An Ontology for Major Histocompatibility Complex (MHC) Alleles and Molecules

Elena Beisswanger1  David S. DeLuca2  Rainer Blasczyk2  Udo Hahn1

1Jena University Language and Information Engineering (JULIE) Lab (http://www.julielab.de)
2Department of Transfusion Medicine, Hannover Medical School (http://www.mh-hannover.de/217.html)

Abstract
We present a formally coherent and consistent multi-species MHC ontology which includes all human MHC alleles and serological groups. The ontology is part of StemNet, a knowledge management system for hematopoietic stem cell transplantation with an integrated semantic search engine. The OWL-encoded MHC ontology contributes to the system in a threefold manner. First, it supports query formulation and query processing as well as mapping onto external terminological resources, second, it eases the interaction with the search engine when navigating through search results, and finally, it provides a formal language for text annotation, a methodological prerequisite for state-of-the-art natural language text processors which are increasingly based on machine learning methods and hence require annotated text corpora.

Introduction
The long tradition of on-going activities related to the development of numerous shared terminologies in the life sciences are unmatched in most other scientific fields. More than hundred terminologies are currently incorporated in the Unified Medical Language System (UMLS)1, while more than fifty terminological resources reside within the Open Biomedical Ontologies (OBO)2 framework. In terms of design principles (reflecting different purposes for usage), however, most of the concept systems being developed for medicine and biology differ markedly. While thesauri and classifications in medicine (as the MESH3 or ICD-10 [6]) were mostly designed for bibliographic search or billing and accountancy matters, (some more recently introduced resources such as the FMA4 [11] or even the SNOMED CT5 are commendable exception since they provide taxonomically ordered terms with explicitly defined semantics), terminologies for biologists reflect a deep concern with biological databases, i.e., the access to and curation of factual data. The latter type of functionality requires biological concept systems not only to be much more fine-grained and detailed than those for medical use but also to maintain a clean semantics of the terms being used. The enforcement of semantic rigidity and formality has become a community-wide shared methodological goal imposed on the design and (re-)engineering of biological terminologies. The corresponding commitments are reflected in the creation of the OBO Foundry6 initiative, which aims at transforming informal nomenclatures, thesauri and classification systems into formal ontologies [2, 3] based on a semantically strict Relation Ontology (RO) [13].

Besides representing a controlled vocabulary for the domain, a formal ontology has several advantages over a purely lexical or thesaurus-like format for terminologies. First, an ontology is an effective means to represent domain knowledge in a coherent and comprehensive way. Grouping data base entries into explicitly defined classes provides a clean interface between terminological and conceptual knowledge. Second, due to its formal rigor it allows for automated reasoning about the domain classes (e.g., by means of description classifiers). Finally, its clearly defined structure makes it an intuitive means to communicate the information related to the application domain to the users of the evolving knowledge base and integrated search engine. In particular, users can browse and navigate through search results more easily (e.g., they

1http://www.nlm.nih.gov/research/umls/
2http://obo.sourceforge.net
3http://www.nlm.nih.gov/mesh/
4http://sig.biostr.washington.edu/projects/fm/
5http://www.snomed.org/
6http://obofoundry.org
may exploit the hierarchical structure of the ontology for zooming, extending, and refining the search results) and thus end up with, hopefully, better and more appropriate results. Apart from the front end perspective (user support) such ontologies will also be used at the back end to increase the quality and performance of natural language text processors. Entities and relations such systems recognize in the documents have their direct correlate at the ontology level in terms of classes and conceptual relations. Hence, information extraction and text mining are also in need of properly defined concept systems for annotation (training data for machine learners) and subsequent semantic search in the knowledge base generated by such an NLP system [4].

We present here an ontology covering major histocompatibility complex (MHC) genes and alleles, MHC molecules and related concepts in terms of a hierarchy of classes and their conceptual interrelations. The emerging ontology is intended to become the domain knowledge backbone of StemNet, a knowledge management system for hematopoietic stem cell transplantation (HSCT). It is intended to increase security and quality of clinical treatment based on enhanced, web-based information supply. The core of the knowledge management system will be a high quality literature search engine dedicated to advancing the localization and exchange of information on HSCT.

**Ontology Design**

The biomedical background of the MHC ontology is anchored in the allogeneic HSCT. This has become an accepted therapeutic method to treat patients with leukemia and other hematological malignancies. The goal of the HSCT is to eliminate the patient’s cancer cells. First, the patient undergoes total body irradiation and high-dose chemotherapy to eradicate as many leukemia cells as possible. As a side effect, the patient’s hematopoietic system is also eradicated. In this situation, the HSCT has both the function of replacing the hematopoietic and immune system, as well as providing allogeneic T cells, which help to eliminate residual leukemia cells (an immunotherapeutic effect). This is called the graft-versus-tumor (GVT) effect. Stem cells can be extracted from cord blood, or from the bone marrow or peripheral blood of the donor. The key to successful allogeneic transplantation is finding a genetically compatible donor because it decreases the risk of graft rejection and graft versus host disease (GVHD). Both depend on how well certain cell-surface antigen-presenting proteins match between donor and recipient. These proteins are called human leukocyte antigens (HLAs). The HLA encoding genes form the major part of the MHC. The MHC is a family of genes involved in immune defense and autoimmunity. MHC genes are classified into class-I, class-II, and class-III genes. HLA genes are human MHC class-I and class-II genes. Their gene products present antigens on the cell surface, which are recognized by T cells. Variations in these genes determine the compatibility of donor and recipient, and affect the risk of graft rejection and GVHD. MHC genes are some of the most genetically variable regions in the mammalian genomes, i.e., for most HLA genes there exist many different variants, the HLA alleles. In order to find a compatible donor for HSCT, potential donors are tested with respect to their HLA types at the allelic level.

In many cases when patients need a hematopoietic stem cell transplant no HLA allele matched donor can be identified. However, an increasing number of patients are successfully transplanted even with a mismatched donor. Thus, the question of which kinds of HLA mismatches are permissive, and which have a negative effect, is the subject of research all over the world. The StemNet system is supposed to assemble evidence from publications that may be of help for clinical decision making in the cases of planning concrete stem cell transplantations on a larger and more systematic scale than the current workflow permits – extensive database search, quantitative scoring of most likely donor-recipient pairs, qualitative, mostly Google- and PubMed-based literature search for relevant documents (all these activities lack systematic support for interoperability).

It is this kind of fine-granular biomedical knowledge which we want to capture in the MHC ontology under development. The current version of the ontology comprises more than 3,000 classes and uses five fundamental semantic relation types (excluding the taxonomic is-a relation, and excluding inverse/reciprocal relations), both dealt with in more depth in the subsequent subsections. It should clearly be pointed out that the ontology exclusively holds classes and relations linking classes, though no instances of classes, since only

---

7http://www.stemnet.de/
knowledge about classes is ‘generalizable knowledge’ valid for inclusion in an ontology (instance knowledge is, e.g., the data of particular donors or cancer patients usually kept in clinical databases) [3].

The ontology is implemented in OWL DL, i.e., it consists of classes, subclasses, object properties (relations), and property restrictions (class definitions) [5]. It has been developed using Protége, an open source ontology editor and knowledge base framework, written in Java.

To guarantee interoperability with already existing resources and also to support the OBO Foundry policy, the MHC ontology is created in a way compliant to the OBO criteria and principles: The content of the ontology is clearly specified and delineated and does not overlap with already existing resources in the OBO library. It comes with textual definitions of classes and also formally described classes, uses unambiguously defined relations importing the OBO Relation Ontology and extends the RO slightly (the newly proposed relations are currently under negotiation with RO representatives). Once the consent of the OBO Foundry is achieved, the MHC ontology will be made publicly available.

**Ontology Classes**

The MHC ontology consists of two major parts, one dealing with classes on the nucleic acid level, the other dealing with classes on the protein level. The ontology branch which is dedicated to nucleic acids covers MHC genes and their alleles. The root class is nucleotide sequence, with subclass MHC gene subordinated to it. MHC gene has the subclasses MHC class-I gene, MHC class-II gene, and MHC class-III gene. MHC class-I gene has in turn the subclasses MHC class-Ia gene (also called classical MHC class-I gene), and MHC class-Ib gene (also called non-classical MHC class-I gene). As already pointed out, MHC class-I and class-II genes are relevant to hematopoietic stem cell transplantation. To group these classes MHC gene encoding antigen presenting protein was introduced as their parent class and as direct subclass of MHC gene.

The ontology is multi-species, by design. The current version encompasses MHC genes from human, mouse and dog, subsumed by the classes human MHC gene, mouse MHC gene, and dog MHC gene (which themselves are subclasses of MHC gene). Class names are based on established organism specific nomenclatures (nomenclature for factors of the HLA system [10], nomenclature for factors of the DLA system [7], mouse MHC H2 loci from the IMGT database).

Generally, each gene has several variants, so-called alleles. In addition to genes, the MHC ontology also covers alleles of genes. The HLA-A allele, e.g., represents the alleles of the HLA-A gene. In many cases differences on the nucleic acid level can be detected on the protein level by serological recognition. For HLA-A molecules, e.g., HLA-A1, HLA-A2, HLA-A3, and HLA-A9 are some of the serological groups that have been identified. In the ontology, alleles that encode, e.g., HLA-A1 are represented by the class HLA-A1 allele. Alleles are named by the order of their discovery. The first allele found to encode an HLA-A1 protein, e.g., is named HLA-A*0101 allele, the second HLA-A*0102 allele, etc. Alleles with synonym mutations are distinguished on the identifier level by two additional digits (e.g., HLA-A*0101 allele and HLA-A*0102 allele). Digit seven and eight distinguish between alleles that only differ in intron regions (e.g., HLA-A*010101 allele and HLA-A*010102 allele).

Alleles of non-functional genes are marked as allele of pseudogene, a subclass of nucleotide sequence. Alleles of MHC genes that do not encode functional proteins are subclassed by null allele. These additional parent classes allow to distinguish between (functional) protein encoding alleles and non-protein encoding ones.

The second ontology branch deals with proteins encoded in MHC genes. With protein being the root class, it has the subclass MHC protein subsuming proteins encoded in MHC genes. MHC proteins are further classified into class I, class II, class III proteins and also comprise the subunits of proteins encoded in MHC genes. The MHC protein hierarchy is based on the hierarchy of alleles of MHC genes. In addition, proteins are classified either as protein strand or as multi strand protein, dependent on whether they consist of different subunits or not.

Because of the rapid rate of new allele discovery, allele incorporation is automated using the IMGT/HLA and IPD-MHC databases.

---

Ontology Relations

The ontology has a taxonomic backbone based on the *is-a* (subclass) relation. In addition it imports the OBO Relation Ontology [13], of which it uses the fundamental relation part-of and its reciprocal relation has-part. Further on, the MHC ontology comes with the semantic relations (reciprocal relations are given in brackets) encodes (encoded-in), from-species, variant-of (has-variant), and position-of (has-position). The relation part-of (has-part) is used to relate parts of objects to objects, such as multi strand protein has-part (at least two) protein strand. The relation encodes relates genes (their nucleotide sequence) to gene products. The relation from-species is used to specify the source species of particular classes. For example, the class HLA allele is related to the class human (subclass of organism) by a from-species relation. The variant-of relation is used to link alleles to their corresponding genes, while genes are associated with the locus on which they reside by has-position relations.

Class Restrictions

The fine art of ontology design is trying to describe ontology classes not only by their name and the corresponding verbal definition, but also define them in a formal way. This can be done by introducing necessary, or even necessary and sufficient, conditions whenever possible in terms of OWL class restrictions. Supplied with this type of information description classifiers are able to dynamically compute subsumption relations (*is-a*) between concepts rather than relying on primitive, i.e., hard-wired encodings of classificatory relations between these concepts [9]. The example in Figure 1 illustrates class restrictions based on the has-part relation. The conceptual relation encoded-in is used.

Example 1: The class HLA-DP protein is specified by the restrictions that it *is-a* MHC class-II protein and that each HLA-DP protein has-part exactly 1 alpha subunit of HLA-DP protein as well as exactly 1 beta subunit of HLA-DP protein (cf. Figure 1).

Example 2: The class alpha subunit of HLA-DP protein is specified by the restrictions that it *is-a* alpha subunit of MHC class-II protein and it is encoded-in some HLA-DP A1 allele (cf. Figure 2).

Related Work

In the field of immunogenetics, two terminological resources deal with related topics: the IMGT ontology for immunogenetics [1, 8] and the IEDB’s ontology describing immune epitopes [12]. The IMGT ontology serves as a controlled vocabulary to describe the data in the international ImmunoGeneTics (IMGT) database in a standardized way. The field of immune epitopes is covered by another ontology, built to characterize the data in the Immune Epitope Database and Analysis Resource14 (IEDB). The database contains multi-species data related to antibody and T-cell epitopes.

Both ontologies were built to organize huge data bases and, thus, are strictly bound to the database structure similar to relational schemata. They thus lack a level of abstraction that relates various classes by explicit taxonomic, partonomic and additional conceptual relations (such as the MHC ontology’s encodes and from-species relations). For the MHC ontology, this also leads to the creation of intermediary classes which capture biomedical generalizations relevant for formal reasoning. So our contribution is mainly one of reformulating immunogenetic knowledge under the premises of proper ontological design considerations.

14http://immuneepitope.org/home.do
In order to shed light on this claim consider the class **MHC molecule** from the IEDB’s ontology. It is represented in terms of a frame which consists of the slots ‘Source Species’, ‘MHC Class’, ‘MHC Allele(s)’, ‘has MHC Chain 1’, and ‘has MHC Chain 2’. There are no subclasses of **MHC molecule**, the frame is instantiated instead by directly filling data base entries into the slots. In the MHC ontology, we also describe source species, MHC class, encoding alleles, included chains. However, instead of slots we use OWL class restrictions, e.g., for a slot like ‘MHC allele’ we introduce a class restriction **MHC molecule encoded-in MHC allele**. Apart from the different ways of representation (frames versus description logics), in the MHC ontology the MHC molecule includes detailed subclasses such as **HLA-DP protein** with class restrictions, whereas in the IEDB ontology no class ‘HLA-DP protein’ exists, instead the IEDB database holds information about HLA-DP in terms of a filled frame.

Generalization in terms of class hierarchies and formal restrictions on class-defining attributes yield a level of abstraction of biomedical knowledge that allows for computations needed at the front-end (user interaction) and back-end (information extraction system) of the StemNet system. For our application scenario (use of the ontology to navigate through search results, use of the ontology to improve the mapping from conceptual to terminological knowledge, use of the ontology to improve advanced NLP applications such as relation extraction) a formal ontology is required that is easily accessible, intuitive, coherent, and comprehensively represents domain knowledge. These requirements are neither fulfilled by the IEDB ontology nor by the IMGT ontology, since their purpose is to serve as a knowledge organization framework for databases only. Not to mention the fact that major parts of both ontologies deal with ‘meta’ information related to ‘content’ classes, such as references (journal ID, article title, authors, date, etc.), though this is not really medical domain knowledge.

To embed the MHC ontology in already existing resources on the conceptual level, its high-level classes such as **MHC Gene** could easily be mapped to IMGT and IEDB ontology classes, apart from technical representation issues. However, it would be even more worthwhile to link the IMGT and IEDB data base entries to MHC ontology classes, since the huge coverage of these data bases would then be complemented by the additional functionalities that are unique for the MHC ontology.

**Conclusions**

The MHC ontology provides a formal, coherent and consistent representation of multi-species MHC alleles and encoded proteins. In contrast to already existing ontologies in the field of immunogenetics that were exclusively built in support of immunogenetics data bases, the MHC ontology achieves conceptual abstraction and higher expressiveness in terms of explicit taxonomic, partonomic and additional conceptual relations and the incorporation of class restrictions. The formal encoding of the MHC ontology in the description logics language OWL DL allows to run a description classifier which, unlike database-oriented conceptual schemata, computes (i.e., makes explicit) formally implicit knowledge as needed for advanced user interaction with query processors and information extraction systems.

**Acknowledgement**

This work was carried out within the StemNet project (http://www.stemnet.de), which is supported by a grant from the Federal Ministry of Education and Research (BMBF), Germany.

**References**