Towards the reconstruction of Brassica napus seed development FA metabolism dynamic regulatory map

Tchagang, Alain; Bérubé, Hugo; Liu, Ziying; Phan, Sieu; Famili, Fazel; Pan, Youlian

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/GENSIPS.2009.5174338
IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2009), pp. 1-4, 2010-05-21

TOWARDS THE RECONSTRUCTION OF *BRASSICA NAPUS* SEED DEVELOPMENT FA METABOLISM DYNAMIC REGULATORY MAP

Alain Tchagang, Hugo Bérubé, Ziying Liu, Sieu Phan, Fazel Famili, and Youlian Pan

Knowledge Discovery Group, Institute for Information Technology, NRC 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
{alain.tchagang, hugo.berube, ziying.liu, sieu.phan, fazel.famili, youlian.pan}@nrc.gc.ca (3998)

ABSTRACT

The increasing demand for canola (*Brassica napus*) for both food (e.g. vegetable oil) and non-food (e.g. biofuel) applications presents significant socio-economic benefits. While genetic engineering offers great potential to speed up the process of canola improvement, such an effort relies on a good understanding of the molecular mechanisms underlying seed development, fatty acid (FA) metabolism, and oil content. Applying a well-defined algorithm to a time-series gene expression dataset of *B. napus* during seed development, and a well selected dataset of interactions between transcription factors and their target genes, we derive a dynamic regulatory map that is able to recover many of the known aspects of these responses. Predictions made in this study are further validated through literature search, leading to potential new roles for LEC1, LEC2, WRI1, FUS3, MYB30, and ABI3 in controlling *B. napus* seed development and FA metabolism related genes, thus potential targets for genetic improvement of oil production.

1. INTRODUCTION

Seed of canola (*Brassica napus*) contains fatty acids (FA) such as oleic acid and linoleic acid, which are not only used for vegetable oil as food supply to the growing population, but also very demanded by expanding markets for values beyond nutrition. For example, vegetable oil is the important lipid source for the production of lubricants, inks, paints, and biofuel. Because of its composition (high in monounsaturated fats and low in saturated fats), canola oil has a positive health effect via reduction of cardiovascular diseases. Canola's unique characteristics make it an ideal feedstock for biodiesel production, thus contributing to a positive environmental impact. The increasing demand for canola in both food and non-food applications presents significant socio-economic benefits. In order to capture these potential benefits, it is essential to increase canola productivity. While genetic engineering offers great potential to speed up the process of crop improvement, such an effort relies on a good understanding of the molecular mechanisms underlying *B. napus* seed development, FA metabolism, and oil content. In higher plants, the biosynthesis of most FA is physiologically coupled with seed development [1].

In an attempt to understand the molecular mechanisms underlying seed development, FA metabolism, and oil content in crops, several studies have been performed using the *Arabidopsis thaliana* plant [1-8]. Despite all these studies and technological advances, the regulatory mechanism of FA metabolism in *A. thaliana* has not yet been well explored. Although the overall biochemical pathways producing storage lipids have been extensively described [1-11], factors regulating FA synthesis and controlling total oil content in oilseed crops are still poorly understood. Likewise, questions remain regarding the interplay between this metabolism and the developmental progression of embryogenesis. Furthermore, when regard to *B. napus*, the genome sequencing is not yet complete even though a large collection of its expressed sequence tags (ESTs) have been accumulated [9-11]. This makes computational analysis much challenging. In other words, the unavailability of the *B. napus* genome sequences and its detailed genomic information counteract several aspects of its computational analysis.

In this study, we derive a preliminary dynamic regulatory map of the *B. napus* FA metabolism during seed development. The dynamic regulatory map derived in this study is able to recover many of the known aspects of these responses. This is done by applying the Dynamic Regulatory Events Miner (DREM) algorithm [12] to a well-defined *B. napus* FA metabolism time-series gene expression data (obtained from a collection of well-defined uni-ESTs and full-length cDNAs) during seed development [12], and a well selected dataset on TF-gene interactions obtained through a carefully selected literature search bias towards experimental evidence [1-8]. Predictions made by our analysis are further validated through literature search, leading to potential new roles for the transcription factors, LEAFY COTYLEDON1 (LEC1), LEAFY COTYLEDON2 (LEC2), WRINKLED1 (WRI1), FUSCA3 (FUS3), MYB DOMAIN PROTEIN 30 (MYB30), and ABA INSENSITIVE 3 (ABI3) in controlling *B. napus* seed development FA and lipid metabolism, thus potential genetic targets for oil production improvement. The rest of this paper is organized as follows: in Section 2, we present the materials and methods used in this study. In

Section 3, we present the results. Finally, we conclude in Section 4.

2. MATERIALS AND METHODS

We describe below the materials and methods used for the reconstruction of *B. napus* FA metabolism dynamic regulatory map during seed development.

2.1. *Brassica napus* gene expression data

We used a publicly available *B. napus* gene expression data (obtained from a collection of well-defined uni-ESTs and full-length cDNAs), and downloaded from an NIH site (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/). This dataset was made public on February 26th 2008 [11] with accession numbers: GSM173081, GSM173082, GSM173084, GSM173087, GSM173090, GSM173092, GSM173100, GSM173101, GSM173102, GSM173103, GSM173104, GSM173105, GSM173106, GSM173107, GSM173108, and GSM173109. It corresponds to the gene expression profiles during seed development and FA metabolism, as well as the relevant regulation of *B. napus.* It has 8 times points: 7, 9, 12, 17, 19, 21, 25, and 31 days after pollination (DAP). Each time point has 2 biological replicates. We refer the reader to [11] for more details regarding the experimental procedures and other technical details associated with this dataset.

2.1.1. Arabidopsis thaliana orthologs

Brassica and *Arabidopsis* are in the same family and closely related. *A. thaliana* is a model plant and its genome is well studied. Since the *B. napus* genome is not yet complete, we use the orthologs in *A. thaliana* for computational analysis. The search for orthologs in *A. thaliana* is done by blasting the sequences of interest *B. napus* genes against the TAIR (http://www.arabidopsis.org/) database (version 8) and keeping only the matches with scores better than 1e-20. The same analysis is also done with TIGR Gene Index databases for *Arabidopsis* (AGI) and *B. napus* (BNGI) annotations of ESTs which gives us additional type of annotation like gene ontology (GO).

2.1.2. Gene expression matrix

The original *B. napus* time-series gene expression dataset downloaded from the GEO database had 8343 probes. After discarding probes with missing values, we obtained 5588 significant probes, among which 3362 appear to have orthologs in *A. thaliana* with *p-value* less than 1e-20.

2.2. Interactions matrix between transcription factor and gene

The TF-gene interactions matrix used in this study was inferred through a carefully selected literature search, bias towards experimental evidence [1-8]. More precisely, we compiled from the literature a set of *A. thaliana* TFs that have been shown by microarray experiments and/or quantitative RT-PCR to promote, suppress, or induce genes related to seed development, FA metabolism and other related biological processes such as lipid metabolism and biosynthetic process, with a *p-value* less than 0.01. The end product corresponds to a matrix where the rows correspond to the genes and the columns to TFs. The entries of the matrix are either 0 or 1, with 1 for a known interaction between a TF and its target gene and 0 for an unknown interaction.

2.3. Computational tool for dynamic regulatory map

We used the Dynamic Regulatory Event Miner (DREM) (http://www.sb.cs.cmu.edu/drem), which has been successfully applied to analyze the *Saccharomyces cerevisiae* [12] and *Escherichia Coli* [13] time-series gene expression data under several biological perturbations. DREM is an algorithm for modeling and analyzing the dynamics of transcriptional gene regulation. It takes time-series gene expression data and TF-gene interaction data as input. The TF-gene interaction data could come from Chromatin Immunoprecipitation combined with microarray experiments (ChIP-chip), TF binding site motif information, or through literature search as in this study. After execution, DREM outputs an annotated dynamic regulatory map that highlights bifurcation events in the time-series that is places in the time-series where sets of genes which previously had roughly similar expression level diverge. Often these bifurcation events can be explained by TFs that selectively regulate a certain subset of genes. DREM annotates these events with TFs potentially responsible for them, thus provides a global map of the gene regulation of the time-series.

The DREM algorithm is an extended version of a procedure for learning input-output hidden Markov models (IOHMMs) [14]. IOHMMs are an extension of hidden Markov models (HMMs) in that they allow for an additional input set (TF-gene interactions in this case) that does not necessarily change over time to control the transition probabilities of genes from state to state [12]. In HMMs and IOHMMs, hidden states are used to group genes by associating a cluster with each path through the hidden states over time. In DREM, each hidden state is associated with one time point and represents a Gaussian distribution of the expression values for genes associated with it [12]. Since our focus in this paper is to reconstruct the *B. napus* FA metabolism dynamic regulatory map during seed development, we refer the reader to [12] for more information regarding the computational aspect behind DREM.

3. RESULTS AND DISCUSSION

Using the above described materials and methodologies, we obtained the following results.

3.1. TF-gene interactions matrix

We identified a set of 12 *A. thaliana* transcription factors ($p < 0.01$) that have been shown by microarray experiments and/or quantitative RT-PCR to play major roles in seed development, FA, and FA-derived complex lipid metabolism and biosynthesis (**Table** 1).

Table 1. *Arabidopsis thaliana*'s FA metabolism TFs and the number of their target genes obtained through literature search.

TFs	# of target genes in *A. thaliana*	# of target genes in *B. napus**
LEC1	425	57
LEC2	750	31
WRI1	26	6
FUS3	30	6
ABI3	14	5
ABI5	12	3
MYB30	18	5
FAE1	5	0
SERK	2	0
PKL	3	0
AtbZIP10	9	3
AtbZIP25	9	3

* The number of target genes of each TFs in *B. napus* is derived based on the identified orthologs between the two species.

3.2. Fatty acid metabolism dynamic regulatory map

Using the preprocessed time-series gene expression data and the static TF-gene interactions matrix as input to the DREM algorithm, we obtained the following *B. napus* seed development FA metabolism dynamic regulatory map (**Figure 1).** We used the default parameters in DREM.

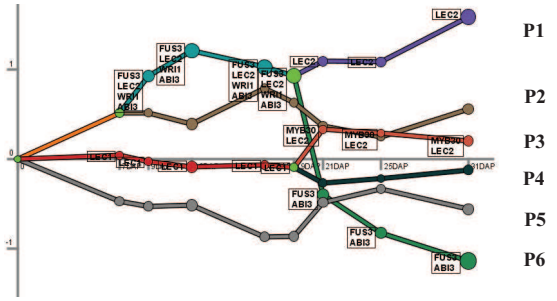

Figure 1. Predicted FA dynamic regulatory map in *Brassica napus* during seed development. The y-axis corresponds to the mean of the expression values of genes that belong to a path, and the x-axis the time in DAP

Our analysis reveals 6 significant paths (P1, P2, P3, P4, P5, and P6). Gene ontology (GO) biological process analyses of each path reveal that high expressions of biosynthesis processes (GO:0009058), lipid biosynthesis (GO:0008610) and metabolism (GO:0006629), FA biosynthesis (GO:0006633) and metabolism (GO:0006631) related genes and transition of FA components are mainly at stages 21-31_DAP, with *p-value* between 1e-05 and 1e-01. Important effects of seed development (GO:0048316), regulation of developmental process (GO:0050793), tissue development (GO:0009888), jasmonic acid stimulus (GO:0009753), jasmonic acid biosynthetic process (GO:0009695), starch metabolism, amino acid biosynthetic process (GO:0008652), carbon flux, oxidative pentose phosphate pathway (OPPP), photosynthesis (GO:0015979), carbohydrate biosynthetic process (GO:0016051) and several other biological processes related to FA and FA-derived complex lipid metabolism and biosynthesis are also observed (**Table 2)**.

Table 2. GO biological process analysis of each path.

Path	Relevant GO biological process*	Fraction of genes**
P1	- Carbohydrate metabolic process	9/93
	- Protein transport	5/67
	- Response to biotic stimulus	8/78
	- Biosynthetic process	19/371
P2	- Photosynthesis	9/29
	- Seed development	11/60
	- Response to abscisic acid stimulus	6/34
	- Lipid metabolic process	12/80
	- Lipid biosynthetic process	7/52
P3	- Fatty acid metabolic process	13/41
	- Cellular lipid metabolic process	21/71
	- Fatty acid biosynthetic process	8/25
	- Lipid biosynthetic process	15/52
	- Lipid metabolic process	22/80
	- Response to jasmonic acid stimulus	5/16
	- Response to abscisic acid stimulus	7/34
	- Response to abiotic stimulus	40/183
	- Seed development	12/60
P4	- Biosynthetic process	111/371
	- Lipid biosynthetic process	18/52
	- Response to abiotic stimulus	54/183
	- Lipid metabolic process	25/80
	- Cellular lipid metabolic process	22/71
	- Response to jasmonic acid stimulus	6/16
	- Amino acid biosynthetic process	10/30
	- Seed development	18/60
	- Response to abscisic acid stimulus	10/34
	- Fatty acid biosynthetic process	7/25
	- Fatty acid metabolic process	11/41
	- Photosynthesis	7/29
P5	- Amino acid biosynthetic process	10/30
	- Photosynthesis	9/29
	- Fatty acid metabolic process	10/41
	- Fatty acid biosynthetic process	6/25
	- Response to abscisic acid stimulus	8/34
	- Seed development	13/60
	- Lipid metabolic process	16/80
	- Cellular lipid metabolic process	14/71
	- Lipid biosynthetic process	9/52
P6	- Seed development	5/60
	- Protein complex	11/191

* Only GO biological processes related to seed development FA and FA-derived complex lipid metabolism and biosynthesis are shown: *p-value* between 1e-05 and1e-01.

** Fraction is relative to the available number of genes annotated under the GO term.

The dynamic regulatory map reveals that the TFs FUS3, LEC1, LEC2, WRI1, and ABI3, are active early during seed development, several of which, such as LEC1, LEC2, ABI3, and FUS3 continue to be active during seed maturation. This observation is consistent with the fact that the LEC1 function is partially dependent on ABI3, FUS3, and WRI1 in the regulation of FA and FA-derived complex lipid [1]. It also relates with the fact that LEC2 has been shown by previous researchers

to act during early stage of seed development to promote FA metabolism [2], and later on during seed maturation to control transcriptional program [2, 8]. Furthermore, LEC2, FUS3, and ABI3 have been implicated to be major regulators of the maturation phase [2, 8], to play positive roles on storage protein gene transcripts [2], and have also been suggested to play a major role during embryogenesis [2, 8]. These observations are consistent with the GO biological process analysis of the dynamic regulatory map: paths P1-P5, which show that most genes controlled by LEC2 are predominantly expressed after 21 DAP, which corresponds to the maturation phase. WRI3 has also been shown to play a significant role during oil accumulation in maturing seed, to be a prerequisite for FA synthesis, and is important for normal embryo development [3]. Consistent with this literature evidence, our analysis shows strong expression of genes regulated by WRI1 at the onset of the maturation phase.

The over-expression of genes related with FA and FA-derived complex lipid metabolism, biosynthesis and the transition of most FA components which are mainly observed during the period of 19-31 DAP. The transition at 19 DAP is somehow correlated with the timing of activation of LEC1, LEC2 and MYB30. This observation is consistent with several studies [1, 2, 5]; the over-expression of LEC1 gene causes globally increased expression of FA biosynthetic genes, which are involved in key reactions of condensation, chain elongation, and desaturation of FA biosynthesis [1]. An interesting observation is between P1 and P6, P3 and P4 (Figure 1), respectively. These paths behave coherently up until 19 DAP, time point at which they diverge significantly, thus 19 DAP may correspond to a transition stage of seedling to sink tissue. Another interesting observation is described in the dynamic regulatory map by P6, which shows a significant decrease in the expression level of genes controlled by FUS3 and ABI3 TFs after 19 DAP, suggesting novel role of FUS3 and ABI3 in *B. napus*. Further experimental analysis is needed to confirm this.

4. CONCLUSION

In this study, we apply a well defined dynamic regulatory map algorithm on a time-series gene expression dataset on *Brassica napus* FA metabolism during seed development together with TF-gene interaction data compiled through a carefully selected literature search bias towards experimental evidence. Results obtained suggest that the transcription factors: ABI3, MYB30, LEC1, LEC2, FUS3, and WRI3 tested in this study act as key regulators to coordinate the expression of fatty acid biosynthesis and metabolism related genes in *Brassica napus*. Thus they may represent potential targets for genetic improvement of oil production in *Brassica napus*. We are currently validating some of the results obtained using experimental procedures such as quantitative RT-PCR. Also, we intend to improve the DREM algorithm by introducing convergent points within the dynamic maps. That is the time points where sets of genes with different expression profiles merge and behave coherently from there on.

5. ACKNOWLEDGMENTS

This work is partially supported by Genomics and Health Initiative of the National Research Council Canada. This is publication NRC 50780 of National Research Council Canada.

6. REFERENCES

[1] J. Mu, H. Tan, Q. Zheng, et al., "LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis". *Plant Physiology*, vol. 148, pp. 1042-1054, 2008.

[2] S.A. Braybrook, S.L. Stone, S. Park, et al., "Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis". *PNAS*, no. 9, vol. 103, pp. 3468-3473, 2006.

[3] S. Baud, M.S. Mendoza, A. To, et al., "WRINKLED1 specifies the regulation action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis". *The Plant Journal*, vol. 50, pp. 825-838, 2007.

[4] H. Wang, J. Guo, K.N. Lambert, Y. Lin, "Developmental control of Arabidopsis seed oil biosynthesis". *Planta*, vol. 226, pp. 773-783, 2007.

[5] S. Raffaele, F. Vailleau, A. Léger, et al., "A MYB transcription factor regulates very-long-chain fatty acid biosynthesis for activation of hypersensitive cell death response in Arabidopsis". *The Plant Cell*, vol. 20, pp. 752-767, 2008.

[6] D.W. James Jr., E. Lim, J. Keller, et al., "Directed tagging of the Arabidopsis FATTY ACID ELONGATION1 (FAE1) gene with the maize transposon activator". *The Plant Cell*, vol. 7, pp. 309-319, 1995.

[7] J.B. Ohlrogge, "Regulation of fatty acid synthesis". *Annu. Rev. Plant Physiol. Plant Mol. Biol.* vol. 48, pp. 109–136, 1997.

[8] J.L. Harada, "Role of Arabidopsis LEAFY COTYLEDON genes in seed development". *J. Plant Physiol*., vol. 158, pp. 405-409, 2001.

[9] J. Dong, W.A. Keller, W. Yan, F. Georges, "Gene expression at early stages of Brassica *napus* seed development as revealed by transcript profiling of seed-abundant cDNAs". *Planta*, vol. 218, pp. 483-491, 2004.

[10] M. Hajduch, J.E. Casteel, K.E. Hurrelmeyer, et al., "Proteomic analysis of seed filling in Brassica *napus*. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis". *Plant Physiology,* vol. 141, pp. 32-46, 2006.

[11] G.Z. Wu, Q.M. Shi, Y. Niu, M.Q. Xing et al. "Shanghai RAPESEED Database: a resource for functional genomics studies of seed development and fatty acid metabolism of Brassica". *Nucleic Acids Res.,* vol. 36, pp. D1044–D1047, 2008.

[12] J. Ernst, O. Vainas, C.T. Harbison, et al., "Reconstructing dynamic regulatory maps". *Nature-EMBO Mol Syst Biol*, vol. 3, a.n.74, 2007.

[13] J. Ernst, Q.K. Beg, K.A. Kay, et al., "A Semi-Supervised Method for Predicting Transcription Factor-Gene Interactions in Escherichia coli". *PLoS Computational Biology*, vol. 4, e1000044, Mar. 2008

[14] Y. Bengio, P. Frasconi, "An input output HMM architecture". *Adv Neural Inform Process*, vol. 7, pp. 427–434, 1995.