A BCHC genetic algorithm model of cotemporal hierarchical *Arabidopsis thaliana* gene interactions

Bree Ann LaPointe*, David J. John†, James L. Norris‡, Edward E. Allen‡, Alexandria F. Harkey‡,
Joëlle K. Muhlemann‡ and Gloria Muday‡
*Qualtrix, Salt Lake City, Utah
†Dept. of Computer Science, Wake Forest University, Winston-Salem, North Carolina, Email: djj@wfu.edu
‡Wake Forest University, Winston-Salem, North Carolina

***Abstract*—Gene interaction network models from time course gene transcript abundance data are algorithmically created using a new aggressive genetic algorithm denoted by BCHC. The BCHC algorithm rigorously integrates probabilistic hierarchical likelihood and Bayesian methodology to produce accurate posterior probabilities of interactions between genes after observance of hierarchical gene transcript abundance data. Forbidden pairwise gene relationships are incorporated into the modeling process. This gene interaction model is compared to a previous gene interaction model utilizing the same data and Bayesian likelihood, however based upon an exponentially slower, less aggressive, and less adaptive Metropolis-Hasting search algorithm. In addition for a smaller data set, our gene interaction model is compared to less rigorous non-probabilistic Lasso estimated partial correlation models which do not fully incorporate the hierarchical structure. A comparison is also made between the smallest Bayesian model and tests for edges based on a restricted non-Bayesian hierarchical technique. The BCHC algorithm performs well when the number of genes is moderately increased, both in terms of execution time and model quality.**

***Index Terms*—Bioinformatics, Biological system modeling, Computational systems biology, Genetic algorithms, Probability Bayes methods**

INTRODUCTION

Time course transcriptomic data sets provide important information on how transcriptional changes drive development, signaling, and other important biological processes. The challenge with these data sets is finding patterns in transcript abundance and relationships between genes that can be experimentally tested. This manuscript introduces a novel algorithm–based on fundamental statistical tools and a modified genetic algorithm–that produces testable biological hypotheses.

There are many different techniques for modeling non-hierarchical single replication abundance data over a sparse number of time points [13], [14]. Several of these involve strictly deterministic, non-random techniques, and thus ignore the random variabilities that commonly exist in natural systems. These include strictly algebraic techniques like [1], [19], [33] which look for mathematical associations. Others use deterministic (non-probabilistic) techniques such as Boolean modeling [22] or differential equations [4]. Non-probabilistic methods which allow for randomness include correlation or partial correlation. With sparse gene data, complete partial correlation usually cannot be determined, because the number of genes exceeds the number of time points. However, both regularized partial correlation [17], [21], and low-order partial correlation (which adjusts for one or two other variables) [5], [36] have been utilized, with some degree of success.

This current work is built on previous work by the current authors on Bayesian posterior probabilities for a single replicate (single level) [15] and for multiple replicate models, having both independent [28] and hierarchical structures [29]. In these previous studies, the search technique was based upon a Metropolis Hastings (MH) algorithm. In this paper, we use a Bayesian version of the *Cross generational elitist selection, Heterogeneous recombination, Cataclysmic mutation* algorithm, traditionally denoted as *CHC* [7]. In its essence, CHC is a type of a genetic algorithm that does not allow the *crossing* of parents that are too similar. Our version, which we denote by *BCHC*, uses Bayesian hierarchical statistical methods in the evaluation of the models. The BCHC genetic algorithm is exponentially faster, more adaptive, and more rigorous for our network setting than the MH analogue. These MH results are shown in [25] and [30], respectively. The advantage of the MH algorithm is a body of mathematical understanding of convergence. However, the disadvantage is its doubly exponential execution time as a function of the numbers of time points and genes.

Any effective algorithmic biological interaction modeling technique must: provide reliable biological information; require reasonable run time resources; and yield probabilistic interaction models that lead to testable biological hypotheses. Such techniques should scale up linearly with the size of the data set and the number of replicates. Genetic algorithms provide such scalable and reliable procedures.

Our hierarchical posterior probability estimate models, which use a genetic algorithm, is innovative and important in several significant ways. First, the use of slope parameters of multiple replicates of rigorously obtained laboratory data in an hierarchical fashion allows for multiple examinations of biological processes and accounts for similarities and differences. Many different research disciplines emphasize the important value of multilevel hierarchical modeling [9, Section 1.3]. The utilization of (hierarchical statistical) likelihood in our work embodies the well-established likelihood principle of statistics, which both Bayesian and non-Bayesian statisticians agree is fundamental for obtaining optimal inferences. Even

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

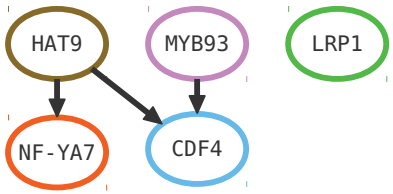


Fig. 1: A directed acyclic graph (DAG) showing 5 vertices representing genes and 3 directed edges.

for a moderate number of genes, it is impossible to search over all potential models; thus, we developed and utilized a genetic algorithm, adapted from the CHC genetic algorithm. With this new technique as well as other computational efficiencies, we can easily examine 37 genes and with adaptive fine-tuning we fully expect to scale up to an even larger set.

Overview of the Data Sets

The long term biological focus of the time course data sets examined in this manuscript is to understand the hormone signaling networks that control root architecture in the model plant species of *Arabidopsis thaliana*. This work builds on published data sets that report the time course of transcriptional response to the two hormones auxin (indole-3 acetic acid; IAA) [20] or the ethylene precursor (1-aminocyclopropane-1-carboxylic acid; ACC) [12]. The challenge with these data sets–that yield $1,246$ or 449 transcripts with consistent and reproducible changes in response to the hormone treatment, respectively–is identification of patterns of transcriptional changes that predict functional relationships that drive changes in root development.

Mathematical & Statistical Preliminaries

Two important mathematical concepts undergird this modeling approach. The directed acyclic graph $G = (V, E)$, or DAG, is the mathematical structure that provides the building blocks for the possible sets of gene interactions (i.e. networks). Fig. 1 is an example of a DAG. The vertex set V represents genes and the directed edges E are one-way relationships between pairs of vertices. Secondly, a probabilistic likelihood is used to compare two DAGs. This likelihood–and related Bayesian model averaging–were developed in [25], [28]–[30] and have been successfully used with a modified MH search for small numbers of genes and times. Essentially, each DAG D represents potential gene interactions, and the likelihood of D is an indication of the agreement of D with the data.

In general, three types of temporal relationships between vertices can be considered: *cotemporal*, *next state one-step* and *next state one-and-two-steps*. A cotemporal relationship between two vertices holds at every point in time. Traditionally, correlation and estimated partial correlation computed across all the measured data are non-rigorous measures of cotemporal relationships. A next state one-step relationship holds between the parents at time $t-1$ and child genes at time t for every pair of adjacent times. Similarly a next state one-and-two steps relationship holds between the parents at times $t-1$ and $t-2$ and the respective child at time t. Only cotemporal relationships will be considered in this manuscript.

Throughout this development, it is assumed that all the DAGs have the same prior probability; ongoing and future work will relax this uniformity assumption. The Norris-Patton likelihood (NPL) is the density function of a set of gene transcript abundance data $d_1, \ldots, d_k$ (with k replicates) being described by a DAG D, [25], [29], [30]. Both independent and hierarchical models over replicates of time course data can be accommodated by the cotemporal NPL. For this work, all the transcript data is assumed to be hierarchical; every set of transcription measurements reflects, to some extent, the underlying true biochemistry as well as the hierarchical nesting of the data. Given two DAGs, D_1 and D_2, we say that D_1 more likely reflects the gene interactions as measured in the transcript data than D_2 if and only if the posterior probability of D_1 is greater than the posterior probability of D_2. In this work, since we assume there are equal priors on the DAGs, the posterior probability for a DAG is proportional to the likelihood for the DAG [6, page 390]. Thus, it suffices to examine the DAG likelihoods.

Genetic Algorithms

A genetic algorithm is a heuristic computational procedure used to search through mathematical spaces in order to identify potential optimal results [10], [24]. Traditionally, the algorithmic search is guided by the genetic algorithm operators of *selection*, *crossover* and *mutation*. Furthermore, information about a candidate solution is provided through a *fitness function*. At time t–1, a population of a fixed number of candidates uses the selection, crossover, and mutation operators to produce, at time t, a new population of candidates.

For this research, the specialized BCHC algorithm is developed. This genetic algorithm is tuned to search through the space of candidate DAGs to determine posterior probabilities for associations between genes; its aim is to discover those DAGs which have the highest hierarchical cotemporal NPLs [18]. In the original CHC genetic algorithm, the crossover operator is extensively and exclusively applied in the production of the children from the parents. The next generation is collected from the most fit of the current children and the current generation. The mutation operator is only employed when the members of the current population become too similar, then *cataclysmic mutation* is performed which resets the class of parents based on which parents are the best fit.

The flow of the BCHC algorithm is shown in Alg. 1. The BCHC algorithm selection and crossover operators play an important role in the creation of a new population from the current population. Under the assumption of equal priors on DAGs, the fitness of a DAG D is computed as the relative NPL of D over the current population of DAGs. Selection (lines 10-11) randomly pairs candidates in the current population; every DAG in the current population is paired with another. This BCHC selection operator is significantly different from

```
1:  procedure BCHC(Bayesian-CHC)
2:      t ← 0
3:      Archive ← {}
4:      multi-step initialization of 200 DAG(s) for P(0)
5:      d ← 50
6:      while t < 250 do
7:          t ← t + 1
8:          X(t) ← P(t − 1)
9:          Y ← {}
10:         randomly reorder X
11:         for all parent pairs (X(2i), X(2i + 1)) do
12:             if parent pair are disimilar then
13:                 Y ← Y∪ crossover-repair of parent pair
14:             end if
15:         end for
16:         d ← d − (|P(t − 1)| − |Y|)
17:         P(t) ← fittest |P(t − 1)| of P(t − 1) ∪ Y
18:         if d < 0 then
19:             P(t) ← cataclysm(P(t))
20:             d ← 50
21:         end if
22:         Append new DAG(s) in P(t) to Archive
23:     end while
24:     return Archive
25: end procedure
```

Alg. 1: The algorithmic flow of the BCHC algorithm which searches the DAG space, explained by the phases of the computation. The first phase, lines 2-5, initializes variables. The next two phases, selection and crossover, lines 8-15, and cataclysmic mutation, lines 16-21, are the basis for constructing the new population of DAGs from the current population.

the selection operator in a simple genetic algorithm (SGA), where pairing is most often based exclusively on fitness. The BCHC crossover operator (line 13) allows for a pair of dissimilar DAGs to exchange genetic information. The two DAGs exchange edge connectivity information subject to a probability of crossover. If two DAGs are overly similar (i.e., the Hamming distance between them is too small) then those two DAGs are barred from exchanging information (line 12); this is a significant difference from the SGA since the next candidate class may have fewer members than the current population. The result of the crossover of two DAGs are two offspring. These offspring certainly will be directed graphs (DGs), but not necessarily DAGs.

The BCHC mutation operator (cataclysmic mutation) is applied only when the current population has reached a point when many of the candidates are very similar, based on the variable d in Alg. 1 (line 18). This mutation operator selects the most fit 5% of the candidates and then creates a new population of DAGs from these by changing the relationships between genes. Existing directed edges can be removed or reversed and currently non-existent edges can be added. As with the crossover operator, when mutation is applied to a DAG the result is a DG, but not necessarily a DAG.

A repair operator is often necessary to transform a DG into a DAG. This operator is required since both the specialized crossover and mutation operators have the potential to produce DGs which are not DAGs. The repair operator uses the Johnson-Tarjan algorithm [16] to identify the number of cycles incident with each edge. Next, an edge occurring in the largest number of cycles is removed from the DG. These edge identification and removal processes continue until no cycles exist. Intuitively, removing an edge belonging to the largest number of cycles simultaneously breaks the maximum number of cycles. This repair operator is algorithmically expensive.

Every execution of the BCHC algorithm requires certain parameters to be specified. There are rules of thumb for a reasonable assignment of parameter values; however, finding best parameter values must be done on a case by case basis. In our particular situation, the number of simultaneous executions is 20, the number of generations is 250, the initial population size is 200, and the probability of crossover is 0.3. Within a DAG, the maximum number of parents is 3. The small maximum number of parents is an extension of the concept of low-order partial correlation. When cataclysmic mutation is invoked, the top 5% of the population is used to repopulate with 200 DAGs. For this research, these genetic algorithm parameters are not varied. This allows for the study of how the number of genes affects the consistency of solutions found.

Each run of the BCHC algorithm consisted of 20 simultaneous executions. This algorithm was implemented in *python 3.0* and used the *NetworkX* package [11]. Parallel execution was implemented using the *dispy* package [31].

GENE INTERACTION MODEL

The gene interaction model is a DG whose nodes represent the genes and whose directed edges are labeled with the posterior Bayesian probability of the parent node having an edge going into the child node, as shown in Figs. 3, 5 and 6. This gene interaction model is produced from the likelihoods of the unique DAGs encountered during the BCHC search. It is essential that the BCHC algorithm visit numerous distinct DAGs during the search of the DAG space. Since we are assuming the priors of all the DAGs are equal, the relative posterior probability of a DAG is proportional to its likelihood. The gene interaction model is created from all the unique DAGs discovered during the parallel executions of the BCHC algorithm, with the highest weights going to those with the largest likelihoods. Specifically, the posterior probability $M(e)$ of a directed edge e is given by

$$M(e) = \frac{\sum_{D \in AR} \chi_D(e) L(d_1, d_2, d_3 \mid D)}{\sum_{D \in AR} L(d_1, d_2, d_3 \mid D)}$$

where AR is the Archive described in the BCHC algorithm, D is a DAG, $L(d_1, \ldots, d_k \mid D)$ is the NPL of abundance data $d_1, \ldots, d_k$, and $\chi_D(e) = 1$ if and only if e is a directed edge in D, otherwise $\chi_D(e) = 0$. This is classical Bayesian model averaging under equal priors [13].

Arabidopsis thaliana DATA SETS

All three sets of the *Arabidopsis thaliana* transcript abundance data were collected in a single laboratory and were

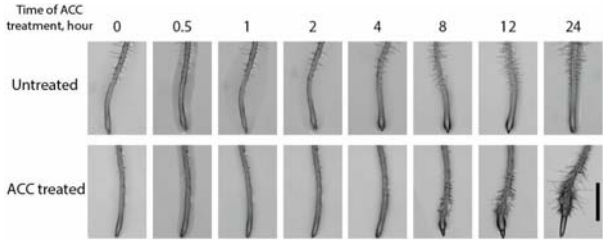


Fig. 2: Micrographs of *Arabidopsis thaliana* roots illustrate the induction of root hair formation after treatment with the plant hormone ethylene (ACC). This image is modified from [12, Figure 1] (Copyright American Society of Plant Biologists, *www.plantphysiol.org*)

previously published [12], [20]. Each set of transcript abundance measurements is derived from *Arabidopsis thaliana* genes measured across 8 time points: 0, 0.5, 1, 2, 4, 8, 12 and 24 hours after treatment by either the plant hormone auxin (IAA) or the ethylene precursor (ACC). These time points were selected as they can be overlaid on root developmental responses, including auxin-increased initiation of lateral roots [20] or ethylene induction of root hairs [12], as shown in Fig. 2. For these two data sets, transcript abundance was measured in triplicate using an Affymetrix microarray. This manuscript focuses on analysis of three subgroups of transcripts from these two data sets, each with increasing size complexity: IAA12, ACC26, and IAA37, with the abbreviation indicating the dataset and the number indicating the number of transcripts used for modeling.

The IAA12 transcript abundance data consists of 12 transcripts which were previously selected and used in Bayesian MH modeling [29] from the set of $1,246$ *Arabidopsis thaliana* genes that reproducibly respond to IAA treatment [20]. The *Arabidopsis thaliana* genes, IAA1, IAA2, IAA3, IAA4, IAA16 and IAA19, are transcriptional repressors from the AUX/IAA family. The genes ARF19 and WRKY23 are transcriptional activators. XLG1 is a G-protein involved in root morphogenesis. The PINOID kinase regulates auxin dependent root growth and development and cyclin CYCB2 participates in the control of the cell cycle by auxin.

The ACC26 transcript abundance data set consists of 26 genes measured across the same 8 time points as the IAA12 data set, in a parallel treatment with the ethylene precursor, ACC [12]. This microarray identified 449 transcripts with differential expression. This set of 26 transcripts is the complete set of transcription factors (TFs) whose transcript abundance changes after ACC treatment passed our rigorous filtering approach. TFs are proteins that bind to DNA and turn on expression of genes making RNA transcripts and the set of ACC-responsive TFs define a gene regulatory network that functions to control developmental changes. Unlike the IAA12 data set, which can be compared to previous MH based models and partial correlation, there is no MH model available for comparison. No MH model is included since the MH algorithm requires too much execution time for 26 genes. To further refine the modeling, biological prior knowledge was integrated. For the ACC26, forbidden relationships were based on experimental data on transcription factor binding partners from a method called DAP-Seq [27]. For TFs, for which this data was available, genes not on the list of potential targets were not allowed.

The labeled IAA37 transcript abundance data is a different subset of $1,246$ IAA responsive transcripts [20]. These IAA37 transcripts were identified as their IAA-dependent transcriptional changes are dependent on expression of the Auxin Response Factor19 (ARF19) transcription factor. ARFs are transcription factors that act as transcriptional activator or repressors by binding to auxin-responsive promoter elements [31]. The gene ARF19 plays an important role in the control of root architecture [25], [36] and its transcript abundance in roots increases rapidly after treatment of plants with the hormone auxin [20]. In an unpublished RNA-Seq experiment, the abundance of the IAA37 transcripts are no longer IAA responsive in a mutant with a defective ARF19 gene (Muhlemann and Muday, unpublished data). The selected transcripts are in two functional groups. The first is predicted to encode either transcription factors (TF), which are proteins that bind to DNA and turn on expression of genes making RNA transcripts. The second group are enzymes that can remodel the cell wall (CW) to allow IAA-mediated developmental changes. These transcripts are labeled TF or CW, to denote the functional group to which they belong.

For the IAA37 dataset, forbidden edges were identified based on the fact that the gene ARF19 can never be a child, while cell wall remodeling genes can only be children of transcription factors and are never parent nodes.

Cotemporal Gene Interaction Models and Comparisons

For each of the three sets of *Arabidopsis thaliana* transcript abundance data, the BCHC algorithm produces an hierarchical cotemporal interaction model. The hierarchical cotemporal NPL used the three replicates of transcript abundance measurements across 8 time points. All three cotemporal models were created using identical parameter settings for the BCHC algorithm. Also, for all three the distribution of the priors on the DAGs was uniform. However, for both the ACC26 and IAA37 data sets, additional edge information, relating to the forbidden edges, was incorporated into the execution parameters of the BCHC algorithm.

A cotemporal gene interaction model of the IAA12 genes is shown in Fig. 3. The edges of the cotemporal model are labeled with $a+b$. The a represents the posterior probability following the directed edge and the b represents the posterior probability of the reversed directed edge. The cotemporal relations should be modeled using an undirected graph; however, viewing each undirected edge as two directed edges provides some insight into the possible biological relationship between the corresponding pair of genes. The actual cotemporal probability is $a+b$. Some of the values of $a+b$ displayed in Fig.

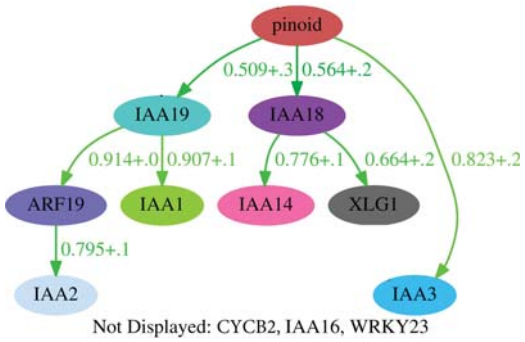


Fig. 3: The Bayesian hierarchical cotemporal model from the BCHC algorithm of the IAA12 genes where the edge posterior probabilities exceed 0.5. There are two posterior probabilities associated with each directed edge, the first is in the direction of the edge, and the second is in the opposite direction. The sum is the posterior probability of the edge.

Edge	MH	BCHC	Alt. BCHC path
IAA16-WRKY23	0.975		
ARF19-IAA2	0.913	0.895	
IAA3-PINOID	0.904	1.0	
ARF19-IAA19	0.895	0.914	
IAA18-IAA14	0.849	0.876	
IAA1-IAA19	0.784	1.0	
IAA16-IAA2	0.675		
IAA18-XLG1	0.620	0.864	
IAA14-XLG1	0.576		IAA14-IAA18-XLG1
CYCB2-IAA1	0.544		CYCB2-IAA19-IAA1
PINOID-IAA19	0.525	0.809	
CYCB2-PINOID	0.505	0.425	

Tab. 1: The Bayesian cotemporal posterior probabilities from the MH based technique [30] and those based upon the BCHC algorithm are shown. As well, for two MH edges for which no BCHC edge is shown, IAA14-XLG1 and CYCB2-IAA1, a BCHC path of length two is shown. The three bolded edges also have estimated partial correlation values that exceed 0.45.

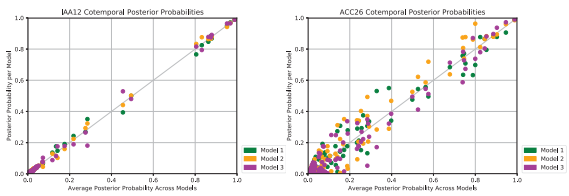
3 exceed 1.0; this is due to numerical roundoff. The total execution time to create each IAA12 model is about 5 minutes.

Fig. 3 strongly suggests, with posterior edge probability greater than 0.9, that, individually, the following pairs of genes have cotemporal relations: IAA19 with ARF19, IAA19 with IAA1, ARF19 with IAA2, and PINOID with IAA3.

The computational consistency of three hierarchical cotemporal IAA12 models is shown in Fig. 4a. Independently, three hierarchical cotemporal gene interaction models were constructed using the BCHC algorithm from the same data. In Fig. 4a, the individual model's posterior probabilities are plotted against their averages. In the perfect world, all three values would match their average. It is seen that there is high agreement in all three cotemporal models. High agreement does not ensure that the models are biologically correct, but it does demonstrate the consistency of our rigorous complex network development.

Comparison to MH Cotemporal Models

In a previous study of this IAA12 data set of *Arabidopsis thaliana* genes, a MH search approach found hierarchical cotemporal edge posterior probabilities [30, C_H column of Table 7]. The total execution time for the MH approach was 3.5 weeks. For these 12 MH edges, the hierarchical cotemporal edge posterior probabilities for the respective MH and BCHC studies are shown in Tab. 1. Of these 12 edges, 7 are also BCHC edges with posterior probability greater than 0.5, and 2 MH edges are cotemporally similar to short BCHC paths.



(a) IAA12 (b) ACC26

Fig. 4: Consistency plots for IAA12 and ACC26

Comparison to Partial Correlation Models

The IAA12 BCHC edges were also compared with estimated partial correlations for the same data set. All non-trivial adaptive Lasso estimates of the partial correlations of the three replicates are found in [30, C_H column of Table 8]. From Tab. 1, the cotemporal edges with absolute partial correlation exceeding 0.45 for at least two of the three replicates are ARF19-IAA2, ARF19-IAA19 and IAA18-XLG1. The MH-based and BCHC-based algorithms also predicted these edges.

Comparison to SAS's PROC MIXED

Furthermore, the IAA12 BCHC edges were compared to non-Bayesian hierarchical mixed models, namely those utilized in the Statistical Analysis System (SAS) PROC MIXED procedure [23]. Like the BCHC procedure, PROC MIXED allows for hierarchical slopes from replicate to replicate, but does not utilize a fundamental multiple child and parent network base. Many separate independent executions for each possible child would be required by PROC MIXED to derive a reasonable semblance of a gene interaction network.

For the IAA12 data set, comparisons between the BCHC model of Fig. 3 with the PROC MIXED model were made. First, separately for each of the eight directed edges of Fig. 3, PROC MIXED was applied. A fixed p-value for the predictor was less than 0.05 for five of these pairs, matching our Bayesian claim of an edge. Two of the remaining edges had a fixed p-value greater than 0.05, thus non-matching. For the last edge, the PROC MIXED iterative algorithm did not converge, resulting in no PROC MIXED conclusion. Then 25 one-predictor models were computed, where each did not occur in the IAA12 BCHC model in Fig. 3. Of these, 13 were also not claimed to have a significant relationship under PROC

MIXED, 2 were claimed by PROC MIXED, while the other 10 PROC MIXED models did not converge.

Some two predictor (parent) IAA12 models were computed using PROC MIXED. In Fig. 3 there are not two edges going into any single gene; thus, there were no BCHC claimed two predictors for a given gene. Nine PROC MIXED analyses of two predictors were conducted. For 8 of these 9, PROC MIXED did not converge; for the lone converging execution, one predictor was claimed by PROC MIXED while the other was not. For the other set of two predictor executions, one of the predictors was also claimed by our BCHC model, while the other was not. Eight such analyses were made. Of the 6 executions that converged, in 7 of the 12 (= 6∗2 predictors per run) executions there was agreement between the two models whether or not there was an edge.

Based on the 50 total PROC MIXED analyses and the 74 edge comparisons explored, the most consistent result was that when a directed edge was not indicated in Fig. 3, then either the PROC MIXED would not converge or would not claim the edge. There are 12 genes associated with the IAA12 data set, which yields a total of $\binom{12}{2} = 66$ possible edges. We compared a non-trivial fraction of Bayesian claimed edges and non-edges, with their corresponding PROC MIXED one and two predictor models. However, for the ACC26 and IAA37 data sets, with substantially larger number of total genes, this PROC MIXED analysis would be extremely labor intensive.

Biological Analysis of the IAA12 Models

Two cotemporal relationships predicted by both the BCHC and MH modeling are particularly statistically and biological well supported. ARFs and IAA proteins function in modules of interacting proteins to control auxin-dependent developmental processes, with ARFs acting as transcriptional regulators and IAA proteins binding to ARFs to block their function. ARF19 and several of the IAA proteins illustrated in Fig. 3 have largely overlapping expression patterns across different root tissues, suggesting that these proteins control similar auxin-dependent developmental processes in root tissues [2],

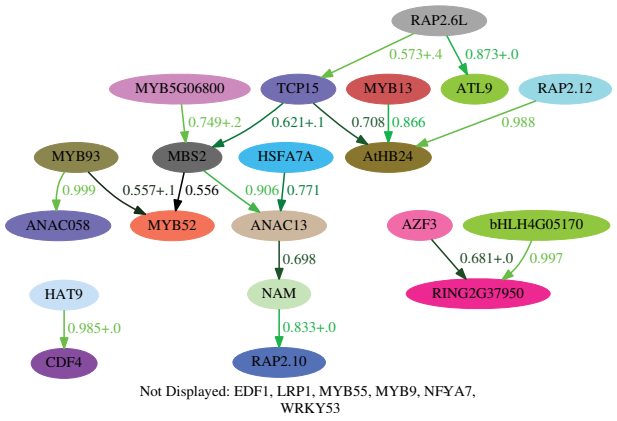


Fig. 5: A cotemporal model of the ACC26 genes showing edges whose cotemporal posterior probability exceeds 0.5.

[3]. In addition to having similar localization, ARF19 and multiple IAA proteins, including IAA2 and IAA19, were shown to interact in yeast-two hybrid assays, indicating that these proteins function together to regulate auxin-dependent transcriptional networks [8], [35]. Of particular interest is the cotemporal nature ARF19 and IAA19 transcript abundance after IAA treatment (Fig. 3). Mutants with a defect in ARF19 and with stabilized IAA19 (in a mutant named Massugu2 or MSG2) have similar defects in lateral root development [26], [34]. ARF19 was shown to be an activator of IAA19 in a yeast synthetic biology model for measuring ARF transcription factor function [32]. Examination of a mutant that does not make ARF19 showed that auxin-dependent IAA19 expression is regulated by ARF19, as well as ARF7 [37].

ACC26 Data Set Models

The ACC26 data set incorporates gene transcript abundance information on 26 *Arabidopsis thaliana* genes, more than double the number of genes in the IAA12 data set. The MH approach cannot be applied to this data set, due to prohibitive execution time. Since the number of genes is so much larger than the number of time points, partial correlation estimates would be even more unreliable. The hierarchical cotemporal ACC26 models were created by the BCHC algorithm in about 20 minutes. Overall the posterior probabilities shown in Fig. 5 are less than those in the IAA12 models, reflecting the increase in the number of genes. However, with posterior probabilities as high as 0.997, the ACC26 cotemporal model inspires confidence for some edges. The ACC26 forbidden edges do change the labeling of the directed edges in Fig. 5. An edge is labeled with a single posterior probability, such as the edge from NAM to RAP2.10, when the reverse edge is one of the forbidden edges. As seen in Fig. 4b, three independently generated hierarchical cotemporal models for ACC26 are generally in agreement but less so than the IAA12 models. Recall that the 3 ACC26 data sets were independently run with the BCHC parameters tuned for IAA12. With additional parameter tuning for ACC26 data sets, it is likely the BCHC genetic algorithm will have improved performance.

Interestingly, an ACC26 interaction model was compared to an ACC26 model with no forbidden edges. The posterior probabilities of the two models had a correlation of 0.77. Even without the specified forbidden edges, the BCHC algorithm derived a model very similar to a regular ACC26 model.

IAA37 Data Set Models

The IAA37 data set consists of 37 genes, a significant increase from the 26 genes in ACC26, and even more so from the 12 genes of IAA12. As well, the IAA37 data set incorporates forbidden edges, chosen according to the criteria described above. A hierarchical cotemporal model produced by the BCHC algorithm is shown in Fig. 6. As with the ACC26 model, the edge labels indicate the presence or absence of the forbidden edges. There are a number of edges labeled with posterior probabilities larger than 0.9. The consistency plot for IAA37 (not shown) shows considerable variation across the

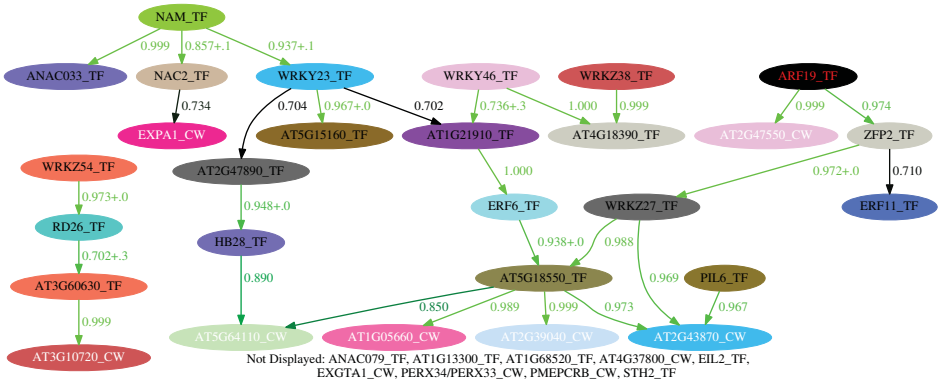


Fig. 6: A cotemporal model of the IAA37 genes. For those edges labeled with exactly one posterior probability, the reverse edge was forbidden.

three IAA37 models. Re-tuning the BCHC parameters should significantly improve the performance of BCHC on IAA37.

Conclusions

The exploration of the DAG space by the BCHC algorithm has been successful from several points of view. First, the BCHC algorithm executes much faster than the Metropolis-Hastings approach previously developed. For 12 genes, BCHC runs in minutes instead of the weeks required for MH. For 26 genes, BCHC executes in 20 minutes, and execution of MH is not feasible. It also aggressively explores many different regions of the DAG space in search of diverse edges and high posterior probability edges. Second, as the size of the measured data increases (increased number of genes, additional replicates, additional time points) the execution time of the BCHC algorithm scales up polynomially. The three gene interaction models for IAA37 show the variation across the models has increased significantly with the number of genes. Thus, the adjusting of the BCHC parameters for data sets based on larger numbers of genes is mandatory future work. The execution time, however, remains reasonable, and the memory demands are also manageable. Third, the BCHC algorithm operators engineered to work with DAGs accomplish their tasks computationally efficiently. Certainly, more work is needed to improve the performance of these operators, especially the repair operator. However, even in their current states, the selection, crossover and mutation operators implement the DAG operations correctly.

Next Steps

As the work on this gene interaction modeling project continues, there are at least four areas of future focus: the repair operator; problems involving scaling up including BCHC parameter tuning; gene relationship guarantees and forbiddance; and, the incorporation of nonuniform prior probabilities. The use of the BCHC algorithm has now made it feasible to consider creating gene interaction models using much larger data sets. The comparisons of the gene interaction models from the BCHC algorithm to those from other techniques, when available, is most encouraging. Computational work will continue to ensure that this approach can scale up comfortably, in terms of execution time and model quality. As the biologists find that certain edge relationships must be present or cannot occur, it is important that the computational model incorporate this information. It is planned to incorporate *guaranteed* edges in addition to *forbidden* edges as is currently done in ACC26 and IAA37. Biologists may also have experimental evidence that could influence the prior probabilities. In the work presented here all the DAG priors are uniform (except for the forbidden edges associated with ACC26 and IAA37); nonuniform priors can be introduced and investigations are continuing with these. In effect, the algorithm should have the capability to learn an edge probability that may differ from a given prior probability, unless it is known to be 0, or forbidden.

The creation of next state models is important as the next state paradigms capture the case-effect relationships between pairs of genes. Both next state paradigms affect the dimension of the data available for the creation of models. For many biological experiments in which gene transcript abundance data are collected, the number of time points is small. Reducing this number presents a challenge. The parameters of the BCHC algorithm must be modified for next state modeling.

In this cotemporal modeling, all the BCHC algorithm parameters were fixed, based on what settings worked well for the ACC12 data set. However, the consistency information for ACC26 and IAA37 strongly suggest that the parameter settings should adjust to the size and other properties of the data sets. For example, the genetic algorithm population size and the number of generations should increase as a function of the number of genes. Effort must be invested in better understanding what are reasonable parameter settings for the modeling paradigms.

The BCHC algorithm has demonstrated that it has the potential to produce high quality gene interaction models in a reasonable amount of time. As the number of genes increases certainly the genetic algorithm parameters must be adjusted. It has demonstrated the flexibility to successfully handle additional connection information such as the forbidden

edges. The BCHC algorithm is a tool that is worthy for more study and development.

ACKNOWLEDGMENTS

The authors thank Kenneth Meza, a visiting research assistant from Tecnológico de Costa Rica, for his efforts in creating a prototype genetic algorithm whose chromosomes are DAGs. The authors thank Qiwen *Wendy* Gao for her running and synthesis of the SAS PROC MIXED models. Bree Ann LaPointe thanks the Wake Forest Graduate School of Arts and Sciences and the Wake Forest University Center for Molecular Signaling for supporting her as a graduate research assistant. The authors thank the National Science Foundation for their support with a grant, NSF#1716279.

REFERENCES

[1] E. E. Allen, J. S. Fetrow, L. W. Daniel, S. J. Thomas, and D. J. John. Algebraic dependency models of protein signal transduction networks from time-series data. *Journal of Theoretical Biology*, 238(2):317–330, January 2006.

[2] B. O. Bargmann, S. Vanneste, G. Krouk, T. Nawy, I. Efroni, E. Shani, G. Choe, J. Friml, D. C. Bergmann, M. Estelle, and K. D. Birnbaum. A map of cell type-specific auxin responses. *Molecular Systems Biology*, 9(688), 2013.

[3] S. Brady, D. Orlando, J. Lee, J. Wang, J. Koch, J. Dinneny, D. Mace, U. Ohler, and P. Benfey. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318(5851):801–806, 2007.

[4] J. Cao, X. Qi, and H. Zhao. Modeling gene regulation networks using ordinary differential equations. In *Next Generation Microarray Bioinformatics*, volume 802 of *Methods in Molecular Biology*, pages 185–197. Springer, 2012.

[5] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[6] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, 4th edition, 2012.

[7] L. J. Eshelman. Genetic algorithms. In T. Bäck, D. B. Fogel, and T. Michalewicz, editors, *Evolutionary Computation 1 - Basic Algorithms and Operators*, volume 1, chapter 8, pages 64–80. Institute of Physics Publishing, 2000.

[8] H. Fukaki, Y. Nakao, Y. Okushima, A. Theologis, and M. Tasaka. Tissue-specific expression of stabilized SOLITARY-ROOT/IAA214 alters lateral root development in arabidopsis. *Plant J.*, 44(3):382–395, 2005.

[9] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[10] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.

[11] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[12] A. F. Harkey, J. M. Watkins, A. L. Olex, K. T. DiNapoli, D. R. Lewis, J. S. Fetrow, B. M. Binder, and G. K. Muday. Identification of transcriptional and receptor networks that control root responses to ethylene. *Plant Physiology*, 176(3):2095–2118, 2018.

[13] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E.I. George, and a rejoinder by the authors). *Statistical Science*, 14(4):382–417, 1999.

[14] P. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[15] D. J. John, J. S. Fetrow, and J. L. Norris. Continuous cotemporal probabilistic modeling of systems biology networks from sparse data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(5):1208–1222, September/October 2011.

[16] D. B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84, 1975.

[17] N. Krämer, J. Schäfer, and A.-L. Boulesteix. Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics*, 10(384), 2009.

[18] B. A. LaPointe. Arabidopsis thaliana gene interaction exploration with CHC genetic algorithm. Master's thesis, Wake Forest University, Department of Computer Science, December 2017.

[19] R. Laubenbacher and B. Stigler. A computational algebra approach to the reverse engineering of gene regulatory networks. *Journal of Theoretical Biology*, 229(4):523–537, August 2004.

[20] D. R. Lewis, A. L. Olex, S. R. Lundy, W. H. Turkett, J. S. Fetrow, and G. K. Muday. A kinetic analysis of the Auxin transcriptome reveals cell wall remodeling proteins that modulate lateral root development in *Arabidopsis*. *The Plant Cell*, 25:3329–3346, September 2013.

[21] H. Li and J. Gai. Gradient directed regularization for sparse Gaussian, concentration graphs with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317, 2008.

[22] J. Liang and J. Han. Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. *BMC Systems Biology*, 6(113):1–20, 2012.

[23] SAS Institute Inc. *SAS/STAT(R) 14.1 User's Guide*. SAS Institute Inc., Cary, North Carolina, 2015.

[24] M. Mitchell. Genetic algorithms: An overview. *Complexity*, 1(1):31–39, 1995.

[25] J. Norris, K. Patton, S. Huang, D. John, and G. Muday. First and second order markov posterior probabilities on multiple time-course data sets. In *SoutheastCon 2015*, pages 1–8, Norfolk, Virginia, April 2015. IEEE.

[26] Y. Okushima, P. J. Overvoorde, K. Arima, J. M. Alonso, A. Chan, C. Chang, J. R. Ecker, B. Hughes, A. Lui, D. Nguyen, C. Onodera, H. Quach, A. Smith, G. Yu, and A. Theologis. Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in Arabidopsis thaliana: Unique and overlapping functions of ARF7 and ARF19. *Plant Cell*, 17:444–463, 2005.

[27] R. C. O'Malley, S. C. Huang, L. Song, M. G. Lewsey, A. Barlett, J. R. Nery, M. Galli, A. Gallavotti, and J. R. Ecker. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165(5):1280–1292, May 2016.

[28] K. L. Patton, D. J. John, and J. L. Norris. Bayesian probabilistic network modeling from multiple independent replicates. *BMC Bioinformatics*, 13(Supplement 9):1–13, June 2012.

[29] K. L. Patton, D. J. John, J. L. Norris, D. Lewis, and G. Muday. Hierarchical Bayesian system network modeling of multiple related replicates. *BMC Bioinformatics*, 7:803–812, 2013.

[30] K. L. Patton, D. J. John, J. L. Norris, D. R. Lewis, and G. K. Muday. Hierarchical probabilistic interaction modeling for multiple gene expression replicates. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(2):336–346, March/April 2014.

[31] G. Pemmasani. dispy: Distributed and parallel computing with/for python. http://dispy.sourceforge.net, 2016.

[32] E. Pierre-Jerome, B. L. Moss, A. Lanctot, A. Hageman, and J. L. Nemhauser. Functional analysis of molecular interactions in synthetic auxin response circuits. *PNAS*, 113:11354–11359, 2016.

[33] B. Stigler. Polynomial dynamical systems in system biology. *2006 AMS Proceedings of Symposia in Applied Mathematics*, 64:59–84, 2007.

[34] K. Tatematsu, S. Kumagai, H. Muto, A. Sato, M. K. Watahiki, R. M. Harper, E. Liscum, and K. T. Yamamoto. MASSUGU2 encodes Aux/IAA19, an auxin-regulated protein that functions together with the transcriptional activator NPH4/ARF7 to regulate differential growth responses of hypocotyl and formation of lateral roots in *Arabidopsis thaliana*. *The Plant Cell*, 16(2):379–393, February 2004.

[35] T. Vernoux, G. Brunoud, E. Farcot, V. Morin, H. V. den Daele, J. Legrand, M. Oliva, P. Das, A. Larrieu, D. Wells, Y. Guédon, L. Armitage, F. Picard, S. Guyomarćh, C. Cellier, G. Parry, R. Koumproglou, J. H. Doonan, M. Estelle, C. Godin, S. Kepinski, M. Bennett, L. D. Veylder, and J. Traas. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Mol Syst Biol.*, 7(508), 2011.

[36] A. Wille, P. Zimmermann, E. Vranova, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biology*, 5(92):1–13, October 2004.

[37] J. C. Wilmoth, S. Wang, S. B. Tiwari, A. D. Joshi, G. Hagen, T. J. Guilfoyle, J. M. Alonso, J. R. Ecker, and J. W. Reed. NPH4/ARF7 and ARF19 promote leaf expansion and auxin-induced lateral root formation. *The Plant Journal*, 43(1):118–130, 2005.