Computational modeling of *Caenorhabditis elegans* vulval induction

Xiaoyun Sun and Pengyu Hong*
Department of Computer Science, National Center of Behavioral Genomics, 415 South Street, Waltham, MA 02454, USA

**ABSTRACT**

**Motivation:** *Caenorhabditis elegans* vulval development is a paradigmatic example of animal organogenesis with extensive experimental data. During vulval induction, each of the six multipotent vulvar precursor cells (VPCs) commits to one of three fates (1°, 2°, 3°). The precise 1°-2°-3° formation of VPC fates is controlled by a network of intercellular signaling, intracellular signal transduction and transcriptional regulation. The construction of mathematical models for this network will enable hypothesis generation, biological mechanism discovery and system behavior analysis.

**Results:** We have developed a mathematical model based on dynamic Bayesian networks to model the biological network that governs the VPC 1°-2°-3° pattern formation process. Our model has six interconnected subnetworks corresponding to six VPCs. Each VPC subnetwork contains 20 components. The causal relationships among network components are quantitatively encoded in the structure and parameters of the model. Statistical machine learning techniques were developed to automatically learn both the structure and parameters of the model from data collected from literatures. The learned model is capable of simulating vulval induction under 36 different genetic conditions. Our model also contains a few hypothetical causal relationships between network components, and hence can serve as guidance for designing future experiments. The statistical learning nature of our methodology makes it easy to not only handle noise in data but also automatically incorporate new experimental data to refine the model.

**Contact:** hong@cs.brandeis.edu

**Supplementary information:** Supplementary data are available at http://combio.cs.brandeis.edu/vpc

1 INTRODUCTION

Cellular activities and animal developments are governed by complex biological systems. The functions and behaviors of such systems are decided by dynamically interacting components, which are coupled via inter- and intra-cellular signaling activities and transcriptional regulations. Studies over the last few decades from genetic and biochemical approaches have identified core components of many such biological systems. The wealth of information accumulated over years has made the quantitative studies of those systems possible and necessary. In this work, we chose to model the molecular network governing vulval induction in *Caenorhabditis elegans*. The *C. elegans* hermaphrodite vulva normally derives from three of the vulva precursor cells (VPCs) (named P3.p, P4.p, P5.p, P6.p, P7.p and P8.p) during the mid-third larval stage (Fig. 1A). Each of the six VPCs is multipotent capable of adopting one of three fates, termed primary (1°), secondary (2°), or tertiary (3°) (Sternberg and Horvitz 1989). The final fates of VPCs are precisely regulated by a long-range gradient epidermal growth factor (EGF) signal from the anchor cell and a cell-to-cell lateral signal mediated by the Notch-like receptor LIN-12 (Fig. 1B). EGF-receptor (EGFR) signaling promotes the primary fate and the lateral signaling promotes the secondary fate. These two signaling pathways are cooperative and antagonistic during the VPC differentiation process. EGFR signaling up-regulates the ligands for the receptor LIN-12, but at the same time inhibits the activity of LIN-12. On the other hand, LIN-12 signaling induces the inhibitors of EGFR signaling. Both EGFR signaling and LIN-12 signaling prevent VPCs from adopting fate 3°.

Sternberg and Horvitz (1989) proposed a diagrammatic model describing gene interactions during vulval induction based on the results from various mutation experiments (Fig. 1C). The diagrammatic model mainly contains the following components: Vul (genes resulting in a vulvalless phenotype if mutated), Muv (genes resulting in a multivulva phenotype if mutated), 1° (genes with 1°-specific functions), 2° (genes with 2°-specific functions), 3° (genes with 3°-specific functions), LS (lateral signaling ligands) and LIN-12 (the receptor for the lateral signal). Each component is represented as a single node in the model. The interactions between components are represented as edges with either arrows indicating activation or bars representing inhibition. This diagrammatic model is very useful for summarizing a qualitative understanding of observations and provides a foundation for later research on *C. elegans* vulval development. However, it only offers a limited view about the dynamics of the underlying system and cannot be used in quantitative reasoning, which is needed to automatically integrate new experimental data and generate hypotheses.

Based on the above diagrammatic model, Fisher et al. (2005) used statecharts, a visual language (Harel, 1987), to construct a computational dynamic model of vulval induction. This work is significant because it is the first attempt to build a mathematical model for simulating the dynamic behaviors of the molecular network governing vulval induction. Their initial model consists of three statecharts representing the anchor cell, the VPCs and the organizer used to set the initial conditions.

*To whom correspondence should be addressed.*

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
A great deal of data related to vulval induction has been published since Sternberg and Horvitz proposed their diagrammatic model in 1989. The integration of those data in a principled manner will help derive a more comprehensive model for vulval induction and the diagrammatic model [adopted from Sternberg and Horvitz (1989)]. (A) Six multipotent VPCs P3.p, P4.p, P5.p, P6.p, P7.p and P8.p, are located just ventral to the gonad. (B) Precise formation of the 1°–2°–3° pattern in the wildtype C. elegans. The graded EGF signal (black arrows) produced by the anchor cell promotes gene expression specific to fate 1°. The lateral signal (white arrows) up-regulates gene expression specific to fate 2°, including expression of genes inhibiting EGFR signaling. The thickness of arrows indicates the strength of the interaction. (C) The diagrammatic model for gene interactions during vulval induction. Two cells are shown.

Fig. 1. Vulval induction and the diagrammatic model [adopted from Sternberg and Horvitz (1989)]. (A) Six multipotent VPCs P3.p, P4.p, P5.p, P6.p, P7.p and P8.p, are located just ventral to the gonad. (B) Precise formation of the 1°–2°–3° pattern in the wildtype C. elegans. The graded EGF signal (black arrows) produced by the anchor cell promotes gene expression specific to fate 1°. The lateral signal (white arrows) up-regulates gene expression specific to fate 2°, including expression of genes inhibiting EGFR signaling. The thickness of arrows indicates the strength of the interaction. (C) The diagrammatic model for gene interactions during vulval induction. Two cells are shown.
Fig. 2. Data types. (A) Complete phenotypic data [examples from Sternberg and Horvitz (1989)]. +/- denotes with/without anchor cell (ac) or gonad; # indicates the number of observed animals with the indicated lineage. (B) Incomplete phenotypic data [examples from Berset et al. (2001)]. Average induction denotes the average number of VPCs adopting 1° or 2°. (C) Incomplete phenotypic data [examples from Simske et al. (1996)]. The vulvaless column lists the percentage of animals with vulvaless phenotype (i.e. all VPC adopt 3°). (D) Qualitative gene expression pattern. The temporal expression pattern of a gene group consisting of lst-1, lst-2, and lst-4 in different VPCs during vulval induction in wildtype C. elegans (Yoo et al., 2004). (E) Re-illustration of (D). The expressions levels of lst-1, lst-2 and lst-4 are high (=3) in all VPCs at the beginning. Upon receiving the inductive signal, the levels drop to medium (=2) in P5.p and P7.p and low (=1) in P6.p, but remain high in P3.p, P4.p, and P8.p. After a while, the levels in P5.p and P7.p increase and stabilize at high levels while remaining the same in other VPCs. The numeric levels only indicate relative strength.

for biologists to design future experiments. Finally, our mathematic model encodes stochastic control mechanisms that can explain the fate variability observed in experiments.

2 METHODS

2.1 Data

The training data was collected from Beitel et al. (1990), Berset et al. (2001), Simske et al. (1996), Sternberg and Horvitz (1989) and Yoo et al. (2004). In many experiments, genetic mutations, which include lose-of-function (lf), reduce-of-function (rf), gain-of-function (gf) or the combinations of the above, were applied to one or more genes. Some training samples contain complete phenotypic information that includes the final fates of all VPCs. Some training samples contain incomplete phenotypic information that only includes the average number of induced VPCs (i.e. adopt fates 1° or 2°) or coarse phenotype categories (i.e. vulvaless or multivulva). In total, we had about 2000 phenotypic data samples. The rest training samples are qualitative descriptions of temporal gene expression patterns which only contribute to a very small portion of the training data, however, are important for learning the dynamics of the model. Yoo et al. (2004) described the patterns of two groups of genes. One of the gene groups consists of lst-1, lst-2 and lst-4, and the other group consists of dpy-23 and let-3. The expression level of the first group is uniformly high in all six VPCs prior to vulval inductive signaling, but forms a gradient that is the inverse of the fate 1° reporter expression pattern in response to inductive signal (Fig. 2D). The expression level of the second group is very faint in all VPCs prior to the mid-L3 stage, but then become strong in P5.p and P7.p and persists in their daughters. Examples of the above data types are shown in Figure 2. The complete data set is provided in the Supplementary information.

2.2 Data preprocessing

The data was preprocessed to produce a set of rules used in the model learning procedure which will be described later in this article. The main goal of learning is to find a model that is capable of simulating vulval induction under various genetic backgrounds to match the experimental observations as precisely as possible. The goodness-of-match is defined by those rules. For complete phenotypic data, we directly used the final fate of each VPC as the desired simulation results of the model. Incomplete phenotypic data is difficult to utilize, however, quite informative. We managed to use some of such kind data based on our best knowledge in the following way. If the average number of induced VPCs is close to 1, we let P6.p adopt fate 1°/2° and other VPCs adopt fate 3°. If the average induction value is close to 3, we let P5.p and P7.p adopt fate 2°, P6.p adopt fate 1° and the rest of VPCs adopt fate 3°. The fate information was further converted into rules describing the states of fate markers. For example, if a VPC adopts fate 1°, its fate 1° marker should be the only fate marker that stabilizes at its highest level at the end of the vulval induction process. Qualitative gene expression patterns provide valuable information about the dynamics of the system. Based on the descriptions of Yoo et al. (2004), we designed several rules describing the temporal expression patterns of five genes lst-1, lst-2, lst-4, dpy-23 and let-3 during vulval induction in wildtype C. elegans. The rules only specify the up-down trends of the expression patterns (Fig. 2E), however, do not define when exactly the expression levels should change because no such accurate temporal information was available. Due to the limited space, the complete details about data preprocessing and the rules are provided in the Supplementary Materials.

2.3 Dynamic bayesian network

DBN is a form of statistical graphical model. In this application, it represents molecular components as nodes and the causal relationships between molecular components as arcs (i.e. edges between nodes). The relationships are quantitatively encoded in the parameters representing the conditional probabilities (e.g. the probability of a gene being up/down regulated given the status of other components connecting to the gene). DBN generalizes both Hidden Markov model (HMM) (Churchill 1989; Rabiner 1989) and Bayesian Network (BN)
Fig. 3. A DBN for the diagrammatic model in Figure 1C. (A) Only two cells and the first two time-slices are shown. Network components are grouped by the dashed rectangles according to their cell identity. The subscription indices denote cell IDs and time. For example, LIN-12,1,2 represents LIN-12 in cell 1 at time 2. The arrows only show the directions of relationships whose quantitative effects are encoded in the associated parameters (see Table 1 for an example). (B) A control node example. Some network components have control nodes denoting the genetic status such as wildtype or mutations (ff, of or gf). For clear illustration purpose, control nodes were hidden in (A) and only one example with its local structure is shown in (B).

(Pearl, 1988) to represent probability distributions of discrete-time stochastic processes. BN has been shown to be a powerful tool for modeling biological networks (Friedman et al., 2000; Friedman, 2004; Imoto et al., 2006; Lee et al., 2004; Pena et al., 2005; Sachs et al., 2002, 2005; Woofll et al., 2005). However, BN restricts the network structure to be a directed acyclic graph (DAG) and is not capable of dealing feedback loops. For example, 1°/C14 to be a directed acyclic graph (DAG) and is not capable of dealing feedback loops. For example, 1°/C14/C14 is modeled as an interslice arc so that the cycle diagrammatic model of Sternberg and Horvitz (1989). In this example, 1°/C14/C14 is modeled as an interslice arc so that the cycle...
training data, a control node is added to be one of its parents and if a network component was ever mutated to produce some of our components, which together decide the dynamic behaviors of the network. The inductive signal up-regulates 1. The lateral signal inhibits 1. When 2 is weak (i.e. 2 = 1), 1 is primary decided by Vul1,2. When both signals are strong (Vul1,2 = 3 and 2 = 2), they compete to control 1,2. And the chances of 1,2 being median (i.e. 2) or high (i.e. 3) are 40 and 60% respectively.

\[
\begin{array}{cccccc}
\text{Vul1,2} = 1 & \text{Vul1,2} = 2 & \text{Vul1,2} = 3 & \text{Vul1,2} = 1 & \text{Vul1,2} = 2 & \text{Vul1,2} = 3 \\
2,1,1 = 1 & 2,1,1 = 1 & 2,1,1 = 1 & 2,1,1 = 2 & 2,1,1 = 2 & 2,1,1 = 2 \\
\end{array}
\]

When both signals are strong (Vul1,2 = 3 and 2 = 2), they compete to control 1,2. And the chances of 1,2 being median (i.e. 2) or high (i.e. 3) are 40 and 60% respectively.

Figure 3A, [Vul1,2, Vul2,1, Vul1,2, Vul2,2, [1,1,1, 1,2,1] and [1,1,2, 1,2,2] are three equivalent classes.

Our training data were collected from the published experiments under various genetic manipulations (e.g. f/f, r/f, g/g). The structure as well as the dynamics of the network cannot be correctly inferred without appropriately modeling the effects of these genetic manipulations. If a network component was ever mutated to produce some of our training data, a control node is added to be one of its parents and represent its genetic status (Fig. 3B). The control nodes do not have any parents. Our DBN model uses discrete states and assumes \( P(X_t^o | \Omega X_{t-1}^o) \) to be multinomial with non-informative Dirichlet priors. \( P(X_t^o | \Omega X_{t-1}^o) \) is represented as a conditional probability table (CPT) associated with the node \( X_t^o \). Table 1 shows an exemplar CPT of the node 1,2 in Figure 3.

The details about the number of possible states for each network component in our DBN model are provided in the supplementary information.

We can simulate vulval induction by sampling our DBN. To do this, we first unroll the vulval induction DBN to have \( T \) time-slices, compute the ordering (the parents of a particular variable must appear before the variable in the ordering), assign values to observed variables, and then sequentially sample unobserved variables according to the ordering. Currently, we empirically set \( T = 40 \) so as to give enough time for the in silico induction processes to reach steady states under the effects of two collaborative while competing signaling—EGFR signaling and LIN-12/Notch signaling. The states of the fate markers in the final time slice are used to decide the fates of VPCs. For example, when simulating the vulval induction in the wildtype \( C.\ elegans \), we first set up the graded inductive signals for \( P3.p, P4.p, P5.p, P6.p, P7.p \) and \( P8.p \) as 1, 2, 3, 4, 3 and 2, respectively, which reflect the distances from VPCs to the anchor cell. Higher values denote higher inductive signal levels. It should, however, be noted that the levels only indicate the relative strength instead of the absolute strength. The inductive signal vectors remain unchanged during the whole induction process. The control nodes are all set as wildtype. The unobserved variables are then sequentially sampled based on the ordering. Upon finishing the simulation, we say that a VPC commits to a specific fate if the corresponding fate marker is the only marker stabilizes at its highest level. The fate of a VPC is set as undecided if more than one fate marker takes its highest value.

### 2.5 Learning a vulval induction DBN model

The goal of learning a vulval induction DBN model is to find one that is as simple as possible and is capable of explaining the training data and simulating vulval induction. Let \( M = \langle G, \theta \rangle \) denote the model, where \( G = \langle B_1, B_2, \ldots \rangle \) and \( \theta \) are the structure and the parameter set of the DBN respectively. The parameter set \( \theta \) includes the CPTs associated with all network components and encodes the detailed information about the causal relationships between the network components, which together decide the dynamic behaviors of the model. The training data is partially observed because none of the experiments reported the complete dynamics of all network components (e.g. the detailed temporal profiles of gene expression and protein activity during vulval induction). Hence, we need to deal with the hidden variable problem in learning the model. Let \( O \) and \( H \) denote the observed and hidden variable subset respectively. We used the Metropolis-Hastings algorithm (Hastings 1970) which starts with an initial model and then repeats the following steps until converges: (a) sample a new structure; (b) estimate the parameters; (c) estimate the likelihood of the training data given the new model; and (d) accept the new model based on the Metropolis-Hastings ratio. The overall algorithm is outlined in Figure 4 and is explained in detail as below.

The following biological knowledge is used as prior to decide the initial structure \( G_1 \). Four proteins \( \text{LET-60 (RAS)}, \text{LIN-45 (RAF)}, \text{MEK-2 (MAPKK) and MPK-2 (MAPK)} \) form a canonical linear RTK/Ras/MAP Kinase signaling cascade that is conserved across many organisms including mammals. The inductive signal LIN-3 and the lateral signal (LS) are respectively mediated by the EGF-receptor \( \text{LET-23} \) and the Notch-like receptor \( \text{LIN-12} \). Given the initial structure \( G_1 \) and the observed variables, we estimate the parameters of the model by the \( \text{EstimateParameter} \) function, which also generates a population of complete data \( \{C_i\}_{i=1}^{S} = \langle O_i, H_i \rangle_{i=1}^{S} \) as the side product, where \( S \) is the population size and \( H_i \) is the \( i \)-th sample of the hidden variable set. The likelihood \( P(O|M) \) is needed to compare different models. The computation of the likelihood requires integrating out the hidden variables, i.e. \( \text{P}(O|M) = \text{P}(O,H|M)\text{d}H \), which is computationally prohibitive. We can approximate the likelihood by using the sampled complete data set \( P(O|M) = \frac{1}{S} \sum_{i=1}^{S} P(C_i|M) = \frac{1}{S} \sum_{i=1}^{S} \text{P}(O,H_i|M) \). In the subsequent model searching procedure, a new structure \( G' \) is randomly selected from the neighborhood of \( G_k \) (\( k \geq 1 \)) so that \( G_k \) and \( G_k \) differ only by one arc. We then again estimate the parameter set \( \theta' \) of \( G' \) using the \( \text{EstimateParameter} \) function and approximate the likelihood \( P(O|M) \) using the sampled complete data set. The approximated likelihoods are used to compute the Bayes factor, which is then used to compute the Metropolis–Hastings ratio. The new model will be stochastically accepted based on the Metropolis–Hastings ratio.

The \( \text{EstimateParameter} \) function is outlined in Figure 5. It first initializes the parameters of a model as below. The CPT of a node is initialized so that all cases are equally possible except for the \( \text{f/f} \) mutation. A network component can only take its lowest value (i.e. 1) if it completely loses its function due to the \( \text{f/f} \) mutation. The \( \text{EstimateParameter} \) function then iterates the following three steps until converges: (a) Sample \( S \) copies of the hidden variables using the current model and the \( \text{probabilistic logic sampling} \) method (Henrion, 1988); (b) The sampled hidden variable value sets and the observed data \( O \) form a current population of complete data that is used to compute the parameter posterior via Bayesian updating; (c) Update the parameter set \( \theta \) by the parameter posterior means. The \( \text{probabilistic logic sampling} \) method was slightly modified to meet our need. The modified one first unrolls the DBN to have \( T \) time-slices, computes
(a) Initialize the model $M$ with the structure $G$, and set $T = 40$.
(b) Estimate the parameter set $\Theta$, of $M$, given its structure $G$, and the observed variable set $O$ by calling the function $[\Theta_{G}, \{C_i\}_{i=1}^{5}] = \text{EstimateParameter}(G, O, S, T)$.
(c) Estimate the likelihood of the observed variables $P(O | M) = \frac{1}{S} \sum_{s=1}^{S} P(C_i | M)$.
(d) for $k = 1, 2, \ldots$, until converge
   (d.1) Randomly select a structure $G'$ from $N(G)$ – the neighborhood of $G$.
   (d.2) $[\Theta', \{C_i\}_{i=1}^{5}] = \text{EstimateParameter}(G', O, S, T)$. Let $M' = <G', \Theta'>$.
   (d.3) $P(O | M') = \frac{1}{S} \sum_{s=1}^{S} P(C_i | M)$.
   (d.4) Bayes factor $F = \frac{P(O | M')}{P(O | M)}$; Ratio $R = \frac{F \cdot N(G')}{|N(G')| \cdot P(G')}$
   (d.5) Accept $M'$ with the probability of $\min(1, R)$. If accepted, let $M_{new} = M'$, otherwise $M_{new} = M$.
(e) return $M_{new}$.

Fig. 4. The Metropolis–Hastings algorithm for learning a vulval induction DBN.

(a) Initialize the parameter set $\Theta_{G}$ for $G$.
(b) for $n = 1, 2, \ldots$, until converge
   (b.1) Use the probabilistic logic sampling method with
        the argument set $(G, \Theta_{G}, O, S, T)$ to sample $S$ copies
        of complete data sets $\{C_i\}_{i=1}^{5} = \langle O, H_i \rangle_{i=1}^{S}$.
   (b.2) Compute $P(\Theta_{G}, O, G, \{C_i\}_{i=1}^{5})$ using Bayesian
        updating.
   (b.3) Compute $\Theta_{G}$ from $P(\Theta_{G}, O, G, \{C_i\}_{i=1}^{5})$.
(c) return $[\Theta_{G}, \{C_i\}_{i=1}^{5}]$.

Fig. 5. $[\Theta, \{C_i\}_{i=1}^{5}] = \text{EstimateParameter}(G, O, S, T)$.

the ordering of nodes, instantiates the network based on the genetic
conditions, and then sequentially samples unobserved variables
according to the ordering. It discards trials whenever a variable
instantiation conflicts with the rules (see the Data preprocessing
Section). For example, a trial will be discarded, if a VPC is supposed to
adopt fate 1/C14, however, the trial shows that its fate 1/C14 marker either
does not stabilize at its highest value or is not the only fate marker
stabilizing at its highest value. When evaluating simulation results of
a wildtype C. elegans, we not only checked the fate markers but also
make sure that the temporal patterns of lst-1, lst-2, lst-4, dpy-23 and
let-3 match the descriptions of Yoo et al. (2004). The probabilistic logic
sampling method is simple but inefficient. However, given the nature of
our training data, it was a reasonable choice. The value of $S$ was set so
that 200 complete data samples were generated for each observation.
For example, the lin-12(t) experiment has 12 observations in our
training data. The sampling method should generate $12 \times 200$ complete
data samples for the lin-12(t) experiment. Assuming the parameters
follow multinomial distributions with Dirichlet priors, we can efficiently
compute the parameter posteriors and the parameter posterior means in
closed forms (Cooper and Herskovits, 1992). Note that we need to pool
the expected sufficient statistics for all nodes that share the same
parameters when computing the posterior means.

Since the number of potential network structures is super exponential
to the number of nodes, it is impractical for us to search in a complete
structure space defined by 20 nodes. To make the task feasible,
we limited the maximum number of parents of a node to be 3
(control nodes are not counted). The LIN-12 receptor of a VPC receives three lateral
signals from two neighboring VPCs and itself. A signal integration node was added as a parent of the LIN-12 node to integrate three lateral signals. The signal integration node simply maps different combinations of lateral signals into a set of numerical values without considering the order of lateral signals. For example, the signal integrator of a VPC maps the following signal combinations to the same value: $[1, 2, 3]$, $[1, 3, 2]$, $[2, 1, 3]$, $[2, 3, 1]$, $[3, 1, 2]$ and $[3, 2, 1]$, where the numbers in the brackets represent the signal strength from the left neighboring VPC, the VPC itself, and the right neighboring VPC respectively. In addition, the arcs between the lateral signal nodes and the LIN-12 nodes are the only allowable direct connections between VPC subnetworks. Complicate models were penalized by setting the prior $P(G) = e^{-\delta}$, where $c$ is a constant, $\psi$ is empirically set as 0.8 and $\delta$ is the number of arcs in $G$.

### 3 RESULTS

Figure 6 shows one of the most likely models found. Due to the space limit, we only show one VPC subnetwork and hide genetic control nodes. We grouped $lst-1$, $lst-2$ and $lst-3$ together for the following reasons. First, they share the same expression pattern (Berset et al., 2001). Second, there is no additional information in our training that can be used to further clarify their roles in the network. Finally, from a practical point of view, this treatment reduces the computational complexity. For the same reasons, $dpy-23$ and $lst-3$ were grouped together. The model contains complicate feedback controls such as those through $lip-1$, $dpy-23$, $lst-3$ and $lst-1_2_4$. Close examination of the corresponding parameters revealed that those feedbacks are negative controls with various strengths. Such highly redundant mode by which LIN-12/Notch signaling antagonizes EGFR pathway make C. elegans vulval development robust to individual perturbations. This model is capable of simulating vulval induction under 36 genetic conditions (Table 2). Detailed simulation results are provided in the supplementary data.

Here we selectively discuss one of the most interesting cases: the $lin-15$ $l/f$ in silico experiment. The basal activity of LIN-15 inhibits the EGFR pathway. The $l/f$ mutation of $lin-15$ will result in a Muv phenotype even though the inductive signal LIN-3 is eliminated. Our simulation results are consistent with the observations in $lin-15$ $l/f$ experiments: it is rare for two neighboring VPCs to adopt fate 1. This phenomenon is the result of the collaboration and competition between EGFR signaling and LIN-12/Notch signaling. When $lin-15$ undergoes $l/f$ mutation, all VPCs are affected by the strong activities of EGFR pathway to highly express genes specific to fate 1. However, in the mean time, they simultaneously send out strong lateral signals to activate LIN12/Notch pathways in their neighboring VPCs, which in return suppress EGFR pathways in the corresponding VPCs. Both pathways are now competing to control the fate decision-making processes. If EGFR signaling in one of VPCs is successfully suppressed by its LIN-12/Notch signaling, it will down-regulate genes specific to fate 1 and at the same time reduce its lateral signal level as suggested by our model. Its neighboring VPCs will then receive weaken lateral signals that may not be strong enough to suppress their EGFR pathways so that they will tend to adopt fate 1. Such an example is illustrated by $P3.p$, $P4.p$ and $P5.p$ in Figure 7. It is also possible that the EGFR pathways of three adjacent VPCs are suppressed almost simultaneously. All of them then send out weaken lateral signals which may result in the restoration of strong EGFR signaling in them. Such an example is illustrated as $P6.p$, $P7.p$ and $P8.p$ in Figure 7. The complicate in silico competition between these two pathways is stochastically regulated by the parameters of our vulval induction DBN. As the result, the expression levels of fate markers fluctuate in our simulations, but eventually all of them reach steady states. Nevertheless, such stochastic properties make our work the first one able to reproduce such fate variability observed biologically. If there is no such stochastic element, the activities of both EGFR pathway and LIN-12/Notch pathway in all VPCs will be either up-regulated or down-regulated simultaneously so that the whole vulval

### Table 2. Simulation results of the model shown in Figure 6. Genotype indicates the genetic conditions. Phenotype indicates the simulation results. The simulated fates of VPCs are listed in the order of P3.p, P4.p, P5.p, P6.p, P7.p and P8.p. Muv means multivulva phenotypes. A complete list of simulated multivulva phenotypes is provided in the Supplementary Material.

| Genotype | Phenotypes |
|----------|------------|
| wildtype | 3 3 2 1 2 3 |
| $lin-3$ (ff) | 3 3 3 2 3 3 |
| $lin-3$ (ff) + $lin-12$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-2$ (ff) | 3 3 3 3 3 3 |
| $lin-7$ (ff) | 3 3 3 3 3 3 |
| $lin-10$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-2$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-7$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-10$ (ff) | 3 3 3 3 3 3 |
| $lin-15$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-15$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-15$ (ff) + $lin-12$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-2$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-3$ (ff) + $lin-7$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-3$ (ff) + $lin-10$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-2$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-7$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-10$ (ff) + $lin-12$ (gf) | 2 2 2 2 2 2 |
| $lin-2$ (ff) + $lin-12$ (ff) | 3 3 3 3 3 3 |
| $lin-7$ (ff) + $lin-12$ (ff) | 3 3 3 3 3 3 |
| $lin-10$ (ff) + $lin-12$ (ff) | 3 3 3 3 3 3 |
| $lin-3$ (ff) + $lin-15$ (ff) + $lin-12$ (gf) | 3 3 3 3 3 3 |
| $lin-15$ (ff) + $lin-12$ (gf) | 3 3 3 3 3 3 |
| $lin-15$ (ff) + $lin-12$ (ff) | 1 1 1 1 1 1 |
| let-23 (ff) | 3 3 3 2 3 3 |
| let-23 (ff) + $lip-1$ (ff) | 3 3 2 1 2 3 |
| sem-5 (ff) | 3 3 3 2 3 3 |
| sem-5 (ff) + $lip-1$ (ff) | 3 3 2 1 2 3 |
| $mpk-1$ (ff) | 3 3 3 2 3 3 |
| $mpk-1$ (ff) + $lip-1$ (ff) | 3 3 2 1 2 3 |
| let-23 (gf) + $lin-2$ (ff) | 3 3 2 1 2 3 |
| let-23 (gf) + $lin-7$ (ff) | 3 3 2 1 2 3 |
| let-23 (gf) + $lin-10$ (ff) | 3 3 2 1 2 3 |
| let-60 (ff) | 3 3 3 3 3 3 |
and Horvitz (1989), in which all VPCs adopt fate 3. Since the anchor cell is absent, i.e. ac(-), there is no inductive signal to activate EGFR signaling and suppress genes specific to fate 3. Hence, the levels of fate 3 markers should be high and all VPCs should tend to commit to fate 3. With the gain-of-function of LIN-12, all VPCs are affected by strong LIN-12/Notch signaling to highly up-regulate genes specific to fate 2. If 2 does not inhibit 3, all VPCs will commit to a mixture of 2 and 3, i.e. an undetermined state. The 2 → 3 relationship enables our model to simulate results that match the experimental observations.

Our learning procedure found several other models that can explain the data almost as well as the one shown in Figure 6. Some of those models differ only locally. For example, several models differ only at the subnetwork consisting of LIN-2, LIN-7 and LIN-10, which form linear paths in those models. There are six possible such linear paths. This means the learning procedure is not able to decide the order of LIN-2, LIN-7 and LIN-10 given the training data we collected. Searching the literature, we found that these three proteins was shown to form a protein complex by yeast two-hybrid, in vitro binding, and in vivo co-immunoprecipitation experiments (Kaech et al. 1998). Hence, one tip learned from this kind of scenario is that a linear local structure with such properties could imply a protein complex.

Our model is not able to completely correctly simulate lin-3 mutants. Sternberg and Horvitz (1989) reported that lin-3 allele resulted in incomplete penetrance. However, our simulation results indicate that lin-3 allele generates complete penetrance in which all VPCs adopt fate 3. We speculate that there are additional mechanisms contributed by other molecules that were not interrogated in our training data and hence were not modeled by us.

4 DISCUSSION

Collaborative cell differentiation and organ patterning are delicate processes that are precisely regulated by networks with redundancy and complex feedback controls. It offers distinct advantages to describe the underlying networks in a computational model, especially one that is capable of modeling the dynamic behaviors of the systems. With the emerging huge amount animal organogenesis data, there is a need to automatically make sense of those data. Our study is a small, however, promising step towards automatically modeling and simulating of full-scale animal organogenesis (i.e. the development of an adult multicellular animal from a single cell). Computational models can also be used to generate hypotheses or provide guidelines for biomedical researchers. A hypothesis may be either verified or disapproved by experiments. Both results are quite informative for computationally re-evaluating other hypotheses.

Our MCMC learning procedure iteratively improves the model. This makes it easy to incorporate new data to refine the current model by simply continue the MCMC process with the new data included. Given experimental data of higher resolution in time and space, we can also progressively increase the complexity and modeling capability of our current discrete model, e.g. by increasing the number of its discrete states.
to asymptotically approximate the detailed dynamics of the underlying systems.

We only managed to use a portion of the published data related to \textit{C. elegans} vulval induction due to the following reasons. Many published data sets were presented in ways extremely difficult for us to compute quantitatively. Our study would have greatly benefited if biologists could report the final fate of each VPC instead of the average inductions or the coarse phenotypes observed in their experiments. A more comprehensive complete phenotypic data set will also make computational modeling of penetrance feasible. Temporal profiles of network components will be ideal for tuning the dynamics of a model. In addition, a gene can be mutated in several different ways by different labs, which could cause noticeable differences. In this study, we tried our best to avoid mixing data from different mutations of the same gene. This can be done by expanding the corresponding control nodes. However, it will increase the computational complexity and require more data to train a model. Finally, the computational complexity of our model learning procedure is very high. Future endeavors will be devoted into developing faster learning algorithms. We will also collect more data sets to expand our vulval induction model to include more molecules.

Ten-fold cross-validation tests were carried out to address the potential overfitting problem of our model. Each time, about 10% data were randomly selected as the test data from those genetic conditions that have no less than 10 samples. The rest data was used to in training. It is prohibitive for us to re-run the whole model training process for each test run because of the high computational complexity. Hence, we fixed the model structure as shown in Figure 6 and retrained its parameters using the training data. The retrained models were able to reproduce the test cases. However, this was expected because our data preprocessing step had greatly simplified the problem by removing most of variances in the incomplete phenotypic data, which contributes to the majority of our training data. To fully address the overfitting problem, we will need a large complete phenotypic data, which unfortunately is not available now. Alternatively, new data preprocessing methods should be developed in collaboration with biological experts to preserve variances in the incomplete phenotypic data, and the model learning procedure may also need to be modified correspondingly. We will investigate these issues in the future.

\textit{Conflict of Interest}: none declared.

\section*{REFERENCES}

\begin{thebibliography}{100}
\bibitem{Beitel:1990} Beitel, G. J. \textit{et al.} (1990) \textit{Caenorhabditis elegans} ras gene let-60 acts as a switch in the pathway of vulval induction. \textit{Nature}, \textbf{348}, 503–509.
\end{thebibliography}

Berset, T. \textit{et al.} (2001) Notch inhibition of RAS signaling through MAP kinase phosphatase LIP-1 during \textit{C. elegans} vulval development. \textit{Science}, \textbf{291}, 1055–1058.

Churchill, G. A. (1989) Stochastic models for heterogeneous DNA sequences. \textit{Bull. Math. Biol.}, \textbf{51}, 79–94.

Cooper, G. F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. \textit{Mach. Learn. J.}, \textbf{9}, 308–347.

Dean, T. and Kanazawa, K. (1989) A model for reasoning about persistence and causation. \textit{Comput. Intell.}, \textbf{5}, 142–150.

Dojer, N. \textit{et al.} (2006) Applying dynamic Bayesian networks to perturbed gene expression data. \textit{BMC Bioinformatics}, \textbf{7}, 249.

Fisher, J. \textit{et al.} (2005) Computational insights into \textit{Caenorhabditis elegans} vulval development. \textit{Proc. Natl Acad. Sci. USA}, \textbf{102}, 1951–6.

Friedman, N. (2004) Inferring cellular networks using probabilistic graphical models. \textit{Science}, \textbf{303}, 799–805.

Friedman, N. \textit{et al.} (2000) Using Bayesian networks to analyze expression data. \textit{J. Comput. Biol.}, \textbf{7}, 601–620.

Harel, D. (1987) Statecharts: a visual formalism for complex systems. \textit{Sci. Comput. Program.}, \textbf{8}, 231–274.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. \textit{Biometrika}, \textbf{57}, 97–109.

Henrion, M. (1988) Propagating uncertainty in Bayesian networks by probabilistic logic sampling. \textit{In Uncertainty in Artificial Intelligence}, 149–163.

Husmeier, D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. \textit{Bioinformatics}, \textbf{19}, 2271–2282.

Imoto, S. \textit{et al.} (2006) Analysis of gene networks for drug target discovery and validation. \textit{Methods Mol. Biol.}, \textbf{360}, 33–56.

Kaelb, S. M. \textit{et al.} (1998) The LIN-2/LIN-7/LIN-10 complex mediates basolateral membrane localization of the \textit{C. elegans} EGF receptor LET-23 in vulval epithelial cells. \textit{Cell}, \textbf{94}, 761–771.

Kim, S. \textit{et al.} (2004) Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. \textit{Bioinformatics}, \textbf{75}, 57–65.

Lee, I. \textit{et al.} (2004) A probabilistic functional network of yeast genes. \textit{Science}, \textbf{306}, 1555–1558.

Murphy, K. P. (2002) Dynamic bayesian networks: representation, inference and learning. \textit{Computer Science}. University of California, Berkeley.

Pearl, J. (1988) \textit{Probabilistic Inference in Intelligent Systems}. Morgan Kaufman, San Mateo, CA.

Pena, J. M. \textit{et al.} (2005) Growing Bayesian network models of gene networks from seed genes. \textit{Bioinformatics}, \textbf{21} (Suppl. 2), i224–i229.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. \textit{Proceedings of IEEE}, \textbf{77}, 257–286.

Sachs, K. \textit{et al.} (2002) Bayesian network approach to cell signaling pathway modeling. \textit{Sci. STKE}, \textbf{2002}, PE38.

Sachs, K. \textit{et al.} (2005) Causal protein-signaling networks derived from multi-parameter single-cell data. \textit{Science}, \textbf{308}, 523–529.

Simsek, J. S. \textit{et al.} (1996) LET-23 receptor localization by the cell junction protein LIN-7 during \textit{C. elegans} vulval induction. \textit{Cell}, \textbf{85}, 185–204.

Sternberg, P. W. and Horvitz, H. R. (1989) The combined action of two intercellular signaling pathways specifies three cell fates during vulval induction in \textit{C. elegans}. \textit{Cell}, \textbf{58}, 679–693.

Woolf, P. J. \textit{et al.} (2005) Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. \textit{Bioinformatics}, \textbf{21}, 741–753.

Yoo, A. S. \textit{et al.} (2004) Crosstalk between the EGF-R and LIN-12/Notch pathways in \textit{C. elegans} vulval development. \textit{Science}, \textbf{303}, 663–666.

Zou, M. and Conzen, S. D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. \textit{Bioinformatics}, \textbf{21}, 71–79.