Volume: 06 Number: 01 Category: 01 Page: 1-17 ISSN No: 1303-6319 Cilt: 06 Sayı: 01 Kategori: 01 Sayfa: 1-17

BAYESIAN INFERENCE OF THE COMPLEX MAPK PATHWAY UNDER THE STRUCTURAL DEPENDENCY

Vilda PURUTÇUOĞLU[*] Ernst WIT[**]

ABSTRACT

The MAPK pathway is one of the main signal transaction system in all eukaryotes which regulates the cellular growth control. Because of its vital role, the regulation of the pathway is conducted via many proteins, thereby constitutes a complex structure. In inference of this system via MCMC techniques based on the Euler approximation, we have observed that there are many proteins which indicate high structural dependencies on other proteins and these species have caused singular diffusion matrices, hereby resulted in infeasible acceptance probabilities. Therefore, we have discarded these problematic substrates at the beginning of the inference and estimated the parameters by using merely linearly independent species in the system. However in that case, the accuracy of the estimation has been highly affected by the underlying exclusion, particularly, when the number of dependent species was big. The elimination of those proteins has led to a significant rise in the number of current missing components in MCMC. In this study, we implicitly include these proteins in our computation via an alternative approach which simulates dependent terms as a linear combination of linearly independent species. In that way, we can add the effect of dependent species in the calculation of acceptance probabilities of reaction rates and states. From the analysis, we conclude that the highlighted innovation decreases the average error of estimates and suggests less computational cost in inference of the MAPK pathway.

Keywords: Bayesian inference, Diffusion approximation, MAPK pathway.

1. INTRODUCTION

All cellular activations are regulated by various signal transduction pathways. The MAPK (mitogen-activated protein kinase) pathway is one of the main pathway structure which regulates the growth control in all eukaryotes, i.e. the organisms whose cells contain a nucleus, thereby it is the system of interest, particularly, in oncogene researches (Kolch, 2005; Orton et al., 2005).

Coming from the importance of the pathway in the cellular life cycle from the cell proliferation, i.e. the reproduction of the cell, to the apoptosis, i.e. the cell death, the activation of the MAPK pathway uses a number of proteins whose main components are Ras, Raf, MEK, and ERK proteins (Figure 1). This activation begins by an external stimulus which causes the binding of the signal to the Epidermal Growth Factor (EGF) receptor and is ended up by the production of the target c-Fos gene after a sequence of recruitments, phosphorylations, and inhibitions.

[*] Dr., Middle East Technical University, Faculty of Art and Science, Department of Statistics, e.mail: vpurutcu@metu.edu.tr

[**] Prof. Dr., University of Groningen, Institute of Mathematics and Computing Science, e.mail: e.c.wit@rug.nl

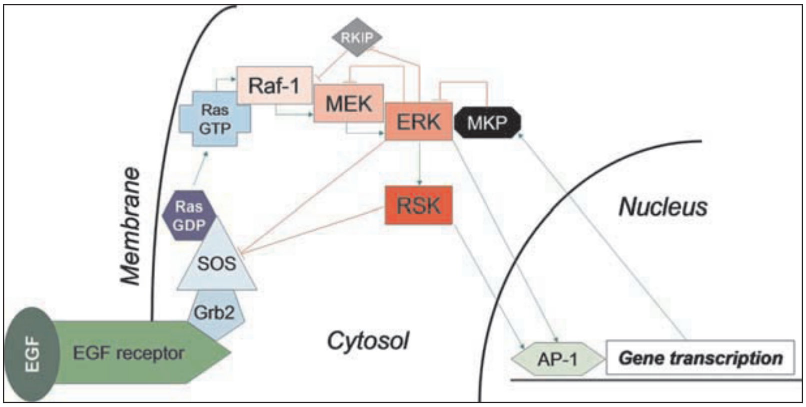

Figure 1. Main components of the MAPK pathway (Kolch, 2005)

In this study, we estimate the stochastic rate constants of quasi reactions of the MAPK pathway which is described by 51 proteins and 66 reactions (Purutçuoğlu and Wit, 2006 and 2008b). In the inference of the model parameters, i.e. reaction rates, from a simulated dataset, we implement the discretized version of the diffusion approximation known as the Euler-Maruyama approximation (Eraker, 2001; Golightly and Wilkinson, 2005).

In the estimation via the Euler technique, we overcome the problems of the missing data and sparse measurements, which are typical challenges in complex systems, by using the MCMC (Markov Chain Monte Carlo) framework. Accordingly we choose the Metropolis-within-Gibbs algorithm with the data augmentation technique (Golightly and Wilkinson, 2005; Purutçuoğlu and Wit, 2008a and 2008b) for the computation. From our previous analysis (Purutçuoğlu and Wit, 2008a and 2008b), we have seen that although the underlying MCMC methods are promising to estimate the reaction rates, the dependency between proteins causes singular diffusion matrices in implementations. Therefore, we have eliminated the proteins which lead to singularities and the algorithms have run by merely linearly independent terms. However, when the total number of excluded species became bigger, the estimation had to be conducted under a large number of missing information. In this study, to unravel the challenges caused by those large missing data, we develop an innovation in the current scheme such that the new plan uses these problematic substrates in the estimation.

We present a brief explanation about biochemical reactions and the diffusion approximation in Section 2. The details of MCMC updates and the new plan are given in Section 3.1 and Section 3.2, respectively. We evaluate the performance of the algorithm in Section 4 by comparing our outcomes with previous findings. Finally Section 5 concludes the results and discusses possible extensions.

2. BIOCHEMICAL PROCESS AND STOCHASTIC MODELLING

A biochemical reaction is a quantitative and qualitative description of a biochemical process. If we have *r* number of equations which explain a biochemical activation, this set of reactions presents a system. A simple biochemical system can be described as the following:

$$
\begin{array}{ll}
\text{Reaction 1: } & m_{11}Y_1 + m_{12}Y_2 + \ldots + m_{1n}Y_n \xrightarrow{c_1} s_{12}Y_1 + s_{12}Y_2 + \ldots + s_{1n}Y_n \\
\text{Reaction 2: } & m_{21}Y_1 + m_{22}Y_2 + \ldots + m_{2n}Y_n \xrightarrow{c_2} s_{22}Y_1 + s_{22}Y_2 + \ldots + s_{2n}Y_n \\
 & \qquad\qquad \vdots \quad \vdots \quad \vdots \quad \vdots \\
\text{Reaction } r\text{: } & m_{r1}Y_1 + m_{r2}Y_2 + \ldots + m_{rn}Y_n \xrightarrow{c_r} s_{r2}Y_1 + s_{r2}Y_2 + \ldots + s_{rn}Y_n
\end{array} \qquad (1)
$$

In that expression, $Y = (Y_1,...,Y_n)$ denotes the n-dimensional vector of current states of the system and n indicates the total number of species. The coefficients m_{ji} and s_{ji} display the stoichiometric coefficients associated with the ith reactant of the jth reaction and the ith product of the jth reaction, respectively, for $i = 1,...,n$ and $j = 1,...,r$. Finally c_j is the reaction rate constant which denotes the speed of the reaction dependent on the temperature of the system and physical properties of reactants.

Equation (1) can be also shown by a matrix form such that $MY \rightarrow SY$ where,

$$
M = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \vdots & \vdots \\ m_{r1} & \cdots & m_{rn} \end{bmatrix} \quad \text{and } S = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \vdots & \vdots \\ s_{r1} & \cdots & s_{rn} \end{bmatrix}
$$

are the (rxn) - dimensional matrix of stoichiometries of reactants and the (rxn) - dimensional matrix of stoichiometries of products, respectively. The interpretation of this description is that when the rth reaction occurs, the number of molecules of Y_i $(i = 1,...,n)$ decreases by m_{ri} and increases by s_{ri} amount. As a result the molecular transfer causes a net change in the system with $V_{ri} = s_{ri} - m_{ri}$ where $V = S - M$ is called the (rxn) - dimensional matrix of net effects and V_{ji} is the corresponding net chance of the ith species after the execution of the rth reaction. More details about the formulation of biochemical processes and the network structure can be found in Wilkinson (2006) and Bower and Bolouri (2001). On the other side, the implementation of this description in a prokaryotic autoregulation gene network and in the MAPK pathway are given in Golightly and Wilkinson (2005) and Purutçuoğlu and Wit (2008b), respectively.

There are several approaches in order to capture the stochastic behaviour of the biochemical system (Gillespie, 1977; Gibson and Bruck, 2000; Turner et al., 2004). The Gillespie algorithm (Gillespie, 1977; Gillespie, 1992) is the most common exact method to simulate a biochemical network, whereas, it is computationally inefficient in inference of the realistic complexity (Golightly and Wilkinson, 2005; Wilkinson, 2006; Boys et al., 2008). The diffusion approximation is an efficient technique as an alternative estimation in place of Gillespie (Golightly and Wilkinson, 2005). In this research, we use the discretized version of the diffusion approximation, known as the Euler-Maruyama approximation, since the observed measurements are collected in discrete time. The Euler method explaines the change of states at time t by the following equation.

$$
\Delta Y_t = \mu(Y_t, \theta)\Delta t + \beta^{\frac{1}{2}}(Y_t, \theta)\Delta W_t \qquad (2)
$$

here ΔY_t stands for the change in state $Y=(Y_1,Y_2...,Y_n)$ at time t to $[t+\Delta t]$. $\theta=(c_1,c_2,...,c_r)$ represents the parameter vector while n and r are the total number of substrates and the total number of reactions in the system, respectively, as mentioned beforehand. $\mu(Y_t,\theta)$ displays an n-dimensional mean or drift vector and is computed by $\mu(Y_t,\theta)=V'h(Y_t,\theta)$. On the other hand, $\beta(Y_t,\theta)$ shows an (nxn) diffusion or variance matrix and is found via $\beta(Y_t,\theta)=V'diag\{h(Y_t,\theta)\}V$. Both μ and β terms are the functions of Y and θ, and are calculated from the hazard $h(Y_t,\theta)$ as well as the net effect matrix V in which V' implies the transpose of V. $diag\{h(Y_t,\theta)\}$ in β is an (rxr) dimensional matrix whose diagonal terms set to $h(Y_t,\theta)$ and off-diagonals are equated to zero (Wilkinson, 2006). Finally ΔW_t denotes an n-dimensional independent identically distributed Brownian random vector generated from the normal distribution with mean zero and covariance-variance as the product of the identity matrix I and the discrete time interval Δt, i.e. $\Delta W_t \sim N(0,I\Delta t)$.

3. INFERENCE OF THE SYSTEM

In the inference of the reaction rates, we consider that the observation matrix Y is composed of both observed and unobserved measurements as used in the studies of Eraker (2001); Golightly and Wilkinson (2005). We denote observed and unobserved terms by n-dimensional X and Z vectors, respectively. Moreover in order to get more precise estimates from the Euler, we use the data augmentation by putting latent states within each pair of time-course measurements. More details about the implementation of the data augmentation can be found in Roberts and Stramer (2001) and Elerian et al. (2001). So every time state of the system Y_i ($i=1,...,T$), where $i=1$ indicates the initial time point and $i=T$ is the final time point after the data augmentation, is presented as $Y_i \equiv (X_i,Z_i)'$. Here $(A)'$ stands for the transpose of any vector (A). If the state has observed measurements, then X_i is set to x_i, which means the observed data by observed components.

In the update of the system via MCMC techniques we implement the Gibbs sampling seeing that the number of unobservable values, i.e. the number of reaction rates and missing data, are large. However as the dimension of the system for every time point is high and each state Y_i is updated via a different Gibbs sampler given the previous Y_{i-1} and the next Y_{i+1} state, we use the Metropolis-within-Gibbs (M-W-G) algorithm. Accordingly the candidate value for the ith state Y_i^* is proposed from the following multivariate normal distribution N. In this expression $\beta(Y_{i-1},\theta)$ displays the diffusion matrix of the previous state Y_{i-1} for the given θ.

$$Y_i^* \sim N\left(\frac{1}{2}(Y_{i-1}+Y_{i+1}),\quad \frac{1}{2}\Delta t\beta(Y_{i-1},\theta)\right) \tag{3}$$

Eraker (2001) shows that the transiton kernel, $q(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$, formulated in Equation (3) converges to the true distribution of Y_i, $\pi(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$, when $\Delta t \to 0$. If the state has additional observed measurements x_i, we consider to generate merely the candidate Z_i, Z_i^*, by further conditioning Y_i^* on $X_i = x_i$ since each Y_i^* can be decomposed as

$$Y_i^* \equiv \begin{pmatrix} X_i \\ Z_i^* \end{pmatrix} \tag{4}$$

Then for each state, the acceptance probability is computed for the candidate Y_i^* by

$$\alpha(Y_i^* \mid Y_i) = \min\left\{1, \quad \frac{p(Y_i^* \mid Y_{i-1}, Y_{i+1}, \theta) q(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)}{p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta) q(Y_i^* \mid Y_{i-1}, Y_{i+1}, \theta)}\right\} \tag{5}$$

where

$$\begin{aligned} p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta) = &\mid \beta(Y_{i-1}, \theta)^{-1} \mid^{\frac{1}{2}} \mid \beta(Y_i, \theta)^{-1} \mid^{\frac{1}{2}} \\ &\exp\left\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)'(\Delta t \beta(Y_{i-1}, \theta))^{-1}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)\right\} \\ &\exp\left\{-\frac{1}{2}(Y_{i+1} - Y_i - \mu(Y_i, \theta)\Delta t)'(\Delta t \beta(Y_i, \theta))^{-1}(Y_{i+1} - Y_i - \mu(Y_i, \theta)\Delta t)\right\} \end{aligned} \tag{6}$$

Here $p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$ is directly proportional to $\pi(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$. More details about candidate generators and associated acceptance probabilities can be found in Golightly and Wilkinson (2005).

Once the updates of missing states are completed, the system executes the updates of reaction rates by the random walk algorithm. In this method the candidate rates are generated from the normal distribution and the acceptance probability is calculated by

$$\alpha(\theta, \theta^* \mid Y) = \min\left\{1, \frac{L(\theta^* \mid Y)}{L(\theta \mid Y)}\right\} \tag{7}$$

in which

$$L(\theta \mid Y) = \prod_{i=1}^{T} \pi(\theta) f(Y_i \mid Y_{i-1}, \theta) \tag{8}$$

In Equation (7), θ^* indicates the proposal rates which are produced via $\theta_j^* = \theta_j + \varphi_j$ ($j = 1, ..., r$) where $\varphi_j \sim N(0, \delta_j)$. The variance of each rate δ_j is called the "tuning parameter" and significantly affects the mixing property of the algorithm (Golightly and Wilkinson, 2005). For a good mixing in univariate random walk chains it is suggested that an acceptance ratio *p* of around 24% is optimal (Roberts et al., 1997). On the other hand for the multivariate inference, the optimal *p* is found as 0.574 (Roberts and Rosenthal, 1998). However, when the complexity of the network structure increases, very low ratios such as 5% can be tolerable since it is difficult to produce a candidate value for particular reaction rates. Thus, in our estimation to get a sensible value for the variance of each rate δ_j, we define δ_j adaptively during the burn-in period of MCMC runs. We multiply δ_j by 1.1 if the acceptance ratio p at every 100th iteration in the burn-in is greater than 60% and we divide δ_j by 1.1 if p is less than 5%. Whereas if p lies between 5% and 60%, we keep the current δ_j. At the end of the burn-in, the final set of δ's is taken as constants and used until the end of the inference.

Indeed, apart from these highlighted optimal acceptance ratios, there are a number of other methods which can assess the convergence of the chain. For instance the sample autocorrelation function (Golightly and Wilkinson, 2005 and 2006b) and the posterior density of each parameter (Gelman et al., 2004), the potential scale reduction (Gelman et al., 2004), and the value of the convergence diagnostic (Geweke, 1992) are some of the methods used for monitoring the convergence. In Section 4 to control the convergence of our estimates, we choose the autocorrelation function and the posterior density besides the evaluation of results via acceptance ratios.

On the other hand $\pi(\theta)$ in Equation (8) shows the prior distribution of reaction rates which is taken as exponential with rate 1 seeing that it satisfies the positivity condition of our model parameters and $f(Y_i \mid Y_{i-1}, \theta)$ displays the transition density of the ith state given the previous state and reaction rates. Therefore $L(\theta \mid Y)$ in Equation (8) can be formulated as

$$L(\theta \mid Y) = \prod_{i=1}^{T} \exp\left\{-\sum_{k=1}^{r} \theta_k\right\} \mid \beta(Y_{i-1},\theta) \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1},\theta)\Delta t)'(\Delta t \beta(Y_{i-1},\theta))^{-1} (Y_i - Y_{i-1} - \mu(Y_{i-1},\theta)\Delta t)\right\} \qquad (9)$$

Further discussion about the updates of rates can be found in Purutçuoğlu and Wit (2008b).

3.1 MCMC Steps for the MAPK Pathway

In the update of the system via MCMC algorithms and data augmentation schemes, we observe that the singularity of diffusion matrices and the dependency between substrates are the main challenges. In order to unravel the first problem, we suggest to work with only nonsingular matrices. Therefore, in every stage of updates we check the corresponding candidate diffusion terms whether they would cause singularities in the system if they are accepted. If the candidate generator leads to singular diffusions, then we simply reject the candidate states and rates without computing the acceptance probability α. Otherwise, α is calculated in order to decide on the next step. We call this kind of dependency the "incidental dependence" (Purutçuoğlu and Wit, 2008b). The second problem, on the other side, is originated from the dependency of V matrix where V is the (rxn) - dimensional net effect matrix. In that case, the substrates are dependent on each other from the description of the system since V is directly produced from the quasi list of systems' reactions. We name this dependency as the "structural dependence". In the update of the system under the structural dependence, we can exclude the problematic species at the beginning of MCMC algorithms (Purutçuoğlu and Wit, 2008b). Because any linear dependence in this matrix, *V*, affects the rank of $V'V$, which is used in the computation of $\beta(Y_t, \theta)$ as stated in Section 2 and leads to singular diffusions. On the other hand, this elimination implies that we lose some of the observed *X* and unobserved *Z* components. Thus the exclusion can rise the average errors of estimates as the number of dependent substrates increases. For the MAPK pathway the number of structurally dependent proteins is 17 over 51 proteins, which correspond to a large proportion (around 33%) with respect to the total number of substrates in the system.

As an alternative approach for the elimination of these species, we consider to include them implicitly in our computation. We suggest that if we preserve the linear relationships between dependent and independent proteins, these relations can be used to generate dependent substrates after the updates of linearly independent terms. Then if these values are added in the calculations of diffusion and drift terms, then the system can be updated under both incidental and structural dependencies. Indeed we have implemented this idea in the simulation of the complex MAPK pathway via the diffusion approximation and we have observed that the method successfully deals with the singularity, which is particularly seen in the steady-state phase of the simulation (Purutçuoğlu and Wit, 2006). Therefore, we develop a new updating scheme for the inference which is based on that implicit computation. We list the steps of the underlying plan as follows:

1. The system is initialized by assigning values for missing states and reaction rates and the counter of iterations *g* is set to zero.
2. After the initialization, all *n* columns of *V* are checked from left to right whether there is any linearly dependent column, denoted as s, which indicates structurally dependent species. For simplicity we assume that we have totally |*s*| dependent columns, thereby (*n*-|*s*|) independent proteins. Then for each dependent species, the vector, which displays the coefficients of the linear relationship between dependent and independent substrates, λ_{jl} ($j \in s$ and $l = 1,\ldots,n- \mid s \mid$), is preserved.
3. The system begins the updates from the states, whose substrates are linearly independent, Y_i^{indep} ($i = 1,\ldots,T$). The candidate value of Y_i^{indep}, Y_i^{indep*}, is generated from the multivariate normal distribution given in Equation (3) and Equation (4). If the proposal state maintains the singularity of the candidate diffusion matrix β_i^*, that is the incidental dependence is not observed in β_i^*, as well as the positivity of the state is satisfied, then it is accepted as the generator for the linearly dependent proteins. Here as the candidate generators (Equation (3) and Equation (4)) are used for the linearly independent terms, which do not indicate neither the incidental nor structural dependence, the transiton kernels given in the study of Eraker (2001) and performed in our research still maintain the convergent properties to the true distribution. Indeed, from our reference study of Golightly and Wilkinson (2005), we also observe a structural dependence in the net effect matrix of a small prokaryotic autoregulation system. In order to unravel the singularity of the diffusion term, that particular dependent subsrate is excluded from the beginning of the estimation and the generators are produced from the remaining linearly independent species. With respect to that system of interest, our network is significantly complex, accordingly, the dependency is observed very often. Although we believe that

the missing data and the underlying high dependency between species can lead to biased estimates, the problems of inaccuracies of estimates can be improved by alternative approaches. More details about the possible solutions of the problems by using the same transition kernels of Eraker (2001) can be found in Section 5 and Purutçuoğlu and Wit (2008b). On the other hand, other alternative solutions to decline the dependency on the estimates can be seen in the studies of Golightly and Wilkinson (2006a and 2006b). In those works, basically, they suggest to update the missing data in block of random size and to implement the method of particle filterings.

4. To produce totally |*s*| linearly dependent species, initially (*n*-|*s*|) increments ψ are generated from the Brownian motion, i.e. the normal distribution with mean zero and variance Δt. These increments are multiplied by the square root of the diffusion term obtained from the previous time step $\beta_{i-1}^{1/2}$ of linearly independent species. Therefore, our computed $\beta_{i-1}^{1/2}$ matrix has the dimension of (*n*-|*s*|)x(*n*-|*s*|). In that way, we get the error term $\varepsilon = \psi \beta_{i-1}^{1/2}$ for linearly independent substrates which corresponds to $\beta^{1/2}(Y_t, \theta)\Delta Wt$ in Equation (2). Then the change in the state of new dependent substrates from $i = t$ to $i = t + \Delta t$ is simulated via $\Delta Y_i^{dep*} = \mu(Y_{i-1}^{indep}, \theta)\Delta t + \varepsilon$ similar to Equation (2) in which $\mu(Y_{i-1}^{indep}, \theta)$ refers to the (*n*-|*s*|)-dimensional drift vector of the previous state whose substrates are linearly independent. Hence, ΔY_i^{dep*} gives an (*n*-|*s*|) - dimensional vector. Accordingly the candidate state for dependent species, Y_i^{dep*}, is generated as $\Lambda^{dep*} = \sum_{l \notin s, l<j}^{n-|s|} \Delta Y_l^{dep^*} \lambda_{jl}$ and $Y_i^{dep*} = Y_{i-1}^{dep} + \Lambda^{dep*}$ when $j \notin s$, $l \in s$, and $l = 1, \ldots, n - |s|$. Λ^{dep*} corresponds to an |*s*|-dimensional vector and represents the change in the state Y_i^{dep*} that is computed by the linear relation within dependent and independent proteins. On the other hand, Y_{i-1}^{dep} stands for the updated state Y at time $t = i - 1$, whose proteins are linearly dependent. Finally, a complete proposal state Y_i^* is produced by combining Y_i^{indep*} with Y_i^{dep*} as a vectoral form.

5. The drift μ_i and the diffusion β_i of the updated state are computed from the hazard function of each reaction based on Y_i^*, i.e. $h(Y_i^*, \theta)$. If we do not observe a new inner dependence between linearly dependent substrates from the computation of the recent β_i, in other words, if we do not write any of the linearly dependent substrate in terms of other linearly dependent substrates, then the acceptance probability $\alpha(Y_i^* \mid Y_i)$ is calculated by (*n*x*n*) - dimensional diffusion matrices of Y_{i-1} and Y_i^*. Y_{i-1} shows the updated state at time $t = i - 1$. Otherwise, $\alpha(Y_i^* \mid Y_i)$ is found from only linearly independent species. For the MAPK pathway, since we riddle with an inner linear dependence within linearly dependent substrates, α is derived from lower dimensional diffusion matrices whose components are linearly independent proteins. Whereas the computation of hazards is executed on both dependent and independent species as described beforehand.

6. From the result of $\alpha(Y_i^* \mid Y_i)$, if the move is accepted, $Y_i^{(g)} = Y_i^*$ at the *g*th iteration. Otherwise, the system preserves the current state. Then we return to Step 3 to update the $(i+1)$th state by M-W-G algorithm and repeat the process until $i = T - 1$. In the final column, i.e. when $i = T$, we perform the Gibbs sampling in place of M-W-G and directly accept the proposal state Y_T^* without computing $\alpha(Y_T^* \mid Y_T)$.
7. The model parameters of the system, i.e. reaction rates, are updated via the random walk algorithm by *d*-dimensional blocks. The *d*-number of deviance terms is generated from the normal distribution with mean zero and variance δ_j ($j = 1,\ldots,r$) and is added to the current θ to produce a candidate θ, θ^*. The new θ^* for each *d*-dimensional group is controlled whether it causes a new source of incidental dependences when it is used in the diffusions of Y_i ($i = 1,\ldots,T$). If θ^*does not lead to any singularity, the acceptance probability given in Equation (7), $\alpha(\theta,\theta^* \mid Y)$, is computed. If the candidate reaction rates increase the likelihood, the move is accepted and $\theta^{(g)} = \theta^*$ at the *g*th iteration, otherwise, the chain does not move. On the other hand, if θ^* results in an incidental dependence, then a new θ^* is proposed until the nonsingularity of all diffusion terms is satisfied for every state.
8. When all states and reaction rates are updated, the counter of the algorithm goes from g to $(g+1)$. The algorithm is repeated from Step 2 until the system converges to the stationary distribution.

4. APPLICATION OF THE METHOD

In order to evaluate the performance of MCMC algorithms, we use a simulated dataset which we previously applied in our analysis (Purutçuoğlu and Wit, 2008b). This dataset is generated from the Gillespie algorithm and has 28 observed and linearly independent substrates, and 23 unobserved substrates in which 6 of them are linearly independent and the remaining 17 terms are dependent species. We choose 50 time points from the underlying data and accept that these are our time-course measurements. Then we extent the dataset by adding 3 augmented states between each pair of 50 time points. Therefore, we generate an observation matrix *Y* which has 197 instead of 50 columns, i.e. $i = 1,\ldots,197$. The complete list of observed and unobserved substrates and more details about the simulated data can be found in Purutçuoğlu and Wit (2008b).

In this study, all the computational work is carried out in the programme language R and our codes are executed on Dual Core Xeon 3.00 GHz processor. To estimate the reaction rates of the MAPK pathway, we iterate the algorithm 200,000 times and take the mean of the last 50,000 MCMC outputs as the estimated values of our model parameters. The lists of estimated rates with true values are presented in Table 1 and Table 2. The first table shows the results from the new algorithm and the second one illustrates the outputs obtained by MCMC algorithms which are conducted by merely linearly independent substrates. From both tables, it is found that most of the acceptance ratios of estimated values lie between 0.05 and 0.60 which display good mixing properties in the inference. Figure 2 and Figure 3 are drawn as an example from the posterior distributions of selected reaction rate constants and their autocorrelation functions after the burn-in. From the figures it is seen that the selected parameters indicate convergent distributions supporting their acceptance ratios given in Table 1 and Table 2. But the new plan typically offers lower acceptance ratios than the old plan produces. On the other side, from the comparison of the average error of each estimate calculated by the following Equation (10),

$$\text{Average error} = |\text{True value-Estimated value}| \,/\, \text{True value} \tag{10}$$

we observe that the new algorithm considerably decreases the error (Table 3). Moreover, from the evaluation of the CPU (Central Processing Unit) time, it is seen that the new scheme also offers a less computational cost (Table 3). Indeed, with respect to the complexity of algorithms, the new scheme has more computational steps, thereby it is expected that this scheme should be computationally more demanding. From our results although, at first sense it seems to be a contradiction, we explain this situation as follows: As stated in Section 3.1, the MAPK pathway can use the dependent substrates solely in the calculation of hazards functions, rather than during the calculation of acceptance probabilities of both rates and states. Hence, the complete computation of dependent substrates according to the new plan cannot be performed in our system. In other words, the steps of both the new and previous algorithms are run for $(n\text{-}|s|)$ terms

Table 1. Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants found by the MCMC plan which includes structurally dependent substrates

Reaction	True rate	μ	σ	*p*	Reaction	True rate	μ	σ	*p*
c_2	0.010	0.020	0.000	0.291	c_{35}	0.010	4.853	0.255	0.450
c_3	0.010	0.051	0.007	0.303	c_{36}	0.010	0.235	0.002	0.468
c_4	0.010	0.130	0.002	0.271	c_{37}	0.010	0.576	0.004	0.510
c_5	1.000	0.596	0.007	0.294	c_{38}	1.000	0.130	0.002	0.476
c_6	1.000	0.996	0.001	0.029	c_{39}	1.000	0.130	0.002	0.456
c_7	1.000	1.001	0.001	0.021	c_{40}	1.000	0.015	0.000	0.499
c_8	1.000	1.028	0.001	0.021	c_{41}	1.000	0.001	0.000	0.014
c_9	0.010	0.001	0.000	0.029	c_{42}	0.010	0.000	0.000	0.014
c_{10}	0.010	0.000	0.000	0.029	c_{43}	0.010	0.252	0.005	0.014
c_{11}	1.000	2.761	0.060	0.548	c_{44}	1.000	0.257	0.001	0.014
c_{12}	0.015	1.619	0.077	0.554	c_{45}	0.015	0.354	0.006	0.014
c_{13}	0.010	0.060	0.001	0.574	c_{46}	0.010	0.002	0.000	0.058
c_{14}	0.010	0.082	0.001	0.595	c_{47}	0.010	0.024	0.006	0.058
c_{15}	0.010	0.083	0.001	0.613	c_{48}	0.010	0.319	0.040	0.058
c_{16}	0.010	4.456	0.117	0.820	c_{49}	0.010	0.119	0.039	0.058
c_{17}	1.000	0.294	0.004	0.776	c_{50}	1.000	0.001	0.000	0.058
c_{18}	0.010	5.404	0.175	0.848	c_{51}	0.010	3.643	0.100	0.782
c_{19}	1.000	0.337	0.006	0.817	c_{52}	1.000	0.126	0.001	0.533
c_{20}	1.000	3.334	0.224	0.866	c_{53}	1.000	0.097	0.001	0.821
c_{21}	0.010	0.041	0.004	0.426	c_{54}	0.010	0.078	0.003	0.850
c_{22}	0.010	4.913	0.235	0.420	c_{55}	0.010	3.875	0.130	0.801
c_{23}	0.015	1.264	0.007	0.283	c_{56}	0.015	0.013	0.000	0.019
c_{24}	0.010	0.010	0.000	0.390	c_{57}	0.010	0.000	0.000	0.020
c_{25}	0.010	0.066	0.001	0.420	c_{58}	0.010	0.004	0.000	0.019
c_{26}	0.010	0.003	0.000	0.345	c_{59}	0.010	0.637	0.007	0.019
c_{27}	0.010	0.416	0.002	0.324	c_{60}	0.010	0.000	0.000	0.020
c_{28}	0.010	0.058	0.001	0.355	c_{61}	0.010	0.000	0.000	0.428
c_{29}	0.010	0.090	0.001	0.353	c_{62}	0.010	0.004	0.000	0.422
c_{30}	0.010	0.016	0.000	0.333	c_{63}	0.010	1.028	0.011	0.317
c_{31}	0.010	0.014	0.010	0.461	c_{64}	0.010	0.705	0.034	0.420
c_{32}	0.010	0.019	0.000	0.415	c_{65}	0.010	0.418	0.006	0.404
c_{33}	1.000	0.223	0.003	0.415	c_{66}	1.000	9.448	0.484	0.775

Table 2. Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants found by the MCMC plan which excludes structurally dependent substrates

Reaction	True rate	μ	σ	p	Reaction	True rate	μ	σ	p
c_2	0.010	0.050	0.001	0.540	c_{35}	0.010	2.465	0.096	0.489
c_3	0.010	0.040	0.001	0.535	c_{36}	0.010	0.269	0.004	0.457
c_4	0.010	0.031	0.001	0.547	c_{37}	0.010	0.422	0.006	0.478
c_5	1.000	5.228	0.187	0.541	c_{38}	1.000	0.720	0.019	0.492
c_6	1.000	1.350	0.034	0.247	c_{39}	1.000	1.011	0.015	0.410
c_7	1.000	1.245	0.020	0.232	c_{40}	1.000	0.019	0.000	0.460
c_8	1.000	0.956	0.017	0.265	c_{41}	1.000	0.002	0.000	0.153
c_9	0.010	0.252	0.005	0.257	c_{42}	0.010	0.002	0.004	0.156
c_{10}	0.010	0.000	0.000	0.273	c_{43}	0.010	0.008	0.005	0.156
c_{11}	1.000	2.165	0.040	0.532	c_{44}	1.000	0.483	0.003	0.149
c_{12}	0.015	1.120	0.045	0.544	c_{45}	0.015	1.053	0.007	0.125
c_{13}	0.010	4.719	0.168	0.597	c_{46}	0.010	0.890	0.034	0.779
c_{14}	0.010	0.040	0.000	0.582	c_{47}	0.010	0.052	0.002	0.777
c_{15}	0.010	0.056	0.001	0.593	c_{48}	0.010	9.416	0.205	0.761
c_{16}	0.010	3.723	0.140	0.807	c_{49}	0.010	5.577	0.308	0.773
c_{17}	1.000	0.266	0.004	0.730	c_{50}	1.000	1.155	0.016	0.572
c_{18}	0.010	3.589	0.139	0.816	c_{51}	0.010	7.070	0.280	0.803
c_{19}	1.000	0.288	0.004	0.773	c_{52}	1.000	0.160	0.006	0.671
c_{20}	1.000	1.789	0.090	0.790	c_{53}	1.000	0.175	0.003	0.673
c_{21}	0.010	0.003	0.002	0.437	c_{54}	0.010	0.217	0.005	0.738
c_{22}	0.010	1.317	0.044	0.381	c_{55}	0.010	4.301	0.209	0.794
c_{23}	0.015	0.638	0.004	0.403	c_{56}	0.015	0.014	0.001	0.228
c_{24}	0.010	0.004	0.000	0.424	c_{57}	0.010	1.968	0.053	0.220
c_{25}	0.010	0.181	0.006	0.380	c_{58}	0.010	1.964	0.008	0.188
c_{26}	0.010	4.244	0.151	0.459	c_{59}	0.010	0.239	0.016	0.229
c_{27}	0.010	0.449	0.004	0.433	c_{60}	0.010	0.000	0.000	0.232
c_{28}	0.010	0.070	0.001	0.467	c_{61}	0.010	0.033	0.032	0.578
c_{29}	0.010	0.116	0.002	0.412	c_{62}	0.010	0.012	0.001	0.535
c_{30}	0.010	0.008	0.000	0.459	c_{63}	0.010	1.198	0.010	0.368
c_{31}	0.010	0.008	0.005	0.526	c_{64}	0.010	0.632	0.039	0.570
c_{32}	0.010	0.008	0.000	0.506	c_{65}	0.010	1.119	0.010	0.374
c_{33}	1.000	4.102	0.065	0.473	c_{66}	1.000	9.286	0.475	0.796

Table 3. Mean and standard deviation of average errors of results presented in Table 1 and Table 2 and corresponding CPU used in inference

	Mean of average errors	Standard deviation of average errors	CPU
Including structurally dependent proteins	60.848	144.652	404.74
Excluding tructurally dependent proteins	90.572	185.176	549.38

in place of *n* species for the new plan and (*n*-|*s*|) subsrates for the previous scheme. On the other hand, the inclusion of dependent species in hazards enables to produce more sensible drift and diffusion terms in the updates of rates and missing states. Thus, as understood from the findings, these highlighted improvements in hazards cause less number of singularities in the system, hereby accelerates the speed of computations by the new plan.

As a result, we consider that our innovated algorithm is more advantageous in the inference of complex systems in terms of the accuracy of estimates. However, it cannot be seen as a better algorithm regarding to the computational cost in the estimation of every complex structure, rather it can be evaluated as a computationally efficient method for the network having an inner dependence like the MAPK pathway.

5. CONCLUSION AND DISCUSSION

We have presented a new MCMC scheme which includes structurally dependent substrates in the estimation of reaction rates of a complex biochemical system. In the inference, we have implemented Bayesian methods based on the Euler approximation and data augmentation techniques due to the fact that the former is computationally more efficient than the exact algorithm and the latter can decrease the bias on estimates caused by the discretization of the diffusion approximation via the Euler method.

In our new algorithm, we have generated candidate values of structurally dependent substrates by using their linear relationships with linearly independent proteins. To capture the underlying linear links, we have investigated the singularity of the net effect matrix.

In our system, since we have observed an additional structural dependency within linearly dependent substrates, we have used all proteins in the calculation of hazard functions which are the base components of drifts and diffusion terms. Whereas all acceptance probabilities α have been computed solely by linearly independent terms, because of the fact that the highlighted inner-structural dependence within dependent substrates has led to infeasible likelihoods in α. However, this is a particular problem in the MAPK description. Therefore, we suggest that indeed thanks to this new algorithm, the calculation of α can be easily implemented by dividing it into two parts. In the first part of the calculation, we can compute the likelihood of linearly independent proteins and in the second part, we can only consider the likelihood found by linearly dependent proteins. Then we can multiply these two terms since the application of our new plan enables to factorize the likelihood. This process can be performed for both the update of reaction rates and missing states.

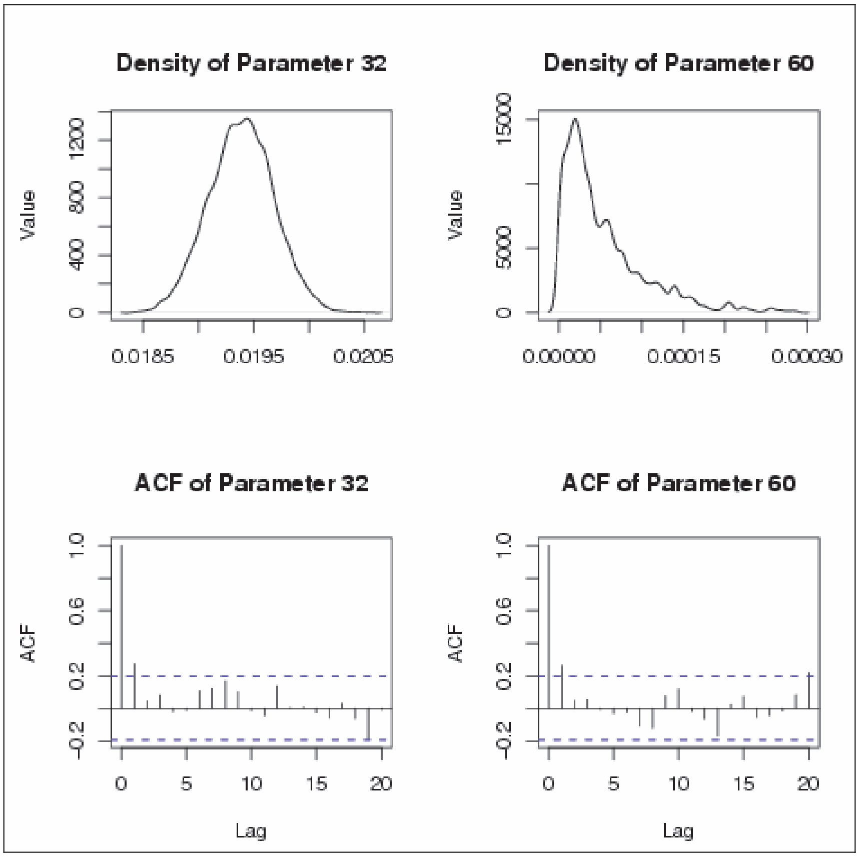

Figure 2. Posterior distributions and autocorrelation functions (ACF) of reaction rate constants 32 and 60 after burn-in via the MCMC plan which includes structurally dependent substrates

As an extension of our study, seeing that we have investigated an inner dependence within structurally dependent species, we propose to develop a sub-algorithm for merely linearly dependent substrates. In that plan, our new scheme can be repeated within these terms iteratively until each linearly dependent protein can be generated in terms of its associated linearly independent species within that particular group. Under this condition the complete likelihood is factorized as a number of independent parts, and thereby can be computed as the product of underlying independent pieces of information. We consider that such an iterative calculation can further improve the accuracy of estimates even though it can also increase the computational cost of the inference. However, we believe that this additional computational demand can considerably decline if the codes are executed on an efficient programme language.

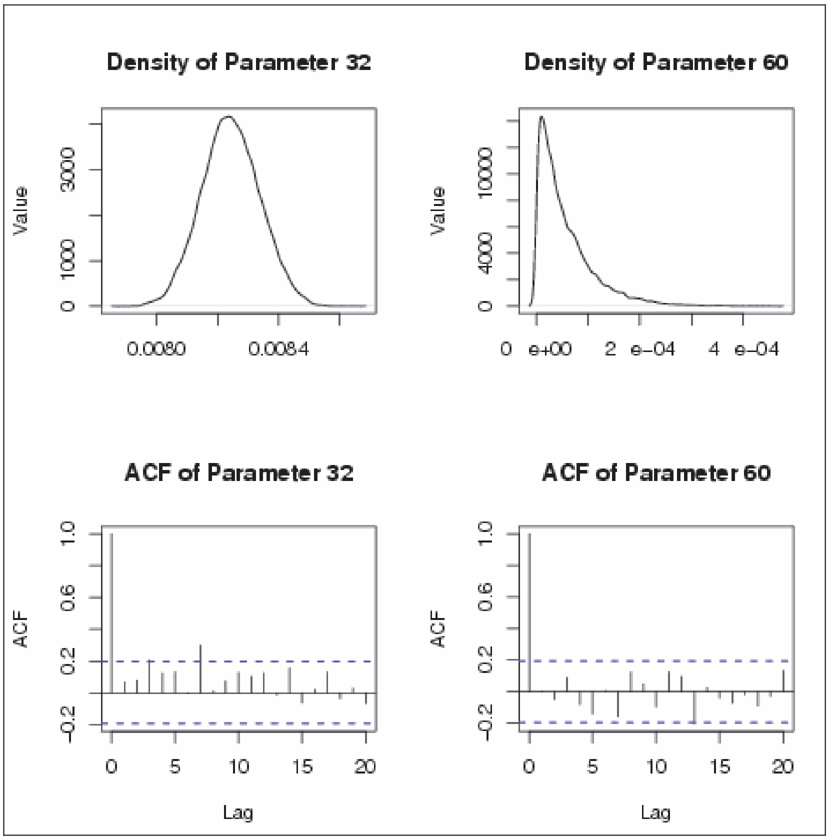

Figure 3. Posterior distributions and autocorrelation functions (ACF) of reaction rate constants 32 and 60 after burn-in via the MCMC plan which excludes structurally dependent substrates

6. REFERENCES

Bower, J.M., and Bolouri, H., 2001. Computational modelling of genetic and biochemical networks (Second edition). Massachusetts Institute of Technology. Cambridge. Massachusetts.

Boys, R.J., Wilkinson, D.J., and Kirkwood, T.B.L., 2008. Bayesian inference for a discretely observed stochastic kinetic model. Statistical Computing, 18, 125-135.

Elerian, B.O., Chib, S., and Shephard, N., 2001. Likelihood inference for discretely observed nonlinear diffusions. Econometrica, 69 (4), 959-993.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. Journal of Business and Economic Statistics, 19 (2), 177–191.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., 2004. Bayesian data analysis. Chapman and Hall/CRC. Florida. U.S.A.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Bayesian Statistics 4 in Bernando, J.M., Berger, J.O., Dawid, A. P., and Smith, A.F.M. (Eds), 169-193. Oxford University Press. Oxford.

Gibson, M.A., and Bruck, J., 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. Journal of Physical Chemistry, A(104), 1876–1889.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry, 81 (25), 2340–2361.

Gillespie, D.T., 1992. A rigorous derivation of the chemical master equation. Physica, A 188, 404–425.

Golightly, A., and Wilkinson, D.J., 2005. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics, 61 (3), 781–788.

Golightly, A., and Wilkinson, D.J., 2006a. Bayesian sequential inference for nonlinear multivariate diffusions. Statistics and Computing, 16, 323-338.

Golightly, A., and Wilkinson, D.J., 2006b. Bayesian sequential inference for stochastic kinetic biochemical network models. Journal of Computational Biology, 13 (3), 838-851.

Kolch, W., Calder, M., and Gilbert, D., 2005. When kinases meet mathematics: the systems biology of MAPK signaling. FEBS Letters, 579, 1891–1895.

Orton, R., Sturm, O.E., Vyshemirsky, V., Calder, M., Gilbert, D.R., and Kolch, W., 2005. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. Biochemical Journal, 392, 249–261.

Purutçuoğlu, V., and Wit, E., 2006. Exact and approximate stochastic simulations of the MAPK pathway and comparisons of simulations’ results. Journal of Integrative Bioinformatics, 3, 231-243.

Purutçuoğlu, V., and Wit, E., 2008a. Inclusion of convoluted measurements in Bayesian inference of the MAPK/ERK pathway via multivariate diffusion model. Proceeding of the Third International Symposium on Health, Informatics and Bioinformatics in Sezerman, U. (Ed), Sabancı University, İstanbul, Turkey, CD-Rom.

Purutçuoğlu, V., and Wit, E., 2008b. Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis, 3 (4), 851-86.

Roberts, G.O., Gelman, A., and Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. The Annals of Applied Probability, 77 (1), 110-120.

Roberts, G.O., and Rosenthal, J.S., 1998. Optimal scaling of discrete approximations to Langevin diffusions. Journal of Royal Statistical Society, Series B, 60 (1), 255-268.

Roberts, G.O., and Stramer, O., 2001. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. Biometrika, 88 (3), 603–621.

Turner, T.E., Schnell, S., and Burrage, K., 2004. Stochastic approaches for modelling in vivo reactions. Computational Biology and Chemistry, 28, 165–178.

Wilkinson, D.J., 2006. Stochastic modelling for systems biology. Chapman and Hall/CRC. Florida. U.S.A.

YAPISAL BAĞIMLILIK ALTINDA KARMAŞIK MAPK YOLUNUN BAYESCİ TAHMİNİ

ÖZET

MAPK yolu, tüm ökaryotlarda bulunan hücresel büyüme kontrolünü düzenleyen başlıca sinyal iletim sistemlerinden biridir. Hayati görevinden dolayı sistemin idaresi çok sayıda protein vasıtasıyla yürütülür, buna bağlı olarak karmaşık bir yapı oluşturur. Çalışmada, Euler yaklaşımına dayalı MCMC teknikleriyle bu sistemin tahmininde diğer proteinlerle yüksek yapısal bağımlılıklar gösteren bir çok proteinin varolduğu gözlenmiştir. Bu proteinler kabul edilme olasılıklarını imkansız yapan tekil difüzyon/varyans matrislerine neden olmuşlardır. Bu nedenle bu sorunlu proteinler tahmin hesabının başında çıkarılmış ve parametreler sadece sistemdeki doğrusal bağımsız türler kullanarak tahmin edilmiştir. Ancak bu durumda da özellikle bağımlı türlerin sayısı arttıkça, tahminin doğruluğu bahsedilen eliminasyondan oldukça etkilenmektedir. Bu proteinlerin elenmesi MCMC'deki mevcut kayıp terim sayısının belirgin derecede artmasına neden olmaktadır. Bu çalışmada dolaylı yoldan bu proteinler, bağımlı terimlerin bağımsız türlerin doğrusal kombinasyonu şeklinde simülasyon eden alternatif bir yaklaşımla hesaplamaların içine katılmaktadır. Bu şekilde reaksiyon oranlarının ve durumlarının kabul edilme olasılıklarını hesaplamada bağımlı türlerin etkileri ilave edilebilmektedir. Analizlerden, bahsedilen yeniliğin tahminlerin ortalama hatalarını azalttığı ve MAPK yolunun tahmininde daha az hesaplama maliyeti önerdiği sonucuna varılmıştır.

Anahtar Kelimeler: Bayesci tahmin, Difüzyon yaklaşımı, MAPK yolu.