# CSCG: Conceptual Schema of the Citrus Genome

Technical Report

Alberto García S., Oscar Pastor

UNIVERSITAT POLITÈCNICA DE VALÈNCIA




| | |
|---|---|
| Ref # | PROS-TR-2020-1 |
| Title | CSCG: Conceptual Schema of the Citrus Genome |
| Author(s) | Alberto García S., Oscar Pastor |
| Corresponding author(s) | algarsi3@pros.upv.es, opastor@pros.upv.es |
| Document version number | 1 |
| Final version | - |
| Release date | - |
| Keywords | CSCG, Conceptual Schema, Evlution, Citrus Genome |

CONTENTS



We describe our proposed Conceptual Schema (CS) to work with Citrus genome information (CSCG). The presented CS is being used in a real-world industrial case to validate it and gather expert domain feedback. The CSCG is divided into two different perspectives, namely, the biological and the technological perspective.

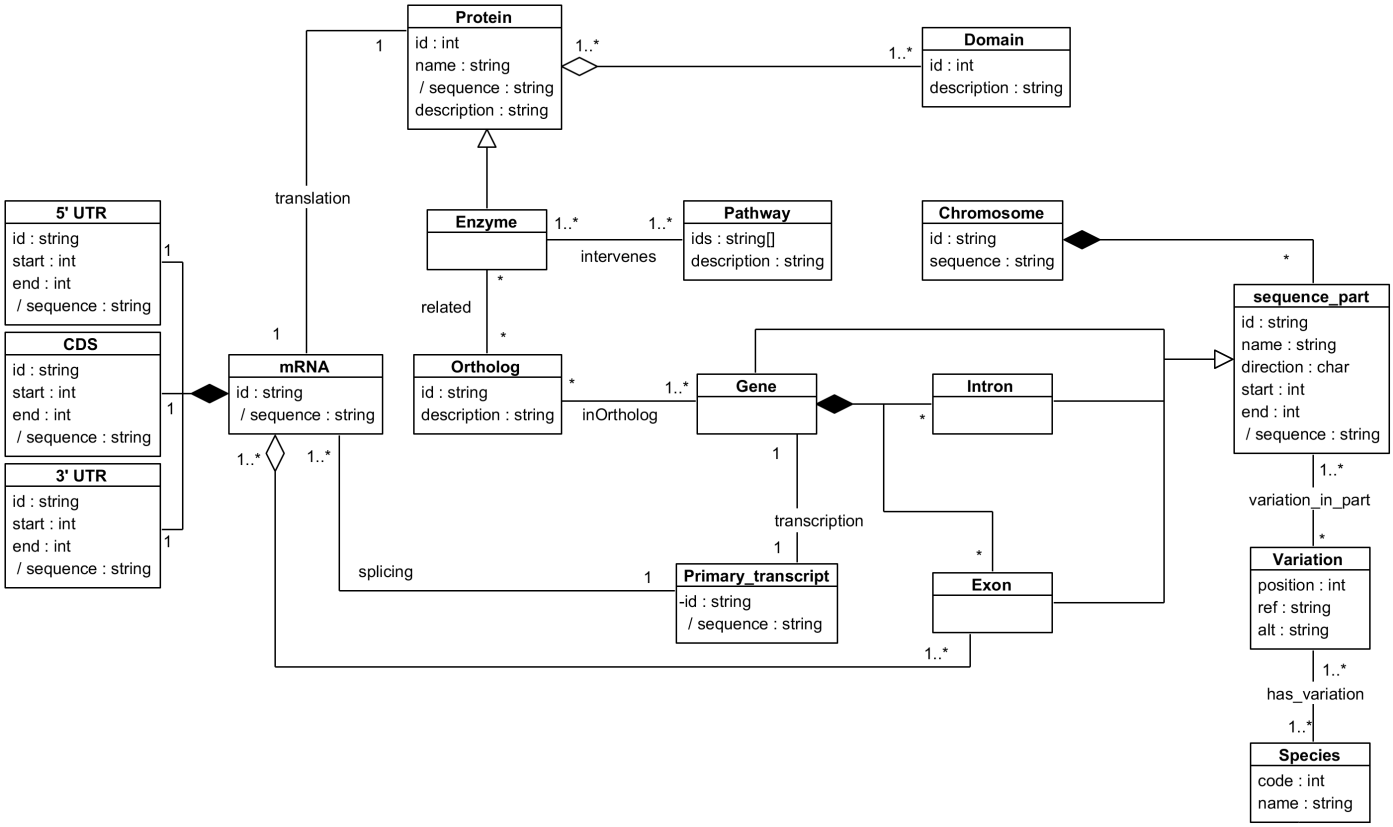
The biological perspective deal with pure biological concepts and its mission is to understand how life works and what are its internals. 17 classes have been defined in the biological perspective: Chromosome, Sequence_part, variation, Species, Gene, Intron, Exon, Primary_transcript, mRNA, 5' UTR, CDS, 3' UTR, Protein, Domain, Enzyme, Ortholog, Pathway. The biological perspective can be found in Annex A.

The technological perspective tries to integrate the biological perspective with the real world. Technology has limitations, knowledge is very limited and real-world genomic information needs to be analysed and understood as a separate problem. We define independent knowledge blocks that can be plugged into the biological perspective to enrich it. The first version of the CSHG defines four different blocks:

- Scaffold Module: real-life sequencing is far from perfect. It is not possible to correctly obtain the genome sequence. Therefore, gaps are encountered.
  - The scaffold chromosome represents the chromosome whose sequence is stored in the linked scaffold.
- Gene Otology (GO) Module: GO ontology tries to provide a standard way to define biological processes, molecular functions that are part of biological processes and cellular components where they are carried on. Then, links genes or gene products to the previously defined concepts.
  - GO consider genes and proteins functional elements, therefore, they are specialisation types of functional elements.
- Variant Call Format (VCF) Module: VCF files are used to report identified variations when comparing a given sequence to a sequence of reference. Along with variations, there is plenty of additional information like quality attributes or genotype information.
  - Variations of the VCF are linked to biological variations. This way, a variation can be identified in multiple VCF files and with different quality attributes.
  - Each sample of a VCF file will be a specie.
- SnpEff Module: variations can be automatically annotated with the predicted effect that will cause.
  - Since annotations are included to VCF files, we link each annotation to a VCF variation. A variation can have zero or multiple annotations and an annotation will belong to only one VCF variation. An VCF variation present in different VCF files may have different annotations depending on the parameters of SnpEff when annotating the VCF file.
  - ANN value can be linked to primary transcripts and genes. The reason is that the SnpEff standard has two field to link the annotation (effect caused by the variation) to a element. The element has to be either a gene or a primary transcript. Gene cardinality is two at most because a VCF variation is linked to two genes when it is annotated as intergenic.

The technological perspective can be found in Annex B. By joining the biological perspective with the defined knowledge blocks in the technological perspective, the final schema arises. It can be found in Annex C

# APPENDIX A
## CSCG: BIOLOGICAL PERSPECTIVE



## APPENDIX B
## CSCG: TECHNOLOGICAL PERSPECTIVE

### *A. Scaffold Module*

ID: 0 Description: Scaffold sequencing technology.

Fig. 1. Scaffold

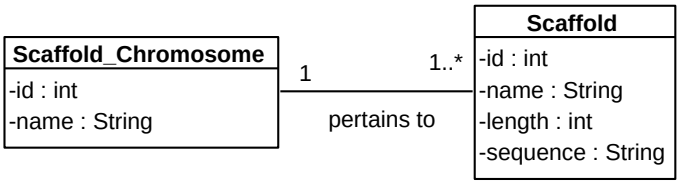



TABLE I. SCAFFOLD KNOWLEDGE ELEMENTS

| **Knowledge chains** | | | | |
|---|---|---|---|---|
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| - | - | - | - | - |
| **Knowledge equalities** | | | | |
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| 0.e.0 | Scaffold_Chromosome is ontologically equivalent to biological chromosome | Scaffold_Chromosome | Biological Oracle | Chromosome |

*B. GO Module*

ID: 1 Description: Gene Ontology knowledgebase.

Fig. 2. Gene Ontology

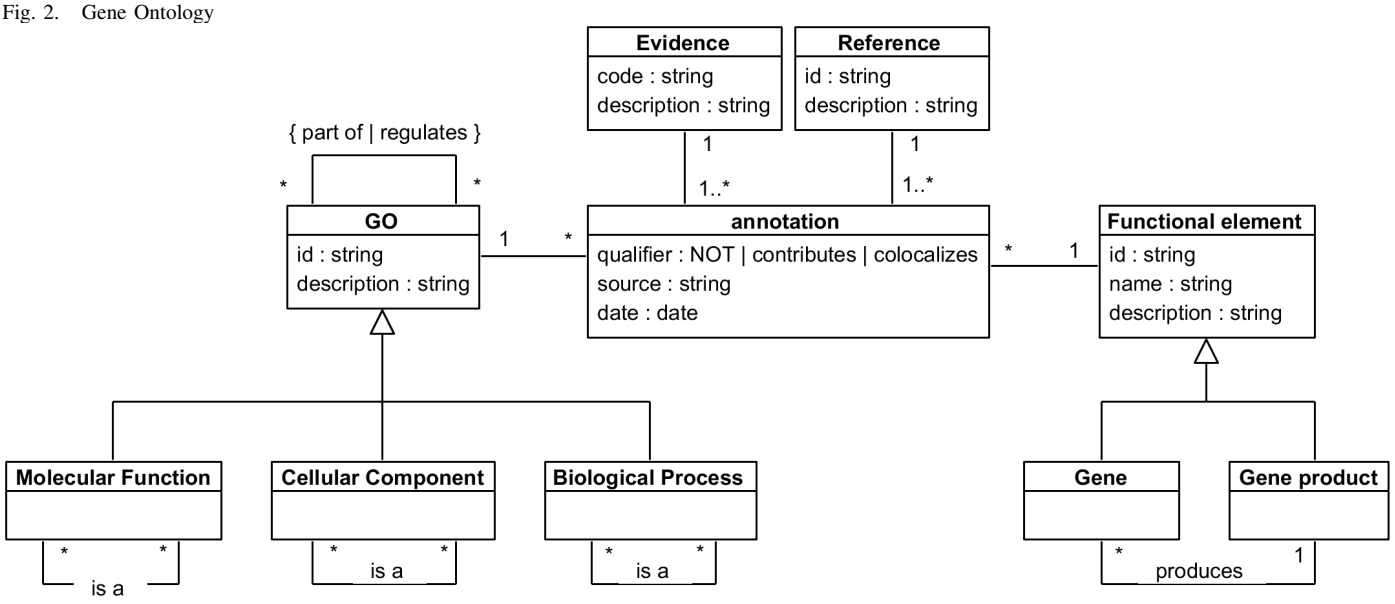



TABLE II. GENE ONTOLOGY KNOWLEDGE ELEMENTS

| **Knowledge chains** | | | | |
|---|---|---|---|---|
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| - | - | - | - | - |
| **Knowledge equalities** | | | | |
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| 1.e.0 | Go Gene is ontologically equivalent to biological gene | GO Gene | Biological Oracle | Gene |
| 1.e.1 | Gene Product is ontologically equivalent to proteins | Gene Product | Biological Oracle | Protein |

### C. VCF Module

ID: 2 Description: Variant Call Format file format.

Fig. 3. VCF

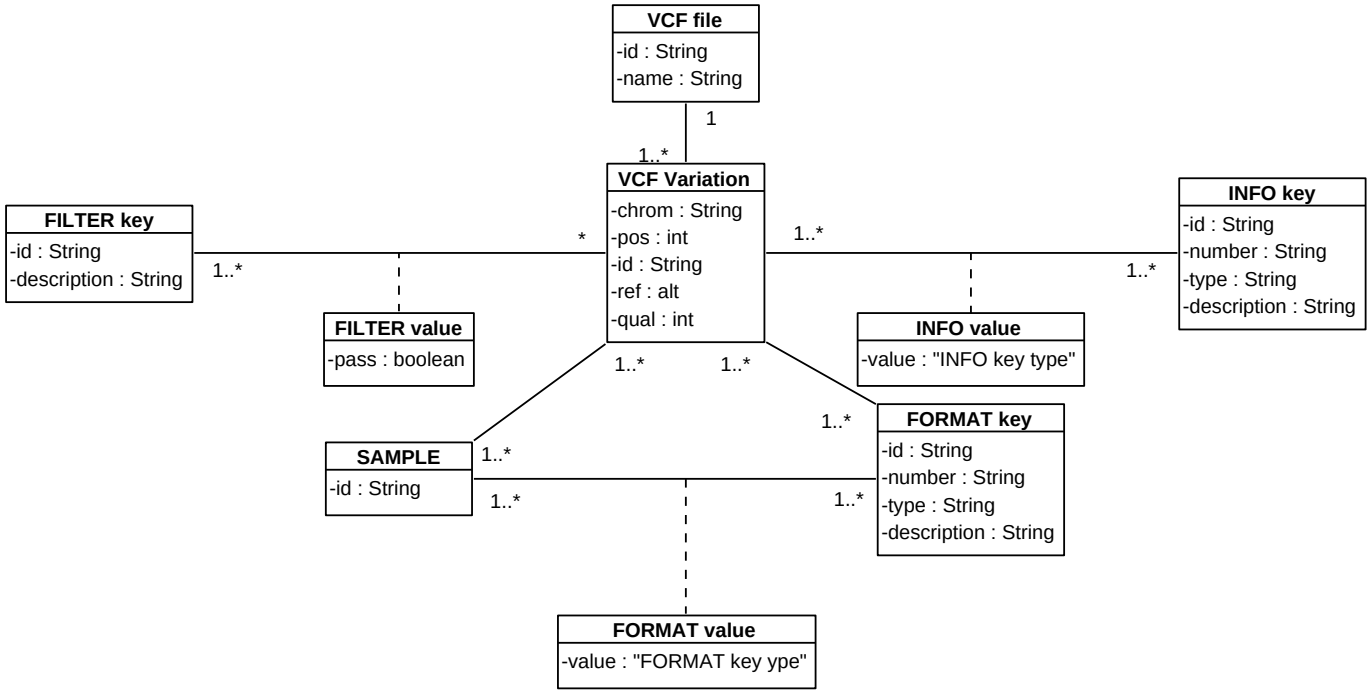



TABLE III. VCF KNOWLEDGE ELEMENTS

| **Knowledge chains** | | | | |
|---|---|---|---|---|
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| 2.c.0 | VCF Variation is linked to biological variation | VCF Variation [1..*] | Biological Oracle | Variation [1] |
| 2.c.1 | Sample is linked to biological species | Sample [*] | Biological Oracle | Species [1] |
| **Knowledge equalities** | | | | |
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| 2.e.0 | VCF Variation is ontologically equivalent to ANN VCF Variation | VCF Variation | ANN Knowledge Block | ANN VCF Variation |
| 2.e.1 | VCF file is ontologically equivalent to ANN VCF file | VCF file | ANN Knowledge Block | ANN VCF file |

*D. ANN Module*

ID: 3 Description: ANN field inside Variant Call Format file format.

Fig. 4. ANN

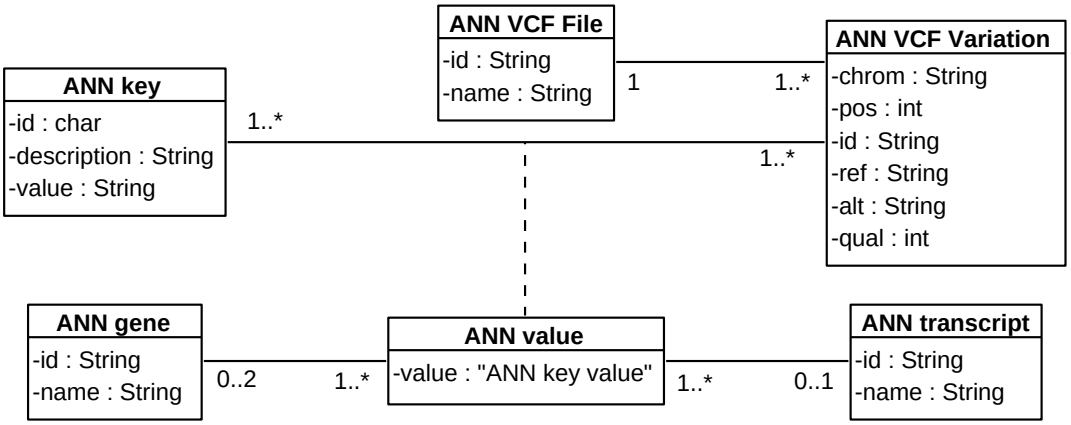



TABLE IV. ANN KNOWLEDGE ELEMENTS

| **Knowledge chains** | | | | |
|---|---|---|---|---|
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| - | - | - | - | - |
| **Knowledge equalities** | | | | |
| Knowledge Block | | | Output CS | |
| ID | Description | Input Class | CS | Class |
| 3.e.0 | ANN VCF Variation is ontologically equivalent to VCF Variation | VCF Variation | VCF Knowledge Block | VCF Variation |
| 3.e.1 | ANN VCF file is ontologically equivalent to VCF file | VCF file | VCF Knowledge Block | VCF file |
| 3.e.2 | ANN gene is ontologically equivalent to biological gene | ANN gene | Biological Oracle | gene |
| 3.2.3 | ANN transcript is ontologically equivalent to biological Primary_transcript | ANN transcript | Biological Oracle | Primary_transcript |

# APPENDIX C
# CSCG: FINAL CS