Differential functional analysis and change motifs in gene networks to explore the role of anti-sense transcription

Marc Legeay[1,2], Béatrice Duval[1], and Jean-Pierre Renou[2]

[1] LERIA - Université d'Angers - UNAM, 2 bd Lavoisier 49045 Angers FRANCE
marc.legeay@univ-angers.fr,
[2] INSTITUT DE RECHERCHE EN HORTICULTURE ET SEMENCES (IRHS), UMR1345 INRA-Université d'Angers-AgroCampus Ouest, Centre Angers-Nantes, 42 rue Georges Morel - BP 60057, 49071 Beaucouzé FRANCE

Abstract. Several transcriptomic studies have shown the widespread existence of anti-sense transcription in cell. Anti-sense RNAs may be important actors in transcriptional control, especially in stress response processes. The aim of our work is to study gene networks, with the particularity to integrate in the process anti-sense transcripts. In this paper, we first present a method that highlights the importance of taking into account anti-sense data into functional enrichment analysis. Secondly, we propose the differential analysis of gene networks built with and without anti-sense actors in order to discover interesting change motifs that involve the anti-sense transcripts. For more reliability, our network comparison only studies the conservative causal part of a network, inferred by the C3NET method. Our work is realized on transcriptomic data from apple fruit.

1 Introduction

Understanding the regulation mechanisms in a cell is a key issue in bioinformatics. As large-scale expression datasets are now available, gene network inference (GNI) is a useful approach to study gene interactions [1], and a lot of methods have been proposed in the literature for this reverse engineering task [2,3,4]. Going a step further, the field of differential network analysis [5,6] proposes to decipher the cellular response to different situations. In medicine the comparison of interaction maps observed in cancerous tissues and healthy tissues may reveal network rewiring induced by the disease [7]. In these approaches, the comparative analysis is performed on networks that involve the same set of actors, namely the genes or proteins of the studied organism.

The aim of our work is to study gene networks, with the particularity to integrate in the process anti-sense transcripts. Anti-sense RNAs are endogenous RNA molecules whose partial or entire sequences exhibit complementarity to other transcripts. Their different functional roles are not completely known but several studies suggest that they play an important role in stress response

mechanisms [8]. A recent study with a full genome microarray for the apple has detected significant anti-sense transcription for 65% of expressed genes [9], which suggests that a large majority of protein coding genes are actually concerned by this process.

The work described in this paper proposes a large-scale analysis of apple transcriptomic data, with measures of anti-sense transcripts. To highlight the impact of anti-sense transcription, we propose to compare context-specific gene networks that involve different kinds of actors, on one hand the sense transcripts that are usually used in gene networks and on the other hand the sense and anti-sense transcripts. GNI methods generally find many false positive interactions, and some authors have proposed to study the core part of a gene network [10], by only computing for each gene the most significant interaction with another gene. We follow this line in order to discover which interactions of the core network are modified when we integrate in our GNI method the anti-sense transcripts. To characterize these modifications, we define the notions of change motifs for the comparison graph. Our preliminary results on the apple datasets show that relevant information is provided by this approach.

In section 2, we present the motivations of this work and the apple datasets that are used in our study. In section 3, we present a differential functional analysis that reveals the interest of taking into account anti-sense data. In section 4, we present our method to compare two core gene networks and to detect motifs that underline the role of anti-sense transcripts.

2 Motivations and biological material

Several studies have revealed the widespread existence of anti-sense RNAs in many organisms. Anti-sense transcripts can have different roles in the cell [8]. A significant effect is the post-transcriptional gene silencing: the self-regulatory circuit where the anti-sense transcript hybridizes with the sense transcript to form a double strand RNA (dsRNA) that is degraded in small interfering RNAs (siRNA). Previous studies on *Arabidopsis Thaliana* showed that sense and anti-sense transcripts for a defense gene (RPP5) form dsRNA and generate siRNA which presumably contributes to the sense transcript degradation in the absence of pathogen infection [11].

In [9], the authors have combined microarray analysis with a dedicated chip and high-throughput sequencing of small RNAs to study anti-sense transcription in eight different organs (seed, flower, fruit, ...) of apple (*Malus* $\times$ *domestica*). Their atlas of expression shows several interesting points. Firstly, the percentage of anti-sense expression is higher than that reported in other studies, since they identify anti-sense transcription for 65% of the sense transcripts expressed in at least one organ, while it is about 30% in previous *Arabidopsis Thaliana* studies. Secondly, the anti-sense transcript expression is correlated with the presence of short interfering RNAs. Thirdly, anti-sense expression levels vary depending on both organs and Gene Ontology (GO) categories. It is higher for genes belonging to the "defense" GO category and on fruits and seeds.

In order to study the impact of anti-sense transcripts, we use data of apple fruit during fruit ripening. The fruit ripening is a stress-related condition involving "defense" genes. We analyse RNA extracted from the fruit of apple thanks to the chip AryANE v1.0 containing 63011 predicted sense genes and 63011 complementary anti-sense sequences. This chip allows us to study the role of anti-sense transcripts at the genome-wide level by supplying transcriptional expression on both sense and anti-sense transcripts. We study the fruit ripening process described by two conditions: harvest (`H`) and 60 days after harvest (`60DAH`), and for each condition, 22 samples of apple fruit have been analysed. We first identify transcripts displaying significant differences between the two conditions (p-val<1%). With a further threshold of 1 log change between the two conditions, we found 931 sense (S) and 694 anti-sense transcripts (AS) differentially expressed, with among them, 200 transcripts ($S \cap AS$) for which both sense and anti-sense fulfil the condition. In the following, these 1625 transcripts will be called *transcripts of interest* for our study of apple ripening.

3 Differential functional analysis

A lot of tools are available to identify which GO categories are statistically over-represented in a set of genes. The Cytoscape plugin BiNGO [12] performs this task in a flexible and interactive way and moreover, the output of BiNGO is a graph where nodes represent GO categories and arcs represent the hierarchy between categories. In this visualisation, the size of a node is proportional to the number of genes in the test set annotated by this category, and the color of a node codes the over-representation: dark orange categories are most significantly over-represented, whereas white nodes are not significant but are included to show the hierarchy linking the dark categories.

In our experiment about apple ripening, the analysis of probes that are differentially expressed shows an important proportion of anti-sense actors: 694 AS probes for 931 S probes. Therefore it is relevant to question the role of these anti-sense actors in the ripening process. To look at this point, we propose a differential functional analysis where we compare the functional categories over-represented in the set of S probes and the functional categories over-represented in the set of probes $S \cup AS$. Anti-sense probes are not associated with a GO category. We decided to associate an anti-sense probe with the category of its corresponding sense probe[3]. This decision is based on the fact that due to its sequence complementarity, an anti-sense transcript may interact with the corresponding sense transcript, or at least with a very close member of the gene family.

The differential functional analysis is performed as follows. We apply BiNGO on the set S containing 931 genes, and we apply BiNGO on the set SAS ($S \cup AS$) containing 494 supplementary genes. These 494 genes are the genes associated with the AS probes of interest for which the corresponding sense probe is not

[3] Because there is no GO category for apple transcripts, we use *Arabidopsis Thaliana* orthologs in order to associate apple transcripts with GO categories.

differentially expressed. We thus obtain two sets of GO categories represented as two sub-graphs of the GO. Our proposal is then to compute the difference of these two sets; this provides a set of functional terms that are over-represented only when we include the AS probes in the functional analysis and we call these terms the *revealed-by-AS* terms.

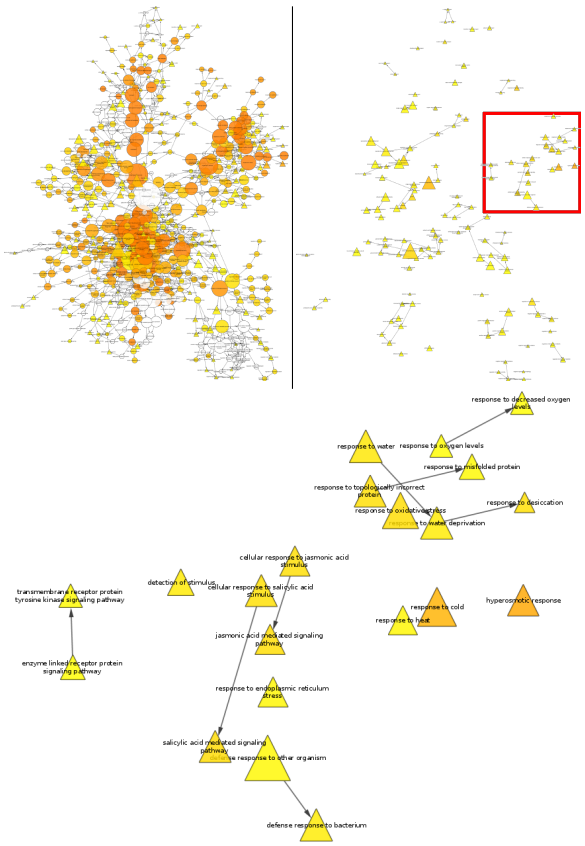

Fig. 1: BiNGO outputs for SAS probes (left), the revealed-by-AS terms (right) with a zoom (bottom) of the red square of the revealed-by-AS terms. Node size denotes the number of genes in the GO category. Node color denotes the over-representation of the GO category (yellow : low, orange : high). Arcs denote the hierarchy between nodes.

Figure 1 shows the SAS ontology and the difference with the S ontology that gives the revealed-by-AS terms. We can see on this representation that the nodes of the difference occur on many branches of the SAS ontology, meaning that the revealed-by-AS terms are not specific to a GO category.

GO category	p-value	# genes	Most specific
hyperosmotic response	4.4644e-05	30	yes
response to cold	5.4256e-05	56	yes
multicellular organismal process	1.1794e-04	225	no
response to high light intensity	5.2641e-04	25	yes
growth	1.6007e-03	56	no
cellular biosynthetic process	1.6007e-03	329	no
cell growth	1.8010e-03	51	no
regulation of response to stimulus	1.8376e-03	58	no
salicylic acid mediated signaling pathway	2.0524e-03	30	yes
jasmonic acid mediated signaling pathway	2.2286e-03	27	yes

Table 1: Top 10 of revealed-by-AS terms, sorted by p-values. For each term, we indicate the number of genes associated to transcripts of interest, and if the term is a most specific revealed-by-AS term.

This differential analysis gives us 125 revealed-by-AS terms, associated with their p-values. We present the top 10 terms in Table 1, where we report how many genes are associated with the terms. As pointed out by the authors of BiNGO, due to the interdependency between GO categories in the hierarchy, the most relevant terms of the output are the terms located farthest down the hierarchy, that correspond to more specific functions. Therefore the interpretation will focus on the most specific terms according to the ontology hierarchy, as indicated in Table 1. These revealed-by-AS terms highlight biological functions that are over-represented in our probes of interest only when we include AS informations. Our experiment concerns the complex process of apple ripening. In this experiment, between harvest (`H`) and 60 days after harvest (`60DAH`), fruits are stored in cold rooms and have to react to cold stress. We notice that the `response to cold` term is a revealed-by-AS term. Therefore, if we do not consider the anti-sense data, we loose important information for a functional analysis. Moreover, the `response to cold` term is represented by 56 sense or anti-sense actors in our probes of interest; if we examine the differential expression of these transcripts, we notice that 24 of them are anti-sense probes with a diminution of their expression between `H` and `60DAH`, while the corresponding sense probes have no differential expression. The differential functional analysis that we propose is thus a way to focus on interesting anti-sense transcripts that deserve a further biological study.

4 Network comparison

4.1 Inference of the core part of a gene network

Many models have been proposed to infer gene networks from transcriptomic data. Reviews of the reverse engineering methods can be found in [2,3,13]. A

family of inference methods reconstruct pairwise gene interaction networks by measuring with a statistical criterion whether two genes are co-expressed or co-regulated. This statistical measure can be the Spearman or Pearson correlation [14], or the mutual information [10,15]. Mutual information enables the detection of non-linear relationships. These methods need a step of thresholding to decide which values of the statistical measure are significant. One major drawback of these methods is that many of the predicted interactions are false positives. We can differentiate two types of false positive interactions: an interaction that does not biologically exist, and an indirect interaction. If two genes g_2 and g_3 are regulated by g_1, then mutal information (as well as correlation) between g_2 and g_3 is high and an indirect interaction is put in the inferred network. Indirect interactions lead in large gene networks difficult to interpret by biologists and they must be pruned from the output networks [15,16]. To avoid this pitfall, the method C3NET proposes to compute the conservative causal core of a gene network, by selecting for each gene a unique interaction. Figure 2 decomposes the C3NET algorithm, that we use in this work. From the mutual information matrix, for each line corresponding to a gene g, the algorithm identifies the maximal mutual information which defines the best neighbor that will be connected to g in the network. Experiments have shown the good ability of this conservative method to capture the causal structure of a regulatory network.

MI	g_1	g_2	g_3	g_4
g_1	1.0	0.7	0.8	0.7
g_2	0.7	1.0	0.6	0.5
g_3	0.8	0.6	1.0	0.1
g_4	0.7	0.5	0.1	1.0

1 ⇒

MI	g_1	g_2	g_3	g_4
g_1	0	0.7	0.8	0.7
g_2	0.7	0	0.6	0.5
g_3	0.8	0.6	0	0
g_4	0.7	0.5	0	0

2 ⇒

MI	g_1	g_2	g_3	g_4
g_1	0	0.7	**0.8**	0.7
g_2	**0.7**	0	0.6	0.5
g_3	**0.8**	0.6	0	0
g_4	**0.7**	0.5	0	0

3 ⇒

Adj	g_1	g_2	g_3	g_4
g_1	0	0	**1**	0
g_2	**1**	0	0	0
g_3	**1**	0	0	0
g_4	**1**	0	0	0

4 ⇒

Adj	g_1	g_2	g_3	g_4
g_1	0	**1**	**1**	**1**
g_2	**1**	0	0	0
g_3	**1**	0	0	0
g_4	**1**	0	0	0

Fig. 2: C3NET procedure from the mutual information matrix to the network adjacency matrix. The mutual information matrix is computed from transcriptomic data. Step 1: non-significant and diagonal values are suppressed. Step 2: the maximal mutual information is identified for each row. Step 3: the matrix is transformed into a boolean matrix. Step 4: the resulting adjacency matrix is made symmetric because mutual information does not provide directional information. This is the adjacency of the computed network.

4.2 Comparison of Core Networks : Change motifs

To study the role of anti-sense transcripts in gene regulation networks, we propose to compare two networks obtained by C3NET, the N_S network using only sense actors, and the N_{SAS} network using sense and anti-sense actors. Our goal is to identify which direct interactions are modified in the core network when we consider anti-sense transcripts. To achieve this, we need the first three steps of C3NET (we do not need the symmetric matrix). Figure 3 illustrates modifications of interactions in the matrix. In N_S for s_3 the direct interaction occurs with s_1 but in N_{SAS}, the direct interaction occurs with as_2. This is this type of modification that we want to identify.

MI	s_1	s_2	s_3	s_4
s_1	0	0.7	**0.8**	0.7
s_2	**0.7**	0	0.6	0.5
s_3	**0.8**	0.6	0	0
s_4	**0.7**	0.5	0	0

MI	s_1	s_2	s_3	s_4	as_1	as_2	as_3	as_4
s_1	0	0.7	**0.8**	0.7	0.6	0.5	0.7	0.6
s_2	**0.7**	0	0.6	0.5	0.4	0.6	0	0.5
s_3	**0.8**	0.6	0	0	0.8	**0.9**	0.3	0
s_4	**0.7**	0.5	0	0	0.6	0	0.5	0.6
as_1	0.6	0.4	**0.8**	0.6	0	0.6	0.4	0.7
as_2	0.5	0.6	**0.9**	0	0.6	0	0.6	0.4
as_3	**0.7**	0	0.3	0.5	0.4	0.6	0	0.5
as_4	0.6	0.5	0	0.6	**0.7**	0.4	0.5	0

Fig. 3: Mutual information matrices from S (left) and SAS (right). With anti-sense data, the maximal mutual information changes (green values). s_3 is connected with s_1 in N_S (red value) and with as_2 in N_{SAS}.

When we integrate anti-sense actors in the core network computation, we focus on sense nodes which become connected with an anti-sense node. It means that an arc from N_S between two sense nodes is now an arc from a sense to an anti-sense in N_{SAS}.

In order to highlight those modifications, we construct a comparison graph G by adding N_S arcs to N_{SAS}. We visualize this graph with Cytoscape [17], where we color arcs of G depending on the network they belong to: an arc is green if it only exists in N_{SAS}, red if it only exists in N_S and grey if it exists in both networks. With this color code, an interaction from N_S replaced in N_{SAS} is represented by a sense node with a green and a red arc (Figure 4a).

Around this elementary motif, named M_0, we observe richer configurations represented in Figure 4b. M_1 motif denotes a strong link between a sense and an anti-sense. M_2 motif reveals that the interaction between $S1$ and $S2$ observed in N_S is in fact an indirect one that involves the anti-sense $AS3$.

We have constructed comparison graphs from `H` and `60DAH` experiments. The `60DAH` comparison graph can be visualized in Figure 4c. This visualization allows biologists to explore what gene links are impacted by the anti-sense transcripts. In a more quantitative way, we give in Table 2 the count of the motifs existing in each of the graphs. The most important information is the number of M_0

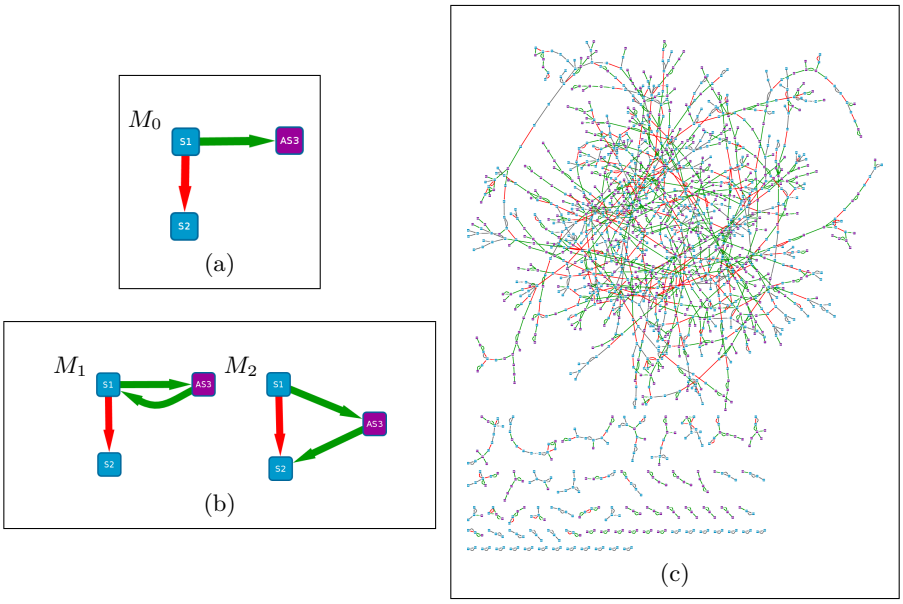

Fig. 4: Motifs and comparison graph of a Sense network with a Sense and Anti-sense network. Blue nodes denote sense nodes and purple nodes denote anti-sense nodes. **(4a)** Elementary motif. **(4b)** Richer motifs observed. **(4c)** Comparison graph between N_S and N_{SAS} networks from `60DAH` experiment.

motifs that indicates how many sense actors are connected to an anti-sense. We notice that there are about 380 M_0 motifs, which means that 40% of the 931 S transcripts are involved in a M_0 motif, for 50% of the 694 AS transcripts. Among these M_0 motifs, about 30% are M_1 motifs, where sense and anti-sense are strongly connected. Richer motifs M_2 are less represented. As the core network tries to capture the most important gene interactions, the fact that 40% of S network is impacted by anti-sense actors shows that anti-sense transcripts play a role in fruit ripening.

4.3 Change motifs and functional analysis

In Section 2 we have proposed a differential functional analysis and defined the revealed-by-AS terms. We now combine the information provided by the functional analysis with the information provided by motifs. To illustrate this, we select in Table 1 the most specific GO categories from the top 10 revealed-by-AS terms and study what kind of motifs are related to these terms. Table 2 counts, for each term, the number of motifs which contain at least one transcript associated to the term, and the number of transcripts associated to the term present in the motifs. We notice that in both experiments, for each term, around 40% of the transcripts are involved in a M_0 motif. This observation encourages

Experiment		H			60DAH		
Motif		M_0	M_1	M_2	M_0	M_1	M_2
Global	# motifs	371	107	18	384	116	19
`hyperosmotic response`	# motifs	18	3	1	21	8	0
(37 transcripts)	# transcripts	17	3	1	18	8	0
`response to cold`	# motifs	25	4	2	32	12	1
(63 transcripts)	# transcripts	24	4	2	29	13	1
`response to high light`	# motifs	14	5	1	13	7	0
`intensity` (31 transcripts)	# transcripts	12	5	2	12	7	0
`salicylic acid mediated`	# motifs	17	3	0	14	3	1
`signaling pathway` (36 transcripts)	# transcripts	15	3	0	13	3	1
`jasmonic acid mediated`	# motifs	14	2	0	13	3	0
`signaling pathway` (31 transcripts)	# transcripts	13	2	0	12	3	0

Table 2: Number of motifs, number of motifs containing at least one revealed-by-AS transcript and number of these transcripts being in motifs. The number of revealed-by-AS transcripts associated with the term is noted in parentheses.

us to study the gene regulatory networks related to the revealed-by-AS terms, which will be the next step of this work.

5 Conclusion

The aim of our work is to study gene networks, with the particularity to integrate in the process anti-sense transcripts. Firstly we propose a method that highlights biological functions impacted by anti-sense transcription. Biological functions are identified by computing the difference between two ontologies. Secondly we propose a differential gene network analysis allowing to identify which direct interactions are modified. We combine these two methods to submit limited sets of anti-sense transcripts to the biological interpretation.

The field of differential network analysis is certainly a promising approach to study context-specific regulation networks. For example, the method proposed in [18] compares two networks computed by C3NET corresponding to different cell conditions (disease versus normal). The aim is to identify the disease network, that is the interactions that only appear in disease-related cells. This method can not be applied in our case. In fact, we rely on the same algorithm to compute a network of direct interactions. But in our case, we compare two networks that involve different actors, the sense transcripts in one hand and the sense and anti-sense transcripts in the other hand. These two networks concern the same experimental condition and the change motifs that we compute aim to highlight potential anti-sense actions.

References

1. Brazhnik P., de la Fuente A., Mendes P.: Gene networks: how to put the function in genomics. Trends in Biotechnology, 20(11):467 – 472 (2002)
2. Bansal M., Belcastro V., Ambesi-Impiombato A., di Bernardo D.: How to infer gene networks from expression profiles. Molecular Systems Biology, 3(1) (2007)
3. Marbach D., Costello J.C., Küffner R., Vega N.M., Prill R.J., Camacho D.M., Allison K.R., The DREAM5 Consortium, Kellis M., Collins J.J., Stolovitzky G.: Wisdom of crowds for robust gene network inference. Nature Methods, 9(8):796–804 (2012)
4. Emmert-Streib F., Glazko G., Altay G., de Matos Simoes R.: Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Bioinformatics and Computational Biology, 3:8 (2012)
5. Sharan R., Ideker T.: Modeling cellular machinery through biological network comparison. Nature biotechnology, 24(4):427–433 (2006)
6. Ideker T., Krogan N. J: Differential network biology. Molecular systems biology, 8(1):565 (2012)
7. Barabási A.-L., Gulbahce N., Loscalzo J.: Network Medicine: A Network-based Approach to Human Disease. Nature reviews. Genetics, 12(1):56–68 (2011)
8. Pelechano V., Steinmetz L.M.: Gene regulation by antisense transcription. Nature Reviews Genetics, 14(12):880–893 (2013)
9. Celton J.-M., Gaillard S., Bruneau M., Pelletier S., Aubourg S., Martin-Magniette M.-L., Navarro L., Laurens F., Renou J.-P.: Widespread anti-sense transcription in apple is correlated with siRNA production and indicates a large potential for transcriptional and/or post-transcriptional control. New Phytologist, 287–299 (2014)
10. Altay G., Emmert-Streib F.: Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology, 4(1):132 (2010)
11. Yi H., Richards E.J.: A Cluster of Disease Resistance Genes in Arabidopsis Is Coordinately Regulated by Transcriptional Activation and RNA Silencing Plant Cell, 19:2929–2939 (2007)
12. Maere S., Heymans K., Kuiper M.: BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics, 21(16):3448–3449 (2005)
13. Friedel S., Usadel B., von Wiren N., Sreenivasulu N.: Reverse Engineering: A Key Component of Systems Biology to Unravel Global Abiotic Stress Cross-Talk. Frontiers in Plant Science, 3 (2012)
14. Langfelder P., Horvath S.: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1):559 (2008)
15. Margolin A.A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera R.D., Califano A.: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics, 7(Suppl 1):S7 (2006)
16. Zhang X., Liu K., Liu Z.-P., Duval B., Richer J.-M., Zao X.-M., Hao J.-K., Chen L.: NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference Bioinformatics, 29(1):106–113 (2012)
17. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 13(11):2498–2504 (2003)
18. Altay G., Asim M., Markowetz F., Neal D. E: Differential C3NET reveals disease networks of direct physical interactions. BMC Bioinformatics, 12(1):296 (2011)