Reverse Engineering of A. thaliana Gene Regulatory Networks

Blagoj Ristevski

Faculty of Information and Communication Technologies – Bitola
"St. Kliment Ohridski" University – Bitola, Republic of Macedonia
blagoj.ristevski@fikt.edu.mk

Abstract. In this paper, biological background of cells' regulatory mechanisms as well as gene regulatory networks are described. The models applied to infer gene regulatory networks such as Boolean networks, dynamic Bayesian networks, graphical Gaussian models and the novel two-stage model based on integration of *a priori* biological knowledge are described. These inference models are applied on Arabidopsis thaliana time series gene expression data subset. To compare models inference capabilities, ROC and AUC value as validation criteria are used. Some further directions for inference of gene regulatory networks as well as microRNA-mediated networks and model development are given at the end of the paper.

Keywords: reverse engineering, gene regulatory networks, bioinformatics, computational biology, model validation.

1 Introduction

Regulation of cells' biological, biochemical, physiological and molecular processes is carried out by inherited information contained in the organisms' genome. The numerous components of biological systems such as DNA, RNA, proteins and metabolites mutually interact composing complex networks named as **gene regulatory networks** (GRNs).

Genes, as fundamental physical and functional inheritance units of every living organism, can be coding or non-coding genes. The coding genes are templates for protein synthesis, while other genes might specify RNA templates as machines for production of different types of RNAs. The process in which DNA is transcribed into mRNA and a protein is produced by translation represents the well-known central dogma in molecular biology. The first stage is transcription and in the second stage, mRNAs translate into a sequence of amino acids composing the proteins. When a protein is produced, the corresponding coding gene is expressed. The gene expression levels indicate the approximate number of produced RNA copies from corresponding

gene, which means that gene expression level corresponds to the amount of produced proteins.

To obtain gene expression data experimentally of many genes in a sample, high-throughput technologies such as DNA microarray, serial analysis of gene expression (SAGE), quantitative polymerase chain reaction (qPCR), as well as RNA-Sequencing (RNA-Seq) are used. Besides gene expression data, other data such as protein-DNA, protein-protein interaction data and microRNAs should be included in reverse engineering of gene regulatory mechanisms.

One of the most important protein regulatory functions is transcription regulation. Transcription factors (TFs) are proteins that bind to DNA sequences and regulate the DNAs transcription and thus gene expression level. TFs can have inhibitory or activator role of expression of the target genes [1].

In the last decade, several small noncoding RNAs such as microRNAs and siRNAs are disclosed [3]. The length of a nucleotide thread in microRNAs is about 18-25 nucleotides [5]. MicroRNAs cause transcription cleavage or translation repression by connecting to their target mRNA [6] and they regulate expression level by more than 30% of coding genes [2] [4]. Besides TFs, microRNAs mutually interact with more cis-regulatory elements making microRNAs important components in the gene regulation.

In order to reveal the TF binding sites (TFBSs) in genome for particular proteins and to reveal protein-DNA interactions, chromatin immunoprecipitation (ChIP) is used [8]. ChIP-chip technology uses ChIP with hybridization microarrays (chip) to identify the protein binding sides and their locations throughout the genome. In ChIP-chip technology, short DNA sequences as probes are used [7]. Differently from ChIP-chip technology, ChIP-Sequencing (ChIP-Seq) technology uses secondary sequencing of DNA instead of microarray [8]. By integration of abovementioned types of biological data, the GRNs inference significantly can be improved [10].

There are several types of regulatory relationships between genes, TFs and microRNAs: TF regulates gene expression (TF-gene), TF regulates microRNA (TF-microRNA), microRNA regulates TF (microRNA-TF), microRNA regulation on gene expression (microRNA-gene), gene-gene regulatory interactions (GRNs) and protein-protein interactions (PPIs, TF-TF) [16].

Regarding GRNs models, there are two types of models: models based on mechanistic (physical) level and influence-based models. The first approach uses protein-DNA and protein-protein interactions data, while the second models infer GRNs from gene expression data [9]. The structure of GRNs is depicted by graphs consisted of genes, proteins, metabolites, their complexes and modules represented by **nodes**, and **edges** representing existing interactions among nodes [30].

To compare the inference capabilities of most commonly used models: Boolean networks, Dynamic Bayesian Networks, Graphical Gaussian models as well as the two-stage inference model based on integration of *a priori* knowledge, A. thaliana time series gene expression data are employed for GRNs inference. As validation criteria, ROC and AUC value are used.

The rest of this paper is organized as follows. In Section 2, experimental time series gene expression data from A. thaliana and applied models for GRNs inference are

described. The following section depicts inferred GRNs and discussion about models used for inferring of GRNs. Finally, Section 4 provides concluding remarks and further directions in the GRNs inference.

2 Materials and Methods

2.1 A. thaliana Gene Expression Data

Arabidopsis thaliana is the first plant whose genome was completely sequenced. As input dataset, A. thaliana gene expression datasets with 30185 genes were employed. Gene expression data were obtained from 41 individuals and measured in four different experimental conditions. These datasets contain 660 missing values. More information about experimental conditions for these experimental datasets is provided in [31]. After preprocessing of experimental data, transformation and normalization was applied on gene expression data. To be suitable for GRNs inference, a subset of gene expression data was extracted. Based on available knowledge about genes involved in floral organ specification [32], the following 13 genes are selected: AP3, PI, AP1, UFO, FT, FUL, TFL1, AP2, EMF1, AG, WUS, LFY1 and SEP4. The names of selected genes, their DAVID (Database for Annotation, Visualization and Integrated Discovery) names [33] and their TAIR (The Arabidopsis Information Resource) identifiers [34] are shown in Table 1.

Table 1. A subset of 13 A. thaliana genes involved in floral organ specification.

Genename	DAVID gene name	TAIR_ID
AP3	Floral homeotic protein APETALA 3	AT3G54340
PI	Floral homeotic protein PISTILLATA	AT5G20240
AP1	Floral homeotic protein APETALA 1	AT1G69120
UFO	Protein UNUSUAL FLORAL ORGANS	AT1G30950
FT	Protein FLOWERING LOCUS T	AT1G65480
FUL	Agamous-like MADS-box protein AGL8	AT5G60910
TFL1	Protein TERMINAL FLOWER 1	AT5G03840
AP2	Floral homeotic protein APETALA 2	AT4G36920
EMF1	AT5G11530	AT5G11530
AG	Floral homeotic protein AGAMOUS	AT4G18960
WUS	Protein WUSCHEL	AT2G17950
LFY1	Protein LEAFY	AT5G61850
SEP4	Agamous-like MADS-box protein AGL3	AT2G03710

2.2 Models for GRNs Inference

The emergence of experimental technologies for studying cell regulatory mechanisms such as DNA microarrays, ChIP-chip and ChIP-Seq has provided large amounts of protein-protein, protein-DNA, gene expression and other -omics data. Because the experimental technologies cannot measure mutual influences among all genes from

one organism's genome simultaneously, computational methods are applying to reverse engineer and uncover mutual gene relationships. In the past decade, several models for GRNs inference have been developed, which are based on the basic reverse engineering methods.

Boolean networks

One of the simplest models of GRNs is the model based on Boolean networks. The genes are represented by nodes and the edges representing the interactions among genes. In this model, gene expression levels are discretized and presented by two-state levels. The state of the genes that have expression levels above a certain threshold is 1, otherwise 0. Boolean networks model simplify the structure and dynamics of gene regulation. Several extended models based on Boolean networks have been proposed. A REVerse Engineering ALgorithm (REVEAL) constructs a Boolean network of given expressed gene data by setting the in-degree value of genes [15]. This algorithm extracts minimal network structures by using the mutual information approach from the state transition tables of the Boolean network. REVEAL can be applied to gene expression data, discretized on multiple discretization levels. On the other hand, multiple discretization levels increase the number of possible state transitions. For greater in-degree value, it is necessary to perform parallel processing or to increase the efficiency in space searching of all possible networks [15].

Bayesian networks

Bayesian networks (BNs) are among the most effective models for GRNs inference. A Bayesian network is a special graph model defined as a triple (G, F, q), where G denotes the graph structure, F is a set of probability distributions and q is the set of parameters for F [12]. The graph structure G consists of a set of n nodes and a set of directed edges between them. The nodes correspond to the random variables and directed edges show the conditional dependences between the random variables. Nodes and edges together create a directed acyclic graph (DAG). One directed graph is acyclic if there is no pathway that starts and ends at the same node. The joint probability distribution of all nodes is computed by multiplying of the local probability distributions. This kind of factorization of the joint probability distribution on multipliers provides its easier computation as a product of simpler conditional probability distributions.

The GRNs inference is accompanied by structure and parameter learning. The aim of structure learning is finding network structure that fits best the real regulatory interactions. For a given network structure, the parameter learning includes estimation of the unknown model parameters for each gene. This learning is performed by determining of the conditional dependencies between nodes in the network. However, in these networks, there is a super-exponential dependence of the number of all possible DAGs on the number of nodes n. Because the BNs inference is an NP-hard problem, BNs are the most suitable when applied to small networks consisted of tens to hundred genes [26]. N. Friedman et al. in [14] had introduced a framework for discovering of genes' interactions based on microarray data employing BNs.

BNs can deal with incomplete and noisy data, as well with stochastic nature of gene expression data. The main disadvantages of Bayesian network learning are the bigger number of genes regarding to the number of conditions and incapability to capture time nature of gene regulation and feedback loops that exist in the real GRNs.

These shortcomings make BNs unsuitable for GRNs inference from time series gene expression data, where dynamic features of gene regulation are included. Therefore, to model time features, BNs are extended by introducing dynamic Bayesian networks (DBNs). For probabilistic inference of DBNs, standard methods used in BNs inference can be used, too. Nevertheless, in the case of large networks, DBNs learning comes to be too complex. The DBNs are effective for GRNs inference when they are combined with other types of biological data. An example of integration of gene expression data with *a priori* biologic knowledge is presented in [16].

The concentration of RNAs, proteins and other metabolites is time changeable. Therefore, in order to describe gene regulatory mechanisms, differential equations can be used, too [11]. Ordinary differential equations (ODEs) systems use continual gene expression data and can easily cover positive and negative feedback loops [9]. Beside ODEs, difference equations model for GRNs inference can be used, too. Unlike the differential equations models that deal with continuous variables, in the difference equations model the variables are discrete. Discretization of the gene expression data leads to information loss [11]. These equations are more suitable, when they are applied on time series gene expression data.

Graphical Gaussian models (GGMs)

Association networks can be applied for steady-state and time series gene expression data. Association networks are represented by undirected graph. To determine which genes are co-expressed and between which genes should be an edge, it is necessary to apply similarity metrics such as Pearson coefficient or mutual information, and additionally to set a threshold.

Although the association networks do not determine the directions of the edges in the networks, they are suitable for inference of large GRNs because of their low computational complexity. Graphical Gaussian models (GGMs) use partial correlation coefficients to determine the conditional dependencies between genes and can determine directed and undirected edges in the network [18]. GGMs can distinguish direct or indirect interactions between genes, unlike the correlation networks where the edges present correlation between genes.

The main disadvantage of the described classical GGMs is that they can be applied when the number of experimental conditions n is greater than the number of genes p. If $p>n$ covariance matrix is not positive definite, its inverse matrix cannot be found. Therefore, an estimation of the covariance matrix is performed by shrinkage estimators to obtain positive definite covariance matrix [18].

Novel two-stage model for GRNs inference using *a priori* knowledge

The GRNs inference based on gene expression data is a very complex and difficult task. Beside inherent noise of these data, additionally the number of experi-

ments/conditions is less than the number of considered genes. Such shortcomings lead to lower precision and accuracy of inferred GRNs. To increase the accuracy and precision, usage of other types of biological data and *a priori* knowledge is needed [20] [26] [27] [28].

Based on comparison of the inference capabilities in [24] [25], Ristevski and Loskovska in [29] have proposed a novel two-stage model for GRNs inference. They have chosen the GGMs in the first phase of the proposed model, because they are a good base for uncovering the hub genes. The GRNs structure *G* can be represented by an adjacency matrix. The adjacency matrix entries *Gij* can be either 1 or 0, which refers to the presence or absence of a directed edge between *i*-th and *j*-th node of the network G, respectively. As a result of the first stage of the proposed model, a matrix of *a priori* knowledge *Gprior* is obtained, whose elements are computed by:

$$Gprior_{ij} = \begin{cases} \dfrac{1}{2} \cdot \dfrac{\left| pcor_{ij} \right| - pcor_{\min}}{pcor_{\max} - pcor_{\min}} + \dfrac{1}{2} \\ 0, if \left| pcor_{ij} \right| < pcor_{\min} \text{ or if edge is from } j \to i \end{cases} \tag{1}$$

where $pcor_{\max}$ and $pcor_{\min}$ are the minimal (set threshold) and maximal partial correlation coefficient, respectively [29]. Obtained matrix of *a priori* knowledge *Gprior*, presents a basis for the second stage of the proposed model.

To integrate the *a priori* knowledge obtained from first phase, the second phase defines a function *Gprior*' as a measure of matching between the given network G and the obtained *a priori* knowledge *Gprior* [28]. In this stage, structure Bayesian learning is carried out using Markov chain - Monte Carlo (MCMC) simulations.

3 Results and discussion

Validation of inferred GRNs represents an assessment of the inferred network quality, matched to the available knowledge in so-called "gold standard" networks. To validate inferred gene regulatory interactions using computational models, results obtained from wet-lab biological experiments are required. As validation criteria, receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) are used.

As GRNs inference models, Boolean networks, DBNs, GGMs and the novel two-stage model are used. For inference using Boolean networks, GGMs and DBNs, the following R packages were used: BoolNet [12], GeneNet [18] and G1DBN [22], respectively. For network reconstruction and MCMC simulations in the second stage of the novel model and the plotting of the ROC curves, BNSL MATLAB toolbox [21] was used and to obtain *a priori* knowledge, GeneNet was employed.

The ROC curves and their corresponding AUC values are shown in Fig. 1. The novel two-stage model shows the best inference capabilities (AUC=0.598), then follows inference using GGMs and DBNs, respectively, while AUC value for inference by Boolean networks was the lowest (AUC=0.487).

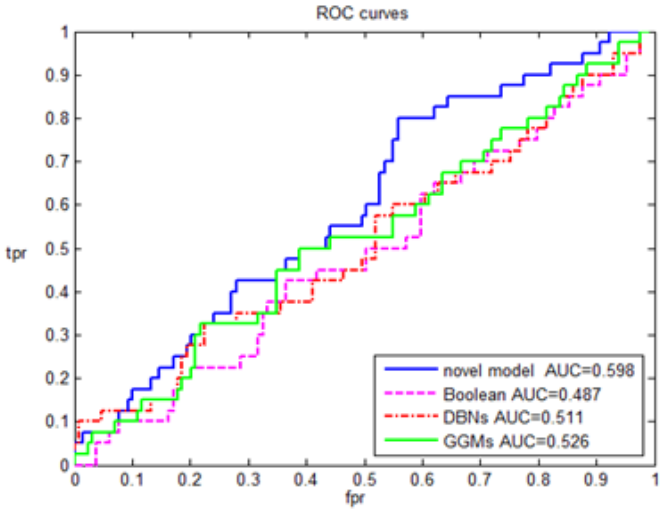

Fig. 1. ROC and AUC values comparison when different GRNs inference models are applied.

Reconstructed networks when novel two-stage model is applied are shown in Fig. 2 and Fig. 3. When the threshold is set on higher value 0.5 (Fig. 2), reconstructed network is sparser than when threshold is set on value 0.33 (Fig. 3). In the first case, the number of inferred interactions is 13, while in the second case, the inferred network contains 19 edges. The following gene-gene interactions LFY1→SEP4, EMF1→PI, UFO→AP2, FT→EMF1, PI→UFO and AP2→AG appear in the second case, while in the first case when the threshold is set on 0.5, they miss.

4 Conclusion and further work

In this paper, several GRN inference models are applied on A. thaliana gene expression time series data: Boolean networks, GGMs, DBNs and the two-stage model including *a priori* knowledge. As a result, GRNs are inferred and additionally, the inference capabilities of these models are compared and validated.

Although the two-stage inference model that integrates *a priori* knowledge has shown best inference capabilities compared to other models, the validation of the inferred networks has shown that there is a need of upgrading and updating of the "gold standard" networks with knowledge obtained from experimental -omics regulatory mechanisms. Beside gene expression data, the availability of various transcriptomics, proteomics, interactomics and metabolomics data, makes the network inference to become more challenging task.

By validation of the inferred networks, the main problem is the lack of “gold standard” networks to whose edges the presence/absence of inferred edges is confirmed. Inferred directed edges, which are not present in the available biological regulatory databases, are clues for further experimental research to confirm their presence

or absence in the real gene regulatory mechanisms. Furthermore, greater efforts should be made toward upgrading of existing regulatory databases with confirmed regulatory relationships between genes, microRNAs, TFs and the other components involved in the cell regulatory processes.

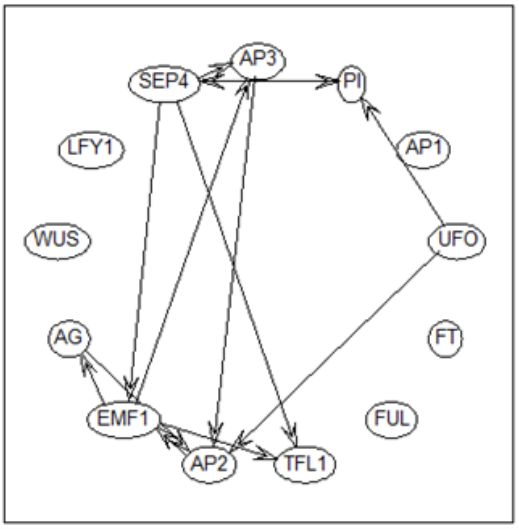

Fig. 2. Inferred network using the novel two stage model when the threshold is set on 0.5.

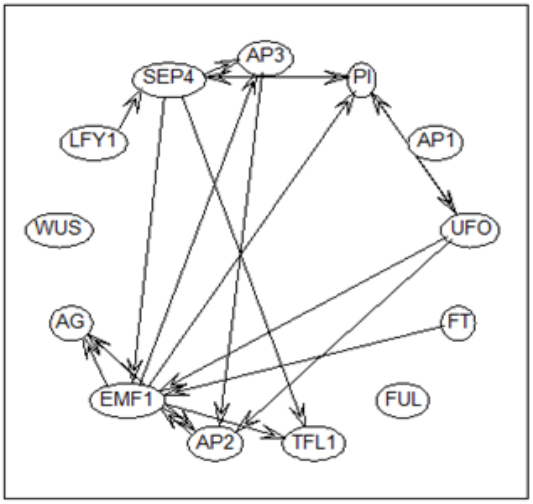

Fig. 3. Inferred network using the novel two stage model and the threshold is set on 0.333.

Different -omics data reveal different perspectives of regulatory networks. Integration of these data and using prior knowledge can discover a more reliable comprehension of the cell regulatory mechanisms. In order to significantly improve the accuracy of the inferred networks, there is still a need of development of models that can integrate the available biological *a priori* knowledge and other data such as ChIP-chip, ChIP-Seq and microRNA data.

References

1. N. A. Kolchanov, T. I. Merkulova, E. V. Ignatieva et al., Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes, Briefings in Bioinformatics, Vol. 8 No.4 pp. 266-274, 2007.
2. Z. Wang, MicroRNA Interference Technologies, New York: Springer-Verlag, 2009.
3. Z. Shuang and L. Mo-Fang, Mechanisms of microRNA-mediated gene regulation, Science in China Series C: Life Science, 2009, pp. 1111-1116.
4. C. Li, Y. Feng, G. Coukos and L. Zhang, Therapeutic MicroRNA Strategies in Human Cancer, AAPS Journal, Vol. 11, No. 4, 2009, pp.747-757.
5. Y. Nikolsky and J. Bryant (eds.), Protein Networks and Pathway Analysis, Humana Press, 2009, pp. 303-352.
6. J. G. Joung and Z. Fei, Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model, Bioinformatics, Data and text mining, Vol. 25 no. 3, 2009, pp 383-393.
7. P.Collas (ed.), Chromatin Immunoprecipitation Assays, Humana Press, 2009, pp. 133-143.
8. P. J. Park, ChIP-seq: Advantages and Challenges of a Maturing Technology, Nature Reviews Genetics, 2009, pp. 669-680.
9. M. Hecker, S. Lambeck, S. Toepfer. E. van Someren and R. Guthke, Gene regulatory network inference: Data integration in dynamic models – A review, BioSystems 96, 2009, pp. 86-103.
10. W. Zhao, E. Serpedin and E. R. Douherty, Recovering Genetic Regulatory Networks from Chromatin Immunoprecipitation and Steady-State Microarray Data, EURASIP Journal on Bioinformatics and Systems Biology, Vol. 2008:248747.
11. L. F. A. Wessels, E. P. Van Someren and M. J. T. Reinders, A comparison of genetic network models, Pacific Symposium on Biocomputing, 6, 2001, pp. 508-519.
12. C. Nuessel, M. Hopfensitz, D. Zhou and H. Kestler, Package 'BoolNet' – Generation, reconstruction, simulation and analysis of synchronous, asynchronous and probabilistic Boolean networks, CRAN, 2010-06-16.
13. N. Friedman and M. Goldszmidt, Learning Bayesian Networks with Local Structure, Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence, 1996, pp. 252-262.
14. N. Friedman, M. Linian, I. Nachman and D. Pe'er, Using Bayesian Networks to Analyze Expression Data, Journal of computational biology 7, pp. 601-620.
15. S. Liang, S. Fuhrman and R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, Pacific Symposium on Biocomputing 3, 1998, pp. 18-19.
16. Ristevski, B., Overview of Computational Approaches for Inference of MicroRNA-Mediated and Gene Regulatory Networks. Advances in Computers, 97, 2015, pp.111-145.

17. Y. Zhang, Z. Deng, H. Jiang and P. Jia, Inferring Gene Regulatory Networks from Multiple Data Sources Via a Dynamic Bayesian Networks with Structural EM, Springer – Verlag Berlin Haidelberg, pp. 204-214, 2007.
18. J. Shäfer, R. Opgen-Rhein and K. Strimmer, Package 'GeneNet' – Modeling and Inferring Gene Networks, CRAN, 2008-11-17.
19. J. Schäfer and K. Strimmer, Learning Large – Scale Graphical Gaussian Models from Genomic Data, CPP776, Science of Complex Networks: From Biology to the Internet and WWW; CNET 2004, pp. 263-276.
20. A. V. Werhli and D. Husmeier, Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Priori Knowledge, Statistical Applications in Genetics and Molecular Biology, Vol. 6, 2007, Article 15.
21. D. Eaton and K. Murphy, Bayesian Network Structure Learning (BNSL) – A Software Package for Matlab, 2007, http://www.cs.ubc.ca/~deaton/struct/bnsl.html.
22. S. Lebre and J. Chiquet, Package 'G1DBN' – A package performing Dynamic Bayesian Network inference, CRAN, 2008-01-23.
23. W.-Po Lee and K.-Cheng Yang, A clustering-based approach for inferring recurrent neural networks as gene regulatory networks, Neurocomputing 71, 2008, pp. 600-610.
24. B. Ristevski and S. Loshkovska**,** A Comparison of Models for Gene Regulatory Networks Inference, 2th International Conference ICT Innovations 2010, Ohrid, R. Macedonia, 2010, pp. 59-68.
25. Ideker, T., J. Dutkowski, and L. Hood, Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power. Cell, 2011. 144(6): pp. 860-863.
26. M. Grzegorczyk and D. Husmeier, Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move, Machine Learning, 71: 265-305, 2008.
27. A. V. Werhli and D. Husmeier, Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Priori Knowledge, Statistical Applications in Genetics and Molecular Biology, Vol. 6, 2007, Article 15.
28. F. Jaffrezic and G. T.-Klopp, Gene network reconstruction from microarray data, BMC Proceedings, 2009, 3(Suppl 4):S12.
29. B. Ristevski and S. Loskovska, A Novel Model for Inference of Gene Regulatory Networks, HealthMed Journal 2011, Vol.-5 No-6, pp. 2024-2033.
30. B. Ristevski, A survey of models for inference of gene regulatory networks, Nonlinear Analysis: Modelling and Control, 2013, Vol. 18, No. 4, pp. 444–465.
31. R.V.L. Joosen, D. Arends, L.A.J. Willems, W. Ligterink, R.C. Jansen, H.W.M. Hilhorst Visualizing the genetic landscape of Arabidopsis seed performance. *Plant Physiology*, 2012 vol. 158 no. 2, pp. 570-589.
32. Y. E. Sánchez-Corrales, E. R. Álvarez-Buylla, L. Mendoza, The Arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process Journal of Theoretical Biology, 2010, Vol. 264 no. 3, pp. 971-983.
33. http://david.abcc.ncifcrf.gov/
34. www.arabidopsis.org