INTUITION Tutorial

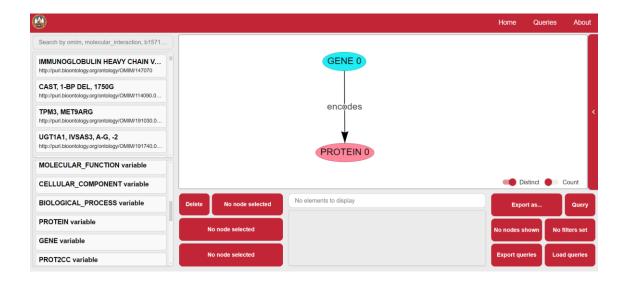
Intro	oduction	1
Des	sign	2
Var	iables and properties	3
Hov	v to build a query	4
Filte	ers and other clauses	6
5.1.	DISTINCT and COUNT	6
5.2.	VALUES	6
5.3.	Filters of properties	7
5.4.	Creating and filtering variables	7
5.5.	UNION	8
Use	Cases	10
7. U	seful notes	10
7.1.	Filtering by taxon	10
7.2.	Filtering by chromosome	11
7.3.	Numbering of variables	11
	Des Var Hov Filte 5.1. 5.2. 5.3. 5.4. 5.5. Use 7.1.	Introduction

1. Introduction

INTUITION 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 (http://ssb4.nt.ntnu.no:23122/sparql), and allows building queries graphically by representing nodes (entities) and edges (properties).

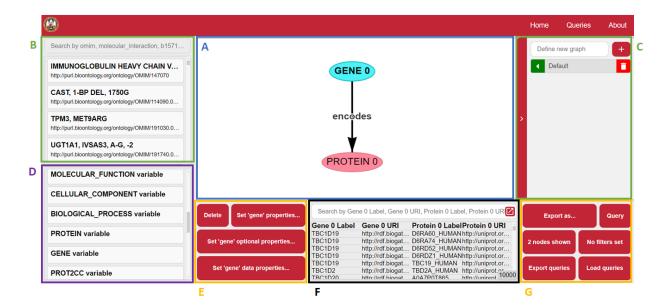
RDF knowledge graphs represent entities through uniform resource identifiers (URIs), and information through triples or statements that represent a directional relationship between two entities, similar to a sentence: <subject> cyredicate> <object>. For example: <Gene> <encodes> <Protein>. The SPARQL 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 screen.
- B- Entity browser (Currently deactivated).
- C- Graph builder (Used for queries that use union clauses).
- D- Variable selection (Main types of entities in the network).
- E- Properties selection:
 - 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 screen.
- 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.

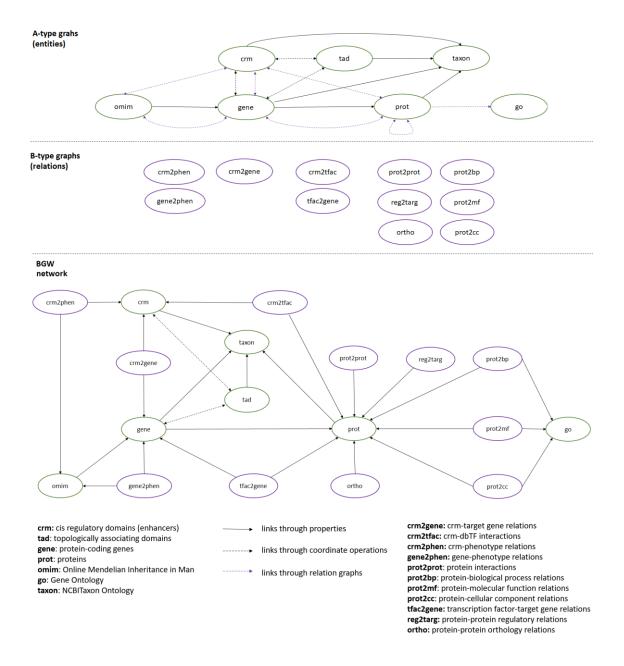


3. Variables and properties

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.

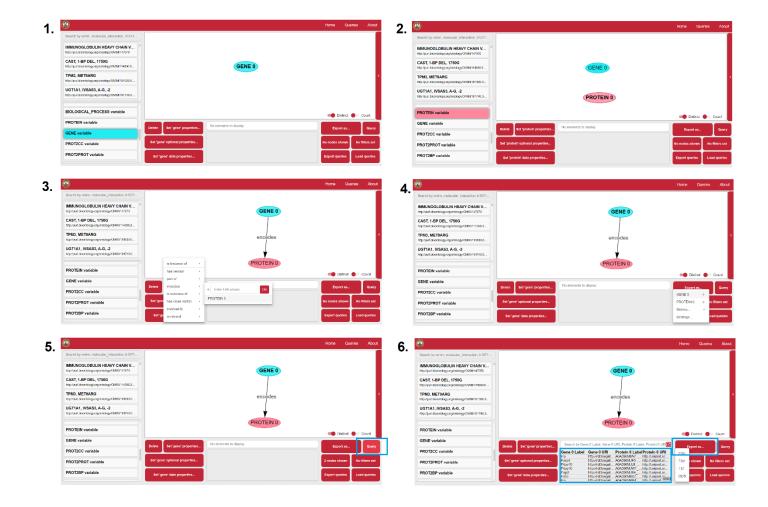
The properties are detailed with examples and their domains here.



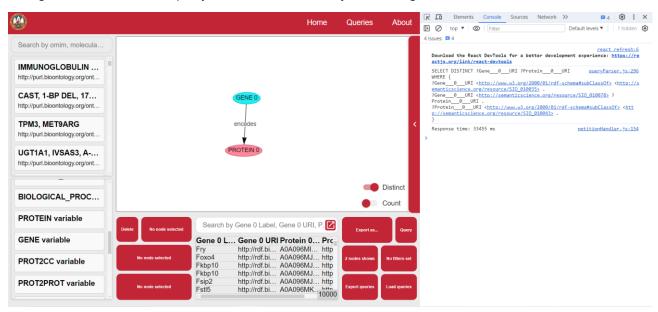
4. How to build a query

We take as an example the previous case: <Gene> <encodes> <Protein>.

- 1. Select the first node or subject, in this case Gene, in the variables section.
- 2. Select the second node or object, in this case Protein, in the variables section.
- 3. Select the object property that relates both entities, in this case "encodes".
- 4. Select in "Nodes shown" the data we want to show in the output, in this case Gene and Protein.
- 5. Click on "Query" to launch the query.
- 6. Click on "Export as" to download the data obtained. Click on "Export query" if you want to save the query as well.



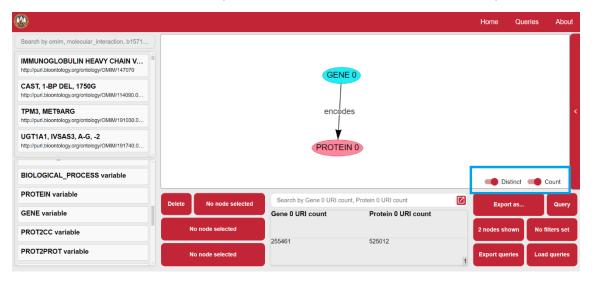
The generated SPARQL query can also be found by accessing the Console.



5. Filters and other clauses

5.1. DISTINCT and COUNT

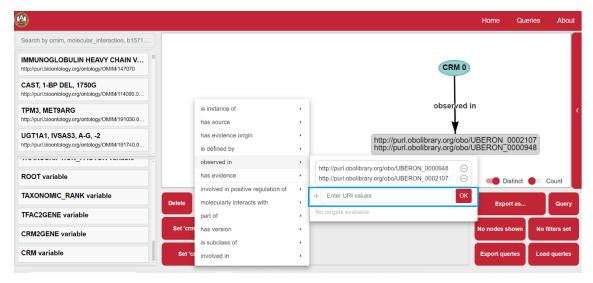
By default, the queries include the DISTINCT clause to avoid duplicate results. This can be desactivated clicking on the Distinct button, in the Query building screen. You can also activate the Count button if you want to include this clause in the query.



5.2. VALUES

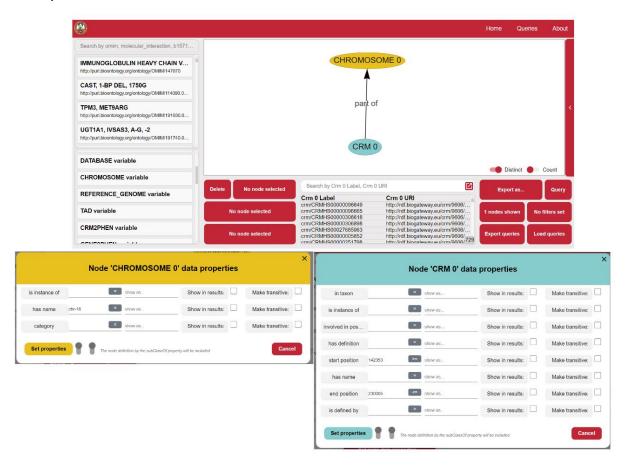
This clause permits the assignment of one or more values to a specific variable. For example, cis regulatory modules identified in two tissues of interest. First we indicate the subject node (CRM variable), then we include in the corresponding object property (observed in) the values to be included. 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.



5.3. Filters of properties

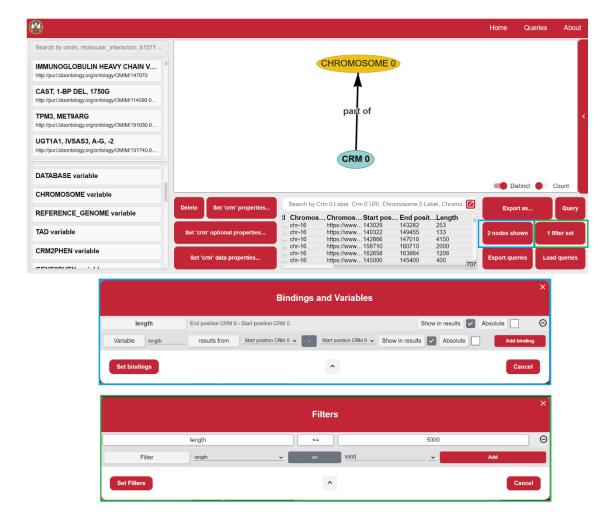
Clicking on "data properties" we can also filter attributes. If the value is a string we can use the operator "=" to indicate an exact value, or the operator " \subseteq " to indicate a substring contained in the string. If the value is numeric we can use the operators =, >, \geq , <, \leq . For example, we filter the variables chromosome and CRM to obtain the enhancers that overlap at certain coordinates.



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

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

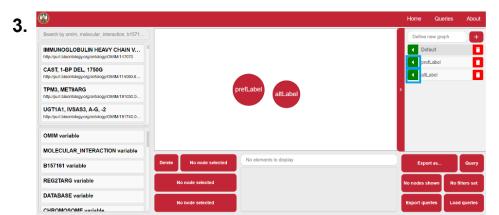
INTUITION also allows the use of the UNION clause of SPARQL. UNION merges subqueries through common variables in both queries. We illustrate its use through a use case. For example, we retrieve the OMIM entities that contain the string "breast cancer" as a name or synonym. To do that:

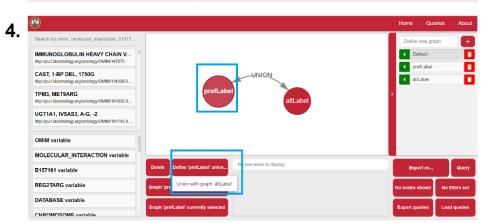
- 1. In the Graph builder section, we create the graphs belonging to each of the subqueries, and we include an OMIM node in each of them. In the example, the graphs are "prefLabel" for the main label query and "altLabel" for the query of the synonym.
- 2. In each of the graphs we define the variable "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.
- 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).



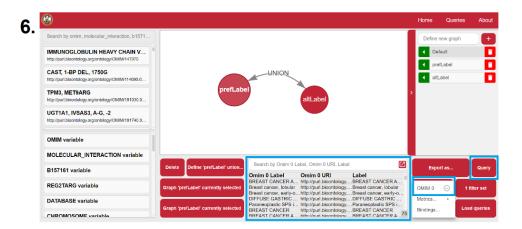












6. Use Cases

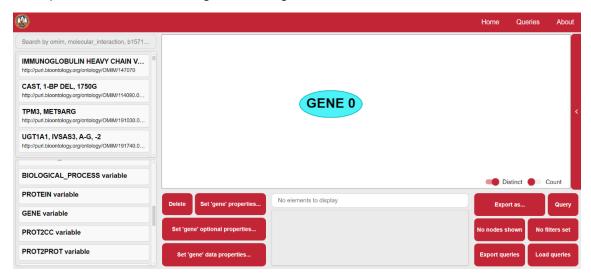
The following Use Cases were developed in the paper "Analysis of the landscape of human enhancer sequences in biological databases". The corresponding queries are attached for reproducibility and as examples of use.

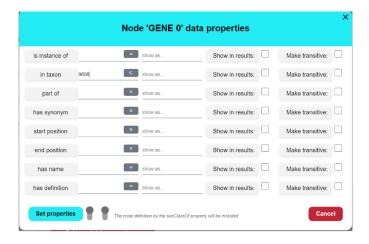
- 1. Use case 1: json files to load here.
- 2. Use case 2: json files to load here.
- 3. Use case 3: json files to load here.

7. Useful notes

7.1. Filtering by taxon

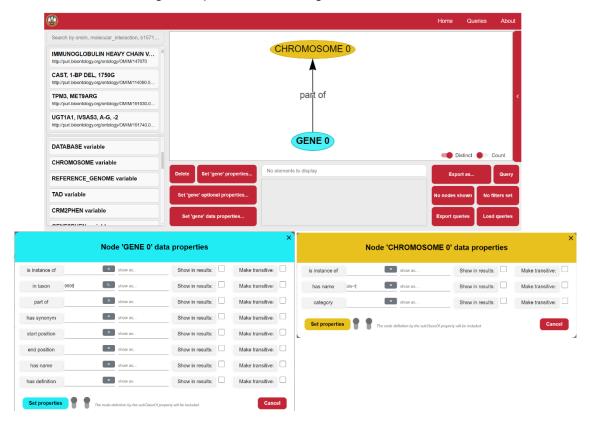
The resources in RDF are represented by URIs, so they must be used when the filtering involves a resource. This is the case for filtering by taxon (in taxon property). Since operating with URIs can be tedious, e.g. http://purl.obolibrary.org/obo/NCBITaxon_9606, the "content in" or "⊆" operator makes it easier to work with identifiers. The following example illustrates the filtering of human genes.





7.2. 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 casesensitive. The following example filters human genes on chromosome 1.



7.3. 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.

