BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 16 2009, pages 2064–2070
doi:10.1093/bioinformatics/btp306

Databases and ontologies

MaHCO: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining[†]

David S. DeLuca[1,2,*], Elena Beisswanger[3], Joachim Wermter[3], Peter A. Horn[4], Udo Hahn[3] and Rainer Blasczyk[1]

[1]Institute for Transfusion Medicine, Hannover Medical School, Hannover, Germany, [2]Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA, [3]Jena University Language & Information Engineering (JULIE) Lab, Jena and [4]Institute for Transfusion Medicine, University Hospital Essen, Germany

Received on February 12, 2009; revised on May 1, 2009; accepted on May 4, 2009

Advance Access publication May 7, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: The high level of polymorphism associated with the major histocompatibility complex (MHC) poses a challenge to organizing associated bioinformatic data, particularly in the area of hematopoietic stem cell transplantation. Thus, this area of research has great potential to profit from the ongoing development of biomedical ontologies, which offer structure and definition to MHC-data related communication and portability issues.

Results: We introduce the design considerations, methodological foundations and implementational issues underlying MaHCO, an ontology which represents the alleles and encoded molecules of the major histocompatibility complex. Importantly for human immunogenetics, it includes a detailed level of human leukocyte antigen (HLA) classification. We then present an ontology browser, search interfaces for immunogenetic fact and document retrieval, and the specification of an annotation language for semantic metadata, based on MaHCO. These use cases are intended to demonstrate the utility of ontology-driven bioinformatics in the field of immunogenetics.

Availability and Implementation: The MaHCO Ontology is available via the BioPortal: http://www.bioontology.org/tools/portal/bioportal.html, and at: http://purl.org/stemnet/

Contact: david_deluca@dfci.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Ontologies, in the recent past, have increasingly become an established methodology to structure and to communicate the knowledge in the field of the life sciences. Their use, e.g. for functional annotation in databases and as a vehicle for query and retrieval processes is now broadly acknowledged in the community. The *Open Biomedical Ontologies* (OBO, http://www.obofoundry.org/) (Smith *et al*., 2007) provide a common access point for a large number of biomedical ontologies, currently about 80 (as of February, 2009). The ontologies within the OBO framework vary under many dimensions—the topics being covered (from human anatomy via cell structure, molecular functions and processes to experimental methods and chemical substances), their size (from several hundred thousands of terms and relations to some hundreds), descriptional granularity and representational perspective. A broad range of biomedical topics have already been covered. Overlapping areas, as well as thematic gaps are encountered. One such gap could be filled by a terminological system that focuses extensively on the major histocompatibility complex (MHC) (Beisswanger *et al*., 2007). This is the subject of MaHCO (Major Histocompatibility Complex Ontology) whose design principles, contents and use cases will be introduced in this article.

The MHC poses many challenges to researchers of pathology, transplantation and immunology resulting from the high level of polymorphism in this genomic area, in human located on chromosome six. In particular, genetic typing of human leukocyte antigen (HLA) alleles to ensure patient/donor compatibility for solid organ or hematopoietic stem cell transplantation is a complex, data intensive task that relies heavily on computers for storage, organization and interpretation of information. HLA typing can be performed with various levels of precision (Little, 2007). This leads to a hierarchical categorization of HLA alleles. The classic method of HLA typing is via serological testing. In this method antibodies are used to recognize particular structural domains of HLA proteins classifying the proteins into serological groups. For a more precise grading into ‘serological splits’ (i.e. subclasses of serological groups) antibodies of higher specificity are used that react with subsets of HLA antigens. Still serological typing is the least precise typing method, because several HLA proteins share the relevant structural domains. In contrast, genetics-based HLA typing can determine which HLA allele is present in a more specific manner. It is also performed with various levels of precision, resulting in so-called low, medium or high resolution results.

The HLA nomenclature has been designed to make the classification of alleles immediately apparent from their names using a two to eight digit code. The first two digits in the name indicate the serological group an allele belongs (Marsh, 2003). This means that alleles sharing the first two-digits in their names encode proteins belonging to the same serological group. If they share also the third

*To whom correspondence should be addressed.

[†]In this article, ontology classes are written in **bold**; class relations are written in *italics*; annotation elements are written in ***bold italics.***

and forth digit, they encode the same protein based on the amino-acid sequence. If, in addition, they share the fifth and sixth digit, they do not contain any synonymous mutations in the coding sequence. Finally, if they also share the seventh and eighth digit, even the non-coding parts of the nucleotide sequence are the same for both alleles.

Determination of the serological group (e.g. A2) as well as 'two-digit' genetics-based typing results (e.g. A*02) are considered low resolution typing. Donor and recipient alleles are considered to be a low resolution match when each allele has been determined to belong to the same serological or two digit group. Therefore, the definition of allelic groups is highly clinically relevant when dealing with HLA alleles and is thus incorporated into MaHCO, in parallel to the more fine-grained grading of HLA alleles relying on genetics-based typing.

Recently, as the number of determined HLA alleles has increased, and for some serological groups has exceeded 100, the third and fourth digit of HLA allele names turned out to be inadequate to distinguish all different known alleles. This has resulted in the A*02 group 'spilling over' into the A*92 group (Marsh *et al.*, 2002). While those users who have years of experience with the HLA system know that an A*92 allele is actually an A*02 allele, this is likely to create much confusion for the next generation of researches and clinicians. Currently, the only further example of a group extension is B*15 spilling over into B*95.

In a ontology, domain-specific concepts are arranged with respect to content instead of names. Thus the development of MaHCO offers a means to render a consistent view on HLA alleles, which may help to get over the present shortcomings and idiosyncrasies of the HLA nomenclature.

2 METHODS

MaHCO has been manually built by a team of biologists and bioinformaticians who relied on previous knowledge structuring efforts in the field of immunogenetics and in particular the MHC. The development and design was strongly influenced by principles advocated by the OBO Foundry, providing a publicly available ontology implemented in a standardized description language and using formally founded relations in the style of the Relation Ontology (RO) (Smith *et al.*, 2005), in particular.

2.1 Source data

MaHCO was designed to cover MHC genes, alleles and encoded proteins. A major goal for the construction of the ontology was to integrate as much external knowledge as possible about the targeted domain and to ensure compatibility with already existing ontologies covering related domains. To reach the goal of compatibility the top-level classes of MaHCO, such as **Allele** and **Gene** and **Chain** were created in equivalence with similar classes in the Sequence Ontology (SO) (Eilbeck *et al.*, 2005). In addition to species-independent classes, we have chosen to incorporate species-specific branches. Terms for human, mouse and dog were implemented in this version for their respective clinical and experimental relevance. Human MHC related classes are based on files provided via FTP by the IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla/), as well as the HLA Dictionary for serological definitions (Schreuder *et al.*, 2005). Definitions of serological splits were adopted from the website of the HLA Informatics Group at the Anthony Nolan Trust (http://www.anthonynolan.org.uk/HIG/lists/broad.html). Subclasses for the **Canine_MHC_Allele** class were adapted from the DLA Nomenclature Reports (DLA = dog leukocyte antigen), as provided by the Immuno Polymorphism Database (IPD) (http://www.ebi.ac.uk/ipd/) (Kennedy *et al.*, 2001; Robinson *et al.*, 2005). The listing of murine MHC genes found at the IMGT web resource (http://www.ebi.ac.uk/imgt/) became the basis of the **Mouse_MHC_Allele** class and subclasses (Lefranc, 2001).

The quantitative distribution of these sources is summarized as about 72% from the IMGT/HLA DB, 27% from the HLA Dictionary and 1% from the IDP and IMFT resources. The quantity of data is based upon the number of classes in the ontology (totaling around 6700) which relied on the given data sources.

2.2 OWL as a formal basis of conceptual design

MaHCO is formalized in OWL DL, a description logics-based sublanguage of OWL. It was chosen because it is highly expressive, on the one hand, and it still retains computational completeness and decidability, on the other hand. In addition, several reasoning systems are available for OWL DL.

The basic structure of MaHCO is a directed acyclic graph (DAG) based on *rdfs:subClassOf* relations (a.k.a. *is-a*) between classes and their parent classes. MaHCO classes are provided with URIs holding its namespace and local name. In addition, for each class the unformatted, human-readable version of the class name is provided in terms of an *rdfs:label* annotation.

Textual definitions of classes, synonyms of class names and references to similar classes in other terminologies and ontologies are provided by the custom OWL annotation properties ***definition***, ***synonym*** and ***reference***, respectively, which we have defined. Metadata describing the resource MaHCO as such are provided using DCMI (Dublin Core Metadata Initiative) Metadata Terms (http://dublincore.org/) such as ***dc:creator***, ***dc:date***, ***dc:subject*** and ***dc:title***. The term ***dc:source*** is used as holder for literature or database references.

2.3 Knowledge editing and reasoning framework

While the upper level of MaHCO was created manually, including domain knowledge from sources mentioned above, and using the Stanford University ontology editor Protégé (Protégé-OWL, version 3.4) (http://protege.stanford.edu), subclasses of the **Human_MHC_Allele** class were automatically generated extracting terms and taxonomic relations from external databases. These terms have been put into an external ontology we call MaHCO-HLA. The MaHCO-HLA, stored in the file MaHCO_HLA.owl, is included in MaHCO, stored in the file MaHCO owl using an ***owl:imports*** statement. The generation algorithm for the MaHCO-HLA was written in Java. The extracted data was exported into the OWL format using the

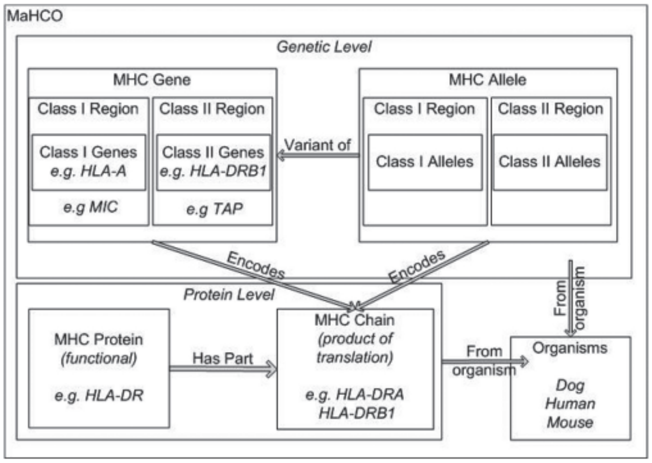


Fig. 1. Bird's eye view of the MaHCO domain. Generalized domains of MaHCO are depicted as boxes and comments are given in italics. Arrows represent relations of the ontology connecting classes. Note: *Genetic/Protein Level* are descriptors for this diagram, but not terms in the ontology.

Jena 2 Ontology API (Carroll *et al*., 2004). An additional version of the MaHCO-HLA was created in XML which offers simplicity while reflecting the hierarchical nature of the HLA system. These versions of the ontology were exported in XML using the JDOM API (http://www.jdom.org/). To check the consistency of MaHCO we run the OWL DL reasoner Pellet (Sirin *et al*., 2007), version 1.5.1.

3 RESULTS

In the following section, we describe the basic classes of MaHCO, discuss the relations it includes, present the hierarchical structure of the MaHCO-HLA which is automatically imported into MaHCO and present how MaHCO references classes in other external knowledge repositories. A conceptualized general overview of MaHCO can be seen in Fig. 1.

3.1 Classes of MaHCO

MaHCO consists of 106 classes interlinked by taxonomic *rdfs:subClassOf* relations and seven semantic relation types. Additionally, the imported MaHCO-HLA contains 6649 *rdfs:subClassOf* linked classes. The top classes of MaHCO are **Organism**, **Nucleotide_Sequence (**with subclasses **Allele** and **Gene)**, **Polypeptide (**with subclass **Chain)** and **Protein** (see Fig. 2). **Organism** subsumes the species for which MaHCO provides MHC gene and molecule descriptions, namely **Human**, **Mouse** and **Dog**. For a more detailed schematic of the upper levels of MaHCO, see Supplementary Figure 1.

The main distinction between **Chain** and **Protein** is that **Chain** contains contiguous peptide sequences, while a **Protein**, on the other hand, consists of single or multiple folded chains. The class **HLA_DP_Molecule**, for example, is specified as a subclass of Protein that *has_part* **MHC_ClassII_Alpha** and *has_part* **MHC_ClassII_Beta,** both of which are subclasses of **Chain**.

Allele and **Gene** subsume MaHCO classes **MHC Allele** and **MHC Gene**, respectively. Further subclasses make distinctions based on MHC class and species membership. The class **MHC Allele**, for example, has the subclasses **MHC_Class_I_Allele**, **MHC_Class_II_Allele** and **MHC_Class_III_Allele**, subdividing alleles corresponding to the family of MHC genes they belong to. Additional subclasses **Human_MHC_Allele**, **Mouse_MHC_Allele** and **Dog_MHC_Allele** allow for an orthogonal, organism-centered classification of MHC alleles. MaHCO also provides the cross-product classes, in a pre-coordinated form, such as **Human_MHC_Class_I_Allele**. These classes are linked to the respective parent classes, in the example **Human_MHC_Allele** and **MHC_Class_I_Allele**. This makes MaHCO to a multi-hierarchy, meaning that a class in the MHC can have multiple parent classes. The subdivision of MHC alleles in MaHCO is shown in Figure 3 and in greater detail in Supplementary Figure 2.

Composing MaHCO required the introduction of new immunogenetic terminology in order to adhere to the strict ontological design principles we subscribed to at the outset of our work. First, there is an additional subclass of **MHC_Allele** called **MHC_Allele_Encoding_Peptide_Presenting_Protein**. It was introduced to distinguish true MHC class I and class II alleles (namely those encoding proteins presenting peptides on the surface of cells, in human called HLA alleles, H2 in mouse and DLA in dog) from all other alleles that are contained in the MHC class I or II region (e.g. such that encode proteins with another functionality, e.g. being involved in the proteasome). For example, in human the class I region is a portion of the MHC which contains true MHC class I genes such as HLA-A, B, C, etc., but also genes called MIC that are not real class I genes (MIC is considered 'class I-similar'). **MHC_Allele_Encoding_Peptide_Presenting_Protein** is defined as the union of **MHC_Class_I_Allele** and **MHC_Class_II_Allele**.

Second, in order to group all alleles by the region of the MHC in which they are encoded, regardless of the kind of protein they encode, the intermediary classes **MHC_Class_I_Region_Allele** and **MHC_Class_II_Region_Allele** have been introduced to MaHCO that are direct subclasses of **MHC_Allele** and direct parent classes of **MHC_Class_I_Allele**, **MHC_Class_II_Allele**, **MIC_Allele** and **TAP_Allele** (TAP genes are encoded in the MHC class II region but are not real class II genes).

MaHCO comes with further subclasses of **Gene** and **Allele**, such as **Protein_Coding_Gene**, **Pseudogene** and **Null_Allele**. The

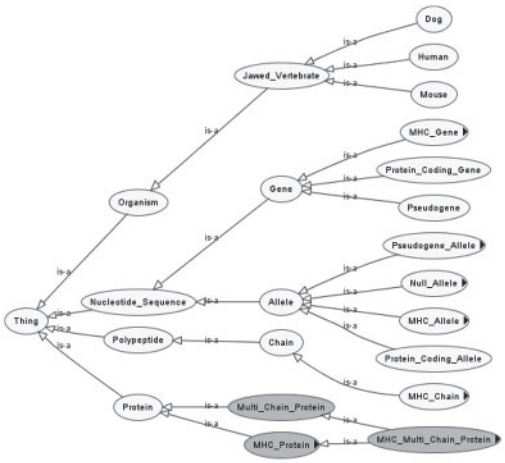


Fig. 2. Top level classes of MaHCO. The top level classes of MaHCO are depicted in the circles and connected via *is_a* relations, as represented by the arrows. For example, **Gene** *is_a* **Nucleotide_Sequence.**

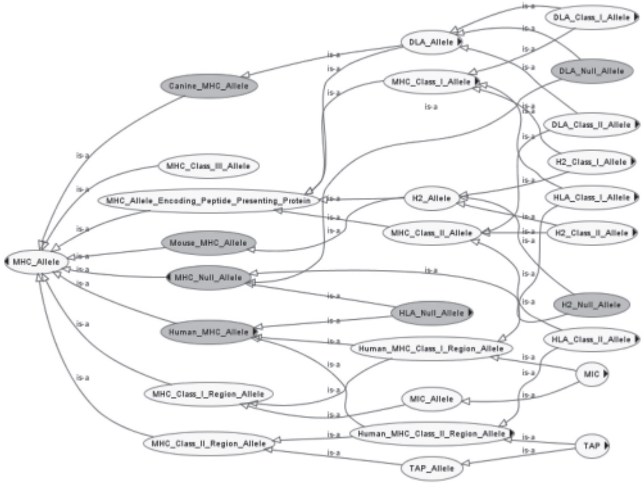


Fig. 3. Multihierarchical classification of MHC alleles in MaHCO. The mid-level classes of MaHCO are depicted in the circles and connected via *is_a* relations, as represented by the arrows. For example, **MHC_Class_I_Allele** *is_a* **MHC_Allele_Encoding_Peptide_Presenting_Protein.**

Table 1. Relations of MaHCO

Relation (inverse)	Domain	Range
encoded_in (encodes)	Polypeptide, Protein	Nuceleotide_ Sequence
has_part (part_of)	owl:Thing	owl:Thing
variant_of (has_variant)	Allele	Gene
from_species	Protein, Nucleotide_Sequence, Polypeptide	Organism

class **Null_Allele**, e.g. subsumes alleles that are not expressed or are expressed but encode functionless peptide chains which would soon be digested. Accordingly, there are no members of the class **MHC_Chain**, encoded by **Null_Allele** or one of the subclasses of it.

Classes subsumed by **MHC_Chain** in the polypeptide branch of MaHCO represent the gene products encoded in MHC alleles. For most subclasses of **MHC_Allele** in the nucleotide sequence branch of MaHCO corresponding subclasses of **MHC_Chain** were created. Since the HLA nomenclature refers only to alleles, there is no distinct terminology for HLA proteins or chains. To be still able to distinguish between the two in MaHCO we created the missing names by appending the word 'Chain' to each appropriate allele name (e.g. **B_4402_Chain** denotes the class of MHC chains encoded in **B_4402** alleles). However, the relevant natural language class name (provided by an ***rdfs:label*** annotation) does not contain the word chain (e.g. the label of **B_4402_Chain** is **'**B*4402'). Outside of the hierarchical context it is indistinguishable from the allele label.

For the subclasses of **Alternatively_Expressed_Allele** we only created the correlatives **HLA_Cytoplasm_Chain**, **HLA_Low_Chain** and **HLA_Secreted_Chain** under the class **Alternatively_Expressed_Chain** because there is no conclusive experimental evidence describing chain products of alleles included in **HLA_Aberrant_Allele** and **HLA_Questionable_Allele**.

Examples of relation use include **MHC_Chain** *encoded_in* **MHC_Allele, MHC_Protein** *has_part* **MHC_Chain, MHC_Allele** *variant_of* **MHC_Gene,** and **HLA_Class_I_Allele** *from_species* **Human.**

3.2 Class restrictions and relations of MaHCO

Wherever possible, we formally specified the meaning of MaHCO classes by means of OWL class restrictions. For example, the class **Human_MHC_Allele** is formally defined as being a subclass of **MHC_Allele** that is linked by a *from_species* relation to the class **Human**. In addition it inherits the restriction 'is *variant_of* some **MHC_Gene'** from its parent class **MHC_Allele.**

The relations provided by MaHCO are summarized in Table 1 along with their respective inverse relations. They are used to link classes within and across MaHCO ontology branches. The relation *encoded_in* relates gene products (classes from the peptide and protein branches of MaHCO) to genes or alleles of genes (classes of the nucleotide sequence branch of MaHCO), such as **MHC_Chain** to **MHC_Allele**. The relation *has_part* relates parts to the whole, such as the class **MHC_Protein** to **MHC_Chain**. The relation *variant_of* links alleles to the respective genes, and *from_species* links species-specific classes to the species, such as **HLA_Class_I_Allele** to **Human**.

3.3 Structure of MaHCO-HLA

The hierarchical structure of HLA alleles, as described in the introduction, was implemented in MaHCO-HLA. Alleles belonging to an HLA locus were divided into serological groups and then subdivided into serological splits. In parallel, the other typical categorization of HLA was implemented as well, namely a categorization based upon the number of digits of the allele name known to the observer. Two-digit alleles are siblings of serological groups. Two-digit alleles as well as serological groups are then subdivided into four-digit, six-digit and finally eight-digit allelic classes. The following exceptions to this rule were made: (i) alleles whose names begin with A*92 were placed in the A*02 group, (ii) alleles whose names begin with B*95 were placed within the B*15 group and (iii) the two-digit group levels for MIC, TAP and DP were excluded.

The HLA nomenclature accounts for five forms of alternatively expressed alleles. These are represented in MaHCO-HLA as subclasses of **Alternatively_Expressed_Allele**, namely **HLA_Aberrant_Allele**, **HLA_Cytoplasm_Allele**, **HLA_Low_Allele**, **HLA_Questionable_Allele** and **HLA_Secreted_Allele.** For the textual definitions see Supplementary Table 1.

In addition to the MaHCO_HLA.owl file that is imported by MaHCO a file MaHCO_HLA_twodigit.owl is provided that contains the MaHCO-HLA without the serological group classes. Because serological typing is becoming a legacy technology, the serological groups in the MaHCO-HLA are necessary in some applications, but obsolete in others.

Our experience has shown that additional formats can be very convenient for implementing the ontological design. Thus, additional files, MaHCO_HLA.xml, MaHCO_HLA_twodigit.owl and MaHCO_HLA_twodigit.xml can be found under http://purl.org/stemnet/. These files provide non-OWL XML representations of the MaHCO-HLA, which might be practical for bioinformatians and programmers which are not interested in carrying the overhead of OWL-parsing libraries.

3.4 Availability

The major characteristics of MaHCO, and the MaHCO-HLA extension, are summarized in Table 2. The first release of this ontology is given the version number 1.0. MaHCO's namespace is http://purl.org/stemnet/MHC; the namespace of the MaHCO-HLA is http://purl.org/stemnet/HLA, and has been given the prefix 'HLA' when imported into MaHCO. MaHCO-HLA represents the HLA system as of the IMGT/HLA Database Release 2.21.0, 08 April 2008. The MaHCO Ontology can be accessed online either under http://purl.org/stemnet/ or at the American National Center for Biomedical Ontology's BioPortal under http://www.bioontology.org/tools/portal/bioportal.html.

To ensure compatibility with other established ontologies, and to prevent redundancy, most of the basic classes of MaHCO are linked to external ontology entries via ***reference*** annotation statements. MaHCO classes and their corresponding classes in external ontologies are listed in Table 3. Currently, the only reference to an immunogenetic resource is **MHC Molecule** in the ontology of the Immune Epitope Database (Sathiamurthy *et al*., 2005) which is referenced by the class **MHC_Protein** in MaHCO. Other central MaHCO entries correspond to

Table 2. MaHCO Ontology fact sheet

Ontology Name	MaHCO	MaHCO-HLA
Namespace	http://purl.org/stemnet/MHC	http://purl.org/stemnet/HLA
Prefix	MHC	HLA
Scope	MHC alleles, genes and proteins in human, mouse, and dog	HLA alleles
Format	OWL DL	OWL DL
Number of classes	106	6649
Dependencies	Dublin core[a], MaHCO-HLA	Dublin core[a]
Data sources	IPD[b], IMGT[c]	IMGT/HLA[d], HLA Dictionary[e], Anthony Nolan Trust[f]
Relations	*rdfs:subClassOf, encoded_in, encodes, has_part, part_of, variant_of, has_variant, from_species*	
Annotations	***rdfs:label, dc:creator, dc:date, dc:publisher, dc:source, dc:subject, dc:title, definition, synonym, reference***	
Additional files	MaHCO_HLA.xml, MaHCO_HLA_twodigit.xml[g]	MaHCO_HLA_twodigit.xml

[a] Dublin core: http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl.
[b] Immuno Polymorphism Database (Robinson *et al*., 2005).
[c] The international ImMunoGeneTics database (Lefranc, 2001).
[d] Official source of HLA nomenclature and sequences (Marsh, 2003).
[e] Source of serological associations (Schreuder *et al*., 2005).
[f] Definitions of serological splits were provided by the website of the HLA Informatics Group at the Anthony Nolan Trust (http://www.anthonynolan.org.uk/HIG/lists/broad.html).
[g] These files can be accessed at http://purl.org/stemnet/

Table 3. References to external knowledge repositories

MaHCO class	External Ontology Class(es)
MHC Protein	IEDB: **MHC Molecule**
Allele	NCI:C16277 **Allele**, SO:0001023 **allele**
Gene	SO:0000704 **gene**
Pseudogene	SO:0000336 **pseudogene**
Protein	CHEBI:36080 **proteins**
Polypeptide	SO:0000104 **polypeptide**
Chain	CHEBI:16541 **protein polypeptide chains**, SO:0001063 **immature_peptide_region**
Jawed_Vertebrates	TaxonomyID:7776 **Gnathostomata**
Human	TaxonomyID:9606 **Homo sapiens**
Dog	TaxonomyID:9615 **Canis lupus familiaris**
Mouse	TaxonomyID:10090 **Mus musculus**
Organism	NCI:C14250 **Organism**

IEDB, Immune Epitope Database; NCI, The National Cancer Institute Thesaurus; SO, Sequence Ontology; ChEBI, Chemical Entities of Biological Interest; TaxonomyID, These are IDs given to entries in the NCBI Taxonomy database.

terms in the Sequence Ontology (SO) (Eilbeck *et al*., 2005). Further reference terms were found in the NCI Thesaurus (http://www.obofoundry.org/cgi-bin/detail.cgi?id=ncithesaurus), as well as in the ontology of Chemical Entities of Biological Interest (ChEBI) (http://www.ebi.ac.uk/chebi/) (Degtyarenko *et al*., 2008). The organism classes of MaHCO are linked to entries in NCBI's Taxonomy database (www.ncbi.nlm.nih.gov/Taxonomy/) (Wheeler *et al*., 2000).

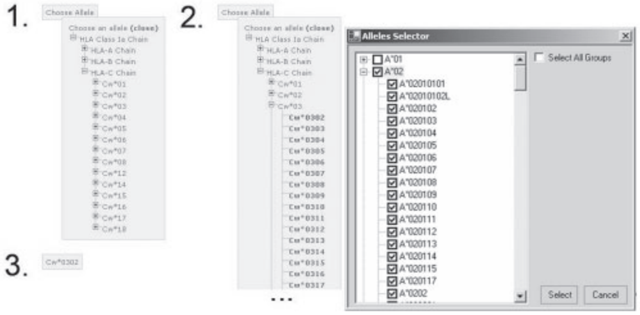


Fig. 4. User interfaces utilizing MaHCO. On the left, a web popup control panel for choosing ontology elements at www.peptidecheck.org. The allele chooser receives a top class as a parameter. In this case, Class I chains are of interest and so the top element is **HLA_Class_Ia_Chain**. In step 1, the user clicks on the "Choose Allele" button to reveal a tree structure in which subclasses can be opened. In step 2, the user has reached the leaf level and may select the desired chain. Step 3 shows that the user's selection has been assigned, and the program can use this input for whatever purpose. On the right, a visual basic control panel for selecting multiple classes in an ontology.

4 USE CASES FOR MaHCO

MaHCO has already proven useful in various use cases. They range from standard database applications via browsing-based search scenarios for biological data as well as documents, up to infrastructure purposes for text mining systems. The major use cases will be briefly outlined below.

4.1 Focusing Browser for MaHCO

Once the relations between HLA concepts had been formally defined in the computer-readable OWL format, creating user interfaces for HLA-centered programs became much easier. We demonstrated this by creating a new webpage dialog window which displays the content of a ontology so that the user may choose one entity (see Fig. 4).

This dialog solved the problem of choosing a specific HLA allele or protein from a list of thousands of entries. The hierarchical nature of ontologies allows them to be represented as a tree structure. In this way, the user is not overwhelmed by a long list of allele or protein names, but simply opens the categories of interest, revealing only the relevant subset. This also has performance benefits with respect to the time it takes to load a webpage. Since the dialog is based upon AJAX, only the relevant information is loaded with each click by communicating with the server in the background. This saves the significant lag time required to load the entire ontology into the browser, especially when dealing with large ontologies. The dialog is not only functional for MaHCO-HLA, but can load any ontology which is represented in OWL. Other formats could be incorporated without requiring major modifications. This dialog can be seen in use on the www.peptidecheck.org webpage, under the Module Explorer page.

In some applications, the user wants to perform an action on a group of alleles. This is particularly true when working at low resolution—or on the 'two-digit' group level. For this purpose we have developed a Visual Basic-based dialog interface which uses MaHCO-HLA to allow the user to conveniently select multiple alleles (Fig. 4).

The advantage of this ontology-driven dialog becomes clear when one wants to perform an action involving all A*02 alleles. Because the A*92 alleles are also part of the A*02 group as defined in the ontology, they will also be included when choosing the whole of A*02.

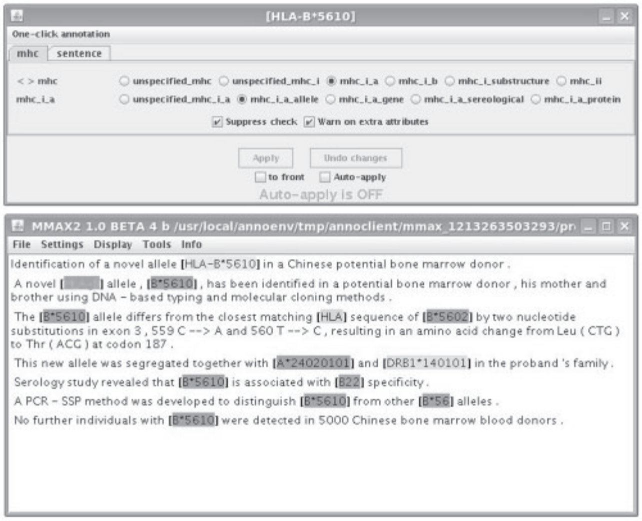
4.2 Annotation vocabulary for text mining engines

Due to the overwhelming amount of knowledge buried in the ever-growing volumes of biomedical literature, text mining techniques, in particular named entity recognition and information extraction, have come to play a major role in bioinformatics applications, as they help to find information in scientific publications more easily and completely (see, e.g. Krallinger *et al*., 2008; Altman *et al*., 2008). A crucial step here is the recognition of named entities in literal natural language text and their subsequent mapping to formal database or ontology identifiers. Hence, when given the sentence '*Serological study revealed that B*5610 is associated with B22 specificity*',[1] a text mining system with an MHC named entity module (such as described, e.g. in Hahn *et al*., 2007) should recognize that the text string 'B*5610' denotes a **MHC_Class_I_Region_Allele** and, subsequently, should map this text string to a respective ontology or database identifier, as e.g. provided in MaHCO-HLA. In addition, such a text mining system should also recognize that the text string '*B*22' refers to the respective HLA serological group type.

While a seemingly straightforward approach in this case would be to compile a dictionary containing the known serological group types and thus match the dictionary entries against natural language text documents, such a naïve approach falls short due to several reasons. For example, querying the PubMed title and abstract field with the string 'B22' yields many abstracts not related to MHC-/HLA-pertinent issues.[2] In order to cope with such notorious word ambiguity problems, current text mining engines employ machine learning (ML)-based named entity recognizers or taggers which are trained on human-annotated text data. In such a way, human annotators mark up the entities of interest in real-world text and these annotations, in turn, serve as training examples for ML-based taggers (Tomanek *et al*., 2007).

One of the key challenges in devising such an annotation task is to define the entities of interest which should be annotated (marked up) in text, i.e. to define the so-called *annotation vocabulary*. Crucially, the annotation vocabulary should be defined at an appropriate level of semantic granularity, which means that the entity types to be annotated should neither be too general nor too specific. In order to find such an appropriate semantic granularity level, a domain-specific terminological resource based on sound conceptual design principles, such as MaHCO, offers valuable guidance. In particular, the selection of MaHCO classes which could serve as annotation entity types needs to reflect their linguistic manifestations in biomedical texts. A screenshot of the MHC Annotation Environment is given in Figure 5 showing the text annotations performed on the PubMed abstract from which the above example sentence was taken.

In total, we annotated ∼300 000 word tokens with the MHC annotation scheme and used these as training data for an ML-based entity tagger. To evaluate the tagger's performance quality, we annotated around 300 abstracts as a gold standard on which we ran the trained entity tagger. In terms of F-measure, our MHC tagger achieves 82.8% (Precision: 83.1%, Recall 82.5%), which is an excellent result given the current state-of-the-art performance of biomedical entity taggers (Hirschman *et al*., 2007) and the inherent difficulty of the MHC annotation and tagging task. The tagged MHC entities, in turn, are then subject to a mapping procedure which links them to their respective MaHCO class identifiers. In this way, the MHC tagger is fully integrated as one of several semantic modules of the high-performance text mining tool suite for biomedicine currently under development at the JULIE Lab (Hahn *et al*., 2008).



[HLA-B*5610]

One-click annotation

mhc | sentence

< > mhc: ○ unspecified_mhc ○ unspecified_mhc_i ● mhc_i_a ○ mhc_i_b ○ mhc_i_substructure ○ mhc_ii

mhc_i_a: ○ unspecified_mhc_i_a ● mhc_i_a_allele ○ mhc_i_a_gene ○ mhc_i_a_sereological ○ mhc_i_a_protein

☑ Suppress check ☑ Warn on extra attributes

Apply | Undo changes

□ to front □ Auto-apply

Auto-apply is OFF

MMAX2 1.0 BETA 4 b /usr/local/annoenv/tmp/annoclient/mmax_1213263503293/pr

File Settings Display Tools Info

Identification of a novel allele [HLA-B*5610] in a Chinese potential bone marrow donor .

A novel [[illegible]] allele , [B*5610] , has been identified in a potential bone marrow donor , his mother and brother using DNA - based typing and molecular cloning methods .

The [B*5610] allele differs from the closest matching [HLA] sequence of [B*5602] by two nucleotide substitutions in exon 3 , 559 C --> A and 560 T --> C , resulting in an amino acid change from Leu (CTG) to Thr (ACG) at codon 187 .

This new allele was segregated together with [A*24020101] and [DRB1*140101] in the proband 's family .

Serology study revealed that [B*5610] is associated with [B22] specificity .

A PCR - SSP method was developed to distinguish [B*5610] from other [B*56] alleles .

No further individuals with [B*5610] were detected in 5000 Chinese bone marrow blood donors .

Fig. 5. Text annotation using MaHCO-based annotation vocabulary scheme. As can be seen, the highlighted text span ('*HLA-B*5610*') denotes a **MHC_Class_I_a_allele.**

5 DISCUSSION

5.1 Related ontological resources

Further immunogenetically relevant ontologies include the Immune Epitope Database (IEDB) (Sathiamurthy *et al*., 2005) ontology and the IMGT Ontology (Giudicelli *et al*., 2005). After inspection of these ontologies, it is clear that the overlap with MaHCO is minimal (visualized in Supplementary Fig. 3). The IEDB ontology focuses on terminology necessary to classify epitopes. The IMGT Ontology focuses on high level annotation concepts, such as identification, classification and description among others, and uses them to organize a range of immunogenetic concepts including immune globulin, T cell receptor. Neither ontology is publicly available in a standard format for download at the time of this publication.

While these Ontologies complement each other effectively at the moment, as they grow, vigilance is required to avoid redundancy and ensure compatibility.

5.2 Conclusions

Our efforts show that ontologies, indeed, improve the organization of complex concepts in immunogenetics. In the long term, development of an international standard could lead to homogeneous database structures in centers and labs across the world. Utilizing such ontologies has a high potential for ensuring efficient networking and

[1] This sentence is taken from PubMed ID (PMID) 12694576.

[2] For instance, 'B22' may refer to plant populations or chromosome positions, among others.

collaboration. A spectrum of immunoinformatic tools for the MHC system has been established, and will continue to grow. Applications include typing software, bone marrow registry analysis (Muller, 2002), histocompatibility prediction (Elsner *et al*., 2004), T cell and B cell epitope prediction (Buus *et al*., 2003; Duquesnoy and Askar, 2007), and minor histocompatibility antigen prediction algorithms (Halling-Brown *et al*., 2006; Schuler *et al*., 2005). When serving as an infrastructure upon which further immunoinformatic tools can be built, the MaHCO ontology will allow such tools to be easily integrated for use in research institutes and clinical laboratories.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the productive discussions provided by Richard Scheuerman, Steve Marsch and Bjoern Peters of the HLA Subcommittee of Data Interoperability Steering Committee of DAIT, NIAID, NIH as well as Nektarios Ladas of Hannover Medical School for his programming contributions.

Funding: German Federal Ministry of Education and Research, grant number 01DS001A-1C.

Conflict of Interest: none declared.

REFERENCES

Altman,R.B. *et al.* (2008) Text mining for biology – the way forward: opinions from leading scientists. *Genome Biol.*, **9** (Suppl. 2), S7.

Beisswanger,E. *et al.* (2007) An ontology for major histocompatibility complex (MHC) alleles and molecules. In *AMIA'07 – Proceedings of the 2007 Annual Symposium of the American Medical Informatics Association*, USA, pp. 41–45.

Buus,S. *et al.* (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, **62**, 378–384.

Carroll,J.J. *et al.* (2004) Jena: implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters*, Association for Computing Machinery, New York, pp. 74–83.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Duquesnoy,R.J. and Askar,M. (2007) HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. V. Eplet matching for HLA-DR, HLA-DQ, and HLA-DP. *Hum. Immunol.*, **68**, 12–25.

Eilbeck,K. *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*, **6**, R44.

Elsner,H.A. *et al.* (2004) HistoCheck: rating of HLA class I and II mismatches by an internet-based software tool. *Bone Marrow Transpl.*, **33**, 165–169.

Giudicelli,V. *et al.* (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. *Stud Health Technol. Inform.*, **116**, 3–8.

Hahn,U. *et al.* (2008) An overview of JCoRe, the JULIE Lab UIMA Component Repository. In *Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, pp. 1–7.

Hahn,U. *et al.* (2007) StemNet – an evolving service for knowledge networking in the life sciences. In *GES 2007 - Proceedings of the German e-Science Conference.* Max Planck Digital Library (www.ges2007.de), p. 7.

Halling-Brown,M. *et al.* (2006) SiPep: a system for the prediction of tissue-specific minor histocompatibility antigens. *Int. J. Immunogenet.*, **33**, 289–295.

Hirschman,L. *et al.* (2007) *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid: CNIO Centro Nacional de Investigaciones Oncologicas.

Kennedy,L.J. *et al.* (2001) Nomenclature for factors of the dog major histocompatibility system (DLA), 2000: Second report of the ISAG DLA Nomenclature Committee. *Tissue Antigens*, **58**, 55–70.

Krallinger,M. *et al.* (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Lefranc,M.P. (2001) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **29**, 207–209.

Little,A.M. (2007) An overview of HLA typing for hematopoietic stem cell transplantation. *Methods Mol. Med.*, **134**, 35–49.

Marsh,S.G. (2003) HLA nomenclature and the IMGT/HLA sequence database. *Novartis Found Symp*, **254**, 165–173; discussion 173–166, 216–122, 250–162.

Marsh,S.G., *et al.* (2002) Nomenclature for factors of the HLA system. *Human Immunology*, **63**, 1213–1268.

Muller,C.R. (2002) Computer applications in the search for unrelated stem cell donors. *Transpl Immunol*, **10**, 227–240.

Robinson, J. *et al.* (2005) IPD – the Immuno Polymorphism Database. *Nucleic Acids Res.*, **33**, D523–D526.

Sathiamurthy,M. *et al.* (2005) An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. *Immunome Res*, **1**, 2.

Schreuder,G.M. *et al.* (2005) The HLA Dictionary 2004: a summary of HLA-A, -B, -C, -DRB1/3/4/5 and -DQB1 alleles and their association with serologically defined HLA-A, -B, -C, -DR and -DQ antigens. *Tissue Antigens*, **65**, 1–55.

Schuler,M.M. *et al.* (2005) SNEP: SNP-derived epitope prediction program for minor H antigens. *Immunogenetics*, **57**, 816–820.

Sirin,E. *et al.* (2007) Pellet: a practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, **5**, 51–53.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

Smith,B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biology*, **6**, R46.

Tomanek,K. *et al.* (2007) An approach to text corpus construction which cuts annotation costs and maintains corpus reusability of annotated data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Association for Computer Lingueistics, Stroudsburg, pp. 486–495.

Wheeler,D.L. *et al.* (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.