Retrieving factual data and documents using IMGT-ML in the IMGT information system®

Authors : Chaume D.*, Combres K.*, Giudicelli V.*, Lefranc M.-P.*

** Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine,*
141 rue de la Cardonille 34396 Montpellier Cedex 5 - France
Tel: +33 (0)4 99 61 99 65 - Fax: +33 (0)4 99 61 99 01 - email: Denys.Chaume@igh.cnrs.fr

Introduction

IMGT, the international ImMunoGeneTics information system® [1] (http://imgt.cines.fr), is a high quality integrated knowledge resource specialized in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system (RPI) that belong to the immunoglobulin superfamily and to the MHC superfamily. IMGT provides a common access to standardized data from genome to proteome. It contains 5 databases, 9 specific interactive tools and 8000 HTML pages of synthesis and knowledge.

IMGT has developed a formal specification of terms to be used in the domain of immunogenetics and immunoinformatics to ensure accuracy, consistency and coherence. This has been the basis of IMGT-ONTOLOGY [3], the first ontology in the domain, which allows the management of the immunogenetics knowledge for human and other vertebrate species. IMGT-ONTOLOGY includes five main concepts: IDENTIFICATION, CLASSIFICATION, DESCRIPTION, NUMEROTATION and OBTENTION. Standardized keywords, standardized sequence annotation, standardized IG and TR gene nomenclature, the IMGT unique numbering and standardized origin/methodology where defined, based on these five main concepts, respectively. The controlled vocabulary and the annotation rules for data and knowledge management of IG, TR, MHC and RPI of human and other vertebrate species constitute the IMGT Scientific chart (http://imgt.cines.fr/textes/IMGTScientificChart.html).

In order to turn IMGT-ONTOLOGY useful for computing, the structures and relations of the concepts have been written with the XML Schema language. XML Schema language has been defined by the W3C consortium (http://www.w3.org/XML/Schema). The resulting **IMGT-ML** markup language [4, 5] includes, for each IMGT-ONTOLOGY concept, a XML Schema:

- IDENTIFICATION: the "identification" tag, composed of one or more "partIdent" tags, each of them introducing, as attribute, the molecule type (DNA, cDNA ...), the configuration (germline, rearranged ...), gene type (variable, diversity, constant or junction), species, functionality, etc.

- CLASSIFICATION: the "classification" tag, composed of one or more "group", "subgroup", "gene", "allele" tags. The "classification" tag contains the "collection" tag in order to formalize loci with their genes.

- DESCRIPTION: "description" and "annotation" tags gather sequence features with their labels and qualifiers.

- NUMEROTATION: "numerotation" tags introduce "nucSystem" and "proSystem" tags for, nucleotide sequences and amino acid sequences, respectively, within a frame, according to a standardized numbering with gaps and mutations.

- OBTENTION: at the moment the formalization of the "obtention" concept is in progress.

In addition to IMGT-ONTOLOGY tags, tags for factual data, sequences and knowledge have been developed. These tags aggregate IMGT-ONTOLOGY tags, sequence metadata tags (for date, external database references, keywords ...) and literature reference tags. These last tags have been developed here "by default" and can be replaced by any standardized literature reference XML Schema.

IMGT-ML a "query language"

IMGT-ML Schema defines the possible values or domains for each component and subcomponent of the `<seqData>` element. Let us consider the set **U** of all `<seqData>` potential elements compliant with IMGT-ML and **A** the set of all `<seqData>` actually present in the IMGT/LIGM-DB database. As the database is consistent, **A** is included in **U**. Now let us consider a `<seqData>` element, where only some given components are populated and **Q** the set of all elements of **U** that have the same component values as this element. The intersection of the two sets **A** and **Q** is the set of all sequence entries in IMGT/LIGM-DB having the given component values. Therefore an uncompleted `<seqData>` element can be used to query the database and the result of the request is a list of `<seqData>` elements, that all have the given component values. So, IMGT-ML can be seen as a "query language". The `<querySeqData>` tag has been added to IMGT-ML to define the use of `<seqData>` elements for query the IMGT/LIGM-DB database.

Logical connectors: If the `<querySeqData>` element contains more than one `<seqData>` element then the result is the union of all individual `<seqData>` query results (it is the OR operator). At the opposite, the internal components of a `<seqData>` are interpreted as being connected by the AND operator.

Domain restrictions: The `<domain>` element, allows the restriction of attribute or body values of any IMGT-ML element. The restriction elements have been borrowed among XML Schema "restrictions". However, at the moment, the "pattern" restriction values are limited to those compatible with the "like" SQL instruction. The `<domain>` element attribute "complement" indicates if the domain is the complementary of the attribute or body value domain. This allows the "NOT" operator.

Here is some examples :

- *Querying for an unique sequence giving its accession number*:

  ```
  <seqData id="M26678"/>
  ```

  In this case the result is one complete `<seqData>` element.

- *Querying for human Ig-Ligth-Kappa sequences in germline configuration*:

  ```
  <seqData>
  <identification>
  <partIdent chainType="Ig-Ligth-Kappa"
  taxonName="Homo sapiens"
  configuration="germline"/>
  </identification>
  <seqData>
  ```

- *Querying for mouse IGKV8-16 gene sequences*:

  ```
  <seqData>
  <classification>
  <gene name="IGKV8-16" taxonName="Mus musculus"/>
  </classification>
  <seqData>
  ```

- *Querying for sequences containing both "V-J-GENE"* ***and*** *"C-GENE" features*:

  ```
  <seqData>
  <annotation>
  <entity labelName="V-J-GENE"/>
  <entity labelName="C-GENE"/>
  </annotation>
  <seqData>
  ```

- *Querying for sequences containing either a "V-J-GENE"* ***or*** *a "C-GENE" feature*:

  ```
  <seqData>
  <annotation>
  <entity labelName="V-J-GENE"/>
  </annotation>
  <seqData>
  <seqData>
  <annotation>
  <entity labelName="C-GENE"/>
  </annotation>
  <seqData>
  ```

- *Querying for huge sequences having a length greater than 100,000 base pairs*:

  ```
  <seqData>
  <nucSequence>
  <sequence>
  <domain on="@length">
  <minInclusive value="100000"/>
  </domain>
  ```

```
</sequence>
</nucSequence>
<seqData>
```

- *Querying for all sequences*:

```
<seqData/>
```

Restricting the output

By default, the result is a list of complete `<seqData>` elements, that is, a list of elements with all their components populated. It is useful to get only a part of these components and not all of them. The needed components are indicated by a `<seqData>` element where only these component are present, but void. The `<resultTemplate>` tag has been added to IMGT-ML to define the use of `<seqData>` elements to indicate the wanted components. Examples:

- *To get just the nucleotide sequence*:

```
<seqData>
  <nucSequence><sequence/></nucSequence>
<seqData>
```

- *To get the literature references:*

```
<seqData><literature/><seqData>
```

- *To get the list of genes associated with the selected sequences:*

```
<seqData>
  <classification><gene/></classification>
</seqData>
```

- *To get the sequence with its annotation:*

```
<seqData>
  <annotation/>
  <nucSequence><sequence/></nucSequence>
</seqData>
```

IMGT Web services

IMGT/LIGM-DB access

IMGT-ML is used by IMGT Web services to exchange IMGT data. Therefore, the output of any Web service can be used as an input for any other one, provided that it makes sense. Any IMGT information system module is candidate to become a Web service. The first one is LIGMWserver (http://kappa.igh.cnrs.fr:8080/axis/LIGMWserver.jws) built to access the IMGT/LIGM-DB database. This Web service includes the remotely callable method "querySeqData" which has two parameters. The first one is a `<querySeqData>` element that indicates which sequences are wanted, the second one is a `<seqData>` element that indicates which components of these sequences are wanted.

Use of IMGT/JunctionAnalysis tool

IMGT/JunctionAnalysis is a tool made to analyse the "junction" of immunoglobulin and T cell receptor sequences. It analyses a given set of nucleotide sequences, gives their description, determines the V-GENE, D-GENE and J-GENE locations and their IMGT unique numerotation. The Web service parameter is a `<querySeqData>` element that contains the list of `<seqData>` elements having a `<nucSequence>` component.

Full text indexing using IMGT controlled vocabulary

This approach has been used to try to answer the question: *How can the IMGT-ONTOLOGY vocabulary be used to index literature documents?* Relations between groups of IMGT terms and groups of terms coming from literature references have been established using sequence data. The method consists in chaining the following steps:

1. use IMGT-ML query to get sequences from IMGT/LIGM-DB database
2. get literature references (Medline number) from EMBL database (EBI SOAP server)
3. get literature abstracts from NCBI (eFetch service encapsulated in a Web server)
4. extract nominal syntagmas from abstracts (home developed Web server)
5. choose eligible nominal syntagmas (home developed Web server)
6. use these syntagmas to query general purpose search engines (for example the Google Web server)

Each step is assumed by a specific Web service, either home developed or publicly provided. The data are exchanged using specific XML data streams (not IMGT-ML, of course). The Web service SEFID (Search Engine For Immunology Domains) has been developed to chain each of them using the Orchestration paradigm (http://www.w3.org/2001/03/WSWS-popa) as its basis.

Conclusions

IMGT-ML is a formalization of IMGT-ONTOLOGY which appears to be useful, not only to distribute sequence data and knowledge, but also as a query language to retrieve sequence data, literature references and documents using IMGT controlled vocabulary.

Acknowledgements

Many thanks to the IMGT team for its expertise and its help during the construction of IMGT-ML. This work has been funded by the European Union's 5th PCRDT programmes: QLG2-2000-01287 and IST-2001-32688 (ORIEL project), the "Centre National de la Recherche Scientifique" (CNRS), the "Ministère de la Recherche" and the "Ministère de l'Education Nationale", ACI IMPBIO (IMPB82), Réseau National des Génopoles.

References

1. Lefranc M.-P., Giudicelli V., Kaas Q., Duprat E., Jabado-Michaloud J., Scaviner D., Ginestoux C., Clément O., Chaume D., Lefranc G. (2005) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res, 33, D593-D597.

2. Lefranc M.-P., Clément, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D. and Lefranc, G. (2005). IMGT-Choreography for Immunogenetics and Immunoinformatics. E pub In Silico Biology 5 0006 http://www.bioinfo.de/isb/2004/05/0006/ 24 December 2004.

3. Giudicelli V., Lefranc M.-P. (1999) Ontology for immunogenetics: IMGT-ONTOLOGY. Bioinformatics, 15, 1047-1054.

4. Chaume, D., Giudicelli, V. and Lefranc, M.-P. (2001). IMGT-ML a language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. In: CORBA and XML: Towards a bioinformatics integrated network environment, Proceedings of NETTAB 2001, Network tools and applications in biology, pp. 71-75.

5. Chaume, D., Giudicelli, V., Combres, K. and Lefranc, M.-P. (2003). IMGT-ONTOLOGY and IMGT-ML for Immunogenetics and immunoinformatics. In: Abstract book of the Sequence databases and Ontologies satellite event, European Congress in Computational Biology ECCB'2003, pp. 22-23.

6. Lefranc M.-P. Clément O., Kass Q., Duprat E., Chastellan P., Coelho I., Combres K., Genestoux C., Giudicelli V., Chaume D. and Lefranc G. (2005). IMGT-Choreography for Immunogenetics and Immunoinformatics . In Silico Biology 5 (2005) 1-16. http://www.bioinfo.de/isb/2004/05/0006, Ontology Workshop Göttingen 2004.

7. Chaume D., Giudicelli V., Combres K., Genestoux C. and Lefranc M.-P. (2005). IMGT-Choreography: processing of complex immunogenetics knowledge, CMSB 2004, Paris May 26-28 2004, Lecture Notes in Computer Science Springer-Verlag GmbH, ISSN: 0302-9743, vol 3082/2005, pp. 73-84.