Prediction of genes involved in lignification

Utz J. Pape[1,2], Stephane Rombaults[3], Lieven Sterck[3],
Wout Boerjan[3], Yves Van de Peer[3], and Martin Vingron[1]

1 Introduction

We try to gain deeper insights into the regulatory mechanisms of lignification. Lignin, the second most abundant polymer in nature, is critical for plant growth, development and fitness. In addition, according to Jung & Ni (1998) it influences paper production as it has to be removed from wood for high-quality paper. In the review by Sandermann (2004) it becomes clear that lignification pathways are interconnected with disease- and stress-responses. Getting deeper insights into disease- and stress-responses is crucial for weed control (Basu *et al.*, 2004) and resistance improvement of economical and oecological important plants (Stuiver & Custers, 2001). Therefore, deeper understanding of the pathways leading to lignification in plants is a fundamental task.

We focus on the tree *Populus trichocarpa* as its genome is sequenced (DOE Joint Genome Institute (JGI), 2005). Anyway, only few experimental data is available for *P. trichocarpa*. This is due to the fact that experiments to uncover genes involved in lignification are difficult to perform in wooden plants as they have a high generation time. Thus, it is reasonable to carry out experiments in model organisms and to transfer the results to wooden plants. In plant biology, the main model organism is *Arabidopsis thaliana* as it is small, has a rapid lifecycle and it is much known about its physiological and biochemical processes (Goodman *et al.*, 1995) as well as its sequence (The Arabidopsis Genome Initiative, 2000). In addition, there are many gene homologies and pathway similarities to *P. trichocarpa* especially in relation to lignification (Allona *et al.*, 1998; Hertzberg *et al.*, 2001).

Raes *et al.* (2003) give a genome-wide characterization of genes involved in lignification. One characteristic is an AC element in the upstream region of the genes. The AC element consists of 10bp. Due to the small size of the AC element, one expects to find many occurences by chance, thus, further restrictions are necessary. Experimental results suggest that the AC element is located on the + strand within 500 bp of the promoter. Therefore, we apply a statistically sound method to find the AC element in upstream regions of *P. trichocarpa*. To decrease the number of false positives, we only consider orthologous genes with a predicted AC element in *P. trichocarpa* and *A. thaliana*. A further decrease of non-functional AC elements is done by integrating tissue-specific microarray expression data from *A. thaliana*.

2 Methods

2.1 Motif Finding

We use the approach Rahmann *et al.* (2003) to find the AC element, which we call motif in the subsequent text. The position weight matrix (PWM) is constructed based on the five experimentally verified AC elements Raes *et al.* (2003), see table 1.

One has to select a fraction of false positives (type I error) or false negatives (type II error) to define a score threshold. We choose both fractions to be equal. The threshold can be calculated based on the score distributions of the null model (background) and motif model (signal) without further parametric distributional constrains as the score is simply the sum of independent random variables. Figure 1 shows both score distributions. Obviously, most scores can uniquely be assigned

[1]Dept. of Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany, E-mail: utz.pape@molgen.mpg.de
[2]Dept. of Mathematics and Computer Science, Free University of Berlin, Berlin, Germany.
[3]Dept. of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Ghent University, Ghent, Belgium.

Eucalyptus I	C	C	C	A	C	C	T	A	C	C
Phaseolus I	C	C	C	A	C	C	T	A	C	C
Phaseolus II	-	C	C	A	C	C	A	A	C	C
Petroselinum II	C	T	C	A	C	C	A	A	C	C
Populus	C	T	C	A	C	C	A	A	C	C

Table 1: Alignment of experimentally confirmed AC elements.

to one of the models. This demonstrates the large power of the AC element matrix to detect motif occurences. Figure 2 contains the type I error with respect to the type II error. The intersection with the diagonal is the point where both errors are equal. Hence, we select a scoring threshold of 0.044 such that type I and type II errors are equal.

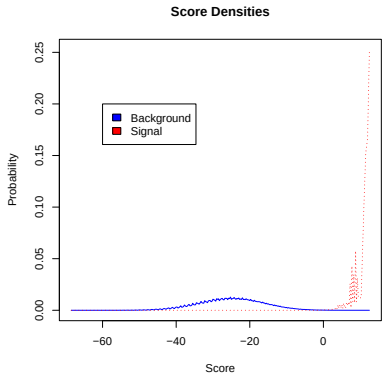

Figure 1: Score Densities for Background and Signal Model.

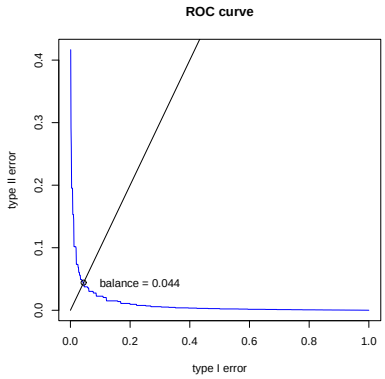

Figure 2: ROC curve for AC element.

2.2 Orthologs

We want to predict genes with functional AC elements. Therefore, we apply the assumption that functionally relevant sequence regions are under selective pressure. Thus, we assume genes with functional AC elements to have orthologs which have an AC element as well. A standard approach to find pairs of orthologs is using BLAST (Altschul *et al.*, 1990) reciprocal best hits. In case of recent genome duplications, the algorithm does not work due to the fact that two highest scoring genes occur. Hence, we have to use a more sophisticated method.

According to Sonnhammer (2002), in-paralogs are paralogs where the duplication occured after the speciation event. Based on this notion of in-paralogs, Remm *et al.* (2001) present a method to identify groups with in-paralogs and their orthologous partners in the other species, which form a group of in-paralogs as well. First, BLAST searches are performed between and within the species. After finding the best reciprocal hit, the corresponding pairs of genes form the main orthologs. BLAST hits with the main ortholog within the species are called in-paralogs if the bit score is higher than the bit score between the main orthologs. We apply this algorithm to all peptide sequences of *A. thaliana* and *P. trichocarpa*.

2.3 Expression Data

Lignification occurs mainly in root and stem. Thus, we assume that genes involved in lignification are highly expressed in these tissues in comparison to other tissues like flower and leaves, which

form the baseline expression. Therefore, we filter the list of genes for differentially expressed genes in root and stem under the condition that the expression is higher in the selected tissues.

Differentially expressed genes can be found by using the empirical Bayes approach (Smyth, 2004). An advantage of this approach is that the variance of gene expression is estimated using partly the estimate of all genes. The regularization decreases the variance of the estimates on the cost of introducing a low bias.

Let's assume that there are n_0 microarrays with baseline expression and n_1 microarrays for root and stem. Thus, we get the total number of microarrays by setting $n = n_0 + n_1$. We denote with $Y_i^{(g)}$ the expression of gene g in the ith microarray. In the following, we neglect the gene index g. The baseline expression is then modeled by

$$Y_i = \alpha + \epsilon \tag{1}$$

for all i. Here, ϵ denotes the error of the model. Apparently, we assume each gene to have a distinct base line expression α. In the root and stem microarrays, the expression is the sum of the baseline expression and a differential expression coefficient β:

$$Y_j = \alpha + \beta + \epsilon \tag{2}$$

We combine these two models by:

$$Y = X(\alpha, \beta)^T + \epsilon$$

where $X \in \{0,1\}^{n \times 2}$ is the design matrix where the first column contains only ones. The first n_0 elements of the second column contain zeros while the remaining elements contain ones. This gives us a multivariate model where the expression on root and stem microarrays is modeled by (2) and on the other microarrays by (1). With this model, we can easily retrieve differentially expressed genes by testing the null hypothesis:

$$H_0 : \beta = 0$$

The null hypothesis H_0 is true for all genes which do not change from the baseline expression in the root and stem tissues. Upregulated genes are the ones with a positive coefficient β. As we test thousands of genes, we run into a multiple testing problem. Thus, we correct the p-values according to Benjamini & Hochberg (1995) and control the false discovery rate (FDR). We control the FDR at a level of 0.01. That is the expected proportion of false positives among the significant genes.

The data is taken from a subset of the comprehensive AtGenExpress Atlas generated by Schmid *et al.* (2005). Variance Stabilizing Normalization (Huber *et al.*, 2002) is used for normalization. We select 21 microarrays from flower tissues each with one replicate resulting in $n_0 = 21 \cdot 2$ microararys for the null model. The number of microarrays for root and stem tissues is $n_1 = 15 \cdot 2$. The original atlas contains more microarrays. Still we do not take knock-outs into account as the knock-outs are not related to root and stem tissues. In addition, the unrelated knock-outs would increase the expression variability within the background model such that α gets less weight while β might explain knock-out effects instead of differential expression. This would result in a higher number of false positives.

3 Results

3.1 Prediction based on Sequences

In the 500bp upstream regions of *A. thaliana*, we detect 2208 hits for an AC motif. In the 1000bp upstream regions of *P. trichocarpa*, the number is 6824. To exclude functionally irrelevant genes, we only take predicted AC elements into account if they are present in the upstream regions of at least two orthologous genes. This results in 363 *A. thaliana* genes and 432 *P. trichocarpa* genes.

3.2 Prediction Refinement with Expression Data

Among the set of 363 *A. thaliana* genes, we find 207 differentially expressed genes between root and stem microarrays and flower microarrays. Still, only 127 of these genes are upregulated.

We compare our prediction to the literature. Raes *et al.* (2003) provide a list of 61 genes involved in lignification. We identify genes of those which where known to contain an AC element (6 genes). In addition, there are 13 genes related to stress response in the prediction set. According to Sandermann (2004), genes involved in lignification are often related to stress response since detoxification is a results of binding to the cell wall. Furthermore, 4 Myb genes are detected which are likely involved in lignification expression (Newman *et al.*, 2004). Another 3 genes belong to the cytochrome P450 family, which was already described to have a relation to lignification by Meyer *et al.* (1996). The impact for lignification of 3 genes related to auxin is shown by Ljung (2002). Another large fraction of the resulted set comprises transcription factors and more generally DNA binding proteins. Many of the remaining predicted genes have no annotated function.

4 Discussion

In principle, we give preliminary hints that it is possible to utilize experimental data from a model organism to get results for a related species. The advantage is that expensive experiments can be reused. In our case, we use experimental data from *A. thaliana* and the sequences from *A. thaliana* and *P. trichocarpa*. We retrieve a set of genes related to lignification. So far, we focus on the predicted *A. thaliana* genes as the annotation for the *P. trichocarpa* genome is not yet finished. In addition, much more is known about *A. thaliana* such that it is worth to give evidence for the prediction based on this species. In the near future, we will make use of orthology relations to predict *P. trichocarpa* genes for lignification based on this analysis.

The results for *A. thaliana* are promising as we find many genes which are very likely involved in lignification. We have given some evidence by assigning the predicted genes to functional groups related to lignification. As this was done manually, we are aware that further evidence is required. Therefore we are planning to integrate the GO Annotation (The Gene Ontology Consortium, 2000) for *A. thaliana* (Berardini *et al.*, 2004). With this data by hand, we are able to compute p-values for overrepresented categories. Ideally, we would find the category 'lignin metabolism' to be overrepresented. In addition, many of our predicted genes should be annotated with 'lignin catabolism'. Unfortunately, this category contains nothing. Other predicted genes should also fall into the stress response categories. In this case, analysis will be difficult as there does not exist one main category for this function but many nodes which are located in many different branches. Further on, the correctness of prediction has to be approved by biological experiments.

References

Allona, I., Quinn, M., Shoop, E., Swope, K., Cyr, S. S., Carlis, J., Riedl, J., Retzel, E., Campbell, M. M., Sederoff, R. & Whetten, R. W. (1998) Analysis of xylem formation in pine by cDNA sequencing. *PNAS,* **95** (16), 9693–9698.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215** (3), 403–410.

Basu, C., Halfhill, M. D., Mueller, T. C. & Jr., C. N. S. (2004) Weed genomics: new tools to understand weed biology. *Trends in Plant Science,* **9** (8), 391–398.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series,* **57**, 289–300.

Berardini, T. Z., Mundodi, S., Reiser, R., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. M., Yoon, J., Doyle, A., Lander, C., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M.,

Miller, N., Weems, D. & Rhee, S. Y. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.,* **135** (2), 1–11.

DOE Joint Genome Institute (JGI). The Populus trichocarpa Genome. http://genome.jgi-psf.org/poplar/.

Goodman, H. M., Ecker, J. R. & Dean, C. (1995) The genome of Arabidopsis thaliana. *PNAS,* **92**, 10831–10835.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T. T., Lundeberg, J., Sundberg, B., Nilsson, P. & Sandberg, G. (2001) A transcriptional roadmap to wood formation. *PNAS,* **98** (25), 14732–14737.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics,* **18** (1), S96–S104.

Jung, H.-J. G. & Ni, W. (1998) Lignification of plant cell walls: impact of genetic manipulation. *Proc. Natl. Acad. Sci.,* **95**, 12742–12743.

Ljung, K. (2002). *Auxin biosynthesis and homeostasis in Arabidopsis thaliana in relation to plant growth and development.* PhD thesis, Acta Universitatis agriculturae Sueciae. Silvestria.

Meyer, K., Cusumano, J. C., Somerville, C. & Chapple, C. C. S. (1996) Ferulate-5-hydroxylase from Arabidopsis thaliana defines a new family of cytochrome P450-dependent monooxygenases. *Proc. Natl. Acad. Sci.,* **93**, 6869–6874.

Newman, L. J., Perazza, D. E., Juda, L. & Campbell, M. M. (2004) Involvement of the r2r3-myb, atmyb61, in the ectopic lignification and dark-photomorphogenic components of the det3 mutant phenotype. *Plant J,* **37** (2), 239–239.

Raes, J., Rohde, A., Christensen, J. H., de Peer, Y. V. & Boerjan, W. (2003) Genome-Wide Characterization of the Lignification Toolbox in Arabidopsis. *Plant Physiology,* **133**, 1051–1071.

Rahmann, S., Müller, T. & Vingron, M. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.,* **2** (7).

Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.,* **314**, 1041–1052.

Sandermann, H. (2004) Molecular ecotoxicology of plants. *TRENDS in Plant Science,* **9** (8).

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. & Lohmann, J. U. (2005) A gene expression map of Arabidopsis development. *Nature Genetics,* **doi:10.1038/ng1543**.

Smyth, G. K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology,* **3** (1).

Sonnhammer, E. L. L. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *TRENDS in Genetics,* **18** (12), 619–620.

Stuiver, M. H. & Custers, J. H. H. C. (2001) Engineering disease resistance in plants. *Nature,* **411**, 865–868.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* **408**, 796–815.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics,* **25**, 25–29.