BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 8 2011, pages 1172–1173
doi:10.1093/bioinformatics/btr078

Systems biology

Advance Access publication February 10, 2011

SiGN-SSM: open source parallel software for estimating gene networks with state space models

Yoshinori Tamada[1,*], Rui Yamaguchi[2], Seiya Imoto[1], Osamu Hirose[3], Ryo Yoshida[4], Masao Nagasaki[5] and Satoru Miyano[1,2,6]

[1]Laboratory of DNA Information Analysis, [2]Laboratory of Sequence Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, [3]Bioinformation Engineering Laboratory, Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, [4]Department of Statistical Modeling, Institute of Statistical Mathematics, Research Organization of Information and Systems, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, [5]Laboratory of Functional Genomics, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and [6]Data Analysis Fusion Team, RIKEN Computational Science Research Program, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Associate Editor: Martin Bishop

ABSTRACT

Summary: SiGN-SSM is an open-source gene network estimation software able to run in parallel on PCs and massively parallel supercomputers. The software estimates a state space model (SSM), that is a statistical dynamic model suitable for analyzing short time and/or replicated time series gene expression profiles. SiGN-SSM implements a novel parameter constraint effective to stabilize the estimated models. Also, by using a supercomputer, it is able to determine the gene network structure by a statistical permutation test in a practical time. SiGN-SSM is applicable not only to analyzing temporal regulatory dependencies between genes, but also to extracting the differentially regulated genes from time series expression profiles.

Availability: SiGN-SSM is distributed under GNU Affero General Public Licence (GNU AGPL) version 3 and can be downloaded at http://sign.hgc.jp/signssm/. The pre-compiled binaries for some architectures are available in addition to the source code. The pre-installed binaries are also available on the Human Genome Center supercomputer system. The online manual and the supplementary information of SiGN-SSM is available on our web site.

Contact: tamada@ims.u-tokyo.ac.jp

Received on December 1, 2010; revised on January 21, 2011; accepted on February 6, 2011

1 INTRODUCTION

Analyzing the dynamical regulatory mechanisms of gene expressions in a cellular system is a challenging problem in systems biology. To this end, many computational methods have been proposed to estimate dynamical systems of regulatory dependencies between gene expressions from temporal gene expression profiles. The major difficulty of these studies comes from insufficient data time points as opposed to the number of variables (genes) in a computational model. A state space model (SSM) (Hirose *et al*., 2008; Kitagawa and Gersch, 1996; West and Harrison, 1997) is a statistical model that is applicable to small time-point temporal datasets because it can reduce the number of parameters to be estimated. The SSM decomposes the temporal gene expressions into a dynamical system of modules called the *system model* (or *state space*) and a mapping from the modules to the particular genes called the *observation model*. There are a number of gene network studies using the SSM (Beal *et al*., 2005; Hirose *et al*., 2008; Rangel *et al*., 2004).

SiGN-SSM is a re-implemented, new version of the previously released one called TRANS-MNET (Hirose *et al*., 2008). In addition to TRANS-MNET, SiGN-SSM has the following improvements: (i) it implements a novel constraint on the model parameters effective to stabilize the estimated models for the short time series data with irregular time intervals; (ii) runs in parallel as a multithreaded program exploiting multicore CPUs, as a bulk (array) job through a job dispatching system such as Sun (Oracle) Grid Engine (SGE) on PC cluster systems, and a multi-process MPI (Message Passing Interface) application on massively parallel supercomputers; (iii) is an open-source software so that everyone can freely access the source code and improve, modify and distribute it, and (iv) implements a statistical permutation test to determine a gene network structure as proposed in Hirose *et al*. (2008) and differentially regulated gene extraction presented in Yamaguchi *et al*. (2008) on the Human Genome Center (HGC) supercomputer system.

2 STATE SPACE MODEL

We introduce the SSM briefly. For detailed definitions, see Hirose *et al*. (2008). Let y_n be a vector of p variables representing gene expression profiles for p genes observed at time $n \leq N$, where N is the total number of time points. In the SSM, the observable expression profiles y_n are assumed to be generated from the k-dimensional hidden state variables (vector) x_n. Here, we assume that $k \ll p$. The SSM is defined by the following two formulae:

$$x_n = Fx_{n-1} + v_n, \quad v_n \sim N(0, Q), \quad \text{[System model]}$$

$$y_n = Hx_n + w_n, \quad w_n \sim N(0, R), \quad \text{[Observation model]}$$

*To whom correspondence should be addressed.

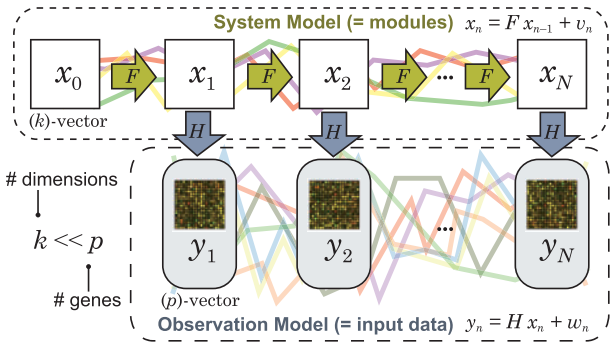

Fig. 1. Conceptual view of the state space model.

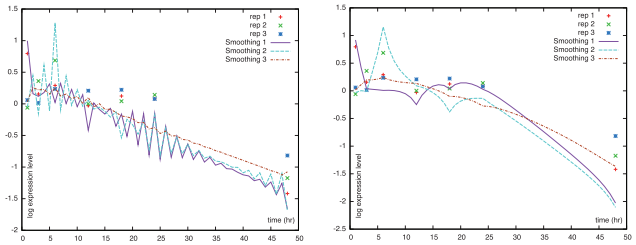

Fig. 2. Comparison without (left) and with (right) the proposed constraint on F. The plotted lines are the estimated smoothing observation variables for the very short, triplicate sample data.

where F is the (k,k)-state transition matrix and H the (p,k)-observation matrix that maps from the state variables x_n to the observation variables y_n. The two vectors v_n and w_n are the system and the observation noise, respectively, where Q and R are the covariance matrices of the normal distributions. The initial state variables x_0 is required to be estimated from the data and we assume $x_0 \sim N(\mu_0, \Sigma_0)$. The SSM estimation problem is to estimate the unknown parameters $\{H, F, Q, R, \mu_0\}$ from the observed temporal gene expression data. The dimensions of the state vector (k) is also an unknown parameter to be determined. The parameter estimation from the observed data is realized by the expectation-maximization (EM) algorithm. The dimensions of the state vector, k, can be determined by comparing the Bayesian Information Criterion values of the estimated models with different k. Figure 1 shows the conceptual view of the SSM.

3 IMPLEMENTATION AND PARALLELIZATION

SiGN-SSM is written in C, using the BLAS/LAPACK library. Since the EM algorithm finds only locally optimal parameters, it is required to run the algorithm many times to obtain better estimate for a single k. To speed up the parameter estimation and the optimal k determination, SiGN-SSM supports multiprocess parallelization using MPI, multithread parallelization using OpenMP and parallelization by bulk jobs on PC clusters. By using MPI, we confirmed that SiGN-SSM can optimize multiple k in parallel very efficiently with up to 256 CPU cores. See Supplementary information on our web site for detailed results. The permutation test determines the gene network structure from the estimated model parameters. However, it requires much more computational time than the parameter estimation. To solve this problem, we parallelized it on the HGC supercomputer using SGE.

4 NEW CONSTRAINT ON PARAMETERS

When the algorithm estimates the model parameters from short time series data measured for irregular time intervals, they often oscillate undesirably (Fig. 2). To suppress such spurious patterns, we propose a novel constraint on the system transition matrix F along with the smoothness prior approach (Kitagawa and Gersch, 1996). With the constraint, we assume that the value of the state vector at time n is similar to that at time $n-1$. The constrained version of F, denoted by $\tilde{F} = \{\tilde{f}_{ij}\}$, has its diagonal elements $\tilde{f}_{ii} = g$ for $i = 1, \ldots, k$ where $0 \leq g \leq 1$ is a constant to control the smoothness, which user can choose (we set 0.8 as the default value). The off-diagonal components of $\tilde{F}$ are estimated by the EM algorithm with the constraint of the diagonal components, in which we utilize the general framework of Wu *et al.* (1996) to constrain parameters of an SSM in the EM algorithm. The value of g can be determined by such as comparison of BIC values, cross-validation, etc.

5 CONCLUSION

SiGN-SSM is a highly-scalable, open-source implementation of an SSM estimation algorithm. The newly proposed constraint on the parameters can significantly stabilize the estimation of the parameters for the very short time point temporal data with irregular time intervals. Users can analyze their time series expression datasets using SiGN-SSM, and the estimated model can be applied to other data to extract differentially expressed genes.

ACKNOWLEDGEMENT

Computational resources required for the development of SiGN-SSM was provided by the HGC Supercomputer System, Human Genome Center, Institute of Medical Science, The University of Tokyo and RIKEN Supercomputer system RICC.

Funding: ISLiM (Next-generation integrated simulation of living matter) project in RIKEN Computational Science Research Program.

Conflict of Interest: none declared.

REFERENCES

Beal,M.J. *et al.* (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **21**, 349–356.

Hirose,O. *et al.* (2008) Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, **24**, 932–942.

Kitagawa,G. and Gersch,W. (1996) *Smoothness Priors Analysis of Time Series*, Springer, New York.

Rangel,C. *et al.* (2004) Modelling T-cell activation using gene expression profiling and state space models. *Bioinformatics*, **20**, 1361–1372.

West,M. and Harrison,J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn., Springer, New York.

Wu,L.S.-Y. *et al.* (1996) An algorithm for estimating parameters of state-space models. *Stat. Probab. Lett.*, **28**, 99–106.

Yamaguchi,R. *et al.* (2008) Predicting differences in gene regulatory systems by state space models. *Genome Inform.*, **21**, 101–113.