



Inference of the genetic network regulating lateral root initiation in *Arabidopsis thaliana*

by

**Daniele Muraro
Ute Voß
Michael Wilson
Malcolm Bennett
Helen Byrne
Ive De Smet
Charlie Hodgman
John King**

Inference of the genetic network regulating lateral root initiation in *Arabidopsis thaliana*

Daniele Muraro, Ute Voß, Michael Wilson, Malcolm Bennett, Helen Byrne, Ive De Smet, Charlie Hodgman, and John King

Abstract—Regulation of gene expression is crucial for organism growth, and it is one of the challenges in Systems Biology to reconstruct the underlying regulatory biological networks from transcriptomic data. The formation of lateral roots in *Arabidopsis thaliana* is stimulated by a cascade of regulators of which only the interactions of its initial elements have been identified. Using simulated gene expression data with known network topology, we compare the performance of inference algorithms, based on different approaches, for which ready-to-use software is available. We show that their performance improves with the network size and the inclusion of mutants. We then analyse two sets of genes, whose activity is likely to be relevant to lateral root initiation in *Arabidopsis*, by integrating sequence analysis with the intersection of the results of the best performing methods on time series and mutants to infer their regulatory network. The methods applied capture known interactions between genes that are candidate regulators at early stages of development. The network inferred from genes significantly expressed during lateral root formation exhibits distinct scale-free, small world and hierarchical properties and the nodes with a high out-degree may warrant further investigation.

Index Terms—Reverse engineering, gene expression data, *Arabidopsis thaliana*.

1 INTRODUCTION

Extensive branching of lateral roots significantly influences the efficiency of nutrient uptake and anchorage by plants [1], [2], [3]. Understanding the molecular basis for the initiation of lateral roots is therefore of significant agronomical interest. Among plants, *Arabidopsis thaliana* is the best characterised organism and it is widely adopted as a model organism to investigate developmental processes. Genes that are involved in the first stages of a signalling cascade which promotes the formation of lateral roots have been identified and some of their regulatory interactions have been determined [3].

A comparative analysis that combines transcriptomic datasets specific for lateral-root initiation shows that about two hundred genes are significantly expressed during this process [4]; nevertheless, the underlying genetic network is still unknown. Recently, the Centre for Plant Integrative Biology (CPiB) at the University of Nottingham has collected a time-course transcriptomic

dataset from tissues involved in lateral root development in wild type roots. The integration of such novel experiments and their analysis by applying reverse engineering algorithms may provide valuable preliminary information to identify and further examine potential novel regulatory interactions.

Numerous algorithms based on different approaches have been proposed for reverse engineering genetic networks from expression data. They can be grouped into three main categories: information theoretic approaches, dynamic models, and Bayesian networks. Although correlation based approaches are generally applied to the clustering of data, they can also be used to infer undirected graphs [5] and considered as a fourth category. Nevertheless, the absolute and comparative performance of these methods is still poorly understood.

A concerted effort to address this problem has been proposed with the DREAM (Dialogue on Reverse Engineering Assessment and Methods) project [6], [7], [8]. In particular, in the DREAM3 challenge the performance of 29 algorithms of different categories has been analysed on synthetic networks of different size (10, 50, 100 nodes). The authors observed that the algorithm performance on the different datasets was not clearly dependent on the type of inference approach applied and that the results were dependent on the type of dataset chosen.

In order to determine which methods are most appropriate for analysing our experimental dataset, we test the performance of different categories of inference algorithms, for which ready-to-use software is available, on simulated data that mirror our experimental time courses and mutants. We quantify how the outcome of these algorithms is affected by the availability of

- M. Bennett, H. Byrne, I. De Smet, C. Hodgman, J. King, D. Muraro, U. Voß, M. Wilson are with the Centre for Plant Integrative Biology, School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough LE12 5RD, UK. E-mails: Name.Surname@nottingham.ac.uk (Ive De Smet's e-mail: Ive.De_Smet@nottingham.ac.uk)
- H. Byrne is also with the Oxford Centre for Collaborative Applied Mathematics and Department of Computer Science, University of Oxford, Oxford, OX1 3LB, UK and with the School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK. E-mail: Helen.Byrne@maths.ox.ac.uk
- J. King is also with the School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK. E-mail: John.King@nottingham.ac.uk
- D. Muraro is also with the Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, OX3 9DS, UK. E-mail: Daniele.Muraro@ndm.ox.ac.uk

knock-out mutants and we show that their performance on simulated data is enhanced when the network size increases.

Further improvement can be obtained by integrating inference algorithms with the information extracted from genome sequence data, as proposed in [9]. Once the interactions are inferred, it is possible to align the promoter regions of the predicted targets with each other to detect motifs that may correspond to a binding site of a common transcription factor. Applying this approach to the network inferred from experimental datasets such as ours may allow the reduction of the number of false positive predictions.

This article is organised as follows. In section 2 we review the four main categories for reverse engineering genetic networks, focusing, in particular, on the algorithms that we have tested. In section 3.1 we compare the performance of these algorithms on simulated data with known network topology and select the best performing approaches. In section 3.2 we integrate sequence analysis with the intersection of the best performing methods to genes which are potentially involved in lateral root initiation to infer their interaction network. We then analyse the network topological properties and show that it exhibits scale-free, hierarchical and small world properties.

2 METHODS

Four main categories of approaches are commonly applied when inferring genetic networks from transcriptomic data. In what follows, we will review these methods focusing, in particular, on the algorithms that we have tested on networks with known topology. The dimensionality of the network provides a first criterion for the algorithm choice. Correlation and information theoretic methods are more appropriate for networks of large dimension because they require a shorter computational time by limiting the analysis to pairwise interactions. By contrast, dynamical models and dynamic Bayesian networks are more conveniently applied to small networks since they enable the investigation of the joint action of each set of parents on the corresponding set of children.

2.1 Correlation-based approaches

Co-expression networks are a common framework within which to analyse gene expression data [10], [11], [12], [13]. Genes are grouped in different clusters according to the similarity of their expression profiles, which is measured using a metric such as the Pearson correlation coefficient. The underlying assumption is that genes that are highly co-expressed, and are grouped in the same cluster, have a higher probability of regulating each other. In [5] hierarchical clustering has been applied as an inference method assuming that each gene is connected to all the other genes of the same cluster. In

this way the authors were able to infer an undirected graph.

2.2 Information theoretic approaches

Mutual information (MI) is a measure of the degree of independence between two random variables [14], [15]. Assuming M expression levels for N genes, we associate the random variable X_i with gene i and denote by $x_{i,j}$ for $i = 1, \dots, N$ and $j = 1, \dots, M$ the j -th state of gene i . Defining the entropy of gene i as

$$H(X_i) := - \sum_{j=1}^M p(x_{i,j}) \log_M p(x_{i,j}),$$

the mutual information of genes X_i and X_j is

$$MI(X_i; X_j) := H(X_i) + H(X_j) - H(X_i, X_j)$$

From the definition it can be easily seen that mutual information is a symmetric quantity and that the result of the inference procedure is an undirected graph. Well-known algorithms based on mutual information include ARACNE [16], CLR [17] and DTI [9].

ARACNE evaluates the mutual information of all gene pairs and computes a statistical MI threshold by shuffling randomly the expression of genes across microarray profiles. Then, it performs a first selection based on a p -value associated with this statistical threshold. Finally, ARACNE applies a second selection which attempts to eliminate cascade errors by examining all the triplets of genes that interact with each other and by eliminating, in each triplet, the interaction with lowest mutual information. This deletion is justified by appealing to the Data Processing Inequality (DPI) which is defined as follows

$$MI(X_1; X_3) \leq \min[MI(X_1; X_2); MI(X_2; X_3)],$$

under the assumption that X_1 is indirectly regulating X_3 via X_2 .

The CLR algorithm was introduced in [17] and extended in [18], [19] for the DREAM competitions. CLR increases the contrast between direct and indirect relationships using the local network context to compute a statistical metric of similarity between expression profiles. Let us assume that M is an $N \times N$ matrix whose elements are defined as $M_{i,j} := MI(X_i, X_j)$, where $i \in \{1, \dots, N\}$. We compute the positive Z-scores for $M_{i,j}$ with respect to the entries of the i -th row and j -th column as follows

$$z_i(X_i, X_j) := \max \left(0, \frac{M_{i,j} - \frac{\sum_{j'} M_{i,j'}}{N}}{\sigma_i} \right),$$

$$z_j(X_i, X_j) := \max \left(0, \frac{M_{i,j} - \frac{\sum_{i'} M_{i',j}}{N}}{\sigma_j} \right)$$

The CLR likelihood estimate is then defined as the pseudo Z-score

$$z(X_i, X_j) := \sqrt{z_i(X_i, X_j)^2 + z_j(X_i, X_j)^2}$$

Under the assumption that the biological regulatory network is sparse, most of the $M_{i,j}$ scores in each row and column of the mutual matrix M represent a random background. In CLR this random background is approximated as a joint normal distribution with the i -th row and j -th column of M as random variables. By imposing a threshold on the Z-score, it is possible to recover an undirected graph.

DTI quantifies the information flow between the expression time series of two genes and, in such a way, allows the reconstruction of a directed graph. Directed mutual information is defined as

$$MI(X_i^T \rightarrow X_j^T) = \sum_{t=2}^T \left(H(X_i^t | X_j^{t-1}) - H(X_i^t | X_j^t) \right)$$

where $X_i^t := (X_{i,1}, \dots, X_{i,t})$ is a segment of the realisation of the random sequence X_i . A combination of DTI and CLR has been applied in [9].

Mutual information provides a more general estimate of statistical dependence than the correlation coefficient, the latter quantifying only linear relationships between variables. It also permits extensions to analyse cooperative regulations [18], [20]. For ARACNE and DTI ready-to-use software is available and will be applied in section 3.

2.3 Dynamic models

Dynamic models are generally based on ordinary differential equations that model the transcription of each gene as a function of its regulators. In order to limit the number of equations and parameters, transcription and translation are typically grouped in a single variable whose concentration is regulated by a transcription function. In a group of N genes, denoting by x_i and by f_i respectively the mRNA concentration transcribed by the i -th gene and its transcription rate function, the network is modelled by the system

$$\frac{dx_i}{dt} = f_i(\vec{x}(t), \vec{p}), \quad i = 1, \dots, N$$

where \vec{p} is a set of parameters describing the interactions between the genes (activations and repressions). The inference problem is thus reduced to the optimisation problem of calibrating the parameters \vec{p} against the experimental data. Several transcription functions have been used in the literature. One of the simplest choices is linear transcription, as considered for example in the TSNI algorithm [21], where the possibility of the inclusion of an external perturbation is also taken into account. In this case the transcription function is assumed to be of the form

$$\frac{dx_i}{dt} := \sum_{j=1}^N a_{ij} x_j(t) + \sum_{l=1}^P b_{il} u_l(t), \quad i = 1, \dots, N$$

where a_{ij} represents the influence of gene j on gene i ; b_{il} represents the effect of the l -th perturbation on x_i and $u_l(t)$ represents the l -th external perturbation at time t .

Other approaches that use linear models can be found in [19], [22]. Other transcription functions are of sigmoidal type [22], [23], e.g.

$$\frac{dx_i}{dt} := \frac{b_{i1}}{1 + \exp(-a_{i0} - \sum_{j \in S} a_{ij} x_j)} - b_{i2} x_i$$

or account for stochastic effects

$$\frac{dx_i}{dt} := \alpha_i + f_i(\vec{x}_i(t)) - \lambda_i x_i(t),$$

where α_i is the basal transcription rate, λ_i is the mRNA decay rate, $\vec{x}_i(t)$ denotes the expression levels of genes that regulate gene i , and f_i is a Gaussian process

$$f_i(\vec{x}_i) \sim GP(m(\vec{x}_i), k(\vec{x}_i, \vec{x}_i')),$$

$m(\vec{x}_i)$ and $k(\vec{x}_i, \vec{x}_i')$ being the mean and covariance functions [24]. The hybrid method described in [22] ranked first on all network sizes in the DREAM3 challenge. The method described in [24] ranked fifth in the DREAM3 challenge on networks of size 50. Since ready-to-use software is available both for this last method, which we will denote for brevity as GP, and TSNI, we focus on applying these two methods in section 3.

2.4 Dynamic Bayesian networks

A Bayesian network is a probabilistic graphical model that represents conditional dependencies between random variables in a directed acyclic graph and can be defined as follows [25]. Let us consider a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ and associate each variable to a node of a directed acyclic graph. We denote by \mathbf{Pa}_i the set of parents of node X_i as well as the variables corresponding to those parents. Given any set of random variables \mathbf{X} , the joint probability distribution of any member $\mathbf{x} = \{x_1, \dots, x_n\}$ can be written according to the chain rule as follows

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1) \cdot \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_1 = x_1).$$

In particular, associating \mathbf{X} with the graph and denoting by \mathbf{pa}_i the set of values of \mathbf{Pa}_i , the chain rule becomes

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n P(X_i = x_i | \mathbf{Pa}_i = \mathbf{pa}_i)$$

The absence of any possible directed edges in the particular graph considered is encoded as a conditional independency. Bayesian networks are based on static data.

In order to analyse time series data it is necessary to consider an extension to dynamic Bayesian networks (DBNs) [26]. Let us suppose that the microarray data measure the time-series expression of N genes at T data points. We denote the microarray dataset by $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)^T$ where \mathbf{x}_t corresponds to the expression level of the N genes at the time point t . We assume that

the set of data follows a first-order Markov model in which the state vector at time t depends only on that at time $t - 1$. Under this assumption, together with one of stability of the network structure at all time points, it is possible to model any graph as a directed acyclic graph. Assuming that gene j has p_j parents, the gene regulation can thus be modelled by the conditional probability

$$P(\mathbf{X}_t|\mathbf{X}_{t-1}) = P(X_{t,1}|\mathbf{Pa}_{t-1,1}) \times \dots \times P(X_{t,N}|\mathbf{Pa}_{t-1,N}),$$

for $t = 1, \dots, T$, where

$$\mathbf{Pa}_{t-1,j} = (P_{t-1,1}^{(j)}, \dots, P_{t-1,p_j}^{(j)})^T$$

is a vector of random variables of the genes which are parents of the j -th gene at time $t - 1$. By applying the chain rule the joint probability can be calculated as

$$P(\mathbf{X}) = P(\mathbf{X}_1)P(\mathbf{X}_2|\mathbf{X}_1) \times \dots \times P(\mathbf{X}_T|\mathbf{X}_{T-1}).$$

The goal is now to select the graph that best fits the data by optimising a convenient objective function.

Bayesian network algorithms vary according to the knowledge about the observability of the data (full, partial) and to the nature of the random variables (discrete, continuous).

Here, we consider two dynamic Bayesian network algorithms. The first one assumes complete knowledge of the data (full observability) and is applied to their discretization. The second one presumes partial observability and continuous random variables which are subject to linear dynamics.

A discrete model has been proposed by [26] based on the REVEAL algorithm [27]. The method applies a greedy search of the best possible set of parents for each node and optimises the score function

$$S(G) := \frac{MI(X, \mathbf{Pa}(X))}{\max\{H(X), H(\mathbf{Pa}(X))\}}$$

by using the equivalence between the maximum likelihood and the sum of the mutual information between each node and its parents, [28]. An implementation of this algorithm can be found in the Bayesian Networks Toolbox (BNT) in Matlab [29]. Extensions of the REVEAL algorithm have been proposed in [30], [31] and a comparison with Boolean networks has been investigated in [32].

A linear-Gaussian state-space model (VBSSM) has been proposed in [33], [34]. This algorithm attempts to discover gene-gene interactions by assuming that hidden variables are modelling effects which are not measured by the available data, these effects being, for example, levels of regulatory proteins. The equations governing the model are

$$\begin{aligned} \mathbf{x}_t &= A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t, \end{aligned}$$

where \mathbf{y}_t are the levels of gene expression at time t , \mathbf{x}_t are the hidden variables, and \mathbf{w}_t and \mathbf{v}_t are variables representing Gaussian noise. The hyperparameters of the

model are optimised by applying a Variational Bayesian Expectation - Maximization algorithm. In section 3 we apply and compare REVEAL and VBSSM.

3 RESULTS

We report in the following sections the results of our evaluations. In section 3.1 we compare the performances of the inference algorithms described above on simulated and experimental data with known network topology. In section 3.2 we present the results of the integration of the best performing algorithms with sequence analysis on datasets of interest for lateral root initiation.

3.1 Evaluation on networks with known topology

Simulated data were generated using the GreenSim simulator [35] which generates network topologies with scale-free out-degree and an exponential in-degree. It also applies empirical tests to check that the sampled network and individual time series do not produce unrealistic behaviour that is uninteresting to infer. Gene expression is updated at discrete intervals according to

$$Y_{t+1} = Y_t + f(A, Y_t - T) + \epsilon,$$

where Y_t is a vector denoting the dimensionless expression level of each gene at time t , f is a function controlling gene regulation, A is a matrix which specifies the gene functional relationships, T is a vector denoting the base expression levels of each gene, and ϵ is Gaussian noise. This modelling approach has been shown to be able to generate an interesting variety of patterns, such as damped oscillations or limit cycles [35].

Steady state data for wild type and knock-out mutations are simulated by assuming that transcription is regulated by Hill functions and assigning a null value to the transcription rate of the mutated gene. Gene interactions are then inferred by analysing the relative changes of the wild type and the mutants. In the following simulations, when including knock-outs, we follow [22], giving more confidence to predictions obtained from mutants than from time series and we select the interactions inferred from the knock-out simulations.

We test the algorithms on time series with a number of replicates similar to the experimental data available for lateral root initiation. We then evaluate how the inclusion of a relatively small number of mutants affects reconstruction.

We evaluate the performances of the algorithms by using the Positive Predictive Value (PPV) and Sensitivity (Se) [36], [37]:

$$PPV := \frac{TP}{(TP + FP)}, \quad Se := \frac{TP}{(TP + FN)},$$

where TP, FP, FN are respectively the number of true positives, false positives and false negatives. The random PPV has been computed as in [38] considering the expected value $E[X]$ of a hypergeometrically distributed random variable which is normalised by the number of

N. GENES	VBSSM		TSNI		RANDOM
	PPV	Se	PPV	Se	
5 undirected	1.0	0.4	0.5	0.4	0.5
5 directed	0.5	0.2	0.4	0.4	0.25
5 signed	0.5	0.2	0.2	0.2	0.13
10 undirected	0.67	0.1	0.45	0.72	0.4
10 directed	0.67	0.1	0.33	0.72	0.2
10 signed	0.67	0.1	0.23	0.5	0.1
20 undirected	0.29	0.06	0.15	0.93	0.16
20 directed	0.14	0.03	0.08	0.87	0.08
20 signed	0.14	0.03	0.06	0.7	0.04
50 undirected	0.62	0.18	0.10	0.61	0.1
50 directed	0.56	0.16	0.05	0.37	0.05
50 signed	0.56	0.16	0.04	0.29	0.02
100 undirected	0.53	0.12	0.05	0.34	0.05
100 directed	0.49	0.11	0.03	0.26	0.02
100 signed	0.49	0.11	0.03	0.24	0.01

(a)

N. GENES	VBSSM		TSNI		RANDOM
	PPV	Se	PPV	Se	
10 undirected	0.75	0.17	0.47	0.78	0.4
10 directed	0.75	0.17	0.38	0.78	0.2
10 signed	0.75	0.17	0.27	0.56	0.1
20 undirected	0.29	0.07	0.15	0.94	0.16
20 directed	0.14	0.03	0.08	0.87	0.08
20 signed	0.14	0.03	0.07	0.67	0.04
50 undirected	0.67	0.24	0.11	0.6	0.1
50 directed	0.62	0.22	0.06	0.41	0.05
50 signed	0.62	0.22	0.05	0.33	0.02
100 undirected	0.44	0.26	0.07	0.44	0.05
100 directed	0.44	0.26	0.05	0.37	0.02
100 signed	0.44	0.26	0.05	0.35	0.01

(b)

TABLE 1: Table summarising the Positive Predictive Values (PPV) and Sensitivity (Se) of VBSSM and TSNI. In bold we represent p -values minor than 0.1. The algorithms were applied in Table 1a to simulated time series with 20 replicates and in Table 1b to simulated time series with 20 replicates and 10% of mutants.

predicted edges. Analogously, the p -value of the inferred network has been computed by a hypergeometric distribution.

We apply and compare algorithms that enable the investigation of the joint action of each set of parents on the corresponding set of children on small networks (< 100 genes), whereas we evaluate the performance of information theoretic methods on larger networks since these algorithms limit their analysis to pairwise interactions.

On a small scale we apply VBSSM, TSNI, REVEAL and GP. Because of their limitations in evaluating replicated time series, we compare their outcome on a single time course. In Table 1a and in Supplementary Table 1 we show the results of the application of VBSSM, TSNI, REVEAL and GP to time series with 20 data points. We represent in bold the results of the networks which have been reconstructed with a p -value < 0.1 . TSNI and VBSSM performed better when the network size was increased, whereas GP and REVEAL returned random results with few exceptions and were more limited in their ability to infer the dimensions and type of the network (the sign of the interactions was not evaluated).

In Table 1b and in Supplementary Table 2 we show how these results are affected by including a number of mutants comprising 10% of the network dimension, this being a percentage of mutants analogous to that experimentally available in the lateral root initiation datasets. The general performance of the algorithms improved, and a larger number of significant predictions were found.

On a larger scale, we estimate the outcome of DTI and ARACNE. DTI can be applied to replicated time courses and ARACNE evaluates each replicates

independently of time. In Table 2a we show the results of the application of ARACNE to networks whose dimension is at least 100 genes. The performance of this algorithm depends on a reliable estimate of mutual information which, in turn, relies on a sufficient number of replicates. In our tests, ARACNE always performed reliably with 70 replicates, although its results were poorer with a smaller amount of replicates. Meaningful results can also be obtained by applying DTI with an analogous number of data points (Table 2b).

3.2 Inference of the genetic network regulating lateral root initiation

We briefly summarise here the main events that control the initiation of lateral roots in *Arabidopsis thaliana* [2], [3]. Lateral root formation is controlled by well-balanced signalling pathways that involve different hormones in specific regions of the main root. The determination of the region in which the primordium is initiated is triggered by the local accumulation of the hormone auxin in a population of founder cells located in the pericycle, a cylindrical tissue between the stele and the endodermis. Initiation is limited to a small number of founder cells and is detected when these cells undergo several rounds of asymmetric divisions. The founder cells are primed for division in response to elevated levels of the hormone auxin via a signalling cascade that comprises some known and mostly unknown regulators [39], [40].

A spreadsheet tool which combines different microarray datasets involved in lateral root development has recently been made publicly available [4]. By applying filters on such combination of data, the authors selected

N. GENES	ARACNE 20 replicates		ARACNE 70 replicates		RANDOM
Wild Type	PPV	Se	PPV	Se	PPV
100 undirected	0.05	0.1	0.09	0.12	0.05
200 undirected	0.018	0.036	0.05	0.07	0.025
500 undirected	0.01	0.02	0.02	0.04	0.01
1000 undirected	0.01	0.08	0.01	0.04	0.01
1% Mutants					
100 undirected	0.08	0.19	0.09	0.12	0.05
200 undirected	0.019	0.038	0.05	0.07	0.025
500 undirected	0.01	0.03	0.02	0.05	0.01
1000 undirected	0.02	0.15	0.02	0.05	0.01

(a)

N. GENES	DTI 4 replicates 18 time points		RANDOM
Wild Type	PPV	Se	PPV
100 undirected	0.063	0.714	0.055
100 directed	0.033	0.502	0.028
200 undirected	0.026	0.676	0.024
200 directed	0.013	0.447	0.012
500 undirected	0.012	0.700	0.010
500 directed	0.007	0.519	0.005
1% Mutants			
100 undirected	0.063	0.711	0.055
100 directed	0.034	0.502	0.028
200 undirected	0.026	0.674	0.024
200 directed	0.013	0.447	0.012
500 undirected	0.012	0.699	0.010
500 directed	0.007	0.519	0.005

(b)

TABLE 2: Tables summarising the Positive Predictive Values (PPV) and Sensitivity (Se) of ARACNE (Table 2a) and DTI (Table 2b) on simulated time series. In bold we represent *p*-values minor than 0.1.

211 genes that are significantly regulated during the initiation of lateral roots. By applying a further filter to select only genes with pericycle-specific expression, they obtained a list of 19 genes which are likely to be involved in the first stages of the initiation.

We analyse the possible interactions between these two sets of potential regulators, and investigate their role in concert with five proteins whose regulatory function in the signalling cascade is already known (IAA14, ARF7, ARF19, LBD16, LBD29). When these known genes are included, the two lists comprise 24 and 212 elements, the former including candidate regulators at an early stage of lateral root initiation, the latter candidate genes that are also involved in later stages.

The Centre for Plant Integrative Biology (CPIB) at the University of Nottingham has collected a time course transcriptomics dataset to investigate gene expression profiles from tissues involved in lateral root development in wild type roots, this dataset consisting of four replicates of time series, each with 18 data points. We combine experimental data for the mutants of IAA14, ARF7, and ARF19 [41], [42] with significantly expressed time series from the CPIB dataset. We apply the methods which performed best to these sets of genes and selected potential interactions obtained by the intersection of their results, following [8]; we then integrated an analysis of motif occurrences to reduce false positives, as suggested in [9]. We investigated the intersection of TSNI and VBSSM to the set of 24 genes and the intersection of ARACNE and DTI to the set of 212 genes. Motif occurrence was examined using the sequence alignment tool *cosmo* [43]. Selection was performed by analysing conserved motifs in each set of potential children and by selecting those that have posterior probabilities of motif occurrences along a given sequence that exceeds

a threshold which is fixed at 0.9. This further selection reduces the size of the interaction networks to 19 and 113 nodes respectively.

In Figure 1 we show the interaction network of 19 genes. Some of the gene interactions are known and compatible with those inferred. In particular, LBD16, LBD29 and GATA23 are known to act downstream of ARF7, ARF19 and to regulate lateral root formation [44], [45]. LBD29 is activated by ARF19 which binds it directly [44]; however, this effect may be obtained by a direct repression through ARF19 and a stronger and indirect activation by ARF19 via LBD16. In addition, 1-kb promoter regions of four genes acting downstream the ARFs have TGTCTC or GAGACA sequences that may play the role of auxin response elements (AuxREs) and indicate a direct regulation of ARF7 or ARF19 [44], [46]. ARF7 is also known to activate ARF19 and LBD29, but these interactions were not inferred since the target genes are not significantly expressed in the ARF7 mutant dataset [3]. The associated edges are represented in bold in Figure 1.

Other inferred interactions may impact the formation of lateral roots. ACS8 is a gene involved in the biosynthesis of ethylene, whose enhanced levels promote the initiation of lateral root primordia. Mutations that block auxin responses, *slr1* and *arf7 arf19*, render initiation of lateral root primordia insensitive to the promoting effect of ethylene [47], [48]. Activation of ACS8 by ARF7 could play a role in this hormonal interaction, see Figure 1. CKX6 is a member of the cytokinin oxidase / dehydrogenase (AtCKX) gene family which catalyses the irreversible degradation of cytokinins and in many plant species is responsible for the majority of metabolic cytokinin inactivation [49]. Cytokinin influences lateral root initiation interacting with auxin signalling [50]; so,

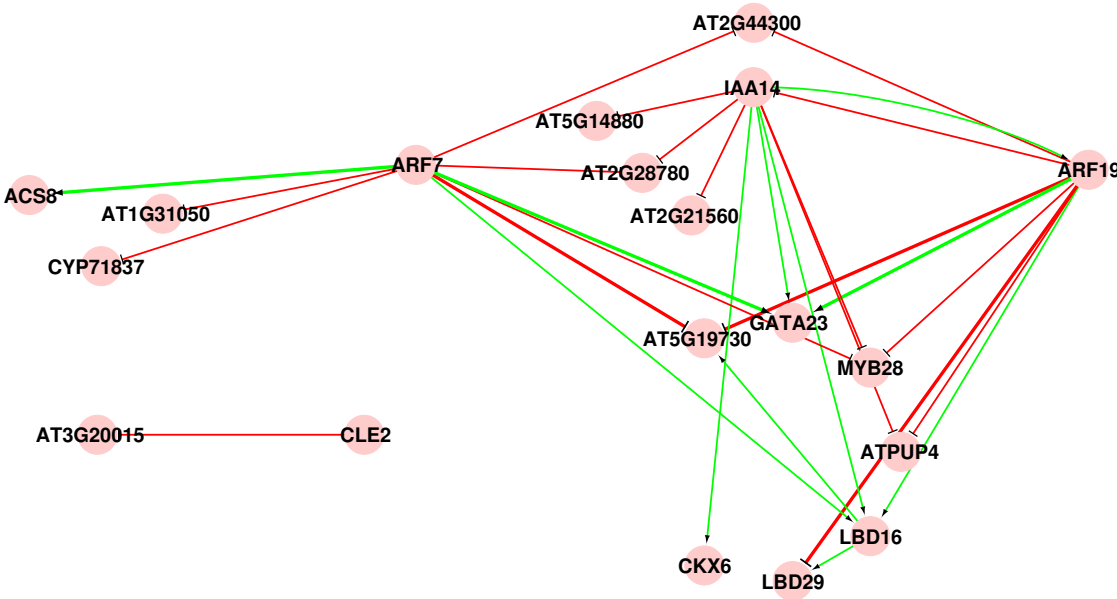


Fig. 1: Inferred network for the set of 19 genes involved in lateral root initiation. Green arrows represent activations, whereas red arrows represent repressions. In bold are presented transcriptional regulations in which an AuxRE motif is present in the target promoter region.

it is possible that part of such cross-talk occurs through an auxin response gene (IAA14), see Figure 1.

In Figure 2 we show the inferred network of 113 genes. A functional categorisation by molecular function highlights a higher presence of genes encoding known transcription factors in this set (18.69%) than in the genome (6.12%), see Supplementary Figure 1. The larger dimension of the full set of candidate genes involved in lateral root initiation permits an analysis of the topological properties of the inferred network. We investigate these properties and fit the datapoints of the associated topological distributions using the Cytoscape plugin Network Analyzer [51]. A network layout which groups nodes with the same out-degree highlights the presence of few hubs. Edge thickness and darkness are proportional to the edge betweenness and reveal that a high number of shortest paths pass through few edges.

Typical parameters and distributions that characterise the network are reported in Table 3 and Figure 3. The network is composed of a single connected component but presents a low density value, this indicating that the number of edges is low relatively to the network dimension, see Table 3. Both the in-degree and out-degree distributions can be fitted with a power law function exhibiting a scale free topology, see Figures 3a and 3b. The significance of the fits has been estimated using R-squared values, which measure the correlation between the data points and the corresponding points on the fitted curve. The quality of our fits follows the general trend of the analysis of gene expression networks with scale-free properties, in which the R-squared values

are above 0.6 – 0.7 [52], [53]. Small world properties are highlighted by the characteristic path length, this being of the magnitude order of the logarithm of the network dimension [54], see Table 3. The clustering coefficient distribution follows a power law indicating a hierarchical organisation, see Figure 3c, whereas the decrease of the topological coefficient distribution with the increase in number of neighbours indicates a modular network organisation [52], [55], see Figure 3d.

The node with the highest out-degree (AT1G54690) encodes a histone protein that plays a role in nuclear structure. Hence, its position as a hub is not surprising. It has high betweenness edges with AT5G54490 (PBP1, a regulatory calcium binding protein) and AT4G35220 (unknown function). Other nodes with high out-degrees include an RNA binding protein (AT5G41190), cytochrome oxido-reductase (AT1G26100), CTP-synthetase (AT3G12670, normally associated with seed dormancy), GT2 transcription factor (AT1G76890), a lipid-transport protein (AT1G48750) and, notably, PIN1 auxin efflux carrier (AT1G73590). PIN1 relocates at the site of LRI upon gravitropic curvature [56] and its inhibition by cytokinin disrupts PIN-dependent formation of an auxin maximum during lateral root development [50]. PIN1 first neighbours include a xylem development protein (KNAT7, AT1G62990), a cell wall protein (GulLO5, AT2G46740) and the RNA binding protein AT5G41190 which is another hub above mentioned. A hierarchical layout of PIN1 first neighbours is presented in Figure 2.

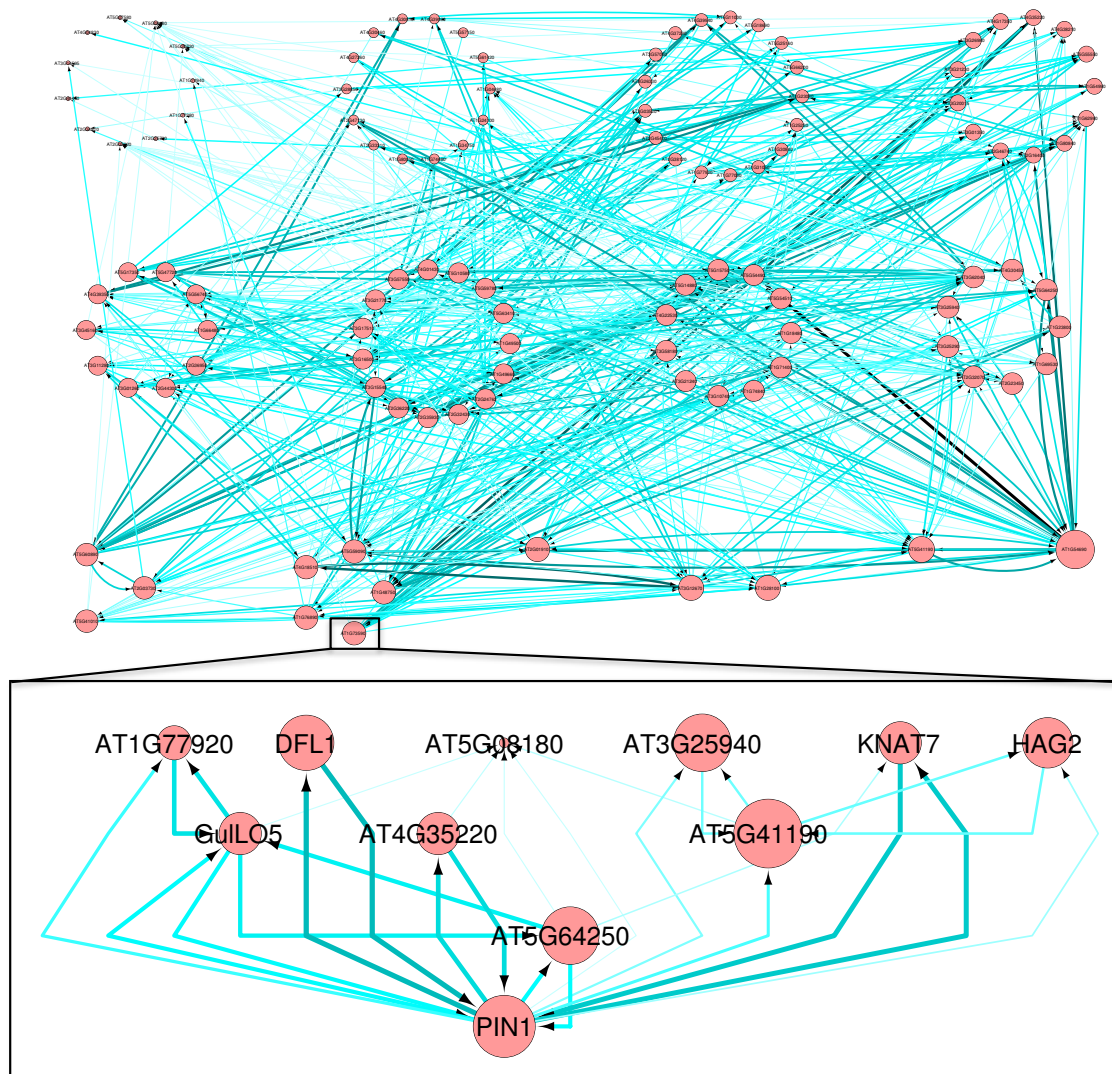


Fig. 2: Inferred network for the set of 113 genes involved in lateral root initiation. Arrows represent a directed regulation which can be either an activation or a repression. The nodes are grouped basing on their out-degree, which is proportional to the node size; notably, PIN1 is among the major hubs and is highlighted with a box. Edge width and darkness is proportional to the edge betweenness. In the box below: PIN1 and its first neighbours are visualised with a hierarchical layout.

4 CONCLUSIONS

Although considerable progress has been made in recent years on the understanding of lateral root development, the growing complexity of the molecular mechanisms under analysis requires new tools that may provide more detailed insight [3]. Increasing knowledge has been gained on the role of different hormones on this biological process but the genetic network comprising these signals is still to be determined. Mathematical modelling may help identify and analyse regulatory networks from gene expression experiments.

Many reverse-engineering methods have been proposed in the literature, but it is still unclear which approach will perform best on a particular dataset [8]. We have compared the performances of different inference

algorithms, for which ready-to-use software is available, by applying them to simulated data of networks with known topology under conditions that mirror a dataset involved in the initiation of lateral roots in *Arabidopsis thaliana*. We have quantified how the algorithms' performance improves with the network size and the inclusion of mutants. We have integrated sequence analysis with the intersection of approaches that performed more reliably and we have proposed a network of candidate interactions for genes potentially involved in lateral root initiation. Some of the inferred interactions capture the known regulatory activity of early components of the transcriptional cascade triggering the development of a new organ and the nodes with a high out-degree may warrant further investigation.

Parameter	Value	Parameter	Value
Number of nodes	113	Connected components	1
Avg. number of neighbours	7.363	Network density	0.06
Characteristic path length	3.482	Shortest paths percentage	89%

TABLE 3: Table summarising the topological properties of the network presented in Figure 2.

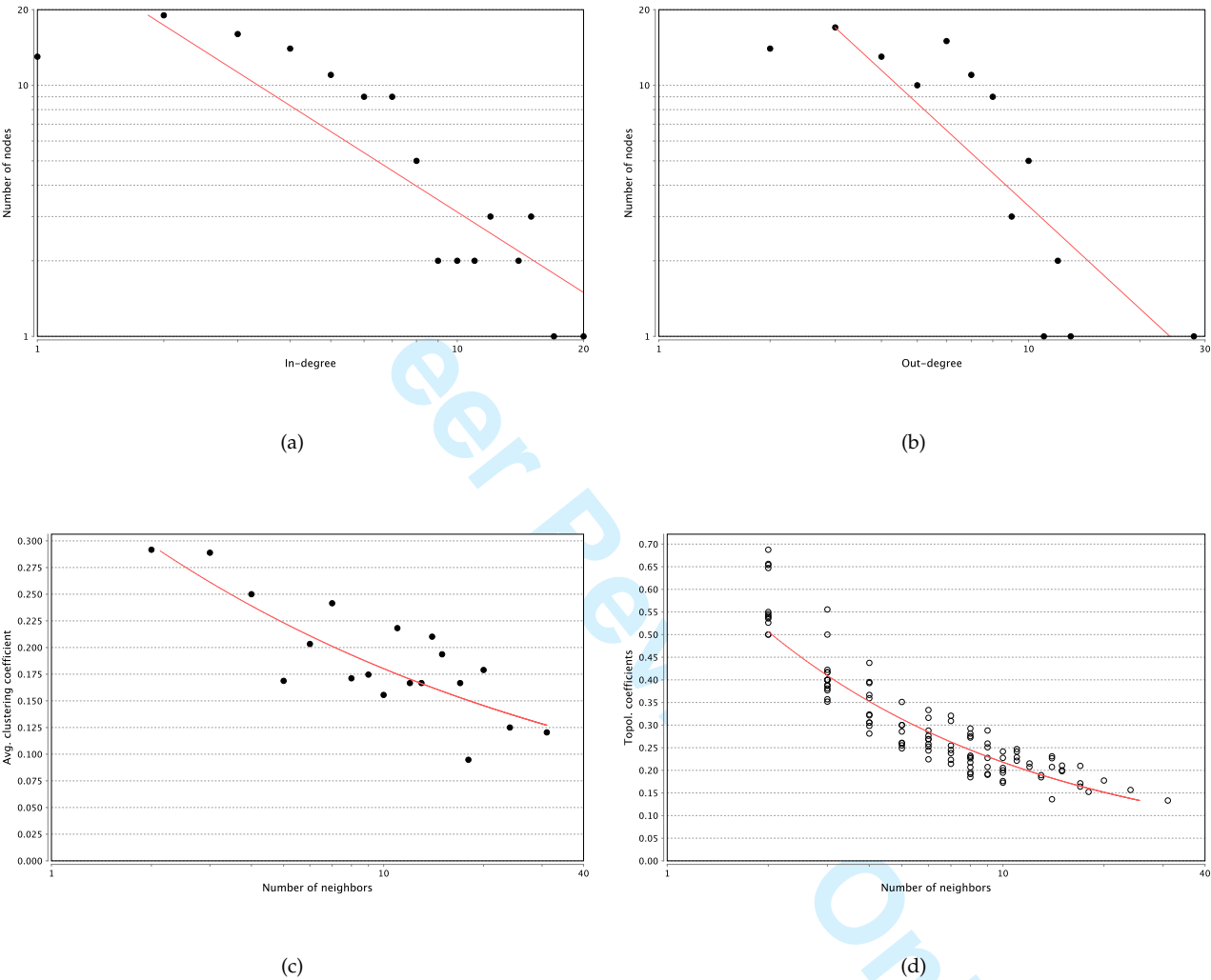


Fig. 3: Topological properties of the network inferred from the set of 212 genes. In Figures 3a and 3b the in-degree and out-degree distributions are fitted by the power-law function $y = ax^b$ with $a = 36.3, b = -1.1$ in 3a (R-squared value: 0.75) and with $a = 75.4, b = -1.4$ in 3b (R-squared value: 0.70), this showing a scale-free topology. In Figure 3c, the average clustering coefficient follows the power-law distribution $y = ax^b$ with $a = 0.37, b = -0.31$ (R-squared value: 0.60) suggesting a hierarchical organisation in the network. In Figure 3d, the topological coefficients shows a gradual decrease when increasing the number of neighbours following the power-law $y = ax^b$ with $a = 0.73, b = -0.52$ (R-squared value: 0.85) indicating a modular network organisation.

The network inferred from genes significantly expressed during such process presents topological properties that are typical of biological networks, such as scale-free, small world and hierarchical properties. We conclude that integrating reverse-engineering methods with the analysis of genome sequences may provide valuable preliminary information for further experimen-

tal validation of the genetic network promoting lateral root initiation.

ACKNOWLEDGMENTS

We gratefully acknowledge the Biotechnology and Biological Research Council and the Engineering and Sciences Research Council for financial support as part of

the CISB programme award to CPIB. The work of H. Byrne was supported in part by Award No. KUK-013-04, made by King Abdullah University of Science and Technology (KAUST). I. De Smet is supported by a BB-SRC David Phillips Fellowship (BB_BB/H022457/1) and a Marie Curie European Reintegration Grant (PERG06-GA-2009-256354). J.R. King gratefully acknowledges the funding of the Royal Society and Wolfson Foundation. The authors also thank Kim Kenobi for helpful comments.

REFERENCES

- [1] S. Smith et al. Root system architecture: insights from Arabidopsis and cereal crops. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 5;367(1595): 1441-52, 2012.
- [2] I. De Smet. Lateral root initiation: one step at a time. *New Phytologist*, 193(4): 867-73, 2012.
- [3] B. Péret et al. Arabidopsis lateral root development: an emerging story. *Trends in Plant Science*, Volume 14, Issue 7, 399-408, 2009.
- [4] B. Parizot et al. VisualRTEC: A New View on Lateral Root Initiation by Combining Specific Transcriptome Data Sets. *Plant Physiology*, 153:34-40 (2010).
- [5] M. Bansal et al. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3:78, EMBO and Nature Publishing Group 2007.
- [6] G. Stolovitzky et al. Dialogue on reverse-engineering assessment and methods: The dream of high-throughput pathway inference. *Annals of the New York Academy of Sciences* 1115:122, (2007).
- [7] G. Stolovitzky et al. Lessons from the DREAM2 challenges. *Annals of the New York Academy of Sciences* 1158:159195, (2009).
- [8] D. Marbach et al. Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, vol. 107, no. 14, 6286-6291, April 6, 2010.
- [9] C. Kaleta et al. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. *BMC Systems Biology*, 2010, 4:116.
- [10] J. Ruan et al. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biology*, 4:8, 2010.
- [11] A. Presson. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Systems Biology*, 2:95, 2008.
- [12] V. van Noort et al. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, 5(3):280-4, 2004.
- [13] S. Carter et al. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20:2242-50, 2004.
- [14] C.E. Shannon et al. The mathematical theory of communication. *University of Illinois Press*, Urbana, Illinois, 1949.
- [15] T.M. Cover et al. Elements of information theory. *John Wiley & Sons*, New York, NY, 1991.
- [16] A.A. Margolin et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* 2006, 7(Suppl 1):S7.
- [17] J. J. Faith et al. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology*, Vol. 5, Issue 1, e8, January 2007.
- [18] Watkinson et al. Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information, *The Challenges of Systems Biology* Ann. N.Y. Acad. Sci. 1158: 302-313, 2009.
- [19] A. Madar et al. DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator. *PLoS ONE*, Vol. 5, Issue 3, March 2010.
- [20] D. Anastassiou Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3:83, EMBO and Nature Publishing Group 2007.
- [21] M. Bansal et al. Inference of gene networks from temporal gene expression profiles. *IET Syst. Biol.*, 1, (5), 306-312, 2007.
- [22] Yip et al. Improved Reconstruction of In Silico Gene Regulatory Networks by Integrating Knockout and Perturbation Data. *PLoS ONE* Vol. 5, Issue 1, January 2010.
- [23] Vu et al. Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of *Saccharomyces cerevisiae*. *Nucleic Acids Research* Vol. 35, no. 1, 279-287, 2006.
- [24] T. Äijö et al. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, Vol. 25, no. 22, 2937-2944, 2009.
- [25] J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann*, 1988.
- [26] N. Friedman et al. Learning the structure of dynamic probabilistic networks. *Proceedings of the uncertainty in AI (UAI98)*, Morgan Kaufman.
- [27] Liang et al. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing* 3:18-29 (1998).
- [28] K. Murphy et al. Modelling gene expression data using dynamic Bayesian networks. *Technical Report*, University of California, Berkeley, 1999.
- [29] K. Murphy. Bayes Net Toolbox for Matlab. <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>, 2007.
- [30] Zou et al. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* Vol. 21, no. 1, 2005.
- [31] Needham et al. From gene expression to gene regulatory networks in Arabidopsis thaliana. *BMC Systems Biology*, 3:85, 2009.
- [32] P. Li et al. Comparison of probabilistic Boolean network and dynamic Bayesian approaches for inferring gene regulatory networks. *BMC Bioinformatics* 8(Suppl 7):S13, 2007.
- [33] M.J. Beal. Variational Algorithms for Approximate Bayesian Inference. *PhD. Thesis*, Gatsby Computational Neuroscience Unit, University College London, 2003, (281 pages).
- [34] M.J. Beal et al. A Bayesian Approach to Reconstructing Genetic Regulatory Networks with Hidden Factors. *Bioinformatics* 21:349-356, 2005.
- [35] C. Fogelberg et al. GreenSim: A Genetic Regulatory Network Simulator. *Technical report n. PRG-RR-08-07*, Computing Laboratory, Oxford University, url = <http://syntilect.com/cgf/pubs/greensimtr>.
- [36] D.G. Altman et al. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 308 (6943): 1552, 1994.
- [37] D.G. Altman et al. Diagnostic tests 2: Predictive values. *BMJ* 309 (6947): 102, 1994.
- [38] I. Cantone et al. A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, 137, 172-181, 2009.
- [39] I. De Smet et al. Auxin-dependent regulation of lateral root positioning in the basal meristem of Arabidopsis. *Development*, 134(4):681-90, 2007.
- [40] M.A. Moreno-Risueno et al. Oscillating gene expression determines competence for periodic Arabidopsis root branching. *Science*, 329(5997):1306-11, 2010.
- [41] Okushima et al. Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in Arabidopsis thaliana: unique and overlapping functions of ARF7 and ARF19. *Plant Cell* 17: 444463.
- [42] Vanneste et al. Cell cycle progression in the pericycle is not sufficient for SOLITARY ROOT/ IAA14-mediated lateral root initiation in Arabidopsis thaliana. *Plant Cell* 17: 30353050.
- [43] O. Bembom et al. Supervised detection of conserved motifs in DNA sequences with cosmo. *Stat. Appl. Genet. Mol. Biol.* 6: Article8.
- [44] Y. Okushima et al. ARF7 and ARF19 Regulate Lateral Root Formation via Direct Activation of LBD/ASL Genes in Arabidopsis. *The Plant Cell* January 2007 vol. 19 no. 1 118-130.
- [45] B. De Rybel et al. A Novel Aux/IAA28 Signaling Cascade Activates GATA23-Dependent Specification of Lateral Root Founder Cell Identity. *Current Biology* 20, 16971706, October 12, 2010.
- [46] T. Ulmasov et al. Dimerization and DNA binding of auxin response factors. *The Plant Journal*, 19(3), 309-319, 1999.
- [47] M.G. Ivanchenko et al. Ethylene-auxin interactions regulate lateral root initiation and emergence in Arabidopsis thaliana. *The Plant Journal*, 55(2):335-47, 2008.
- [48] F. Vandenbussche et al. Ethylene and auxin control the Arabidopsis response to decreased light intensity. *Plant Physiology*, 133(2):517-27, 2003.
- [49] T. Werner et al. Cytokinin-deficient transgenic Arabidopsis plants show multiple developmental alterations indicating opposite func-

tions of cytokinins in the regulation of shoot and root meristem activity. *Plant Cell*, 15(11):2532-50, 2003.

[50] L. Laplace et al. Cytokinins Act Directly on Lateral Root Founder Cells to Inhibit Root Initiation *The Plant Cell*, 19:3889-3900, 2007.

[51] Y. Assenov et al. Computing topological parameters of biological networks. *Bioinformatics*, Vol. 24 no. 2, pages 282-284, 2008.

[52] U. Sengupta et al. Expression-Based Network Biology Identifies Alteration in Key Regulatory Pathways of Type 2 Diabetes and Associated Risk/Complications. *PLoS ONE* 4(12): e8100. doi:10.1371/journal.pone.0008100.

[53] M.R. Carlson et al. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7: 40, 2006.

[54] D.J. Watts et al. Collective dynamics of small-world networks. *Nature*, 393, 440-442, 1998.

[55] E. Ravasz et al. Hierarchical organization of modularity in metabolic networks. *Science*, 297 1551-1555, 2002.

[56] F.A. Ditengou et al. Mechanical induction of lateral root initiation in *Arabidopsis thaliana*. *PNAS*, vol. 105 no. 48 18818-18823, 2008.

Supplementary Tables

N. GENES	REVEAL		GP		RANDOM
	PPV	Se	PPV	Se	PPV
5 undirected	0.75	0.6	1.0	0.2	0.5
5 directed	0.25	0.2	0.5	0.2	0.25
10 undirected	0.83	0.28	0.4	0.1	0.4
10 directed	0.33	0.11	0.0	0.0	0.2
20 undirected	0.2	0.1	0.38	0.1	0.16
20 directed	0.13	0.06	0.0	0.0	0.08
50 undirected			0.19	0.09	0.1
50 directed			0.02	0.01	0.05

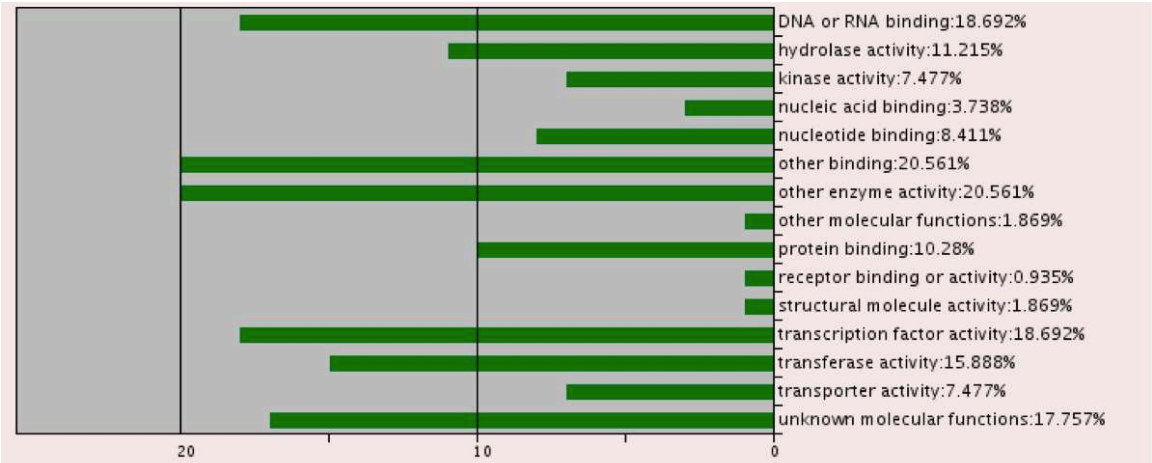
Table 1 – Table summarising the Positive Predictive Values (PPV) and Sensitivity (Se) of REVEAL and GP. In bold we represent p -values minor than 0.1. Simulated time series with 20 replicates.

N. GENES	REVEAL		GP		RANDOM
	PPV	Se	PPV	Se	PPV
10 undirected	0.86	0.33	0.5	0.17	0.4
10 directed	0.43	0.17	0.17	0.06	0.2
20 undirected	0.14	0.07	0.29	0.07	0.16
20 directed	0.14	0.07	0.0	0.0	0.08
50 undirected			0.29	0.13	0.1
50 directed			0.14	0.07	0.05

Table 2 – Table summarising the Positive Predictive Values (PPV) and Sensitivity (Se) of REVEAL and GP. In bold we represent p -values minor than 0.1. Simulated time series with 20 replicates and 10% of mutants.

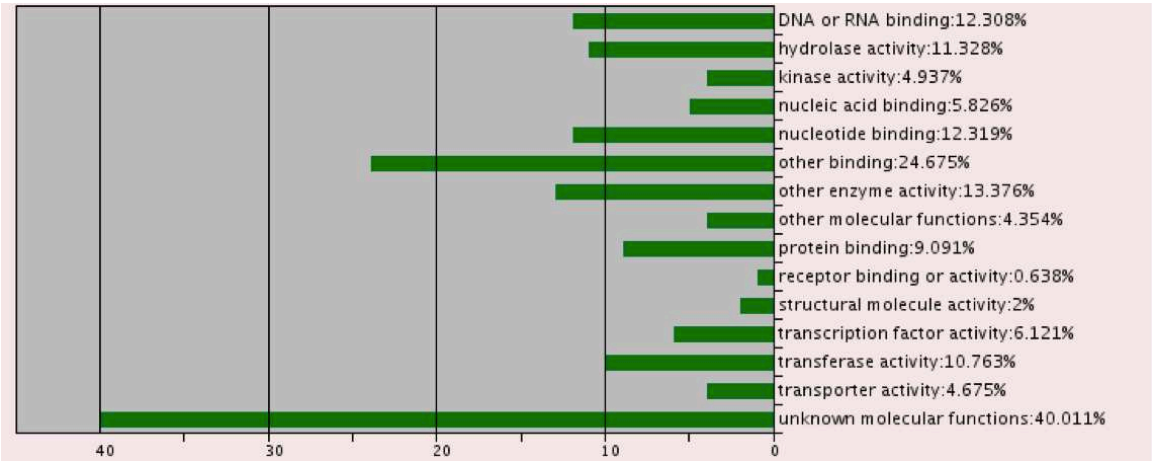
Supplementary Figures

Molecular Function



(a)

Molecular Function



(b)

Figure 1 – Bar charts representing the functional categorisation by molecular function of the set of 113 genes involved in lateral root initiation (Supplementary Figure 1a) and of the genome (Supplementary Figure 1b). The estimates were obtained by applying the functional categorisation tool of The Arabidopsis Information Resource (TAIR) [1].

References

- [1] P. Lamesch et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 2011 doi: 10.1093/nar/gkr1090.

RECENT REPORTS

12/74	Static and dynamic stability results for a class of three-dimensional configurations of Kirchhoff elastic rods	Majumdar Goriely
12/75	Error estimation and adaptivity for incompressible, nonlinear (hyper)elasticity	Whiteley Tavener
12/76	A note on heat and mass transfer from a sphere in Stokes flow at low Péclet number	Bell Byrne Whiteley Waters
12/77	Effect of disjoining pressure in a thin film equation with non-uniform forcing	Moulton Lega
12/78	A Review of Mathematical Models for the Formation of Vascular Networks	Scianna Bell Preziosi
12/79	Fast and Accurate Computation of Gauss-Legendre and Gauss-Jacobi Quadrature Nodes and Weights	Hale Townsend
12/80	On the spectral distribution of kernel matrices related to radial basis functions	Wathen Zhu
12/81	Inner product computation for sparse iterative solvers on distributed supercomputer	Zhu Gu Liu
12/82	A new pathway for the re-equilibration of micellar surfactant solutions	Griffiths Breward Colegate Dellar Howell Bain
12/83	Object-Oriented Paradigms for Modelling Vascular Tumour Growth: a Case Study	Connor Cooper Byrne Maini McKeever
12/84	Chaste: an open source C++ library for computational physiology and biology	Mirams Arthurs Bernabeu Bordas Cooper Corrias Davit Dunn Fletcher Harvey Marsh Osborne Pathmanathan Pitt-Francis Southern Zemzemi

12/87	Asymptotic solutions of glass temperature profiles during steady optical fibre drawing	Taroni Breward Cummings Griffiths
12/88	The kinetics of surfactant desorption at the airsolution interface	Morgan Breward Griffiths Howell Penfold Thomas Tucker Petkov Webster
12/89	An experimental and theoretical investigation of particlewall impacts in a T-junction	Vigolo Griffiths Radl Stone
12/90	Transitions through Critical Temperatures in Nematic Liquid Crystals	Majumdar Ockendon Howell Surovyatkina
12/91	Biaxial defect cores in nematic equilibria: an asymptotic result	Majumdar Pisante Henao
12/92	The Three Sphere Swimmer in a Nonlinear Viscoelastic Medium	Curtis Gaffney
12/93	Diffusion of multiple species with excluded-volume effects	Bruna Chapman
12/94	The Mechanics of a Chain or Ring of Spherical Magnets	Hall Vella Goriely
12/95	On-Lattice Agent-based Simulation of Populations of Cells within the Open-Source Chaste Framework	Figueredo Joshi Osborne Byrne Owen
12/96	Mathematical Biomedicine and Modeling Avascular Tumor Growth	Byrne

Copies of these, and any other OCCAM reports can be obtained from:

**Oxford Centre for Collaborative Applied Mathematics
Mathematical Institute
24 - 29 St Giles'
Oxford
OX1 3LB
England
www.maths.ox.ac.uk/occam**