



HAL
open science

Knowledge model of regulatory networks involved in Arabidopsis seed development for information extraction and integration from text

Estelle Chaix, Bertrand Dubreucq, Dialekti Valsamou, Abdelhak Fatihi, Robert Bossy, Louise Deleger, Pierre Zweigenbaum, Philippe Bessieres, Loic Lepiniec, Claire Nédellec

► To cite this version:

Estelle Chaix, Bertrand Dubreucq, Dialekti Valsamou, Abdelhak Fatihi, Robert Bossy, et al.. Knowledge model of regulatory networks involved in Arabidopsis seed development for information extraction and integration from text. BioCreative 5, Sep 2015, Madrid, Spain. hal-01512197

HAL Id: hal-01512197

<https://hal.science/hal-01512197v1>

Submitted on 2 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.

A KNOWLEDGE MODEL OF REGULATORY NETWORKS INVOLVED IN ARABIDOPSIS SEED DEVELOPMENT FOR INFORMATION EXTRACTION AND INTEGRATION FROM TEXT

Estelle Chaix¹, Bertrand Dubreucq², Dialekti Valsamou^{1,3}, Abdelhak Fatihi², Louise Deléger¹, Robert Bossy¹, Pierre Zweigenbaum³, Philippe Bessières¹, Loïc Lepiniec², Claire Nédellec¹

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

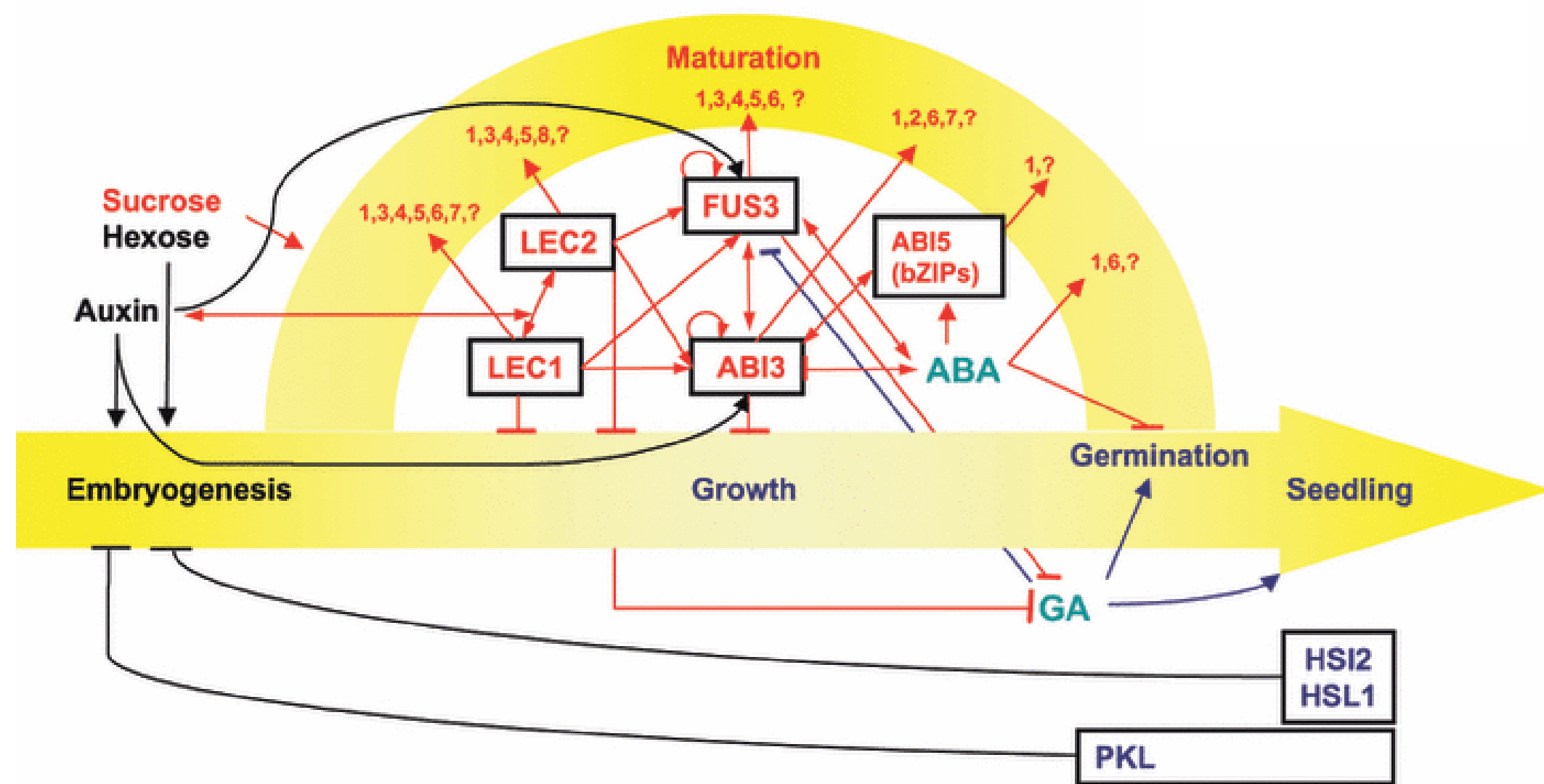


Figure 1: A model of genetic (framed) and molecular (in blue cyan) interactions involved in the control of seed development and maturation in *Arabidopsis thaliana* (from Santos-Mendoza *et al.*, 2008) (Numbers are biological processes).

Motivation

- *Arabidopsis thaliana* is a plant model in biology
- Understanding the molecular network underlying seed development regulations is important for fundamental research, agriculture and industry
- Regulatory networks for reserve accumulation and seed maturation are complex (Figure 1)
- Modeling the overall seed development process requires the reconstruction of regulatory networks at different scales (*e.g.*, genetic level, environmental factors and phenotypes)
- A large part of the information needed is spread throughout thousands of articles

➤ We propose a knowledge model of *Arabidopsis* regulatory networks to extract information from text with Natural Language Processing and Machine Learning.

Arabidopsis knowledge model

- Collaboration between information extraction specialists and biology experts
- Fined-grained and complex model : 16 biological entities and 10 relations (Figure 2)

1°) Entity description

Entities	Description
Biological Processes	
DNA-type	
Gene	"LEC1", "APETALA2"
GeneFamily	"LEC genes", "AP2-like"
Box	"WRI1 targets", "5'-GCATCG-3'"
Promoter	"BCCP2 promoter", "5' flanking regions"
DNA Product	
RNA	"CLV3 mRNA", "transcript of FLC"
Protein	"CLV1", "LEAFY COTYLEDON 1"
ProteinFamily	"MYB", "B3 proteins"
ProteinComplex	"core-binding factor", "HDAC1 complex"
ProteinDomain	"B3 domain", "basic helix-loop-helix"
Hormone	"abscisic acid", "GA"
Process	
Pathway	"fatty acid biosynthetic pathway", "glycolysis"
RegulatoryNetwork	"sensitivity to ABA", "embryonic programs"
Observed conditions	
Species and Genotype	
Genotype	"overexpression of WRI1", "Fus3 mutant"
Cell, Tissues and Organ	
Tissue	"cotyledon", "petioles of the rosette leaves"
Growth stage	
Development phase	"embryogenesis", "transition to flowering"
External factor	
Environmental factor	"long-day conditions", "cold stress"

2°) Types of relations

10 relation types were defined :

→ Regulation :

- *RegulatesActivityOf*
- *RegulatesAccumulationOf*
- *RegulatesExpressionOf*

→ Interaction :

- *InteractWith*
- *BindTo*

→ Localisation :

- *IsFoundIn*
- *IsFoundDuring*

→ Similarity :

- *Comparison*
- *Belongs to*
- *Encodes*

1 relation to define n-ary events

- *Condition*

3°) Relation arguments

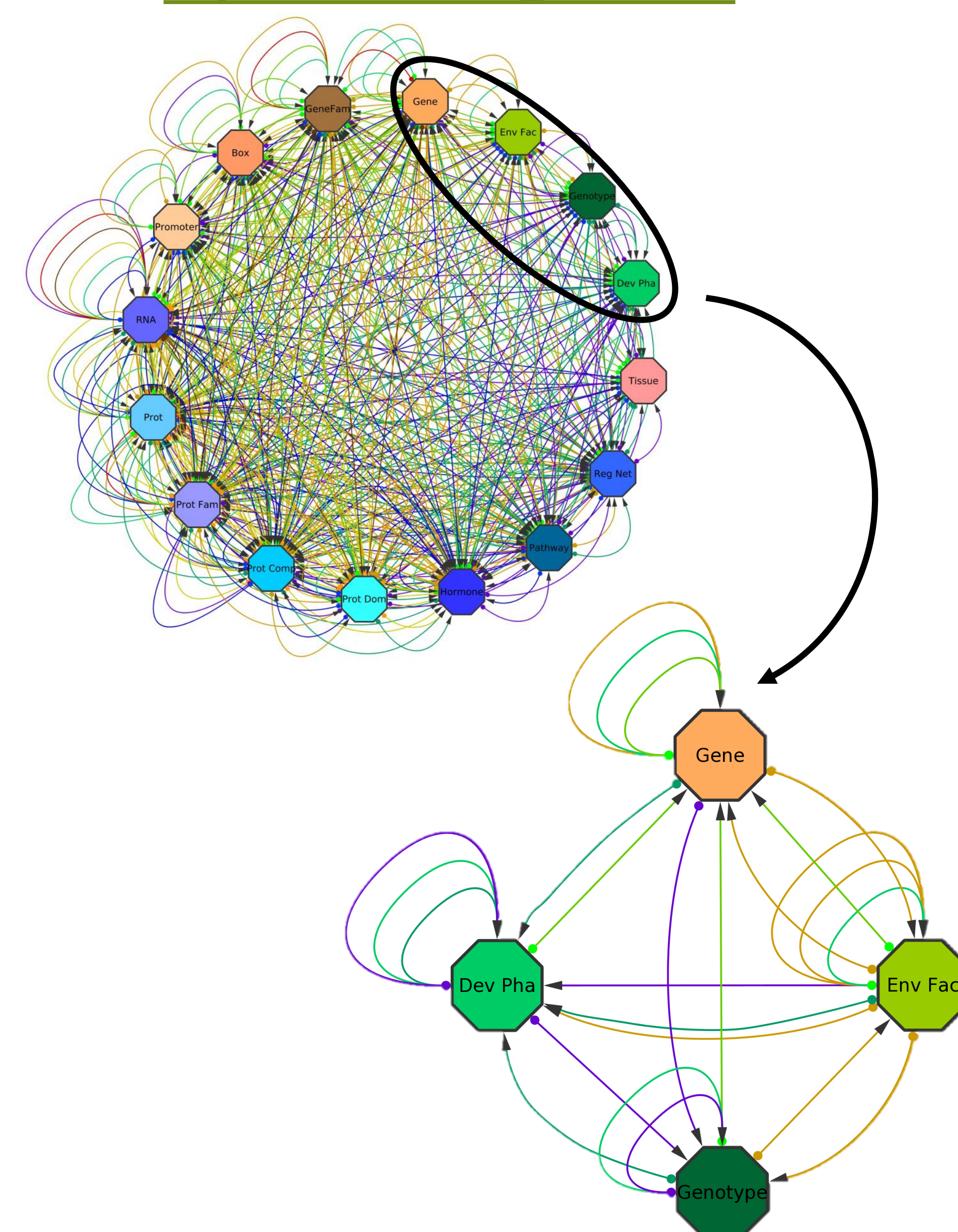


Figure 2: Schematic representation of the knowledge model

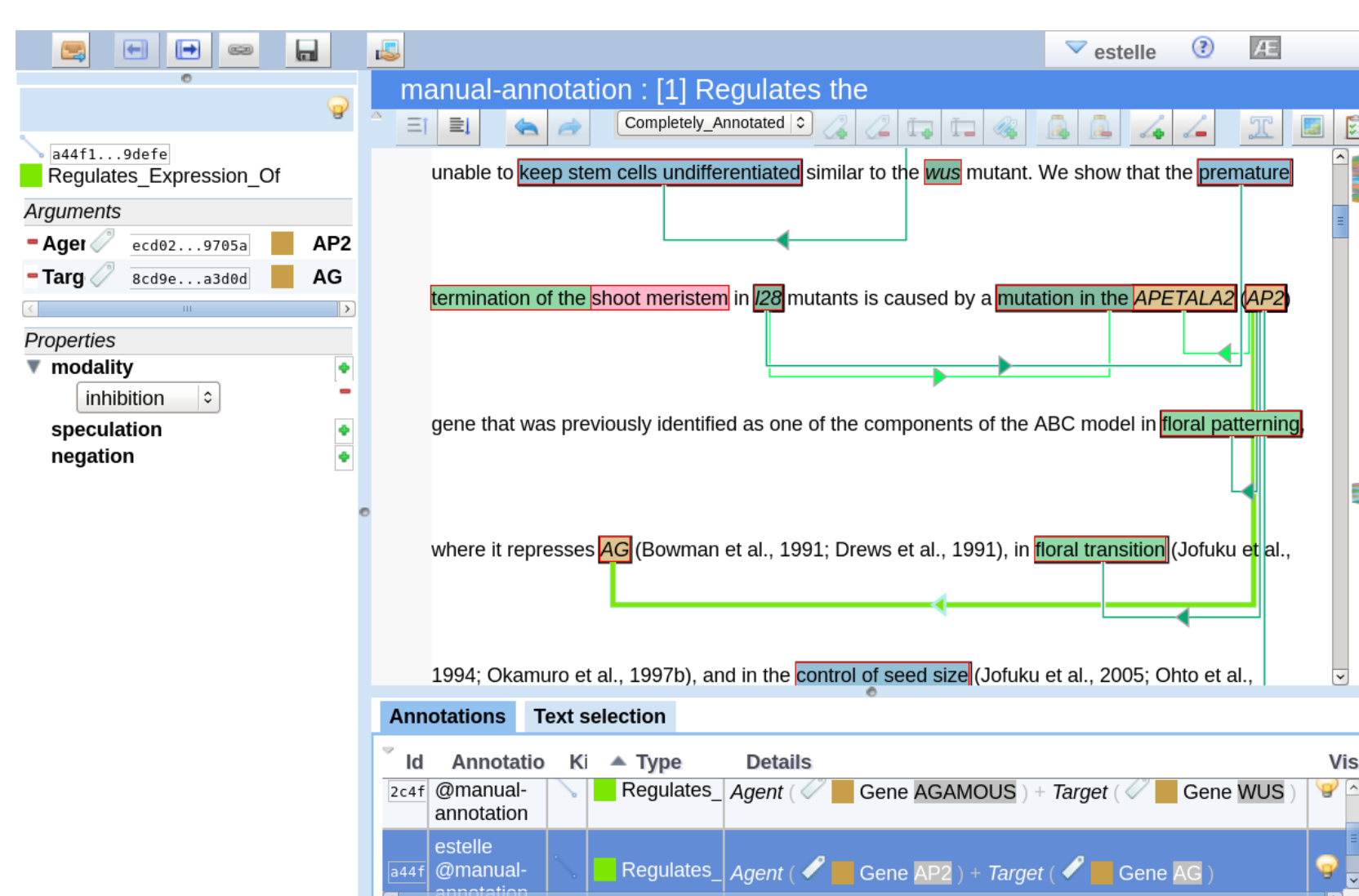


Figure 3: Annotation with AlvisAE

Annotation

- Integration of the knowledge model in the corpus annotation editor, AlvisAE (Figure 3, Papazian *et al.*, 2012)
- Annotation of a corpus of scientific articles (Figure 4) by biology experts (on-going work) : currently 4,444 entities and 1,421 n-ary events.

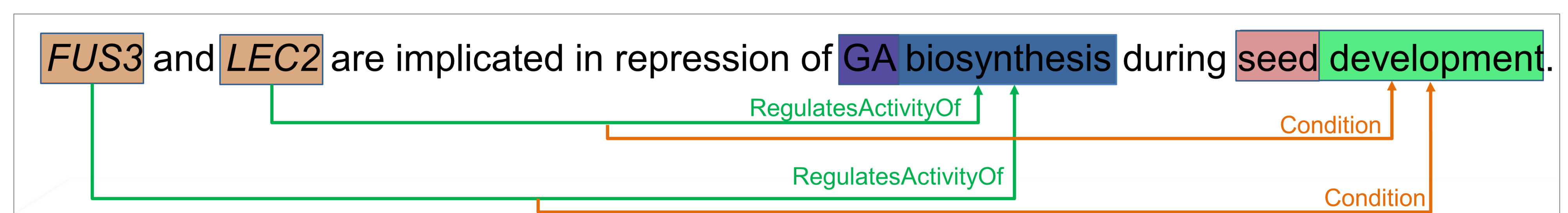


Figure 4: Example of an annotated sentence

Conclusion & Perspectives

- An accurate and complex knowledge model was developed to describe seed development of *Arabidopsis thaliana*
- We are currently training Machine-Learning methods using the *Arabidopsis thaliana* corpus
- We aim at proposing for the first time an information extraction task on *Arabidopsis thaliana* at BioNLP Shared Task 2016
- We plan to use the knowledge model to describe regulation networks in other plants and/or organisms

References

Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., & Lepiniec, L. (2008). Deciphering gene regulatory networks that control seed development and maturation in *Arabidopsis*. *The Plant Journal*, 54(4), 608-620.
 Papazian, F., Bossy, R., & Nédellec, C. (2012). AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop* (pp. 149-152). Association for Computational Linguistics.

{estelle.chaix, dialekti.valsamou, louise.deleger, robert.bossy, philippe.bessieres, claire.nedellec}@jouy.inra.fr ; {bertrand.dubreucq, abdelhak.fatihi, loic.lepiniec}@versailles.inra.fr ; pz@limsi.fr