INFERENCE OF GENE REGULATORY NETWORKS USING GENETIC PROGRAMMING AND KALMAN FILTER

Haixin Wang and Lijun Qian

Department of Electrical Engineering
Prairie View A&M University
Prairie View, Texas 77446
Email: HWang, LiQian@pvamu.edu

Edward Dougherty

Department of Electrical Engineering
Texas A&M University
College Station, TX 77843
Email: edward@ee.tamu.edu

ABSTRACT

In this paper, gene regulatory networks are infered through evolutionary modeling and time-series microarray measurements. A nonlinear differential equation model is adopted and an iterative algorithm is proposed to identify the model, where genetic programming is applied to identify the structure of the model and Kalman filtering is employed to estimate the parameters in each iteration. Simulation results using synthetic data and microarray measurements show the effectiveness of the proposed scheme.

I. INTRODUCTION

DNA microarray and gene chips have allowed biologists to analyze the genetic behaviors among different genes during the gene expression. After image processing is applied to the DNA microarray photos, it becomes possible to discover gene regulatory networks (GRNs) which are complex and nonlinear in nature. This provides great chances and challenges to model GRNs and help discover new drugs.

In general, modeling GRNs is a nonlinear identification problem. Assume that there are N genes of interest, define x_i as the state (such as the gene expression level) of the i^{th} gene, then the dynamics/interactions of the GRN may be modeled as

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \cdots, x_N) \tag{1}$$

where the nonlinear functions f_i need to be determined from time-series microarray measurements. In this paper, assuming that the statistics (correlation matrices) of the system and measurement noise can be obtained, a two-step procedure is proposed to identify f_i. Firstly, Genetic programming (GP) is applied to determine the nonlinear terms; then the corresponding parameters associated with each term are estimated by Kalman filtering.

Compared to discrete models, the proposed approach is computationally more intensive and can require more data. Moreover, whereas some discrete approaches allow inference from steady-state data, the proposed approach requires time-series data. Nevertheless, it has two major advantages. Continuous rather than logical variables allow a more accurate representation of the GRN. The range of nonlinear behaviors exhibited by GRNs can be more thoroughly understood with nonlinear differential equations to model reaction kinetics. In addition, well established dynamical systems theory is available to characterize the dynamics produced by these models.

II. ALGORITHM DESCRIPTION

II-A. Genetic programming

Genetic programming [1] is a type of evolutionary algorithms. All evolutionary algorithms work with a population of individuals, where each individual may be a solution of the optimization problem. GP operates on a tree structure. A tree structure is flexible enough to allow one to represent relationships efficiently. The leaves of a tree represent variables or constants, while the other nodes implement operators. Mutation and crossover operations may be performed to generate offsprings. Selection of better performing individuals (with smaller fitness value) ensures that the population evolves toward better performing individuals, thus solving the optimization problem.

II-B. Kalman filter

The development of efficient linear estimation algorithms, e.g., the Kalman filter, has been based mainly on the minimization of the L_2 -norm of the corresponding estimation error. This type of estimation assumes that the message generating process has a known dynamics and that the exogenous inputs have known statistical properties. The well-known Kalman filter offers optimal filtering when the system model parameters and power spectral density of the noise are known [2].

The implementation of the Kalman filter is given by the following equations

$$G(n) = K(n, n-1)u(n)[u^t K(n, n-1)u(n) + Q_2]^{-1} \tag{2}$$

$$e(n) = d(n) - u^t(n)\hat{w}(n-1) \tag{3}$$

$$\hat{w}(n) = \hat{w}(n-1) + G(n)e(n) \tag{4}$$

$$K(n+1, n) = K(n, n-1) - G(n)u^t(n)K(n, n-1) + Q_1 \tag{5}$$

where $G(n)$ is the Kalman filter gain, K is the correlation matrix of the error, $e(n)$ is the vector of estimation error, $d(n)$ is the output vector, $u(n)$ is the input vector calculated from the nonlinear terms evaluated at time step n, $\hat{w}(n)$ is the vector of estimated parameters. Q_1 and Q_2 are the correlation matrices of the system and measurement noise, respectively. Note that although the GRN model itself is nonlinear, the parameters estimation problem is *linear* in each iteration.

II-C. Proposed iterative algorithm

The task of identifying GRNs may be considered as an optimization problem. The goal is to minimize the identification error and keep the model as simple as possible, which may be achieved by minimizing the following fitness function

$$fitness = \sum_{i=1}^{N}[\sum_{k=1}^{M}(x_i(k) - x_i^{tar}(k))^2 + C_i] \tag{6}$$

where M is the number of data points, x_i^{tar} is the target time series and x_i is the obtained time series given by the obtained differential equation represented by a GP individual. C_i is a penalty term. Since it is a global nonlinear optimization problem, a nested optimization structure is adopted, where GP is applied to determine the nonlinear terms (global optimization) while Kalman filter is employed to estimate the corresponding parameters for each term (local optimization) in each iteration. Such a decomposition of

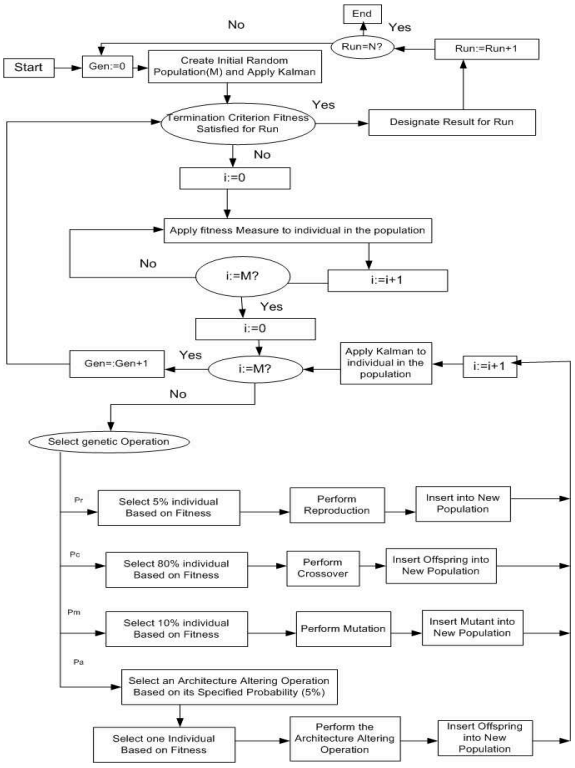

Fig. 1. The genetic programming process with Kalman filter

the problem into a structural part solved by GP and a parameter optimization part solved by Kalman filtering reduces the complexity significantly and speeds up convergence. The detailed procedures of the proposed iterative algorithm is illustrated in Fig. 1. The GP process has four operations: reproduction, crossover, mutation and selection. Kalman filtering is employed to estimate the parameters for every generation.

III. SIMULATION EVALUATION

III-A. Synthetic data

In this part of the simulation, we use data of a metabolic network (a part of the biological phospholipid pathway) that consists of three substances and compare our algorithm with the approach in [3], where GP and recursive least-square (RLS) estimation were used without considering noise. The results of the concentration levels of the three substances are shown in Fig. 2. We observe that under noisy conditions, GP plus Kalman filter performs well and Kalman filtering is a better choice than RLS algorithm.

III-B. Yeast data

We consider time-series gene-expression data corresponding to yeast protein synthesis. Here, the data for three genes (HAP1, CYC7 and CYB2) are picked because the relations among them have been revealed by biological experiments. HAP1 represses the nuclear encoding cytochrome gene CYC7 under the anaerobic condition; CYB2 activates CYC7; HAP1 is a repressor and it represses other genes. The following model is obtained by the proposed algorithm using the time-series data from [4]

$$\begin{aligned} \dot{x_1} &= -0.329x_1 - 0.236x_2 \\ \dot{x_2} &= 0.1057x_2 \\ \dot{x_3} &= 0.1263x_2 - 0.619x_3 \end{aligned} \tag{7}$$

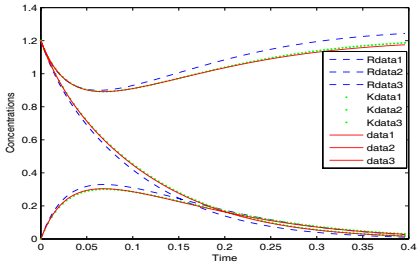

Fig. 2. E-Cell simulation by RLS and Kalman filtering. "data" is the original data without noise. "Rdata" and "Kdata" are from the obtained model by GP+RLS and GP+Kalman filtering, respectively.

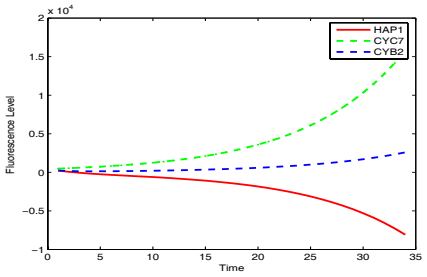

Fig. 3. The dynamics of expression level (in fluorescence level of that gene's mRNA) of CYB2, HAP1 and CYC7

The fluorescence values, which are linearly proportional to the amount of mRNA for a particular gene, are shown in Fig. 3. We observe that HAP1 represses CYC7, and CYB2 activates CYC7. HAP1 behaves as a repressor. These observations are in agreement with the biological experiment findings.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, a joint genetic programming and Kalman filtering approach is proposed to infer gene regulatory networks from time-series data. Simulations with synthetic and yeast data demonstrate the effectiveness of the proposed algorithm. In general, the statistics of the noise in the microarray measurements are not known. Thus, the Kalman filter may not be appropriate for estimating parameters. Instead, a H_∞ filter may be employed to provide robust estimation of parameters even without the knowledge of the noise statistics. This will be one of our future efforts.

V. REFERENCES

[1] J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT press, 1992.

[2] M. Grewal and A. Andrews, *Kalman Filtering: Theory and Practice*, Prentice Hall, 1993.

[3] S. Ando, E. Sakamoto and H. Iba, "Evolutionary Modeling and Inference of Gene Network", *Information Science*, Vol.145, pp.237-259, 2002.

[4] http://sgdlite.princeton.edu/download/yeast_datasets