BMC Bioinformatics

BioMed Central

Research article

Predicting state transitions in the transcriptome and metabolome using a linear dynamical system model

Ryoko Morioka[1], Shigehiko Kanaya[2], Masami Y Hirai[1], Mitsuru Yano[3], Naotake Ogasawara[2] and Kazuki Saito*[1,3]

Address: [1]RIKEN Plant Science Center, Yokohama, Kanagawa, Japan, [2]Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan and [3]Department of Molecular Biology and Biotechnology, Graduate School of Pharmaceutical Science, Chiba University, Inage-ku, Chiba, Japan

Email: Ryoko Morioka - ryoko@psc.riken.jp; Shigehiko Kanaya - skanaya@gtc.naist.jp; Masami Y Hirai - myhirai@psc.riken.jp; Mitsuru Yano - 326-yano@psc.riken.jp; Naotake Ogasawara - nogasawa@bs.naist.jp; Kazuki Saito* - ksaito@faculty.chiba-u.jp

* Corresponding author

Published: 18 September 2007

BMC Bioinformatics 2007, **8**:343 doi:10.1186/1471-2105-8-343

Received: 28 December 2006
Accepted: 18 September 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/343

Abstract

Background: Modelling of time series data should not be an approximation of input data profiles, but rather be able to detect and evaluate dynamical changes in the time series data. Objective criteria that can be used to evaluate dynamical changes in data are therefore important to filter experimental noise and to enable extraction of unexpected, biologically important information.

Results: Here we demonstrate the effectiveness of a Markov model, named the Linear Dynamical System, to simulate the dynamics of a transcript or metabolite time series, and propose a probabilistic index that enables detection of time-sensitive changes. This method was applied to time series datasets from *Bacillus subtilis* and *Arabidopsis thaliana* grown under stress conditions; in the former, only gene expression was studied, whereas in the latter, both gene expression and metabolite accumulation. Our method not only identified well-known changes in gene expression and metabolite accumulation, but also detected novel changes that are likely to be responsible for each stress response condition.

Conclusion: This general approach can be applied to any time-series data profile from which one wishes to identify elements responsible for state transitions, such as rapid environmental adaptation by an organism.

Background

Biochemical systems in living cells are robust and flexible. Investigating the responses of cells (and organisms) to environmental changes typically requires a system-level analysis of the interactions between the various molecular elements (genes, enzymes, and metabolites) that comprise the system. A key step to analyze system responses to environmental changes is identifying large state changes or "transitions". A statistical method that could detect such transitions would be a powerful analytical tool for finding important factors in large-scale profiles, such as variations in gene expression.

Previous analyses of gene expression profiles have often made use of graphical models, such as Bayesian Networks [1,2], Graphical Gaussian Modelling [3], Boolean Networks [4,5], and Auto-Regressive models [6]. However, not many approaches have explicitly modelled observa-

tional noise. In Auto-Regressive analyses, for example, observational vectors $\mathbf{y}_t$ are recursively defined by the following Equation [6]:

$$\mathbf{y}_t = A\mathbf{y}_{t-1} + \varepsilon_t \quad (1)$$

where $\mathbf{y}_t$ is an observational vector of genes or metabolites at time t, A is an observational transition matrix, and ε_t is Gaussian noise. Because this model does not distinguish observational and inherent (e.g. biological) noises, identification of transition states becomes difficult in the presence of substantial noise.

Here we propose an extension of the Auto-Regressive model [6], which has been modified by the addition of reduced set of internal states, as explained in the Results and Discussion section. We chose a mathematical model, the Linear Dynamical System (LDS), as the basis of our method because it does not impose any specific requirements on the data used. LDS is expected to eliminate the confounding influence of observational noise in time series data. The model was applied to detect cellular state transitions in transcriptome and metabolome time series datasets from *Bacillus subtilis* and *Arabidopsis thaliana* maintained under stress conditions.

Results and discussion

Overview of our method

Our method has two steps. First, transition time points for each time series are detected using LDS, which mathematically distinguishes transitional fluctuations from experimental noise. The transition point is detected by the logarithm of the likelihood values. Here "likelihood value" means the generative probability of current data based on the condition of the past datasets. If this value is low, then the current data cannot be adequately explained by past datasets. In other words, a transition has occurred. In the second step, relevant factors such as genes and/or metabolites related to the transitions are extracted by Batch-Learning Self Organizing Mapping (BL-SOM) using changes in expression levels [7]. In summary, the LDS uses compressed information called "internal states", defined as the degenerate parameters of gene expression/metabolite accumulation profiles, to detect transitions, and then BL-SOM generates a 'Feature map', which is a two-dimensional lattice reflecting the similarity among clusters, based on the gene expression/metabolite accumulation profiles in order to visually characterize each state.

Linear Dynamical System (LDS) for time series analyses

LDS uses internal state variables in the generative model for cellular internal state changes. These internal states correspond to the compressed description of the observed biological system prior to adding noise factors.

The total experimental dataset of the time series and the corresponding internal state are denoted by $Y_{1:T} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T\}$ and $X_{1:T} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$, respectively. Each element in these vectors is defined as:

$$\mathbf{y}_t = (\mathbf{y}_{t1}, \mathbf{y}_{t2}, \cup, \mathbf{y}_{tD})' \in R^D \quad (2)$$

$$\mathbf{x}_t = (\mathbf{x}_{t1}, \mathbf{x}_{t2}, \cup, \mathbf{x}_{tN})' \in R^N \quad (3)$$

where $t = 1, 2, ...,$ T is the measurement order of the time series, D is the dimension of vector $\mathbf{y}_t$ representing expression levels of D genes or metabolites, and N is the dimension of vector $\mathbf{x}_t$ representing internal states. To distinguish observational noise from true information on cellular transitions, we focus on two probability densities: the density between internal state variables $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, and the density for evaluation of observational noise $p(\mathbf{y}_t|\mathbf{x}_t)$. The proposed model is further defined as follows:

$$\text{Observational equation: } \mathbf{y}_t = V\mathbf{x}_t + \eta_t \quad (4)$$

$$\text{Transition equation: } \mathbf{x}_t = W\mathbf{x}_{t-1} + \varepsilon_t \quad (5)$$

where V is a $D \times N$ observational matrix, W is an $N \times N$ internal state transition matrix, D-dimensional vector η_t is observational noise, N-dimensional vector ε_t is transition noise. The vectors $\mathbf{x}_1$, ε_t, η_t are generated according to the following equations:

$$\mathbf{x}_1 \sim N_N(\mathbf{x}_1|\mu_1, \sigma_1^2 I_N) \quad (6)$$

$$\varepsilon_t \sim N_N(\varepsilon_t|0_N, \sigma_\varepsilon^2 I_N) \quad (7)$$

$$\eta_t \sim N_D(\eta_t|0_D, \sigma_\eta^2 I_\eta) \quad (8)$$

The next step is to define the relevant probability densities. $N_p(\mathbf{x}|\mathbf{m}, \Sigma)$ is a probability density function when a p-dimensional probabilistic vector $\mathbf{x}$ obeys a Gaussian distribution whose mean vector is $\mathbf{m}$, and covariance matrix Σ (Equation 9).

$$N_p(\mathbf{x}|\mathbf{m}, \Sigma) \equiv (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m})'\Sigma(\mathbf{x}-\mathbf{m})] \quad (9)$$

We assume that the observational and internal transition noises are both Gaussian, and therefore the relationship is a first-order Markov process (Equation 10).

$$p(\mathbf{x}_t, \mathbf{y}_t \mid X_{1:t-1}, Y_{1:t-1}) = p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|x_{t-1}) \quad (10)$$

The model parameter of (4)–(8) is defined as the parameter set θ.

$$\theta = \{\mu_1, \sigma_1, \mathbf{W}, \sigma_\varepsilon, \mathbf{V}, \sigma_\eta\} \quad (11)$$

Note that the model corresponds to a Kalman Filter when θ is known (see also Methods section) [8].

The initial state $\mathbf{x}_1$ is defined as:

$$p(\mathbf{x}_1|\theta) = N_N(\mathbf{x}_1|\mu_1, \sigma_1^2 I_N) \quad (12)$$

From Equations (5) and (7), the following function is obtained:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta) = N_N(\mathbf{x}_t|\mathbf{W}\mathbf{x}_{t-1}, \sigma_\varepsilon^2 I_N) \quad (13)$$

From Equations (4) and (8), the following function is obtained:

$$p(\mathbf{y}_t|\mathbf{x}_t, \theta) = N_D(\mathbf{y}_t|\mathbf{V}\mathbf{x}_t, \sigma_\eta^2 I_D) \quad (14)$$

Using these results, the following joint probability is obtained:

$$p(Y_{1:T}, X_{1:T}|\theta) = p(\mathbf{x}_1|\theta)\left\{\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \theta)\right\}\left\{\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t, \theta)\right\} \quad (15)$$

The parameter optimization follows a standard EM algorithm (see Methods section).

Criterion for detecting internal state transitions

Using the resulting estimated parameter, the log-likelihood with respect to the present time point t when all time points are given is defined by Equation (16):

$$\log L_t = \log p(\mathbf{y}_t|Y_{1:t-1}, \theta) \quad (16)$$

This is calculated using the E-step formula (see Equation 23 in Methods) after parameter estimation using the Kalman filter.

When the log-likelihood value $\log L_t$ becomes much lower than $\log L_{t-1}$, then γ_t cannot be explained by $Y_{1:t-1}$, i.e., the cellular internal state has changed at time t. In this study, the point at which the log-likelihood value becomes relatively low between whole time points is defined as the state transition point. If the log-likelihood value remains low over a certain period, then the cells are changing their states continuously during that period.

Analysis of the Bacillus subtilis *data*

We first analyzed the relationship of cell population to state transition time on transcriptome data of *Bacillus subtilis* (Figure 1). Here, the exponential growth phase and stationary phase are commonly used microbiology terms referring to the state of the cellular population, as measured by the optical density (see also Methods section). The transition from exponential growth to the stationary phase was observed in 8 culture media: Lysogeny Broth (LB), Minimum Glucose Medium (MGM), Glucose Starvation (GS), Phosphate Starvation (PS), Competence Medium (CM), Difco Sporulation Medium (DSM), Competence Sporulation Medium (CSM) and DSM plus Glucose Glutamine (DGG). We confirmed that the log-likelihood index produced by LDS was smaller at the transition time between two phases. Next, we fitted the index calculated by the model to the phase transition data. For cell populations growing under two culture conditions, namely LB (control) and MGM (limited glucose), we found that BL-SOM yielded different classification results for gene expression (Figure 2a, b). This result indicates that expression of the genes responsible for the transition varied between the different environmental conditions, although their transitions appeared similar. For cells grown in either CSM or DSM, two transition points for sporulation were detected. The first was the well-known transition from exponential growth to the stationary phase. However, the second was a novel transition detected by this approach. At the first transition in CSM at around time point 3, log-likelihood values show a sustained drop. The analysis suggests that cells take a long time to adapt to the CSM culture environment. The second transition point in the sporulation media was further investigated by analysis of Feature maps generated by BL-SOM [7]. The candidate genes for the second transition were those activated just before the transition point and repressed soon after the transition point (Table 1). These genes are listed in Table 1 and include those related to lysis of the mother cell, such as *cwlH*. Thus, the second transition corresponded to mother cell lysis [9], a type of apoptosis.

Using the analytical approach described here, we not only succeeded in detecting the well-known transition from exponential growth to the stationary phase, but also identified another, novel transition point. This result suggests the possibility that, even in periods that are assumed to be eventless, cells may be invoking their adaptive systems.

Analysis of the Arabidopsis *data*

As described in the Methods section, we analysed changes in gene expression and metabolite accumulation in *Arabidopsis* plants following their transfer to sulfur-deficient conditions. We detected a transition between 12 and 24 hours in both gene expression and metabolite accumulation profiles in both leaves and roots. In addition, we detected a second transition at the final time point (168 hr) in the metabolite accumulation profile in roots (see Figure 3). At the transition point of 12–24 hr, glucosinolate biosynthesis was decreased in leaves and anthocy-

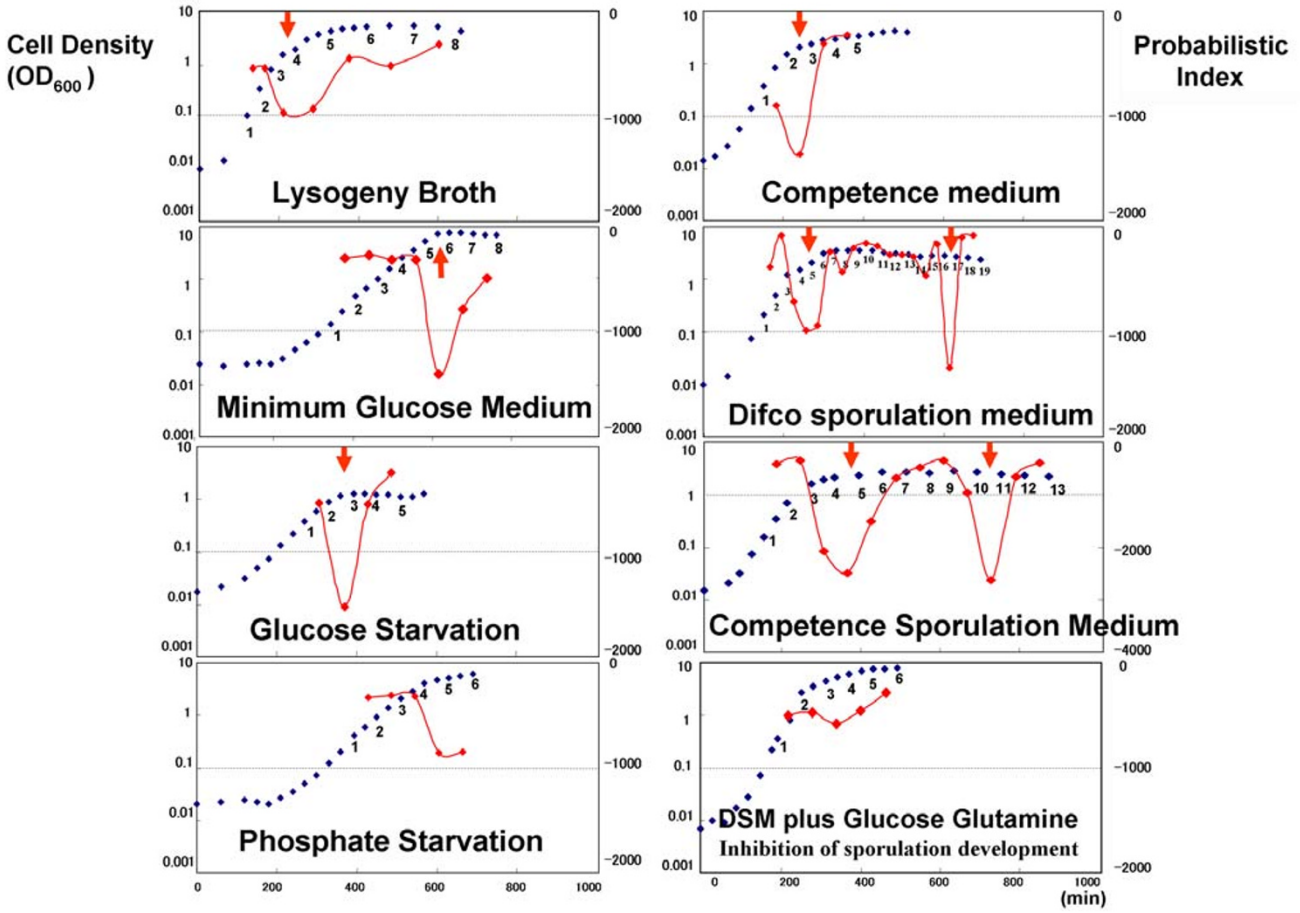

Figure 1
The results of log-likelihood values and experimental conditions. Relationship between optical density values and state transition time. Red plots show the probabilistic index for the evaluation of the state change. Blue plots show the Optical Density (OD) values that represent the cellular populations at each time.

anin biosynthesis was initiated in roots. The predicted transitions obtained by this analysis are consistent with those identified previously [10], indicating that our method can reliably identify candidate genes and metabolites involved in transition points. The transition time point detected for root metabolites at the end of the experiment (168 hr) showed that even after this period of time roots of *A. thaliana* continued to change in response to sulfur deprivation, at least in terms of metabolite accumulation.

On the basis of the estimated transition results, coupled with prior knowledge and the Feature map subtraction obtained by BL-SOM, we identified metabolites whose accumulation profiles showed changes that coordinated with the predicted transition point. These metabolites were found to be involved in biochemical pathways that are critical for the response to sulfur deprivation stress, for example, glucosinolate biosynthesis in leaf and anthocyanin biosynthesis in root [10].

Our results also suggested the presence of lipid metabolic responses in *Arabidopsis* to sulfur stress. The accumulation patterns of detected ion peaks whose mass-to-charge ratio (*m*/*z*) values corresponded to molecular species with various acyl groups, such as phosphatidylglycerol, phosphatidylethanolamine, phosphatidylcholine, phosphatidic acid, and sulfoquinovosyl diacylglycerol, are shown in Figure 4. Because the accumulation profiles of these compounds showed similar patterns, we predict that lipid biosynthesis was also co-ordinately repressed at the transition at 24 hr.

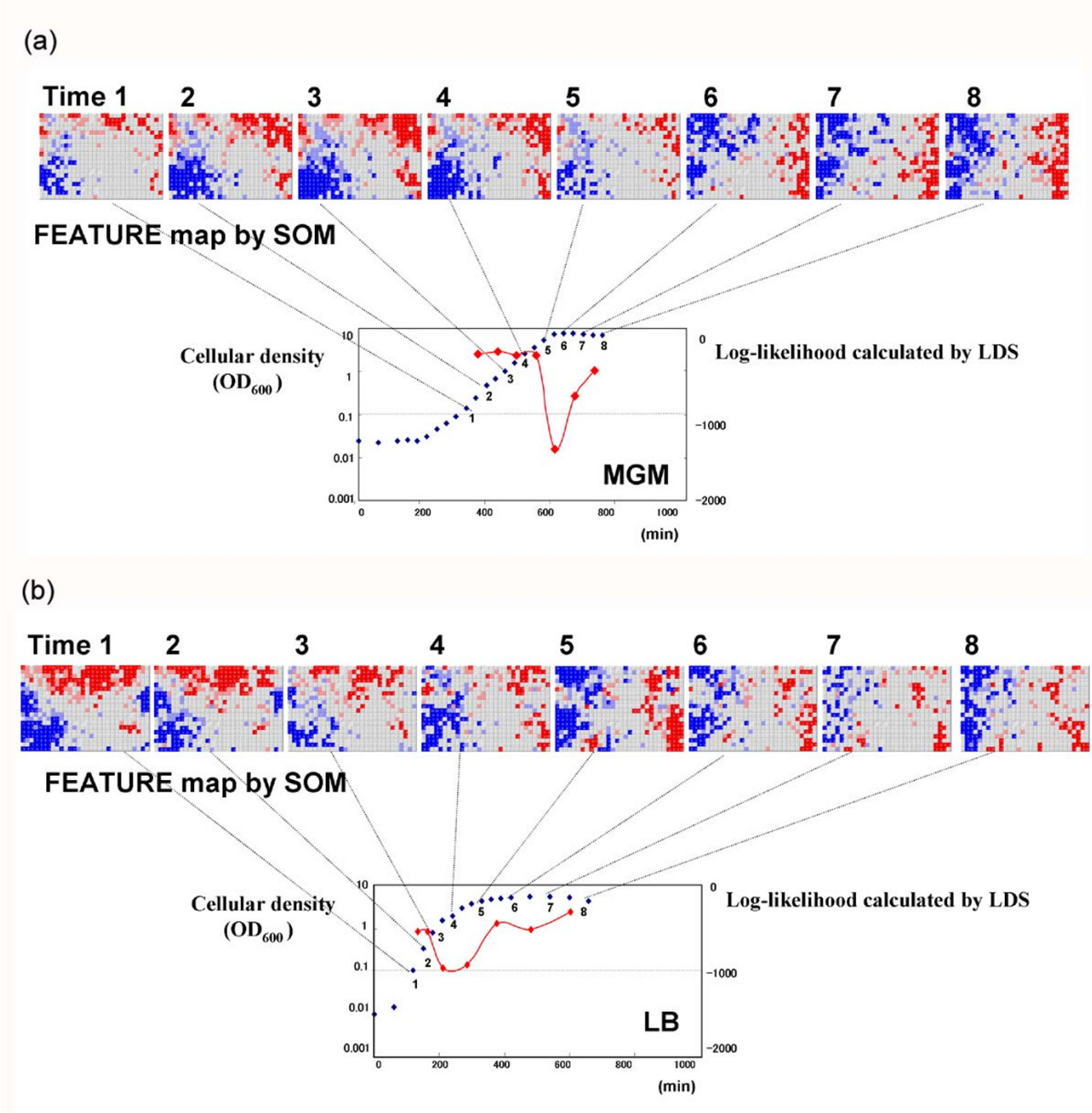

Figure 2

Examples of the results of *Bacillus* data analyses. The horizontal axis shows the culture time in minutes. The vertical axis represents the cellular density values (blue plots) and the probabilistic index for transition (red plots). a) The results of analysis of the LB data (control culture condition). According to the probabilistic index, the first transition was predicted between time points 3 and 4, the period that corresponded to transition from exponential growth phase to the stationary phase as indicated by cellular density values. The probabilistic index identifies another state change between time points 6 and 7 during the "stationary" phase. b) The results of analysis of the MGM data (stress culture condition). According to the probabilistic index, a state change was predicted between time points 5 and 6, a period that corresponded to the transition from the exponential growth phase to the stationary phase. Compared to the results from cells grown in LB, the transition timing was different. This difference was caused by the lack of glucose in MGM.

Table 1: The list of transition driving genes identified in cells grown in CSM and DSM

Functional categories	Genes
Adaptation to typical conditions	*ypeB*
Cell wall	*ytcC, ykuG, ykoT, ywhE, yunA, cwlH*
Germination	*yaaH, yfkQ, yndD, gerBB, gerKB, yndE, gerAC, yfkR*
Membrane bioenergetics	*yhfW*
Sporulation	*spoVFA, spoVAD, spoVAC, spoVAE, spoVAB, spoVK, spoIVCA, cotC, cotA, yaaH, sspE, sspB*
Transport/binding proteins and lipoproteins	*ymfD, araP, yveA, ywcA, ywrK*
Detoxification	*ykoY*
Regulation	*sigG, splA*
Antibiotic production	*yitA, yitC*
Carbohydrates and related molecules	*yqiQ, adhB, yoaI, yesX, yitF, yqiQ*
Metabolism of amino acids and related molecules	*aprX, spoVFA*
Metabolism of lipids	*yngF*
Phage-related function	*yndL, yqbO, yqbQ*
Proteins with unknown function	*yodP, yheD, yhcQ, yvdQ, ytzC, yybC, ydfO, yhcV, yesV, yndM, ydfR, yngD, ykjA, yetA, yusN, yozN, ypjD, ytlB, yqaN, ythQ, yycQ, yurS, yrkS, yxaG, yesJ, ysnE*

A profile of sulfate accumulation was generated using capillary electrophoresis (Figure 4f). In comparison with the

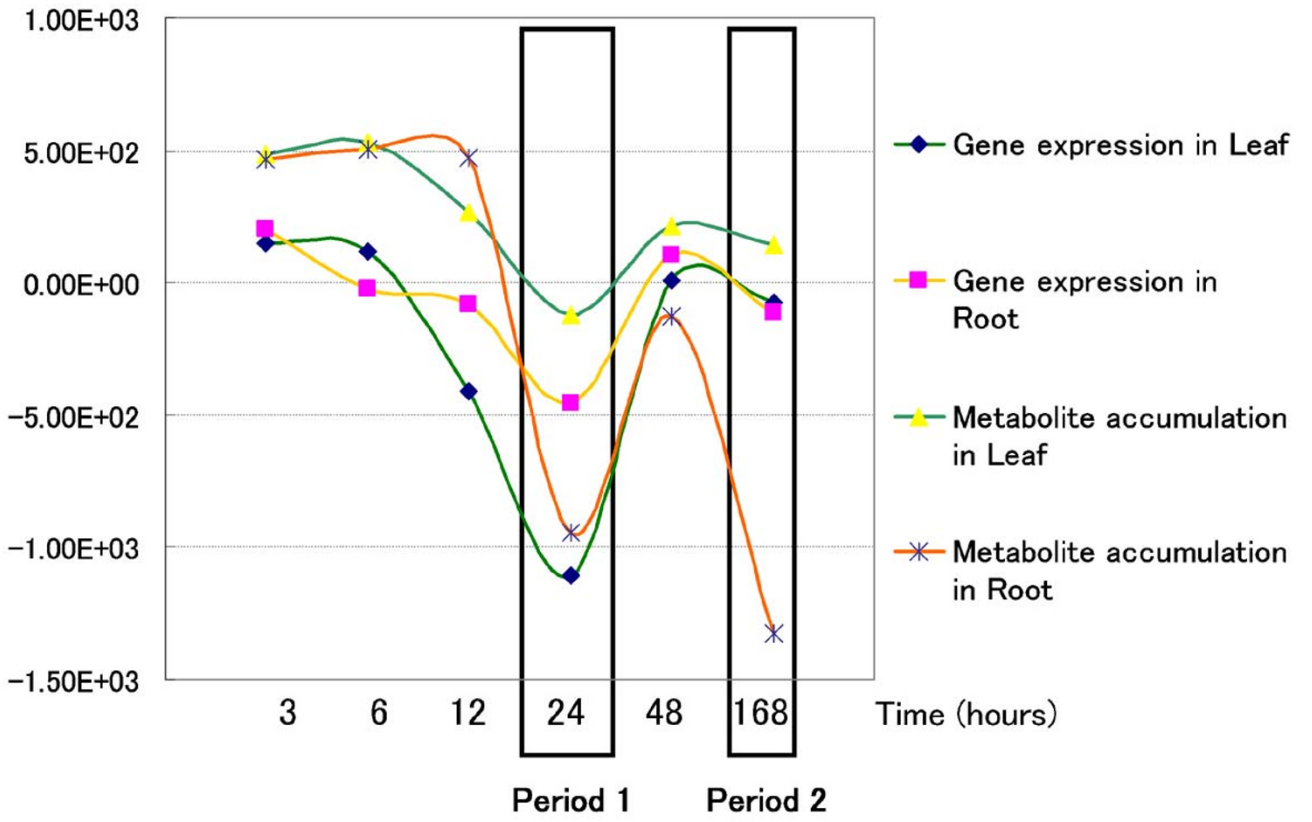

Figure 3
The result of an LDS-based calculation showing a transition in gene expression and metabolite accumulation. The ordinate scale indicates a log-likelihood value calculated by LDS. The transition in gene expression and metabolite accumulation in both leaf and root occurred most often in Period 1, showed by the left bold rectangle. During Period 2, shown by the right bold rectangle, a second transition in metabolite accumulation occurred solely in the root.

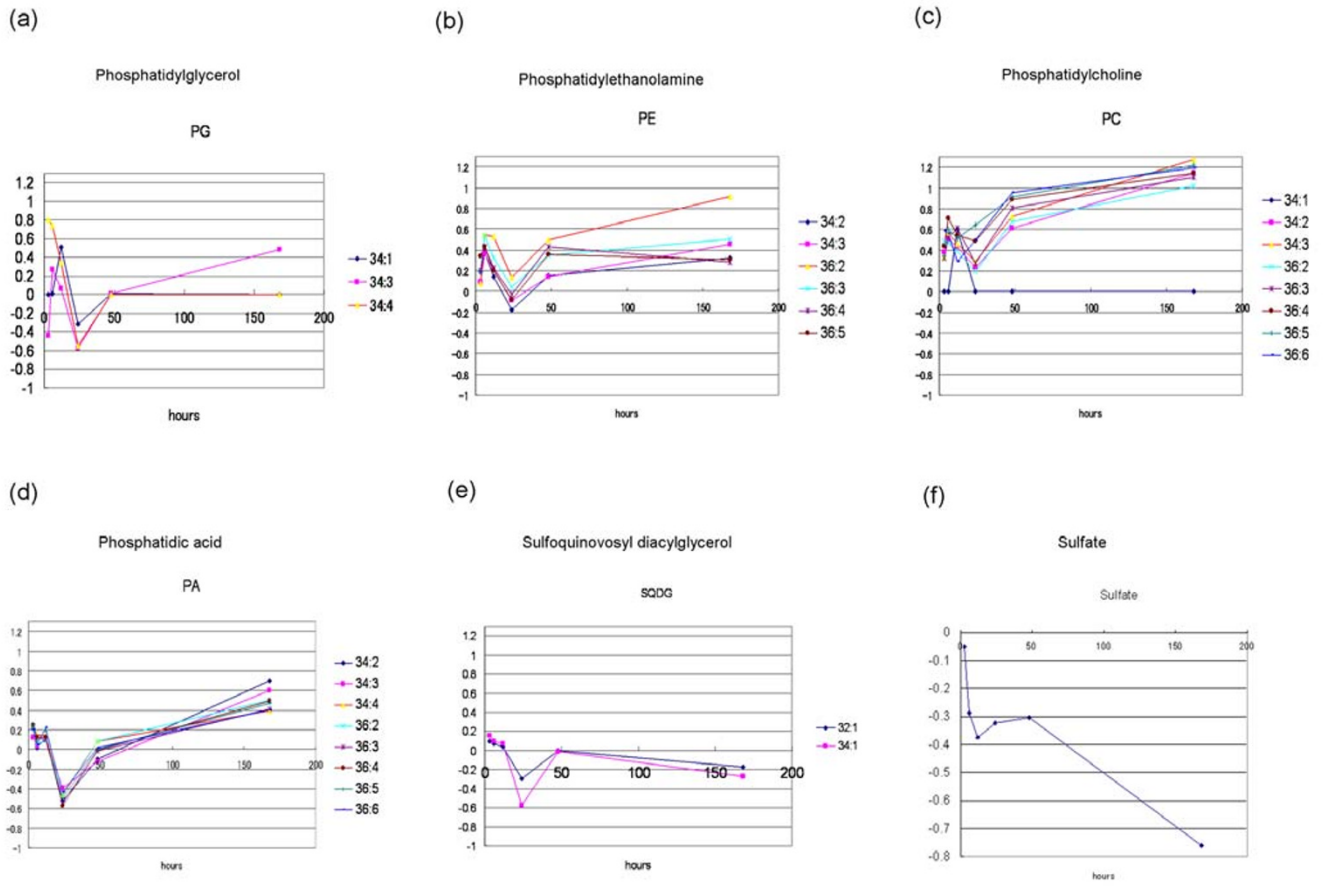

Figure 4

Lipid accumulation profiles. The accumulation profiles of metabolites whose m/z values corresponded to lipids expressing "total acyl carbon: total double bonds in two acyl groups", i.e., phosphatidylglycerol (a), phosphatidylethanolamine (b), phosphatidylcholine (c), phosphatidic acid (d), and sulfoquinovosyl diacylglycerol (e). the vertical axis shows normalized log-ratio values of the sulfur starvation condition to the control condition. (f) Sulfate profile analyzed by capillary electrophoresis. The vertical axis shows the log-ratio values of the sulfur starvation condition to the control condition.

control condition, the accumulation of sulfate was strongly repressed immediately after the shift to sulfur deficiency. Under sulfur deprivation, it was believed that sulfate levels (Figure 4f) would only decrease. During the transition period from 12 to 24 hr after the shift to sulfur deficiency, however, sulfate levels temporarily ceased declining and stayed relatively constant as compared to the control.

From these results, we hypothesize that sulfate is in an active form and is distributed throughout the plant at the transition time. During this period, in order to maintain the intracellular environment, membrane lipids are temporarily degraded and re-synthesized after the transition. This suggestion is consistent with the reported decrease in lipids under conditions of sulfur starvation [11].

Conclusion

In summary, by using a linear dynamical system, we have identified transition times in the adaptation processes of *Bacillus subtilis* and *Arabidopsis thaliana* to environmental stresses. By focusing on transition information based on a well-defined probabilistic index, we obtained novel observations on apoptosis in *Bacillus subtilis* and the regulation of lipid metabolism connected with sulfur-stress responses in *Arabidopsis thaliana*. As this approach uses probabilistic values to detect the transitions, the results are objectively supported without the risk of misinterpretation due to experimental noise. The results of this approach will enable us to more effectively design experiments specifically tailored for functional identification of genes and metabolites. By obtaining time series data with higher temporal resolution around the transition time points, we can obtain more precise information on the

details of the responses. The strategy described here was successful in identifying a small number of candidate genes and metabolites, from the vast number of genes and metabolites in comprehensive "omics" databases.

Methods

Time series data of Bacillus subtilis

The *Bacillus subtilis* time series data used in the present study were obtained from microarray analysis of cells sampled from 8 different experimental conditions. The data were produced using a two-colour fluorescence cDNA microarray that included 3100 *Bacillus subtilis* genes. The LB medium was developed to maximize cellular growth, and cells grown in this medium represented the control, unstressed population. In the initial phase of culture, the cell number increases by binary division – this is called the exponential growth phase in contrast to the stationary phase where the cell number has reached equilibrium. Data were collected from cells grown at 37°C in LB medium; the total length of culture was 12 hr and sampling was performed at 8 time points. Other culture conditions were also used with the aim of inducing stress responses in the cells. Cells were grown in Minimum Glucose Medium (MGM) at 37°C for 13 hr and sampled at 8 time points. Glucose starvation (GS) was achieved by eliminating the sugar from MGM; the cells were cultured in this medium for 10 hr and were sampled at 5 time points. Phosphate starvation (PS) was achieved by eliminating phosphoric acid from the MGM; the cells were cultured in this medium for 11 hr, and were sampled at 6 time points. Some cells were grown in Competence Medium (CM), which increases the ability of the cells to ingest DNA from the external environment. The cells were grown in CM for 9 hr and were sampled at 5 time points. Some cells were grown in Competence-Sporulation medium (CSM) for 15 hr and were sampled at 13 time points. A second sporulation medium, Difco sporulation medium (DSM), was also used. Cells were grown in this medium for 12 hr and were sampled at 19 time points. We also used Difco Glucose Glutamine (DGG) medium in which glucose and glutamine have been added to DSM medium in order to inhibit sporulation. The cells were grown for 9 hr in DSM and were sampled at 6 time points.

Time series data of Arabidopsis thaliana

The *Arabidopsis thaliana* data used in the present study were obtained from DNA microarray experiments and by Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS), as previously described [10]. In brief, *Arabidopsis* was cultured in sulfur-sufficient control medium for three weeks, transferred to control or sulfur-deprived medium, and cultured for up to one more week. Rosette leaves and roots were harvested at 3, 6, 12, 24, 48 and 168 hr after transfer, and subjected to transcriptome and metabolome analyses [10].

Transcriptome data were obtained using the Agilent Arabidopsis 2 microarray (Agilent Technologies, Palo Alto, CA), which carries 21,500 Arabidopsis genes [10]. The data are available on ArrayExpress in EMBL-EBI [12]. Non-targeted metabolome data were obtained by FT-ICR-MS [10], which produces precise mass-to-charge ratio values (m/z values) of metabolites [13]. Metabolites were provisionally identified from their m/z values and the analytical conditions used for FT-ICR-MS.

Parameter estimation

The test distribution is defined as $q(X_{1:T}|Y_{1:T}, \theta)$ and is used to approximate the true posterior distribution. The Kullback-Leibler divergence takes the minimum value of 0 if the two distributions are equivalent.

In Equation (17), maximization of the free energy with respect to q and θ is equal to the calculation of the maximum likelihood estimate θ with respect to $Y_{1:T}$.

$$F[q_x, \theta] \equiv \log p(Y_{1:T}|\theta) - KL[q_x(X_{1:T}) \| p(X_{1:T}|Y_{1:T}, \theta)] = \int dX_{1:T} q_x(X_{1:T}) \log p(X_{1:T}, Y_{1:T}|\theta) - \int dX_{1:T} q_x(X_{1:T}) \log q_x(X_{1:T}) \quad (17)$$

The free energy is maximized using the Expectation-Maximization algorithm [14] consisting of the following steps:

Step 1. Parameter set θ is initialized.

Step 2. E-step(step 2.1) and M-step (step 2.2) are successively repeated until the free energy converges.

Step 2.1. E-step:

k is a repeat loop index. By fixing the parameter $\theta^{(k-1)}$, F in Equation (17) is maximized with respect to q.

According to Equation (17), the solution is

$$q_x^{(k)}(X_{1:T}) = p(X_{1:T}|Y_{1:T}, \theta^{(k-1)}) = p(\mathbf{x}_1|Y_{1:T}, \theta^{(k-1)}) \left\{ \prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, Y_{1:T}, \theta^{(k-1)}) \right\} \quad (18)$$

After the fixation of parameter θ, the calculations needed to calculate the value of Equation (18) are as follows:

When both the data $Y_{1:t-1}$ and the parameter of the prior distribution $p(x_t|Y_{1:t-1}, \theta)$ of x_t are given, the posterior distribution of x_t, given the data $Y_{1:t}$, is

$$p(\mathbf{x}_t | Y_{1:t}, \theta) = \frac{p(\mathbf{y}_t | \mathbf{x}_t, \theta) p(\mathbf{x}_t | Y_{1:t-1}, \theta)}{\int d\mathbf{x}_t p(\mathbf{y}_t | \mathbf{x}_t, \theta) p(\mathbf{x}_t | Y_{1:t-1}, \theta)} \quad (19)$$

Using (19), the prior distribution of x_{t+1}, given the data $Y_{1:t}$, is

$$p(x_{t+1} \mid Y_{1:t}, \theta) = \int d\mathbf{x}_t p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \theta) p(\mathbf{x}_t \mid Y_{1:T}, \theta) \quad (20)$$

By successively iterating Equations (19) and (20), $p(\mathbf{x}_t \mid Y_{1:T}, \theta)$ with arbitrary t is obtained. This repeating method is called the Kalman Filter [8].

If all data are given, the following joint probability is obtained:

$$p(\mathbf{x}_{t+1}, \mathbf{x}_t \mid Y_{1:T}, \theta) = p(\mathbf{x}_{t+1} \mid Y_{1:T}, \theta) \frac{p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \theta) p(\mathbf{x}_t \mid Y_{1:t}, \theta)}{\int d\mathbf{x}_t p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \theta) p(\mathbf{x}_t \mid Y_{1:t}, \theta)} \quad (21)$$

If the parameter of $P(x_{t+1}, \mid Y_{1:T}, \theta)$ is given, the following distribution is obtained:

$$p(\mathbf{x}_t \mid Y_{1:T}, \theta) = \int d\mathbf{x}_{t+1} p(\mathbf{x}_{t+1}, \mathbf{x}_t \mid Y_{1:T}, \theta) \quad (22)$$

By successively iterating Equations (21) and (22), $p(\mathbf{x}_{t+1}, \mathbf{x}_t | Y_{1:T}, \theta)$ and $p(\mathbf{x}_t | Y_{1:T}, \theta)$, which are necessary to calculate the value of Equation (18), are obtained.

This repeating method is called the Kalman smoother [15].

Using the Kalman smoother, the statistical values necessary for parameter estimation are obtained.

If $p(\mathbf{x}_t | Y_{1:t-1}, \theta)$ is given, the following likelihood is calculated:

$$p(\mathbf{y}_t | Y_{1:t-1}, \theta) = \int d\mathbf{x}_t p(\mathbf{y}_t | \mathbf{x}_t, \theta) p(\mathbf{x}_t | Y_{1:t-1}, \theta) \quad (23)$$

Using (23), the log-likelihood is calculated as

$$\log p(Y_{1:T} \mid \theta) = \sum_{t=1}^{T} \log p(\mathbf{y}_t \mid Y_{1:t-1}, \theta) \quad (24)$$

Step 2.2. M-step

In this step, the value of θ that will maximize F under the condition $q_x = q_x^{(k)}$ is calculated using Equation (25):

$$\theta^{(k)} = \max_{\theta} \left\{ \int dX_{1:T} q_{\mathrm{x}}^{(\mathrm{k})}(X_{1:T}) \log p(X_{1:T}, Y_{1:T} \mid \theta) \right\} \quad (25)$$

The objective function to be maximized is defined as

$$J(\theta) = \int dX_{1:T} q_x^{(k)}(X_{1:T}) \log p(X_{1:T}, Y_{1:T} | \theta) \quad (26)$$

which is obtained by the following equation:

$$\frac{\partial J(\theta)}{\partial \theta} = 0 \quad (27)$$

and the solution of parameter θ is calculated that maximizes F.

Parameter θ is then updated, and the process goes back to E-step.

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

RM designed the LDS method and carried out the computer simulations. SK designed the BL-SOM and carried out the computer experiments. MYH supplied the Arabidopsis dataset. MY analyzed FT-MS data with RM. NO supervised the *Bacillus* experiments. KS proposed and supervised the study. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by Grant-in-Aids for Scientific Research from Japan Society for the Promotion of Science. We would like to thank Dr. Kazuo Kobayashi, Graduate School of Biological Sciences, Nara Institute of Science and Technology, who provided the *Bacillus* datasets. We would like to thank Dr. Masanori Arita, RIKEN Plant Science Center, for critical comments.

References

1. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian network to analyze expression data.** *J Comput Biol* 2000, **7:**601-620.
2. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S: **Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection.** *Bioinformatics* 2003, **19:**227-236.
3. Toh H, Horimoto K: **Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling.** *Bioinformatics* 2003, **18:**287-297.
4. Akutsu T, Miyano S, Kuhara S: **Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.** *J Comput Biol* 2000, **7:**331-343.
5. Kim H, Lee JK, Park T: **Boolean network using the chi-square test for inferring large-scale gene regulatory networks.** *BMC Bioinformatics* 2007, **8:**37.
6. Dewey TG, Galas DJ: **Dynamic models of gene expression and classification.** *Funct Integr Genomics* 2001, **1:**269-278.
7. Kanaya S, Kinouchi M, Abe T, Kubo Y, Yamada Y, Nishi T, Mori H, Ikemura T: **Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome.** *Gene* 2001, **276:**89-99.
8. Kalman RE, Bucy RS: **New results in linear filtering and prediction theory.** *Trans ASME, J Basic Eng* 1961, **83D-1:**95-108.
9. Nugroho FA, Yamamoto H, Kobayashi Y, Sekiguchi J: **Characterization of a new sigma-K-dependent peptidoglycan hydrolase gene that plays a role in *Bacillus subtilis* mother cell lysis.** *J Bacteriol* 1999, **181:**6230-6237.
10. Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki H, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K: **Elucida-**

tion of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *J Biol Chem* 2005, **280:**25590-25595.
11. Nikiforova VJ, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R: **Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants.** *Plant Physiology* 2005, **138:**304-317.
12. **EMBL-EBI** [http://www.ebi.ac.uk/arrayexpress/index.html]
13. Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB: **Nontargeted metabolome analysis by use of fourier transform ion cyclotron mass spectrometry.** *Omics* 2002, **6:**217-234.
14. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM algorithm.** *J R Stat Soc* 1977, **39:**1-22.
15. Anderson BDO, Moore JB: **Optical Filtering.** NY: Prentice Hall; 1979.