International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:6, 2009

The Knowledge Representation of the Genetic Regulatory Networks Based on Ontology

Ines Hamdi, and Mohamed Ben Ahmed

***Abstract*—**The understanding of the system level of biological behavior and phenomenon variously needs some elements such as gene sequence, protein structure, gene functions and metabolic pathways. Challenging problems are representing, learning and reasoning about these biochemical reactions, gene and protein structure, genotype and relation between the phenotype, and expression system on those interactions.

The goal of our work is to understand the behaviors of the interactions networks and to model their evolution in time and in space. We propose in this study an ontological meta-model for the knowledge representation of the genetic regulatory networks. Ontology in artificial intelligence means the fundamental categories and relations that provide a framework for knowledge models. Domain ontology's are now commonly used to enable heterogeneous information resources, such as knowledge-based systems, to communicate with each other. The interest of our model is to represent the spatial, temporal and spatio-temporal knowledge. We validated our propositions in the genetic regulatory network of the Aarbidosis thaliana flower.

***Keywords*—**Ontological model, spatio-temporal modeling, Genetic Regulatory Networks (GRNs), knowledge representation.

I. Introduction

THE analysis of genetic regulatory networks, responsible for cell differentiation and development in prokaryotes and eukaryotes, will much benefit from the recent up scaling to the genomic level of experimental methods in molecular biology.

One of the hottest research topics in Genome Science is the interaction between genes. Genetic Regulatory Network (GRN) [1] is one of the recent focuses to understand metabolic pathways and bioprocesses. GRNs act as analog biochemical computers to specify the identity and level of expression of a group of targeted genes.

Its output is the constellation of RNAs (Ribonucleic acid) and proteins encoded by target genes. Time series expression data obtained from DNA microarrays is one of the most useful kinds of data used to construct and test GRNs.

There are numerous techniques to model GRNs or the behavior of a cell: boolean networks, Petri nets, Bayesian networks, cluster analysis etc. There are even genetic/metabolic circuit networks where genes, metabolic enzymes, and proteins are modeled as nodes with the relationship between activation, inhibition and mediation as links.

GRN models can be used to identify genetic diseases and estimate the effects of medications. Actually there few systems that are interested to the GRN modeling. Despite this systems described explicitly the modeling used techniques; they not define how represent the biological knowledge required to model the GRN. The goal of our project is to study the dynamic of the GRN by giving a spatio-temporal model, able to represent the gene expression evolution caused by external factors. Biomedical knowledge is encapsulated in tens of millions of publications with various degrees of coherence and computability. The most difficulty is to identify the biological knowledge and in particular the GRN knowledge [2]. The ontology's are a very powerful formalisms of representation knowledge domains as complex and rich that the cellular biology.

This paper will be focused to describe how resolve the knowledge representation problem in biological domain and spatially of genetic regulatory networks. We used an ontological approach by giving a GRN Ontology Design pattern (ODP). This ODP defines a spatial and temporal attributes to express the spatial and temporal knowledge's. Our ODP is validated actually by the ARABIDOPSIS THALIANA GRN.

This paper is organized as follows: section II describes the Knowledge representation in Biological Domain; section III describes the proposed approach; section IV presents our conclusions.

II. The Knowledge Representation In Biological Domain

Genes are complex structures and they cause dynamic transformation of one substance into another during the whole life of an individual, as well as the life of the human population over many generations. When genes are "in action", the dynamics of the processes in which a single gene is involved are complex, as this gene interacts with many other genes, proteins, and is influenced by many environmental and developmental factors.

The complexity of biological phenomena is primarily caused by interactions of biochemical components in the underlying bimolecular regulatory networks at different layers

Authors are with the Laboratoire de Recherche en Informatique Arabisée et Documentique Intégrée (R.I.A.D.I), Ecole Nationale des Sciences de l'Informatique. Campus Universitaire de Manouba, 2010 Manouba, Tunis, Tunisie (phone: 216 71 600 444; fax: 216 71 600 449; e-mail: ines.hamdi@riadi.rnu.tn, mohamed.benahmed@riadi.rnu.tn).

including gene regulatory networks (Fig. 1), protein interaction networks, and metabolic networks [3] [4].

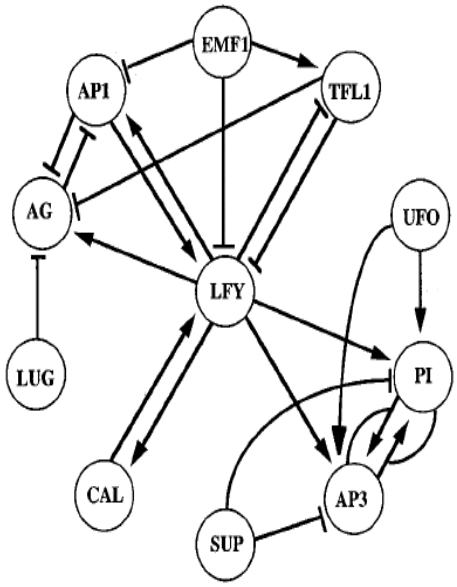


Fig. 1 An example of genetic regulatory network: Arabipdosis thaliana flower morphogenesis. Nodes represents genes, arrows and arcs represent the interactions between genes (activation, inhibition).[5]

Developments in biology and biomedicine are reported in large bibliographical databases (Table I) either focused on a specific types (*e.g.* Flybase, specialized on *Drosophila Melanogaster*) or not (*e.g.* Medline). These types of information sources are crucial for biologists, but there is a lack of tools to explore them and extract relevant information.

While recent named entity recognition tools have gained a certain success on these domains, event-based Information Extraction (IE) is still challenging. Biologists can search bibliographic databases via the Internet, using keyword queries that retrieve a large set of relevant papers.

To extract the requisite knowledge from the retrieved papers, they must identify the relevant abstracts or paragraphs. Such manual processing are repetitive and expensive in time, because of the bibliography size, the relevant data sparseness, and because the database is continually updated.

TABLE I
EXAMPLES OF BIBLIOGRAPHY DATABASES

	Medline	**Flybase**
Data Base Size	> 16 millions of references	> 9500 genes recorded
Abstract length	10 sentences	2 - 3 sentences

Medline database is the more complete than Flybase but isn't easy and simple to extract and represent biological knowledge from it. To resolve the knowledge representation problem we propose to use the ontological formalism. What are the benefits and what are the costs of using ontologies? Using ontologies can have several benefits:

- Interoperability
- Browsing/searching
- Reuse
- Structuring

III. THE ONTOLOGICAL APPROACH

Biological knowledge is evolving so rapidly that it is difficult for most scientists to assimilate and integrate the new information with their existing knowledge. One advantage of ontologies over terminological systems is to support reasoning. The formal structure and rules of inference provided by logic may be coupled with the properties of the relations among things in ontology in order to draw inferences. The uses of bioinformatics ontologies include natural language processing, knowledge discovery, and supporting interoperability among the many knowledge resources now available. In this paper we propose an ontological approach to resolve the knowledge representation of genetic networks. Our approach consists to create an ontological meta-model able to represent the biological knowledge.

This meta-model contains four layers (Fig. 3):

- The foundational ontology: is an attempt to create an ontology which describes very general concepts that are the same across all domains.
- The core ontology: is a basic and minimal ontology consisting only of the minimal concepts required to understand the other concepts.
- The domain ontology: that is an ontology with content focused on a specific aspect or facet of the world/reality or our human Knowledge/representation of it.
- The operational ontology: that is an ontology expressed in an operational language and equipped with an operational semantics.

Our meta-model contains also a spatial and temporal attributes to represent the dynamic knowledge.

Example

The inflammation process can be caused by a virus, the size of inflammation is 4 cm (Spatial data), for the duration of 1 week (Temporal data). To test our meta-model, we validated it on "Arabidopsis Thaliana" genetic network, because it is an organism model for many studies. The interaction network of Arabidopsis Thaliana flower (Fig. 2) found in [8].

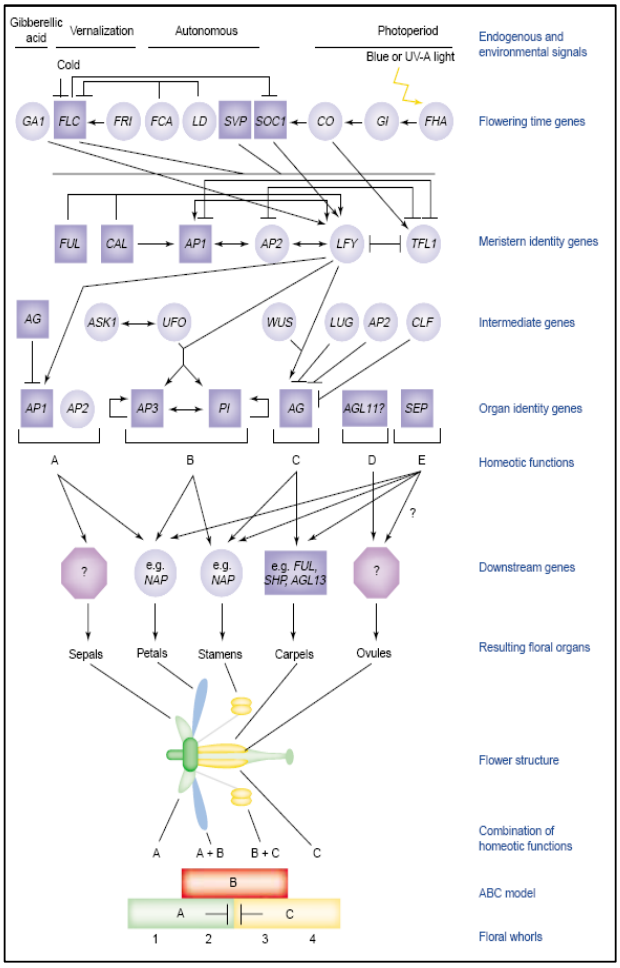


Fig. 2 The interaction network of the flower development for Arabidopsis Thaliana [8]

Floral initiation and development in the model organism *Arabidopsis* are controlled by many genes from many families that exhibit a diversity of functions, including transcription factors (e.g. *FLO/LEAFY*, *HB*. *MADS*and *YABBY*) and signal transduction genes (*CLAVATA1*, *CLAVATA3* and *KAPP*) [8]. Genes encoding bidentate ribonucleases and genes controlling cell division and chromatin structure are also important [9] [10].

MADS-box genes play key roles in specifying floral organ identity and are by far the best understood. During flower development in *Arabidopsis*, the identity of the floral organs is specified by at least three classes of homeotic genes, including the well known A-, B- and C-function genes [11][12].

Fig. 7 shows the global architecture of our system. The model-building process can be described in three main steps:

(1) Extraction and representation of spatio-temporal gene expression data in a quantitative way

(2) Modeling in terms of mathematical equations

(3) Validation with experimental results

The "knowledge extractor" will extract knowledge from TAIR (The Arabidopsis Information Resource) database [7]. TAIR maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana.

Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.

The extracted knowledge will aliment the "knowledge base". By the "ontological model" we will represent the knowledge saved in "knowledge base".

This representation will facilitate the comprehension of biological terms and concepts and be used for the dynamic modelling of the genetic regulatory networks. The "*ontological model*" will be an input for the "*spatio-temporal modeling*" module. The dynamic of the model will be detected by the "*dynamic detector*". This dynamic can be spatial, temporal or spatio-temporal.

	Sepal	Petal	Stamen	Carpel
ag	?	↑	–	–
ap3	↑	–	–	↑
pi	↑	–	–	↑
ap2	–	–	↓	↑
ap1	–	↓	?	?

Fig. 4 The diagram depicts the expected changes in levels of organ-specific transcripts in the floral mutants compared with wild-type plants. (-), absent, (↑) , upregulated; (↓), downregulated; (?), questionable.[13]

In Fig. 5, Fig. 6 and Fig. 7 examples that describe the dynamic of the gene Apetala 1 (AP1), expression took in many experiences (e.g photoperiod). It's visualised by the AtGenExpress Visualisation Tool.

This example shows the spatio-temporal dynamic of the gene expression. A gene is inhibited if its concentration (intensity) is null or lower than a threshold.

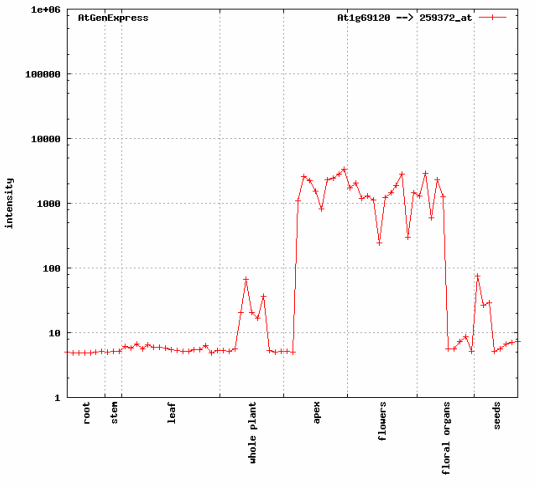
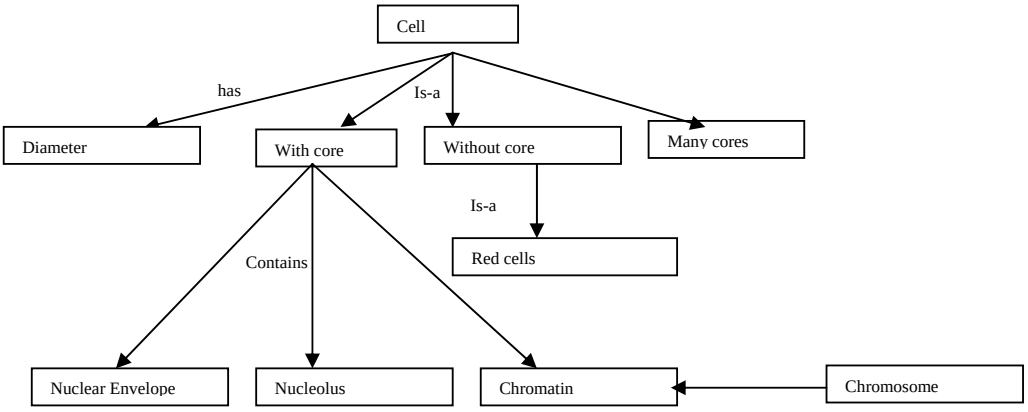
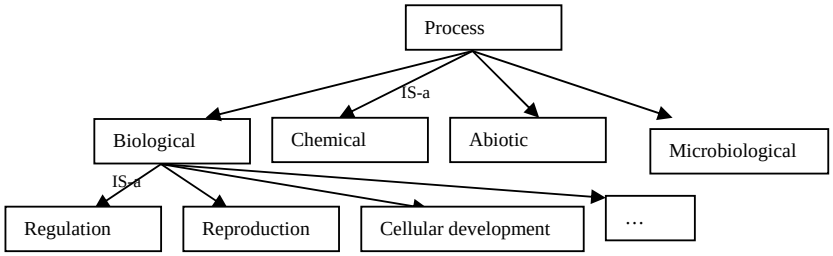


Fig. 5 The spatio-temporal dynamic of Apetala 1 (AP1) gene expression

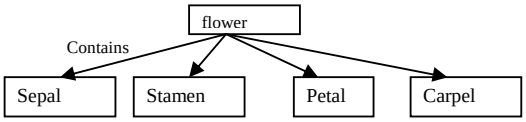
Layer 1: Fondationel ontology:



Layer 2: Core ontology:



Layer 3: Domain ontology:



Layer 4: Operationnel ontology:

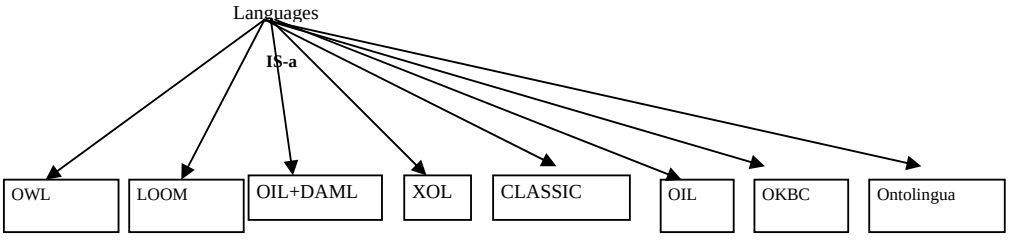


Fig 3 Examples from our proposed meta-model: it contains four layers, the first layer descripes for example the components and the types of the cell, the second layer describes the types of biological process, it can be regulation process, reproduction process…the third layer describes the domain ontology like the flower of the arabidopsis thaliana that is our study domain, the last layer concerns the artefacts and the ontological languages (OWL, XOL, OIL, etc)

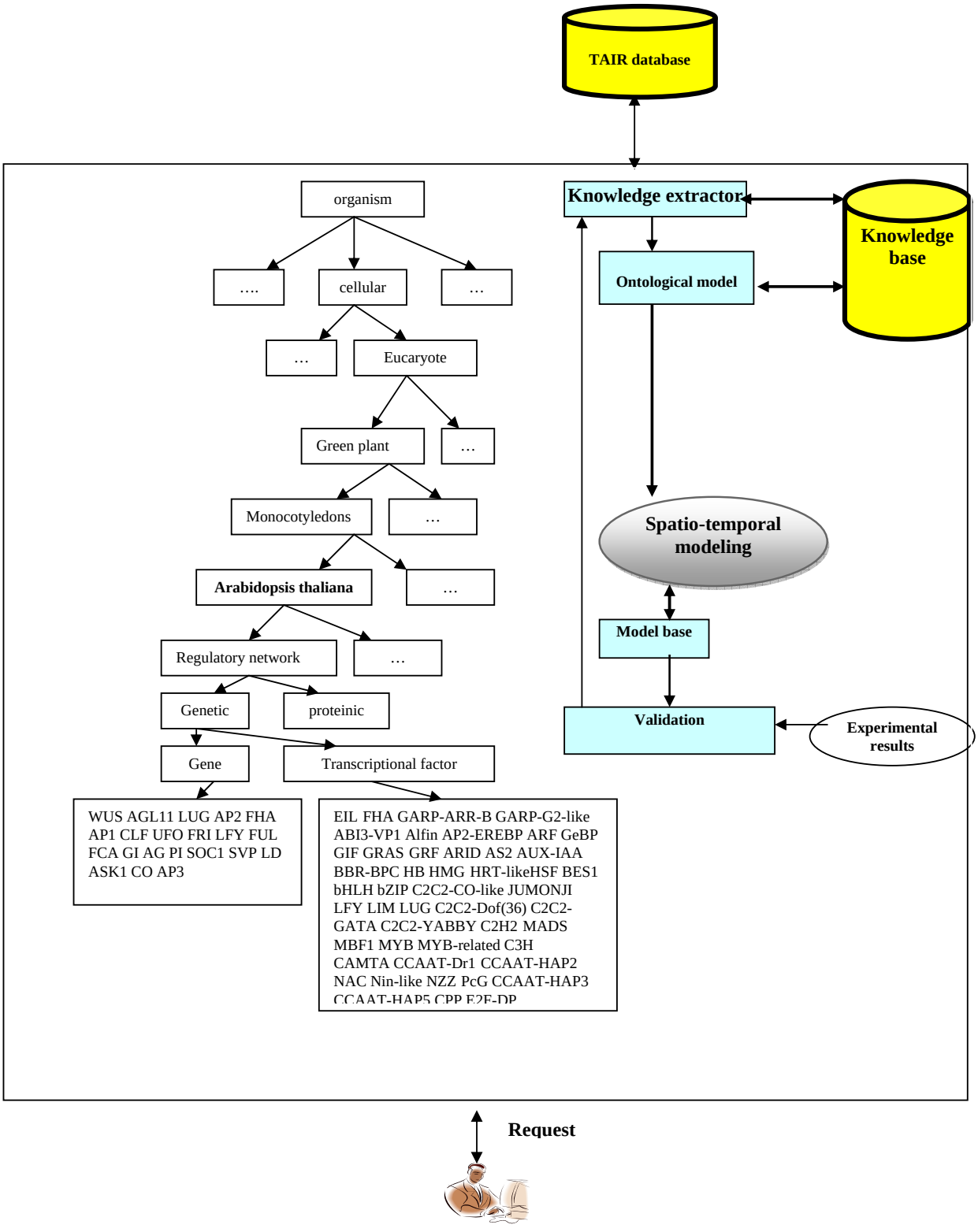


Fig. 8 The global architecture of our system

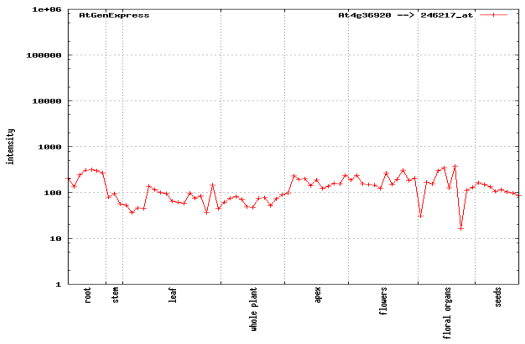


Fig. 6 The spatio-temporal dynamic of Apetala 2 (AP2) gene expression

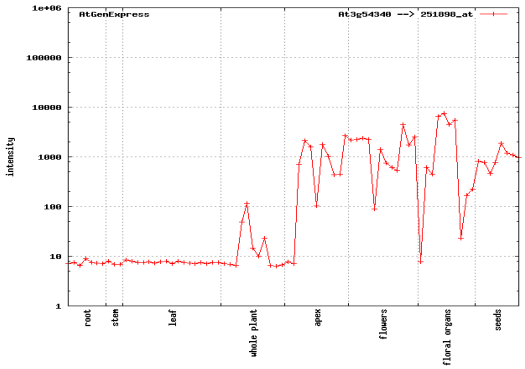


Fig. 7 The spatio-temporal dynamic of Apetala 3 (AP3) gene expression

To learn the network, the resulted model will be added to the "*model base*". Modelling pattern formation in terms of their GRN implies a description of the interactions between the different genes. Although some network structure is known, in most cases very little quantitative information is known about these interactions. Several mathematical rule-based models have been proposed to describe GRNs.

In modelling pattern formation, spatially coupled ordinary differential equations (ODEs) and partial differential equation (PDEs) have been used to describe the temporal and spatial behaviours of the genetic interaction in the system. The goal is to understand the GRNs by quantitative simulation of the model to reproduce a spatial temporal pattern obtained from experimental data.

Quantitative models are in general used to test the GRNs underlying the mechanisms behind the pattern formation and to explore some principles such as evolvability and robustness.

To validate our results we compare it with the "*experimental results*". When the result isn't correct, we must extract again the knowledge from the TAIR database to update the knowledge base and consequently the "ontological model".

IV. Conclusion

Discovering gene regulatory dependencies is fundamental for understanding mechanisms responsible for proper and pathological activity of a cell. As the complexity of gene regulatory networks under study increases so does the need for accurate modeling techniques.

Genetic regulatory networks, once constructed, can be potentially used to model the behaviour of a cell or an organism from initial conditions. Challenges of contemporary molecular biology include predicting how genes are regulated in a network, which proteins participate in metabolic pathways and how they interact. The main obstacle is how to extract and represent the knowledge of the genetic regulatory networks. Ontology's provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and widely spread application systems. Since ontology's have been developed and investigated in artificial intelligence to facilitate knowledge sharing and reuse, they should form the central point of interest for the task of exchanging user models.

This paper presents an ontological approach for semantic retrieving biological knowledge taking into account the representation of spatial and temporal knowledge. Our approach permits to overcome the obstacles of keyword-based approach in the biological domain. This work contributes the methodologies for the semantic representation, and the retrieval processes using ontology approach.

Actually we are validating our propositions on the Arabidopsis Thaliana flower. Our work has many advantages in biology, it resolves the biological knowledge representation. The use of ontology allows also sharing, reasoning, reusing and updating easily the knowledge.

References

[1] Daniel Bryce et. Seungchan Kim, "Planning for Gene Regulatory Network Intervention", IJCAI, 2007.
[2] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat. Genet., 2000, 25: 25-29.
[3] Wolkenhauer, O., Ghosh, B.K. and Cho, K.-H. (2004) Control and coordination in biochemical networks. IEEE Control Syst. Mag. 24, 30–34.
[4] Barabasi, A.-L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.
[5] Mendoza L., Alvarez-Buylla E.R. 1998. Dynamics of the genetic regulatory network for Arabipdosis thaliana flower morphogenesis. Journal of Theoretical Biology 219 :257-267.
[6] Aldo Gangemi, Domenico M. Pisanelli, Geri Steve, Ontology Integration: Experiences with Medical Terminologies, fois (1998).
[7] The official Web site of TAIR database: http://www.arabidopsis.org/
[8] Douglas E. Soltis et al., "Missing links: the genetic architecture of flower and floral diversification", TRENDS in Plant Science Vol.7 No.1 January 2002.
[9] Jacobsen, S.J. et al. (1999) Disruption of an RNA helicase/RNAse III gene in Arabidopsis causes unregulated cell division in floral meristems. Development 126, 5231–5243.
[10] Bernstein, E. et al. (2001) Role for bidentate ribonuclease in the initiation step of RNA interference. Nature 409, 363–366
[11] Bowman, J.L. et al. (1989) Genes directing flower development in Arabidopsis. Plant Cell 1, 37–52.
[12] Coen, E.S. and Meyerowitz, E.M. (1991) The war of the whorls: genetic interactions controlling flower development. Nature 353, 31–37.
[13] FrankWellmer, Jose´ Luis Riechmann, Marcio Alves-Ferreira, and Elliot M. Meyerowitz, "Genome-Wide Analysis of Spatial Gene Expression in Arabidopsis Flowers", The Plant Cell, Vol. 16, 1314–1326, May 2004.