## **INTUITION Tutorial for BioGateway**

1.	Inti	oduction	1
2.	De	sign	1
3.	Va	riables and properties	2
		w to build a query	
5.	Da	ta filtering and other possible operations	5
	5.1.	Filters of properties	5
	5.2.	Count and display unique results	9
	5.3.	Multiple values	10
	5.4.	Creating and filtering variables	10
	5.5.	Union of queries	11
6.	Us	e Cases	13
7.	Us	eful notes	13
	7.1.	Numbering of variables	13
	7.2.	Selection of object properties	14

#### 1. Introduction

INTUITION (<a href="http://semantics.inf.um.es/intuition">http://semantics.inf.um.es/intuition</a>) is a web application for user friendly SPARQL query building. In this way, users can exploit RDF knowledge graphs without advanced knowledge in SPARQL query language.

INTUITION analyses the knowledge network of an accessible endpoint, in this case BioGateway (<a href="http://ssb4.nt.ntnu.no:23122/sparql">http://ssb4.nt.ntnu.no:23122/sparql</a>), and allows building biological queries graphically by representing biological entities (nodes) related through properties (edges).

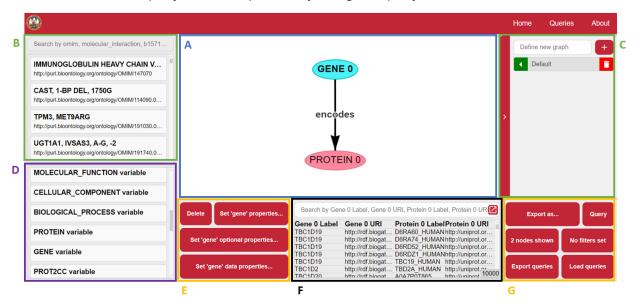
RDF knowledge graphs represent entities through Uniform Resource Identifiers (URIs), and information through triples or statements that represent a directional relationship between two entities, similar to a sentence: <subject> cyredicate> <object>. For example: <Gene> <encodes> <Protein>. The SPARQL query language also uses this pattern to build queries. These queries can be complex by linking multiple triples and including operations. A tutorial for building SPARQL queries in BioGateway (BGW) is available in this repository (https://github.com/juan-mulero/cisreg). INTUITION uses this same design for query development.

#### 2. Design

In INTUITION we distinguish different sections:

A- Query building canvas.

- B- Entity browser (Currently deactivated).
- C- Union buider (Used for queries that use union clauses).
- D- Variable browser (Main types of entities in the network).
- E- Properties selectors:
  - Object properties: relations between entities. Their inclusion acts as a filter.
  - Data properties: attributes of the entities. Their inclusion acts as a filter.
  - Optional properties: relations between entities. Their inclusion does not act as a filter (optional).
- F- Output display.
- G- Query builder:
  - Query: runs the query.
  - Export as: exports the query results.
  - Nodes shown: output editor. Allow to indicate which variables are shown in the output screen, and to create bindings, i.e., to define variables.
  - Filters set: allow adding filters.
  - Export query: export the designed query.
  - Load query: to load a previously designed query.



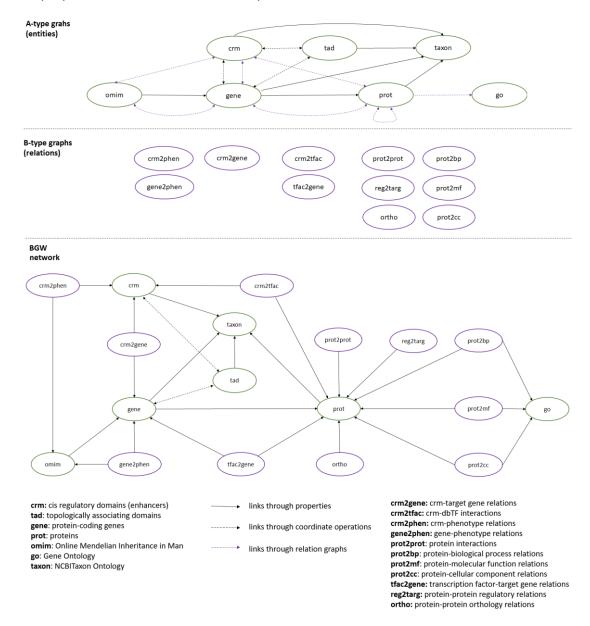
#### 3. Variables and properties

#### Variables:

- CRM variable: cis regulatory module.
- Gene variable: genes.
- Protein variable: proteins.
- OMIM variable: entities from OMIM ontology (mainly phenotypes).
- Molecular interaction: entities from Molecular Interactions ontology (MI).
- crm2gene variable: relation between CRM and gene.
- gene2phen variable: relation between gene and phenotype.
- crm2phen variable: relation between CRM and phenotype.
- crm2tfac variable: relation between CRM and protein (transcription factor).
- Transcription factor variable: transcription factors (currently only proteins that interact with CRM).
- reg2targ variable: regulatory relation between proteins.

- prot2prot: molecular interaction relation between proteins.
- tfac2gene variable: relation between gene and protein.
- Database variable: databases.
- Chromosome variable: chromosomes.
- Reference\_genome variable: genome assembly.
- TAD variable: topologically associated domain.
- Cellular component variable: cellular components from Gene Ontology (GO).
- prot2cc variable: relation between protein and its cellular components.
- Molecular function variable: molecular functions from GO.
- prot2mf: relation between protein and its molecular functions.
- Biological process variable: biological processes from GO.
- prot2bp variable: relation between protein and its biological processes.
- Ortho variable: orthology relation between proteins.
- Root variable: top hierarchically class of NCBITaxon Ontology.
- Taxonomic\_rank variable: top hierarchically class of NCBITaxon Ontology.

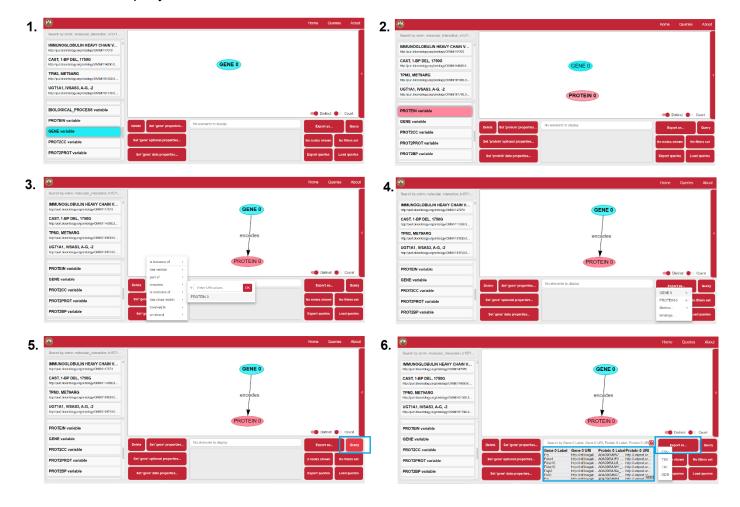
The properties are detailed with examples and their domains here.



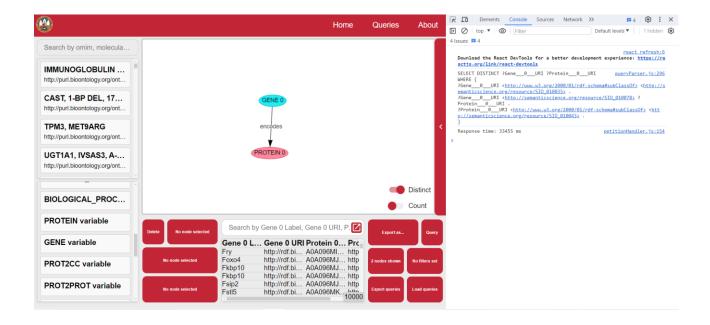
### 4. How to build a query

The query building process involves linking entities (nodes) through properties (edges). We take as an example the previous case: <Gene> <encodes> <Protein>.

- 1. Select the first entity (subject node), in this case, Gene, in the Variable browser.
- 2. Select the second entity (object node), in this case, Protein, in the Variable browser
- 3. Select the relation between both entities in the Properties selectors, in this case, "encodes".
- 4. Select in "Nodes shown" the data you want to show in the output.
- 5. Click on "Query" to launch the query.
- 6. Click on "Export as" to download the data. Click on "Export query" to save the query.



The generated SPARQL query can also be found by accessing the Console.

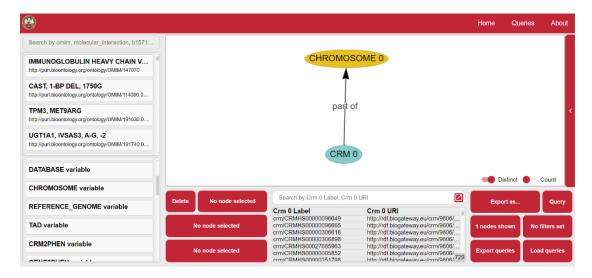


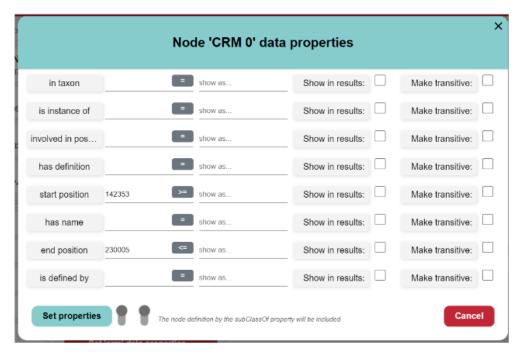
### 5. Data filtering and other possible operations

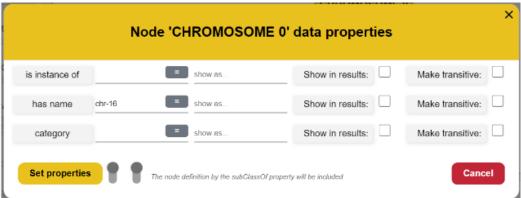
### 5.1. Filters of properties

Linking two biological entities or variables by their relation (object property) is the simplest filtering. However, any biological entity can also be filtered by its characteristics or attributes (data properties). The section "Property selectors" enables this task.

Clicking on "data properties" we can also filter attributes. If the value is a string, we can use the operator "=" to indicate an exact value, or the operator " $\subseteq$ " to indicate a substring contained in the string. In addition, to make filtering easier for the user, the default configuration of INTUITION is not case-sensitive. If the value is numeric, we can use the operators =, >,  $\geq$ , <,  $\leq$ . For example, we filter the variables chromosome and CRM to obtain the enhancers that overlap at certain coordinates.





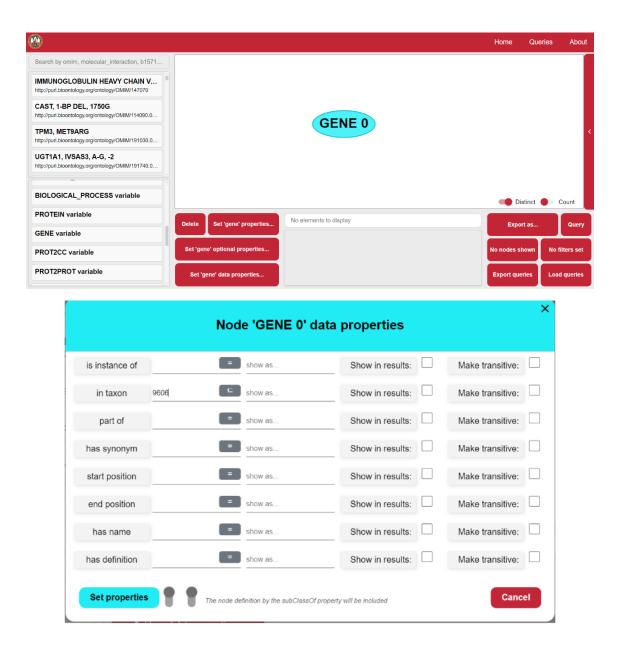


Variables can also be filtered clicking on "Filters set" button.

Some additional examples are given below:

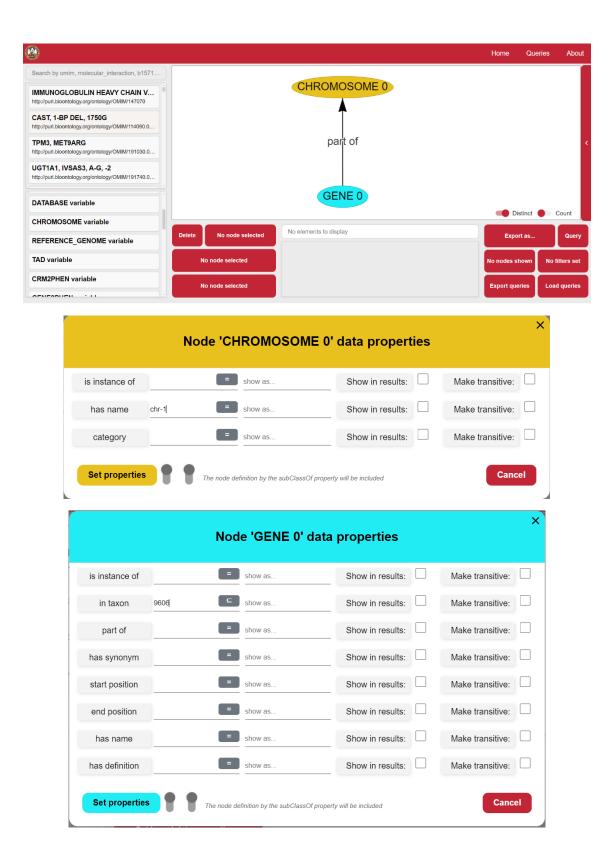
- **Example 1**: Filtering by taxon.

The resources in RDF are represented by URIs, so they must be used when the filtering involves a resource. This is the case for filtering by taxon (*in taxon* property). Since operating with URIs can be tedious, e.g. http://purl.obolibrary.org/obo/NCBITaxon\_9606, the "content in" or "⊆" operator makes it easier to work with identifiers. The following example illustrates the filtering of human genes.



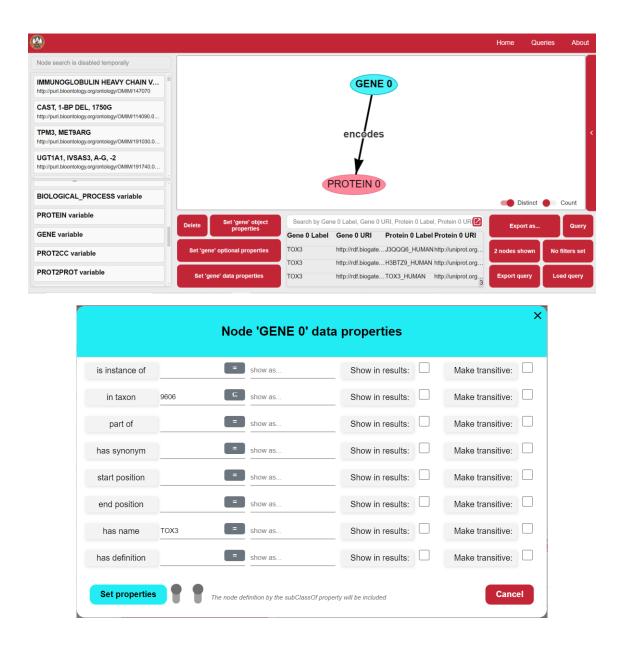
#### - **Example 2**: Filtering by chromosome.

Chromosomes are resources that have labels. Therefore, chromosomes can be filtered through their labels. Strings can be filtered in INTUITION using the "=" (exact value) or " $\subseteq$ " (contained in) operators. The default configuration of string filtering is not case-sensitive. The following example filters human genes on chromosome 1.



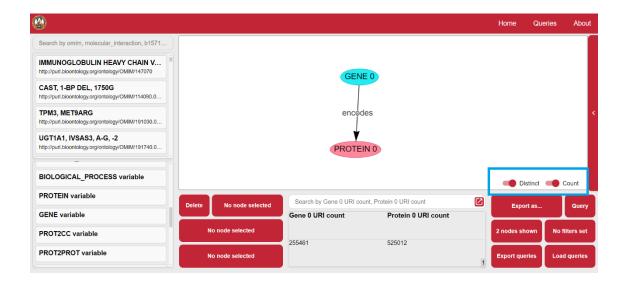
- **Example 3**: Filtering by name.

For this example we illustrate the query building to obtain the proteins encoded by the human TOX3 gene.



### 5.2. Count and display unique results

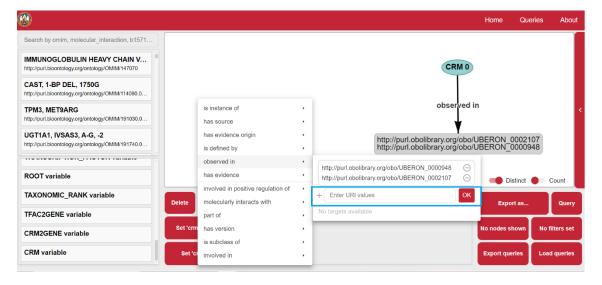
The output table shows the biological entities consulted. Since the user has the freedom to choose which entities want to include in the output ("Nodes shown" button), the result can include duplicate entities. The Distinct button, activated by default, removes these duplicates. The Count button allows you to count the results obtained.



### 5.3. Multiple values

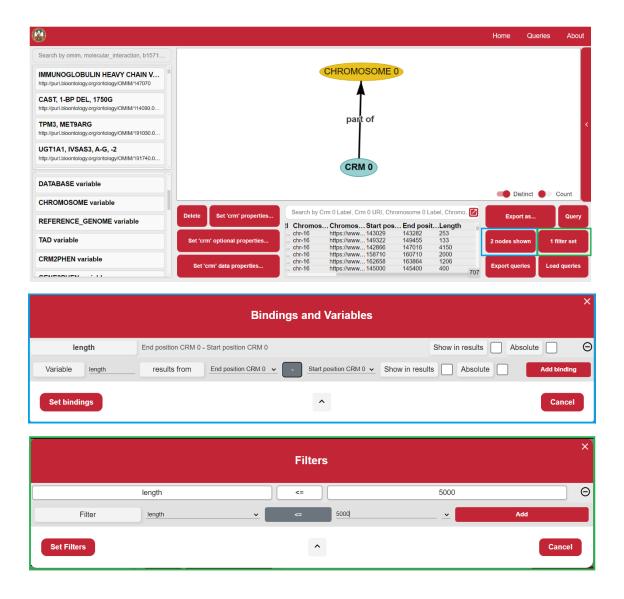
INTUITION also permits the assignment of one or more values to a specific variable. For example, cis regulatory modules identified in two tissues of interest. First we indicate the subject node (CRM variable), then we include in the corresponding object property (observed in) the values to be included. Click on "+" to include values and click on OK when all values are listed.

Since RDF resources are represented through URIs, the input values for filtering must be URIs. We are working on the use of labels for the next release version.



#### 5.4. Creating and filtering variables

INTUITION supports the generation and filtering of new variables. This functionality is implemented in "Nodes shown" > "Bindings" button. For example, by subtracting the end and start positions of the CRMs we obtain the length of the sequences in a new variable. Then, we can filter this new variable in the "Filters set" button.



### 5.5. Union of queries

INTUITION also allows the use of the UNION clause of SPARQL. UNION merges subqueries through common variables in both queries. We illustrate its use through a use case. For example, we retrieve the OMIM entities that contain the string "breast cancer" as a name or synonym. To do that:

- 1. In the Graph builder section, we create the graphs belonging to each of the subqueries, and we include an OMIM node in each of them. In the example, the graphs are "prefLabel" for the main label query and "altLabel" for the query of the synonym.
- 2. In each of the graphs we define the variable "label" according to the appropriate dataproperties ("has name" and "has synonym" properties, respectively). For this, we use the "show as" functionality, which enables to rename the variables, in this case under the common variable called "label".
- 3. We return to the main graph where we will join the two subqueries. For this, we include the subgraphs clicking on the green flap of each of the subgraphs.
- 4. Select one of the subgraphs represented as nodes and click on "Define union".

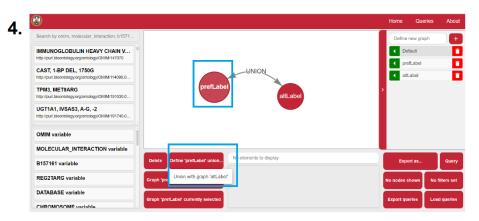
- 5. In "Node shown" we select the variables to be shown and in "Filters set" we filter the variable "label".
- 6. Run the query (Query).



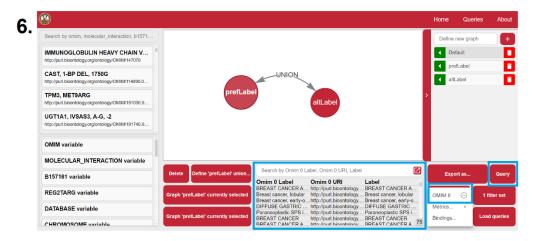












#### 6. Use Cases

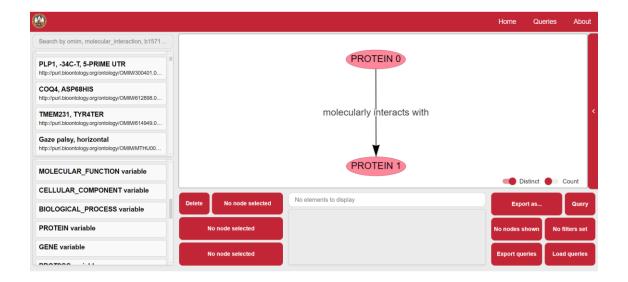
The following Use Cases were developed in the paper "Analysis of the landscape of human enhancer sequences in biological databases". The corresponding queries are attached for reproducibility and as examples of use. These use cases include complex queries that connect multiple nodes, use different filters, create variables, and join queries, so we recommend their consultation for a deeper understanding of the concepts introduced here for the graphical query building.

- 1. Use case 1: json files to load here.
- 2. Use case 2: json files to load here.
- 3. Use case 3: json files to load here.

#### 7. Useful notes

### 7.1. Numbering of variables

Queries can involve different variables. These queries can require the use of an entity as different variables, for example the protein-protein interaction involves two different proteins, and therefore two different variables. INTUITION permits to include the same variable more than once. These nodes are listed starting from 0. The counter is reset when the application is refreshed, not when the query is deleted.



# 7.2. Selection of object properties

In order to help the user select appropriate links between biological entities (nodes introduced by variables), INTUITION only allows to connect two nodes according to their appropriate property (object property). That is, INTUITION does not allow you to link two entities through an incorrect property. For example, if the variables Gene and Protein are introduced, INTUITION only allows connecting both variables through the "encodes" property.

