



## **Annotation Guidelines BIONLP-ST 2016 SeeDev task**

Estelle Chaix, Bertrand B. Dubreucq, Dialekti Valsamou, Abdelhak Fatihi,  
Louise Deleger, Robert R. Bossy, Pierre Zweigenbaum, Philippe Bessières,  
Loïc L. Lepiniec, Claire Nédellec

### **► To cite this version:**

Estelle Chaix, Bertrand B. Dubreucq, Dialekti Valsamou, Abdelhak Fatihi, Louise Deleger, et al..  
Annotation Guidelines BIONLP-ST 2016 SeeDev task. 2016, pp.47. hal-02795594

**HAL Id: hal-02795594**

**<https://hal.inrae.fr/hal-02795594>**

Submitted on 5 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Annotation Guidelines

## BIONLP-ST 2016 SeeDev task

Campaign: Scientific literature on *Arabidopsis Thaliana*.

Estelle Chaix<sup>1</sup>, Bertrand Dubreucq<sup>2</sup>, Dialekti Valsamou<sup>1,3</sup>, Abdelhak Fatihi<sup>2</sup>, Louise Deléger<sup>1</sup>, Robert Bossy<sup>1</sup>, Pierre Zweigenbaum<sup>3</sup>, Philippe Bessi eres<sup>1</sup>, Lo ic Lepiniec<sup>2</sup>, Claire N edellec<sup>1</sup>

<sup>1</sup> MaIAGE, INRA – Jouy en Josas, France; <sup>2</sup> Institut Jean-Pierre Bourgin, INRA–Versailles, France; <sup>3</sup> LIMSI, CNRS, Orsay, France.

### [A. Annotation Scheme](#)

#### [A.1. Entities](#)

##### [A.1.1 Color scheme](#)

##### [A.1.2 Entity annotation](#)

##### [A.1.3 DNA Entities](#)

###### [Gene](#)

###### [Gene Family](#)

###### [Box](#)

###### [Promoter](#)

##### [A.1.4 Gene product entities](#)

###### [RNA](#)

###### [Protein](#)

###### [Protein Family](#)

###### [Protein Complex](#)

###### [Protein Domain](#)

##### [A.1.5 Other molecule entities](#)

###### [Hormone](#)

##### [A.1.6 Dynamic process](#)

###### [Regulatory Network](#)

###### [Pathway](#)

##### [A.1.7 Context](#)

###### [Genotype and Species](#)

###### [Tissue, Cell and Organ](#)

###### [Development Phase](#)

###### [Environmental Factor](#)

#### [A.2. Events](#)

##### [A.2.1 Required arguments of relations](#)

##### [A.2.2 Event type and color scheme](#)

- [A.2.3 Extra-properties of events](#)
- [A.2.4 Event for specific case](#)
- [A.2.5 “Where and When” events](#)
  - [Presence In Genotype](#)
  - [Occurrence In Genotype](#)
  - [Presence At Stage](#)
  - [Occurrence During](#)
  - [Localization](#)
- [A.2.6 “Function” events](#)
  - [Involvement In Process](#)
  - [Transcription Or Translation](#)
  - [Functional Equivalence](#)
- [A.2.7 “Regulation” events](#)
  - [Use of “Regulation” relations when there is a genotype involved](#)
  - [Regulation Of Accumulation](#)
  - [Regulation Of Expression](#)
  - [Regulation Of Development Phase](#)
  - [Regulation Of Molecule Activity](#)
  - [Regulation Of Process](#)
  - [Regulation Of Tissue Development](#)
- [A.2.8 “Composition and Membership” events](#)
  - [Primary Structure Composition](#)
  - [Protein Complex Composition](#)
  - [Protein Domain Composition](#)
  - [Family Membership](#)
  - [Sequence Identity](#)
- [A.2.9 “Interaction” events](#)
  - [Binding](#)
  - [Interaction](#)
- [A.2.10 Optional arguments of relations](#)
- [A.2.1 Extra properties](#)
  - [Negation](#)

## [B. Figures](#)

- [B.1 Entity types and Categories](#)
- [B.2 Relation classification](#)
- [B.3 Schematical representation of Arguments and Events](#)
  - [Presence In Genotype](#)
  - [Occurrence In Genotype](#)
  - [Presence At Stage](#)
  - [Occurrence During](#)
  - [Localization](#)
  - [Involvement In Process](#)
  - [Transcription Or Translation](#)
  - [Functional Equivalence](#)
  - [Regulation Of Accumulation](#)
  - [Regulation Of Development Phase](#)
  - [Regulation Of Expression](#)
  - [Regulation Of Molecule Activity](#)

[Regulation Of Process](#)  
[Regulation Of Tissue Development](#)  
[Primary Structure Composition](#)  
[Protein Complex Composition](#)  
[Protein Domain Composition](#)  
[Family Membership](#)  
[Sequence Identity](#)  
[Interaction](#)  
[Binding](#)  
[Visualisation of all relations](#)  
[Acknowledgement](#)

## A. Annotation Scheme

Parts of the document are declared as completely annotated by the `[Completely_Annotated]` label.

### A.1. Entities

Entities are all terms belonging to relevant predefined semantic types.

#### A.1.1 Color scheme

Here you can see the annotation scheme with the colors used in the Annotation Editor ([AlvisAE](#)). In the following, we use this color coding to highlight entity mentions, and refer to Annotated entities.

`[Completely_Annotated]` : In the `[Completely_Annotated]` text, all entities should be annotated, even if they do not participate in a relation.

#### Entity types:

**Molecule :** DNA : *Gene*, *Gene\_Family*, *Box*, *Promoter*

DNA product : *RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*,  
*Protein\_Domain*

Hormone: *Hormone*

**Dynamic Process :** *Regulatory\_Network*, *Pathway*

**Context :** *Tissue*, *Development\_Phase*, *Genotype*, *Environmental\_Factor*

In the following, we use the conventions defined below to distinguish between relation names, roles in relation and entities:

- Entities are noted in italic font with the color scheme previously described (e.g. *Tissue*), and in sentences are between brackets (e.g. `[Tissue]`)
- Relation names are noted in straight font with the color scheme described below (e.g. *Protein\_Domain\_Composition*),
- Roles of relation are noted in black and bold (e.g. **Tissue**)

#### A.1.2 Entity annotation

When an entity includes another, all possible entities are annotated with the relevant types (which may be different). For example, in “embryo development” we annotate the *[Development\_Phase]* (*embryo development*) as well as the *[Tissue]* (*embryo*). “*WRI1-like group*” is another example where “*WRI1*” is a *[Gene]* and “*WRI1-like*” a *[Gene\_Family]*. Embedded entities may be linked by relationships. The entity span may include in rare cases prepositional phrases or even relative clauses, if they are modifiers of the entity and if they contribute to specify its nature.

**Examples:**

- cotyledons that are initiated outside of the seed environment
- later stages of morphogenesis
- organs with embryonic features
- organs and tissues derived from double fertilization in flowering plants
- proteins which start to be expressed coordinately in the endosperm of maize 8 to 10 days after pollination .

**Annotation of words denoting an entity type:**

When entities are “named” (e.g. [Gene], [Protein]), the general words that denote the type of the entity (e.g. “gene”) should not be included in the annotation (the name itself is sufficient).

**Examples :** BCCP2 gene, AP2 protein

When entities do not have proper names (e.g. [RNA], [Protein Family], [Promoter]), words denoting the entity type (e.g. “mRNA”) should be included in the annotation when they are needed to identify the entity.

**Examples :** AGL16 mRNA (where “AGL16” is a [Protein]), AP2 family (where “AP2” is annotated as a [Protein Domain] and “AP2 family” the [Protein Family] that contain AP2 domain)

Entities should be noun or adjectival phrases, such as “FA biosynthesis” or “FA biosynthetic” (in “FA biosynthetic genes”).

Entities (in particular developmental phases) should include the preposition occurring before when it expresses temporal or spatial information, such as “before” or “after” (but not “during”) (e.g. AGL15 was present before embryogenesis).

**Discontinuity:**

Discontinuous entities are allowed when they are syntactically coherent. In particular, they are used for the annotation of coordinated structures.

**Example :** BCCP2 and PKp-β1 promoters

- “BCCP2 [...] promoters” is a discontinuous entity and should be annotated as [Promoter]

Discontinuity is not allowed between elements that are not in the same sentence or that are *too far* from each other in the sentence.

**Counter example :** “An analysis of genetic factors affecting *Arabidopsis* seed mass used a segregating population from a cross between small-seeded and large-seeded ecotypes to show that both maternal and nonmaternal QTL affect seed mass”

- “Arabidopsis [...] small-seeded” and “Arabidopsis [...] large-seeded” should NOT be annotated as a discontinuous entity

**Inclusion of entities**

All entities must be annotated with the correct type, even if they are included in a larger one.

**Example:** “central cell of female gametophyte”

- The whole term “central cell of female gametophyte” should be annotated as [Tissue]
- but also “female gametophyte” as [Tissue]
- and “gametophyte” as [Tissue]

Entities should be linked by a relation such as “Localization” : “central cell of female gametophyte” Localization “female gametophyte”

### Errors in the text:

When the typography (italic and case) is not consistent with the intended type of entities, entities should be annotated with their intended type (e.g. italic instead of plain text used for protein names as in “interaction between *WRI1* and the *BCCP2* promoter.”, “*WRI1*” is annotated as [Protein] and not as [Gene] )

Some sentences may contain errors from the authors (e.g. gene names used to refer to proteins, protein names used to refer to RNA). **These sentences should be excluded from the annotated paragraphs when possible.**

An exception is made for enumerations where the intended type of the entities is clearly stated in the context (e.g. “RNAs include ...”). In this case, annotation is permitted and entities should be annotated with their intended type (since it can easily be inferred from the context).

**Example :** These RNAs include 12S storage proteins, oleosin and steroleosin

- “12S storage proteins”, “oleosin” and “steroleosin” are annotated as [RNA] even though they are protein names.

## A.1.3 DNA Entities

### Gene

#### Description:

*Definition:* DNA sequence coding for a mRNA.

Gene names are in italic and uppercase. The [Gene] type includes artificial design, for instance with a reporter gene denoted : gene1::gene2 or promoter::gene.

It excludes overgeneral terms such as *these genes*. In this case, the antecedent should be annotated and linked to the relevant arguments (e.g. PHAP2A and AP2 extends throughout the whole protein, suggesting that these genes (...)).

Genes belonging to a set of genes that is not a family are annotated with the [Gene] type. However, plasmids should **not** be annotated (e.g. GAL4-DBD).

Full names of genes (e.g. APETALA2) and locus of gene (e.g. At4g38130) should also be annotated (and should be linked to their abbreviation (e.g. AP2) with a Sequence Identity relation).

**Examples:** *FUS3* ; *FUS3::GUS* ; *ABI3* ; *APETALA2* ; At4g38130

## Gene\_Family

### Description:

A family of genes mentioned by their common function, including coding for a same [Protein\_Family] (e.g. albumin genes), or their common ancestor. It should not be confused with proteins (e.g. B3 transcription factors).

Words that do not designate the [Gene\_Family] should be excluded from the name. For instance, in “three new MADS-box”, the words *three new* should be excluded.

**Examples:** LEC genes ; AP2-like ; albumin genes ; B3 transcription factors genes

### Counter examples:

GA-biosynthesis genes ; glycolytic biosynthetic genes ; genes involved in acquisition of desiccation tolerance ; B3 transcription factors

Databases:

<http://www.arabidopsis.org/browse/genefamily/index.jsp>

## Box

### Description:

**Definition:** A (short) DNA sequence that corresponds to a **binding site** for a “product” (the transcription machinery or a transcription factor). It also includes DNA elements such as “response elements”.

**Examples:** AFL targets ; TT targets ; WRI1 targets ; dehydration response elements ; AACA

Databases:

More on <http://arabidopsis.med.ohio-state.edu> , <http://www.athamap.de/>, or <http://bioinformatics.psb.ugent.be/ATCOECIS/>

## Promoter

### Description:

**Definition:** Upstream region of a gene that binds the polymerase for gene transcription.

It can be designated as a regulatory region of a gene. It may happen that it is not named, such as in the expression “promoters of At2S3” where “At2S3” is a gene name or in “5' flanking regions followed by promoter”. The expression “5' flanking region” **without any gene information** should not be annotated because it is too general and not informative. It may be a set of promoters, such as “albumin promoters”. The promoter name should not be confused with the gene name.

The expression “upstream of” should be annotated as a promoter **only** if it refers to the DNA region located before the gene: e.g. in “The WRI1 binding sites in the upstream regions of PI-PK  $\beta$  1” the expression should be annotated as promoter, but **not** in “GL2 acts upstream of MUM4” which refers to a genetic relationship.



**Examples:** *BCCP2* promoter ; *AGL15* regulatory regions ; 35s; 5' flanking regions of *LEC2*-induced genes ; promoters of seed storage protein genes

**Databases:**

<http://urgv.evry.inra.fr/projects/FLAGdb++/HTML/index.shtml>

### A.1.4 Gene product entities

#### RNA

##### Description:

*Definition:* Gene product.

an RNA entity is generally named with the gene name followed by the word “mRNA” (e.g. *AGL16* mRNA) or transcript (e.g. *TTG1* transcript). In some cases, it may be named with the gene name or protein name (metonymy). For instance, “These RNAs include 2S and 12S storage proteins, oleosin, and steroleosin” In this case, the gene or the protein is annotated with the [RNA] type (see above). This generates ambiguities and should be avoided when possible.

**Examples:** *CLV3* mRNA ; *LEC2*-induced RNAs ; *CBF2* transcript

#### Protein

##### Description:

*Definition:* RNA product

Protein names are in plain letters and uppercase (e.g. *CLV3*). If a protein name is followed by the word “protein”, the word “protein” should not be annotated. (e.g. *CLV3* protein). If a protein is denoted by the gene name followed by the word “expression” or “protein”, the whole phrase is annotated as a protein (e.g. *ABI3* expression)

**Examples:** *CLV3* ; *LEC1* ; *GLABRA3*

#### Protein\_Family

##### Description:

*Definition:* A family of proteins mentioned by their common biologic function (e.g. kinase) or by their common ancestor.

[Protein] should not be confused with [Gene]. Proteins should not be confused with protein domains or families.

A protein family can also be defined by a common domain shared by the members of the family (e.g. *B3* domain), which can be ambiguous with the protein domain type and should be avoided as much as possible.

Proteins mentioned by their common pathway, such as “enzymes of the glycolysis” should not be annotated as a [Protein\_Family].

**Examples:** MADS-domain proteins ; MYB ; bHLH protein ; bZIP transcription factor families ; seed storage proteins

**Counter examples:** enzymes of the glycolysis ; transcription factors ; sequence-specific DNA binding proteins

*Databases:*

<http://planttfdb.cbi.edu.cn/index.php?sp=At>

<http://arabidopsis.med.ohio-state.edu/AtTFDB/>

*Protein\_Complex***Description:**

*Definition:* A group of proteins that physically interact together.

**Examples:** PolycombGroupProtein (PCG) ; MYB-bHLH-WD40 (MBW) ; AFL (ABI3-FUS3-LEC) ; SERK1 complex

*Protein\_Domain***Description:**

*Definition:* A protein domain is a protein sequence and structure that can evolve, function and exist independently, a motif is a sequence that is widespread and has biological significance.

[*Protein\_Domain*] should not be confused with [*Protein*] or [*Protein\_Family*] .

**Examples:** B3 domain ; C-terminal region ; non-LEC1-type B domain ; basic helix-loop-helix (bHLH)

**A.1.5 Other molecule entities***Hormone***Description:**

*Definition:* A hormone is a plant molecule that influences plant physiology and development.

It does not include sensitivity to the hormone, such as in “ABA sensitivity” where the whole expression should be annotated as [*Regulatory\_Network*] , but only “ABA” should be annotated as [*Hormone*]. If a hormone was artificially added, such as in “ethylene treatment”, the expression should be annotated as [*Environmental\_Factor*].

**Examples:** auxin ; ethylene ; abscisic acid (ABA) ; GA Gibberellic ac (GA) ; Cytokinin (CK) ; strigolactone ; Brassinosteroid ; Salicylic Acid (SA), Nitric oxide (NO)

**A.1.6 Dynamic process***Regulatory\_Network***Description:**

*Definition:* A set of DNA products and/or DNA that control the expression of a gene, a pathway.

Generally, regulatory functions should be annotated by [*Regulatory\_Network*].

- Sensitivity to a factor (e.g. sensitivity to abscisic acid),
- Response to a factor (e.g. sugar responses),

- Acquisition of properties (e.g. **acquisition of desiccation tolerance**)
  - Fate (e.g. **embryonic cell fate**) in the sense of determination of the cell destiny.
  - Processes and functionality involving several genes (e.g. **meristem function**, **control of seed size**, **maturation-specific genes**)
  - are regulatory networks.
- ❑ In the case of the specific expression “control of” such as “**control of seed size**”, the whole expression should not be annotated : only the words that follow “control of” should must be annotated as a **[Regulatory Network]**.
- ❑ In some cases, regulatory networks are designated by the genes that achieved a function. The annotation depends on the action verb (regulate, associated to, involved in) and the action (**[Pathway]** / **[Regulatory Network]** / **[Development Phase]**):

- *The genes that regulate* **[Pathway]** / **[Regulatory Network]** / **[Development Phase]**: the whole expression should be annotated as a **[Regulatory Network]**

Ex: **The genes that regulate female reproductive tract development.**

- *The genes regulate* **[Pathway]** / **[Regulatory Network]** / **[Development Phase]**: the genes are related to the **[Development Phase]** by the relation they regulate the activity of (option involvement).

Ex: **The genes regulate female reproductive tract development.**

- *The genes involved in/related to* **[Development Phase]**, **[Regulatory Network]**: the whole expression should be annotated as a **[Regulatory Network]**

Ex: **Genes involved in the female reproductive tract development.**

- *The genes involved in/related to* **[Pathway]**: the expression **should not be annotated** as a **[Regulatory Network]**

Ex: **Genes involved in the accumulation of storage protein [Pathway].**

- ❑ Abnormal development such as “incomplete cytoplasmic rearrangement” and formation of tissues that occurs in abnormal conditions such as “vegetative organs” or “patterning defects” should be annotated as **[Regulatory Network]** because it means that the development and tissue formation has been triggered by regulation (e.g. **embryo formation in vegetative organs, somatic embryo or embryonic mode**)
- ❑ In “**acquisition of desiccation tolerance**” the whole term is annotated by **[Regulatory Network]**, and the term “**desiccation tolerance**” is also annotated as a **[Pathway]**.

**Examples:** germination sensitivity to abscisic acid ; sensitivity to ABA inhibition of

germination ; sensitivity to sugar ; salt responses ; auxin responses ; seedling responses to sugars ; acquisition of desiccation tolerance ; fate of embryonic cells ; vegetative leaf fate ; formation of ectopic trichome on embryos

### Counter examples:

- “the expression of genes involved in **fatty-acid biosynthesis**”
  - This is not a **[Regulatory\_Network]**, only “fatty-acid biosynthesis” should be annotated as a **[Pathway]**
- “Genes related to the **biosynthesis** and **storage of triacylglycerols**”
  - This is not a **[Regulatory\_Network]**, only “biosynthesis...of triacylglycerols” and “storage of triacylglycerols” should be annotated as **[Pathway]**

## Pathway

### Description:

*Definition:* **[Pathway]** means here metabolic pathway, for instance synthesis or degradation.

A pathway represents a group of genes or corresponding products that are involved in a same metabolic, physiological or developmental pathway. In reality, a set of any entities that are involved in such a pathway can be annotated as a pathway here. The biological function of the pathway is also annotated as a pathway (e.g. regulation of **storage protein synthesis**). It may happen that “storage protein synthesis” is abbreviated into “storage protein”.

In the example “**LEC2** regulates the **accumulation of storage protein**”, the word “accumulation” should be annotated in the pathway entity span. The *relation* **Regulation\_Of\_Process** should be used.

In “**acquisition of desiccation tolerance**” the whole term is annotated as regulatory network. The term “**desiccation tolerance**” is a pathway.

The word “pathway” is included in the annotation only when it is part of the name of the pathway. Typically, when it is followed by an adjective:

- **fatty acid biosynthetic pathway** → “pathway” is included
- **glycolysis** pathway → “pathway” is not included

The signaling pathways must be annotated as **[Regulatory\_Network]** and not as **[Pathway]**.

**Examples:** metabolic pathways ; lipid/triacylglycerol/oil pathway ; fatty acid pathways; storage protein pathways ; starch pathway ; carbohydrate pathway ; secondary metabolite pathways ; flavonoid pathway ; tannin pathway

*Database:*

<ftp://ftp.plantcyc.org/Pathways/>

## A.1.7 Context

## Genotype and Species

(Genotype type)

### Description:

**Definition:** A genotype is given part or the whole genetic information (genetic composition) expressed by an organism genome. In this case, for *Arabidopsis thaliana*. It includes **ectopic**, as for example in “**ectopic expression**”, which means the expression of the non-modified organism.

In the corpus, it mainly concerns the mutant organism resulting from a genetic modification (e.g. insertion or deletion of a gene). They are denoted by the gene name with all letters in lowercase and italic (e.g. *fus3* mutant). If the gene name has a gene typography (italic and all letters in uppercase) and it is followed by the word “mutant”, the whole expression should be annotated as a genotype (e.g. *Fus3* mutant), and the gene name should be annotated as a gene.

The non-mutant is called **wild-type** and should be annotated with [*Genotype*].

The entity text span **includes** the terms qualifying a change in the expression as a result of a mutation, like overexpression. For example, in “**plants that overexpress floral MADS factors**, such as **AP1**, **AG**, or **AP3** with **PISTILLATA**” the whole expression “**plants that overexpress floral MADS factors**” should be annotated as [*Genotype*].

However, the entity **excludes** the terms qualifying a change in the expression as a result of an exterior factor.

The entity [*Genotype*] should be annotated with all its modifiers (i.e., the largest text span is annotated), such as in “**Ectopic postembryonic expression of the *LEC1* gene in vegetative cells** induces the expression [..].”

**Species:** referring to the biological nomenclature.

The expression **in planta** should be annotated as [*Genotype*]; but **in vivo** should be annotated as [*Environmental Factor*].

**Examples:** *serk2* null mutant ; variety ; cultivar ; cross ; mutation ; mutant cyanobacteria ; plants ; yeast ; animals ; ciliates ; oilseed rape ; soybean ; maize ; eukariotic organisms ; plants that overexpress *AP1* ; *fus3* ; *fusca3* ; *Fus3* mutant ; wild-type

## Tissue, Cell and Organ

(Tissue type)

### Description:

**Definition:** A tissue is a group of cells, not necessarily identical, that carry out a specific function together. Organs are then formed by the functional grouping together of multiple tissues.

The Tissue type includes organs, as well as entities on the intra-cellular level, such as the “**nuclei of endosperm**” for example.

The entity span includes all characteristics of the tissue. The state of the tissue should be included in the annotation span (e.g. **maturing seed**). The species should be included in the annotation span (e.g. **endosperm of maize**).

When a tissue is described as a part of another tissue, both tissues should be annotated, and linked by a relationship (e.g. in **aleurone layer of mature embryos**, the whole expression is annotated as **[Tissue]**, as well as “**mature embryos**”, and the **two** are also linked by the relation **Localization**)

**Examples:** vegetative tissues ; reproductive tissues ; seedling ; cotyledon ; root ; stem ; leaf ; bud ; flower ; fruit ; silique ; seed ; testa ; integument ; chalaze ; micropyle ; envelope ; embryo ; endosperm ; epiderm ; ovule ; pollen ; green seed ; germinating seed ; embryo-like structures from vegetative tissues ; endosperm of maize ; cytoplasm ; nuclei ; cytosol ; Golgi apparatus

**Counter examples:**

Cell (too general)

Databases:

<http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi> and more specifically for seed:

<http://seedgenenetwork.net/>

## Development\_Phase

**Description:**

*Definition:* A growth stage.

It includes identity (**cotyledon identity**), dormancy (**bud dormancy**), development (**cotyledon development**), growth (**etiolated growth**).

It should not be confused with **[Tissue]**, such as developing tissues (e.g. **embryonic cotyledons**, **nuclei of young embryos**) or maturing tissue (e.g. **maturing A.th seeds**).

It should not be confused with abnormal development such as “**earlier flowering**” or “**precocious germination**” : the whole expression should be annotated as **[Regulatory\_Network]**, and only “**flowering**” and “**germination**” should be annotated as **[Development\_Phase]**

**Examples:** number of days after fertilization (daf) ; flowering ; maturation ; juvenile phase ; globular stage

Databases:

As for the **[Tissue]**, nomenclature can be found at

<http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi> and more specifically for seed at

<http://seedgenenetwork.net/>

## Environmental\_Factor

**(Environmental or experimental conditions)**

**Description:**

*Definition:* Any factor, abiotic or biotic, that influences living organisms (e.g. temperature, light).

They are rare in the corpus. They should not be confused with mutants, such as **tnp**



*mutants* that should be annotated as genotypes.

Expressions such as “laboratory conditions” should not be annotated as

[*Environmental Factor*] because it is too general and not informative.

However, expression “*in vitro*”, “*in vivo*” and “*yeast two-hybrid*” should be annotated as [*Environmental Factor*].

### Examples:

Temperature ; drought ; salt ; heat ; abiotic stress ; biotic stress ; ethylene treatment

## A.2. Events

Events (or relations) are n-ary and directed named (typed) relations between entity arguments of the types defined above. All events have two required arguments and may have between 0 and 6 optional secondary arguments (except for *Presence In Genotype*).

In the following, we use the conventions defined below to distinguish between relation names, roles in relation and entities:

- Entities are noted in italic font with the color scheme previously described (e.g. *Tissue*), and in sentences are between brackets (e.g. [*Tissue*])
- Relation names are noted in straight font with the color scheme described below (e.g. *Protein\_Domain\_Composition*),
- Roles of relation are noted in black and bold (e.g. **Tissue**)

For readability, roles are usually not notified in binary examples, but are detailed in the case of n-ary examples.

### n-ary event example with 2 arguments:

“Three *VP1/ABI3-LIKE genes* [*Gene Family*] encode *B3 proteins* [*Protein Family*]”

- **Relation:**
- **Transcription\_or\_Translation** : Source= *VP1/ABI3-LIKE genes* [*Gene Family*] + Product= *B3 proteins* [*Protein Family*]

### n-ary event example with 4 arguments (2 required arguments and 2 optional arguments)

“In addition to the ectopic localization of *protoderm* [*Tissue*] markers in the *suspensor* [*Tissue*], *globular* [*Developmental Phase*] *dcl1-5* [*Genotype*] *suspensors* [*Tissue*] had reduced levels of *WOX8 transcripts* [*RNA*]”

- **Relation:**
- **Localisation** : Functional\_Molecule= *WOX8 transcripts* [*RNA*] + Target\_Tissue= *suspensor* [*Tissue*] + Organism\_Genotype= *dcl1-5* [*Genotype*] + Developmental\_Stage= *globular* [*Developmental Phase*]

#### A.2.1 Required arguments of relations

The **roles of relation arguments** are also typed. Groupings of entity types have

been created for better readability and are used for naming the argument roles when applicable:

→ **Molecule :**

◆ **DNA :** *Gene*, *Gene\_Family*, *Box*, *Promoter*

◆ **Functional\_Molecule**

● **DNA product :**

○ *RNA*

○ **Amino acid sequence :** *Protein*, *Protein\_Family*,  
*Protein\_Complex*, *Protein\_Domain*

● *Hormone*

→ **Dynamic Process :** *Regulatory\_Network*, *Pathway*

→ **Context :**

◆ **Biological context** *Tissue*, *Development\_Phase*, *Genotype*,

◆ *Environmental\_Factor*

For **required arguments**, not all combinations are possible. [ X | Y ] stands for the union of type X and Y, meaning that the argument can be either of type X or of type Y. In the following, we describe each of the main arguments allowed for each event.

### A.2.2 Event type and color scheme

Here you can see the event types of our model, classified in large groups, as well as the annotation color scheme used in the Annotation Editor (*AlvisAE*).

#### Where and When

- *Presence\_In\_Genotype*
- *Occurrence\_In\_Genotype*
- *Presence\_At\_Stage*
- *Occurrence\_During*
- *Localization*

#### Function

- *Involvement\_In\_Process*
- *Transcription\_Or\_Translation*
- *Functional\_Equivalence*

#### Regulation

- *Regulation\_Of\_Accumulation*
- *Regulation\_Of\_Expression*
- *Regulation\_Of\_Development\_Phase*
- *Regulation\_Of\_Molecule\_Activity*
- *Regulation\_Of\_Process*
- *Regulation\_Of\_Tissue\_Development*

#### Composition and Membership

- *Primary\_Structure\_Composition*
- *Protein\_Complex\_Composition*
- *Protein\_Domain\_Composition*
- *Family\_Membership*



- **Sequence Identity**

### Interaction

- **Interaction**
- **Binding**

## A.2.3 Extra-properties of events

To limit the number of relations, qualifications of given relations are represented by parameters. A default is set for each relation. All relations have parameters. In this task, only one kind of parameters exists:

- Modalities: Negation (Y/N) : describe the status of the knowledge

## A.2.4 Event for specific case

- **Specific case of entities with synonyms**

Generally, synonyms are linked by a "**Sequence Identity**" relation.

These synonyms may be included in an other relation, also to avoid double annotations, the following rules are used, in **this significance order**:

1. The abbreviated form is preferred to the full form  
e.g. **Diacylglycerol acyltransferase 1** (**DGAT1**) belongs to **DGAT**
  - "**Sequence Identity**" relation between **DGAT1** [**Protein**] and **Diacylglycerol acyltransferase 1** [**Protein**]
  - "**Family Membership**" relation between **DGAT1** [**Protein**] and **DGAT** [**Protein\_Family**]
2. The most precise form is preferred to the generic form :  
e.g. **MIKC** contain a second weakly conserved **coiled-coil motif** (the **K domain**)
  - "**Sequence Identity**" relation between **K domain** [**Protein\_Domain**] and **coiled-coil motif** [**Protein\_Domain**]
  - "**Protein\_Domain\_Composition**" relation between **K domain** [**Protein\_Domain**] and **MIKC** [**Protein\_Family**]
3. The entity in text is preferred to that between brackets :  
e.g. **Arabidopsis** (**Arabidopsis thaliana**) has two **HDA19** orthologs, **HDA19** and **HDA6**.
  - "**Sequence Identity**" relation between **Arabidopsis** [**Genotype**] and **Arabidopsis thaliana** [**Genotype**]
  - "**Presence In Genotype**" relation between **HDA19** [**Gene**] and **Arabidopsis** [**Genotype**]
  - "**Presence In Genotype**" relation between **HDA6** [**Gene**] and **Arabidopsis** [**Genotype**]

## A.2.5 "Where and When" events

### Presence\_In\_Genotype

**Description:** A **Molecule** or **Element** is present **in** a given **Genotype**.

**Arguments:**

**Molecule:** Molecule (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*)

**Element:** Biological context (*Tissue*, *Development Phase*, *Genotype*)

**Genotype:** *Genotype*

(**Molecule** et **Element** are mutually exclusive)

The relation **Presence In Genotype** should be used in case of genes that are found specifically in species/genotype (e.g. : “*VP1* of *maize*”), because it means genes are found in plant, and not necessarily expressed

**Vocabulary:** identified, conserved, possesses

**Examples:**

- Proteins containing homologues of the *AP2 domain* [*Protein Domain*] have been identified in *cyanobacteria* [*Genotype*]
- *Yeast* [*Genotype*] has only one *HDAC* [*Protein Family*]

### Occurrence\_In\_Genotype

**Description:** A **Process** occurs **in** a given **Genotype**.

**Arguments:**

**Process:** Dynamic\_Process (*Regulatory Network*, *Pathway*)

**Genotype:** *Genotype*

**Vocabulary:** occurs, observed, described, present

**Examples:**

- It is becoming increasingly apparent that autoregulatory loops are a common phenomenon in the *regulation of MADS-box genes* [*Regulatory Network*] in *plants* [*Genotype*]

### Presence\_At\_Stage

**Description:** A **Molecule** is present **during** a given **Developmental phase**.

However, the relation **Presence At Stage** should not be used in case of a specific expression of genes during a developmental phase (e.g. : “*expression of the LEC1 gene in embryogenesis*”), because **all** genes are found in cells at any developmental stage (but not necessarily expressed).

However, the relation **Regulation Of Expression** should be used :

- “*embryogenesis*” [*Developmental Phase*] **Regulation Of Expression** “*LEC1*” [*Gene*]

**Arguments:**

**Functional\_Molecule:** Functional\_Molecule (*RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*)

**Development :** *Development Phase*

**Vocabulary:** expressed during, accumulates during, present during

**Examples:**

- > We showed that **LEC2 RNA** [**RNA**] accumulates primarily during **seed development** [**Developmental\_Phase**]

### *Occurrence\_During*

**Description:** A **Process** occurs *during* a given **Developmental\_Phase**.

**Arguments:**

**Process:** Dynamic\_Process (**Regulatory\_Network**, **Pathway**)

**Development:** **Developmental\_Phase**

**Vocabulary:** implicated during, established during, occurs during, active during

**Examples:**

- > **Higher plant embryogenesis** [**Developmental\_Phase**] is divided conceptually into two distinct phases: **early morphogenetic processes** [**Regulatory\_Network**] [...]

### *Localization*

**Description :** A **Molecule** is found *in* a **Tissue**

However, the relation **Localization** should not be used in case of a specific expression of genes in tissues (e.g. : “expression of the **LEC1** gene in **vegetative cells**”), because **all** genes are found in cells (but not necessarily expressed).

However, the relation **Regulation\_Of\_Expression** should be used :

- > “**vegetative cells**” [**Tissue**] **Regulation\_Of\_Expression** “**LEC1**” [**Gene**]

**Arguments:**

**Functional\_Molecule:** Functional\_Molecule (**RNA** , **Protein**, **Protein\_Family**, **Protein\_Complex**, **Protein\_Domain**, **Hormone**)

**Process:** Dynamic\_Process (**Regulatory\_Network**, **Pathway**)

**Target\_Tissue:** **Tissue**

(**Functional\_Molecule** et **Process** are mutually exclusive)

**Vocabulary:** accumulate in, within, possess, present in, active in

**Examples:**

- > **FUS3 mRNA** [**RNA**] accumulates in **seed** [**Tissue**]
- > **AGL15** [**Protein**] was initially present in the **cytoplasm of cells** [**Tissue**]

**Counter examples:**

- > expression of the **LEC1** [**Gene**] gene in **vegetative cells** [**Tissue**]

## A.2.6 “Function” events

### *Involvement\_In\_Process*

**Description:** A **Molecule** is involved *in* a **Dynamic Process**.

**Arguments:**

**Participant:** Molecule (**Gene**, **Gene\_Family**, **Box**, **Promoter**, **RNA** , **Protein**,

*Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*, *Hormone*)

**Process:** *Dynamic\_Process* (*Regulatory\_Network*, *Pathway*)

**Vocabulary:** involved in, includes, mediating

**Examples:**

- A complex process of differentiation [*Regulatory\_Network*] that includes the biosynthesis and secretion of pectinaceous mucilage [*Pathway*]
  - biosynthesis of pectinaceous mucilage [*Pathway*] *Involvement In Process* process of differentiation [*Regulatory\_Network*]
  - secretion of pectinaceous mucilage [*Pathway*] *Involvement In Process* process of differentiation [*Regulatory\_Network*]
- *MUM4* [*Gene*] encodes an enzyme involved in RGI biosynthesis [*Pathway*]

### *Transcription\_Or\_Translation*

**Description:** A DNA entities encodes for RNA (Transcription) or RNA entities encodes a Protein (Translation). Often, reference is made to the gene encoding the protein, without mention of the RNA.

**Arguments:**

**Source:** DNA | RNA (*Gene*, *Gene\_Family*, *RNA*)

**Product:** *DNA\_Product* (*RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*)

**Vocabulary:** encodes

**Examples:**

- Three *VP1/ABI3-LIKE* (*VAL*) [*Gene\_Family*] genes encode B3 proteins [*Protein\_Family*]

### *Functional\_Equivalence*

**Description:** A Molecule, Dynamic Process or Context compared to another similar Molecule, Dynamic Process or Context.

**Functional\_Equivalence** relation is to be used to link similar products in different species, such as homolog/ortholog proteins.

- [...] *BnSCL1*, an ortholog of the *Arabidopsis* *SCARECROW-like protein 15* (*SCL15*)

**Arguments:**

**Element1:** All entities (*Gene*, *Gene\_Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*, *Hormone*, *Regulatory\_Network*, *Pathway*, *Tissue*, *Development\_Phase*, *Genotype*, *Environmental\_Factor*)

**Element2:** All entities (*Gene*, *Gene\_Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*, *Hormone*, *Regulatory\_Network*, *Pathway*, *Tissue*, *Development\_Phase*, *Genotype*, *Environmental\_Factor*)

**Vocabulary :** ortholog, homolog

**Examples :**

- **WER** [Gene] and **GL1** [Gene] encode functionally equivalent proteins
- Rice homologs of **ABI3** and **ABI5** (**OSVP1** and **TRAB1**, respectively)
  - **ABI3** [Protein] **Functional Equivalence** **OSVP1** [Protein]
  - **ABI5** [Protein] **Functional Equivalence** **TRAB1** [Protein]

### A.2.7 “Regulation” events

#### *Use of “Regulation” relations when there is a genotype involved*

The **agent** of the **Regulation** relation is always the **genotype**, even when the gene involved in the genotype is specified, **if no information on the direct role of the agent is given**. If direct information on the agent is given, genotype is indicated in the optional argument **Organism\_Genotype** [Genotype].

#### Examples :

- **RNA interference of L1L function** [Genotype] has been shown to cause **embryo arrest** [Regulatory\_Network]
  - “**RNA interference of L1L function**” [Genotype] **Regulation\_Of\_Process** “**embryo arrest**” [Regulatory\_Network], even though the gene “**L1L**” is mentioned, because the subject is “**RNA interference of L1L function**”, and the role of **L1L** is not clearly defined.
- **Ectopic expression of LEC1** [Genotype] is sufficient to induce **embryo formation** [Regulatory\_Network]
  - “**Ectopic expression of LEC1**” [Genotype] **Regulation\_Of\_Process** “**embryo formation**” [Regulatory\_Network], even though the gene “**LEC1**” is mentioned.
- **val1 val2 double-mutant** [Genotype] **seedlings** form no **leaves** [Tissue]
  - “**val1 val2 double-mutant**” [Genotype] **Regulation\_Of\_Tissue\_Development** “**leaves**” [Tissue] (no gene is mentioned).

#### Counter examples:

- **WRI1** [Protein] is able to regulate **in planta** [Genotype] the activity of the **BCCP2 promoters** [Promoter]
  - Here, the direct role of **WRI1** [Protein] is given : **WRI1** **Regulation\_Of\_Expression** **BCCP2 promoters**, with the **Organism\_Genotype** instance “**in planta**”.

#### *Regulation\_Of\_Accumulation*

**Description:** A **Molecule**, **Dynamic Process** or **Context** regulates the accumulation of a **Functional Molecule** (in particular, [Protein], [RNA], [Hormone] entities).

#### Arguments:

**Agent:** All entities (**Gene**, **Gene\_Family**, **Box**, **Promoter**, **RNA**, **Protein**, **Protein\_Family**, **Protein\_Complex**, **Protein\_Domain**, **Hormone**, **Regulatory\_Network**, **Pathway**, **Tissue**, **Development\_Phase**, **Genotype**, **Environmental\_Factor**)



**Functional\_Molecule:** Functional\_Molecule (*RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Hormone*)

**Vocabulary:** activates, induces, enhances, trigger, inhibits, accumulation, steady state-level, concentration, increase

**Examples:**

- AGL15 [*Protein*] induces accumulation of AGL18 transcript [*RNA*]
- Finally, the results of our studies of somatic embryogenesis [*Development\_Phase*] indicate that AGL15 [*Protein*] accumulates even when embryos [*Tissue*] arise de novo from cells in other phases of the life cycle.
  - **Regulation\_Of\_Accumulation** : Agent= somatic embryogenesis [*Development\_Phase*] + Functional\_Molecule= AGL15 [*Protein*] + Tissue =embryos [*Tissue*]

**NB:** the target of this relation **cannot** be a [*Pathway*] or a [*Regulatory\_Network*]. In those cases, the relation “**Regulation\_Of\_Process**” should be used instead.

### Regulation\_Of\_Expression

**Description:** A Molecule, Dynamic Process or Context regulates the expression of a DNA entity. DNA entity includes [*Promoter*] and [*Box*].

**Argument:**

**Agent:** All entities (*Gene*, *Gene\_Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*, *Hormone*, *Regulatory\_Network*, *Pathway*, *Tissue*, *Development\_Phase*, *Genotype*, *Environmental\_Factor*)

**DNA:** DNA (*Gene*, *Gene\_Family*, *Box*, *Promoter*)

**Vocabulary:** activates, induces, enhances, trigger, inhibits, gene expression or gene transcription

**Examples:**

- WRI1 [*Protein*] is a limiting factor of lipogenic gene [*Gene\_Family*] expression in seeds [*Tissue*], directly induces the transcriptional activation of these genes at the onset of the maturation phase [*Development\_Phase*].
  - **Regulation\_Of\_Expression**: Agent= WRI1 [*Protein*] + DNA= lipogenic gene [*Gene\_Family*] + Tissue= seeds [*Tissue*] + Development\_Stage=maturation phase [*Development\_Phase*]
- Expression of the LEC1 [*Gene*] gene in vegetative cells [*Tissue*]
- We suggest that VAL [*Gene\_Family*] targets Sph/RV [*Box*] -containing genes

**Counter example:**

- WRI1 [*Protein*] directly enhances the expression of genes involved in glycolysis [*Pathway*] : even if “genes” is mentioned, the relation was made between [*Protein*] and [*Pathway*] with the “**Regulation\_Of\_Process**” relation

### Regulation\_Of\_Development\_Phase

**Description:** A Molecule, Dynamic Process or Context regulates the activity of a Development phase.

**Arguments:**

**Agent:** All entities (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*, *Regulatory Network*, *Pathway*, *Tissue*, *Development Phase*, *Genotype*, *Environmental Factor*)

**Development:** *Development Phase*

**Vocabulary:** activates, induces, enhances, trigger, inhibits, influences, blocks

**Examples:**

- Biologically active *GAs* [*Hormone*] have been suggested to have a role in *embryogenesis* [*Development Phase*]
- Genetic analysis shows that *termination of the primary shoot meristem* [*Development Phase*] in *128 mutants* [*Genotype*] requires an active *CLV signaling pathway* [*Pathway*]
  - *Regulation Of Development Phase*: **Agent**= *CLV signaling pathway* [*Pathway*] + **Development**= *termination of the primary shoot meristem* [*Development Phase*] + **Organism Genotype**= *128 mutants* [*Genotype*]

*Regulation\_Of\_Molecule\_Activity*

**Description:** An **Agent** (**Molecule**, **Dynamic Process** or **Context**) regulates the activity of a **Molecule**, such as [*Protein*].

**Arguments:**

**Agent:** All entities (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*, *Regulatory Network*, *Pathway*, *Tissue*, *Development Phase*, *Genotype*, *Environmental Factor*)

**Molecule:** Amino acid sequence | Hormone (*Protein*, *Protein Family*, *Protein Complex*, *Hormone*)

**Vocabulary:** control, regulate, phosphorylate

**Examples:**

- *p97/VCP* [*Protein Complex*] can be phosphorylated by the *JAK-2* [*Protein*] kinase

*Regulation\_Of\_Process*

**Description:** A **Molecule**, **Dynamic Process** or **Context** regulates the activity of a **Dynamic Process**.

**Arguments:**

**Agent:** All entities (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*, *Regulatory Network*, *Pathway*, *Tissue*, *Development Phase*, *Genotype*, *Environmental Factor*)

**Process:** Dynamic Process (*Regulatory Network*, *Pathway*)

**Vocabulary:** restrict, regulate, induce

**Examples:**

- The **stem cells** in turn signal back via the **CLV3** **[Protein]** peptide to restrict the **size of the OC** **[Regulatory\_Network]**
- **WRI1** **[Protein]** directly enhances the expression of genes involved in **glycolysis** **[Pathway]**

### *Regulation\_Of\_Tissue\_Development*

**Description:** A **Molecule**, **Dynamic Process** or **Context** regulates the activity of a **Dynamic Process**.

**Arguments:**

**Agent:** All entities (**Gene**, **Gene\_Family**, **Box**, **Promoter**, **RNA**, **Protein**, **Protein\_Family**, **Protein\_Complex**, **Protein\_Domain**, **Hormone**, **Regulatory\_Network**, **Pathway**, **Tissue**, **Development\_Phase**, **Genotype**, **Environmental\_Factor**)

**Target\_Tissue:** **Tissue**

**Vocabulary:** produce, generate, developped, repress, possess

**Examples:**

- The results of our study of **organs** **[Tissue]** produced during **precocious germination** **[Regulatory\_Network]**

### A.2.8 “Composition and Membership” events

#### *Primary\_Structure\_Composition*

**Description:** A specific sequence of nucleotide is found in a molecule of **DNA**

**Arguments:**

**DNA\_Part:** **Box** | **Promoter**

**DNA:** DNA (**Gene**, **Gene\_Family**, **Box**, **Promoter**)

**Vocabulary:** possess, motif, core of, located, sequence, contained

**Examples:**

- We show that mutations in the **AACCCA** **[Box]** element of the **BCCP2** **promoter** **[Promoter]**
- The **WRI1** binding sites in the **upstream region of Pl-PK  $\beta$  1** **[Promoter]** contained the conserved **AW-box** **[Box]**

#### *Protein\_Complex\_Composition*

**Description:** A specific DNA product is found in a **protein complex**.

**Arguments:**

**Amino\_Acid\_Sequence:** Amino acid sequence (**Protein**, **Protein\_Family**, **Protein\_Complex**, **Protein\_Domain**,)

**Protein\_Complex:** **Protein\_Complex**

**Vocabulary:** presence, associated, complex, detected, component of, composition of



**Examples:**

- The identification of two members of the **BR signaling pathway**, the main BR receptor **BRI1** [**Protein**] and its coreceptor **BAK1** [**Protein**] (**SERK3** [**Protein**]) as components of the **SERK1 complex** [**Protein\_Complex**].

*Protein\_Domain\_Composition*

**Description:** A specific **Protein Domain** is found in an Amino acid sequence. It can be used to link products that are part of a factor, such as [**Protein\_Domain**]  
**Protein\_Domain\_Composition** [**Protein**].

**Arguments:**

**Domain:** **Protein\_Domain**

**Product:** DNA\_Product (**RNA** , **Protein**, **Protein\_Family**, **Protein\_Complex**, **Protein\_Domain**)

**Vocabulary:** contains, presence, characterized by

**Examples:**

- Proteins of the **RAV family** [**Protein\_Family**] contain one **AP2 domain** [**Protein\_Domain**]

*Family\_Membership*

**Description:** A DNA, RNA or **Protein** belongs to another DNA, Product or Factor. This relation is to be used between entities of the same nature, to denote members of a set (e.g. [**Gene**] belonging to [**Gene\_Family**], [**Protein**] to a [**Protein\_Family**], sub-families to families, etc.).

**Family\_Membership** should **not** be used for entities that are part of another entity (e.g. [**Protein\_Domain**] that are part of [**Protein**]). The more generic relation **Protein\_Domain\_Composition** should be used instead.

**Arguments:**

**Element:** **Gene**, **Gene\_Family**, **RNA**, **Protein**, **Protein\_Family**, **Protein\_Domain**

**Family:** **Gene\_Family**, **RNA**, **Protein\_Family**

**Vocabulary:** member of, family , belongs

**Examples:**

- **AP2** [**Protein**] belongs to **AP2/EREBP** [**Protein\_Family**] family
- **AGL15** [**Gene**] from the large group of **floral MADS box genes** [**Protein\_Family**]
- **bZIP10** [**Protein**] and **bZIP25** [**Protein**], which have been classified into **group C** [**Protein\_Family**]

**Counter examples:**

- The **LEC2** [**Protein**] protein contains a DNA-binding **B3 domain** [**Protein\_Domain**]  
→ **Protein\_Domain\_Composition** between **B3 domain** and **LEC2**

*Sequence\_Identity*

**Description:** A **Molecule**, **Dynamic Process** or **Context** compared to another similar **Molecule**, **Dynamic Process** or **Context**.

The **Sequence Identity** relation is to be used to link identical products as well as synonyms, full form and abbreviation .

#### Arguments:

**Element1:** All entities (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*, *Regulatory Network*, *Pathway*, *Tissue*, *Development Phase*, *Genotype*, *Environmental Factor*)

**Element2:** All entities (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*, *Regulatory Network*, *Pathway*, *Tissue*, *Development Phase*, *Genotype*, *Environmental Factor*)

**Vocabulary :** identical, synonym

#### Examples :

- The *SUN6* *gene* [*Gene*] is identical to the previously described *ABI4* *gene* [*Gene*]
- *LEAFY COTYLEDON2* [*Protein*] (*LEC2*) [*Protein*]

### A.2.9 “Interaction” events

#### *Binding*

**Description:** A **functional molecule** physically binds to a **molecule**. In most cases, a protein binds to a promoter or a gene.

When the interaction between two protein is specifically performed “*in vitro*” or in “*yeast two-hybrid*”, the relation **Binding** is used (and not **Interaction**, which is less precise)

- **Example:** *FUS3* [*Protein*] interacts with *LEC2* [*Protein*] *in vitro*

In the specific case of a homodimeric interaction, the agent and the target of the **Binding** relation is the same entity:

- **Example:** *BRI1* [*Protein*] can also form *homodimers* in the *plasma membrane* [*Tissue*]
- In this case: **Binding Functional\_Molecule**= *BRI1* [*Protein*] + **Molecule**= *BRI1* [*Protein*] + **Tissue**=*plasma membrane* [*Tissue*]

#### Arguments:

**Functional\_Molecule:** Functional\_Molecule (*RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*)

**Molecule:** Molecule (*Gene*, *Gene Family*, *Box*, *Promoter*, *RNA*, *Protein*, *Protein Family*, *Protein Complex*, *Protein Domain*, *Hormone*)

**Vocabulary:** binds, physically interacts, co-precipitates, co-migrate

#### Examples:

- Interaction between *WRI1* [*Protein*] and the *BCCP2* promoter [*Promoter*], both *in vitro* and in *yeast*

- TT2 [Protein] interacts with TT8-TTG1 [Protein\_Complex] in yeast two-hybrid studies.
- CLV3 [Protein] acts as an extracellular ligand of the CLV1 receptor kinase complex [Protein\_Complex]

## Interaction

**Description:** A molecule interacts with another molecule.

The relation **Interaction** should be used between DNA-DNA, and in the case of indirect interaction. It may be applied to a DNA or Amino acid sequence, but **Binding**, **Regulation Of Molecule Activity** or **Regulation Of Expression** must be used when possible.

It excludes the case where the interaction is direct (physical): the more specific relation **Binding** must be used in this case.

## Arguments:

**Agent:** DNA | Amino acid sequence (*Gene*, *Gene\_Family*, *Box*, *Promoter*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*)

**Target:** DNA | Amino acid sequence (*Gene*, *Gene\_Family*, *Box*, *Promoter*, *Protein*, *Protein\_Family*, *Protein\_Complex*, *Protein\_Domain*)

**Vocabulary:** cooperate, synergistic antagonist, counteract relationships, effects

## Examples:

- SAP18 [Gene] alone repressed LEA and CBF2, possibly through interaction with AGL15 [Gene]
  - Genetic interaction
- AGL15 [Protein] interacts with members of the SIN3 histone deacetylase (HDAC) complex [Protein\_Complex]
  - Physically bind unproved

## Counter example:

- CLV3 [Protein] presumably acts as an extracellular ligand of the CLV1 [Protein] receptor kinase complex
  - Here, the relation **Binding** should be used
- Ternary complex formation between the bZIP [Protein\_Family] heterodimers and ABI3 [Protein]
  - Here, the relation **Binding** should be used

## A.2.10 Optional arguments of relations

6 optional arguments are possible to describe n-ary event. 5 of them require an entity identifier as argument. For these, only one entity type per role is authorized. These arguments can exist in ALL types of relations (except for **Presence In Genotype**).

- **Tissue:**
  - *Tissue*

- **Developmental\_Stage:**
  - **Development\_Phase**
- **Organism\_Genotype:**
  - **Genotype**
- **Environmental\_Factor:**
  - **Environmental\_Factor**
- **Hormone:**
  - **Hormone**

A specific optional argument is **Prerequisite\_Event** which accepts events as argument.

- **Prerequisite event:**
  - **Primary\_Structure\_Composition**
  - **Interaction**
  - **Localization**
  - **Protein\_Domain\_Composition**

Optional arguments are used to handle cases when we have some conditional information of the type:

- “A [**Protein**] accumulates in B [**Tissue**] when there exists C [**Hormone**] .”
- In this fictional example we have a **Localization** relation between A [**Protein**] & B [**Tissue**], and the optional argument **Hormone** is C [**Hormone**].

In a second fictional example we can see the second scenario :

- “A [**Protein**] activates B [**Gene**] in the flower [**Tissue**], if C [**Hormone**] exists.”
  - Here we have a **Regulation\_Of\_Expression** event:
    - **Agent A** [**Protein**] **Regulation\_Of\_Expression** DNA B [**Gene**] , **Tissue** flower [**Tissue**] , **Hormone** C [**Hormone**]

In a third and fourth fictional example, we can see scenarios illustrating the difference between arguments in conjunction vs. arguments in disjunction. When arguments are in conjunction, only one relation is annotated with all the arguments.

- “A [**Protein**] accumulates in B [**Tissue**] when there exists C [**Hormone**] and D [**Environmental\_Factor**] .”
  - Here : we have a **Localization** relation between **Functional\_Molecule** A [**Protein**] & **Target\_Tissue** B [**Tissue**], and the 2 optional arguments: **Hormone** C [**Hormone**] and **Environmental\_Factor** D [**Environmental\_Factor**]

When arguments are in disjunction, two separate relations are annotated.

- “A [**Protein**] accumulates in B [**Tissue**] when there exists C [**Hormone**] or D [**Environmental\_Factor**] .”
- Here : we have two relations :
  - **First relation** : **Localization** relation between **Functional\_Molecule** A [**Protein**] & **Target\_Tissue** B [**Tissue**], and an optional argument:

**Hormone C** [*Hormone*]

- Second relation : *Localization* relation between **Functional\_Molecule** A [*Protein*] & **Target\_Tissue** B [*Tissue*], and an optional argument: **Environmental\_Factor** D [*Environmental\_Factor*]

In a fifth fictional example, we can see the scenario where a prerequisite event is needed.

- “A [*Protein*] binds to B [*Protein\_complex*] if C [*Hormone*] is found in D [*Tissue*].”
  - First relation: Event identifiant: Event1: *Localization* relation between **Functional\_Molecule** C [*Hormone*] & **Target\_Tissue** C [*Tissue*]
  - Second relation: *Binding* relation between **Agent** : A [*Protein*] **Target**: B [*Protein\_complex*] **Pre-requisite event**: Event1

### A.2.1 Extra properties

#### *Negation*

All types of relations have an extra field to mark the negation of an event.

#### **Example :**

- *Trichomes* [*Tissue*] are present on *Arabidopsis* *leaves* and *stems* but **not** on *wild-type* [*Genotype*] *cotyledons* [*Tissue*]
- Notably, **none** of the proposed downstream components of *BRI1*, such as *BES1* [*Protein*], *BZR1* [*Protein*], or *BIN2* [*Protein*] was detected in the *SERK1 complex* [*Protein\_Complex*].

## B. Figures

### B.1 Entity types and Categories

Agent	Molecule	DNA		Gene	"LEC1"	
				Gene_Family	"AP2-like"	
				Box	"5'-GCATCG-3'"	
				Promoter	"BCCP2"	
		DNA Product		Amino acid sequence	RNA	"FUS3 transcript"
					Protein	"WRI1"
					Protein_Family	"SSPs"
					Protein_Complex	"SIN3/HDAC"
					Protein_Domain	"MADS-domain"
					Hormone	"ABA"
	Dynamic Process			Regulatory_Network	"embryonic process"	
				Metabolic_Pathway	"FA biosynthesis"	
	Context		Biological context	Genotype	"fus3 mutant"	
				Tissue	"embryo"	
				Development_Phase	"meristem formation"	
				Environmental_Factor	"in vitro"	

## B.2 Relation classification

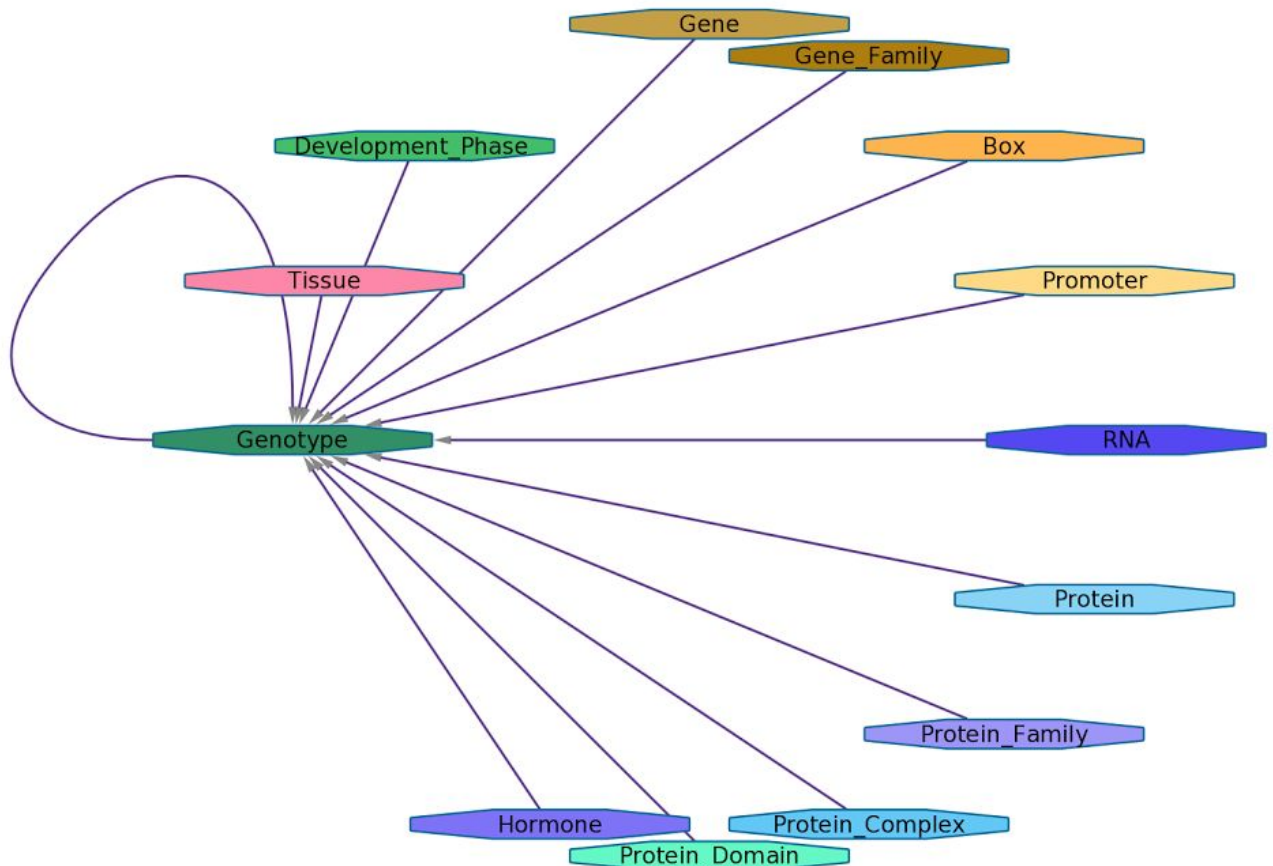
BIONLP-ST task See-Dev have two subtasks: *binary relation extraction* and *full event extraction*. The labels are the same, except **Is\_Linked\_To** relation, which is specific to the binary framework :

Set of relation type	N-ary relation name	Binary relation name
Where and When	<i>Presence_At_Stage</i>	<i>Exists_At_Stage</i>
Where and When	<i>Presence_In_Genotype</i>	<i>Exists_In_Genotype</i>
Where and When	<i>Localization</i>	<i>Is_Localized_In</i>
Where and When	<i>Occurrence_During</i>	<i>Occurs_During</i>
Where and When	<i>Occurrence_In_Genotype</i>	<i>Occurs_In_Genotype</i>
Function	<i>Functional_Equivalence</i>	<i>Is_Functionally_Equivalent_To</i>
Function	<i>Involvement_In_Process</i>	<i>Is_Involved_In_Process</i>
Function	<i>Transcription_Or_Translation</i>	<i>Transcribes_Or_Translates_To</i>
Regulation	<i>Regulation_Of_Accumulation</i>	<i>Regulates_Accumulation</i>
Regulation	<i>Regulation_Of_Development_Phase</i>	<i>Regulates_Development_Phase</i>
Regulation	<i>Regulation_Of_Expression</i>	<i>Regulates_Expression</i>
Regulation	<i>Regulation_Of_Molecule_Activity</i>	<i>Regulates_Molecule_Activity</i>
Regulation	<i>Regulation_Of_Process</i>	<i>Regulates_Process</i>
Regulation	<i>Regulation_Of_Tissue_Development</i>	<i>Regulates_Tissue_Development</i>
Composition and Membership	<i>Primary_Structure_Composition</i>	<i>Composes_Primary_Structure</i>
Composition and Membership	<i>Protein_Complex_Composition</i>	<i>Composes_Protein_Complex</i>
Composition and Membership	<i>Sequence_Identity</i>	<i>Has_Sequence_Identical_To</i>
Composition and Membership	<i>Family_Membership</i>	<i>Is_Member_Of_Family</i>
Composition and Membership	<i>Protein_Domain_Composition</i>	<i>Is_Protein_Domain_Of</i>
Interaction	<i>Binding</i>	<i>Binds_To</i>
Interaction	<i>Interaction</i>	<i>Interacts_With</i>
Interaction	<i>XXXX</i>	<i>Is_Linked_To</i>

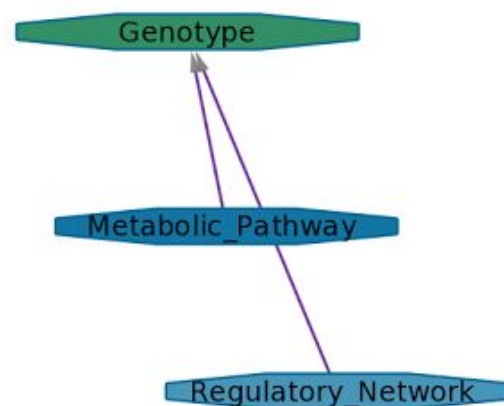


## B.3 Schematical representation of Arguments and Events

### B.3.1 Presence\_In\_Genotype

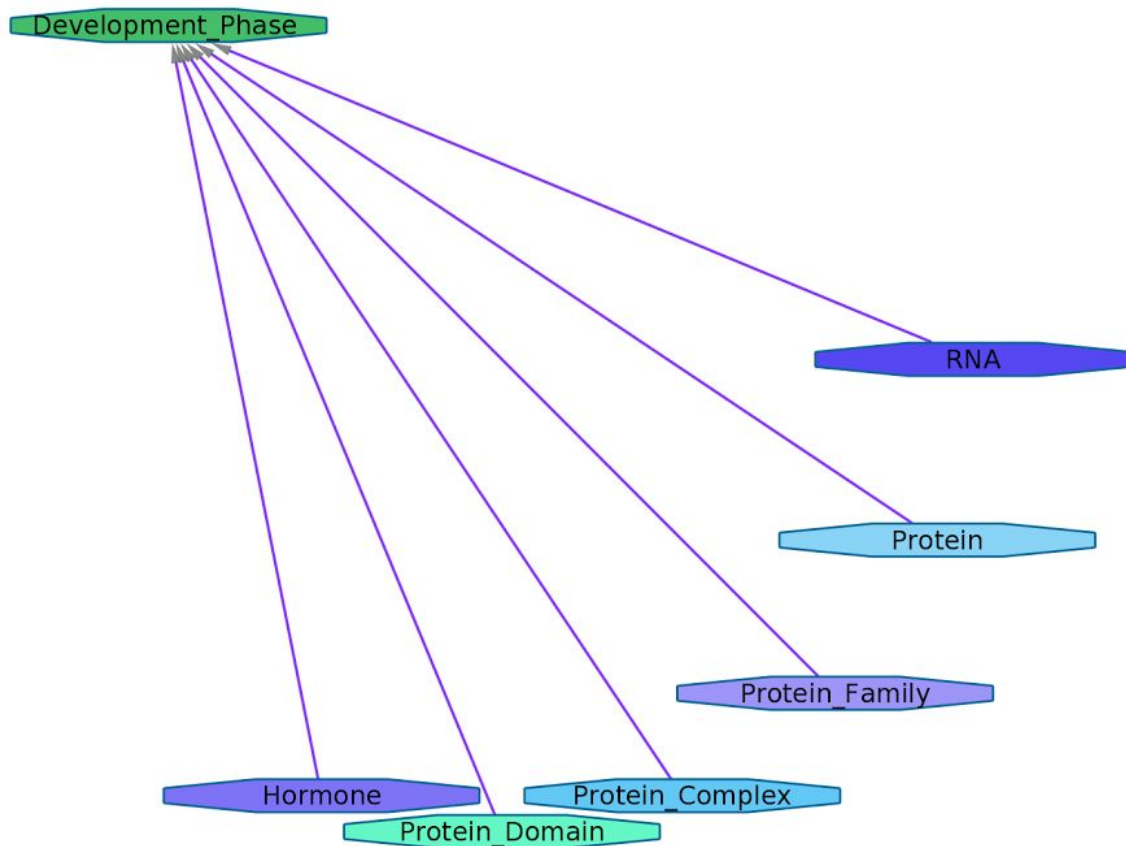


### B.3.2 Occurrence\_In\_Genotype

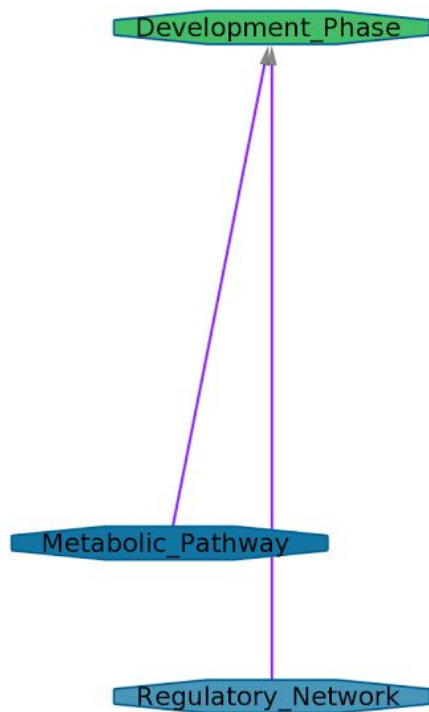




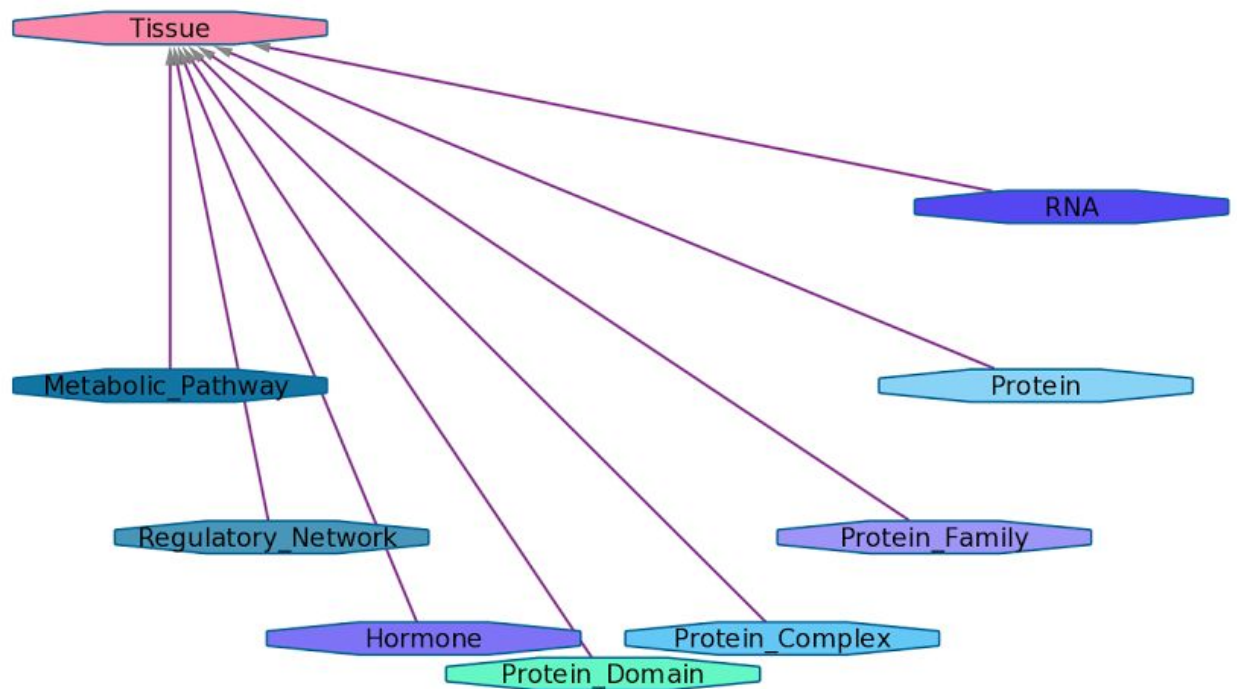
### B.3.3 Presence\_At\_Stage



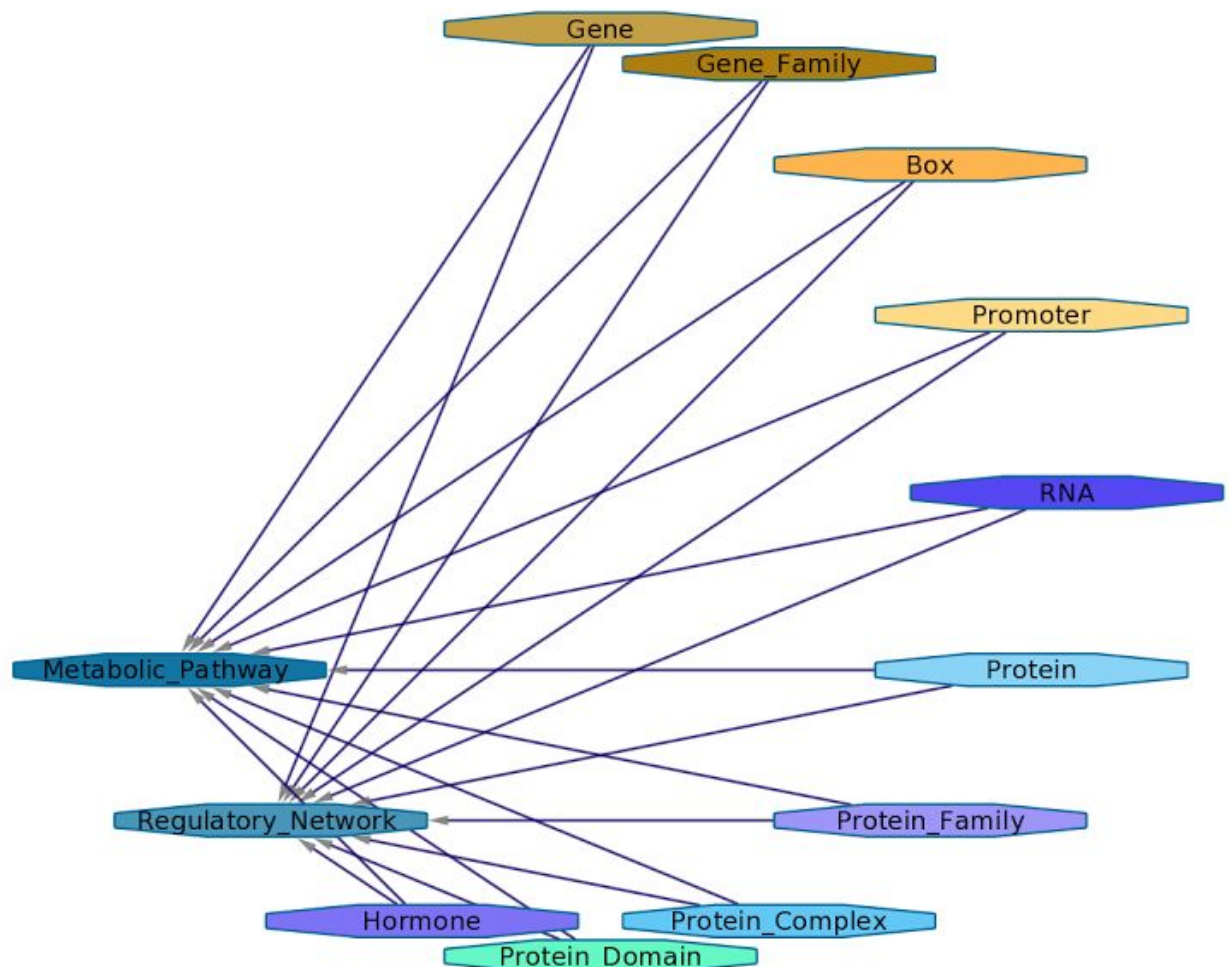
### B.3.4 Occurrence\_During



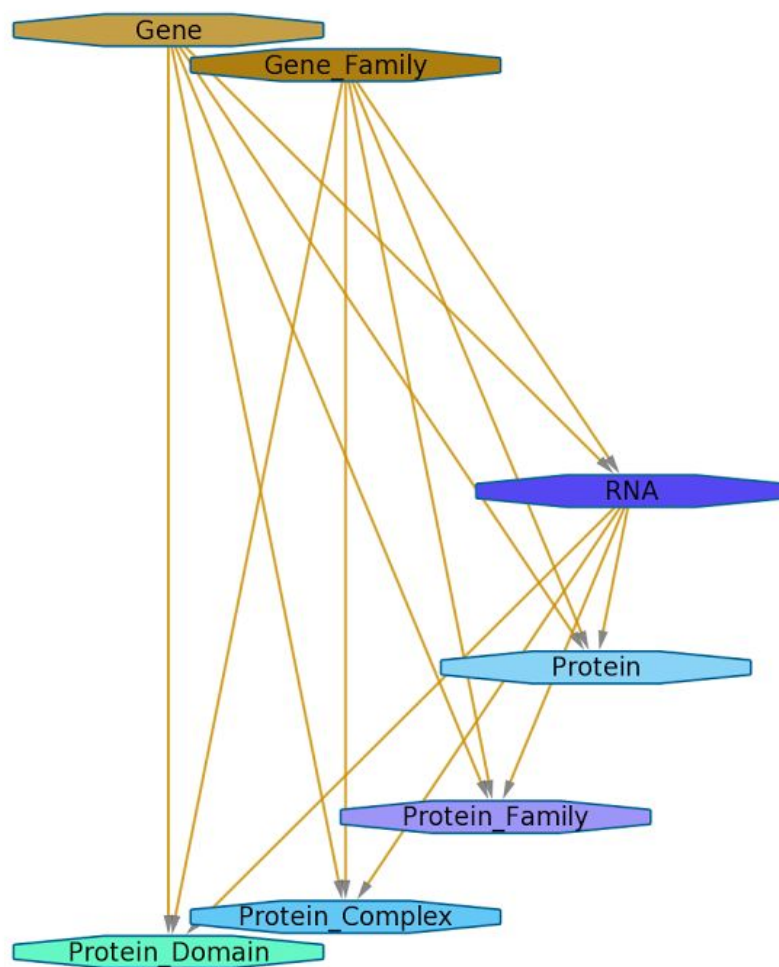
### B.3.5 Localization



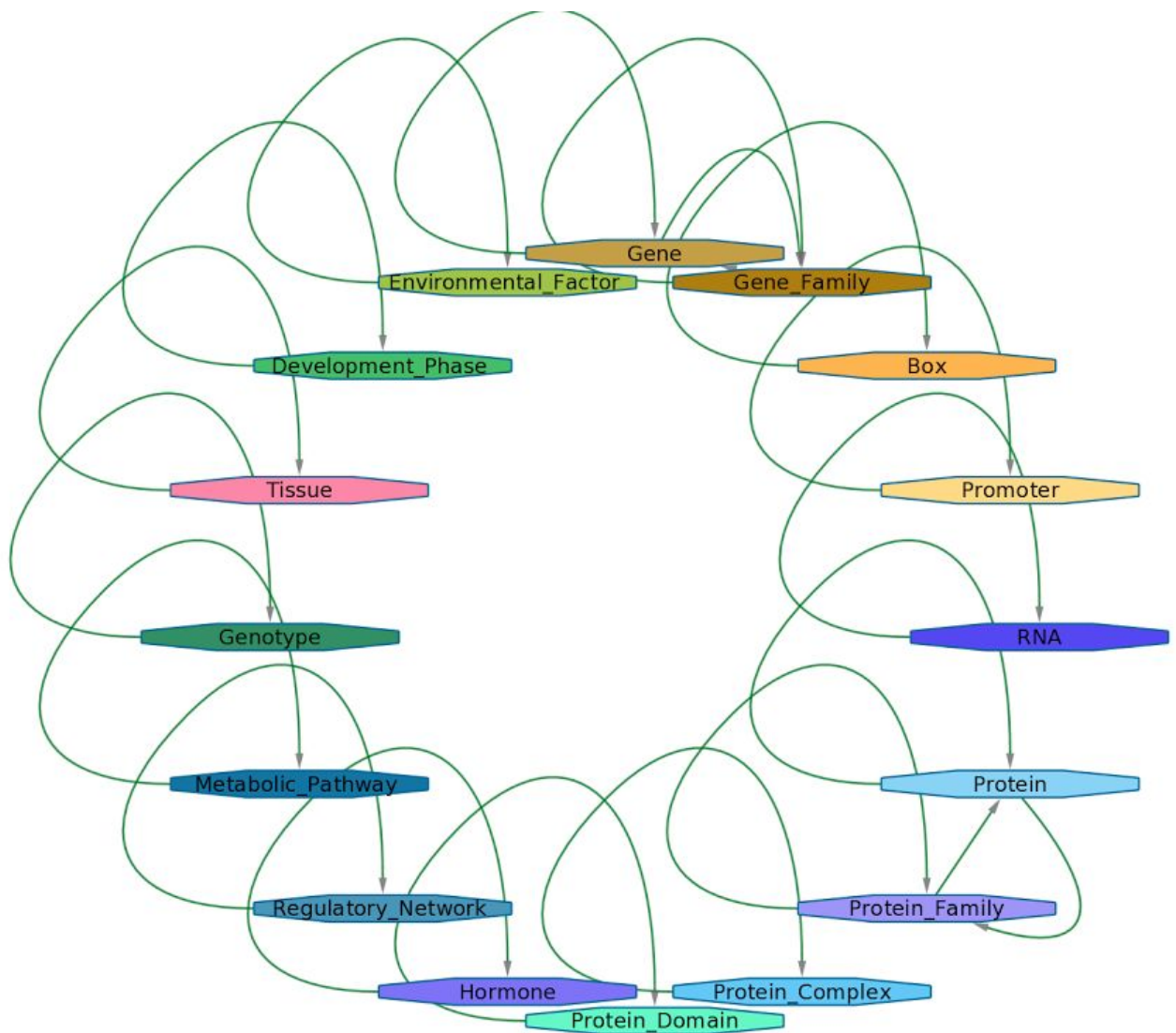
### B.3.6 Involvement\_In\_Process



### B.3.7 Transcription\_Or\_Translation

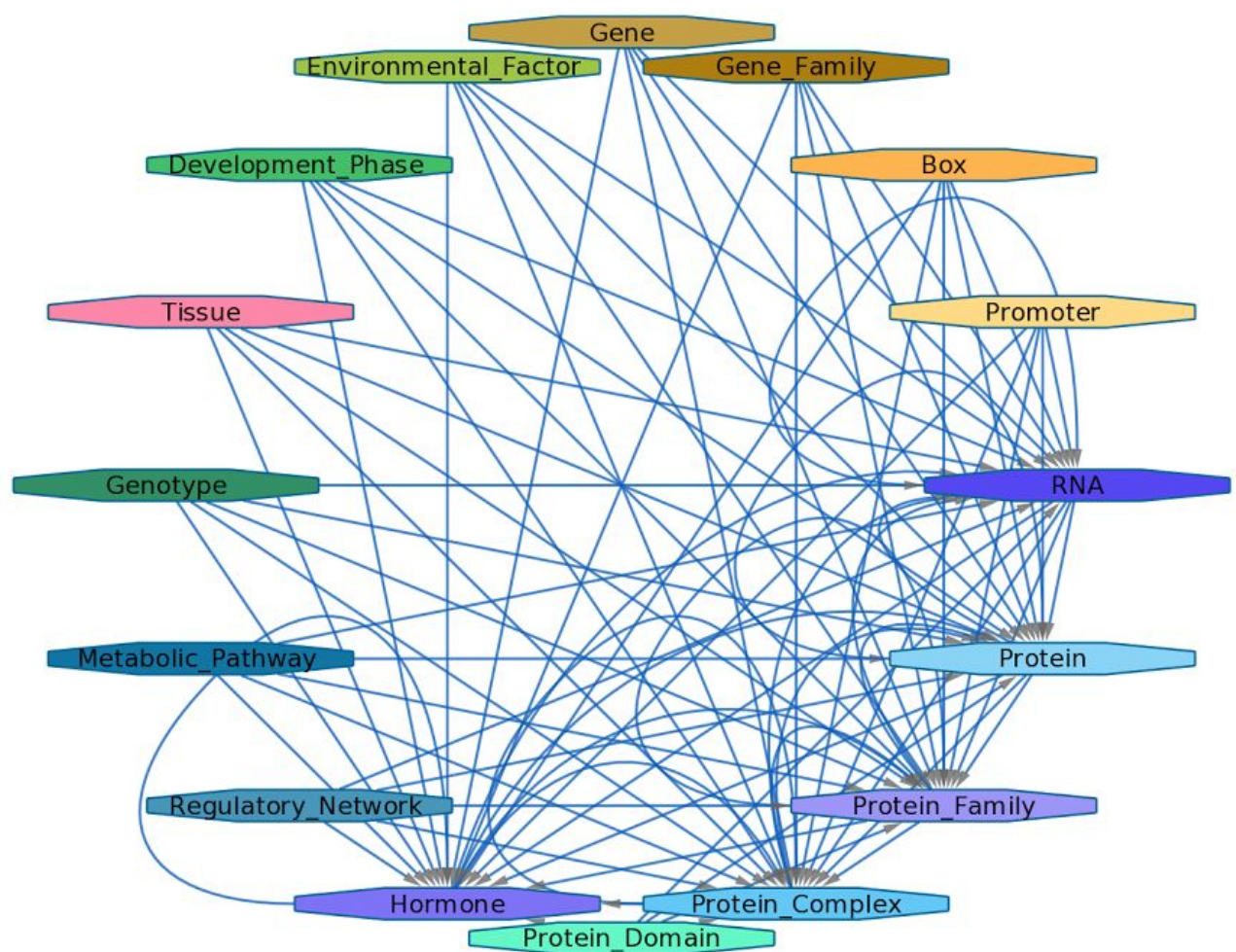


### B.3.8 Functional\_Equivalence

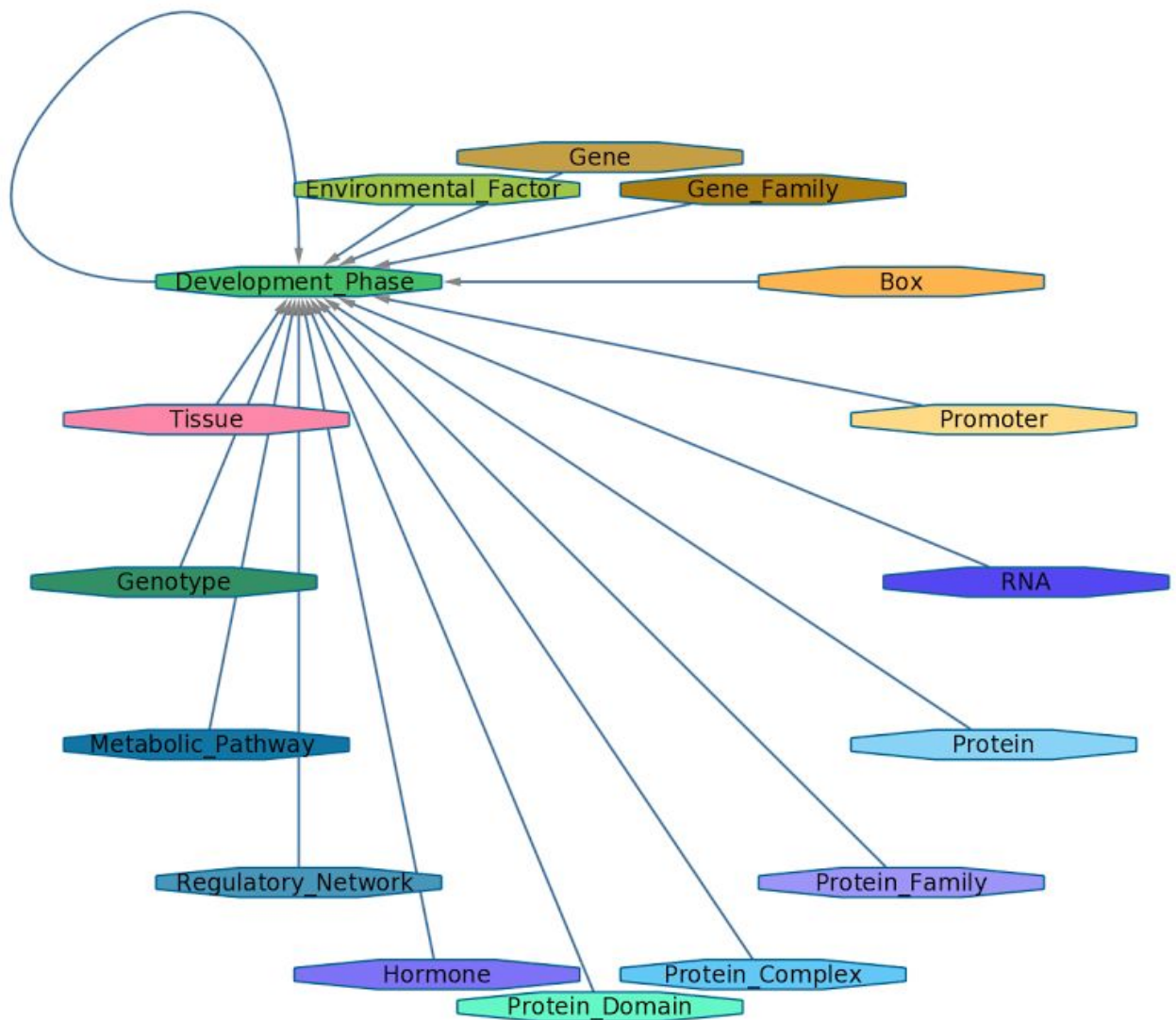




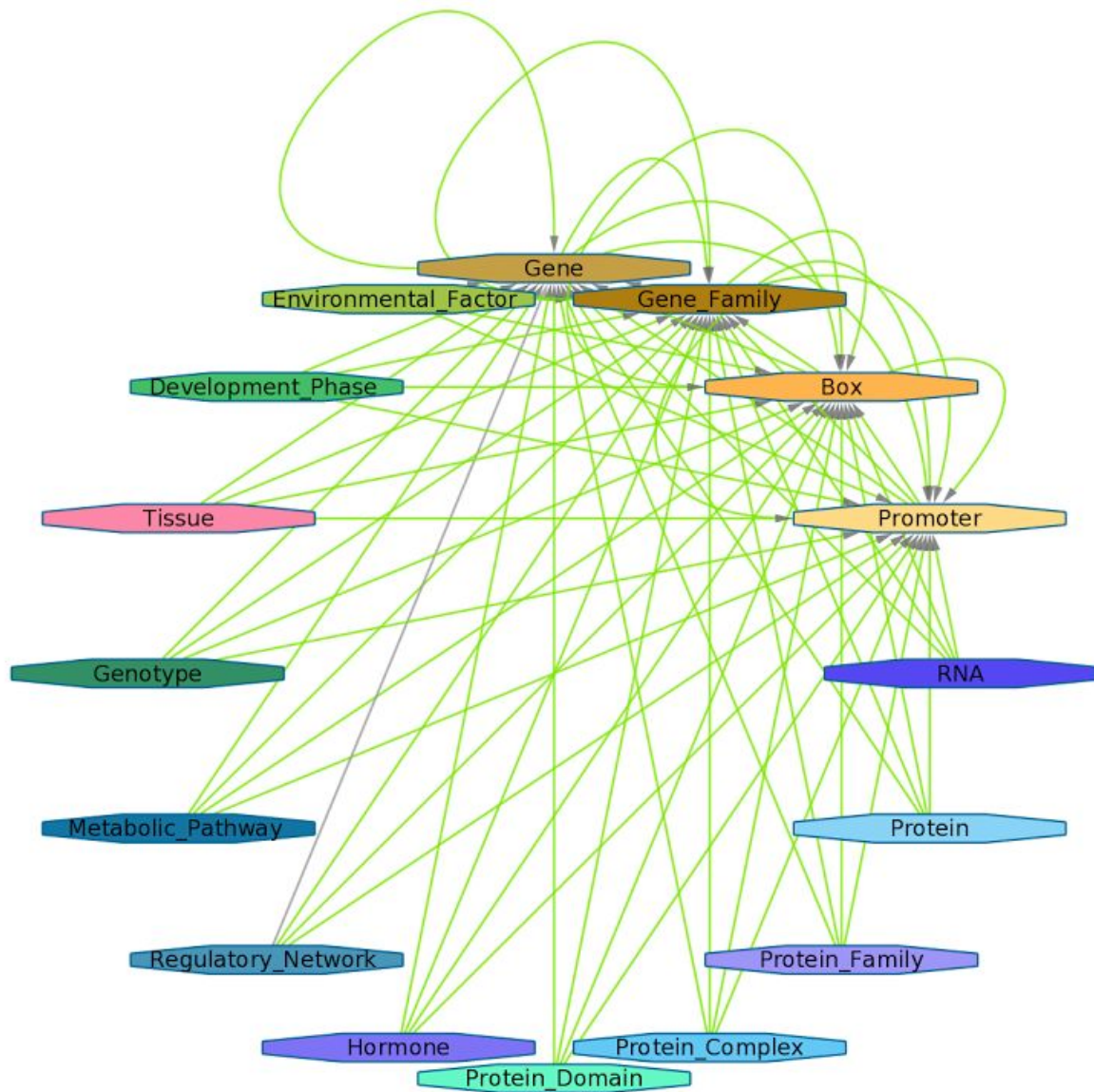
### B.3.9 Regulation\_Of\_Accumulation



### B.3.10 Regulation\_Of\_Development\_Phase

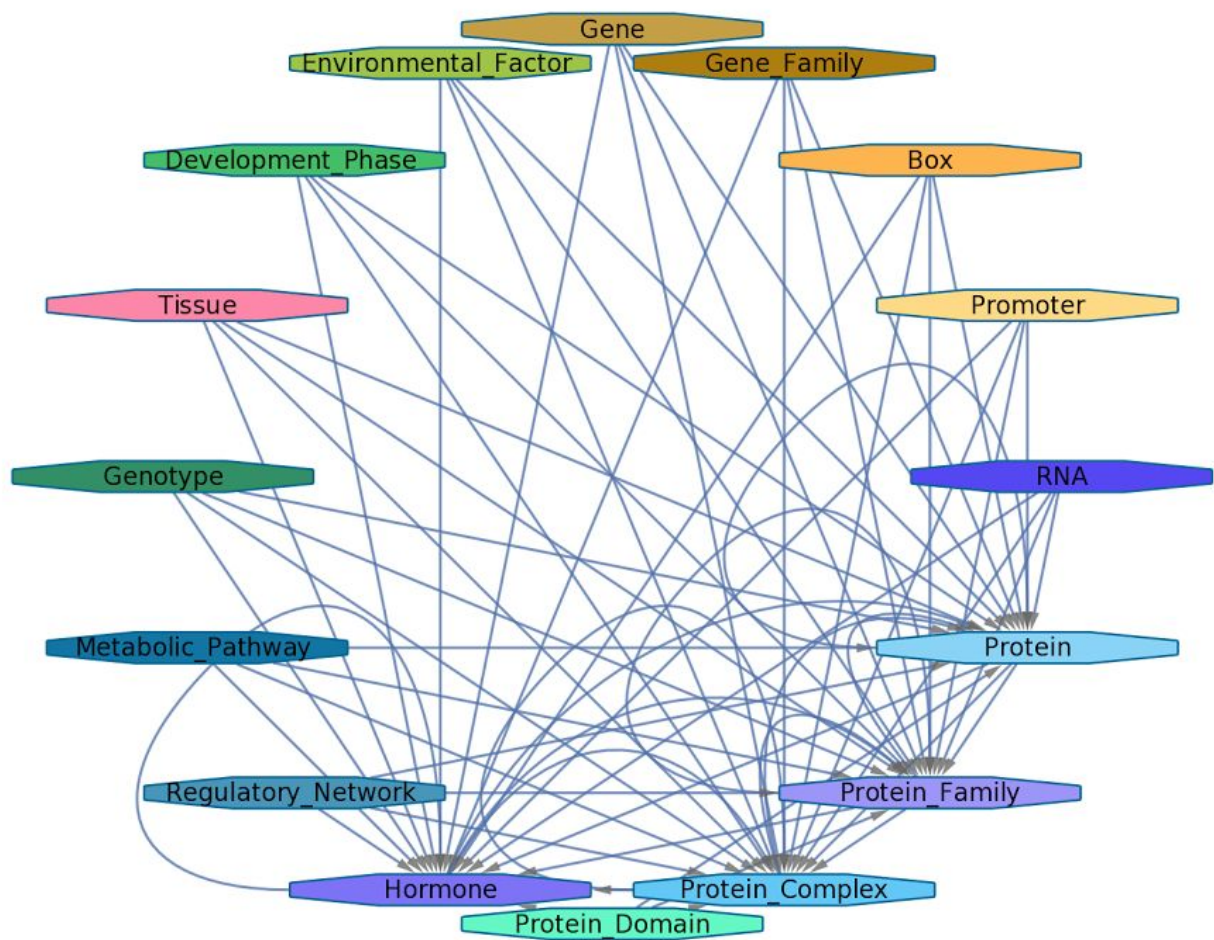


### B.3.11 Regulation\_Of\_Expression



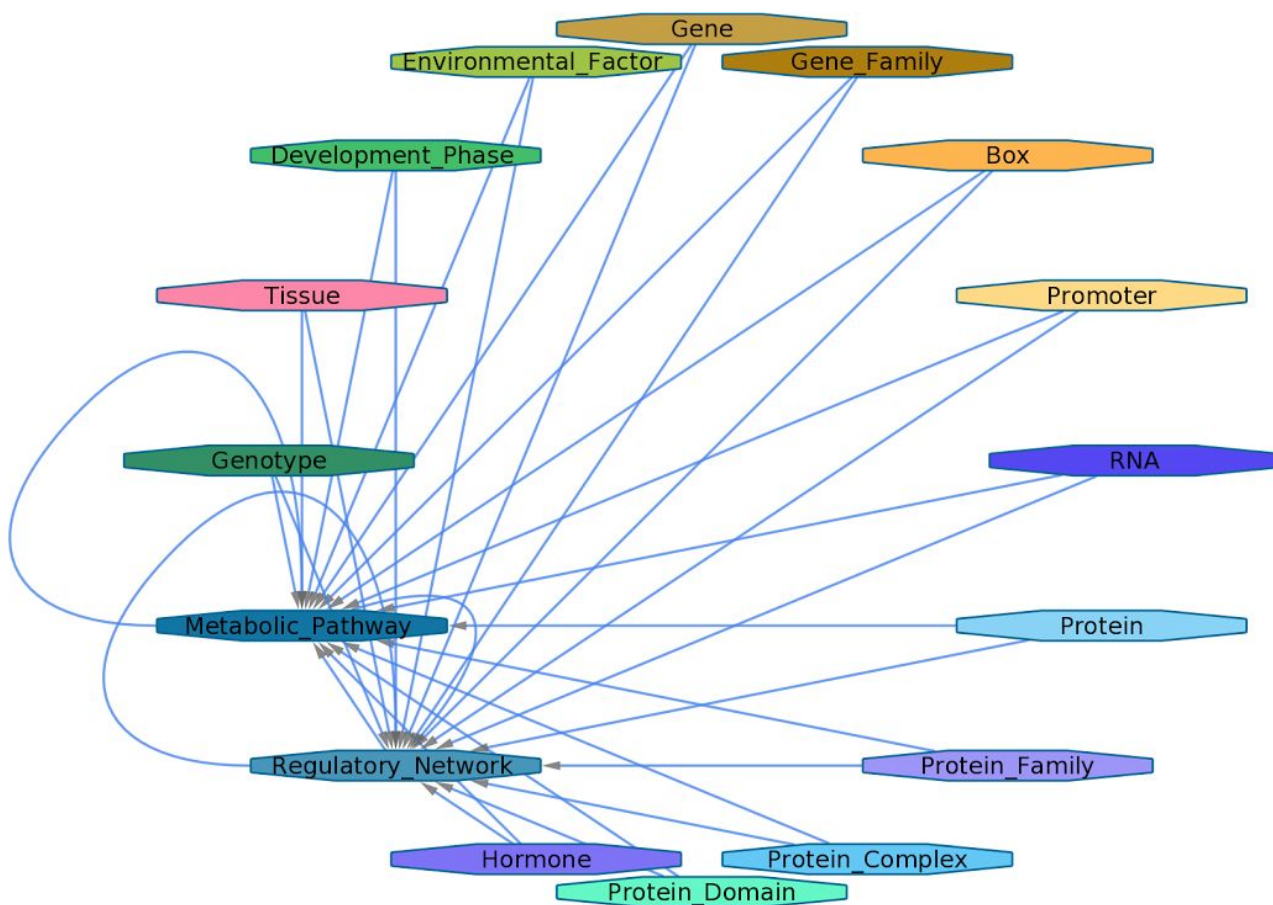


### B.3.12 Regulation\_Of\_Molecule\_Activity

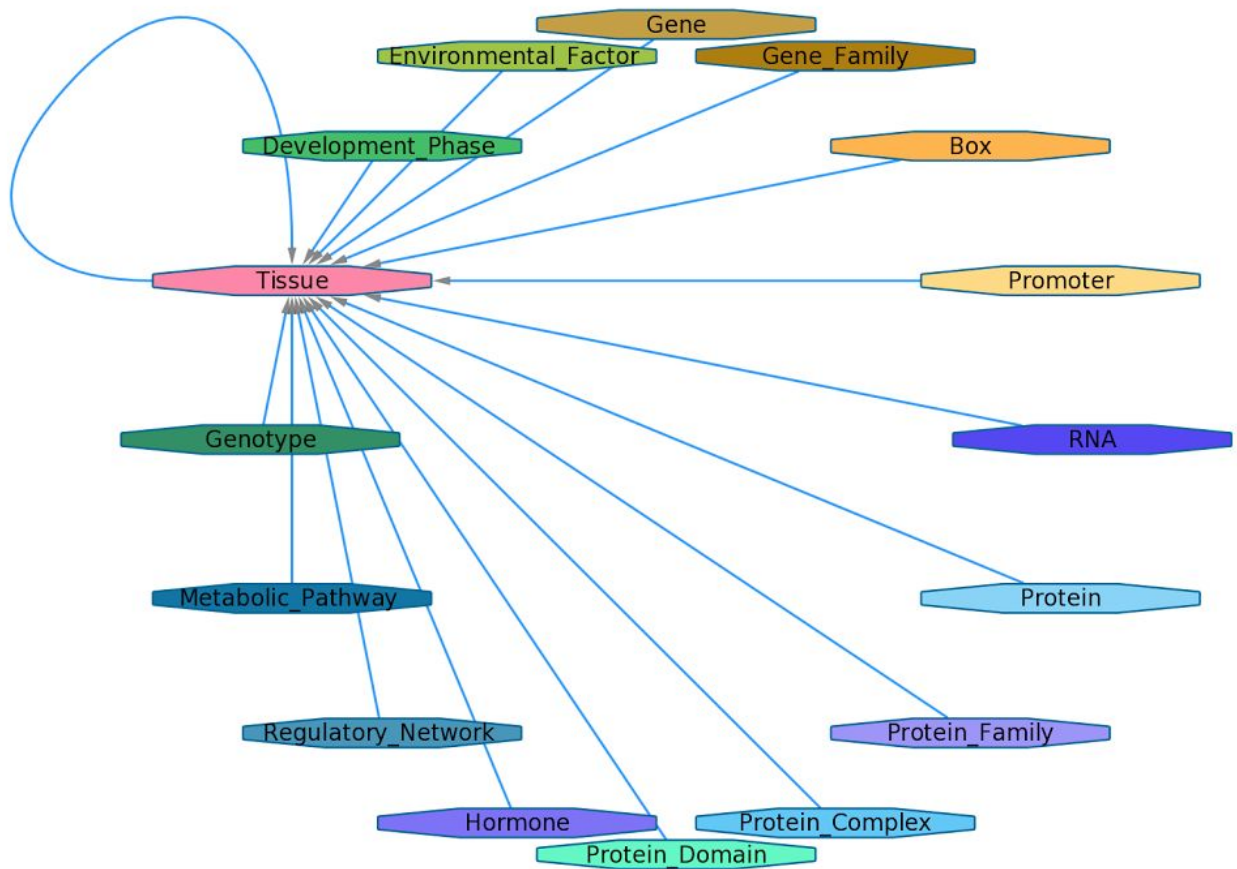




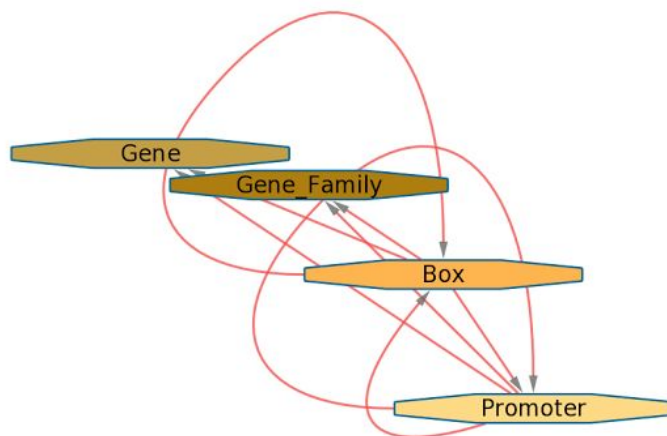
### B.3.13 Regulation\_Of\_Process



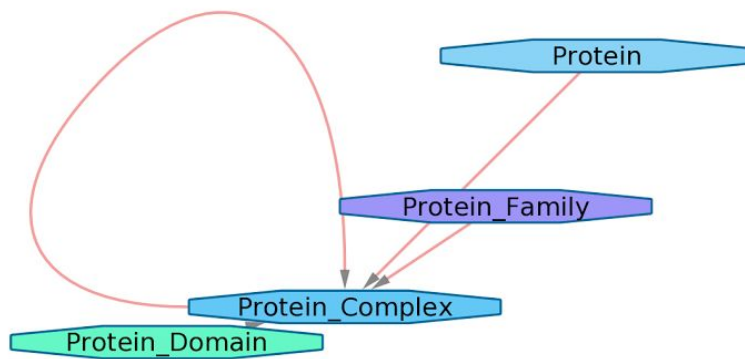
### B.3.14 Regulation\_Of\_Tissue\_Development



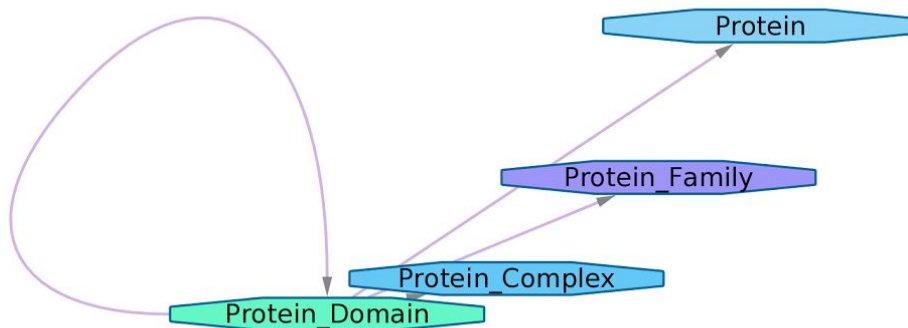
### B.3.15 Primary\_Structure\_Composition



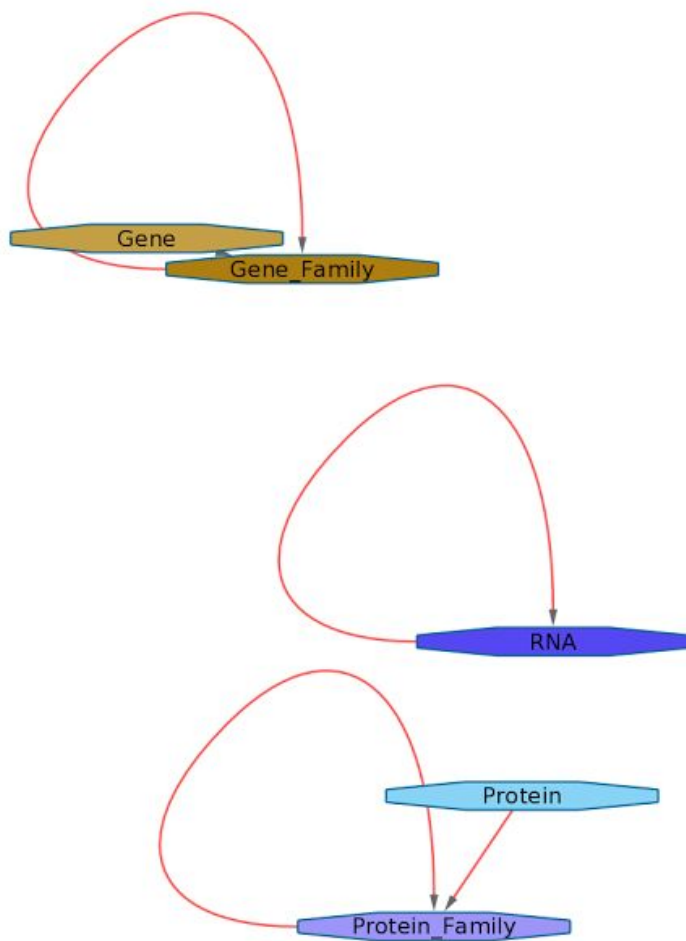
### B.3.16 Protein\_Complex\_Composition



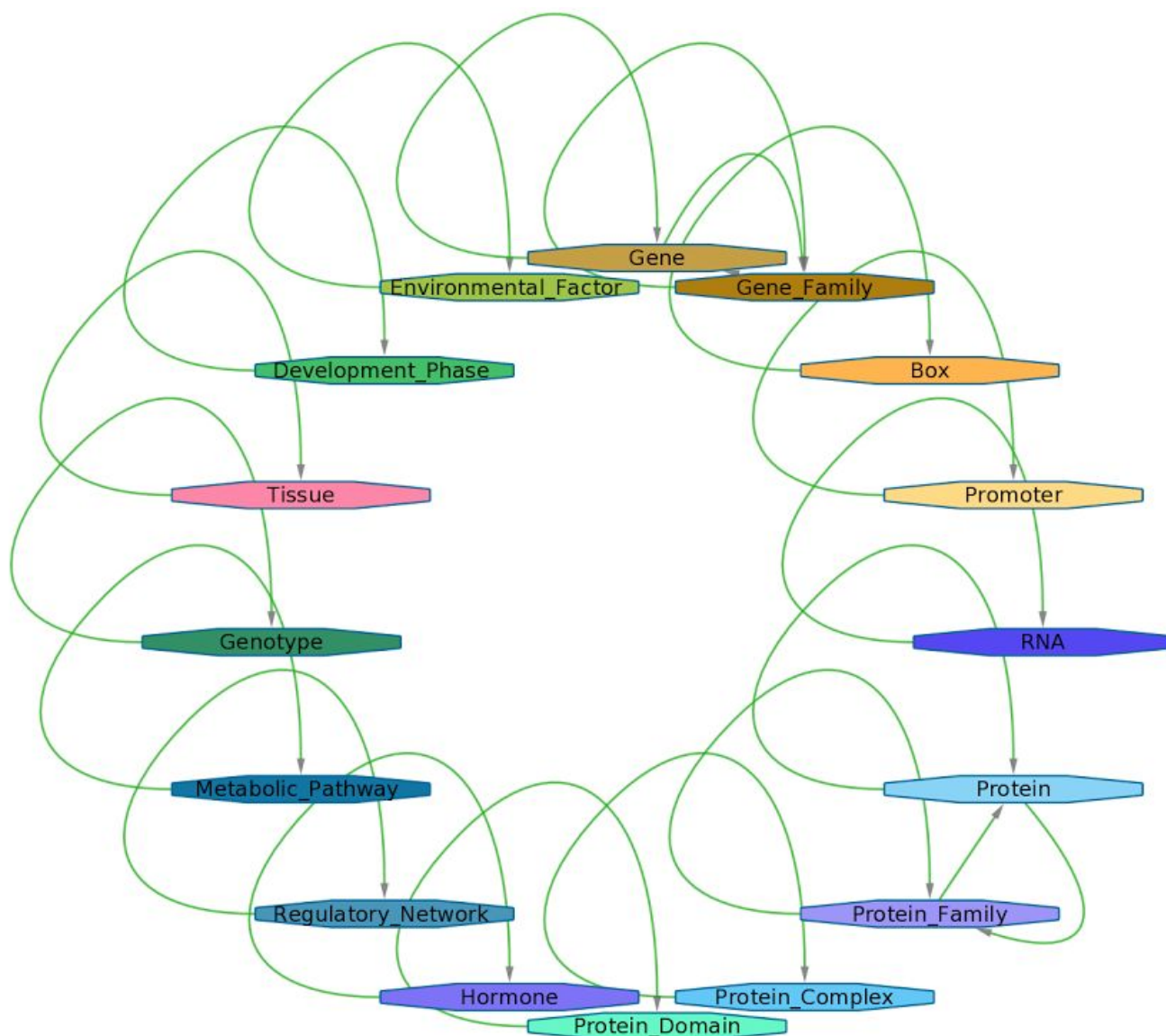
### B.3.17 Protein\_Domain\_Composition



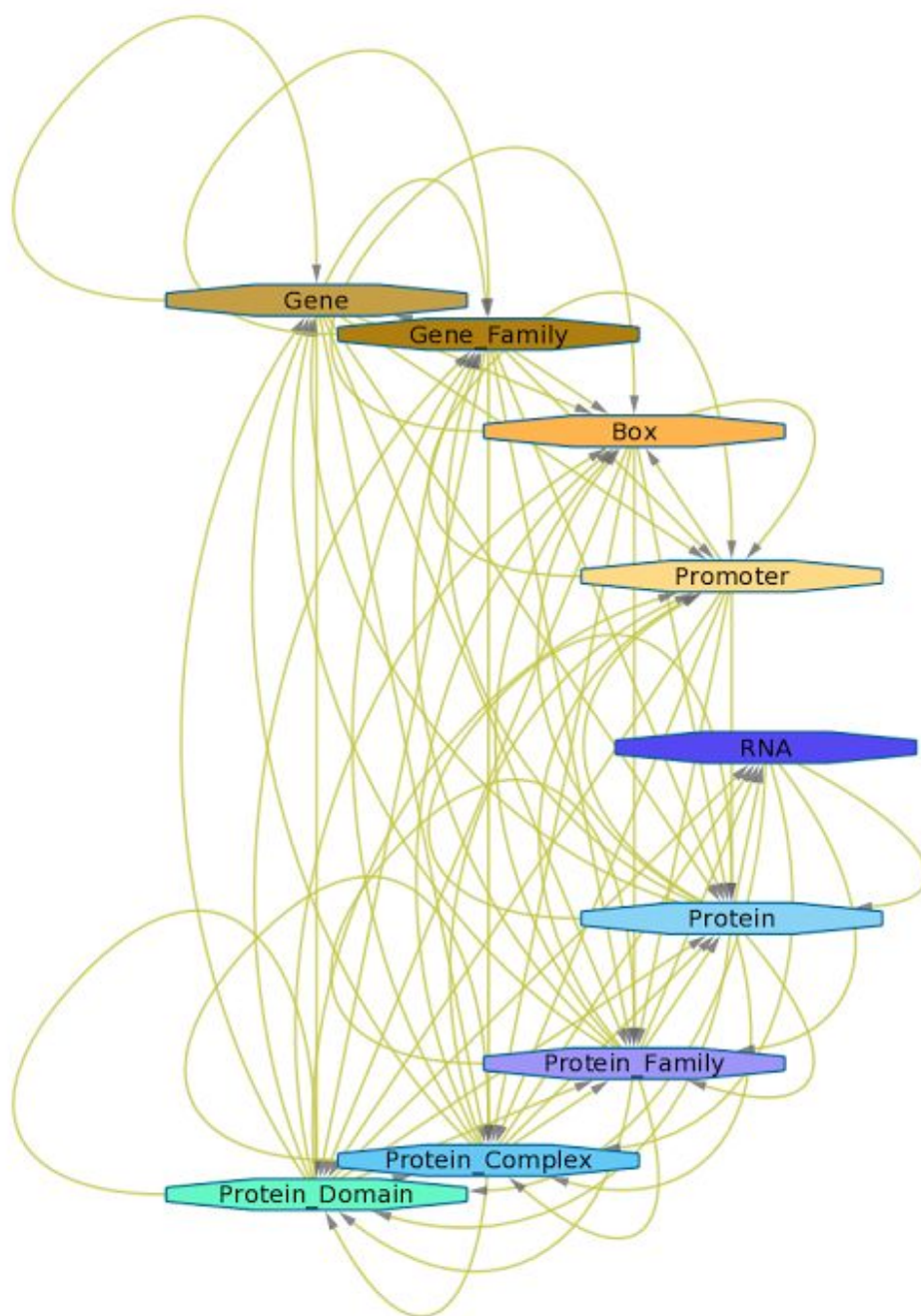
### B.3.18 Family\_Membership



### B.3.19Sequence\_Identity

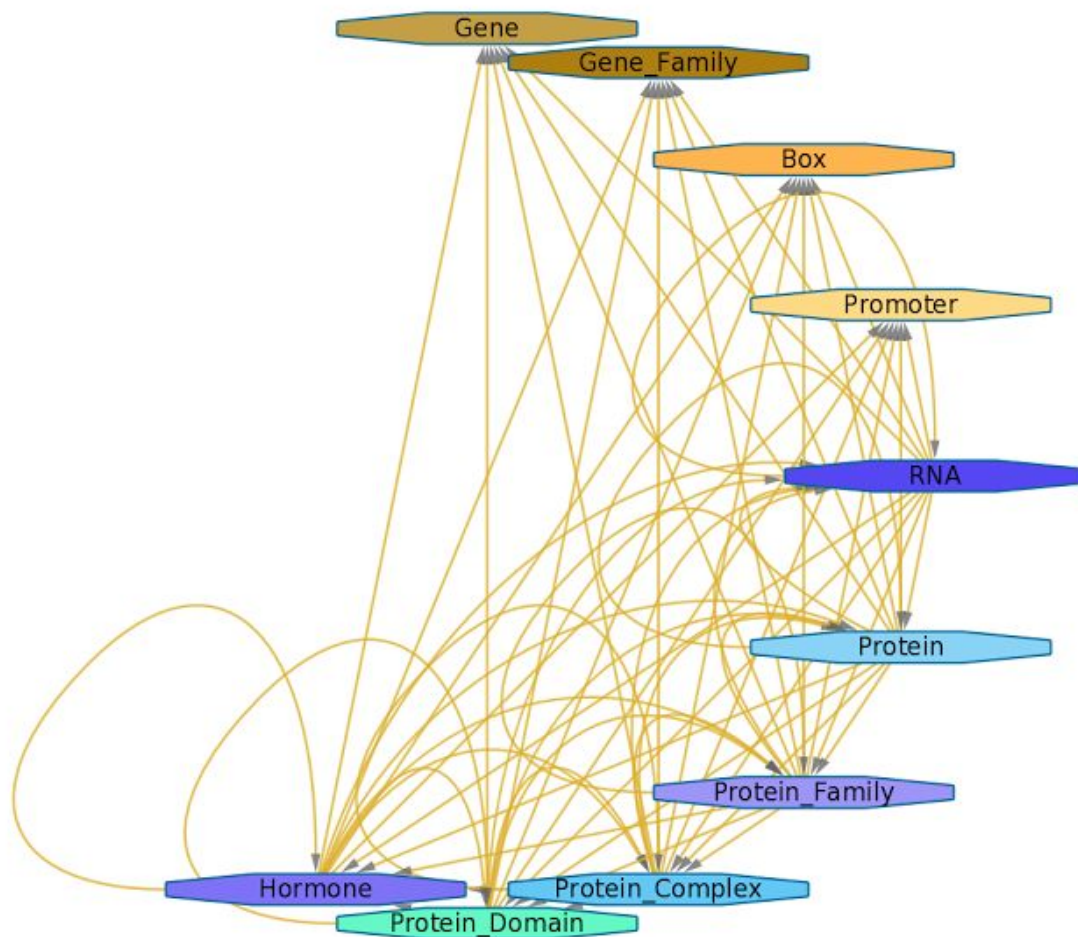


### B.3.20 Interaction



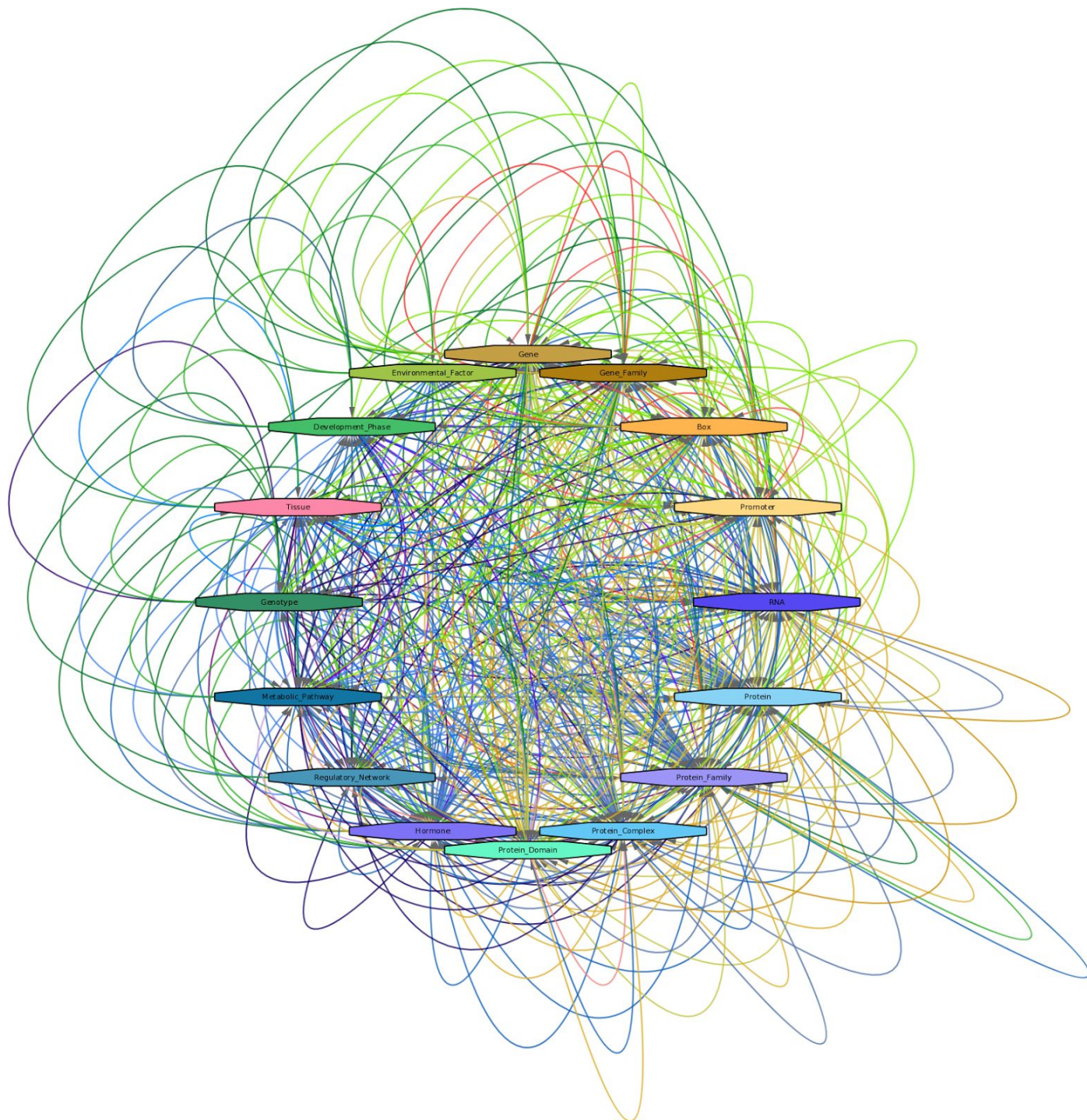


### B.3.21 Binding





### B.3.22 Visualization of all relations



## Acknowledgement

This work is supported by the French National Institute for Agricultural Research (Inra), the Center for Data Science (CDS) and the Institute for Living Systems (IMSV), funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.