon the concept of a taxonomy in that it allows for multiple types of relationships between and among objects. Such additional relationships include part of, derived from, plays role and proxy for, and can
be used to formalize the textual definitions of terms. For example, in the formal ontology depicted in Fig. 1c, the protein immunogen has two relationships; it is a protein and it plays role immunogen.

**Standing on the shoulders of BFO, OBO, and OBI**

One of the benefits of using a formal ontology for data annotation is the potential for re-use of existing ontologies upon which the domain-specific solution can be built. Although this could significantly reduce the amount of work to build the domain-specific ontology, the primary advantage of this approach is the enhanced interoperability between different resources that it enables.

For the 2.0 version of ONTIE, we have integrated our development work into the Ontology for Biomedical Investigations (OBI) effort (Whetzel et al. 2006a), in which we are actively participating. The scope of OBI is to provide the vocabulary necessary to describe any biomedical experiment, instrument, reagent and the like. By integrating the experimental terms needed for ONTIE into OBI, we are ensuring consistency and interoperability with the nearly 20 scientific communities included in the OBI effort.

OBI itself is a candidate for date for the open biomedical ontologies (OBO) foundry (Smith et al. 2007). While OBO is an umbrella group open to all ontology developers adhering to some format and availability requirements, the OBO foundry represents an effort to develop high quality, interoperable, and orthogonal ontologies. In short, many domain-specific ontologies are under development by the respective experts in their fields. The OBO foundry aims to enforce a set of standards to ensure that these ontologies will be interoperable and to avoid unnecessary duplication of work. Other candidate ontologies for the OBO foundry such as the Gene Ontology, Cell Ontology or Disease Ontology deal with the vocabulary necessary to describe different aspects of biological reality. This allows OBI to reference these other foundry ontologies to describe things like the cell types contained in a sample used in an experiment.

OBI is designed with the Basic Formal Ontology (BFO) (Grenon et al. 2004) as its upper level. BFO intends to be a high-level ontology from which more focused domain-specific ontologies can be developed. At its most basic level, BFO divides the world into continuants and occurrents. Continuants are entities that persist through time and are further divided into independent and dependent continuants. Independent continuants are defined as bearers of qualities and include OBI material entities (e.g., instruments, organisms, cells, proteins). Dependent continuants are entities that are intrinsic to or borne by other entities. These include things like qualities (e.g., mass, shape), and roles (drug, placebo) that require an independent continuant to exist. Occurrents make up the second major branch of the BFO hierarchy and are defined as entities that unfold over time. OBI further divides these into planned (e.g., experiments) and unplanned (e.g., diseases) processes. A partial schema of the hierarchy is depicted in Fig. 2.
In the next section, we describe how the integration into this high level structure has effected the development of ONTIE. It is important to point out that OBI and, to a lesser degree, BFO are still under active development and their topology may change. We are actively participating in this development and plan to keep ONTIE in sync with these ontologies.

## Ontology Development for the IEDB

### Objects, Roles, and Processes

The initial ontology for the IEDB (Sathiamurthy et al. 2005) was developed prior to population of the database, and was based upon consultations with domain experts and guidance from expert ontologists. It proved capable of dealing with the vast majority of epitope data encountered in the literature. However, it proved to be hard to extend to deal with novel types of data, contained internal inconsistencies and ambiguities that only became apparent over time, and provided only limited interoperability with other data resources. These experiences gained through several years of curation led to the decision to develop a revised version of the ontology, and integrate it into OBI.

ONTIE 2.0 development required mapping the existing terms used in the IEDB to the BFO/OBI structure. Fig. 3 depicts a representative experiment to characterize a
T cell epitope. The terms captured in the IEDB to describe such experiments include the immunization procedure, ELISPOT assay, host, immunogen, antigen, antigen-presenting cells and effector cells. The first two of these correspond to different steps of the experiment, and are clearly processes as they occur over time. The others are objects playing specific roles. For example, the same type of protein can be the immunogen in the immunization protocol and the antigen in the ELISPOT assay. Similarly, the same type of cell type can play the role of an effector cell and antigen presenting cell. In other words, these are structurally the same objects playing different roles in different processes. Examples for distinct types of objects, on the other hand, are organisms, cells, or proteins. This is the primary breakdown of experimental terms in OBI: Experimental steps correspond to processes of which each is associated with certain roles or functions that can be played by certain types of objects.

This restructuring of the ontology was extended to the database schema. In the previous schema, each object-role combination was stored in a separate table, resulting in a great deal of redundancy. In the new schema, there is one table where all object-related information is stored and it can be referenced by other tables that associate objects with roles in a process. For example, a particular type of peptide can play the role of epitope, the role of immunogen and the role of antigen. Before restructuring, the same exact peptide would have been stored in the epitope table, the immunogen table and the antigen table, but is now only stored once in the object table and is referenced by its different roles. In addition to reducing redundancy, this makes it easier to extend the database when new object types are encountered.

Although the IEDB aims for a complete representation of the data, not all steps of an experiment are captured explicitly. In the T cell epitope identification experiment in Fig. 3, the IEDB will capture information about the immunization and ELISPOT assay protocols. Although other steps exist, these are not as relevant for an accurate representation of the epitope-related information derived from the experiment. Several steps are therefore only implicitly captured. For example, the use of mouse lymph node cells implies that the lymph nodes were removed and apportioned in some fashion. The information about those tissue harvesting and cell isolation protocols is not captured by the IEDB, but their occurrence is implied.

**Relationships**

In the initial IEDB ontology, the only connection between terms was the unspecific “has” relationship, e.g., immunization has a host. In ONTIE 2.0, the more specific BFO/OBI relations are used instead. For example, an organism participates in an immunization, and plays role host. Between objects, the most important relations are has part and derived from. For example, a T cell has part TCR or a cell extract is derived from cells.
In addition to the previously defined BFO relationships, we needed to introduce new ones to model all the terms encountered in the IEDB. A linear peptidic epitope is not necessarily part of or derived from a source protein molecule (Fig. 4). In many cases, the peptides used in an experiment are artificially synthesized and were therefore never part of any protein molecule. In these cases, the peptides have a special relationship termed “proxy for.” A proxy for relationship indicates that the material is used in an experiment, observations are made, and conclusions are drawn on the material or class of materials that it is believed to represent. For example, the peptide in Fig. 4 is a proxy for a region of the SARS coronavirus nucleoprotein in the experiments in Fig. 3. In this experiment, observations are made on the peptide with the assumption that these observations can be generalized to any peptide with the same sequence.

Fig. 3 A typical T cell epitope identification experiment captured by the IEDB. All protocols are described as in the IEDB. This experiment starts with an Immunization protocol with a naïve (in the immunological sense) mouse (playing the role of host) and SARS coronavirus nucleoprotein (playing the role of immunogen) as an input and an immunized mouse as an output. The immunized mouse is input for an organ harvesting protocol that produces mouse lymph nodes as output. Lymph cells are next generated as output from a cell isolation protocol. This yields a mixture of T cells and potential antigen presenting cells (APC) among other cell types. These protocols are labeled N/A for IEDB as they are not explicitly captured. In the next protocol (ELISPOT assay) a peptide that can be derived from the protein administered in the first protocol, is administered as the antigen, to the isolated cell mixture. APCs present bound peptide on their surface via MHC molecules. T cells specific for this peptide will bind to the peptide-MHC complex and become activated. This results in the secretion of interferon-γ (IFN-γ). Next, labeled antibodies specific for IFN-γ are added and bind to IFN-γ bound to the plate via a capture antibody (not depicted) in the vicinity of these activated T cells. In the next step – Spot counting, the plate is scanned by an instrument and spots are counted where labeled antibodies have bound. The final step – Data acquisition – is the transformation of the raw count to a frequency measurement followed by encoding into a digital format.
These proxy for relations form a chain that can be traversed all the way back to the immune response. That is, the spot count is proxy for the SFC, which is proxy for the activation of T cells, which is proxy for the frequency of epitope specific T cells induced by the immunization protocol. The intermediate steps can be subsumed so that the count of spots is ultimately proxy for the frequency of epitope-specific T cells induced by the immunization protocol.

**Summary and Conclusions**

A vast amount of epitope information produced by experimental researchers is available in the scientific literature. Aside from the physical separation of this information in different publications, these experiments employ a wide variety of techniques, obtain different types of measurements, and are reported in divergent formats. Capturing this data in a consistent fashion enhances its utility and is what the IEDB team strives towards. This requires capturing the information in a natural, coherent, and unambiguous data structure. Developing this structure requires expertise from, and communication between, domain-expert curators and data modelers.

By applying the knowledge gained from the curation of experimental contexts that did not easily fit into our previous schema and utilizing concepts drawn from OBO, OBI, and BFO, we developed an ontology of immune epitopes for general use. Its development in accordance with OBI ensures its compatibility with the
plethora of other ontologies developed under the OBO foundry. Simultaneously, we redesigned the IEDB data structure to follow the principles of the ontology and house a representation of the data that will be easier to maintain and extend. The finalized new version of the Immune Epitope Database based on this work, IEDB 2.0, is accessible to the public since 2009.

References

Ashburner M, Ball CA et al (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
Grenon P, Smith B et al (2004) Biodynamic ontology: Applying BFO in the biomedical domain. Stud Health Technol Inform 102:20–38
Peters B, Sidney J et al (2005) The immune epitope database and analysis resource: From vision to blueprint. PLoS Biol 3(3):e91
Rosse C, Mejino JL Jr (2003) A reference ontology for biomedical informatics: The Foundational Model of Anatomy. J Biomed Inform 36(6):478–500
Salimi N, Vita R (2006) The biocurator: Connecting and enhancing scientific data. PLoS Comput Biol 2(10):e125
Sathiamurthy M, Peters B et al (2005) An ontology for immune epitopes: Application to the design of a broad scope database of immune reactivities. Immunome Res 1(1):2
Smith B, Ashburner M et al (2007) The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251–1255
Vita R, Vaughan K et al (2006) Curation of complex, context-dependent immunological data. BMC Bioinform 7:341
Whetzel PL, Brinkman RR et al (2006a) Development of FuGO: An ontology for functional genomics investigations. Omics 10(2):199–204
Whetzel PL, Parkinson H et al (2006b) The MGED Ontology: A resource for semantics-based description of microarray experiments. Bioinformatics 22(7):866–873